

ĐÁNH GIÁ KHẢ NĂNG ỨNG DỤNG MÔ HÌNH HỒI QUY LOGISTIC ĐA THỨC ĐỂ THÀNH LẬP BẢN ĐỒ ĐẤT TỈNH BẮC NINH

Đoàn Thanh Thủy, Lê Thị Giang*

Khoa Quản lý đất đai, Học viện Nông nghiệp Việt Nam

Email: lethigiang@vnua.edu.vn*

Ngày nhận bài: 08.01.2016

Ngày chấp nhận: 29.04.2016

TÓM TẮT

Mô hình Hồi quy logistic đa thức (Multinomial logistic regression - MLR) đang ngày càng được ứng dụng rộng rãi để thành lập bản đồ đất trên thế giới. Mô hình này sẽ giúp tiết kiệm chi phí và thời gian do yêu cầu số lượng các điểm mẫu đầu vào (phần diện đất) ít hơn nhiều so với phương pháp thành lập bản đồ truyền thống. Đề tài nghiên cứu khả năng ứng dụng của MLR để thành lập bản đồ đất ở tỉnh Bắc Ninh bằng việc đánh giá kết quả của mô hình. Mô hình có 9 biến đầu vào bao gồm: loại đất, loại hình sử dụng đất, độ cao, độ dốc, chỉ số thực vật (Normalized difference vegetation index - NDVI), chỉ số thực vật vuông góc (Perpendicular vegetation index - PVI), tỷ số chỉ số thực vật (Ratio vegetation index - RVI), chỉ số ẩm ướt địa hình và chỉ số ẩm ướt SAGA. Mô hình MLR được sử dụng để dự báo loại đất theo 2 mức: mức Nhóm đất chính và mức Loại đất trung gian. Chất lượng bản đồ kết quả được đánh giá dựa trên 3 chỉ số: Độ chính xác phân loại, Chỉ số đa dạng và Chỉ số kết hợp. Kết quả cho thấy ở mức Nhóm đất chính, mô hình MLR dự báo bản đồ có độ chính xác cao nhưng thấp hơn so với mức Loại đất trung gian. Chỉ số kết hợp cho thấy mô hình MLR cho kết quả tốt nhất ở mức Loại đất trung gian.

Từ khóa: Bản đồ đất, chỉ số đa dạng, độ chính xác phân loại, hồi quy logistic đa thức.

Evaluating the Potential of Digital Soil Mapping Method to Map Soil Types in Bac Ninh Province

ABSTRACT

The multinomial logistic regression (MLR) model has been being widely used for soil mapping because it could help save time and costs for collecting and analyzing soil data points compared to conventional methods. This research aimed to assess the potential of mapping soil types in Bac Ninh by assessing the performance of Multinomial Logistic Regression (MLR) model in predicting soil types. Nine predictive variables were derived from the ancillary data including land use, altitude, slope, NDVI, PVI, RVI, Topographic Wetness Index and SAGA Wetness Index. MLR model was constructed to predict soil classes at 2 levels: WRB-Reference Soil Group and intermediate level of Soil Group between Reference Soil Group and the full WRB soil name. The map quality was evaluated by the soil map purity estimated with an independent validation dataset. The diversity indices were calculated to assess the information content of the resultant maps. Assessment of the model's prediction was based on the soil map purity, the Shannon's entropy and a combined index. MLR yielded high map purity at the level of Reference Soil Group. When the taxonomic level changed from Reference Soil Group level to intermediate level, the map purity decreased while the value of the diversity indices increased. Therefore, soil mapping using MLR in predicting Reference Soil Group might be more efficient. However, at intermediate level, the model predicted higher diversity of soil map and thus the informative value estimated by the combined index was higher.

Keywords: Diversity index, multinomial logistic regression, soil map, soil map purity.

1. ĐẶT VẤN ĐỀ

Bản đồ đất là loại bản đồ chuyên đề thể hiện sự phân bố theo không gian các loại đất có

trong một vùng lãnh thổ hay một đơn vị hành chính. Bản đồ đất là tài liệu cơ bản quan trọng, là căn cứ xây dựng kế hoạch phát triển nông nghiệp, lâm nghiệp, nuôi trồng thủy sản. Việc

thành lập các bản đồ đất tỷ lệ lớn và độ chính xác cao bằng phương pháp điều tra lấy mẫu truyền thống đòi hỏi nhiều chi phí và thời gian. Trong điều kiện Việt Nam - một nước nông nghiệp đang phát triển, việc ứng dụng mô hình MLR trong thành lập bản đồ đất có thể trở thành giải pháp hiệu quả thay thế cho các phương pháp truyền thống vì khả năng xây dựng các bản đồ có tỷ lệ lớn với chi phí thấp do yêu cầu ít các điểm mẫu phân tích đất.

