# การค้นหาจุดสนใจโดยใช้ข้อมูลเชิงพื้นที่จากรถแท็กซี่ด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่นที่กำหนดพารามิเตอร์แบบอัตโนมัติ

นางสาวอุไรวรรณ์ อังคะเวย์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

# DETECTING POINT OF INTEREST FROM TAXI GPS DATA USING DBSCAN WITH AUTOMATIC PARAMETER CONFIGURATION

Miss Uriwan Angkhawey



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การค้นหาจุดสนใจโดยใช้ข้อมูลเชิงพื้นที่จากรถแท็กซี่ด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่นที่กำหนดพารามิเตอร์แบบอัตโนมัติ |
| โดย | นางสาวอุไรวรรณ์ อังคะเวย์ |
| สาขาวิชา | วิทยาศาสตร์คอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | ผู้ช่วยศาสตราจารย์ ดร.วีระ เหมืองสิน |

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์

(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(อาจารย์ ดร.ดวงดาว วิชาดากุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.วีระ เหมืองสิน)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร.กาญจนา ศิลาวราเวทย์)

อุไรวรรณ์ อังคะเวย์ : การค้นหาจุดสนใจโดยใช้ข้อมูลเชิงพื้นที่จากรถแท็กซี่ด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่นที่กำหนดพารามิเตอร์แบบอัตโนมัติ (DETECTING POINT OF INTEREST FROM TAXI GPS DATA USING DBSCAN WITH AUTOMATIC PARAMETER CONFIGURATION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.วีระ เหมืองสิน, 59 หน้า.

ความเจริญของเมืองในปัจจุบันทำให้เกิดสถานที่ต่างๆ มากมาย ไม่ว่าจะเป็นห้างสรรพสินค้า สถานที่ท่องเที่ยว โรงแรม ร้านค้า ร้านอาหาร อาคารสำนักงาน หรือแม้แต่ลานกิจกรรมต่างๆ ซึ่งล้วนแล้วแต่เป็นจุดที่ผู้คนมักไปรวมตัว สะท้อนให้เห็นถึงพฤติกรรมของประชากร ข้อมูลเหล่านี้เป็นประโยชน์ในหลายด้านเช่น การวางแผนผังเมือง การวางแผนการจราจร การสำรวจโรคระบาด หรือวิเคราะห์การเกิดอาชญากรรมต่างๆ การสำรวจเพื่อให้ได้มาซึ่งข้อมูลเหล่านี้จำเป็นต้องอาศัยสิ่งที่สะท้อนให้เห็นถึงตำแหน่งการเคลื่อนที่ของประชากร เช่น ข้อมูลจีพีเอสจากโทรศัพท์มือถือ ข้อมูลจากอุปกรณ์จีพีเอสที่ติดตามยานพาหนะ ข้อมูลจากการใช้บริการขนส่งสาธารณะของประชากรในเมืองใหญ่ รถแท็กซี่เป็นการขนส่งสาธารณะอีกประเภทหนึ่งที่ให้บริการอย่างกว้างขวางในเขตกรุงเทพมหานคร ด้วยลักษณะของการเคลื่อนที่ไปยังบริเวณต่างๆ ทุกพื้นที่ พบว่า ข้อมูลจีพีเอสจากการรับส่งผู้โดยสารของรถแท็กซี่ สามารถนำมาใช้เพื่อเป็นข้อมูลในการค้นหาบริเวณที่เป็นพื้นที่ที่น่าสนใจ

งานวิจัยนี้ นำเสนอแนวทางในการค้นพบพื้นที่จุดสนใจโดยใช้เทคนิคการจัดกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น (DBSCAN) ในกรณีของชุดข้อมูลแท็กซี่การกำหนดพารามิเตอร์เป็นเรื่องยากเพราะข้อมูลรับ - ส่ง ผู้โดยสารกระจายไปในพื้นที่ต่างๆ ที่แตกต่างกัน ดังนั้นงานวิจัยจึงพัฒนาวิธีการกำหนดพารามิเตอร์ที่เหมาะสมแบบอัตโนมัติเพื่อใช้ในการทำงานของอัลกอริทึมโดยการพิจารณาจากปริมาณและความหนาแน่นของข้อมูลในแต่ละพื้นที่

ผลการทดลองกับพื้นที่ตัวอย่างทำให้พบว่า สามารถค้นพบสถานที่ที่เป็นจุดสนใจ รวมถึงบริเวณพื้นที่ที่ประกอบด้วยจุดสนใจหลายๆจุดรวมกัน ดังนั้นวิธีการจากงานวิจัยนี้จึงเหมาะสำหรับการนำมาใช้เพื่อค้นหาจุดสนใจในกรุงเทพมหานคร แนวทางการศึกษาในอนาคต เราจะทดลองใช้วิธีนี้กับพื้นที่อื่นเพื่อค้นหาพื้นที่น่าสนใจสำหรับวิเคราะห์ทิศทางของการพัฒนาเมือง

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2560

# # 5970984221 : MAJOR COMPUTER SCIENCE
KEYWORDS: TAXI GPS DATA / DBSCAN / CLUSTERING / POINT OF INTEREST

URIWAN ANGKHAWEY: DETECTING POINT OF INTEREST FROM TAXI GPS DATA USING DBSCAN WITH AUTOMATIC PARAMETER CONFIGURATION. ADVISOR: ASST. PROF. VEERA MUANGSIN, Ph.D., 59 pp.

The current urbanization causes new places in the city such as department store, tourist attraction, hotel, building or a park. They are the place that people lives every day and it reflects the behavior and activity of populations. Analysis of POIs can be useful in many fields, for example, urban planning, traffic planning, epidemic survey or crime analysis. The survey created to conduct these information needs the tool that can reflect the population's moving like the mobile phone GPS, vehicle GPS or information from the public transportation. Taxi is the one that widely provides service in Bangkok. Since this transportation travel to everywhere in town, therefore the information from Taxi GPS can be used to discover the interesting area.

This research presents the method to discover Point of Interest areas by using the clustering technique with the Density-Based Spatial Clustering Algorithm (DBSCAN). The effectiveness of the algorithm depends on the appropriate parameters. In the case of the taxi dataset, determining the parameters is hard because pick-up and drop-off locations are distributed differently in different areas. This research proposes the methods to automatically determine the both necessary parameters for DBSCAN algorithm by considering the density distribution of dataset.

The experimental shows the clustering results with parameters from the proposed method separate small and fine clusters that we are able to identify the interesting area. It is suitable for use to discover the interesting area in Bangkok by using information from Taxi GPS. In the future, we will apply this method to other areas to detect new POIs to analyze the direction of urban development.



Department: Computer Engineering Student's Signature .....
Field of Study: Computer Science Advisor's Signature .....
Academic Year: 2017

# กิตติกรรมประกาศ

ความสำเร็จของวิทยานิพนธ์ฉบับนี้ นอกจากการทำงานของผู้วิจัยแล้ว ยังได้รับความช่วยเหลือจากบุคคลหลายท่านที่ให้การสนับสนุนจนสามารถดำเนินงานให้เกิดผลลุล่วงด้วยดี ผู้จัดทำจึงใคร่ขอกราบขอบพระคุณทุกท่าน

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. วีระ เหมืองสิน ที่กรุณารับเป็นอาจารย์ที่ปรึกษา ให้คำแนะนำ ความช่วยเหลือ รวมทั้งสละเวลาช่วยแก้ปัญหาในทุกขั้นตอนของการดำเนินงานวิจัย

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ทุกท่าน อาจารย์ ดร. ดวงดาว วิชาดากุล และ ดร. กาญจนา ศิลาวราเวทย์ ที่ให้ความกรุณาให้คำแนะนำและแนวคิดเพิ่มเติมในการดำเนินงานวิจัย ซึ่งเป็นประโยชน์อย่างยิ่งต่อการพัฒนางานวิจัยนี้

ขอขอบพระคุณอาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชา อบรมสั่งสอนให้มีความรู้ จนสามารถนำมาประยุกต์ใช้ให้เกิดประโยชน์ในทุกๆ ด้าน

ขอขอบพระคุณบิดา มารดา สมาชิกในครอบครัว และเพื่อนๆ พี่น้องทุกท่าน ที่คอยให้กำลังใจ ให้การสนับสนุนในทุกเรื่อง และเสมอมา

สุดท้ายนี้ ผู้วิจัยหวังเป็นอย่างยิ่งว่า งานวิจัยฉบับนี้จะเป็นประโยชน์แก่ผู้อื่นสืบไป

**สารบัญ**

หน้า

หน้า

หน้า

# สารบัญตาราง

# สารบัญรูปภาพ

# บทที่ 1
# บทนำ

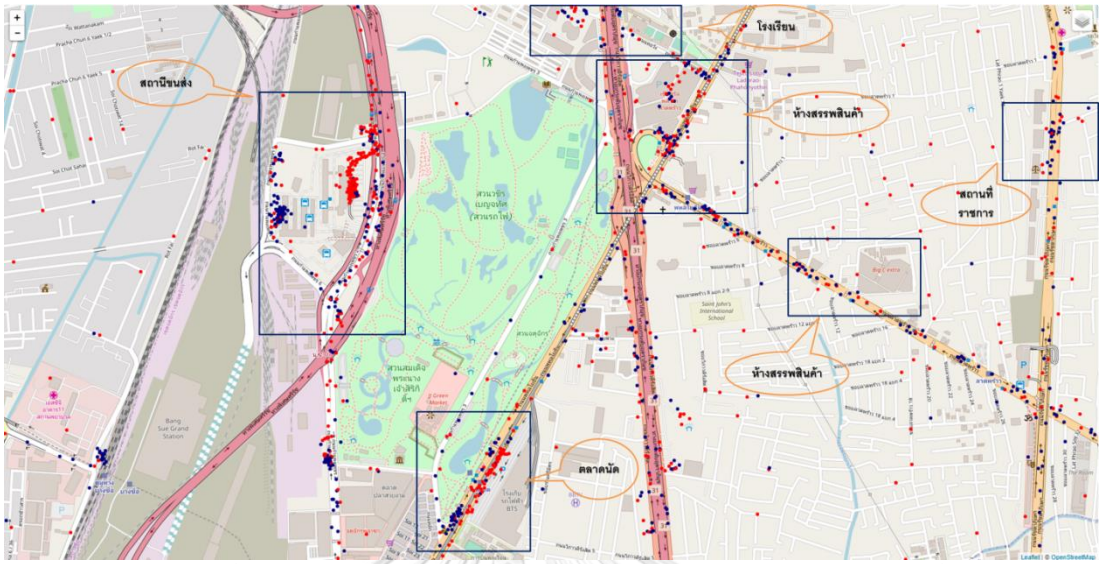
## 1.1 ที่มาและความสำคัญของปัญหา

แท็กซี่ คือการโดยสารสาธารณะที่ให้บริการอย่างแพร่หลายและได้รับความนิยมในเมืองใหญ่ เช่น กรุงเทพมหานคร รถแท็กซี่โดยทั่วไปมีการติดตั้งอุปกรณ์จีพีเอส (Global Positioning System: GPS) ซึ่งเป็นอุปกรณ์ติดตามยานพาหนะแบบเรียลไทม์ (Real-time) ข้อมูลที่ได้จากจีพีเอสสามารถนำมาใช้ในการติดตามการเคลื่อนที่ของยานพาหนะ เนื่องจากประกอบด้วยข้อมูลเชิงพื้นที่ ละติจูด (Latitude) ลองติจูด (Longitude) เส้นทางการเดินทาง ความเร็วในการเคลื่อนที่ รวมถึงการเปลี่ยนแปลงของสถานะแท็กซี่มิเตอร์ จึงเหมาะที่นำจะมาใช้ในการวิจัยงานทางด้านการขนส่ง [1] และศึกษารูปแบบการเดินทางผ่านลักษณะการรับส่งผู้โดยสารทั้งทางด้านเวลา สถานที่ รวมถึงเส้นทางที่ใช้ในการเดินทาง เพื่อค้นหาจุดสนใจ (Point Of Interest) เช่น สถานที่ท่องเที่ยว ห้างสรรพสินค้า โรงแรม ตลาด หรือสถานที่ที่ผู้คนให้ความสนใจและเดินทางไปยังบริเวณดังกล่าว โดยข้อมูลเหล่านี้มีความสัมพันธ์กับพฤติกรรมและเกี่ยวเนื่องกับกิจกรรมของประชากร

เมื่อนำข้อมูลการรับส่งผู้โดยสารของรถแท็กซี่กระจายลงในแผนที่พบลักษณะการกระจายตัวของข้อมูลอย่างน่าสนใจ กล่าวคือ มีการกระจายตัวและเกาะกลุ่มของข้อมูลที่แตกต่างกัน หากขยายดูพื้นที่ จะพบว่าบริเวณนั้นเป็นสถานที่ที่เป็นจุดสนใจ ยกตัวอย่างจากภาพที่ 1 แสดงตัวอย่างของพื้นที่และการกระจายตัวของข้อมูลที่มีความหลากหลาย โดยพบว่าบริเวณที่จุดรับส่งผู้โดยสารเกาะกลุ่มหนาแน่น คือ สถานีขนส่ง ตลาดนัด และห้างสรรพสินค้า เป็นต้น จากลักษณะดังกล่าว แสดงให้เห็นว่าข้อมูลจากรถแท็กซี่สามารถสะท้อนถึงจุดสนใจที่ประชากรมักจะเดินทางไปยังบริเวณนั้นๆ ได้จริง และเป็นข้อมูลที่สามารถนำมาใช้เพื่อจำแนกและค้นหาบริเวณที่เป็นจุดสนใจได้

แต่เนื่องจากข้อมูลที่มีปริมาณมากและบริเวณพื้นที่ที่กว้างใหญ่ การค้นหาด้วยตาเปล่าจึงไม่ใช่วิธีที่สะดวก จำเป็นจะต้องใช้วิธีทางเทคโนโลยีเข้าช่วย ซึ่งวิธีการแบ่งกลุ่มข้อมูล (Clustering Analysis) เหมาะสำหรับนำมาช่วยค้นหาข้อมูลดังที่กล่าวเบื้องต้น วิธีแบ่งกลุ่มข้อมูลมีเทคนิคและการทำงานหลายรูปแบบเช่น การแบ่งแบบตัดเป็นส่วน (Partitioning methods) การแบ่งแบบเป็นลำดับขั้น (Hierarchical methods) การแบ่งแบบตาราง (Grid-based methods) และการแบ่งตามความหนาแน่น (Density-based methods) สำหรับการแบ่งกลุ่มข้อมูลเพื่อค้นหาบริเวณหรือสถานที่ที่เป็นจุดสนใจจะต้องมีความเหมาะสมกับข้อมูลและสะท้อนถึงผลลัพธ์ ดังสังเกตได้จากรูปร่างของการเกาะกลุ่มของข้อมูลตามภาพที่ 1 ลักษณะการกระจายตัวของข้อมูลรับส่งผู้โดยสารของรถแท็กซี่พบว่า

มีลักษณะรูปทรงที่หลากหลาย ขนาดที่แตกต่างกัน และจำนวนของกลุ่มข้อมูลมีการเปลี่ยนแปลงได้เสมอ จึงไม่สามารถที่จะระบุจำนวนได้อย่างตายตัว




ภาพที่ 1 ลักษณะการกระจายตัวของข้อมูลรับส่งผู้โดยสารของรถแท็กซี่

อัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น หรือที่เรียกว่า Density-based spatial clustering of applications with noise (DBSCAN) จึงเป็นอัลกอริทึมที่มีความเหมาะสมเนื่องจากลักษณะการทำงานของอัลกอริทึมสามารถแบ่งกลุ่มข้อมูลได้หลากหลายลักษณะ หลากหลายรูปร่างและขนาดกลุ่มข้อมูลที่ได้มีความเป็นอิสระต่อกัน ไม่จำเป็นต้องกำหนดจำนวนกลุ่มข้อมูลไว้ล่วงหน้า และจัดการกับข้อมูลที่มีความผิดปกติ (Noise)ได้ดี

การทำงานของอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น จำเป็นต้องใช้พารามิเตอร์สองค่าที่มีความสำคัญในการประมวลผล ประกอบด้วย รัศมีระหว่างจุดในกลุ่มข้อมูล (radius epsilon: Eps) และจำนวนจุดขั้นต่ำสุดสำหรับกำหนดจุดศูนย์กลางของกลุ่ม (Minimum Points: MinPts) การกำหนดพารามิเตอร์ให้มีความเหมาะสมเป็นปัญหาอย่างมาก เนื่องจากส่งผลกระทบต่อผลลัพธ์ที่มีประสิทธิภาพ ซึ่งจำนวนจุดขั้นต่ำสุดจะถูกกำหนดโดยผู้เชี่ยวชาญที่มีความเข้าใจในข้อมูล จากนั้น จึงหาค่ารัศมีระหว่างจุดในกลุ่มข้อมูลได้จากการหาระยะห่างระหว่างจุดสองจุดตามจำนวนของจุดขั้นต่ำสุดที่ถูกกำหนดไว้ โดยทั่วไปแล้วจำนวนขั้นจุดต่ำสุดไม่ควรน้อยกว่า 3 [2] เนื่องจากหากจำนวนจุดขั้นต่ำสุดเท่ากับ 1 นั่นหมายถึงทุกๆ จุดจะถูกจัดเป็นกลุ่มๆ และหากจำนวนจุดต่ำสุดเท่ากับ 2 ผลลัพธ์ของการจัดกลุ่มจะมีลักษณะเหมือนการจัดกลุ่มเป็นลำดับขั้นแบบล่างขึ้นบน (bottom-up) แต่หากข้อมูลมีปริมาณมาก การกำหนดพารามิเตอร์ที่ดีที่สุดควรเลือกจากพฤติกรรมและการกระจายของข้อมูล [3] ดังนั้นการกำหนดพารามิเตอร์เพื่อใช้สำหรับแบ่งกลุ่มข้อมูลการรับ-ส่งผู้โดยสารจากรถ

แท็กซี่ควรกำหนดจากปริมาณความหนาแน่นและลักษณะการกระจายตัวของข้อมูลที่แตกต่างกันในแต่ละพื้นที่

งานวิจัยนี้ได้นำเสนอวิธีการกำหนดพารามิเตอร์ที่มีความสำคัญต่อการแบ่งกลุ่มข้อมูลของอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่นหรือ DBSCAN อันได้แก่ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และจำนวนจุดขั้นต่ำสุดสำหรับกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) อย่างอัตโนมัติ โดยพิจารณาจากความหนาแน่นและการกระจายตัวของข้อมูลรับส่งผู้โดยสารจากรถแท็กซี่ในบริเวณพื้นที่นั้นๆ เพื่อสำรวจและค้นหาจุดสนใจในเขตกรุงเทพฯ

## 1.2 วัตถุประสงค์

เพื่อพัฒนาวิธีการกำหนดพารามิเตอร์ที่เหมาะสมแบบอัตโนมัติโดยพิจารณาจากความหนาแน่นและการกระจายตัวของข้อมูลการรับส่งผู้โดยสารจากรถแท็กซี่ในเขตกรุงเทพมหานคร สำหรับใช้ในการแบ่งกลุ่มและค้นหากลุ่มพิกัดหรือบริเวณจุดสนใจด้วยอัลกอทึมการจัดกลุ่มตามความหนาแน่น (Density-based spatial clustering of applications with noise: DBSCAN)

## 1.3 ขอบเขตการดำเนินงาน

งานวิจัยนี้ใช้ข้อมูลการรับส่งผู้โดยสารจากรถแท็กซี่จำนวน 2,375 คันในเขตกรุงเทพมหานคร ระยะเวลา 8 เดือน (กุมภาพันธ์ – กันยายน 2559) โดยเสนอวิธีการเลือกพารามิเตอร์แบบอัตโนมัติเพื่อใช้ร่วมกับอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ในการค้นหาและแบ่งกลุ่มบริเวณที่เป็นจุดสนใจ นอกจากนี้ งานวิจัยทำการเปรียบเทียบผลการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่กำหนดโดยอัตโนมัติกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ดังต่อไปนี้

- DMDBSCAN (A Dynamic Method for Discovery Density Varied Clusters)
- VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise)
- AutoEpsDBSCAN (DBSCAN with Eps Automatic for Large Dataset)

## 1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถกำหนดพารามิเตอร์ที่เหมาะสมแบบอัตโนมัติโดยพิจารณาจากความหนาแน่นและการกระจายตัวของข้อมูลการรับส่งผู้โดยสารจากรถแท็กซี่ในเขตกรุงเทพมหานคร สำหรับใช้ในการ

แบ่งกลุ่มและค้นหากลุ่มพิกัดหรือบริเวณที่น่าสนใจด้วยอัลกอทึมการจัดกลุ่มตามความหนาแน่น และสามารถนำกรอบงานวิจัยนี้ไปประยุกต์ใช้กับเขตพื้นที่อื่น

## 1.5 วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัย สามารถแบ่งออกได้เป็นขั้นตอนดังนี้

1. ศึกษาและรวบรวมข้อมูลการเดินทางด้วยรถแท็กซี่ของประชากรในกรุงเทพมหานคร
2. ศึกษาลักษณะการกระจายตัวของข้อมูลการรับ-ส่งผู้โดยสารของรถแท็กซี่
3. ศึกษางานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลด้วยอัลกอทึมการจัดกลุ่มจากความหนาแน่น
4. ศึกษาเทคนิค วิธีการทำงานของอัลกอริทึมการจำแนกข้อมูล
5. ศึกษาเครื่องมือที่ใช้ในงานวิจัย
6. ออกแบบและทำการทดลอง
7. สร้างแบบจำลองจำแนกข้อมูลจุดสนใจ
8. ทดสอบและประเมินผลความถูกต้อง
9. วิเคราะห์ผลการทดลอง
10. สรุปผลและเรียบเรียงวิทยานิพนธ์
11. สอบและเผยแพร่วิทยานิพนธ์

## 1.6 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของงานวิจัยนี้ได้นำเสนอในการประชุมวิชาการดังนี้

Uriwan Angkhawey and Veera Muangsin, “Detecting Points of Interest in a City from Taxi GPS with Adaptive DBSCAN”, the 2018 Seventh ICT International Student Project Conference (ICT-ISPC), Mahidol University, Nakhon Pathom, Thailand, July 11-13, 2018.

