วิธีการถดถอยเชิงเส้นพหุคูณและโครงข่ายประสาทเทียมโดยใช้องค์ประกอบหลักในการพยากรณ์อินทรีย์วัตถุในดิน

MULTIPLE LINEAR REGRESSION AND ARTIFICIAL NEURAL NETWORKS BASED ON PRINCIPAL COMPONENTS TO PREDICT SOIL ORGANIC MATTER

สิริกัลยา ประมวล
นักศึกษาหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
E-mail : sirikanlaya.pramual@gmail.com

กมลชนก พานิชการ
อาจารย์ประจำภาควิชาสถิติ
คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
E-mail : kamolcha@su.ac.th

นัทธีรา สรรมณี
อาจารย์ประจำภาควิชาวิทยาศาสตร์สิ่งแวดล้อม
คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
E-mail : sanmanee@su.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการสำหรับการสร้างโมเดลในการพยากรณ์ปริมาณอินทรีย์วัตถุในดินสองวิธี คือวิธีการถดถอยเชิงเส้นพหุคูณและวิธีการโครงข่ายประสาทเทียมแบบเชื่อมโยงไปข้างหน้า โดยเก็บตัวอย่างดินมาจาก 3 จังหวัดได้แก่ นครปฐม, สมุทรสาคร และสมุทรสงคราม วัดข้อมูลเกี่ยวกับคุณสมบัติทางเคมีของดิน แล้วใช้เป็นตัวแปรพยากรณ์ทั้งหมด 17 ตัวแปร เพื่อให้ได้จำนวนตัวแปรที่เหมาะสมในโมเดล วิธีวิเคราะห์องค์ประกอบหลักถูกนำมาใช้ เพื่อลดมิติข้อมูลและขจัดความสัมพันธ์ระหว่างตัวแปรพยากรณ์ ซึ่งผลลัพธ์ได้ 5 องค์ประกอบและใช้เป็นตัวแปรในการพยากรณ์ปริมาณอินทรีย์วัตถุ ในการเปรียบเทียบค่าความแม่นยำของโมเดลใช้ค่า RMSE, MAE, MAPE และ R ผลลัพธ์ที่ได้จากวิธีการถดถอยเชิงเส้นพหุคูณมีค่าวัดความแม่นยำเท่ากับ 0.27, 0.23, 9.39 และ 0.91 ตามลำดับ เมื่อเปรียบเทียบผลลัพธ์ที่ได้จากวิธีการโครงข่ายประสาทเทียมแบบที่มีการเชื่อมโยงไปข้างหน้า มีค่าวัดความแม่นยำ 0.24, 0.22, 8.57 และ 0.87 ตามลำดับ แสดงให้เห็นว่าโมเดลที่ได้จากวิธีการโครงข่ายประสาทเทียมให้ผลลัพธ์ดีกว่าโมเดลที่ได้จากวิธีการถดถอยเชิงเส้นพหุคูณ

คำสำคัญ : การถดถอยเชิงเส้นพหุคูณ โครงข่ายประสาทเทียมแบบที่มีการเชื่อมโยงไปข้างหน้า การวิเคราะห์องค์ประกอบหลัก อินทรีย์วัตถุในดิน

ABSTRACT

In this research, multiple linear regression (MLR) and feed-forward artificial neural networks (FANN) were used to fit the models for predicting quantity of organic matter. Soil samples were selected from Nakhon Pathom, Samut Sakhon and Samut Songkram and were measured 17 soil properties using as independent variables. In order to reduce numbers of independent variables and eliminate data collinearity, Principal Component Analysis were used. The models with principal components were performed. The evaluations of performance of models were measured by RMSE, MAE, MAPE and R. The results from MLR model are 0.27, 0.23, 9.39 and 0.91, respectively. The results from FANN model are 0.24, 0.22, 8.57 and 0.87, respectively. The result showed that FANN model is better than MLR model for predicting soil organic matter.

KEYWORDS : Multiple linear regression, Feed-forward artificial neural networks, Principal component analysis, Soil organic mater

บทนำ

เนื่องจากอินทรีย์วัตถุ (Organic Matter: OM) ในดินจัดว่าเป็นแหล่งสำรองของธาตุอาหารในดินและมีส่วนสำคัญในการเสริมสร้างคุณสมบัติของดินทั้งทางกายภาพ ทางเคมี และทางชีวภาพให้เหมาะสมต่อการเจริญเติบโตของพืช บทบาทของอินทรีย์วัตถุที่มีผลต่อทางกายภาพของดิน เช่น ความหนาแน่นและความพรุนของดิน โครงสร้างของดินและความสามารถในการอุ้มน้ำของดิน บทบาททางด้านเคมีของดิน เช่นเป็นแหล่งธาตุอาหารให้แก่พืช ทำให้ดินมีค่าการแลกเปลี่ยนประจุบวก (CEC) ในดินสูงขึ้น ซึ่งจะช่วยในการดูดซับสารอาหารในดิน บทบาททางด้านชีวภาพเช่น การแปรสภาพธาตุอาหาร การปลดปล่อยไนโตรเจนในรูปของแอมโมเนียหรือไนเตรตที่เป็นประโยชน์ รวมทั้งช่วยยับยั้งการเจริญเติบโตของเชื้อโรคในพืช (Stevenson,1994) ซึ่งพื้นที่ในประเทศไทยส่วนใหญ่เป็นพื้นที่ในด้านการเกษตรและส่วนมากเป็นการเกษตรแบบดั้งเดิมที่มีการเพาะปลูกติดต่อกันเป็นเวลายาวนาน โดยไม่มีการปรับปรุงบำรุงดินหรือการเติมอินทรีย์วัตถุที่ไม่มากเพียงพอ เป็นสาเหตุให้ปริมาณอินทรีย์วัตถุลดต่ำลงเรื่อยๆ (คณาจารย์ภาควิชาปฐพีวิทยา, 2548) ดังนั้นระดับของอินทรีย์วัตถุจึงเป็นคุณสมบัติอย่างหนึ่งในการบ่งชี้คุณภาพของดิน การพยากรณ์ปริมาณอินทรีย์วัตถุจึงมีความสำคัญในด้านการตรวจสอบการเปลี่ยนแปลงของดินในอนาคต และส่งผลดีในด้านการพัฒนาการเกษตรเป็นอย่างยิ่ง