MLR sử dụng mô hình số học để dự báo sự phân bố theo không gian và thời gian của các đặc tính đất đai dựa trên cơ sở mối quan hệ giữa đất và các yếu tố hình thành đất bao gồm: khí hậu, sinh vật trong đất (cây trồng, vi sinh vật, con người), địa hình, đá mẹ, thời gian và vị trí trong không gian (Lagacherie and McBratney 2007). Với sự bùng nổ của các công nghệ hiện đại như GIS, viễn thám và GPS, việc xác định các yếu tố ảnh hưởng đến sự hình thành của đất làm đầu vào cho các mô hình dự báo loại đất đang trở nên dễ dàng hơn, dẫn đến việc phát triển mạnh mẽ các mô hình MLR trong lĩnh vực khoa học đất. Đã có nhiều nghiên cứu trên thế giới sử dụng mô hình khác nhau để xây dựng các bản đồ đặc tính của đất hoặc bản đồ loại đất, cập nhật bản đồ đất đã có sẵn, hoặc đánh giá biến động lớp phủ,... (Carré and Girard 2002; Turetta, Mendoca Santos *et al.*, 2006; Kempen, Brus *et al.*, 2009; Yang, Jiao *et al.*, 2011). Vì vậy, nghiên cứu sử dụng mô hình Hồi quy logistic đa thức để xây dựng bản đồ đất ở tỉnh Bắc Ninh có ý nghĩa khoa học và thực tiễn.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Khu vực nghiên cứu

Bắc Ninh là một tỉnh thuộc vùng đồng bằng Bắc Bộ có tổng diện tích khoảng 82.300ha. Bắc Ninh có khí hậu nhiệt đới gió mùa với lượng mưa hàng năm khoảng 1.500mm, nhiệt độ trung bình 23°C. Địa hình của tỉnh tương đối bằng phẳng, ít chia cắt, hướng dốc là Tây Bắc - Đông Nam.

2.2. Vật liệu nghiên cứu

- Dữ liệu phân tích phẫu diện đất
- Ảnh SPOT khu vực Bắc Ninh năm 2010 với độ phân giải 20m.

- Mô hình số độ cao (DEM) độ phân giải 20m khu vực Bắc Ninh được xây dựng từ bản đồ đường đồng mức với khoảng cao đều 20m bằng phần mềm ArcGIS.

- Bản đồ hiện trạng sử dụng đất của tỉnh Bắc Ninh năm 2010 tỷ lệ 1:25.000.

Các bản đồ đầu vào được xử lý và thực hiện chồng xếp trong phần mềm ArcGIS.

2.3. Phương pháp nghiên cứu

2.3.1. Thu thập dữ liệu đầu vào

Để xây dựng mô hình, cần thiết phải thu thập các biến về các yếu tố ảnh hưởng đến sự hình thành đất làm đầu vào bao gồm:

- Biến loại đất: các điểm mẫu phân tích phẫu diện đất: bao gồm 537 điểm. Tại mỗi điểm, mẫu đất được lấy để mô tả và phân tích trong phòng thí nghiệm, từ đó được phân loại theo 2 mức sử dụng hệ phân loại của WRB (WRB 2006):

+ Mức nhóm đất chính: gồm 5 nhóm là Đất phù sa (402 điểm), Đất xám (58 điểm), Đất cát (7 điểm), Đất Glây (15 điểm) và Đất có tầng loang lổ (55 điểm)

+ Mức Loại đất trung gian: được phân loại bởi người thực hiện gồm 15 loại cấu thành bởi tên Nhóm đất chính kèm các đặc tính liên quan đến sử dụng đất như: độ kiềm, độ ẩm, tầng loang lổ và thành phần cơ giới đất.

- Biến địa hình: DEM được sử dụng để tính toán các biến địa hình bằng phần mềm Saga GIS: Độ cao, Độ dốc, Chỉ số ẩm ướt địa hình và Chỉ số ẩm ướt SAGA (Olaya 2004).