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

## 2.1 ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องกับงานวิจัยนี้ประกอบด้วย จุดสนใจ (Point of Interest: POI) ระบบกำหนดตำแหน่งบนโลก (Global Positioning System: GPS) ฟังก์ชั่น Haversine การสร้างฮิสโตแกรม (Histogram) และฟังก์ชั่น Sturge การแบ่งกลุ่มข้อมูล (Clustering Analysis) และการแบ่งกลุ่มข้อมูลตามความหนาแน่น (Density-Based Spatial Clustering of Applications with Noise: DBSCAN)

2.1.1 จุดสนใจ (Point of Interest: POI)

จุดสนใจ [4] คือ สถานที่ บริเวณหรือ จุดที่บุคคลให้ความสนใจ ตัวอย่างเช่น ห้างสรรพสินค้า ร้านอาหาร โรงแรม หรือสถานที่ท่องเที่ยวต่างๆ มักใช้อย่างแพร่หลายเพื่อประโยชน์ในการทำแผนที่หรือระบบนำทางจีพีเอส(GPS) เนื่องจากข้อมูลประกอบด้วยละติจูด ลองติจูด เพื่อระบุสถานที่และเส้นทางที่ผู้คนให้ความสนใจ นอกจากนี้ยังมีการนิยามเพิ่มเติมเกี่ยวกับขอบเขตของความสนใจ (Region Of Interest: ROI) และปริมาณของความสนใจ (Volume of Interest: VOI) ซึ่งประกอบไปด้วยจุดสนใจของแต่ละคนที่มีความแตกต่างกัน

2.1.2 ระบบกำหนดตำแหน่งบนโลก (Global Positioning System: GPS)

หรือรู้จักในชื่อ นาฟสตาร์ (Navstar) [5] คือระบบดาวเทียมนำร่องโลก (Global Navigation Satellite System : GNSS) เพื่อระบุข้อมูลของตำแหน่งและเวลาโดยอาศัยการคำนวณจากความถี่สัญญาณนาฬิกาที่ส่งมาจากตำแหน่งของดาวเทียมต่างๆ ที่โคจรอยู่รอบโลกทำให้สามารถระบุตำแหน่ง ณ จุดที่สามารถรับสัญญาณได้ทั่วโลกและในทุกสภาพอากาศ รวมถึงสามารถคำนวณความเร็วและทิศทางเพื่อนำมาใช้ร่วมกับแผนที่ในการนำทางได้



2.1.3 ฟังก์ชั่น Haversine

เป็นสมการในการหาระยะทางที่สั้นที่สุดระหว่างพิกัดสองพิกัดบนพื้นผิวโลก โดยใช้ละติจูดและลองติจูดคำนวณหาระยะห่าง และเนื่องจากโลกไม่เป็นทรงกลมหรือวงรีที่สมบูรณ์แบบ ผลลัพทธ์ที่ได้จึงเป็นระยะทางโดยประมาณจากระยะทางจริง สูตร Haversine เป็นสูตรการหา cosine ของวงกลม แต่สามารถใช้ประโยชน์สำหรับมุมเล็กๆ และระยะทางได้ดี สมการของ Haversine เป็นดังนี้ [6]

$$\boldsymbol{Haversin\,(\theta) = sin^2\,\frac{\theta}{2}} \qquad (1)$$

ดังนั้น การหาระยะทางระหว่างพิกัดสองพิกัดบนพื้นผิวโลกจึงหาได้จาก

$$\boldsymbol{d = 2r\, arcsin\sqrt{sin^2\left(\frac{\theta_2-\theta_1}{2}\right) + cos(\theta_1)cos(\theta_2)sin^2(\frac{\lambda_2-\lambda_1}{2})}} \quad (2)$$

โดย r คือ รัศมีของโลกมีค่า 6,371

$\lambda$ คือ ลองติจูด

$\boldsymbol{\theta}$ คือ ละติจูด

2.1.4 การสร้างฮิสโตแกรม (Histogram) และฟังก์ชั่น Sturge

ฮิสโตแกรม คือกราฟที่สร้างขึ้นเพื่อใช้ศึกษาการกระจายของข้อมูล การสร้างฮิสโตแกรม ต้องแบ่งข้อมูลออกเป็นช่วงๆ เรียกว่า อันตรภาคชั้น (bin) จากนั้นนับจำนวนค่าที่ตกอยู่ในแต่ละช่วง แต่ละอันตรภาคชั้นระบุเป็นช่วงข้อมูลที่ไม่ทับซ้อนกัน ต้องอยู่ติดกันและควรมีช่วงข้อมูลเท่ากัน ฮิสโตแกรมใช้เพื่อประมาณความหนาแน่นของข้อมูล หรือประมาณความน่าจะเป็นของตัวแปร โดยทั่วไปจะสะท้อนการกระจายตัวของตัวแปรได้อย่างแม่นยำ

การกำหนดอันตรภาคชั้น (bin) เพื่อสร้างฮิสโตแกรม มีหลายฟังก์ชั่นที่ใช้เช่น [7] Square-root choice, Rice rule, Doane's formula, Scott's normal reference rule, Freedman-Diaconis' choice, Minimizing cross-validation extimated squared error, และ Shimazaki and Shinomoto's choice ถึงแม้จะไม่มีวิธีไหนที่ดีที่สุด เนื่องจากต้องขึ้นอยู่กับข้อมูล แต่วิธีที่ถือว่าเป็นที่ยอมรับและใช้อย่างแพร่หลายคือ ฟังก์ชั่น Sturges [8] ใช้สมการดังนี้

$$\boldsymbol{K = 1 + 3.322\ \log N} \quad (3)$$



โดย K คือ อันตรภาคชั้น

N คือ จำนวนของข้อมูลในชุดข้อมูล

อย่างไรก็ตาม สมการของ Sturge ถึงแม้จะทำงานได้ดีกับข้อมูลที่มีความต่อเนื่องและมีการแจกแจงแบบปกติ (Normal Distribution) สามารถแปลงข้อมูลได้อย่างสมมาตร แต่ก็ยังไม่สามารถให้ผลลัพธ์ที่ดีได้เมื่อปริมาณข้อมูลมีมากจนเกินไปและข้อมูลมีความเบี่ยงเบนสูง

2.1.5 การแบ่งกลุ่มข้อมูล (Clustering Analysis)

เป็นกระบวนการจัดกลุ่มข้อมูลที่มีลักษณะเหมือนกันไว้ในกลุ่มเดียวกัน โดยพิจารณาจากความคล้ายคลึง (Similarity) และระยะของความห่าง (Distance measure) ที่ใช้บ่งบอกความใกล้ชิด เนื่องจากรูปแบบการแบ่งกลุ่มข้อมูลจากความคล้ายโดยไม่มีการกำหนดหมวดหมู่ของข้อมูลไว้ก่อน กล่าวคือไม่ทราบจำนวนกลุ่มล่วงหน้า จึงแตกต่างจากการแบ่งประเภทข้อมูล (Classification) และ

จัดเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) สามารถนำไปใช้ทำเหมืองข้อมูลกับงานหลายสาขา เช่น ชีววิทยา, ความปลอดภัย, ระบบธุรกิจอัจฉริยะและโปรแกรมสืบค้นข้อมูลในอินเทอร์เน็ต

วิธีการแบ่งกลุ่มข้อมูลสามารถจัดประเภทได้ดังต่อไปนี้ [9]

การแบ่งแบบตัดเป็นส่วน (Partitioning methods) อัลกอริทึมในกลุ่มนี้ได้แก่ k-means การทำงานต้องกำหนดจำนวนกลุ่มข้อมูลที่ต้องการแบ่ง (k) โดยในขั้นตอนนี้อาจจะต้องอาศัยผู้เชี่ยวชาญในการกำหนด เมื่อกำหนดค่า k แล้ว อัลกอริทึมจะประมวลผลโดยการแบ่งกลุ่มแบบสุ่มก่อนในรอบแรก จากนั้นก็จะคำนวนหาค่ากลาง (centroid) ของแต่ละกลุ่มข้อมูล เมื่อได้ค่ากลางแล้ว จุดที่ถูกกำหนดให้เข้ากลุ่มไปครั้งแรก อาจจะไม่ได้อยู่ใกล้ค่ากลางของกลุ่มที่ถูกกำหนดในรอบแรก ต้องทำการจัดกลุ่มใหม่ไปยังกลุ่มที่จุดนั้นๆ อยู่ใกล้ค่ากลาง และคำนวนหาค่ากลางของกลุ่มข้อมูลอีกครั้ง ทำแบบนี้ไปเรื่อยๆ จนไม่สามารถที่จะจัดกลุ่มข้อมูลใหม่ได้

การแบ่งแบบลำดับขั้น (Hierarchical methods) วิธีการคือ ในแต่ละจุดข้อมูล ถือว่าเป็นกลุ่มข้อมูลของตัวเอง จากนั้นจะทำการรวมกลุ่มข้อมูลโดยหากลุ่มข้อมูล สองอันที่มีระยะทาง ใกล้กันมากที่สุด นำมารวมกัน ทำเช่นนี้ไปเรื่อยๆ ซึ่งในระหว่างทาง จำนวนกลุ่มข้อมูลจะลดน้อยลง และแต่ละกลุ่มข้อมูลที่มีอยู่จะมีขนาดใหญ่ขึ้นเรื่อยๆ ซึ่งถ้าไม่หยุดกระบวนการนี้ จะทำให้มีเหลือเพียงกลุ่มข้อมูลเดียวที่ครอบคลุมข้อมูลทั้งหมด

การแบ่งแบบตาราง (Grid-based methods) นิยมใช้แบ่งข้อมูลที่มีขนาดใหญ่และมีหลายมิติ ทำการแบ่งข้อมูลเป็นตารางตามความหนาแน่นของข้อมูลในขอบเขตชุดข้อมูล ซึ่งมีข้อดีคือช่วยลดความซับซ้อนในการคำนวณ การจัดกลุ่มด้วยวิธีนี้จะคำนึงถึงพื้นที่โดยรอบของข้อมูล

การแบ่งแบบวัดความหนาแน่น (Density-based methods) การแบ่งกลุ่มพิจารณาจากความหนาแน่นภายในกลุ่มข้อมูล โดยแยกกลุ่มที่มีความหนาแน่นน้อย ออกจากกลุ่มที่มีความหนาแน่นมาก ซึ่งหากจุดข้อมูลใด ไม่ได้อยู่ในกลุ่มดังกล่าวจะถือว่าเป็นจุดผิดปกติ (Noise) อัลกอริทึมที่ได้รับความนิยมคือ DBSCAN

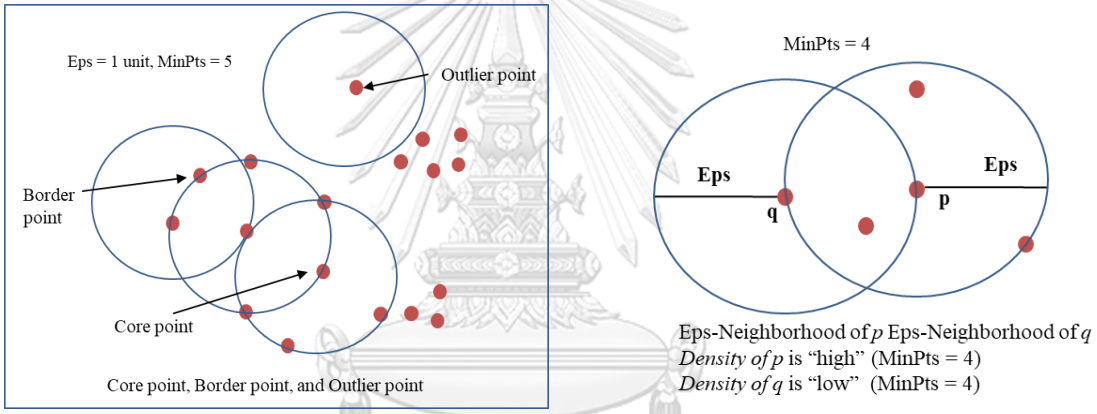
2.1.6 การแบ่งกลุ่มข้อมูลตามความหนาแน่น (Density-Based Spatial Clustering of Applications with Noise: DBSCAN)

เป็นอัลกอริทึมที่ไม่ต้องกำหนดจำนวนกลุ่มข้อมูลที่ต้องการแบ่ง ใช้เพื่อค้นหาความสัมพันธ์และโครงสร้างของข้อมูลที่ยุ่งยากแต่ยังสามารถระบุรูปแบบและคาดการณ์พฤติกรรมของข้อมูลในอนาคตได้ เหมาะสำหรับข้อมูลที่มีการกระจายตัวกันแบบไม่เป็นกลุ่มก้อน มีรูปร่างเป็นรูปทรงต่างๆ ที่ k-means ไม่สามารถจัดกลุ่มได้ และจัดการกับข้อมูลผิดปกติ (Noise) ได้ดี [2] [3] ผลลัพธ์ที่ได้จากการจัดกลุ่มจะเป็นคลัสเตอร์ไม่จำกัดขนาดและรูปทรง

*2.1.6.1 นิยามของอัลกอลิทึม*

จากภาพที่ 2 อธิบายนิยมของ DBSCAN ได้ดังนี้

a. Eps คือ รัศมีระหว่างจุดในกลุ่มข้อมูล ดังนั้นระยะห่างระหว่างจุดสองจุดต้องมีค่าน้อยกว่าหรือเท่ากับ Eps
b. MinPts คือ จำนวนจุดขั้นต่ำสุดสำหรับการสร้างจุดศูนย์กลางของกลุ่มข้อมูล (Core Point)
c. Core Point คือ จุดที่เป็นแกนหลักของกลุ่มข้อมูล มีจำนวนจุดที่ใกล้เคียงอยู่ภายในรัศมี Eps มากกว่าหรือเท่ากับ MinPts
d. Border Point คือ จุดบนเส้นขอบของกลุ่มข้อมูลที่มีจำนวนจุดที่อยู่ใกล้เคียงภายในรัศมี Eps น้อยกว่า MinPts แต่ยังอยู่ในบริเวณใกล้เคียง Core Point
e. Noise Point คือ จุดใดๆ ที่ไม่ได้อยู่ในกลุ่มข้อมูล




ภาพที่ 2 อธิบายนิยามของ DBSCAN

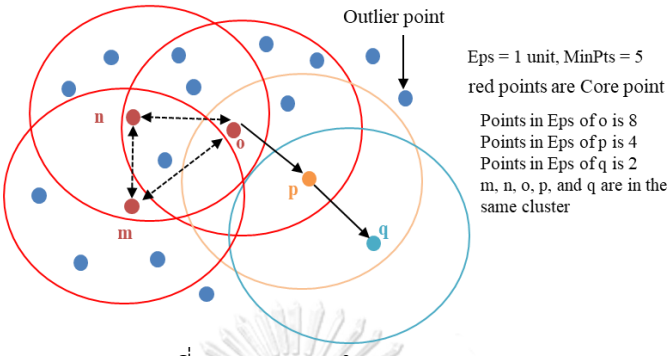


*2.1.6.2 การทำงานของอัลกอลิทึม*

การประมวลผลเริ่มจาก การกำหนดจุดแกนกลาง (Core points) ทุกจุดในชุดข้อมูลจะถูกตั้งค่าเป็นไม่ถูกตรวจสอบ เมื่อจุดใดที่อยู่ภายในระยะของ Eps ถูกเข้าถึงจากจุดแกนกลาง จุดนั้นจะถูกทำเครื่องหมายว่าถูกตรวจสอบ จากนั้นดำเนินการสำรวจทุกจุดที่อยู่ใน Eps กลุ่มข้อมูล (Cluster) จะถูกสร้างเมื่อจำนวนจุดมากกว่าหรือเท่ากับ MinPts จุดที่เหลือจะถูกสร้างเป็นคอลเลกชัน N. ทำซ้ำเพื่อค้นหาจุดแกนกลาง และสร้างกลุ่มข้อมูลจนกว่าจะถึงจุดใด ๆ ที่ไม่สามารถจัดสรรให้กับกลุ่มข้อมูลใดได้ ซึ่งจุดเหล่านั้นจะถูกกำหนดให้เป็นจุดที่อยู่นอกกลุ่มข้อมูล (Noise point)

จากภาพที่ 3 แสดงการสร้างกลุ่มข้อมูล โดยจุดสีแดงคือ จุดแกนกลางของกลุ่ม ซึ่งในหนึ่งกลุ่มข้อมูลสามารถมีจุดแกนกลางได้หลายจุด หากจุดแกนกลางสามารถเข้าถึงกันได้โดยอยู่ภายในรัศมี

Eps ระหว่างกัน จุด p เป็นจุดบนเส้นขอบของ o เพราะอยู่ในรัศมีของ o แต่มีจำนวน point น้อยกว่า MinPts ส่วนจุด q อยู่ในรัศมีของ p ซึ่งทำให้ o สามารถเข้าถึงได้โดยผ่านจุด p ดังนั้น จุด m, n, o, p, และ q จึงจัดอยู่ในกลุ่มข้อมูลเดียวกัน




ภาพที่ 3 แสดงการทำงานของ DBSCAN

DBSCAN สามารถจัดกลุ่มได้อย่างมีประสิทธิภาพด้วยชุดข้อมูลความหนาแน่นเดียว อย่างไรก็ตามจะมีปัญหาเกี่ยวกับชุดข้อมูลที่มีความหนาแน่นที่แตกต่างกัน ดังนั้นจึงมีการพัฒนาอัลกอริทึมเพื่อใช้เพิ่มประสิทธิภาพการทำงานให้แก่ DBSCAN

OPTICS เป็นอัลกอริทึมที่มีแนวความคิดเดียวกับ DBSCAN แต่สามารถแก้ไขจุดอ่อนการทำงานของ DBSCAN กับข้อมูลที่มีความหนาแน่นแตกต่างกันได้ โดยทำการเรียงลำดับข้อมูลเพื่อหาจุดที่ใกล้กัน และพิจารณาระยะทางระหว่างจุดเพื่อแสดงถึงความหนาแน่น และการเป็นสมาชิกของกลุ่มเดียวกัน

VDBSCAN พัฒนาขึ้นเพื่อแก้ไขปัญหาจุดอ่อนของ DBSCAN ที่ไม่สามารถทำงานได้อย่างมีประสิทธิภาพเมื่อข้อมูลมีความหนาแน่นที่แตกต่างกัน โดยใช้การกำหนด Eps ที่เหมาะสมให้แก่ความหนาแน่นในแต่ละระดับ ซึ่งจะกล่าวถึงวิธีการทำงานเพิ่มเติมในหัวข้อที่ 2.2.1.2

HDBSCAN พัฒนาต่อเนื่องจาก DBSCAN โดยใช้เทคนิคการแบ่งกลุ่มข้อมูลแบบลำดับขั้น (Hierarchical methods) มาประยุกต์ใช้ นอกจากนี้ยังให้ความสำคัญกับจุดที่เป็นแกนหลักของกลุ่มข้อมูล โดยไม่สนใจจุดบนเส้นขอบของกลุ่ม ซึ่งผลลัพธ์ที่ได้จากการแบ่งกลุ่ม จะเป็นกลุ่มของคลัสเตอร์ที่มีความเป็นไปได้ว่าจะถูกแบ่งกลุ่มด้วย DBSCAN

ประสิทธิภาพการทำงานของ DBSCAN ขึ้นอยู่กับการเลือกใช้พารามิเตอร์ จำนวนจุดขั้นต่ำสุดสำหรับการสร้างจุดศูนย์กลางของกลุ่มข้อมูล (MinPts) เป็นพารามิเตอร์ที่ง่ายที่สุดในการตั้งค่าโดยทั่วไปมักกำหนดไว้เท่ากับ 2 สำหรับข้อมูลหนึ่งมิติ พารามิเตอร์ที่ยากต่อการตั้งค่าคือ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) แต่มีข้อพิจารณาว่า ควรจะมีขนาดเล็กมากที่สุด โดยขึ้นอยู่กับระยะทางระหว่างข้อมูล หรือผู้เชี่ยวชาญเป็นผู้กำหนด แต่ในกรณีที่ข้อมูลมีปริมาณมากก็ควรกำหนดให้

เหมาะสมตามปริมาณของข้อมูล มีการนำเสนออัลกอริทึมเพื่อเลือกพารามิเตอร์เพื่อใช้สำหรับการทำงานของ DBSCAN [10] ไว้หลากหลายวิธี ดังนี้

ตารางที่ 1 อัลกอริทึมการเลือกพารามิเตอร์เพื่อใช้แบ่งกลุ่มของ DBSCAN

| อัลกอริทึม | พารามิเตอร์ | ประสิทธิภาพ |
|---|---|---|
| DMDBSCAN | k | สามารถแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ และให้ผลลัพธ์ที่ดีในการแบ่งกลุ่ม |
| LDBSCAN | MinPts, LOFUD, pct | สามารถแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ |
| EDBSCAN | $\varepsilon, \delta, \mu, \tau$ | สามารถแยกคลัสเตอร์ออกจากกันได้ชัดเจนแม้จะอยู่ติดกัน และแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ |
| DBSCAN-GM | - | กำหนดพารามิเตอร์ได้อย่างอิสระ การทำงานมีประสิทธิมากกว่า DBSCAN, KMeans และ G-Mean |
| VDBSCAN | - | กำหนดพารามิเตอร์ได้อย่างอิสระ สามารถแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ |
| FNDBSCAN | $\varepsilon 1, \varepsilon 2$ | สามารถแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ ประมวลผลได้เร็วกว่าอัลกอริทึมการแบ่งกลุ่ม FJP และ NRFJP |
| Soft-DBSCAN | $\varepsilon, MinPts$ | สามารถแบ่งกลุ่มข้อมูลได้แม่นยำมากกว่าอัลกอรึทึม FCM และสามารถแบ่งกลุ่มข้อมูลที่มีความหนาแน่นแตกต่างกันได้ |
| Active-DBSCAN | N, b, B | สามารถแบ่งกลุ่มข้อมูลได้แม่นยำมากกว่าอัลกอรึทึมอื่นๆ |
| FDBSCN | $\varepsilon, MinPts$ | สามารถแบ่งกลุ่มข้อมูลได้แม่นยำมากกว่า DBSCAN และ Eps มีผลกระทบอย่างรวดเร็วต่อการทำงาน |
| Fast Parzen-Window | theshold | สามารถทำงานได้เร็วมากกว่า DBSCAN |

## 2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้ แบ่งออกเป็นประเภท 3 ประเภทคือ พื้นที่และการแบ่งกลุ่มข้อมูลตามความหนาแน่น (Density-Based Spatial Clustering of Applications with Noise: DBSCAN) การเลือกพารามิเตอร์ที่เหมาะสมในการทำงานของอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น และการค้นหาจุดที่น่าสนใจ

2.2.1 การแบ่งกลุ่มข้อมูลตามความหนาแน่น

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

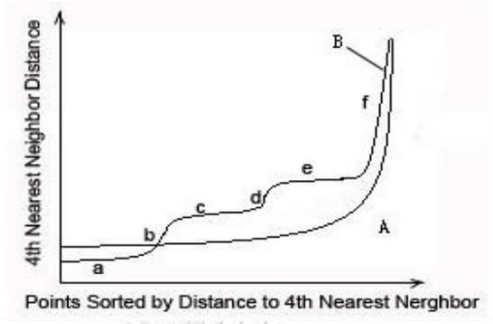
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [2] โดย Martin Ester, Hans-Peter Kriegel, Jörg Sander และ Xiaowei Xu ในปี พ.ศ. 2539 เป็นอีกหนึ่งวิธีในการแบ่งกลุ่มข้อมูลที่ได้รับความนิยม DBSCAN ผลลัพธ์ของการแบ่งกลุ่มจะให้กลุ่มข้อมูลที่มีขนาด รูปร่างอิสระ และหลากหลายรูปแบบ การทำงานของอัลกอริทึมจำเป็นต้องใช้พารามิเตอร์สองค่าที่มีสำคัญเป็นอย่างมากในการประมวลผล ประกอบด้วย รัศมีระหว่างจุดในกลุ่มข้อมูล (radius epsilon: Eps) และจำนวนจุดขั้นต่ำสุดในการสร้างกลุ่มข้อมูล (Minimum Points: MinPts)

การประมวลผลเริ่มจาก การกำหนดจุดแกนกลาง (Core points) ทุกจุดในชุดข้อมูลจะถูกตั้งค่าเป็นไม่ถูกตรวจสอบ เมื่อจุดใดที่อยู่ภายในระยะของ Eps ถูกเข้าถึงจากจุดแกนกลาง จุดนั้นจะถูกทำเครื่องหมายว่าถูกตรวจสอบ จากนั้นดำเนินการสำรวจทุกจุดที่อยู่ใน Eps กลุ่มข้อมูล (Cluster) จะถูกสร้างเมื่อจำนวนจุดมากกว่าหรือเท่ากับ MinPts จุดที่เหลือจะถูกสร้างเป็นคอลเลกชัน N. ทำซ้ำเพื่อค้นหาจุดแกนกลาง และสร้างกลุ่มข้อมูลจนกว่าจะถึงจุดใด ๆ ที่ไม่สามารถจัดสรรให้กับกลุ่มข้อมูลใดได้ ซึ่งจุดเหล่านั้นจะถูกกำหนดให้เป็นจุดที่อยู่นอกกลุ่มข้อมูล (Noise point) DBSCAN สามารถจัดกลุ่มได้อย่างมีประสิทธิภาพด้วยชุดข้อมูลความหนาแน่นเดียว อย่างไรก็ตามจะมีปัญหาเกี่ยวกับชุดข้อมูลที่มีความหนาแน่นที่แตกต่างกัน

*2.2.1.2 Varied Density Based Spatial Clustering of Applications with Noise (VDBSCAN)*

Varied Density Based Spatial Clustering of Applications with Noise (VDBSCAN) [11] พัฒนาขึ้นเพื่อแก้ไขปัญหาจุดอ่อนของ DBSCAN ที่ไม่สามารถทำงานได้อย่างมีประสิทธิภาพเมื่อข้อมูลมีความหนาแน่นที่แตกต่างกัน แนวคิดของอัลกอริทึมนี้คือ การเลือก Eps ในแต่ละระดับความหนาแน่น ซึ่ง Eps คำนวณจากการหาระยะทางระหว่างจุดไปยังจุดเพื่อนบ้านที่ใกล้ที่สุดลำดับที่ k ตามที่ถูกกำหนดไว้ล่วงหน้าโดยผู้เชี่ยวชาญ ซึ่งค่า k ดังกล่าวนี้ถูกนำมาใช้เป็น MinPts เมื่อทำการประมวลผลด้วย DBSCAN อัลกอลิทึมจะดำเนินการแบ่งกลุ่มตาม Eps ในแต่ละระดับความหนาแน่น หากจุดใดถูกกำหนดเข้ากลุ่มแล้ว ในรอบต่อไปจะไม่ถูกนำมาจัดกลุ่มอีก จากภาพที่ 4 แสดงลักษณะ

กราฟ k-dist ที่สร้างจากการเรียงลำดับระยะทางระหว่างจุดไปยังจุดที่ใกล้ที่สุดลำดับที่ k จากน้อยไปมาก โดยเปรียบเทียบให้เห็นข้อแตกต่างของลักษณะเส้นกราฟระหว่างข้อมูลมีความหนาแน่นระดับเดียว ดังแสดงในเส้น A กับข้อมูลที่มีความหนาแน่นหลายระดับดังแสดงในเส้น B




(ก) กราฟ k-dist เปรียบเทียบลักษณะความหนาแน่นที่แตกต่างกัน

ภาพที่ 4 กราฟ k-dist แสดงการเปรียบเทียบลักษณะความหนาแน่นที่แตกต่างกัน [11]

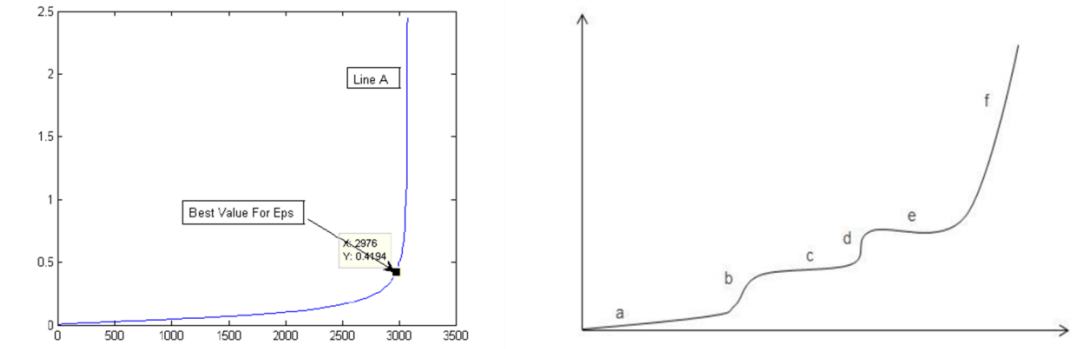
2.2.2 การเลือกพารามิเตอร์

*2.2.2.1 A Dynamic Method for Discovery Density Varied Cluster (DMDBSCAN)*

A Dynamic Method for Discovery Density Varied Clusters (DMDBSCAN) [12] เนื่องจากปัญหาของ DBSCAN คือไม่สามารถจัดการกับการกระจายตัวของข้อมูลที่มีความหนาแน่นมากๆ ได้ ดังนั้นเพื่อทำการปรับปรุงการทำงานของ DBSCAN ให้มีประสิทธิภาพยิ่งขึ้น งานวิจัยจึงนำเสนอวิธีการกำหนดรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) แบบอัตโนมัติ สามารถเปลี่ยนแปลงตามระยะทางระหว่างจุด k ไปยังจุดที่ใกล้ที่สุด โดยลำดับจุดที่ใกล้ที่สุดที่นำมาคำนวณจะถูกกำหนดไว้ล่วงหน้า และนำมาใช้เป็น จำนวนจุดขั้นต่ำสุด (MinPts)

