รายงานสืบเนื่อง

การประชุมวิชาการและนำเสนอผลงานวิจัย ระดับชาติ ครั้งที่ ๑

"นวัตกรรมสร้างสรรค์ ศาสตร์พระราชา สู่การพัฒนาที่ยั่งยืน ไทยแลนด์ ๔.๐"

วันพฤหัสบดี ที่ ๑๗ กรกฎาคม ๒๕๖๐
ณ ห้องประชุมกันเกรา ชั้น ๓
ศูนย์ฝึกประสบการณ์วิชาชีพ ราชภัฏกรีนวิว
มหาวิทยาลัยราชภัฏร้อยเอ็ด

บัณฑิตวิทยาลัย
มหาวิทยาลัยราชภัฏร้อยเอ็ด
GRADUATE SCHOOL
ROI ET RAJABHAT UNIVERSITY

การประเมินแบบจำลองการกระจายของสิ่งมีชีวิตโดยการใช้ตัววัดแบบต่างๆ

Assessment of Species Distribution Models by Difference Measurements

*สุรัสวดี นางแล **
*ธนายุทธ ช่างเรือนงาม, ศิวรี สุดสนิท ***

บทคัดย่อ

การวิจัยครั้ง มีวัตถุประสงค์เพื่อศึกษาการประเมินแบบจำลองการกระจายของสิ่งมีชีวิต (Species Distribution Models) โดยการใช้ตัววัดแบบต่างๆ และเปรียบเทียบการประเมินตามการสุ่มจุดที่ไม่มีการค้นพบสิ่งมีชีวิตเทียม (Pseudo-absence points) ในระยะ 10, 20, 30, 40 และ 50 กิโลเมตร รอบจุดที่มีการค้นพบสิ่งมีชีวิต (Presence points) จากผลการศึกษาพบว่า การสุ่มเลือกจุดการไม่ค้นพบสิ่งมีชีวิตเทียมในระยะที่แตกต่างออกไปจากจุดที่มีการค้นพบสิ่งมีชีวิตมีผลกระทบต่อแบบจำลอง กล่าวคือ การสุ่มจุดในระยะที่ห่างจากจุดการค้นพบสิ่งมีชีวิตไม่ควรสุ่มในระยะใกล้หรือระยะไกลเกินไป ซึ่งระยะที่เหมาะสมคือ 20 – 30 กิโลเมตร รอบจุดที่มีการค้นพบสิ่งมีชีวิต แต่ทั้งนี้ทั้งนั้น ควรพิจารณาตัวปัจจัยที่นำมาสร้างแบบจำลองด้วย โดยเฉพาะปัจจัยทางด้ายภูมิศาสตร์กายภาพ

ABSTRACT

The purposes of this research were to study assessment of species distribution models (SDMs) by difference measurements and compare various measurements on SDMs that had randomly pseudo-absence inside the circles at each radius (10, 20, 30, 40 and 50 km) from presence points. The results showed that the generating pseudo-absence points at each radius affect the model. The random pseudo-absence points should not be random in the near or far radius too. The optimal range is 20 to 30 kilometers around the presence points. However, it should consider the factors that contribute to modeling especially the geographical factors.

คำสำคัญ : *การประเมิน, แบบจำลองการกระจายของสิ่งมีชีวิต, จุดที่ไม่มีการค้นพบสิ่งมีชีวิตเทียม*

Keyword(s): *Assessment, Species Distribution Models, Pseudo-absence Points*

บทนำ

แบบจำลองการกระจายของสิ่งมีชีวิต (Species Distribution Models : SDMs) ถือว่าเป็นเครื่องมือสำคัญในการพยากรณ์ว่าสิ่งมีชีวิตมีการกระจายตัวอย่างไรบ้าง ซึ่งเป็นประโยชน์ต่อการวางแผนการอนุรักษ์ การสำรวจ การศึกษาทางด้านวิวัฒนาการหรือการตรวจสอบผลกระทบของการเปลี่ยนแปลงสภาพภูมิอากาศ (Thuiller et al., 2005; Guisan and Thuiller, 2005; Engler et al., 2004; Marini et al., 2009) โดยแบบจำลองการกระจายของสิ่งมีชีวิตจะแสดงถึงความสัมพันธ์ระหว่างสิ่งมีชีวิตและปัจจัยทางสิ่งแวดล้อม ดังนั้น ในการสร้างแบบจำลองการกระจายของสิ่งมีชีวิตโดยทั่วไปจะอาศัยข้อมูลการค้นพบสิ่งมีชีวิตที่ต้องการศึกษา ซึ่งจัดเก็บในรูปแบบจุดพิกัดที่ค้นพบสิ่งมีชีวิตนั้นๆ (Presence Points) และจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิต (Absence Points) แต่ในทางปฏิบัติส่วนมากมักบันทึกเพียงข้อมูลจุดพิกัดที่ค้นพบสิ่งมีชีวิตเท่านั้น ทำให้การสร้างแบบจำลองที่ต้องอาศัยทั้งจุดพิกัดที่มีการค้นพบสิ่งมีชีวิตและจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิต อาทิ เช่น การสร้างแบบจำลองการกระจายของสิ่งมีชีวิตโดยวิธี Generalized Linear Models (GLMs) หรือ Generalized Additive Models (GAMs) ต้องมีการสร้างจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิตเทียม (Pseudo-absence Points) เพื่อใช้ในการสร้างแบบจำลองดังกล่าว