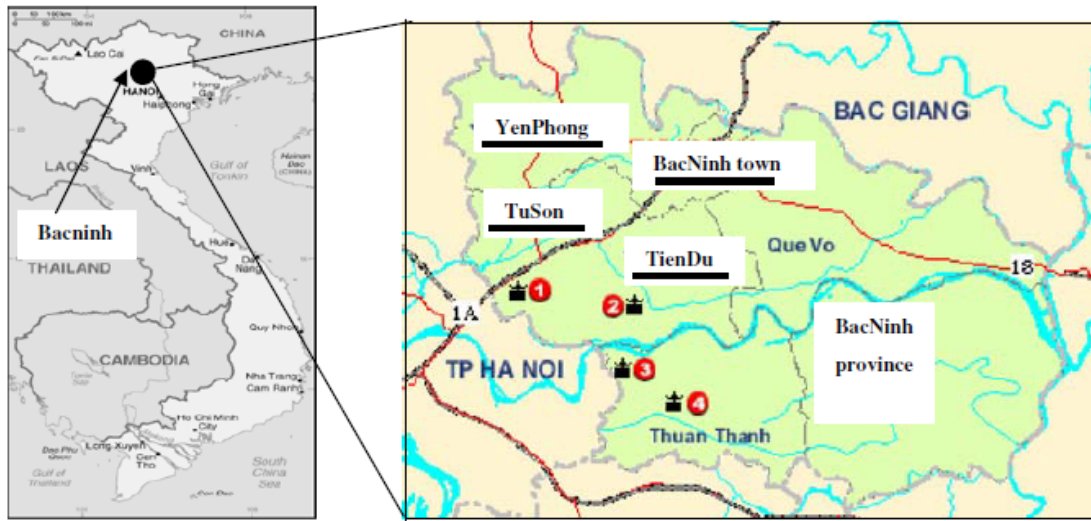
- Biến thực vật: ảnh SPOT được sử dụng để tính toán các chỉ số thực vật bao gồm: NDVI, PVI và RVI sử dụng phần mềm ArcGIS.

- Biến loại hình sử dụng đất: được xác định từ bản đồ hiện trạng sử dụng đất.

Bảng 1 tóm tắt các biến đầu vào cho hai mô hình MLR và ANNs để dự báo loại đất.

2.3.2. Xây dựng mô hình dự báo

Mô hình dự báo là Hồi quy logistic đa thức được ứng dụng để dự báo bản đồ đất của tỉnh Bắc Ninh theo 2 mức phân loại: mức Nhóm đất



Hình 1. Vị trí địa lý của tỉnh Bắc Ninh

Bảng 1. Dữ liệu đầu vào

Bản đồ đầu vào	Tên biến	Độ phân giải/Tỷ lệ
Loại đất	Soil	1:25.000
Mô hình số độ cao	DEM	25m
Bản đồ độ dốc	SLOPE	25m
Bản đồ chỉ số ẩm ướt Saga	SAGAWET	25m
Bản đồ chỉ số ẩm ướt địa hình	WETNESS	25m
Bản đồ NDVI	NDVI	25m
Bản đồ PVI	PVI	25m
Bản đồ RVI	RVI	25m
Bản đồ hiện trạng sử dụng đất	LU	1: 25.0000

chính và mức Loại đất trung gian. Từ tổng số 537 điểm mẫu đất đầu vào, sử dụng 484 điểm làm đầu vào cho mô hình, 53 điểm còn lại sẽ được sử dụng như một bộ điểm mẫu độc lập để đánh giá độ chính xác của mô hình.

Mô hình hồi quy logistic đa thức được sử dụng để xem xét mối quan hệ giữa biến phụ thuộc (Nhóm đất chính hoặc Loại đất trung gian) với các biến độc lập bao gồm: DEM, SLOPE, SAGAWET, WETNESS, NDVI, PVI, RVI và LU. Mô hình thuộc nhóm các mô hình hồi quy tuyến tính nhưng được sử dụng để dự báo các biến thứ bậc.

Gọi π_{ij} là khả năng để 1 điểm ảnh i thuộc nhóm đất j trong m nhóm đất ($j = \dots, m$), π_{i1} là

khả năng để 1 điểm ảnh i thuộc nhóm đất tham chiếu. Trong mô hình MLR để dự báo nhóm đất chính, Đất phù sa ($j = 1$) là nhóm đất tham chiếu vì nó chiếm số lượng lớn nhất trong tổng số mẫu đầu vào (402 trong số 537 mẫu đầu vào là Đất phù sa). Tương tự trong mô hình dự báo loại đất trung gian, loại đất trung gian Đất phù sa 4 là loại đất tham chiếu. Mô hình hồi quy logistic đa thức phát biểu (Fagerland, Hosmer *et al.*, 2008):

$$\ln \frac{\pi_{ij}}{\pi_{i1}} = \alpha_j + \beta_{1j}x_{i1} + \dots + \beta_{kj}x_{ik}, j = 2, \dots, m$$

Trong đó: x_i là các vectơ biến độc lập, α và β là các tham số của phương trình liên hệ. Mô hình được chạy trong môi trường phần mềm R-

project sử dụng gói “nnet”. Khi mô hình đã được thiết lập, chúng ta có thể tính toán khả năng xuất hiện của một Nhóm đất chính hoặc Loại đất trung gian nào đó tương quan với nhóm đất tham chiếu từ các biến môi trường đầu vào.