ค่า Eps ที่เหมาะสมถูกเลือกจากกราฟของระยะทางระหว่างจุด k ไปยังจุดที่ใกล้ที่สุด ที่ถูกเรียงลำดับระยะทางจากน้อยไปหามาก ซึ่งเรียกว่า k-dist ลักษณะกราฟแสดงตามภาพที่ 5 (ก) บริเวณที่แสดงจุดเปลี่ยนแปลงของกราฟจะถูกเลือกให้เป็นค่า Eps เพื่อนำมาให้เป็นพารามิเตอร์สำหรับการแบ่งกลุ่มข้อมูลด้วย DBSCAN ขนาดของ Eps จะเปลี่ยนแปลงไปตามจำนวนของ MinPts แต่ไม่ได้เปลี่ยนแปลงอย่างรวดเร็วหรือทันที

ในกรณีที่ข้อมูลมีความหนาแน่นแตกต่างกัน กราฟ k-dist แสดงดังภาพที่ 5 (ข) ซึ่งการทำงานของ DMDBSCAN จะกำหนดระดับความหนาแน่น และหาค่า Eps ในแต่ละระดับความหนาแน่น จาก (ข) ความหนาแน่นของข้อมูลแบ่งเป็น 3 ระดับ คือ เส้น a และ b กำหนดเป็น Eps1 เส้น c กำหนดเป็น Eps2 และเส้น e กับ f สามารถกำหนดเป็น Eps3




(ก) กราฟ k-dist สำหรับข้อมูลที่มีความหนาแน่นที่ไม่แตกต่างกัน (ข) กราฟ k-dist สำหรับข้อมูลที่มีความหนาแน่นแตกต่างกัน

ภาพที่ 5 กราฟแสดงการเลือกค่า Eps [12]

*2.2.2.2 AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset*

AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset [13] นำเสนอวิธีกำหนดช่วงรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และจำนวนจุดขั้นต่ำสุด (MinPts) ที่เหมาะสมอย่างอัตโนมัติ เพื่อแก้ปัญหาในกรณีที่ผู้เชี่ยวชาญไม่สามารถกำหนดพารามิเตอร์ได้ และสำหรับข้อมูลที่มีความหนาแน่นแตกต่างกัน

ขั้นตอนการกำหนดพารามิเตอร์มีวิธีที่แตกต่างจาก VDBSCAN โดยหาค่าของ Eps จากค่าเฉลี่ยของระยะทางของทุกจุดไปยัง k และจุดที่อยู่ใกล้ที่สุด ในขณะที่ VDBSCAN คำนวณระยะทางเฉพาะตามจำนวนจุดที่กำหนดไว้เท่านั้น นำระยะทางที่ได้สร้างเป็นกราฟ k-dist โดยเรียงลำดับระยะทางจากน้อยไปมาก ลักษณะกราฟที่ได้แสดงเช่นเดียวกับ k-dist ของ VDBSCAN ซึ่งแสดงให้เห็นถึงระดับความหนาแน่นที่แตกต่างกันเป็นชั้นๆ คล้ายขั้นบันได เลือกค่า Eps ที่เหมาะสมในแต่ละชั้น โดยคำนวณหาค่าความลาดชันระหว่างชั้นความหนาแน่นตามเกณฑ์ที่กำหนด ซึ่งจะทำให้ได้ค่า Eps ที่แตกต่างกัน หลังจากนั้นให้คำนวณหาค่า MinPst จากค่าเฉลี่ยของจำนวนจุดภายในแต่ละช่วง Eps

เมื่อกำหนดพารามิเตอร์ได้แล้ว นำไปแบ่งกลุ่มข้อมูลด้วยอัลกอริทึม DBSCAN ในแต่ละช่วงของ Eps และ MinPts ซึ่งข้อมูลจุดใดถูกแบ่งกลุ่มแล้วจะไม่ถูกนำมาจัดกลุ่มสำหรับการประมวลผลในรอบถัดไปเช่นเดียวกับ VDBSCAN

2.2.3 การค้นหาจุดที่น่าสนใจ

*2.2.3.1 Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra*

Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra [14] นำวิธีการเลือกพารามิเตอร์จากแนวคิดของ DMDBSCAN ไปประยุกต์ใช้เพื่อค้นหาสถานที่สำคัญบนเกาะสุมาตรา ซึ่งกำหนด MinPts เท่ากับ 3 จากนั้นคำนวณหาค่า Eps ที่เหมาะสมจากกราฟ k-dist เลือกค่า Eps จากจุดที่ทำให้กราฟเกิดการลาดเอียงขึ้นจากสูตร (y2-y1)/(x2-x1) กำหนดอัตราส่วนการเปลี่ยนแปลงที่ 1% ดังนั้นความความลาดเอียงที่เท่ากับ 1% จะถูกเลือกเป็นค่า Eps

ผลลัพธ์การแบ่งกลุ่มโดยใช้พารามิเตอร์ที่เลือกด้วย DBSCAN พบว่าสามารถจัดแบ่งกลุ่มข้อมูลและจำแนกเป็นจุดฮอตสปอตบนเกาะสุมาตรา พร้อมทั้งพบบริเวณที่มีความหนาแน่นสูงสุดคือจังหวัดเรียว ซึ่งอยู่ในกลุ่มที่ 1

*2.2.3.2 Exploiting Taxi Demand Hotspots Based on Vehicular Big Data Analytics*

Exploiting Taxi Demand Hotspots Based on Vehicular Big Data Analytics [15] นำเสนอการค้นหาจุด hotspots จากข้อมูลการรับผู้โดยสารของรถแท็กซี่ด้วยอัลกอริทึมที่เรียกว่า GD-DBSCAN ซึ่งเป็นอัลกอรีทึมที่พัฒนาขึ้นเพื่อใช้เพิ่มประสิทธิภาพการแบ่งกลุ่มข้อมูลของ DBSCAN โดยประยุกต์ใช้เทคนิคของ Grid ในการแบ่งพื้นที่ ผลลัพธ์ที่ได้ประกอบด้วยพื้นที่นั้นและพื้นที่รอบๆ ที่ทำการแบ่ง จากนั้นจัดโครงสร้างของข้อมูลการรับผู้โดยสารภายในแต่ละพื้นที่ ด้วยอัลกอริทึม Kd-tree อีกครั้ง ก่อนที่จะนำไปแบ่งกลุ่มข้อมูลด้วย DBSCAN

จากการทดลองนำอัลกอริทึม GD-DBSCAN แบ่งกลุ่มข้อมูลการรับส่งผู้โดยสารในเซี่ยงไฮ้จำนวน 10,000 ข้อมูล พบว่าสามารถแบ่งกลุ่มข้อมูลได้เหมือนกับการใช้อัลกอริทึม DBSCAN และยังลดความซับซ้อนของการทำงาน ทำให้สามารถปรับปรุงประสิทธิภาพการประมวลผลได้อย่างน้อย 10% เมื่อเทียบกับด้วย DBSCAN

*2.2.3.3 P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos*

P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos [16] ซึ่งเป็นอัลกอริทึมการจัดกลุ่มตามความหนาแน่นแบบใหม่ที่พัฒนาต่อเนื่องจาก DBSCAN สำหรับการวิเคราะห์สถานที่และเหตุการณ์ โดยใช้ชุดภาพถ่ายที่ติดแท็กทางภูมิศาสตร์ (geo-tagged photos) โดยเมื่อนำไปแบ่งกลุ่ม

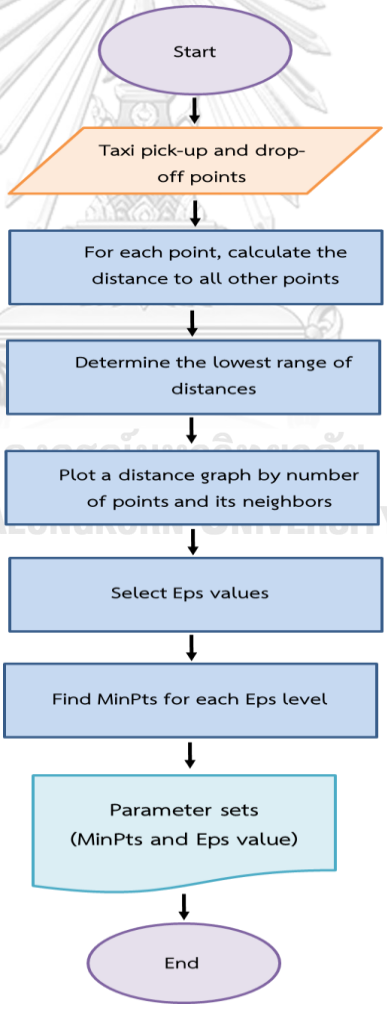
ข้อมูลภาพถ่ายที่มีการติดแท็กทางภูมิศาสตร์ ในพื้นที่ของวอชิงตันดีซี พบว่าสามารถแสดงให้เห็นพื้นที่ที่เป็นจุดฮอตสปอต และแสดงให้เห็นว่าการกำหนดพารามิเตอร์โดยความหนาแน่นกำหนดตามจำนวนประชากรในพื้นที่ และการปรับความหนาแน่นมีผลต่อการแบ่งกลุ่มข้อมูล

*2.2.3.4 Detecting Hotspots from Taxi Trajectory Data Using Spatial*

Detecting Hotspots from Taxi Trajectory Data Using Spatial [17] งานวิจัยนำเสนอการแบ่งกลุ่มข้อมูลด้วยเทคนิคของต้นไม้ตัดสินใจ (Decision tree) เพื่อค้นหาจุดฮอตสปอต (hotspot) จากข้อมูลวิถีการเดินทางของรถแท็กซี่ที่ให้บริการในเมืองอู่ฮั่น (Wuhan) ในช่วงวันหยุด วันธรรมดา และวันหยุดสุดสัปดาห์ โดยผลลัพธ์เปรียบเทียบและวิเคราะห์แผนผังการกระจายฮอตสปอตในวันหยุดและวันหยุดสุดสัปดาห์พบว่ารูปแบบการแพร่กระจายของฮอตสปอตในช่วงเวลาที่เลือกมีความคล้ายคลึงกัน แต่อย่างไรก็ตาม พบว่าวันหยุด วันธรรมดา และวันหยุดสุดสัปดาห์ มีอิทธิพลต่อความแตกต่างของการกระจายตัวของข้อมูล ส่งผลให้มีผลต่อการปรากฏขึ้นของจุดฮอตสปอต

# บทที่ 3
# การออกแบบอัลกอริทึม

การทำงานของอัลกอริทึมแบ่งกลุ่มข้อมูลตามความหนาแน่น (DBSCAN) จำเป็นต้องใช้พารามิเตอร์ 2 พารามิเตอร์คือ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และ จำนวนจุดต่ำสุดเพื่อกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) โดยทั่วไป ผู้เชี่ยวชาญหรือผู้ที่มีความรู้ความเข้าใจในข้อมูลมักระบุจำนวนของจุดต่ำสุดไว้ล่วงหน้า เพื่อที่จะใช้หารัศมีระหว่างจุดในกลุ่มข้อมูล แต่ในกรณีที่ผู้ใช้งานไม่สามารถระบุจำนวนของจุดต่ำสุดได้เหมาะสมหรือเมื่อข้อมูลมีปริมาณมาก พารามิเตอร์ที่นำมาใช้ในการแบ่งกลุ่มควรกำหนดจากลักษณะและปริมาณของข้อมูล โดยงานวิจัยเสนอขั้นตอนกำหนดพารามิเตอร์จากความหนาแน่นและปริมาณของข้อมูลแบบอัตโนมัติ เพื่อใช้สำหรับการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมแบ่งกลุ่มตามความหนาแน่น ดังที่แสดงในภาพที่ 6



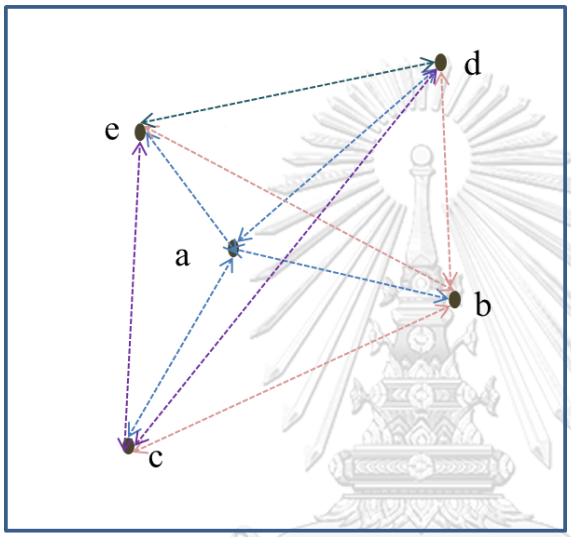

ภาพที่ 6 แสดงขั้นตอนการกำหนดพารามิเตอร์แบบอัตโนมัติ

จากภาพที่ 6 สามารถอธิบายรายละเอียดขั้นตอนของการกำหนดพารามิเตอร์แบบอัตโนมัติดังต่อไปนี้

**3.1 คำนวนหาระยะทางระหว่างพิกัด**

For each point, calculate the distance to all other points.

คำนวนหาระยะทางจากพิกัดไปยังทุกพิกัดที่อยู่ในชุดข้อมูล โดยใช้ฟังก์ชั่น Haversine ซึ่งเป็นฟังก์ชั่นที่ใช้ในการคำนวนหาระยะทางระหว่างจุดสองจุด ผลลัพธ์ที่ได้เป็นชุดข้อมูลของระยะทางระหว่างพิกัดที่มีหน่วยเป็นเมตร




| ตำแหน่งต้นทาง | ตำแหน่งปลายทาง | ระยะทาง (เมตร) |
|---|---|---|
| a | b | 20 |
| a | c | 20 |
| a | d | 60 |
| a | e | 10 |
| b | c | 60 |
| b | d | 20 |
| b | e | 100 |
| c | d | 100 |
| c | e | 30 |
| d | e | 25 |

ภาพที่ 7 แสดงตัวอย่างการคำนวณระยะทางระหว่างพิกัด



**3.2 หาช่วงระยะทางระหว่างพิกัดที่น้อยที่สุด**

Determine the lowest range of distances.

โดยบริเวณที่เป็นจุดสนใจจะต้องแสดงถึงความหนาแน่นและมีการกระจุกตัวของข้อมูล ระยะทางระหว่างพิกัดจึงควรเป็นช่วงระยะทางที่น้อยที่สุด การหาช่วงระยะทางดังกล่าวในการวิจัยนี้หาได้จากสูตร Sturges'rule ตามสูตรที่ (4) แบ่งช่วงชุดข้อมูลระยะทางระหว่างพิกัด จากนั้นเลือกช่วงข้อมูลที่น้อยที่สุด

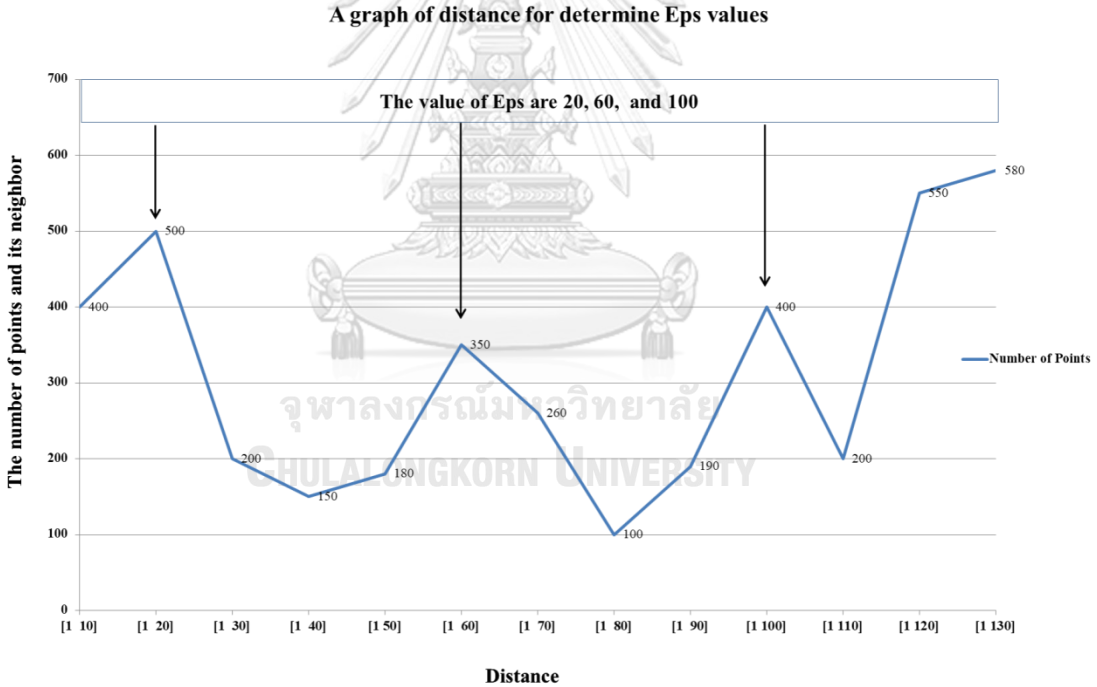
$$\text{Lowest range} = \frac{\text{max distance value} - \text{min distance value}}{1 + 3.322 \log N} \qquad (4)$$

โดย $1 + 3.322 \log N$ คือ Sturge's rule และ N จำนวนข้อมูลในชุดข้อมูล

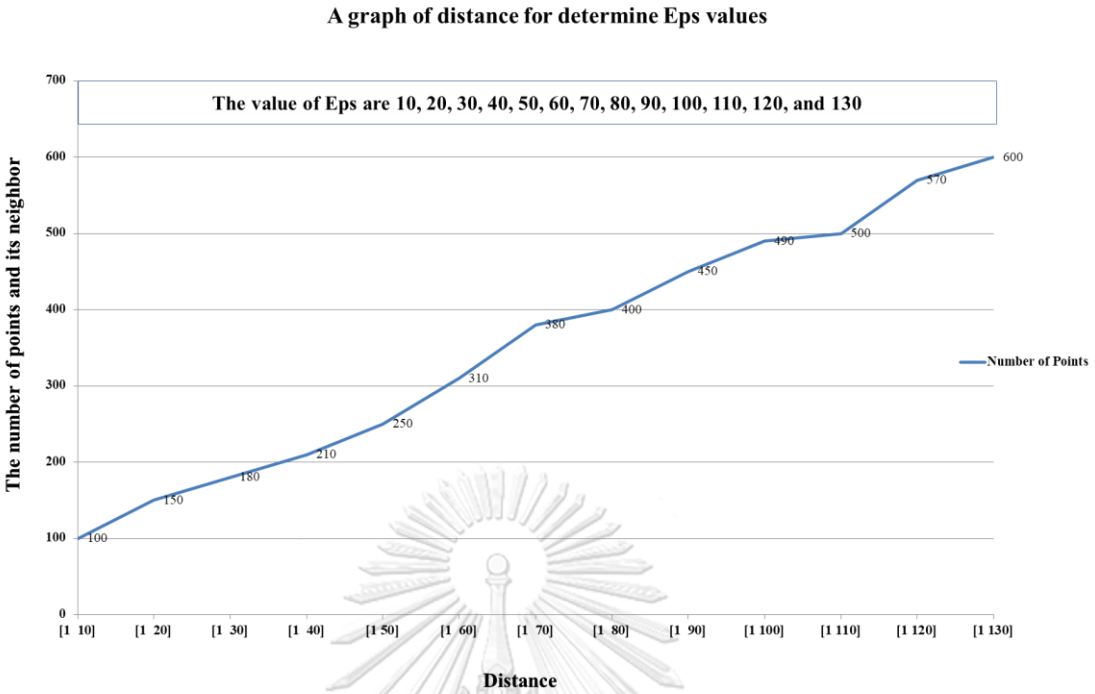
**3.3 นำช่วงข้อมูลระยะทางที่ได้จากผลลัพธ์ในข้อ 3.2 สร้างเป็นกราฟ**

Plot a distance graph by number of points and its neighbors and Select Eps values.

เมื่อได้ผลลัพธ์จากเป็นช่วงข้อมูลที่น้อยที่สุด นำผลดังกล่าวมาสร้างเป็นกราฟ โดยกำหนดช่วงระยะทางโดยใช้สูตรที่ (4) อีกครั้ง ซึ่งลักษณะกราฟจะแสดงแตกต่างกันดังภาพที่ 8 กราฟ (ก) มีความแตกต่างของระยะทางที่ชัดเจน สามารถเลือกค่าของรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) ได้จากตำแหน่งที่แสดงถึงจุดที่มีความหนาแน่นสูง คือบริเวณที่มีการเปลี่ยนแปลงสูงสุดในแต่ละช่วงระยะของเส้นกราฟ ค่ารัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) ที่สามารถระบุได้คือ 20 60 และ 100 ส่วนกราฟแสดงดัง (ข) ไม่สามารถระบุความแตกต่างของระยะทางได้อย่างชัดเจน สามารถเลือกค่าของรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) ได้จากทุกตำแหน่งของช่วงระยะของเส้นกราฟ จากกราฟที่แสดงสามารถระบุค่ารัศมีระหว่างจุดในกลุ่มข้อมูลได้ดังนี้ 10 20 30 40 50 60 70 80 90 และ 100




(ก) กราฟแสดงการหาค่ารัศมีของกลุ่ม (Eps) ที่แสดงความแตกต่างอย่างชัดเจน



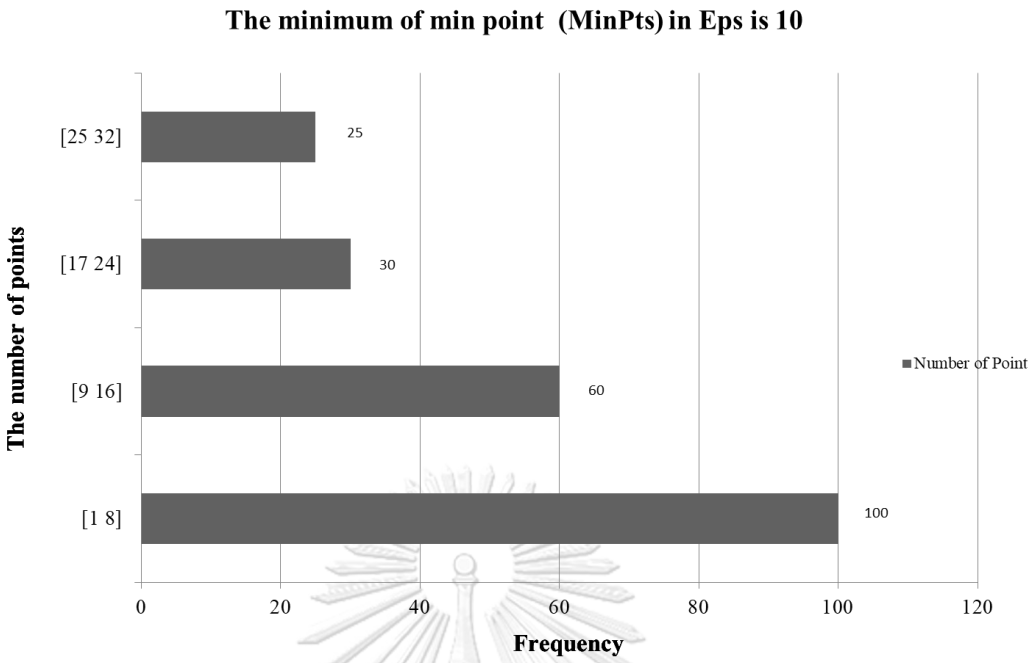

**(ข) กราฟแสดงการหาค่ารัศมีของกลุ่ม (Eps) ที่แสดงความแตกต่างไม่ชัดเจน**

ภาพที่ 8 แสดงกราฟเพื่อใช้หาค่ารัศมีระหว่างจุดในกลุ่มข้อมูล (Eps)

### 3.4 หาจำนวนจุดต่ำสุดเพื่อกำหนดจุดศูนย์กลางของกลุ่ม (MinPts)

Find MinPts for each Eps level.