เนื่องจากอินทรีย์วัตถุในดินเป็นแหล่งให้ธาตุอาหารสำคัญหลายชนิด การวิจัยในครั้งนี้จึงใช้ข้อมูลตัวอย่างดินที่วัดข้อมูลเกี่ยวกับคุณสมบัติทางเคมีของดินตัวอย่าง รวมทั้งหมด 17 ตัวแปร โดยเก็บตัวอย่างดินมาจากพื้นที่เกษตรกรรมประเภทสวนผลไม้ ในภูมิภาคตะวันตกของประเทศไทยจำนวนสามจังหวัดเป็นพื้นที่ศึกษา ได้แก่ นครปฐม, สมุทรสาคร และสมุทรสงคราม มาใช้ในสร้างโมเดลในการพยากรณ์ปริมาณอินทรีย์วัตถุ เพื่อใช้เป็นเครื่องมือในการตรวจสอบคุณภาพของดินในอนาคต

การประยุกต์วิธีวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ในขั้นตอนแรก เพื่อให้ได้ตัวแปรพยากรณ์จำนวนที่เหมาะสมในโมเดลและสามารถขจัดความสัมพันธ์ระหว่างตัวแปรพยากรณ์ ที่อาจนำไปสู่ค่าทำนายที่ผิดพลาดได้ (Sousa และคณะ, 2007) ซึ่งองค์ประกอบหลักใหม่ที่ได้จะใช้เป็นตัวแปรในการพยากรณ์ค่าปริมาณอินทรีย์วัตถุในแต่ละโมเดล ซึ่งจะแบ่งข้อมูลออกเป็นสองชุด คือชุดเรียนรู้ (training set) ที่ใช้ในการสร้างโมเดลและชุดทดสอบ (test set) สำหรับการประเมินผลโมเดล

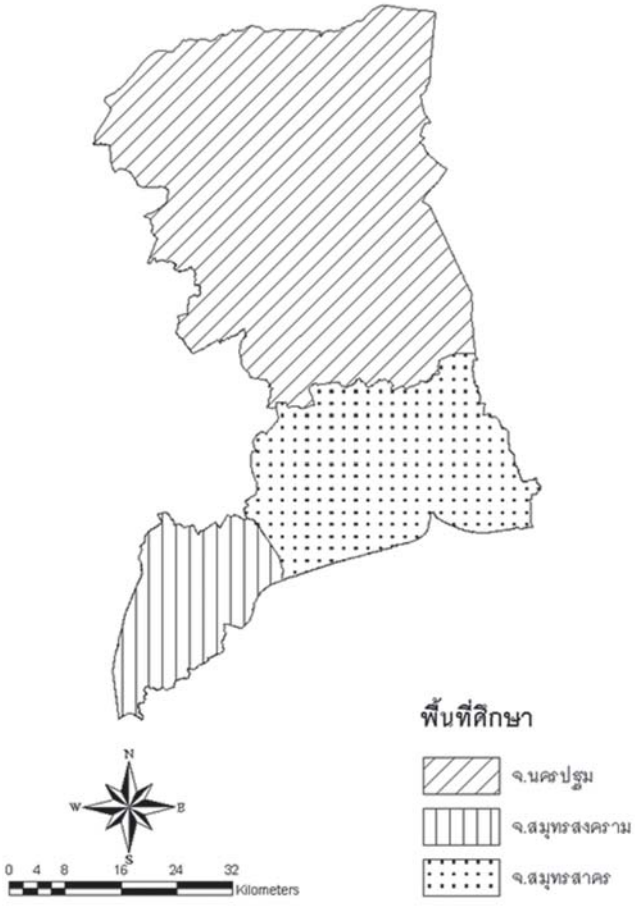
วิธีแรกที่ใช้ในการสร้างโมเดลคือวิธีการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression: MLR) เป็นวิธีที่นิยมใช้กันอย่างกว้างขวางในการพยากรณ์ตัวแปรตามที่มีลักษณะข้อมูลเป็นแบบต่อเนื่อง และหาความสัมพันธ์ระหว่างตัวแปรพยากรณ์หลายตัวกับตัวแปรตาม รวมทั้งมีข้อสมมติว่าการแจกแจง

ของข้อมูลเป็นแบบปกติ ซึ่งข้อมูลตัวแปรตามที่ได้บางครั้งอาจจะมีความไม่เป็นเชิงเส้น ดังนั้นจึงนำเสนอวิธีที่สองคือวิธีการสร้างโมเดลด้วยวิธีโครงข่ายประสาทเทียม (Haykin, 1998) โดยในงานวิจัยนี้ใช้วิธีโครงข่ายประสาทเทียมแบบเชื่อมโยงไปข้างหน้า (Feed-forward Artificial Neural Networks: FANN)