2.3.3. Xác nhận kết quả của mô hình

- Đánh giá chất lượng bản đồ

Chất lượng của bản đồ kết quả từ hai mô hình ANNs và MLR được đánh giá thông qua *Độ chính xác phân loại* (Map purity) được tính toán sử dụng bộ điểm mẫu độc lập (53 điểm) bằng một ma trận nhầm lẫn với phương trình:

$$p = \sum_{u=1}^U \frac{A_{UU}}{A}$$

Trong đó: p là độ chính xác phân loại, A_{UU} là số điểm mẫu được phân loại đúng so với thực tế thuộc loại U và A là tổng số điểm mẫu (53 điểm). Giá trị p càng gần tới 1 thì độ chính xác của bản đồ càng cao (Finke, 2011).

- Đánh giá chất lượng thông tin của bản đồ

Chất lượng thông tin của bản đồ được đánh giá thông qua *Chỉ số đa dạng của bản đồ*

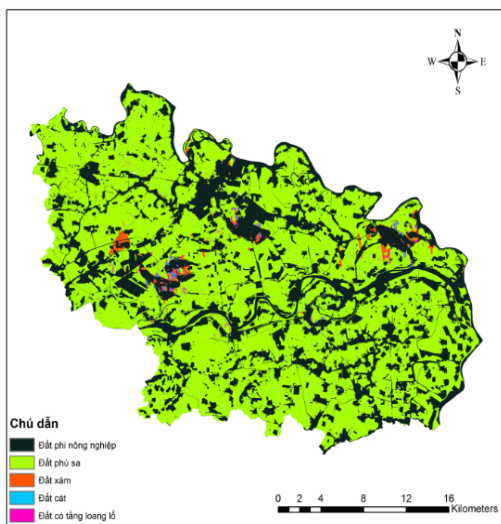
(Shannon's entropy) được xác định bởi phương trình:

$$H = -\sum_{i=1}^S p_i \times \ln p_i$$

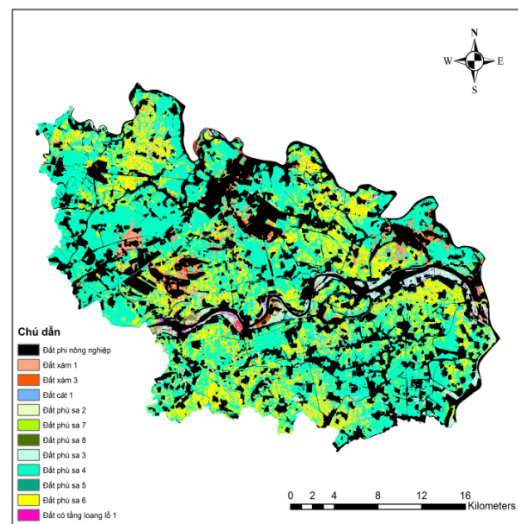
Trong đó: p_i là tỷ lệ diện tích vùng được dự báo thuộc nhóm i trên tổng diện tích của bản đồ. Chỉ số đa dạng được dùng để so sánh sự đa dạng về nhóm đất chính (hoặc loại đất trung gian) được dự báo bởi các mô hình. Giá trị H càng gần với $1/S$ (S: tổng số các nhóm đất trên khu vực nghiên cứu) thì thông tin trên bản đồ càng đa dạng. (Martin and Rey, 2000).

- Chỉ số kết hợp

Trong thực tế, mục tiêu của việc thành lập bản đồ chính là đưa ra được các bản đồ có độ chính xác cao và thể hiện được tính đa dạng của thông tin. Vì vậy, chỉ số kết hợp được tính toán bằng tích số của *Độ chính xác phân loại* với *Chỉ số đa dạng của bản đồ* là một yếu tố quan trọng để đánh giá toàn diện chất lượng của bản đồ kết quả từ mô hình MLR.