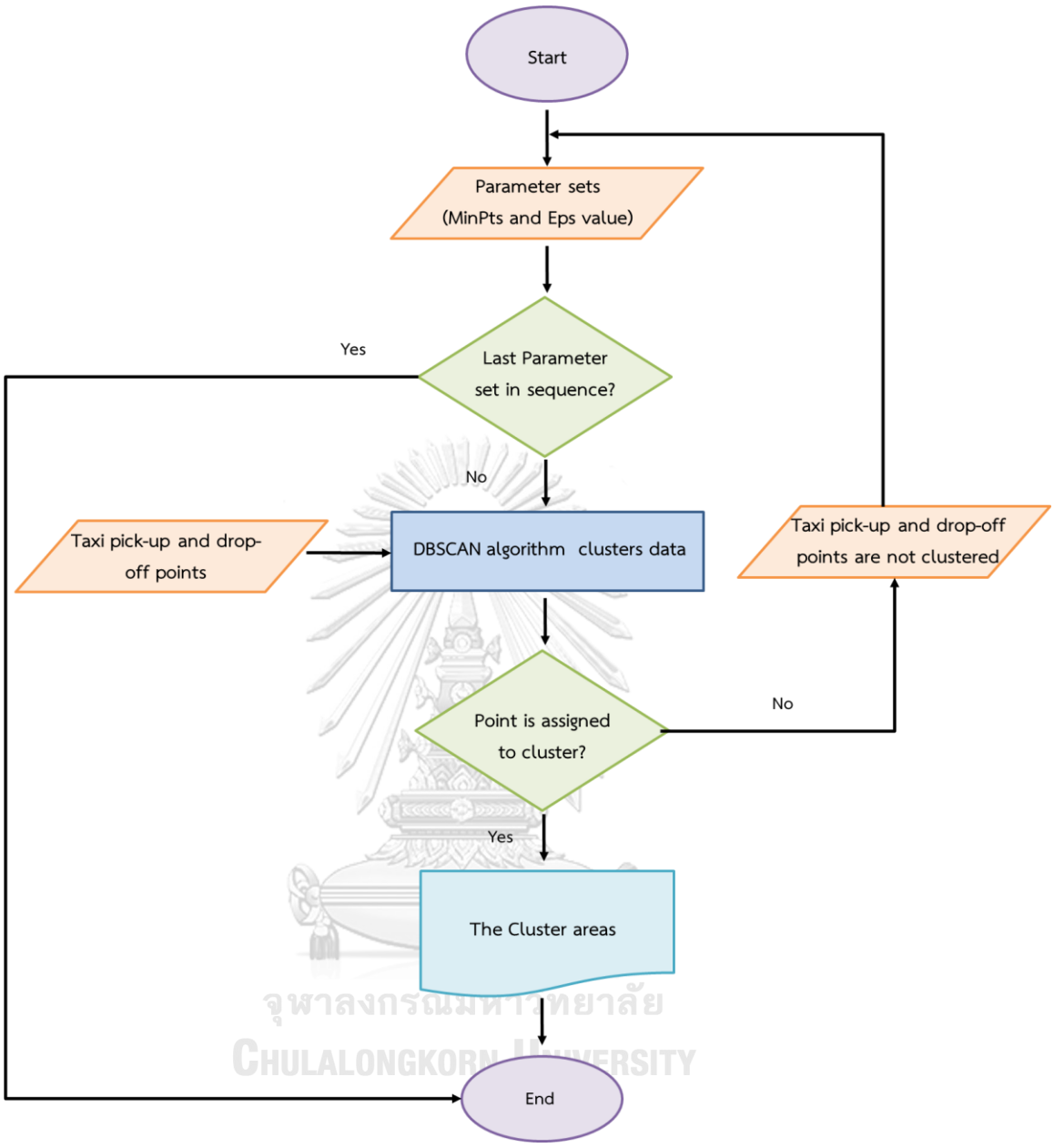
เมื่อสามารถระบุค่ารัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) จากนั้นนำมาใช้หาจำนวนจุดต่ำสุดเพื่อกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) โดยนับจำนวนคู่พิกัดที่อยู่ภายในระยะรัศมีระหว่างจุดในกลุ่มข้อมูลในแต่ละช่วงค่ารัศมีระหว่างจุดในกลุ่มข้อมูลที่หาได้จากข้อ 3.3 แล้วสร้างเป็นฮิสโตแกรม โดยกำหนดการหาช่วงข้อมูลตามสูตร (4) เลือกค่าสูงสุดจากช่วงข้อมูลที่มีจำนวนความถี่ของข้อมูลมากที่สุดเป็นจำนวนจุดต่ำสุดเพื่อกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) ดังภาพที่ 9 เมื่อรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) = 10 จุดต่ำสุดเพื่อกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) = 8




ภาพที่ 9 แสดงฮิสโตแกรมเพื่อใช้หาจำนวนจุดต่ำสุดสำหรับกำหนดจุดศูนย์กลางของกลุ่ม (MinPts)

### 3.5 จัดกลุ่มข้อมูล

เมื่อกำหนดค่าพารามิเตอร์ที่เหมาะสมทั้งสองพารามิเตอร์ได้เรียบร้อยแล้ว กระบวนการต่อมาคือทำการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น (DBSCAN) ตามภาพที่ 10 อัลกอริทึมจะจัดกลุ่มตามรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และจำนวนจุดต่ำสุด (MinPts) พิกัดใดที่ไม่ถูกจัดให้เข้ากลุ่มจะถือว่าเป็น Noise และจะถูกนำไปจัดกลุ่มโดยใช้พารามิเตอร์ในชุดถัดไปดำเนินการไปเรื่อยๆ จนครบจำนวนชุดพารามิเตอร์สำหรับชุดข้อมูลนั้น จึงจะเสร็จสิ้นการแบ่งกลุ่มข้อมูล






ภาพที่ 10 กระบวนการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึม DBSCAN โดยใช้พารามิเตอร์ที่กำหนดแบบอัตโนมัติ

.

# บทที่ 4
# การทดลองและผลการทดลอง

วิทยานิพนธ์ฉบับนี้แบ่งการทดลองเป็น 6 ส่วน เพื่อแสดงให้เห็นถึงผลลัพธ์ของการค้นหาพื้นที่ที่เป็นจุดสนใจ จากการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น โดยใช้พารามิเตอร์ที่นำเสนอในงานวิจัยนี้ แบ่งออกได้ดังต่อไปนี้

1) การทดลองการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีความหนาแน่นและการกระจายข้อมูลที่แตกต่างกัน

2) การทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก DMDBSCAN

3) การทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก VDBSCAN

4) การทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก AutoEpsDBSCAN

5) การทดลองการแบ่งกลุ่มในวันที่แตกต่างกัน

6) การทดลองเปรียบเทียบผลการแบ่งกลุ่มข้อมูลโดยใช้พารามิเตอร์ที่นำเสนอกับแผนที่ออนไลน์

## 4.1 ระบบที่ใช้ในการทดลอง

ระบบที่ใช้ในการทดลอง สามารถสรุปได้ดังนี้

4.1.1 คอมพิวเตอร์ที่ใช้ทำการทดลอง และการพัฒนาโปรแกรม



ดำเนินการทดลองงานวิจัยนี้บนเครื่องคอมพิวเตอร์ที่มีหน่วยประมวลผลกลาง Intel Core i5-7200U ความเร็ว 2.7 Ghz หน่วยความจำขนาด 12 GB ระบบปฏิบัติการ Windows 10 64 bits

พัฒนาโปรแกรมโดยใช้ภาษา R v 3.4.1 บน RStudio v. 1.1.1419

4.1.2 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลดิบที่ได้จากเครื่องบันทึกการรับส่งผู้โดยสารอยู่ในรูปแบบของไฟล์เอกซ์เอ็มแอล แต่เนื่องจากงานวิจัยนี้ต้องการใช้ข้อมูลที่เป็นจุดรับส่งผู้โดยสารเท่านั้น จึงต้องสกัดข้อมูลเพื่อระบุพิกัดการรับส่งผู้โดยสาร [18] ดังนี้

การสกัดเพื่อระบุพิกัดการรับส่งผู้โดยสารของรถแท็กซี่ สามารถระบุได้จากการเปลี่ยนแปลงของมิเตอร์ไฟที่แสดงสถานะรถว่าง โดยเมื่อมีการรับผู้โดยสารขึ้นรถ คนขับรถแท็กซี่จะทำการกดมิเตอร์ไฟให้ดับลง ข้อมูลที่ถูกบันทึกจะระบุสถานะในฟิวส์ข้อมูล passenger_flag = 0 เมื่อถึงจุดหมาย

คนขับรถแท็กซี่จะทำการกดมิเตอร์ไฟแสดงให้สว่างขึ้น ข้อมูลที่ถูกบันทึกจะระบุสถานะในฟิวส์ข้อมูล passenger_flag = 1 ทั้งนี้ยังพิจารณาถึงสถานะของเครื่องยนต์ ที่ระบุในฟิวด์ engine_flag = 1 หมายถึงเครื่องยนต์ติด และ engine_flag = 0 หมายถึงเครื่องยนต์ดับ นอกจากนี้ยังต้องมีเงื่อนไขอื่นดังนี้

- ตลอดการเดินทางสถานะเครื่องยนต์ต้องเป็น 1 ตลอดเวลา
- ข้อมูลการเดินทางจะต้องไม่มีข้อมูลที่สูญหาย โดยพิจารณาจากข้อมูลที่ติดกัน จะต้องมีเวลาห่างกันไม่เกิน 3 นาที
- ข้อมูลการเดินทางนั้นต้องมีระยะห่างระหว่างจุดเริ่มต้นและจุดสิ้นสุดมากกว่า 1 กิโลเมตร
- ข้อมูลการเดินทางจะต้องมีข้อมูลมากกว่า 3 บันทึก
- ข้อมูลการเดินทางที่ได้จะต้องไม่มีข้อมูลพิกัดที่ผิดพลาด โดยพิจารณาจากความเร็ว ที่คำนวณจากระยะห่างและเวลาของข้อมูล โดยเวลาที่ติดกันจะต้องไม่มากกว่า 150 กิโลเมตรต่อชั่วโมง
- เวลารวมตลอดทั้งการเดินทางต้องไม่มากกว่า 6 ชั่วโมง เพื่อทำการกรองข้อมูลการเดินทางที่เดินทางไปต่างจังหวัดหรือเกิดข้อผิดพลาดของสัญญาณแสดงสถานะ ผู้โดยสาร

จากเงื่อนไขเบื้องต้น ได้ผลลัพธ์สถานะการการรับส่งผู้โดยสารดังนี้

- tipstatus = O (engine_flag = 1 and passenger_flag = 0) หมายถึพิกัดแรกที่มีผู้โดยสารอยู่บนรถ
- tipstatus = D (engine_flag = 1 and passenger_flag = 1) หมายถึพิกัดสุดท้ายที่มีผู้โดยสารอยู่บนรถ



ข้อมูลที่ใช้ในการทดลองในงานวิจัยนี้ เป็นข้อมูลการรับส่งผู้โดยสารของรถแท็กซี่ซึ่งผ่านการสกัดจากเงื่อนไขที่กล่าวเบื้องต้น โดยเก็บข้อมูลของรถแท็กซี่ที่ให้บริการในเขตกรุงเทพมหานครจำนวน 2,375 คัน ระยะเวลา 8 เดือน (กุมภาพันธ์ – กันยายน 2559) จำนวน 128,082,024 รายการ ฟิวด์ข้อมูลและตัวอย่างการเก็บข้อมูลแสดงในตารางที่ 2

ตารางที่ 2 ฟิวด์ข้อมูลการรับส่งผู้โดยสารของรถแท็กซี่

| ฟิวด์ | คำอธิบาย |
|---|---|
| vehicle_id | รหัสของรถแท็กซี่ |
| tripid | รหัสของเที่ยวโดยสารของรถแท็กซี่ |
| latitude | พิกัดละติจูด |
| longitude | พิกัดลองจิจูด |

| | |
|---|---|
| gps_timestamp | วันและเวลาที่บันทึก |
| tripstatus | สถานะการรับส่งผู้โดยสารของรถแท็กซี่ ประกอบด้วย O = Origin (จุดรับ) D = Destination (จุดส่ง) |

Filter

| | vehicle_id | tripid | latitude | longitude | gps_timestamp | tripstatus |
|---|---|---|---|---|---|---|
| 1 | AK3954 | AK3954_T_3608 | 13.84314 | 100.4940 | 2016-08-24 11:46:36 | D |
| 2 | AK3954 | AK3954_T_3610 | 13.86316 | 100.5009 | 2016-08-24 11:57:36 | O |
| 3 | AK3954 | AK3954_T_3610 | 13.85978 | 100.5183 | 2016-08-24 12:29:36 | D |
| 4 | AK3954 | AK3954_T_3612 | 13.85308 | 100.5251 | 2016-08-24 12:40:36 | O |
| 5 | AT1982 | AT1982_T_4962 | 13.78442 | 100.5131 | 2016-08-04 11:59:35 | O |
| 6 | AT1982 | AT1982_T_4962 | 13.77580 | 100.5293 | 2016-08-04 12:09:35 | D |
| 7 | AT1982 | AT1982_T_4964 | 13.72819 | 100.5809 | 2016-08-04 12:50:35 | O |
| 8 | AT1982 | AT1982_T_4964 | 13.73103 | 100.5854 | 2016-08-04 12:54:35 | D |
| 9 | AT1982 | AT1982_T_4967 | 13.72746 | 100.5738 | 2016-08-04 13:13:23 | O |
| 10 | AT1982 | AT1982_T_4967 | 13.72812 | 100.5733 | 2016-08-04 13:15:23 | D |
| 11 | AT1982 | AT1982_T_4969 | 13.72727 | 100.5231 | 2016-08-04 14:53:25 | O |
| 12 | AT1982 | AT1982_T_4969 | 13.72638 | 100.5280 | 2016-08-04 15:03:25 | D |
| 13 | AT1982 | AT1982_T_4971 | 13.74579 | 100.5630 | 2016-08-04 15:52:25 | O |
| 14 | AT1982 | AT1982_T_4971 | 13.74803 | 100.5635 | 2016-08-04 15:56:25 | D |
| 15 | AT1982 | AT1982_T_4973 | 13.75014 | 100.5728 | 2016-08-04 16:06:25 | O |
| 16 | AK3615 | AK3615_T_6586 | 13.95200 | 100.5474 | 2016-09-23 10:06:32 | D |
| 17 | AK3615 | AK3615_T_6587 | 13.95191 | 100.5477 | 2016-09-23 10:14:18 | O |
| 18 | AK3615 | AK3615_T_6587 | 13.99145 | 100.5814 | 2016-09-23 10:32:18 | D |
| 19 | AK3615 | AK3615_T_6588 | 13.99145 | 100.5815 | 2016-09-23 11:01:56 | O |
| 20 | AK3615 | AK3615_T_6588 | 13.98975 | 100.6157 | 2016-09-23 11:28:56 | D |
| 21 | AK3615 | AK3615_T_6590 | 13.98666 | 100.6747 | 2016-09-23 11:46:56 | O |
| 22 | AK3615 | AK3615_T_6590 | 14.06752 | 100.6021 | 2016-09-23 12:30:56 | D |
| 23 | AK3615 | AK3615_T_6594 | 14.00086 | 100.6885 | 2016-09-23 13:04:56 | O |
| 24 | AK3615 | AK3615_T_6594 | 14.06547 | 100.6476 | 2016-09-23 13:36:29 | D |
| 25 | AK3615 | AK3615_T_6596 | 14.11782 | 100.6183 | 2016-09-23 13:48:29 | O |
| 26 | AK3615 | AK3615_T_6596 | 14.05834 | 100.6174 | 2016-09-23 14:04:29 | D |
| 27 | AK3615 | AK3615_T_6598 | 14.05655 | 100.6126 | 2016-09-23 14:45:29 | O |

ภาพที่ 11 แสดงตัวอย่างรายการข้อมูลการรับส่งผู้โดยสารของรถแท็กซี่

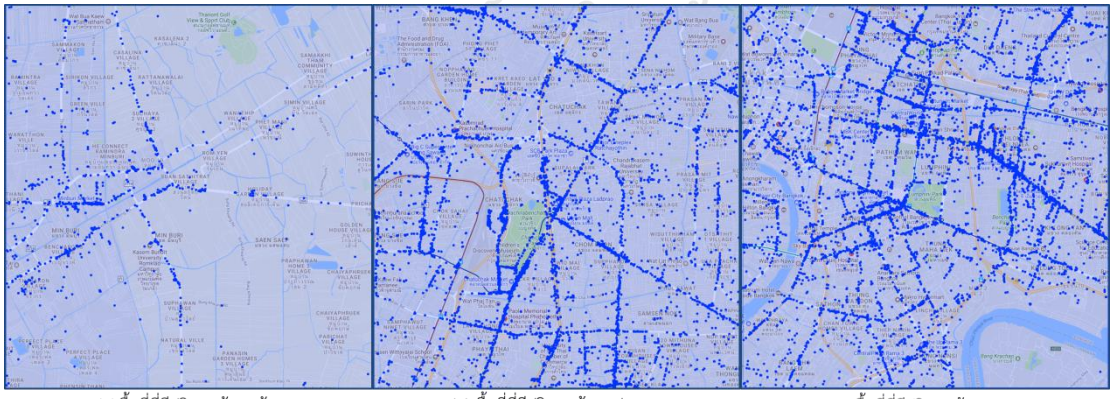
**4.3 ผลการทดลองการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีความหนาแน่นและการกระจายข้อมูลที่แตกต่างกัน**

งานวิจัยทำการทดลองโดยใช้ข้อมูลการรับส่งผู้โดยสารจำนวน 1 วัน โดยพิจารณาเลือกจากปริมาณข้อมูลที่มากที่สุดในพื้นที่ที่ทำการทดลองจากจำนวนข้อมูลระยะเวลา 9 เดือน เพื่อกำหนดพารามิเตอร์แบบอัตโนมัติสำหรับใช้แบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น (DBSCAN) โดยพื้นที่ตัวอย่างมีความหนาแน่นและการกระจายข้อมูลที่แตกต่างกัน สามารถแบ่งระดับการกระจายข้อมูลออกเป็น 2 ประเภทคือ การกระจายตัวและความหนาแน่นระดับเดียว (Single-density distribution) และการกระจายตัวและความหนาแน่นหลายระดับ (Multi-density distributions) พื้นที่ตัวอย่างสำหรับการทดลองแสดงดังต่อไปนี้

1) พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว (Low population and single-density distribution) ประกอบด้วยจำนวนข้อมูลการรับส่งผู้โดยสาร 640 - 882 เที่ยว มีปริมาณมากที่สุดในวันอังคาร เดือนกรกฎาคม

2) พื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ (Medium population and multi-density distribution) ประกอบด้วยจำนวนข้อมูลการรับส่งผู้โดยสาร 4,672– 5,657 เที่ยว มีปริมาณมากที่สุดในวันเสาร์ เดือนมีนาคม

3) พื้นที่ที่มีปริมาณข้อมูลมาก การกระจายตัวและความหนาแน่นหลายระดับ (High population and multi-density distribution) ประกอบด้วยจำนวนข้อมูลการรับส่งผู้โดยสาร 6,260 – 6,534 เที่ยว มีปริมาณมากที่สุดในวันพฤหัสบดี เดือนมีนาคม

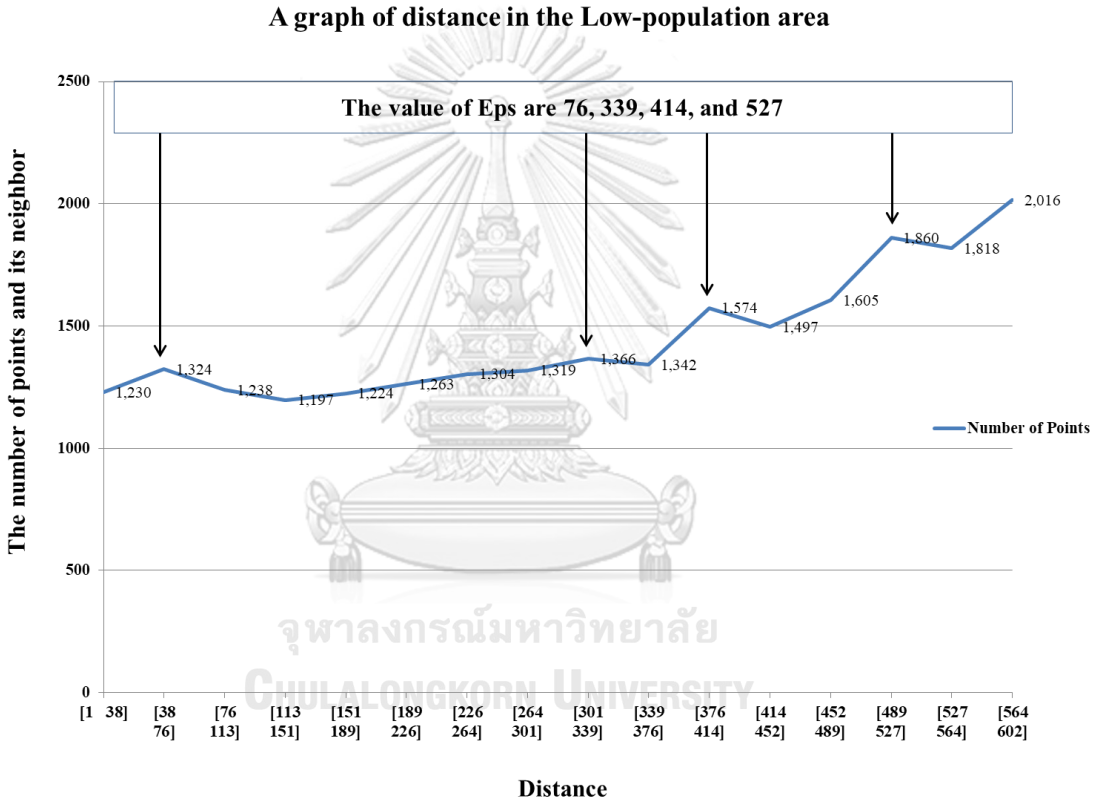


(ก)พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายและความหนาแน่นระดับเดียว | (ข) พื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายและความหนาแน่นหลายระดับ | (ค) พื้นที่ที่มีปริมาณข้อมูลมาก การกระจายและความหนาแน่นหลายระดับ

ภาพที่ 12 แสดงตัวอย่างการกระจายตัวของข้อมูลการรับส่งผู้โดยสารในพื้นที่ที่แตกต่างกัน

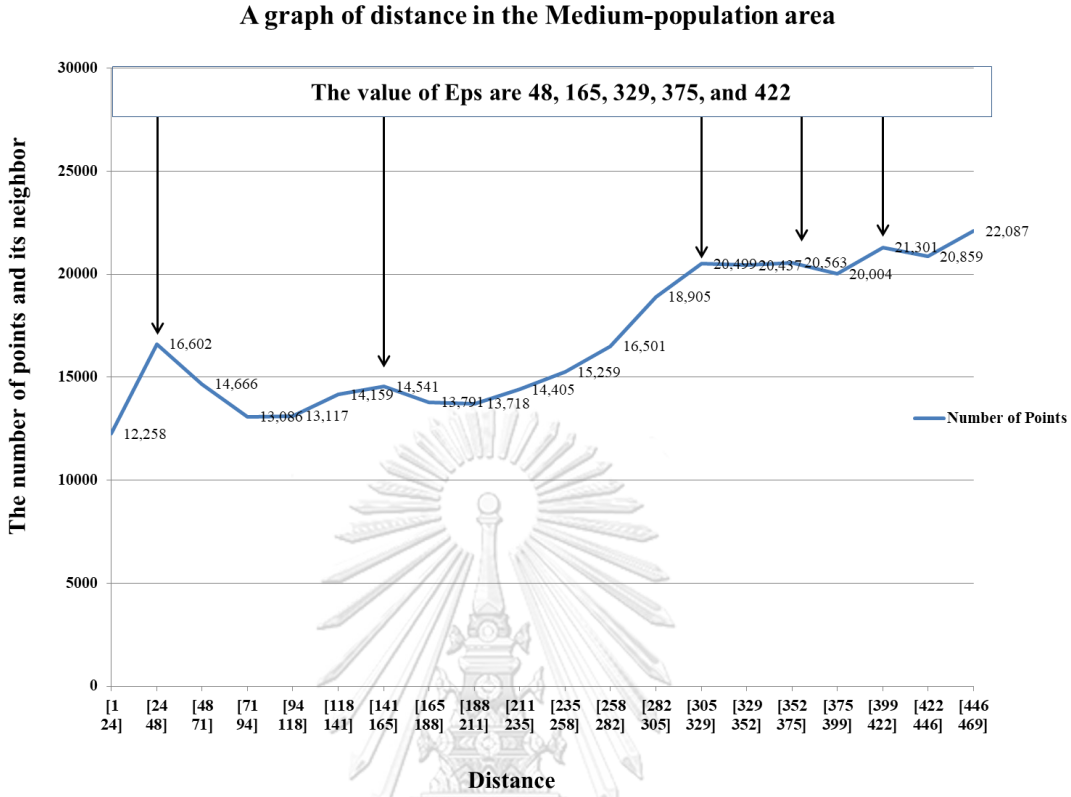
จากวิธีการกำหนดพารามิเตอร์อัตโนมัติในบทที่ 3 สามารถกำหนดพารามิเตอร์คือ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และจำนวนจุดขั้นต่ำสุดสำหรับกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) เพื่อแบ่งกลุ่มข้อมูล กราฟแสดงการหารัศมีระหว่างจุดในกลุ่มข้อมูลในพื้นที่ต่างๆ แสดงดังในภาพที่ 13 โดย (ก) และ (ข) ลักษณะกราฟสามารถระบุความแตกต่างของระยะทางที่ชัดเจน จึงพิจารณาเลือกค่ารัศมีระหว่างจุดในกลุ่มข้อมูลจากจุดที่สูงที่สุดในแต่ละช่วงข้อมูลที่ทำให้กราฟมีการเปลี่ยนแปลง ส่วน (ค) ความแตกต่างของระยะทางมีความลาดชันขึ้นเรื่อย ไม่สามารถระบุจุดเปลี่ยนแปลงของกราฟอย่างชัดเจน จึงพิจารณาเลือกรัศมีระหว่างจุดในกลุ่มข้อมูลจากช่วงระยะทางทั้งหมด

ค่ารัศมีระหว่างจุดในกลุ่มข้อมูล บนพื้นที่ที่มีปริมาณข้อมูลน้อย คือ 76 339 414 และ 527