เนื่องจากเป็นวิธีที่ใช้กันอย่างแพร่หลายโดยเฉพาะเป็นวิธีหลักที่ใช้สำหรับการสร้างโมเดลและการพยากรณ์ ซึ่งโมเดลนี้สามารถใช้ได้กับข้อมูลที่มีลักษณะไม่เป็นเชิงเส้น และสามารถแก้ปัญหาที่ยุ่งยากซับซ้อนได้ (Yang และคณะ, 2006, Ingleby และ Crowe, 2001)

วิธีการศึกษา

1. ตัวอย่างดินที่ใช้ในการศึกษา



ภาพที่ 1 พื้นที่ตัวอย่างดินจากเกษตรกรรมสวนผลไม้

ตัวอย่างดินที่ใช้ในการศึกษา เป็นดินจากพื้นที่เกษตรกรรม ประเภทสวนผลไม้ในภูมิภาคตะวันตกของประเทศไทย ประกอบด้วยพื้นที่สามจังหวัดของประเทศไทย ได้แก่ นครปฐม สมุทรสาครและสมุทรสงครามดังแสดงดังภาพที่ 1 จำนวน 58 จุด ซึ่งนำมาวัดข้อมูลคุณสมบัติทางเคมีในดิน ประกอบไปด้วยตัวแปรพยากรณ์ ทั้งหมด 17 ตัวแปร สถิติพื้นฐานของข้อมูลคุณสมบัติทางเคมีของตัวอย่างดินที่เก็บมาศึกษา ประกอบด้วย Al, Mn, Fe, Cr,Mg, Zn, Cu, Pb, K, Na, Ca ,FA, HA, CEC, %clay, TN และ OC แสดงดังตารางที่ 1

2. การลดจำนวนตัวแปรอิสระ

วิธีการวิเคราะห์องค์ประกอบหลัก (PCA) เป็นเทคนิคการลดจำนวนตัวแปรและขจัดความสัมพันธ์ระหว่างตัวแปรพยากรณ์ โดยการสร้างเซตของตัวแปรใหม่ ให้เป็นฟังก์ชันเชิงเส้นของตัวแปรเดิม และเซตของตัวแปรใหม่นั้นจะมีรายละเอียดของตัวแปรเดิม โดยที่จำนวนตัวแปรใหม่ต้องไม่เกินจำนวนตัวแปรเดิม นั่นคือกรณีมีตัวแปรเดิม p ตัว จำนวนตัวแปรใหม่จะมี m ตัว จะได้ว่า $m \le p$ และ m องค์ประกอบใหม่ที่เลือกมาสามารถ อธิบายความแปรปรวนของข้อมูลได้

ความสามารถของ PCA บนชุดข้อมูลที่ใช้ในการศึกษานี้ถูกวัดโดยใช้การทดสอบ Bartlett's sphericity (กัลยา วานิชบัญชา, 2551, Sousa และคณะ, 2007) ดังแสดงในสมการที่ 1

$$\chi^2 = -\left[(n-1) - \frac{2(p+5)}{6}\right] \mathrm{In}|\mathrm{R}| \qquad (1)$$

โดยที่ χ^2 มีองศาอิสระเท่ากับ $\frac{1}{2}\mathrm{p}(\mathrm{p}-1)$,

$\ln|\mathrm{R}|$ คือค่า log ของดีเทอร์มิแนนท์ของเมตริกซ์สหสัมพันธ์ R

p คือจำนวนตัวแปร,

n คือจำนวนข้อมูล

$$|\mathrm{R}| = \prod_{i=1}^{p} \lambda_i$$

โดยที่ λ_i คือค่าไอเกนของตัวแปรที่ i ; i = 1,2,...,p

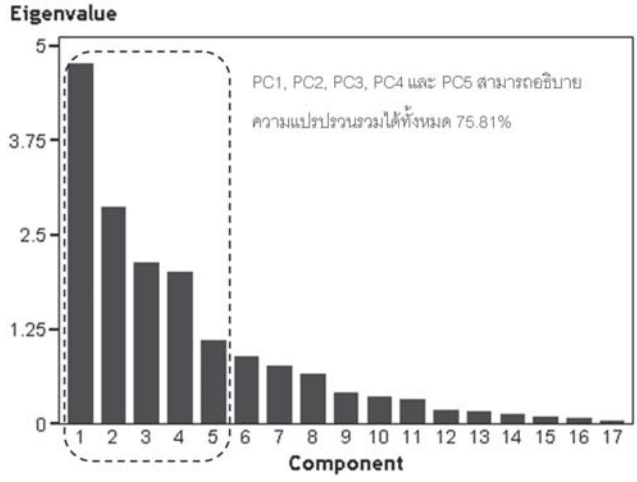
สมมติฐานหลักที่เราพิจารณาคือตัวแปรทุกตัวไม่มีสหสัมพันธ์ต่อกัน ถ้าเราปฏิเสธสมมติฐานหลัก นั่นคือ PCA