** อาจารย์ประจำโปรแกรมวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงราย*
*** อาจารย์ประจำโปรแกรมวิชาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงราย*

นอกเหนือจากการสร้างแบบจำลองการกระจายของสิ่งมีชีวิตโดยการสร้างจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิตเทียมแล้ว ขั้นตอนที่สำคัญอีกขั้นตอนหนึ่งคือการประเมินความถูกต้องของแบบจำลอง (Model Assessment) ซึ่งเป็นขั้นตอนการเปรียบเทียบผลลัพธ์ที่ได้จากการพยากรณ์กับผลลัพธ์ที่แท้จริงว่ามีความถูกต้องมากน้อยเพียงใด และวิธีการที่นิยมใช้สำหรับข้อมูลที่ประกอบไปด้วยจุดพิกัดที่ค้นพบสิ่งมีชีวิตและจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิตคือตาราง Confusion Matrix ซึ่งเป็นตาราง 2 ทางขนาด 2 x 2 ดังแสดงในภาพที่ 1 ซึ่งสามารถคำนวณค่าที่ใช้ในการประเมินความถูกต้องได้หลายแบบ อาทิเช่น Kappa, Commission error, Omission error, Accuracy หรือ Sensitivity (Drew, Wiersma and Huettmann. 2011) นอกจากนี้วิธีการประเมินความถูกต้องของแบบจำลองอีกวิธีหนึ่งที่นิยมใช้คือ การหาพื้นที่ใต้กราฟ ROC (Receiver - operating characteristic) หรือเรียกสั้นๆ ว่าค่า AUC (Area under the ROC curve) ซึ่งกราฟ ROC เป็นกราฟแสดงความสัมพันธ์ระหว่างข้อมูลที่ทำนายถูก (True Positive) ในแนวแกน Y และทำนายผิด (False Positive) ในแนวแกน X ถ้าพื้นที่ใต้กราฟ ROC มีค่าเข้าใกล้ 1 แสดงว่าแบบจำลองมีประสิทธิภาพดี แต่เนื่องจากแบบจำลองการกระจายของสิ่งมีชีวิตมีการสร้างจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิตเทียมขึ้นแทนจุดพิกัดที่ไม่มีค้นพบสิ่งมีชีวิตจริง ดังนั้น ในการศึกษาครั้งนี้จึงเป็นการศึกษาเปรียบเทียบการประเมินแบบจำลองการกระจายของสิ่งมีชีวิตโดยการใช้ตัววัดแบบต่างๆ ซึ่งแบบจำลองการกระจายของสิ่งมีชีวิตที่ใช้ในการศึกษาครั้งนี้คือ วิธี GLMs

		Observed	
		Present	Absent
Predict	Present	a	b
	Absent	c	d

ภาพ 1 ตาราง Confusion Matrix

วัตถุประสงค์

เพื่อศึกษาและเปรียบเทียบการประเมินแบบจำลองการกระจายของสิ่งมีชีวิตโดยการใช้ตัววัดที่แตกต่างกัน

วิธีดำเนินการการวิจัย

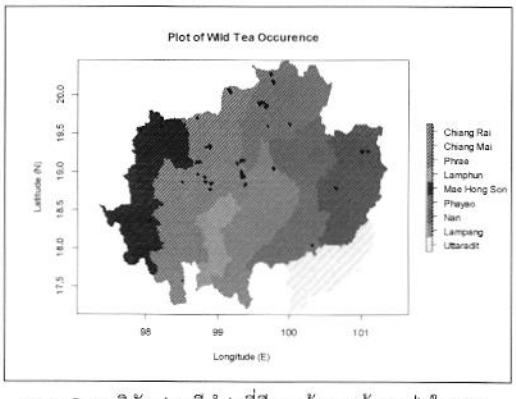
ผู้วิจัยได้ดำเนินการตามลำดับขั้นตอนดังต่อไปนี้

1. ข้อมูลของสิ่งมีชีวิต (Presence data)