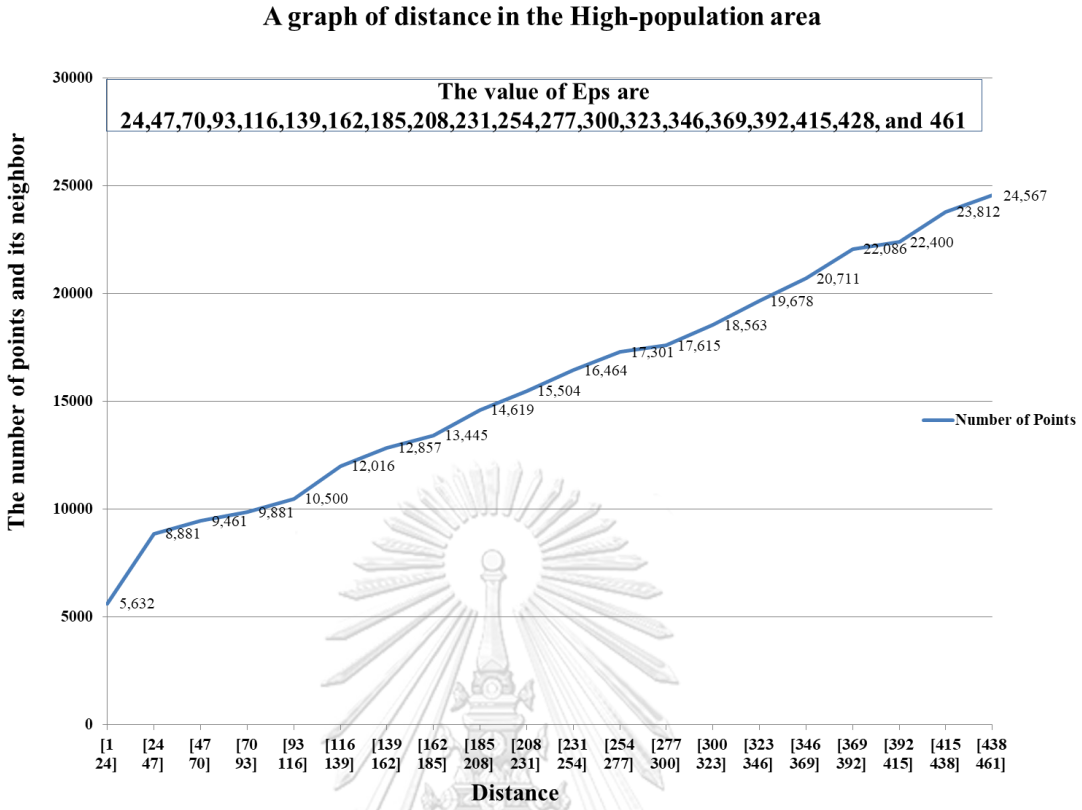

(ก) กราฟแสดงการหาค่ารัศมีของกลุ่ม (Eps) พื้นที่ที่มีปริมาณข้อมูลน้อย

ค่ารัศมีระหว่างจุดในกลุ่มข้อมูล บนพื้นที่ที่มีปริมาณข้อมูลปานกลาง คือ 48 165 329 375 และ 422




(ข) กราฟแสดงการหาค่ารัศมีของกลุ่ม (Eps) พื้นที่ที่มีปริมาณข้อมูลปานกลาง

ค่ารัศมีระหว่างจุดในกลุ่มข้อมูล บนพื้นที่ที่มีปริมาณข้อมูลมาก คือ 24 47 70 93 116 139 162 185 208 231 254 277 300 323 346 369 392 415 428 และ 461




(ค) กราฟแสดงการหาค่ารัศมีของกลุ่ม (Eps) พื้นที่ที่มีปริมาณข้อมูลมาก

ภาพที่ 13 กราฟแสดงค่ารัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) ในแต่ละพื้นที่

โดยผลลัพธ์ที่ได้จากการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่นได้ดังนี้

ตารางที่ 3 ตารางแสดงผลการแบ่งกลุ่มข้อมูลด้วย DBSCAN โดยใช้พารามิเตอร์จากวิธีที่นำเสนอ

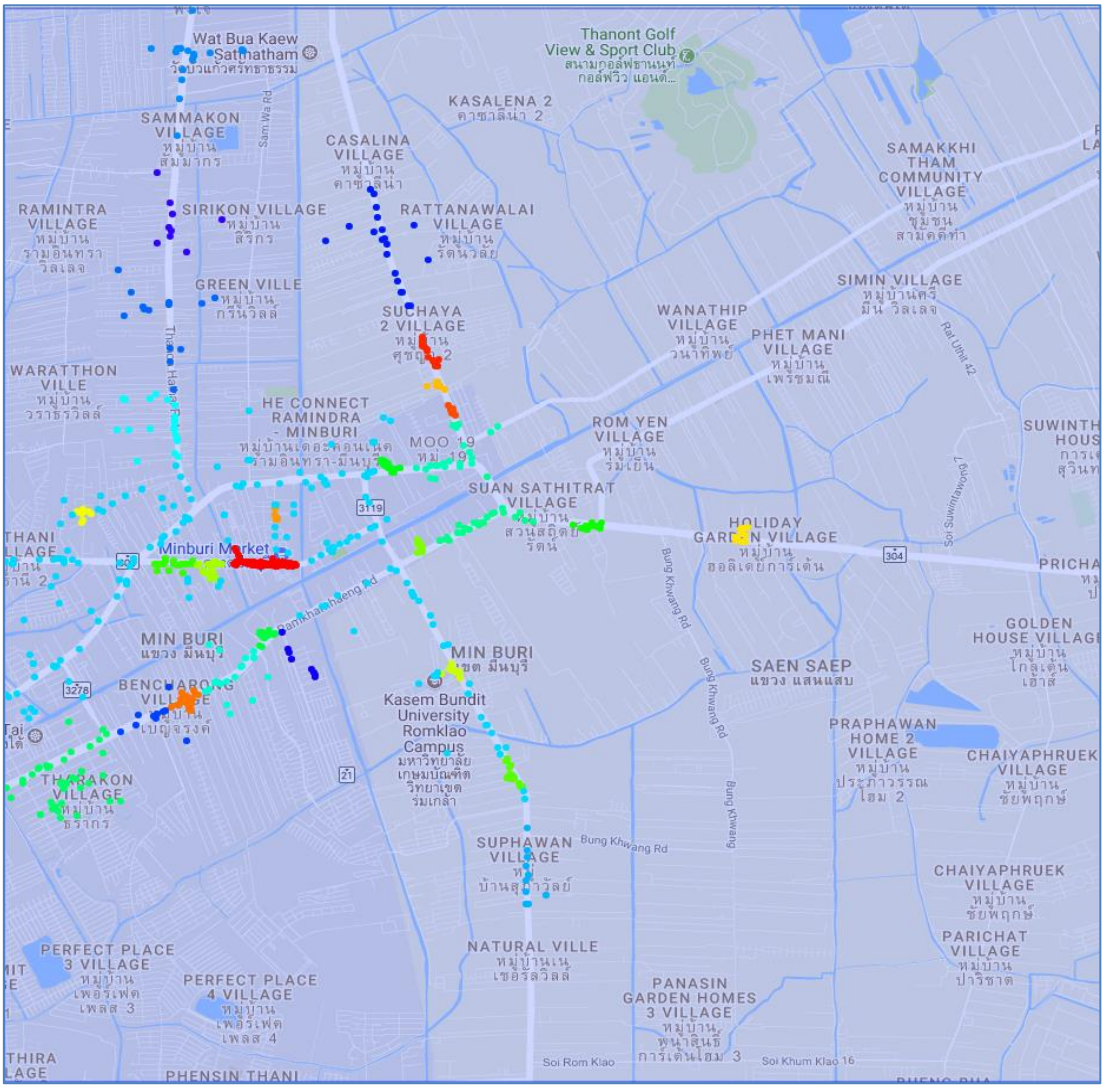
| พื้นที่ | จำนวนข้อมูลรับ-ส่ง ผู้โดยสาร | Eps (เมตร) | MinPts (Points) | Cluster | Noise (Points) |
|---|---|---|---|---|---|
| พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว | 882 | 76 | 8 | 16 | 554 |
| | 554 | 339 | 19 | 5 | 419 |
| | 419 | 414 | 9 | 8 | 154 |
| | 154 | 527 | 11 | 0 | 154 |
| พื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ | 5,657 | 48 | 21 | 18 | 4,432 |
| | 4,432 | 165 | 31 | 18 | 3,270 |
| | 3,270 | 329 | 57 | 1 | 3,213 |
| | 3,213 | 375 | 24 | 18 | 351 |

| | | | | | |
|---|---|---|---|---|---|
| | 351 | 422 | 32 | 0 | 351 |
| พื้นที่ที่มีปริมาณข้อมูลมาก การกระจายตัวและความหนาแน่นหลายระดับ | 6,534 | 24 | 7 | 66 | 5,650 |
| | 5,650 | 47 | 7 | 127 | 4,111 |
| | 4,111 | 70 | 6 | 148 | 2,782 |
| | 2,782 | 93 | 7 | 32 | 2,519 |
| | 2,519 | 116 | 7 | 48 | 2,124 |
| | 2,124 | 139 | 9 | 7 | 2,057 |
| | 2,057 | 162 | 8 | 40 | 1,679 |
| | 1,679 | 185 | 7 | 60 | 1,117 |
| | 1,117 | 208 | 8 | 2 | 1100 |
| | 1,100 | 231 | 9 | 6 | 1,046 |
| | 1,046 | 254 | 10 | 3 | 1,008 |
| | 1,008 | 277 | 9 | 23 | 746 |
| | 746 | 300 | 8 | 19 | 565 |
| | 565 | 323 | 8 | 9 | 490 |
| | 490 | 346 | 10 | 0 | 490 |
| | 490 | 369 | 17 | 0 | 490 |
| | 490 | 392 | 15 | 0 | 490 |
| | 490 | 415 | 9 | 10 | 373 |
| | 373 | 438 | 11 | 0 | 373 |
| | 373 | 461 | 10 | 3 | 345 |

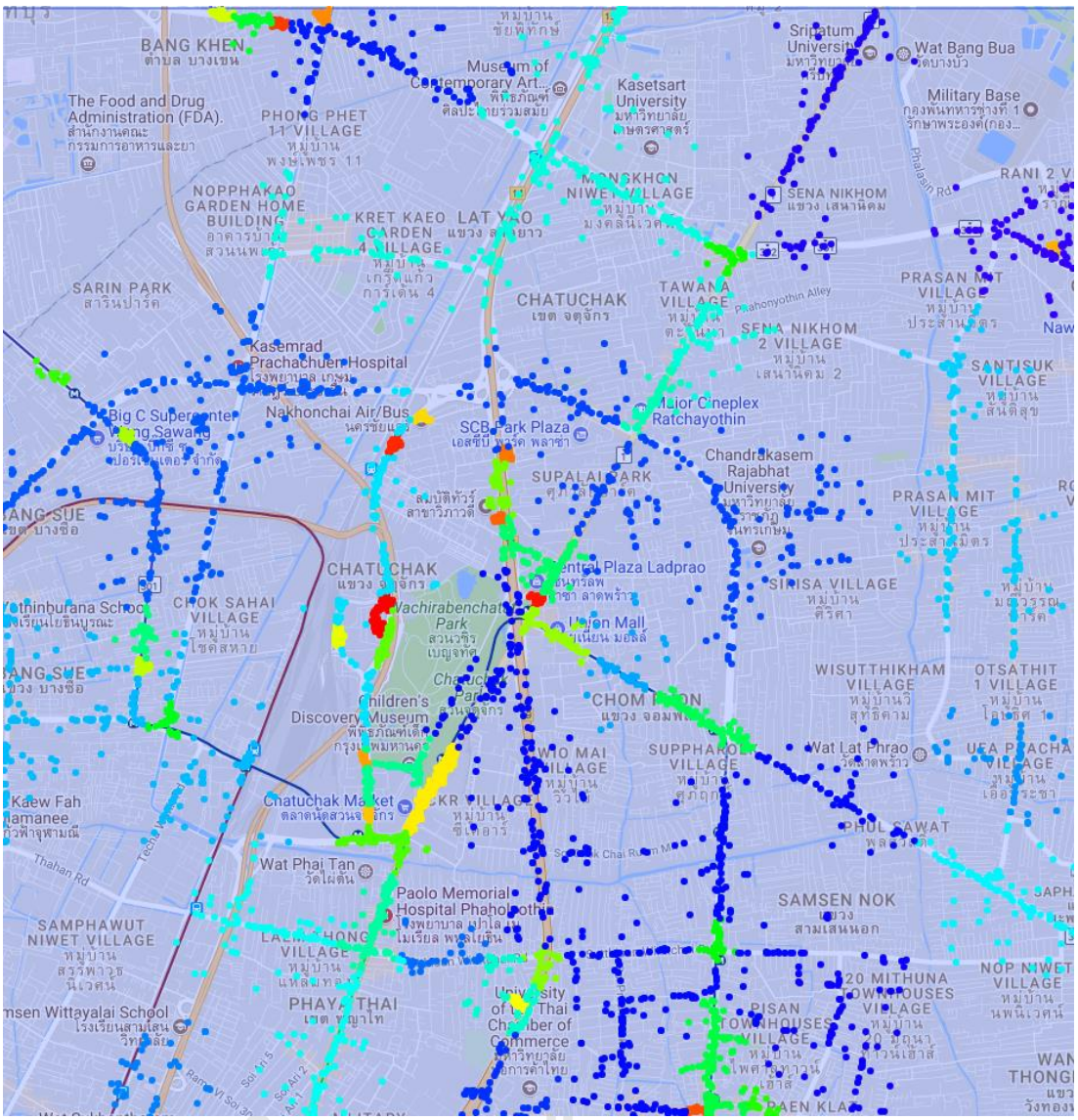


จากผลการทดลองในตารางที่ 3 แสดงให้เห็นว่า แต่ละพื้นที่มีรัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) หลายค่า เนื่องจาก ลักษณะของความหนาแน่นของข้อมูลมีการกระจายตัวที่แตกต่างกัน พื้นที่ที่จัดว่าเป็นจุดที่น่าสนใจมีลักษณะที่หลากหลาย กระจัดกระจาย บางบริเวณข้อมูลกระจุกตัวเป็นกลุ่มๆ บางบริเวณมีการกระจายตัวของข้อมูลเป็นพื้นที่กว้าง ยกตัวอย่างภาพที่ 14 หากดูตามภาพที่แสดงจะพบว่า บริเวณที่เป็นกลุ่มพื้นที่ที่มีความหนาแน่นนั้นเป็นตลาด และบริเวณใกล้เคียงข้อมูลเกาะกลุ่มกันอย่างหนาแน่น แต่ในพื้นที่อื่นๆ ลักษณะข้อมูลในคลัสเตอร์มีการกระจายตัวมากขึ้น สันนิษฐานได้ว่า บริเวณดังกล่าว อาจมีพื้นที่ที่เป็นจุดสนใจอยู่อย่างกระจัดกระจาย ในภาพที่ 15 ซึ่งเป็นผลการแบ่งข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นของ

ข้อมูลหลายระดับ พบว่าค่ารัศมีระหว่างจุดในกลุ่มข้อมูลมีระยะที่สั้นลง นั่นหมายถึงว่า กลุ่มข้อมูลมีการกระจุกตัวมากยิ่งขึ้น และในภาพที่ 16 รัศมีระหว่างจุดในกลุ่มข้อมูลลดลงกว่าในภาพที่ 15 แสดงให้เห็นว่าพื้นที่นั้นมีความหนาแน่นของจุดสนใจ คลัสเตอร์ที่ได้จึงมีลักษณะเป็นหย่อมๆ อยู่ใกล้กัน สามารถแสดงได้ว่าพื้นที่ดังกล่าวเป็นย่านกลางเมือง พื้นที่จุดสนใจจะมีรัศมีขนาดเล็ก แต่มีจำนวนมาก

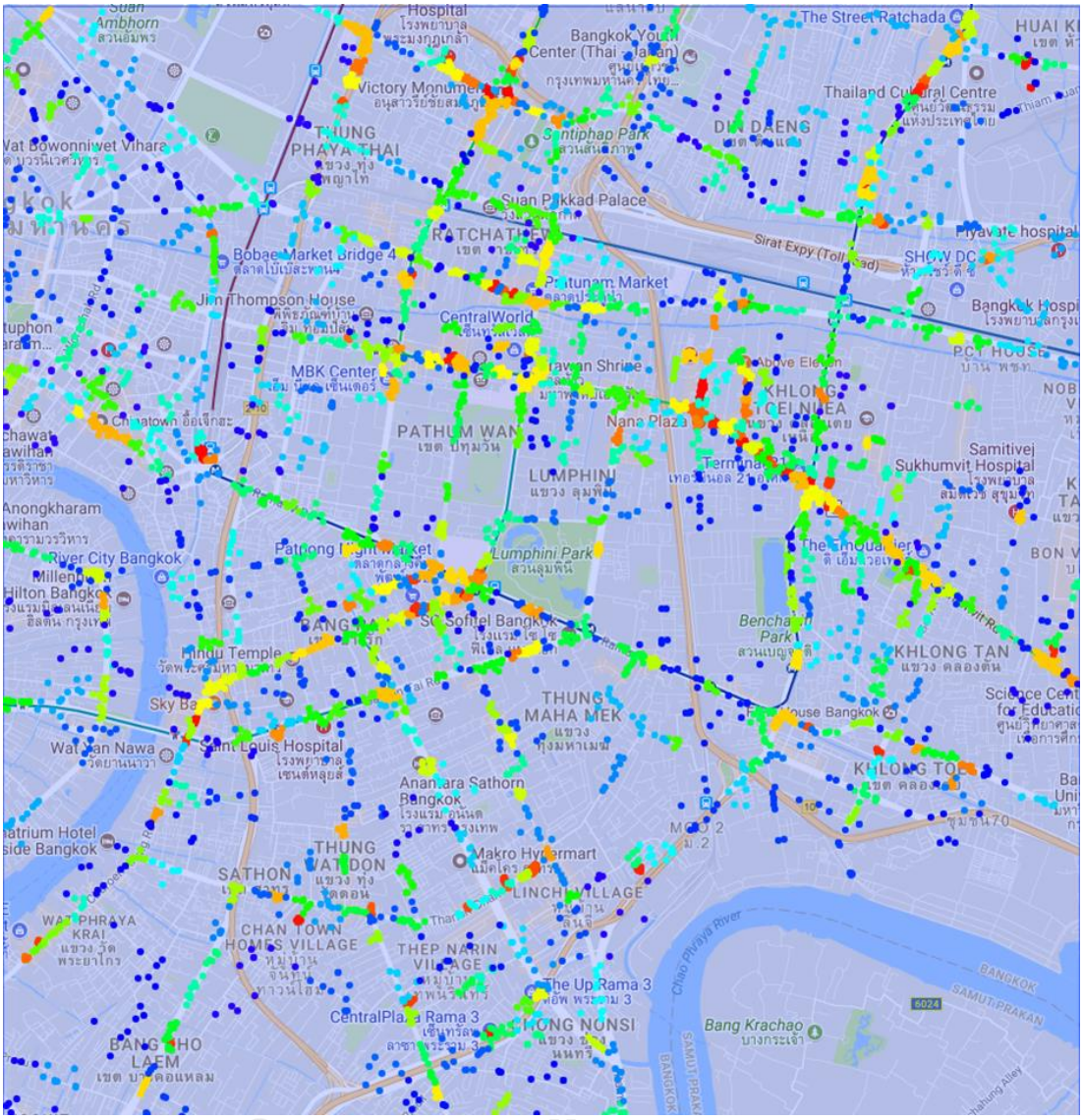



ภาพที่ 14 แสดงผลการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว






ภาพที่ 15 แสดงผลการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ






ภาพที่ 16 แสดงผลการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลมาก และการกระจายตัวและความหนาแน่นหลายระดับ

### 4.4 ผลการทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก DMDBSCAN

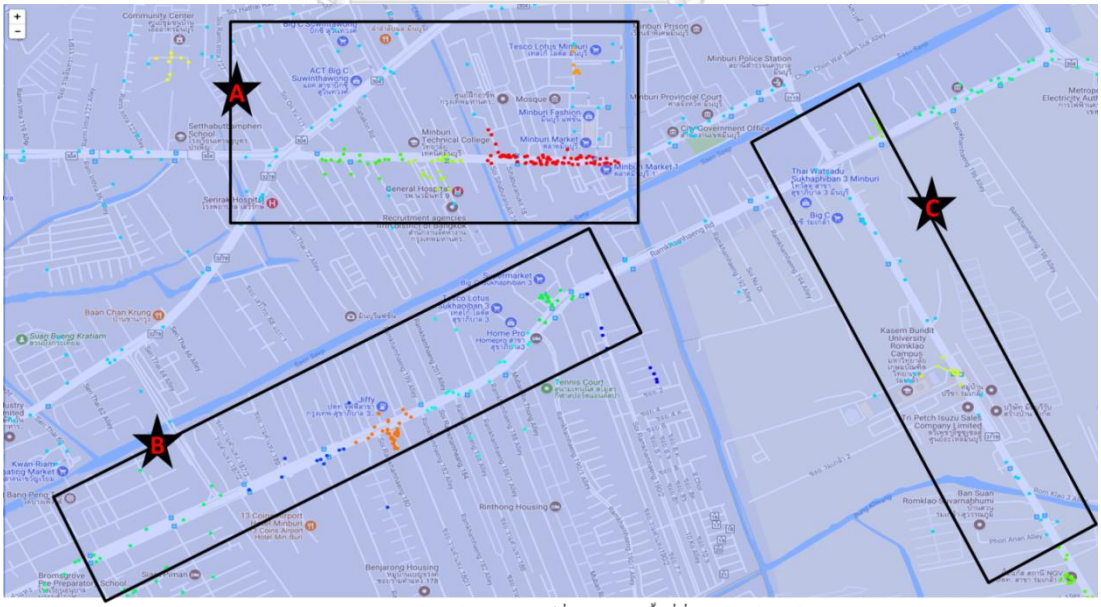
การทดลองเพื่อเปรียบเทียบผลการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น โดยใช้พารามิเตอร์จากวิธีที่นำเสนอและการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จากการหารัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) จากจำนวนจุดขั้นต่ำสุดสำหรับกำหนดจุดศูนย์กลางของกลุ่ม (MinPts) ที่ถูกกำหนดไว้ล่วงหน้า (A Dynamic Method for Discovery Density Varied Clusters: DMDBSCAN) โดยใช้เงื่อนไขการทดลอง ตามข้อ 4.3 ผลการทดลองแสดงดังต่อไปนี้

ตารางที่ 4 ตารางแสดงผลการเปรียบเทียบการแบ่งกลุ่มข้อมูลด้วย DBSCAN โดยใช้พารามิเตอร์จากวิธีที่นำเสนอและ DMDBSCAN

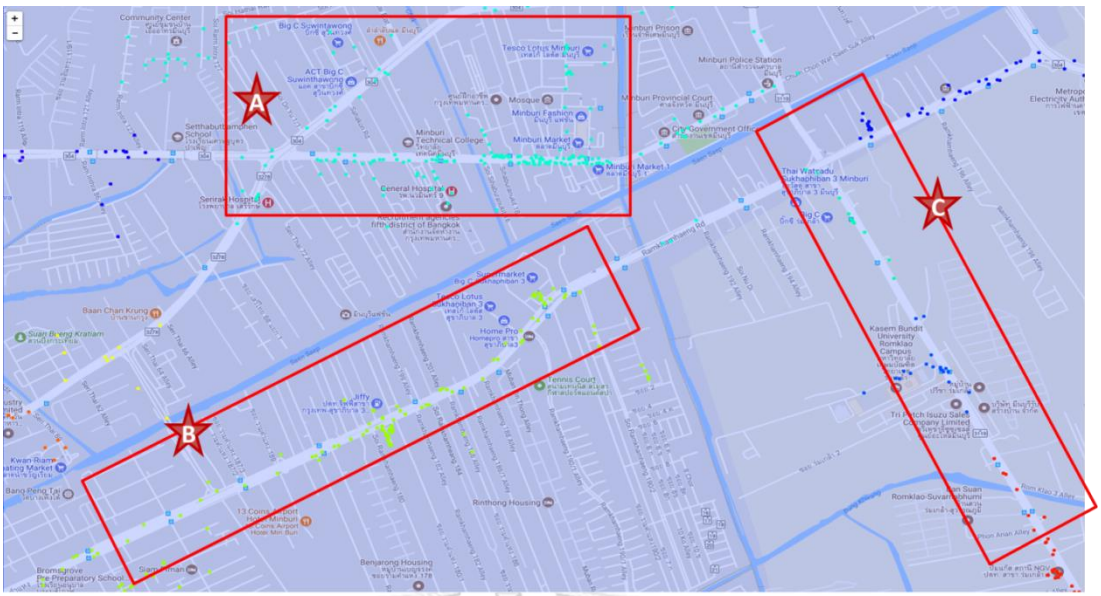
| พื้นที่ | วิธีการ | จำนวนข้อมูลรับ-ส่งผู้โดยสาร | Eps (เมตร) | MinPts (Points) | Cluster | Noise (Points) |
|---|---|---|---|---|---|---|
| พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 882 | 76 | 8 | 16 | 554 |
| | | 554 | 339 | 19 | 5 | 419 |
| | | 419 | 414 | 9 | 8 | 154 |
| | | 154 | 527 | 11 | 0 | 154 |
| | DMDBSCAN | 6,111 | 1,067 | 3 | 34 | 133 |
| พื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 5,657 | 48 | 21 | 18 | 4,432 |
| | | 4,432 | 165 | 31 | 18 | 3,270 |
| | | 3,270 | 329 | 57 | 1 | 3,213 |
| | | 3,213 | 375 | 24 | 18 | 351 |
| | | 351 | 422 | 32 | 0 | 351 |
| | DMDBSCAN | 5,657 | 546 | 3 | 61 | 169 |
| พื้นที่ที่มีปริมาณข้อมูลมาก การกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 6,534 | 24 | 7 | 66 | 5,650 |
| | | 5,650 | 47 | 7 | 127 | 4,111 |
| | | 4,111 | 70 | 6 | 148 | 2,782 |
| | | 2,782 | 93 | 7 | 32 | 2,519 |
| | | 2,519 | 116 | 7 | 48 | 2,124 |
| | | 2,124 | 139 | 9 | 7 | 2,057 |
| | | 2,057 | 162 | 8 | 40 | 1,679 |
| | | 1,679 | 185 | 7 | 60 | 1,117 |
| | | 1,117 | 208 | 8 | 2 | 1100 |
| | | 1,100 | 231 | 9 | 6 | 1,046 |
| | | 1,046 | 254 | 10 | 3 | 1,008 |
| | | 1,008 | 277 | 9 | 23 | 746 |
| | | 746 | 300 | 8 | 19 | 565 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 565 | 323 | 8 | 9 | 490 |
| | | 490 | 346 | 10 | 0 | 490 |
| | | 490 | 369 | 17 | 0 | 490 |
| | | 490 | 392 | 15 | 0 | 490 |
| | | 490 | 415 | 9 | 10 | 373 |
| | | 373 | 438 | 11 | 0 | 373 |
| | | 373 | 461 | 10 | 3 | 345 |
| | DMDBSCAN | 6,534 | 559 | 3 | 69 | 150 |