ตารางที่ 1 ค่าสถิติพื้นฐานของคุณสมบัติทางเคมีของตัวอย่างดินที่นำมาศึกษา

ตัวแปรพยากรณ์	ค่าต่ำสุด	ค่าสูงสุด	ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตรฐาน
1) อลูมินัม (Al)	7929.24	34947.96	14163.40	4137.62
2) แมงกานีส (Mn)	190.15	2180.88	836.92	394.33
3) เหล็ก (Fe)	11810.52	28602.88	19846.78	3394.37
4) โครเมียม (Cr)	13.27	48.71	22.40	4.98
5) แมกนีเซียม (Mg)	1270.10	5717.70	3131.85	820.79
6) สังกะสี (Zn)	31.36	114.10	60.78	19.91
7) ทองแดง (Cu)	11.47	267.44	39.49	42.46
8) ตะกั่ว (Pb)	24.01	61.38	37.83	8.33
9) โพแทสเซียม (K)	747.57	4831.45	2167.00	746.72
10) โซเดียม (Na)	437.32	5101.70	1697.41	643.47
11) แคลเซียม (Ca)	3957.08	16816.74	12475.71	2567.58
12) กรดฟัลวิก (FA)	10.65	609.80	151.60	113.61
13) กรดฮิวมิก (HA)	6.57	49.47	32.91	12.24
14) ค่าการแลกเปลี่ยนประจุบวก (CEC)	14.73	34.29	24.68	4.79
15) เปอร์เซ็นต์ดินเหนียว (% clay)	14.27	87.87	44.28	15.21
16) ไนโตรเจนรวม (TN)	0.07	0.22	0.14	0.04
17) อินทรีย์คาร์บอน (OC)	0.70	2.32	1.41	0.42

ควรถูกนำมาใช้ จากผลลัพธ์การทดสอบสมมติฐานพบว่าค่า Bartlett's sphericity มีค่าเท่ากับ 681.672 (p-value = .000) ดังนั้นปฏิเสธสมมติฐานหลัก นั่นคือตัวแปรมีสหสัมพันธ์ต่อกัน ดังนั้น PCA ควรถูกนำมาใช้ ในการเลือกค่าไอเกน (eigenvalues) จากตารางที่ 2 แสดงค่าน้ำหนักของ 5 องค์ประกอบหลักแรกที่ได้ โดยเลือกตัวแปรที่มีค่าน้ำหนักมากกว่า 0.40 ขึ้นไป

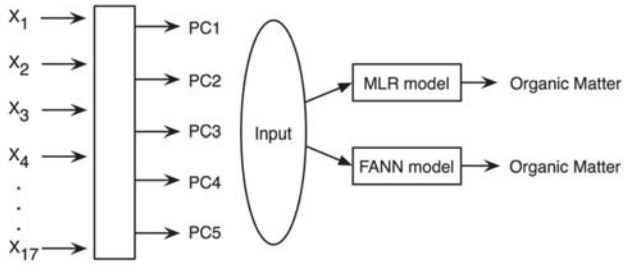
จากภาพที่ 2 แสดงการเลือกค่าไอเกน โดยเลือกองค์ประกอบที่มีค่าไอเกนมากกว่า 1 ซึ่งมีเพียง 5 องค์ประกอบ เมื่อพิจารณาความแปรปรวนรวมที่ถูกอธิบายโดยองค์ประกอบหลักทั้งหมด พบว่าองค์ประกอบหลักที่ 1 (PC1) จนถึงองค์ประกอบหลักที่ 5 (PC5) สามารถอธิบายความแปรปรวนรวมได้ทั้งหมด 75.81% แต่องค์ประกอบหลักที่ 6 (PC6) ถึงองค์ประกอบหลักที่ 17 (PC17) อธิบายความแปรปรวนได้เพียง 14.19% ดังนั้นจึงเลือก 5 องค์ประกอบหลักแรกที่ได้ มาใช้เป็นตัวแปรพยากรณ์ปริมาณอินทรีย์วัตถุในดินในโมเดลการถดถอยเชิงเส้นพหุคูณและโมเดลโครงข่ายประสาทเทียมแบบที่มีการเชื่อมโยงไปข้างหน้า ซึ่งมีโครงสร้างการทดสอบดังแสดงในภาพที่ 3



ภาพที่ 2 ความแปรปรวนที่ถูกอธิบายโดยองค์ประกอบทั้งหมด

ตารางที่ 2 ค่าน้ำหนักของ 5 องค์ประกอบหลักแรกของธาตุอาหารจากตัวอย่างดิน

องค์ประกอบ	PC1	PC2	PC3	PC4	PC5
อลูมินัม (Al)	**0.946**	0.042	0.090	0.022	0.014
แมงกานีส (Mn)	-0.076	-0.071	0.236	**0.788**	-0.114
เหล็ก (Fe)	**0.775**	0.017	-0.339	0.316	0.012
โครเมียม (Cr)	**0.937**	-0.058	-0.086	0.007	0.158
แมกนีเซียม (Mg)	0.109	0.013	-0.166	**0.874**	0.179
สังกะสี (Zn)	0.535	**0.649**	-0.020	-0.061	0.006
ทองแดง (Cu)	-0.042	0.268	**0.547**	-0.278	0.271
ตะกั่ว (Pb)	0.442	0.109	**-0.669**	0.206	0.208
โพแทสเซียม (K)	0.377	0.191	-0.299	0.306	**0.611**
โซเดียม (Na)	0.291	0.009	-0.071	**0.672**	0.293
แคลเซียม (Ca)	-0.240	**0.653**	0.084	0.323	-0.054
กรดฟัลวิก (FA)	0.022	0.310	-0.226	0.016	**0.812**
กรดฮิวมิก (HA)	0.028	0.039	-0.042	0.104	**0.794**
ค่าการแลกเปลี่ยนประจุบวก (CEC)	-0.125	0.022	**0.775**	0.091	-0.302
เปอร์เซ็นต์ดินเหนียว (% clay)	0.093	-0.037	**0.858**	0.157	-0.131
ไนโตรเจนรวม (TN)	0.117	**0.879**	0.014	-0.142	0.324
อินทรีย์คาร์บอน (OC)	-0.006	**0.931**	-0.024	-0.092	0.190
ค่าไอเกน	4.766	2.868	2.135	2.009	1.108
ความแปรปรวนสะสม	0.280	0.449	0.575	0.693	0.758