ชาป่า หรือ ชาพันธุ์พื้นเมือง เป็นชาสายพันธุ์อัสสัม (Camillia sinensis var. assamica) โดยข้อมูลชาป่าในเขตภาคเหนือถูกเก็บรวบรวมโดยสถาบันชา มหาวิทยาลัยแม่ฟ้าหลวง เป็นส่วนหนึ่งของโครงการเก็บและกำหนดพันธุ์ชาที่เหมาะสมในพื้นที่ปลูกภาคเหนือของประเทศไทย (สายลม สัมพันธ์เวชโสภา และคณะ. 2550)

2. พื้นที่ที่ทำการศึกษา

การศึกษาครั้งนี้เป็นการสร้างแบบจำลองการกระจายตัวของต้นชาป่าในเขตพื้นที่ภาคเหนือตอนบน ประกอบด้วย 9 จังหวัด ดังต่อไปนี้ เชียงราย เชียงใหม่ พะเยา แพร่ น่าน ลำพูน ลำปางอุตรดิตถ์และแม่ฮ่องสอน ดังแสดงในภาพที่ 2



ภาพ 2 จุดพิกัด (จุดสีดำ) ที่มีการค้นพบต้นชาป่าในเขตพื้นที่ภาคเหนือตอนบนจำนวน 185 จุด

3. วิธีการสร้างจุด Pseudo-absence

การสร้างจุด Pseudo-absence โดยการสุ่มในพื้นที่ที่กำลังศึกษา เป็นวิธีการสุ่มจุด Pseudo-absence ภายในพื้นวงกลมรอบๆ จุดที่มีการค้นพบสิ่งมีชีวิต (Presence Points) ในรัศมีที่แตกต่างกัน

4. ปัจจัยทางด้านสิ่งแวดล้อม

ในการศึกษาครั้งนี้ เลือกใช้เฉพาะข้อมูลสภาพภูมิอากาศ ซึ่งเป็นข้อมูลออนไลน์จาก http://www.worldclim.org โดยสร้างมาจากอุณหภูมิและปริมาณน้ำฝน มีทั้งหมด 19 ตัวแปรด้วยกัน แต่เลือกใช้ตัวแปร Isothermality (Bio3), อุณหภูมิเฉลี่ยในไตรมาสที่ฝนตกชุกที่สุด (Bio8), ปริมาณน้ำฝนรายปี (Bio12), ปริมาณน้ำฝนในเดือนที่แห้งแล้งที่สุด (Bio14)

5. แบบจำลองการกระจายของสิ่งมีชีวิต

ในการศึกษาครั้งนี้ ใช้แบบจำลองการกระจายของสิ่งมีชีวิตโดยวิธี GLMs ซึ่งในการสร้างแบบจำลองการวิเคราะห์สมการถดถอยโลจิสติก เนื่องจากตัวแปรตาม y ที่เป็นค่าของ Presence Points และ Pseudo-absence Points จะมีค่าเท่ากับ 1 เมื่อ y อยู่ในจุดพิกัดของ Presence Points และค่า y จะมีค่าเท่ากับ 0 เมื่อ y อยู่ในจุดพิกัดของ Pseudo-absence Points โดยมีรูปแบบสมการ คือ

$$E(Y) = \frac{e^{X\beta}}{1+e^{X\beta}}$$

และมีรูปแบบพยากรณ์ คือ

$$\text{Logit}(\mu) = X\beta$$

6. การประเมินความถูกต้องของแบบจำลอง

ในการศึกษาครั้งนี้ได้ใช้ตัวประเมินความถูกต้องดังนี้

1) AUC (Area under the ROC curve) เป็นการหาพื้นที่ใต้กราฟ ROC ซึ่งเป็นกราฟแสดงความสัมพันธ์ระหว่างข้อมูลที่ทำนายถูกในแนวแกน Y และทำนายผิดในแนวแกน X

2) Accuracy เป็นพิจารณาผลของการพยากรณ์ที่ถูกต้องทั้งหมดจากตาราง Confusion Matrix (ภาพ 1) โดยเป็นอัตราส่วนของการทำนายที่ถูกต้องต่อการทำนายทั้งหมด ดังนี้

$$Accuracy = \frac{a+d}{a+b+c+d}$$

3) Sensitivity คือ การพิจารณาจำนวนความถูกต้องของการทำนายต่อจำนวนจุดที่มีการค้นพบสิ่งมีชีวิตทั้งหมด

$$Sensitivity = \frac{a}{a+c}$$

4) Specificity คือ การพิจารณาจำนวนความถูกต้องของการทำนายต่อจำนวนจุดที่ไม่มีการค้นพบสิ่งมีชีวิตทั้งหมด