จากผลการทดลองดังปรากฏในตารางที่ 4 และภาพที่ 17 ภาพที่ 18 และภาพที่ 19 พบว่า การแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่นำเสนอ สามารถแบ่งข้อมูลได้เป็นกลุ่มๆ และมีความละเอียด ระบุเป็นกลุ่มพื้นที่ได้ชัดเจน ในขณะที่ผลลัพธ์ที่ได้จากการแบ่งกลุ่มข้อมูลด้วย DMDBSCAN มีความละเอียดน้อยกว่า เมื่อข้อมูลมีความหนาแน่นมากๆ พื้นที่ที่ได้จะมีลักษณะติดต่อกันขนาดใหญ่ ไม่สามารถระบุพื้นที่แยกจากการชัดเจน ดังพื้นที่ A B และ C ในภาพที่ 17 (ข) ดังนั้น พบว่า การสำรวจหาบริเวณที่เป็นจุดสนใจด้วยวิธีที่นำเสนอในงานวิจัยนี้ มีความเหมาะสม และสามารถระบุขอบเขตบริเวณได้ชัดเจน และมีความละเอียด แม้มีปริมาณการกระจายตัวของข้อมูลอย่างหนาแน่น

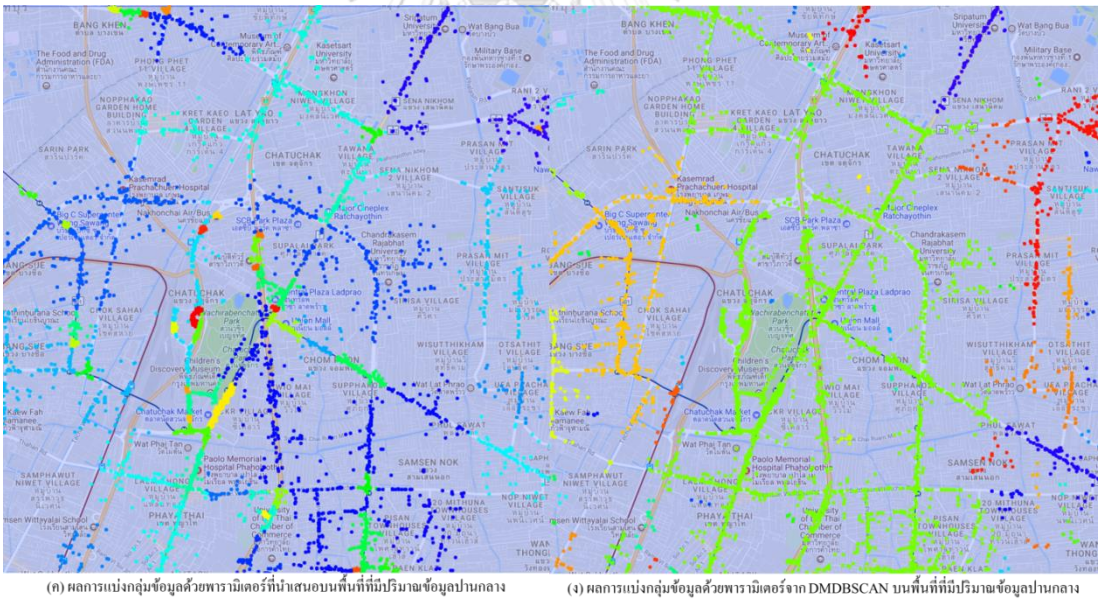



(ก) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่นำเสนอบนพื้นที่ที่มีปริมาณข้อมูลน้อย



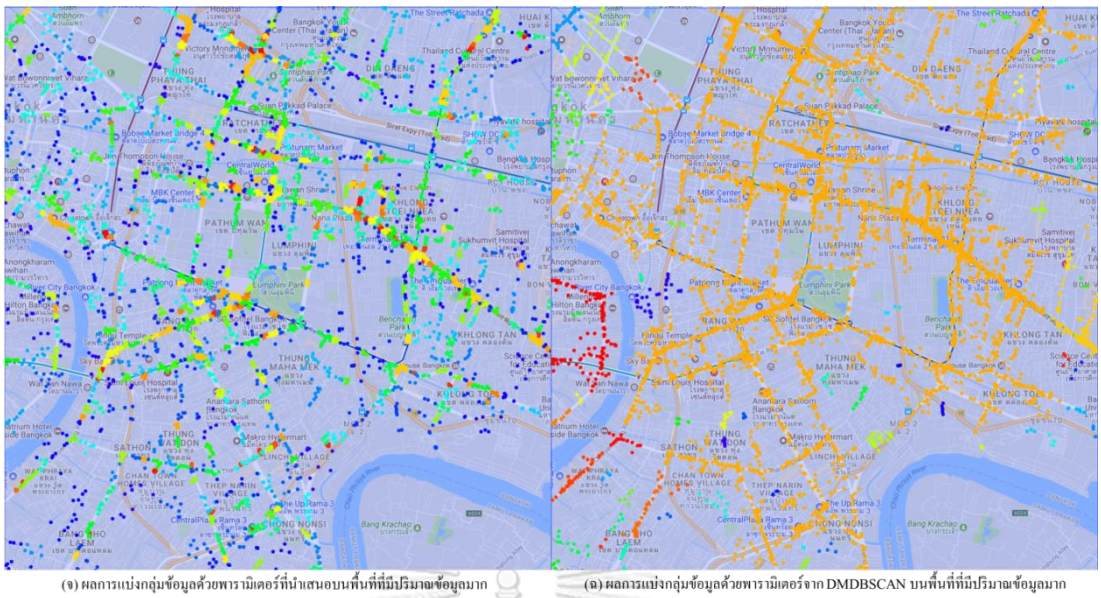

(ข) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์จาก DMDBSCAN บนพื้นที่ที่มีปริมาณข้อมูลน้อย

ภาพที่ 17 แสดงผลเปรียบเทียบการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว ระหว่างพารามิเตอร์ที่ได้จากวิธีที่นำเสนอและ DMDBSCAN



(ค) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่นำเสนอบนพื้นที่ที่มีปริมาณข้อมูลปานกลาง (ง) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์จาก DMDBSCAN บนพื้นที่ที่มีปริมาณข้อมูลปานกลาง

ภาพที่ 18 แสดงผลเปรียบเทียบการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ ระหว่างพารามิเตอร์ที่ได้จากวิธีที่นำเสนอและ DMDBSCAN



(จ) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่นำเสนอบนพื้นที่ที่มีปริมาณข้อมูลมาก (ฉ) ผลการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์จาก DMDBSCAN บนพื้นที่ที่มีปริมาณข้อมูลมาก

ภาพที่ 19 แสดงผลเปรียบเทียบการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลมาก การกระจายตัวและความหนาแน่นหลายระดับ ระหว่างพารามิเตอร์ที่ได้จากวิธีที่นำเสนอและ DMDBSCAN

### 4.5 ผลการทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก VDBSCAN

การทดลองเพื่อเปรียบเทียบผลการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่นโดยใช้พารามิเตอร์จากวิธีที่นำเสนอและการแบ่งกลุ่มข้อมูลเชิงพื้นที่ที่สามารถทำงานได้อย่างมีประสิทธิภาพเมื่อข้อมูลมีความหนาแน่นที่แตกต่างกัน (Varied Density Based Spatial Clustering of Applications with Noise: VDBSCAN) การทดลองใช้พื้นที่ตามข้อ 4.3 และแบ่งข้อมูลเพื่อใช้กำหนดพารามิเตอร์เป็น 2 มิติ กำหนดจำนวน MinPts = 3 ผลการทดลองแสดงดังต่อไปนี้

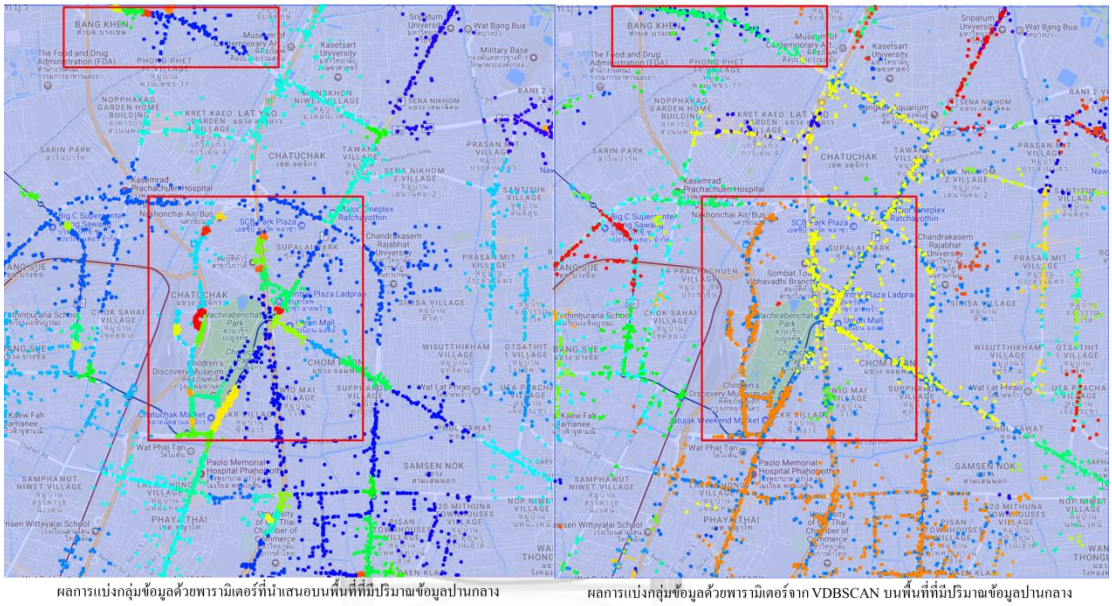
ตารางที่ 5 ตารางแสดงผลการเปรียบเทียบการแบ่งกลุ่มข้อมูลด้วย DBSCAN โดยใช้พารามิเตอร์จากวิธีที่นำเสนอและ VDBSCAN

| พื้นที่ | วิธีการ | จำนวนข้อมูลรับ-ส่งผู้โดยสาร | Eps (เมตร) | MinPts (Points) | Cluster | Noise (Points) |
|---|---|---|---|---|---|---|
| พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัว | การหาพารามิเตอร์แบบอัตโนมัติที่ | 882 | 76 | 8 | 16 | 554 |
| | | 554 | 339 | 19 | 5 | 419 |
| | | 419 | 414 | 9 | 8 | 154 |

| | | | | | | |
|---|---|---|---|---|---|---|
| และความหนาแน่นระดับเดียว | นำเสนอในงานวิจัยนี้ | 154 | 527 | 11 | 0 | 154 |
| | VDBSCAN | 441 | 974 | 3 | 18 | 54 |
| | | 441 | 1,282 | 3 | 24 | 84 |
| พื้นที่ที่มีปริมาณข้อมูลปานกลางการกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 5,657 | 48 | 21 | 18 | 4,432 |
| | | 4,432 | 165 | 31 | 18 | 3,270 |
| | | 3,270 | 329 | 57 | 1 | 3,213 |
| | | 3,213 | 375 | 24 | 18 | 351 |
| | | 351 | 422 | 32 | 0 | 351 |
| | VDBSCAN | 2,828 | 711 | 3 | 38 | 85 |
| | | 2,829 | 754 | 3 | 53 | 112 |
| พื้นที่ที่มีปริมาณข้อมูลมากการกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 6,534 | 24 | 7 | 66 | 5,650 |
| | | 5,650 | 47 | 7 | 127 | 4,111 |
| | | 4,111 | 70 | 6 | 148 | 2,782 |
| | | 2,782 | 93 | 7 | 32 | 2,519 |
| | | 2,519 | 116 | 7 | 48 | 2,124 |
| | | 2,124 | 139 | 9 | 7 | 2,057 |
| | | 2,057 | 162 | 8 | 40 | 1,679 |
| | | 1,679 | 185 | 7 | 60 | 1,117 |
| | | 1,117 | 208 | 8 | 2 | 1100 |
| | | 1,100 | 231 | 9 | 6 | 1,046 |
| | | 1,046 | 254 | 10 | 3 | 1,008 |
| | | 1,008 | 277 | 9 | 23 | 746 |
| | | 746 | 300 | 8 | 19 | 565 |
| | | 565 | 323 | 8 | 9 | 490 |
| | | 490 | 346 | 10 | 0 | 490 |
| | | 490 | 369 | 17 | 0 | 490 |
| | | 490 | 392 | 15 | 0 | 490 |
| | | 490 | 415 | 9 | 10 | 373 |
| | | 373 | 438 | 11 | 0 | 373 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 373 | 461 | 10 | 3 | 345 |
| | VDBSCAN | 3,267 | 608 | 3 | 29 | 49 |
| | | 3,267 | 715 | 3 | 48 | 80 |

ตัวอย่างผลการเปรียบเทียบการแบ่งข้อมูลในพื้นที่ที่มีข้อมูลปริมาณปานกลาง ดังภาพที่ 20 พบว่าผลการแบ่งกลุ่มข้อมูลมีความคล้ายกันในบางบริเวณ แต่กลุ่มข้อมูลที่แบ่งกลุ่มด้วยพารามิเตอร์ที่นำเสนอสามารถให้กลุ่มข้อมูลที่มีขนาดเล็ก เป็นกลุ่มย่อยๆ ได้ดีกว่าการแบ่งกลุ่มข้อมูลด้วยเทคนิคของ VDBSCAN แม้บริเวณพื้นที่ที่มีความหนาแน่นสูง ดังแสดงในพื้นที่กรอบสีแดง




ภาพที่ 20 แสดงผลเปรียบเทียบการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นระดับเดียว ระหว่างพารามิเตอร์ที่ได้จากวิธีที่นำเสนอและ VDBSCAN

### 4.6 ผลการทดลองเปรียบเทียบการแบ่งกลุ่มข้อมูลระหว่างการใช้พารามิเตอร์ที่นำเสนอกับการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จาก AutoEpsDBSCAN

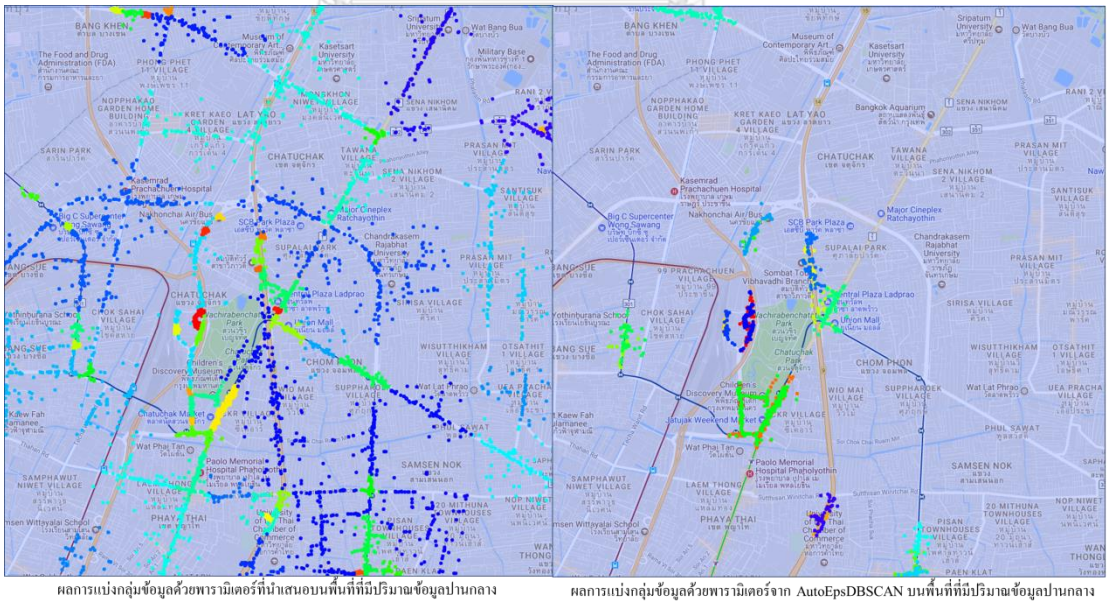
การทดลองเพื่อเปรียบเทียบผลการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น โดยใช้พารามิเตอร์จากวิธีที่นำเสนอและการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์ที่ได้จากการกำหนดพารามิเตอร์แบบอัตโนมัติ (DBSCAN with Eps Automatic for Large Dataset: AutoEpsDBSCAN) โดยใช้เงื่อนไขการทดลอง ตามข้อ 4.3 และแบ่งข้อมูลเพื่อใช้กำหนดพารามิเตอร์เป็น 2 มิติ กำหนดหนดจำนวน MinPts = 3 ผลการทดลองแสดงดังต่อไปนี้

ตารางที่ 6 ตารางแสดงผลการเปรียบเทียบการแบ่งกลุ่มข้อมูลด้วย DBSCAN โดยใช้พารามิเตอร์จากวิธีที่นำเสนอและ AutoEpsDBSCAN

| พื้นที่ | วิธีการ | จำนวนข้อมูลรับ-ส่งผู้โดยสาร | Eps (เมตร) | MinPts (Points) | Cluster | Noise (Points) |
|---|---|---|---|---|---|---|
| พื้นที่ที่มีปริมาณข้อมูลน้อย การกระจายตัวและความหนาแน่นระดับเดียว | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 882 | 76 | 8 | 16 | 554 |
| | | 554 | 339 | 19 | 5 | 419 |
| | | 419 | 414 | 9 | 8 | 154 |
| | | 154 | 527 | 11 | 0 | 154 |
| | AutoEpsDBSCAN | 441 | 969 | 27 | 2 | 372 |
| | | 441 | 1,046 | 33 | 1 | 389 |
| พื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 5,657 | 48 | 21 | 18 | 4,432 |
| | | 4,432 | 165 | 31 | 18 | 3,270 |
| | | 3,270 | 329 | 57 | 1 | 3,213 |
| | | 3,213 | 375 | 24 | 18 | 351 |
| | | 351 | 422 | 32 | 0 | 351 |
| | AutoEpsDBSCAN | 2,828 | 551 | 38 | 11 | 1,786 |
| | | 2,829 | 561 | 38 | 10 | 1,964 |
| พื้นที่ที่มีปริมาณข้อมูลมาก การกระจายตัวและความหนาแน่นหลายระดับ | การหาพารามิเตอร์แบบอัตโนมัติที่นำเสนอในงานวิจัยนี้ | 6,534 | 24 | 7 | 66 | 5,650 |
| | | 5,650 | 47 | 7 | 127 | 4,111 |
| | | 4,111 | 70 | 6 | 148 | 2,782 |
| | | 2,782 | 93 | 7 | 32 | 2,519 |
| | | 2,519 | 116 | 7 | 48 | 2,124 |
| | | 2,124 | 139 | 9 | 7 | 2,057 |
| | | 2,057 | 162 | 8 | 40 | 1,679 |
| | | 1,679 | 185 | 7 | 60 | 1,117 |
| | | 1,117 | 208 | 8 | 2 | 1100 |
| | | 1,100 | 231 | 9 | 6 | 1,046 |
| | | 1,046 | 254 | 10 | 3 | 1,008 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | 1,008 | 277 | 9 | 23 | 746 |
| | | 746 | 300 | 8 | 19 | 565 |
| | | 565 | 323 | 8 | 9 | 490 |
| | | 490 | 346 | 10 | 0 | 490 |
| | | 490 | 369 | 17 | 0 | 490 |
| | | 490 | 392 | 15 | 0 | 490 |
| | | 490 | 415 | 9 | 10 | 373 |
| | | 373 | 438 | 11 | 0 | 373 |
| | | 373 | 461 | 10 | 3 | 345 |
| | AutoEpsDBSC | 6,534 | 501 | 29 | 11 | 2,284 |
| | AN | 6,534 | 562 | 35 | 8 | 2,609 |

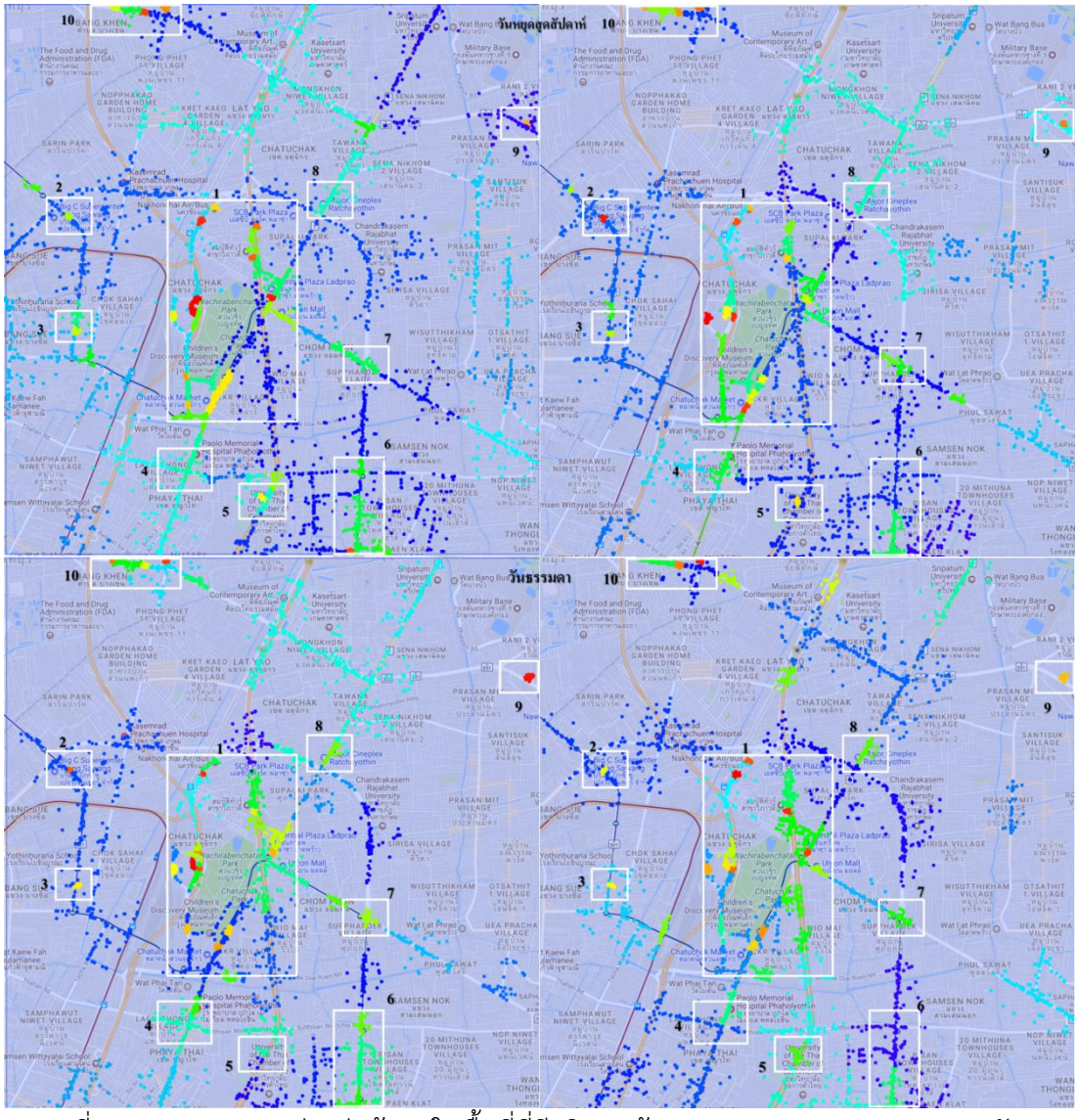
ตัวอย่างการเปรียบเทียบผลการแบ่งกลุ่มข้อมูลบนพื้นที่ที่มีข้อมูลปริมาณปานกลาง แสดงตามภาพที่ 21 โดยกลุ่มข้อมูลที่ได้จากวิธีการกำหนดพารามิเตอร์แบบ AutoEpsDBSCAN สามารถจัดกลุ่มข้อมูลได้เฉพาะพื้นที่ที่มีปริมาณข้อมูลกระจุกตัวอย่างหนาแน่นเท่านั้น ซึ่งแตกต่างจากการแบ่งกลุ่มด้วยพารามิเตอร์จากงานวิจัยที่นำเสนอ ที่สามารถแยกกลุ่มข้อมูลได้ทั้งในบริเวณที่ข้อมูลกระจุกตัวอย่างหนาแน่น และกระจายตัวแบบไม่หนาแน่น




ภาพที่ 21 แสดงผลเปรียบเทียบการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง การกระจายตัวและความหนาแน่นระดับเดียว ระหว่างพารามิเตอร์ที่ได้จากวิธีที่นำเสนอและ AutoEpsDBSCAN

### 4.7 ผลการทดลองการแบ่งกลุ่มในวันที่แตกต่างกัน

การทดลองเพื่อเปรียบเทียบการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น โดยใช้พารามิเตอร์จากวิธีที่นำเสนอในงานวิจัยนี้ เพื่อค้นหาการพื้นที่ที่เป็นจุดน่าสนใจ ในวันธรรมดาและวันหยุดสุดสัปดาห์ จากตัวอย่างพื้นที่ที่ทำการทดลองพบว่า ในเวลาหนึ่งสัปดาห์เมื่อเปรียบเทียบการแบ่งกลุ่มข้อมูล พบกลุ่มข้อมูลที่ปรากฏขึ้นอย่างสม่ำเสมอ ดังพื้นที่ในกรอบสี่เหลี่ยมสีขาว ดังภาพที่ 22 แม้ลักษณะความหนาแน่นของข้อมูลมีความแตกต่างกันบางพื้นที่มีการกระจายตัวของข้อมูลอย่างหนาแน่น และบางพื้นที่มีการกระจายตัวอย่างกระจัดจาย