ภาพที่ 3 กระบวนการ MLR และ FANN โดยใช้องค์ประกอบหลัก

3. ความแม่นยำของโมเดล

ในการเปรียบเทียบความแม่นยำของทั้งสอง โมเดลจะวัดค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE), ค่าเฉลี่ยของความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute error: MAE), ความคลาดเคลื่อนร้อยละสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error: MAPE) ซึ่งค่าวัดที่ดีที่สุดมีค่าเท่ากับ 0 ดังนั้นวิธีการที่ให้ค่าวัดเหล่านี้น้อยกว่า แสดงถึงความแม่นยำที่มากกว่า ส่วนค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient: R ,Sousa และคณะ , 2007) ค่าวัดที่ดีที่สุดมีค่าเท่ากับ 1 ดังนั้นวิธีการที่ให้ค่าวัดมากกว่าแสดงถึงการอธิบายความผันแปร ของตัวแปรตามด้วยตัวแปรพยากรณ์ที่ใช้ได้มากกว่า ค่าวัดที่ใช้ในการเปรียบเทียบความแม่นยำของโมเดลคำนวณจากสมการที่ 2-5

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \quad (3)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\hat{Y}_i - Y_i\right|}{Y_i}\right)\times 100 \quad (4)$$

$$R = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}_i\right)^2 - \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}_i\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}_i\right)^2}} \quad (5)$$

โดยที่ Y_i คือปริมาณอินทรีย์วัตถุที่แท้จริงของข้อมูลลำดับที่ i, $\hat{Y}_i$ คือปริมาณอินทรีย์วัตถุที่ได้จากการพยากรณ์ของข้อมูลลำดับที่ i, $\overline{Y}_i$ คือค่าเฉลี่ยของปริมาณอินทรีย์วัตถุ และ n คือจำนวนข้อมูลทดสอบ

4. วิธีการถดถอยเชิงพหุคูณ

วิธีการถดถอยเชิงเส้นพหุคูณถูกใช้บ่อยในการพยากรณ์ข้อมูลที่มีลักษณะเป็นแบบต่อเนื่อง และหาความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระกับตัวแปรตาม สมการทั่วไปเป็นดังสมการที่ 6

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \quad (6)$$

โดยที่ $\hat{Y}$ คือ ค่าที่ได้จากการพยากรณ์

β_1 คือ พารามิเตอร์ที่ถูกประมาณค่าด้วยวิธีกำลังสองน้อยสุด (i=1, 2,..., p)

X_i คือ ตัวแปรพยากรณ์ (i=1, 2,..., n)

โดยใช้ชุดข้อมูลทั้งหมด 58 ชุด ซึ่งแบ่งออกเป็นสองชุดอย่างสุ่ม คือชุดข้อมูลสำหรับการเรียนรู้ (training set) จำนวน 52 ชุด และชุดข้อมูลทดสอบ (test set) จำนวน 6 ชุด สำหรับข้อมูลชุดทดสอบโมเดลได้แก่ ข้อมูลชุดที่ 29, 33, 34, 48, 50 และ 52 ตามลำดับ จะใช้ข้อมูลที่แบ่งนี้กับโมเดล FANN เช่นกัน

5. วิธีการโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม เป็นที่ยอมรับกันอย่างแพร่หลายในด้านการเป็นเทคโนโลยี สำหรับการแก้ปัญหาที่ยุ่งยากซับซ้อนและใช้กับปัญหาการไม่เป็นเชิงเส้นระหว่างตัวแปรตามและตัวแปรพยากรณ์ รวมถึงสามารถที่จะประยุกต์ในการเรียนรู้กับข้อมูลอิสระใหม่ในการพยากรณ์ตัวแปรตาม โดยทั่วไปแล้วโครงข่ายประสาทเทียม ประกอบด้วยจำนวนของการดำเนินการในขั้นแรกที่เรียกว่า นิวรอน (Neurons) เป็นคอนเนคชั่นระหว่างค่าน้ำหนักที่เชื่อมกันคือไซแนปส์ (Synapses) ค่าน้ำหนักที่เชื่อมต่อกันต้องมีการเรียนรู้และปรับค่าน้ำหนักตามฟังก์ชันของโครงข่าย (Network function) โมเดลโครงข่ายประสาทเทียม (Haykin, 1998) แสดงดังภาพที่ 4 สามารถอธิบายโมเดลโครงข่ายประสาทเทียมในเทอมคณิตศาสตร์ ได้ดังสมการที่ 7, 8 และ 9