$$Specificity = \frac{d}{b+d}$$

5) Positive Predictive Power (PPP) คือ ความสามารถในการทำนายจุดที่มีการค้นพบสิ่งมีชีวิตได้อย่างถูกต้องต่อการทำนายจุดที่มีการค้นพบสิ่งมีชีวิตทั้งหมด

$$PPP = \frac{a}{a+b}$$

6) Negative Predictive Power (NPP) คือ ความสามารถในการทำนายจุดที่ไม่มีการค้นพบสิ่งมีชีวิตได้อย่างถูกต้องต่อการทำนายจุดที่ไม่มีการค้นพบสิ่งมีชีวิตทั้งหมด

$$NPP = \frac{d}{c+d}$$

สรุปผลการวิจัย

จากการสร้างแบบจำลองโดยวิธี GLMs พบว่าการสร้างจุดที่ไม่มีการค้นพบสิ่งมีชีวิตในระยะต่างๆ มีผลต่อแบบจำลอง จากตารางที่ 1 พบว่า ตัวแปรแต่ละตัวมีผลต่อแบบจำลองอย่างมีนัยสำคัญทางสถิติ (p-value < 0.05) แตกต่างกันตามระยะของรัศมี กล่าวคือตัวแปรอุณหภูมิเฉลี่ยในไตรมาสที่ฝนตกชุกที่สุด (Bio8) มีผลต่อแบบจำลองในทุกๆ ระยะ 10 – 50 กิโลเมตร ตัวแปร Isothermality (Bio3) มีผลต่อแบบจำลองในระยะ 10 – 30 กิโลเมตร ตัวแปรปริมาณน้ำฝนรายปี (Bio12) มีผลต่อแบบจำลองในระยะ 40 – 50 กิโลเมตร มีแต่เพียงตัวแปรปริมาณน้ำฝนในเดือนที่แห้งแล้งที่สุด (Bio14) ที่ไม่มีผลต่อแบบจำลองในทุกๆ ระยะ ดังนั้น การเลือกระยะของรัศมีเพื่อสร้างจุดที่ไม่มีการค้นพบสิ่งมีชีวิตเทียม จึงเป็นสิ่งจำเป็นที่ต้องนำมาประกอบการพิจารณา

ตาราง 1 ค่าสัมประสิทธิ์ของแบบจำลอง GLMs ในระยะรัศมีต่างๆ

ตัวแปร	ค่าสัมประสิทธิ์									
	10 กิโลเมตร		20 กิโลเมตร		30 กิโลเมตร		40 กิโลเมตร		50 กิโลเมตร	
Intercept	-1.3810×10	*	-6.6128		-1.9137		-0.7803		10.3390	
Bio 3	3.3850×10^{-1}	*	0.2357	*	0.2034	*	0.1785		-0.0381	
Bio 8	-2.3820×10^{-1}	*	-0.0322	*	-0.0461	*	-0.0483	*	-0.0478	*
Bio 12	9.4330×10^{-3}		0.0001		0.0005		0.0015	*	0.0014	*
Bio 14	-4.8630×10^{-2}		-0.0203		-0.0323		-0.0760		-0.0574	

จากแบบจำลอง GLMs ที่สร้างขึ้นมาเพื่อพยากรณ์พื้นที่ที่เหมาะสมต่อการกระจายของสิ่งมีชีวิตนั้น เมื่อนำมาตรวจสอบความถูกต้องของการพยากรณ์ของแบบจำลองในแต่ละระยะ โดยใช้ตัววัดความถูกต้อง คือค่า AUC, Specificity, Sensitivity, PPP และ NPP ดังแสดงในภาพที่ 3 พบว่า ทุกๆ ค่ามีค่าเพิ่มขึ้น เมื่อระยะของรัศมีเพิ่มขึ้น แต่มีจุดสังเกตคือ ที่ระยะรัศมี 10 กิโลเมตร ค่า AUC จะมีค่าน้อยกว่าที่ระยะ 20 กิโลเมตร ซึ่งสอดคล้องกับค่า NPP และ Sensitivity ที่เป็นไปในทิศทางเดียวกันแต่ค่า Specificity กลับมีทิศทางในตรงกันข้ามกล่าวคือ ค่า Specificity ในระยะ 10 กิโลเมตรมีค่ามากกว่าในระยะ 20 กิโลเมตร ส่วนค่า PPP มีค่าที่ค่อนข้างคงที่ นอกจากนี้จะสังเกตได้ว่า ที่ระยะ 40 กิโลเมตร ค่า Sensitivity จะมีค่าลดลงจากระยะ 30 กิโลเมตร ในขณะที่ค่า AUC, Specificity และ PPP ที่ระยะ 40 กิโลเมตรจะมีค่าเพิ่มขึ้นจากระยะ 30 กิโลเมตร