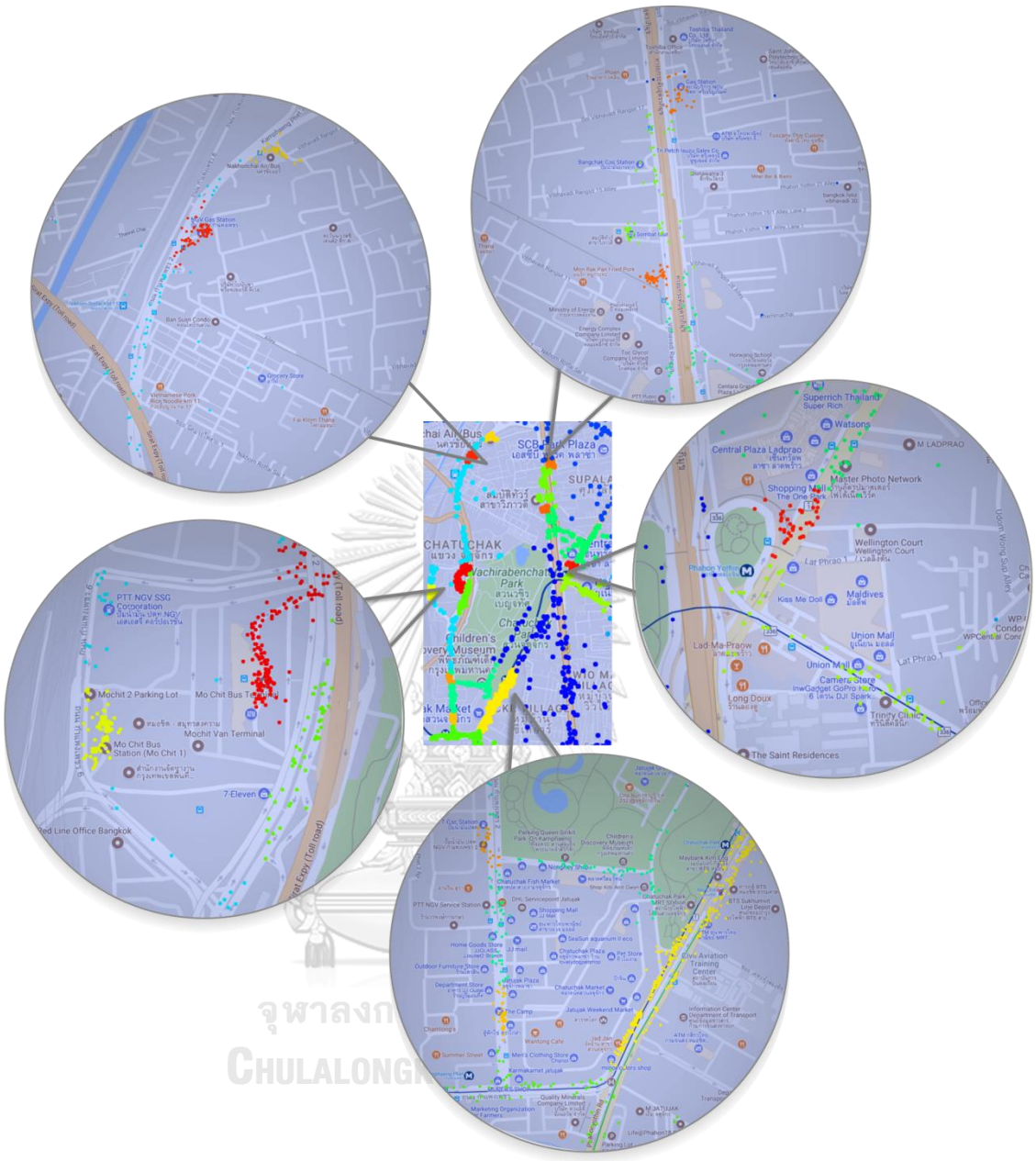

ภาพที่ 22 แสดงผลการแบ่งกลุ่มข้อมูลในพื้นที่ที่มีปริมาณข้อมูลปานกลาง และการกระจายตัวและความหนาแน่นหลายระดับ

**4.8 ผลทดลองการแบ่งกลุ่มข้อมูลโดยใช้พารามิเตอร์ที่นำเสนอเมื่อเปรียบเทียบกับแผนที่ออนไลน์**

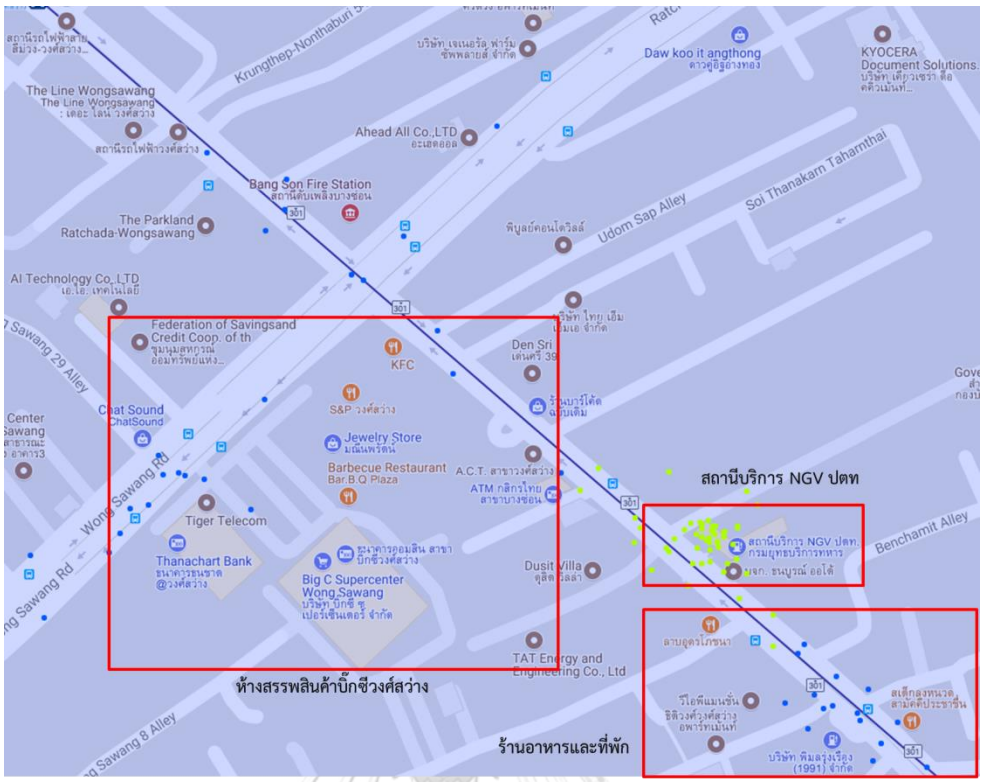
จากภาพที่ 22 ในข้อ 4.7 แสดงให้เห็นกลุ่มข้อมูลที่ปรากฏขึ้นอย่างต่อเนื่อง เมื่อนำไปสำรวจเปรียบเทียบกับแผนที่ออนไลน์ เพื่อระบุถึงบริเวณการเป็นจุดสนใจ ผลปรากฏดังต่อไปนี้ โดยจะพบว่า จุดสนใจที่ปรากฏมีทั้งที่เป็นสถานที่เช่น ห้างสรรพสินค้า โรงแรม ร้านอาหาร และบริเวณที่เป็นย่านจุดสนใจ โดยมีลักษณะเป็นพื้นที่บริเวณกว้างที่ประกอบด้วยจุดสนใจหลายจุดรวมกัน ในกรณีนี้จะพบตัวอย่างได้จากภาพที่ 33 เนื่องจากเมื่อลองเปรียบเทียบกับแผนที่ออนไลน์จะเห็นบริเวณที่เป็นพื้นที่ที่ประกอบด้วยจุดสนใจจำนวนมาก (Area of Interest)โดยมีลักษณะเป็นสีไฮไลน์สีส้ม

ตารางที่ 7 สถานที่ที่เป็นจุดสนใจ (POI) ที่ปรากฏในพื้นที่การแบ่งข้อมูล

| พื้นที่ | จุดสนใจที่ปรากฏ |
|---|---|
| พื้นที่ที่ 1 | สถานีเดินรถนครชัยแอร์ สถานีเดินรถสมบัติทัวร์ โรงเรียนธัชรินทร์วิทยา สถานีขนส่งหมอชิต ตลาดนัดจตุจักร สวนสาธารณจตุจักร บีทีเอชหมอชิต ห้างสรรพสินค้าเซนทรัลลาดพร้าว ห้างสรรพสินค้ายูเนี่ยนมอลล์ ห้างสรรพสินค้าเทสโก้โลตัสลาดพร้าว โรงเรียนหอวัง สำนักงานใหญ่ธนาคาร SCB |
| พื้นที่ที่ 2 | ห้างสรรพสินค้าบิ๊กซีวงศ์สว่าง สถานีบริการ NGV ปตท ร้านอาหารต่างๆ แมนชั่น |
| พื้นที่ที่ 3 | สหกรณ์แท็กซี่สุวรรณภูมิ |
| พื้นที่ที่ 4 | ห้างสรรพสินค้าบิ๊กซีสะพานควาย โรงพยาบาลเปาโลเมโมเรียล ร้านอาหาร โรงแรม แมนชั่น ตึกสำนักงาน (AIS Tower, ESV Tower) |
| พื้นที่ที่ 5 | ร้านอาหาร ตลาดมิ่งขวัญ ปั้มน้ำมัน ปตท. สำนักงาน Voice TV สโมสรทหารบก |
| พื้นที่ที่ 6 | ร้านอาหาร ตลาดห้วยขวาง เทวาลัยพระพิฆเนศ อาคารสำนักงาน วัดกนนทีรุทธาราม โรงเรียนปัญจทรัพย์ คอนโดมิเนียม โรงแรมเพียตร้ากรุงเทพ |
| พื้นที่ที่ 7 | สวนลุมไนท์บาซาร์ ร้านอาหาร คอนโดมิเนียม |
| พื้นที่ที่ 8 | ห้างสรรพสินค้าเมเจอร์รัชโยธิน ร้านอาหาร อาคารสำนักงาน |
| พื้นที่ที่ 9 | โรงแรม ร้านกาแฟ ปั้มน้ำมัน ปตท. |
| พื้นที่ที่ 10 | ห้างสรรพสินค้าเดอะมอลล์งามวงศ์วาน โรงเรียนสอนภาษา ร้านอาหาร ธนาคาร ห้างสรรพสินค้าไอทีซิตี้ โรงพยาบาลนนทเวช |

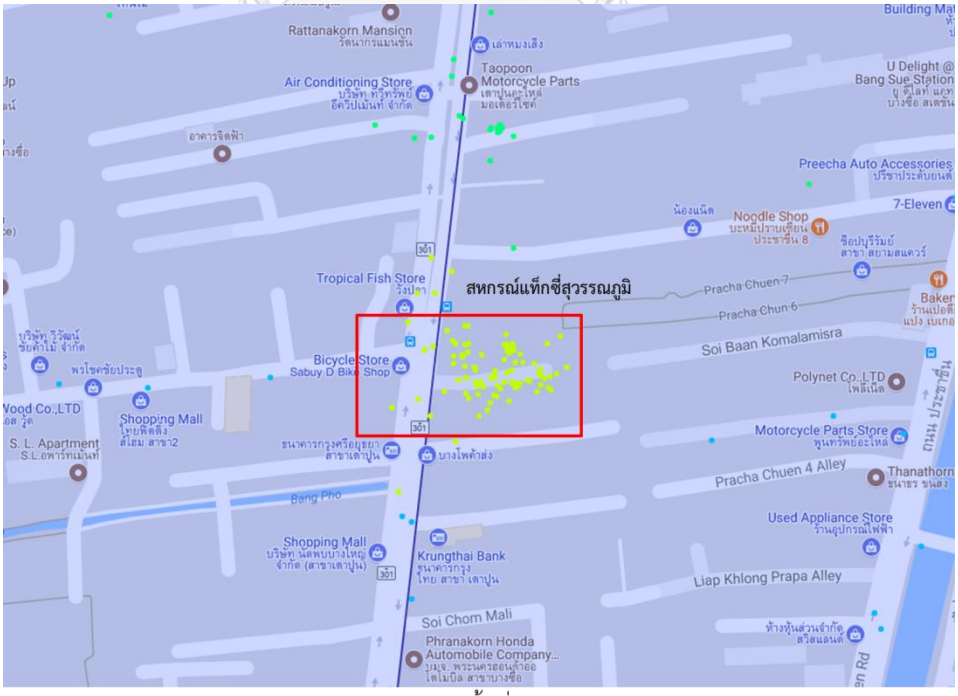



ภาพที่ 23 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 1



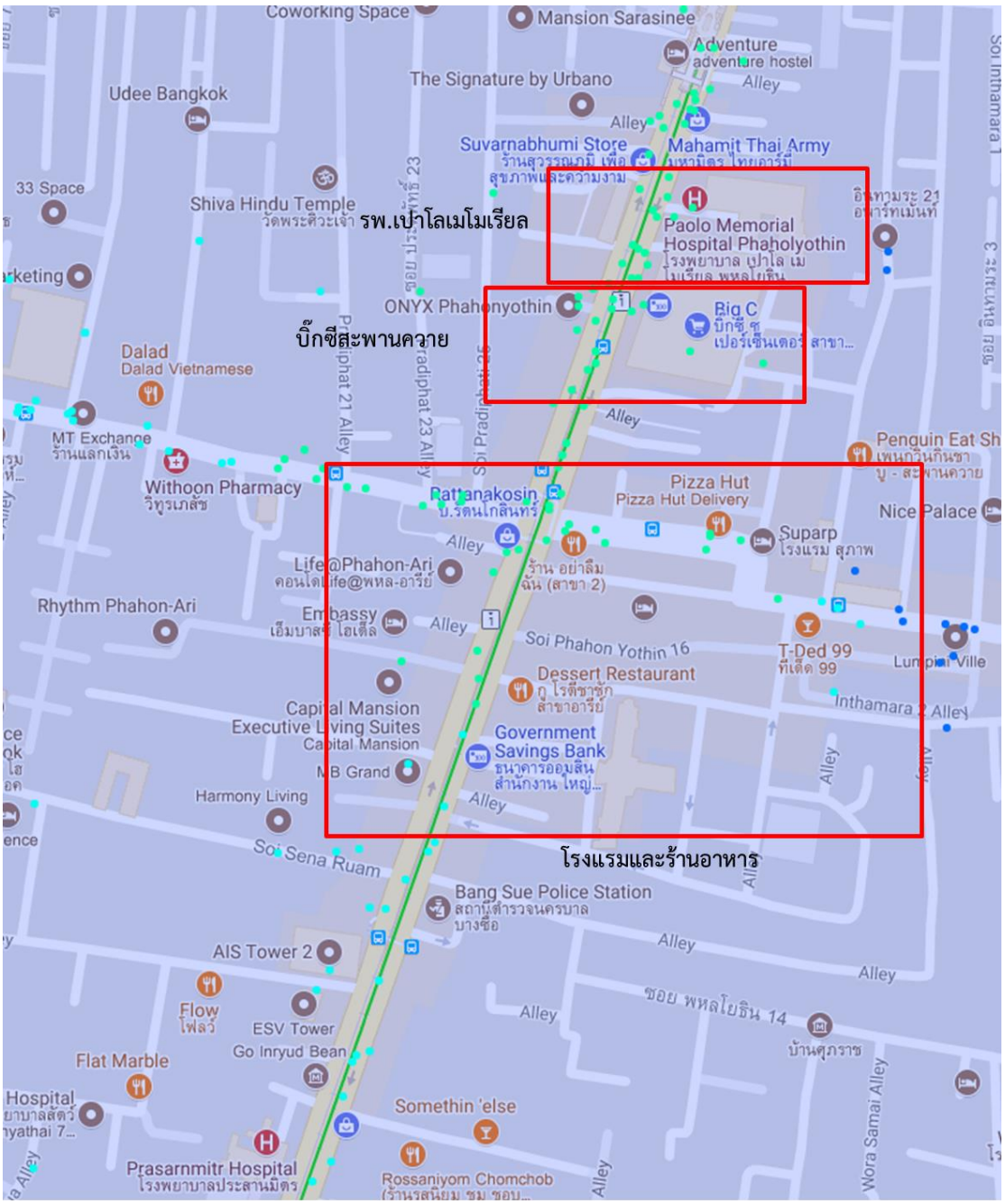

พื้นที่ 2

ภาพที่ 24 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 2




พื้นที่ 3

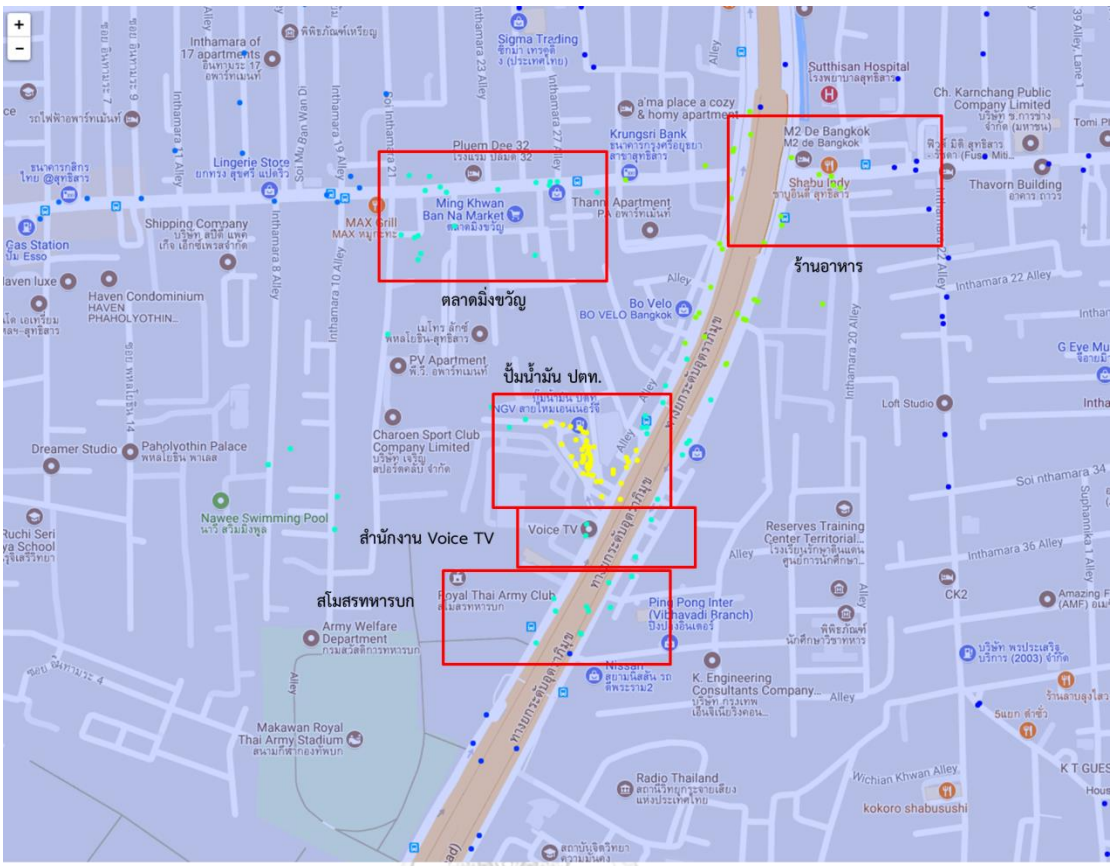
ภาพที่ 25 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 3




พื้นที่ 4

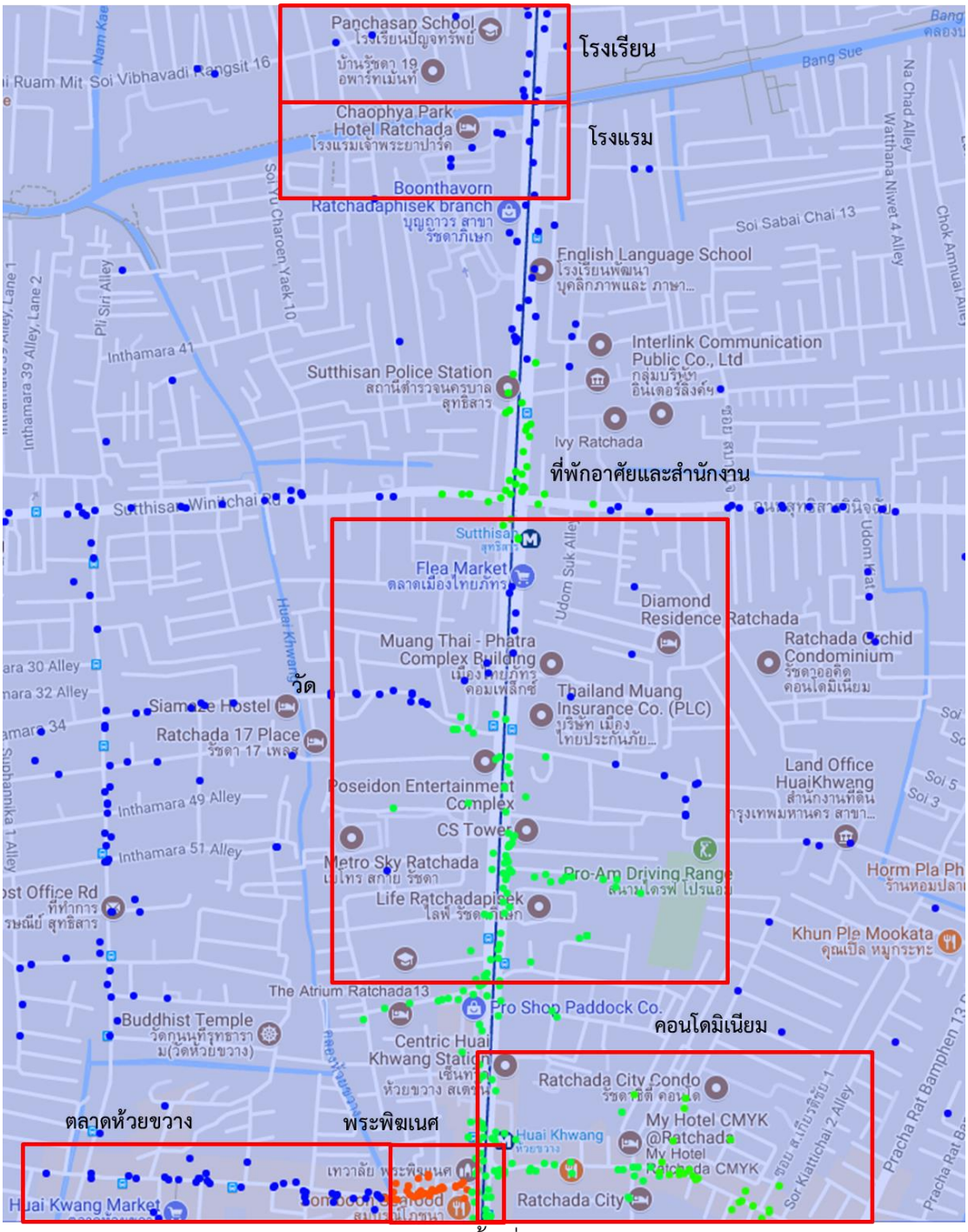
ภาพที่ 26 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 4

เมื่อพิจารณาบนแผนที่ออนไลน์จะสังเกตพบพื้นที่มีสีเหลือง ซึ่งกำหนดเป็นเขตที่มีความหนาแน่นของประชากร จากการเปรียบเทียบกลุ่มข้อมูลที่ได้การแบ่งข้อมูลในงานวิจัยนี้ พบว่ากลุ่มข้อมูลปรากฏแสดงในเขตดังกล่าวอย่างหนาแน่น ทำให้เห็นว่า ผลลัพธ์ที่ได้นอกจากสามารถระบุเป็นสถานที่ที่เป็นจุดสนใจแล้ว ยังสามารถสะท้อนให้เห็นถึงบริเวณที่เป็นย่านจุดสนใจ




พื้นที่ 5

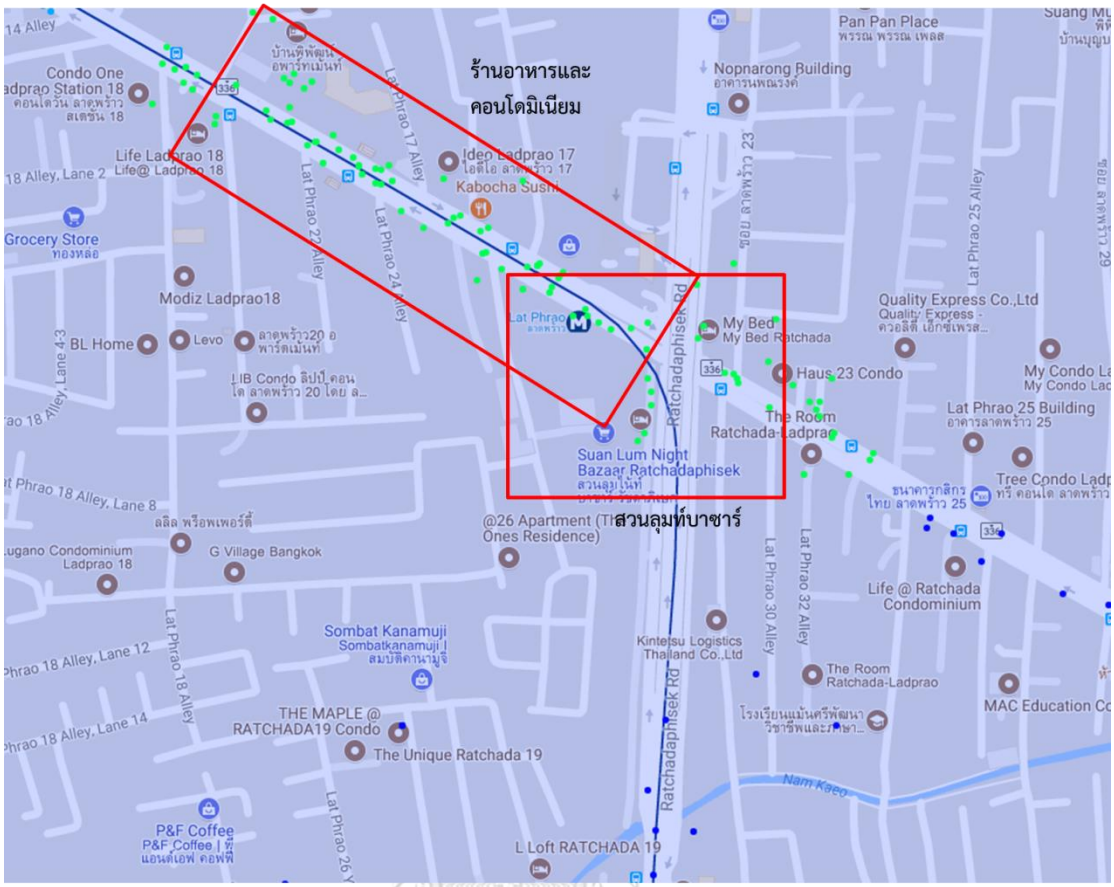
ภาพที่ 27 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 5