$$u_k = \sum_{j=1}^{m} wk_i x_j \quad (7)$$

$$y_k = \varphi(u_k + b_k) \quad (8)$$

$$\upsilon_k = u_k + b_k \quad (9)$$

โดยที่ $x_1, x_2, ...x_n$ คือ ตัวแปรพยากรณ์หรืออินพุต

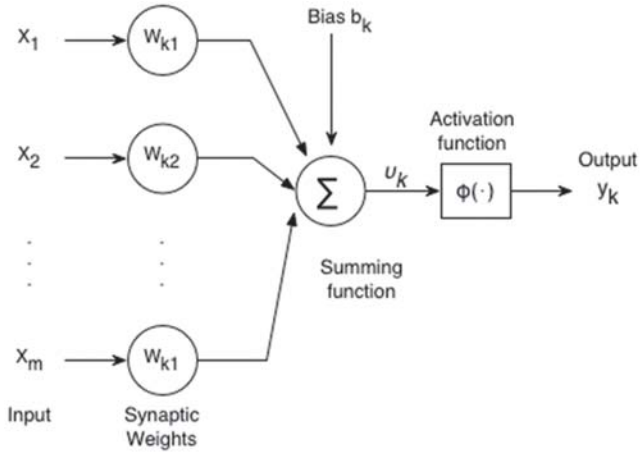
$w_{k1}, w_{k2}, ...w_{km}$ คือ ค่าน้ำหนักของนิวรอนที่ k

b_k คือ ค่าเอนเอียง (bias)

$\varphi(\cdot)$ คือ ฟังก์ชันการกระตุ้น

y_k คือ ตัวแปรตามหรือเอาท์พุต

υ_k คือ ผลรวมของตัวแปรพยากรณ์และค่าเอนเอียง



ภาพที่ 4 โมเดลโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมแบบที่มีการเชื่อมโยงไปข้างหน้า ซึ่งนิยมนำมาใช้มาก เนื่องจากโหนด ในชั้นแรกเชื่อมต่อกับโหนดในชั้นถัดไปเท่านั้น

จำนวนโหนดในชั้นแรกจะถูกกำหนดโดยจำนวนตัวแปรพยากรณ์หรืออินพุตในโมเดล ในขณะที่จำนวนโหนดในชั้นผลลัพธ์หรือเอาท์พุตเท่ากับจำนวนผลลัพธ์ที่ต้องการในโมเดล สิ่งสำคัญในการสร้างโมเดลคือการเลือกจำนวนโหนดและฟังก์ชันการแปลง (transfer function) โดยโครงข่ายประสาทเทียมที่ขึ้นอยู่กับตัวแบบในการศึกษานี้มีรูปแบบทั่วไปดังนี้

Organic Matter =FANN (PC1,PC2,PC3,PC4,PC5)

โดยที่ FANN คือฟังก์ชันของโครงข่ายประสาทเทียม ในการศึกษาเพื่อเลือกตัวฟังก์ชันการแปลง (transfer function) และหาอัลกอริทึมในการเรียนรู้ที่เหมาะสม งานวิจัยนี้จะใช้ hyperbolic tangent transfer function และใช้วิธีการฝึกโครงข่ายแบบ Levenberg-Marquardt algorithm (Matignon, 2007)

เนื่องจากการสร้างและการทดสอบโมเดล FANN ไม่มีข้อกำหนดตายตัวขึ้นอยู่กับรูปแบบของปัญหาหรืองานวิจัย ดังนั้นจึงต้องอาศัยวิธีการลองผิดลองถูก

ในการสร้างโมเดลด้วยวิธีการถดถอยเชิงเส้นพหุคูณและกระบวนการเรียนรู้ข้อมูลของวิธีการโครงข่ายประสาทเทียม ทำโดยใช้โปรแกรม SAS (Statistical Analysis System)

ผลการศึกษา

จากการสร้างโมเดล MLR โดยใช้วิธีการถดถอยพหุคูณแบบเป็นขั้นตอน (Stepwise multiple regression) และผลลัพธ์ที่ได้จากการทดสอบโมเดล จากตารางที่ 3 แสดงค่าประมาณพารามิเตอร์ที่ได้จากโมเดลวิธีการถดถอยเชิงพหุคูณ

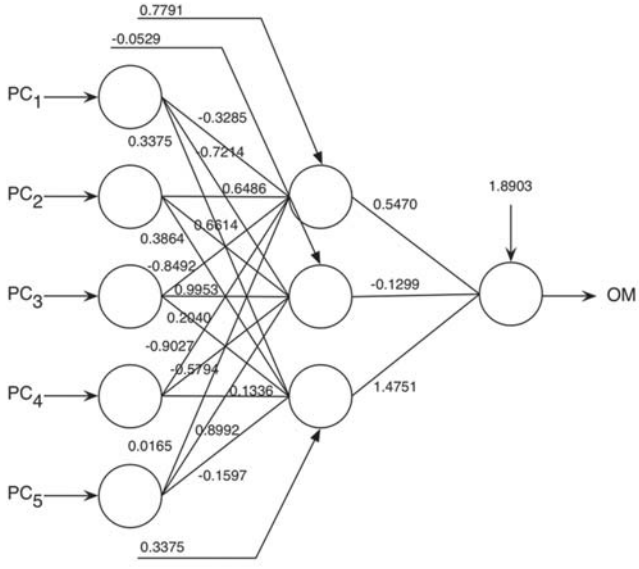
ตารางที่ 3 ค่าประมาณของพารามิเตอร์ของโมเดล MLR

โมเดล	ค่าคงที่	PC1	PC2	PC5
Parameter estimate	2.423	0.135	0.367	-0.184
Standard error	0.034	0.015	0.020	0.030

ผลจากการสร้างโมเดล FANN พบว่าโมเดลที่ให้ผลลัพธ์ดีที่สุดประกอบด้วย ชั้นอินพุต 5 โหนด ชั้นซ่อน 3 โหนด ชั้นเอาท์พุต 1 โหนด แสดงดังภาพที่ 5

ผลลัพธ์จากการพยากรณ์ชุดทดสอบที่เลือกมาอย่างสุ่ม เมื่อคำนวณค่าวัดความแม่นยำกับชุดทดสอบพบว่า ค่า RMSE, MAE, MBE และค่า R เมื่อเปรียบเทียบทั้งสองโมเดล แสดงดังตารางที่ 4 ซึ่งจะพบว่าโมเดล FANN ให้ค่าความแม่นยำในการวัดความคลาดเคลื่อนดีกว่า แต่มีค่า R ต่ำกว่าโมเดล MLR

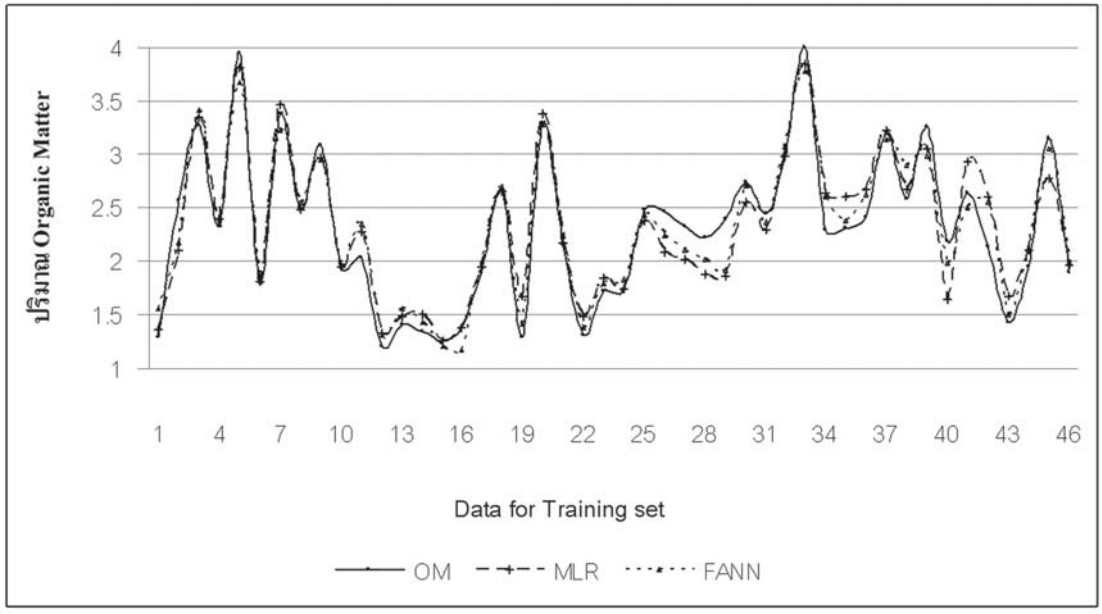
เมื่อลงจุดกราฟความสัมพันธ์ระหว่างค่าจริงของปริมาณอินทรีย์วัตถุกับค่าพยากรณ์ของชุดข้อมูลเรียนรู้ แสดงดังภาพที่ 6 จะเห็นว่าผลลัพธ์ที่ได้ใกล้เคียงกันแต่ค่าพยากรณ์ที่ได้จากโมเดล FANN มีค่าใกล้เคียงค่าจริงของปริมาณอินทรีย์วัตถุมากกว่าค่าพยากรณ์ที่ได้จากโมเดล MLR ซึ่งสอดคล้องกับผลลัพธ์การเปรียบเทียบประสิทธิภาพของโมเดล



ภาพที่ 5 โครงข่ายประสาทเทียมของการพยากรณ์ปริมาณอินทรีย์วัตถุในดิน

ตารางที่ 4 เปรียบเทียบความแม่นยำของโมเดล MLR และ FANN

ประสิทธิภาพ	MLR	FANN
RMSE	0.27	0.24
MAE	0.23	0.22
MAPE	9.39	8.57
R	0.91	0.87



ภาพที่ 6 เปรียบเทียบค่าพยากรณ์กับข้อมูลจริงจากโมเดล MLR และ FANN ในข้อมูลชุดเรียนรู้

อภิปรายผล

ในการสร้างโมเดลให้มีประสิทธิภาพ กระบวนการประมวลข้อมูลขั้นต้น (Data preprocessing) มีความสำคัญมาก เนื่องจากการมีข้อมูลที่ไม่จำเป็นในโมเดลน้อยยิ่งเป็นผลดี ซึ่งงานวิจัยนี้ใช้วิธี PCA เป็นเครื่องมือในการลดจำนวนมิติของข้อมูล และขจัดความสัมพันธ์ระหว่างตัวแปร นอกจากที่จะลดความยุ่งยากและความซับซ้อนของโมเดล แล้วยังทำให้โมเดลสามารถที่จะอธิบายความสัมพันธ์จริงของข้อมูลได้มากยิ่งขึ้น