พื้นที่ 6

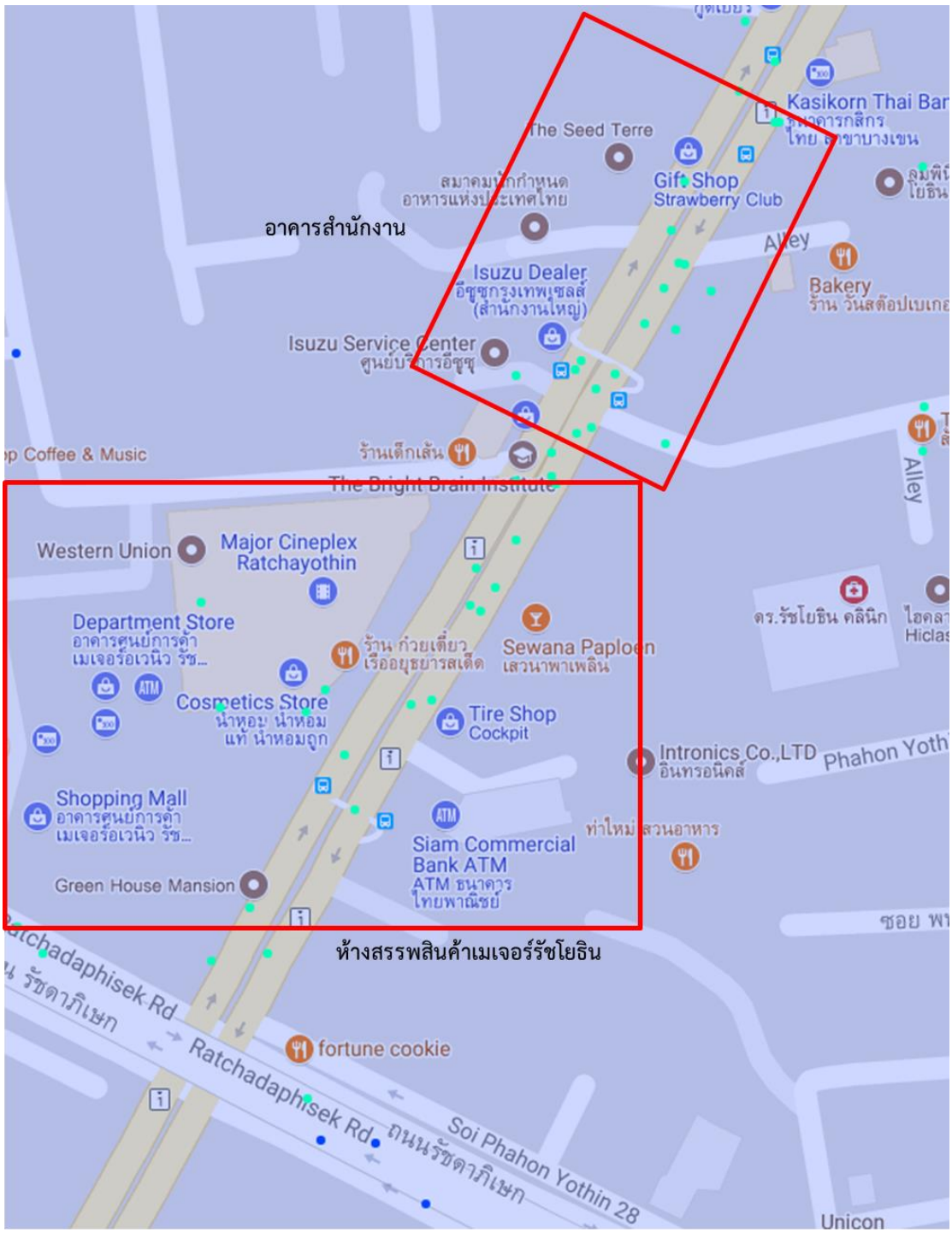
ภาพที่ 28 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 6

บริเวณตลาดห้วยขวาง เทวาลัยพระพิฆเนศ แหล่งคอนโดมิเนียม สะท้อนให้เห็นถึงย่านจุดสนใจ



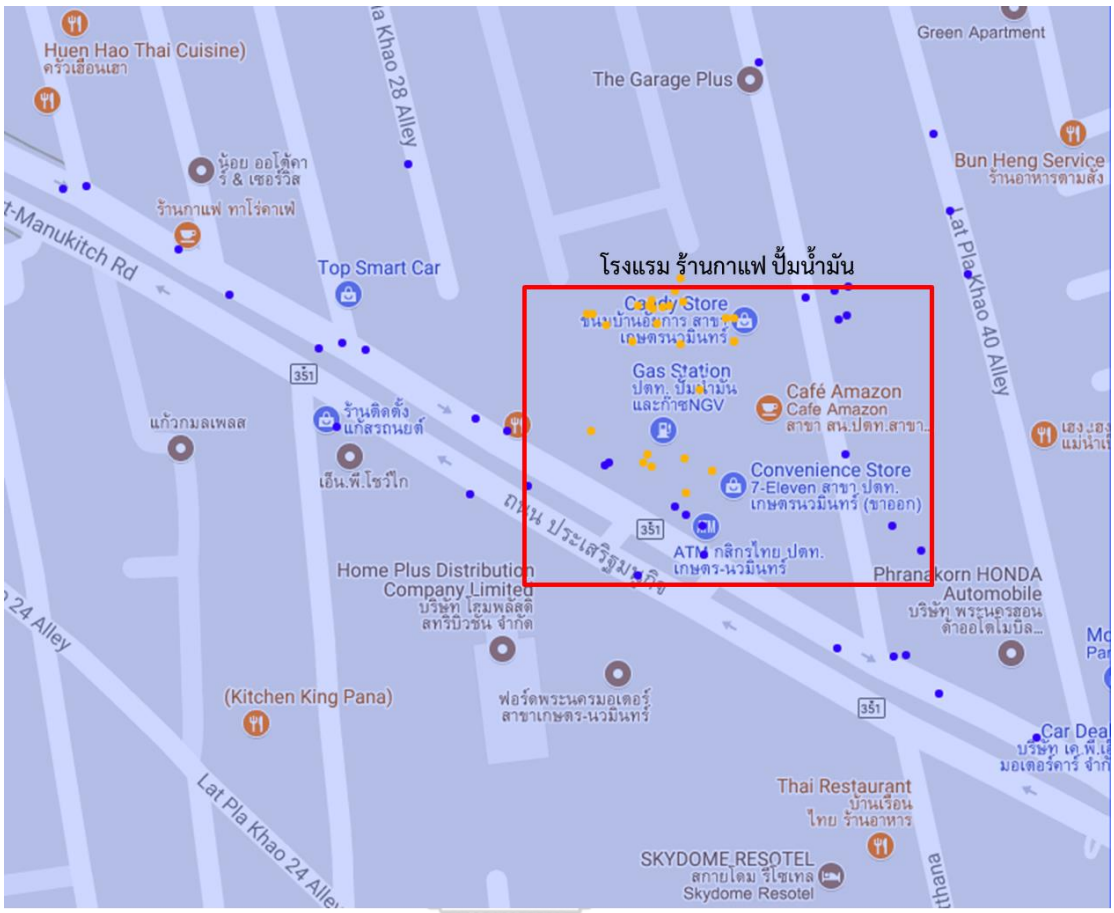

พื้นที่ 7

ภาพที่ 29 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 7



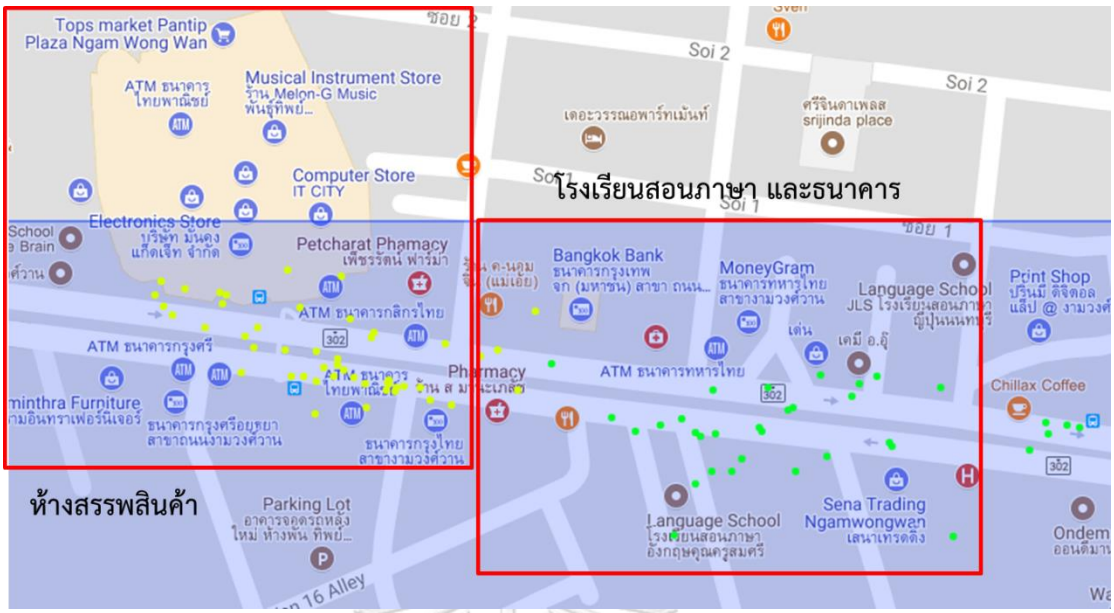

พื้นที่ 8

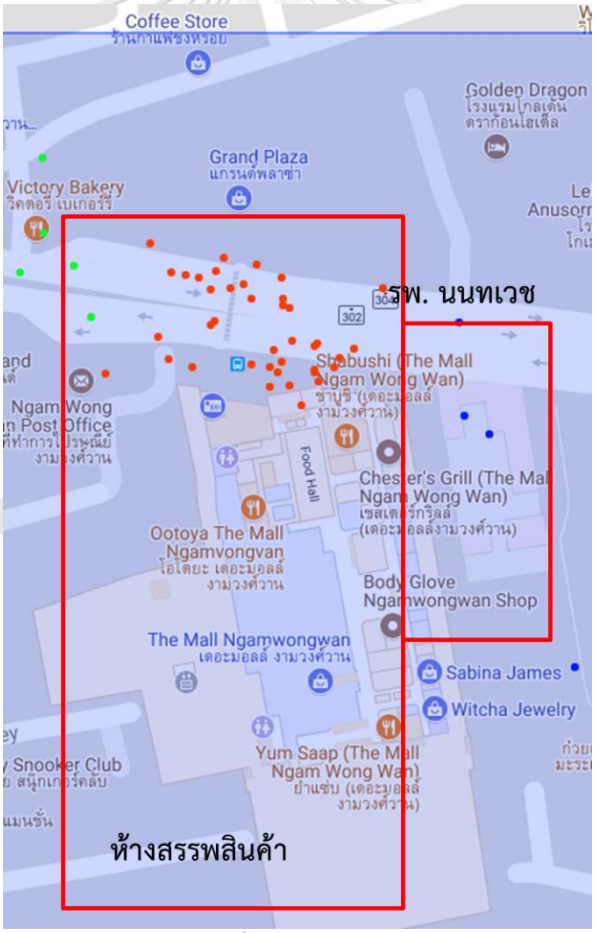
ภาพที่ 30 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 8




พื้นที่ 9

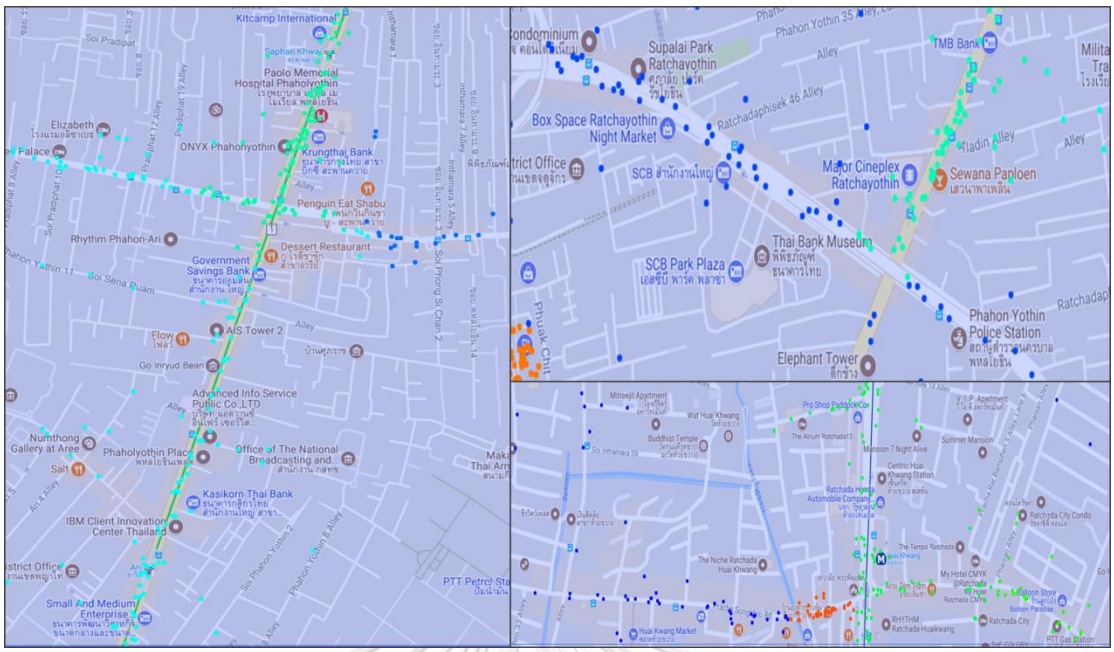
ภาพที่ 31 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 9




พื้นที่ 10

ภาพที่ 32 แสดงสถานที่บนแผนที่ออนไลน์ของพื้นที่กลุ่มที่ 10




ภาพที่ 33 แสดงพื้นที่ที่ประกอบด้วยจุดสนใจจำนวนมาก (Area of Interest)

จากภาพที่ 33 พบว่ากลุ่มข้อมูลตกอยู่ภายในพื้นที่ที่ประกอบด้วยจุดสนใจจำนวนมาก (บริเวณไฮไลน์สีเหลือง) โดยลักษณะของกลุ่มข้อมูลมักมีขนาดใหญ่ กระจายตัวตามแนวของพื้นที่ แสดงให้เห็นว่า นอกจาก การจัดกลุ่มข้อมูลจากพาริเตอร์ที่เสนอในงานวิจัยนี้ นอกจากจะให้ผลลัพธ์ที่สามารถระบุสถานที่ที่เป็นจุดที่น่าสนใจแล้ว ยังแสดงให้เห็นถึงพื้นที่ที่ประกอบด้วยจุดสนใจด้วย

# บทที่ 5
# สรุปการวิจัยและแนวทางการวิจัย

## 5.1 สรุปการวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการกำหนดพารามิเตอร์แบบอัตโนมัติเพื่อใช้ในการค้นหาพื้นที่ที่เป็นจุดจากการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมการแบ่งกลุ่มตามความหนาแน่น (DBSCAN) ซึ่งพารามิเตอร์ที่เป็นส่วนสำคัญในการทำงานของอัลกอริทึมสามารถกำหนดได้จากปริมาณและลักษณะการกระจายตัวของข้อมูล โดยวิธีที่ได้นำเสนอในงานวิจัยนี้ จะทำการหาระยะทางระหว่างพิกัดทุกพิกัดที่อยู่ภายในพื้นที่ที่ทำการทดลอง จากนั้นหาระยะทางที่ใกล้ที่สุด เนื่องจากสามารถแสดงให้เห็นการเกาะกลุ่มของข้อมูลเพื่อระบุพื้นที่ที่มีความหนาแน่นมากๆ จากนั้นเลือกพารามิเตอร์ตัวแรกคือ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) จากช่วงข้อมูลภายในช่วงระยะทางที่ใกล้ที่สุด โดยพิจารณาจุดที่ทำให้ข้อมูลมีการเปลี่ยนแปลงสูงสูด หรือในพื้นที่ใดที่มีความหนาแน่นของข้อมูลค่อนข้างสูงและใกล้เคียงกัน จนไม่สามารถระบุจุดเปลี่ยนแปลงได้อย่างชัดเจน ก็สามารถใช้ค่ารัศมีระหว่างจุดในกลุ่มข้อมูลได้ในทุกช่วงระยะข้อมูล เมื่อได้ค่ารัศมีระหว่างจุดในกลุ่มข้อมูล ซึ่งในแต่ละพื้นที่อาจมีได้หลายค่า ลำดับต่อไปนำไปคำนวณหาค่าจุดต่ำสุด (MinPts) โดยนับจำนวนพิกัดภายในรัศมีข้อมูลแล้วจึงเลือกค่าสูงสูดในช่วงข้อมูลที่มีความถี่มากที่สุดกำหนดให้เป็นจุดต่ำสุดในแต่ละรัศมีกลุ่มนั้นๆ เมื่อนำไปกำหนดให้อัลกอริทึมการแบ่งกลุ่มข้อมูลด้วยความหนาแน่น อัลกอริทึมจะประมวลผลแบ่งกลุ่มตามพารามิเตอร์ที่กำหนดให้ ซึ่งในแต่ละรอบจะปรากฏพิกัดที่ไม่สามารถจัดเข้ากลุ่มข้อมูลใดได้ โดยพิกัดดังกล่าวจะถูกนำไปพิจารณาเพื่อแบ่งกลุ่มให้ชุดพารามิเตอร์ในลำดับถัดไป จนกระทั่งเสร็จสิ้นการทำงาน พิกัดที่ไม่สามารถจัดเข้ากลุ่มใดๆ ได้เลยในรอบสุดท้ายของการทำงาน จึงจะถือให้เป็น Noise

งานวิจัยนี้ สามารถนำเสนอวิธีที่สามารถกำหนดพารามิเตอร์ทั้งสองพารามิเตอร์คือ รัศมีระหว่างจุดในกลุ่มข้อมูล (Eps) และ จำนวนจุดขั้นต่ำสุดสำหรับการสร้างจุดศูนย์กลางของกลุ่มข้อมูล (MinPts) แบบอัตโนมัติ โดยพิจารณาจากความหนาแน่นและการกระจายตัวของข้อมูล

จากผลการทดลองแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์จากวิธีที่นำเสนอดังกล่าว เพื่อใช้ค้นหาจุดสนใจ พบว่า ผลการแบ่งกลุ่มข้อมูลมีความละเอียด สามารถระบุพื้นที่ได้ชัดเจน ทั้งในพื้นที่ที่มีความหนาแน่นน้อย และพื้นที่ที่มีความหนาแน่นสูงเช่น พื้นที่กลางเมือง และเมื่อเปรียบเทียบกับวิธีการแบ่งกลุ่มข้อมูลด้วยพารามิเตอร์จาก DMDBSCAN VDBSCAN และ AutoEpsDBSCAN พบว่าผลลัพธ์จากวิธีที่นำเสนอ สามารถสะท้อนกลุ่มข้อมูลที่แสดงขอบเขตชัดเจน แบ่งแยกกลุ่มข้อมูลที่มีการกระจายตัวหนาอย่างแน่นออกจากกันได้ถึงแม้จะอยู่ในบริเวณใกล้กัน ให้กลุ่มข้อมูลที่มีขนาดและ

รูปร่างตามลักษณะการกระจายตัวของข้อมูล แสดงจุดสำคัญได้ดีกว่า แม้ในพื้นที่ที่มีความหนาแน่นแตกต่างกัน

นอกจากนี้การทดสอบเพื่อระบุความเป็นจุดสนใจ โดยการทดลองรันเปรียบเทียบการแสดงพื้นที่ที่ปรากฏอย่างต่อเนื่องหลายๆ วัน พบพื้นที่ที่มีสถานที่ที่น่าสนใจปรากฏขึ้นอย่างสม่ำเสมอ แม้จะมีความหนาแน่นที่แตกต่างกัน เช่น ตลาดที่เปิดให้บริการในช่วงวันหยุดสุดสัปดาห์ ปริมาณข้อมูลในกลุ่มข้อมูลจะมีความหนาแน่นมากกว่าในวันธรรมดา สุดท้ายเมื่อทำการเปรียบเทียบผลกับแผนที่ออนไลน์ พบว่า บริเวณที่ทำการทดลอง แสดงกลุ่มข้อมูลปรากฏเป็นสถานที่ที่เป็นจุดสนใจต่างๆ เช่น ห้างสรรพสินค้า ร้านอาหาร โรงเรียน สถานีขนส่ง โรงพยาบาล โรแรม อาคารสำนักงาน เป็นต้น นอกจากนี้ บางกลุ่มข้อมูลให้ผลลัพธ์ของเขตพื้นที่ที่สามารถเทียบได้กับพื้นที่ที่น่าสนใจ (Area of Interest) บนแผนที่ออนไลน์ (Google map) ซึ่งพื้นที่ดังกล่าวถือเป็นบริเวณที่ประกอบด้วยสถานที่ที่เป็นจุดสนใจมากมายเช่น ย่านสะพานควาย ย่านรัชดาภิเษก เป็นต้น

## 5.2 แนวทางในการวิจัยต่อ

การทดลองนี้ในขั้นตอนของการกำหนดพารามิเตอร์ ส่วนสำคัญคือการกำหนดช่วงข้อมูล ซึ่งงานวิจัยได้เลือกใช้ฟังก์ชั่น Sturge' rule ซึ่งเป็นสูตรที่ใช้ในการหาช่วงข้อมูลที่เหมาะกับข้อมูลที่มีการกระจายตัวแบบปกติ (Normal Distribution) ดังนั้น หากในกรณีที่พบการกระจายตัวของข้อมูลนอกเหนือลักษณะดังกล่าวอาจมีผลกระทบต่อการกำหนดพารามิเตอร์ ดังนั้นจึงควรทดลองใช้ฟังก์ชั่นอื่นเพิ่มเติมเพื่อให้เหมาะสมกับลักษณะของข้อมูล

ในการทดลองเพื่อค้นหาจุดสนใจควรทำการค้นหากับพื้นที่ต่างๆ เพิ่มขึ้นหรือนำแนวความคิดไปประยุกต์ใช้ร่วมกับข้อมูลที่สามารถนำมาใช้ระบุพฤติกรรมการเดินทางของประชากรเช่น ข้อมูลจีพีเอสจากเครื่องมือสื่อสาร ข้อมูลการใช้โซเชียลเน็ตเวิร์ค เพื่อที่จะสามารถค้นพบสถานที่ที่เป็นจุดสนใจแม้ในบางพื้นที่ที่ข้อมูลจากรถแท็กซี่ไม่สามารถเข้าถึง

# รายการอ้างอิง


1. Zheng, Z., S. Rasouli, and H. Timmermans, *Evaluating the Accuracy of GPS-based Taxi Trajectory Records.* Procedia Environmental Sciences, 2014. 22: p. 186-198.
2. Ester, M., et al., *A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise*, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, AAAI Press: Portland, Oregon. p. 226-231.
3. Schubert, E., et al., *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.* ACM Trans. Database Syst., 2017. 42(3): p. 1-21.
4. interest, P.o., *https://en.wikipedia.org/wiki/Point_of_interest*.
5. Raza, A., S. Hameed, and T. Macintyre. *Global Positioning System – Working and its Applications*. 2008. Dordrecht: Springer Netherlands.
6. Mahmoud, H. and N. Akkari. *Shortest Path Calculation: A Comparative Study for Location-Based Recommender System*. in *2016 World Symposium on Computer Applications & Research (WSCAR)*. 2016.
7. Histogram, *https://en.wikipedia.org/wiki/Histogram*.
8. Sturges, H.A., *The Choice of a Class Interval.* Journal of the American Statistical Association, 1926. 21(153): p. 65-66.
9. Han, J., M. Kamber, and J. Pei, *10 - Cluster Analysis: Basic Concepts and Methods*. 2012. 443-495.
10. Yaşar, F.G. and G. Ulutagay. *Challenges and possible solutions to density based clustering*. in *2016 IEEE 8th International Conference on Intelligent Systems (IS)*. 2016.
11. Liu, P., D. Zhou, and N. Wu. *VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise*. in *2007 International Conference on Service Systems and Service Management*. 2007.

12. Elbatta, M.T.H. and W.M. Ashour., *A Dynamic Method for Discovering Density Varied Clusters.* International Journal of Signal Processing, Image Processing and Pattern Recognition, 2013. 6(1): p. 123-134.
13. Gaonkar, M.N. and K. Sawant, *AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset.* 2013.
14. Rahmah, N. and I. Sitanggang, *Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra*. Vol. 31. 2016. 012012.
15. Zhang, L., et al. *Exploiting Taxi Demand Hotspots Based on Vehicular Big Data Analytics*. in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*. 2016.
16. Kisilevich, S., F. Mansmann, and D. Keim, *P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos*, in *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. 2010, ACM: Washington, D.C., USA. p. 1-4.
17. Zhao, P.X., et al., *Detecting Hotspots from Taxi Trajectory Data Using Spatial Cluster Analysis.* 2015.
18. Tongsinoot, L., *ANALYSIS OF ROAD TRAFFIC PATTERNS USING CDR AND GPS.* 2017.

# ภาคผนวก

## ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอุไรวรรณ์ อังคะเวย์ เกิดวันที่ 10 สิงหาคม 2528 ที่จังหวัดอุดรธานี สำเร็จการศึกษาปริญญาตรีหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ในปีการศึกษา 2552 และเข้าศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2559