โมเดลการถดถอยเชิงเส้นพหุคูณเป็นโมเดลที่นิยมใช้กันอย่างกว้างขวาง เนื่องจากเป็นโมเดลที่ไม่ยุ่งยาก ง่ายต่อการใช้งาน และใช้เวลาในการคำนวณไม่มากนัก ซึ่งให้ผลลัพธ์ใกล้เคียงกับผลลัพธ์ที่ได้จากโมเดล โครงข่ายประสาทเทียมแต่ยังไม่สามารถที่จะให้ผลลัพธ์ที่ดีกว่า เมื่อเปรียบเทียบทั้งสองโมเดล จะพบว่าทั้งสองโมเดลมีข้อได้เปรียบเสียเปรียบแตกต่างกันและความเหมาะสมของการเลือกใช้โมเดลในการแก้ปัญหา ขึ้นอยู่กับลักษณะของข้อมูลที่มี ซึ่งถ้าตัวแปรตามไม่มีความเป็นเชิงเส้น ทำให้โมเดลที่ได้จากวิธีการถดถอยเชิงเส้นพหุคูณให้ผลลัพธ์ไม่เป็นที่น่าพอใจ และโมเดลโครงข่ายประสาทเทียมอาจให้ผลลัพธ์การพยากรณ์เป็นที่น่าพอใจและมีความแม่นยำกว่า แต่การคำนวณในแต่ละโครงข่ายบางครั้งเสียเวลามากขึ้นอยู่กับจำนวนโหนดในแต่ละชั้น

สรุปผล

โมเดลการพยากรณ์ปริมาณอินทรีย์วัตถุในดินที่แม่นยำที่สุดในงานวิจัยนี้คือโมเดลที่ได้จากวิธีการโครงข่ายประสาทเทียม โดยมีชั้นอินพุต ชั้นซ่อนและชั้นเอาท์พุต เป็น 5-3-1 ตามลำดับ ซึ่งวิธีการนี้สามารถที่จะเรียนรู้ข้อมูลใหม่ที่ได้และสามารถเปลี่ยนแปลงไปตามสภาพแวดล้อม เหมาะสำหรับประยุกต์ใช้กับข้อมูลทางด้านการใช้ดินในด้านการเกษตร รวมถึงการตรวจสอบการเปลี่ยนแปลงปริมาณอินทรีย์วัตถุในดิน ซึ่งจะเป็นประโยชน์ในการจัดการและการวางแผนการใช้ดินในอนาคต

เอกสารอ้างอิง

กัลยา วานิชบัญชา. 2551. **การวิเคราะห์ข้อมูลหลายตัวแปร**. กรุงเทพฯ: บริษัทธรรมสาร.

คณาจารย์ภาควิชาปฐพีวิทยา. 2548. **ปฐพีวิทยาเบื้องต้น**. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์.

Haykin, Simon. 1998. **Neural Networks: A Comprehensive Foundation.** Upper Saddle River, NJ: Prentice Hall.

Ingleby, H.R. and Crowe, T.G. 2001. "Neural network models for predicting organic matter content in

Saskatchewan soils." **Canadian Biosystems Engineering.** Vol 43: 7.1-7.5.

Matignon, Randall.2007. **Data Mining Using SAS Enterprise Miner.** Hoboken, NJ: John Wiley & Sons, Inc.

Sousa,S.I.V., Martins, F.G., Alvim-Ferraze, M.C.M., Pereira, M.C.. 2007. "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations." **Environmental Modelling & Software.** vol 22: 97-103.

Stevenson, F.J. 1994. **Humus Chemistry.** 2nd John Wiley & Sons, Inc.

Yang,L., Dawson,C.W., Brown, M.R., Gell, M., 2006. "Neural network and GA approaches for dwelling fire occurrence prediction." **Knowledge-Based Systems.** Vol 19: 213-219.

>> สิริกัลยา ประมวล

จบการศึกษาหลักสูตรปริญญาวิทยาศาสตรบัณฑิต วิชาเอกสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร และกำลังศึกษาหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาสถิติประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

ปัจจุบันทำงานในตำแหน่ง ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร พระราชวังสนามจันทร์ จังหวัดนครปฐม

>> กมลชนก พานิชการ

จบการศึกษาหลักสูตร Ph.D. (Statistics) Montana State University, USA. หลักสูตร MS. (Statistics) Montana State University USA. หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาสถิติประยุกต์ มหาวิทยาลัยธรรมศาสตร์ และหลักสูตรวิทยาศาสตรบัณฑิต สาขาสถิติ มหาวิทยาลัยเชียงใหม่

ปัจจุบันทำงานในตำแหน่ง ผู้ช่วยศาสตราจารย์ ระดับ 8 อาจารย์ประจำภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร พระราชวังสนามจันทร์ จังหวัดนครปฐม ผลงานทางวิชาการ เช่น Model-Based Principal Components of Covariance Matrices and Soil Classification based on their Chemical Composition using Principal Component Analysis

>> นัทธีรา สรรมณี

จบการศึกษาหลักสูตร Ph.D. (Environmental Science) University of North Texas, USA. หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์สิ่งแวดล้อม จุฬาลงกรณ์มหาวิทยาลัย และหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิทยาศาสตร์ทั่วไป จุฬาลงกรณ์มหาวิทยาลัย

ปัจจุบันทำงานในตำแหน่ง ผู้ช่วยศาสตราจารย์ ระดับ 8 อาจารย์ประจำภาควิชาวิทยาศาสตร์สิ่งแวดล้อม คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร พระราชวังสนามจันทร์ จังหวัดนครปฐม ผลงานทางวิชาการ เช่น ผลกระทบของระดับการเกิดฮิวมิกของปุ๋ยหมักต่อการเจริญเติบโตของพืชบางชนิด (ทุนสนับสนุนการวิจัยสถาบันวิจัยและพัฒนา), โครงการเกษตรอินทรีย์: การฟื้นฟูคุณภาพดินด้วยวิธีเศรษฐกิจพอเพียง (ทุนสำนักงานคณะกรรมการการอุดมศึกษา)