Hình 2. Bản đồ nhóm đất chính



Hình 3. Bản đồ loại đất trung gian

Bảng 2. Kết quả đánh giá độ chính xác của bản đồ

Mức phân loại	Độ chính xác	Chỉ số đa dạng	Chỉ số kết hợp
Nhóm đất chính	0,73	0,20	0,15
Loại đất trung gian	0,39	1,21	0,41

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Bản đồ dự báo nhóm đất chính và loại đất trung gian

Hình 2 là bản đồ dự báo nhóm đất chính từ mô hình MLR. Dễ dàng nhận thấy, nhóm đất phù sa chiếm phần lớn diện tích của tỉnh Bắc Ninh vì đây là một tỉnh thuộc vùng đồng bằng sông Hồng. Các nhóm đất còn lại (đất Xám, đất Cát và đất có tầng loang lổ) được dự báo với diện tích rất hạn chế, đặc biệt nhóm đất Glây không xuất hiện theo dự báo của mô hình. Điều này được giải thích bởi số lượng các điểm mẫu đầu vào chủ yếu thuộc nhóm đất phù sa (75%) trong khi các nhóm đất khác chỉ chiếm khoảng (25%). So sánh với mô hình số độ cao của tỉnh Bắc Ninh, nhóm đất Xám được dự báo ở những vị trí có độ cao khá lớn, điều này hợp lý vì nhóm đất này thường được hình thành ở những khu vực có địa hình cao ở vùng nhiệt đới ẩm (FAO 2001).

Hình 3 là bản đồ dự báo loại đất trung gian từ mô hình MLR. Ở mức phân loại thấp hơn với nhiều loại đất hơn, mô hình cho kết quả dự báo 11 trên 15 loại đất. Loại đất phù sa chua (Đất phù sa 4) xuất hiện trên phần lớn diện tích của tỉnh Bắc Ninh. Đây là loại đất phù sa màu mỡ sau thời gian dài canh tác mà không có biện pháp cải tạo, bảo vệ thích hợp khiến cho độ phì của đất bị giảm nghiêm trọng. Một điểm đáng chú ý là mô hình dự báo loại đất phù sa ít chua, cơ giới trung bình (Đất phù sa 3) ở hai bên bờ sông Đuống. Điều này cho thấy khả năng dự báo khá tốt của mô hình vì khu vực này hàng năm đều được bồi đắp một lượng phù sa đáng kể nên đất có độ phì cao và chất lượng tốt. Một lần nữa, mô hình MLR dự báo loại đất xám ở những vị trí cao và không có sự xuất hiện của loại đất thuộc nhóm đất Glây giống như ở mức Nhóm đất chính.

3.2. Đánh giá chất lượng bản đồ

Bảng 2 cho thấy kết quả tính toán độ chính xác bản đồ, chỉ số đa dạng của thông tin trên bản đồ và chỉ số kết hợp.

Ở mức nhóm đất chính, mô hình dự báo cho kết quả độ chính xác của bản đồ khá cao 0,73 có

nghĩa là khoảng 73% của bộ điểm mẫu độc lập đã được phân loại đúng. Trong khi đó, ở mức loại đất trung gian, độ chính xác của bản đồ chỉ còn 0,39 tương đương 39% bộ điểm mẫu độc lập được phân loại đúng. Điều này được lý giải bởi số lượng nhiều hơn của các loại đất trung gian (15 loại) so với nhóm đất chính (5 nhóm) khiến cho các yếu tố ảnh hưởng đến sự hình thành đất sẽ nhiều hơn và có mối liên hệ phức tạp hơn, dẫn đến kết quả dự báo của mô hình sẽ kém chính xác hơn.

Chỉ số đa dạng thông tin của 2 bản đồ kết quả lại cho thấy xu hướng ngược lại với độ chính xác của bản đồ: ở mức nhóm đất chính H khá thấp (0,2) nhưng ở mức loại đất trung gian H tăng lên 1,21. Kết quả này cho thấy mô hình MLR dự báo ở mức phân loại chi tiết sẽ cho mức độ đa dạng thông tin cao hơn nhiều vì ở mức này số lượng các loại đất nhiều hơn so với mức Nhóm đất chính.

Trong thực tế, chúng ta cần các bản đồ có độ chính xác cao đồng thời phải thể hiện được mức độ đa dạng của các đối tượng trên bản đồ. Vì vậy, chỉ số kết hợp là một yếu tố quan trọng để đánh giá toàn diện chất lượng của bản đồ kết quả. Bảng 2 cho thấy mô hình MLR dự báo loại đất trung gian cho chỉ số kết hợp cao hơn nhiều so với dự báo nhóm đất chính.

4. KẾT LUẬN VÀ ĐỀ NGHỊ

- Mô hình hồi quy logistic đa thức cho kết quả dự báo khá chính xác ở mức phân loại nhóm đất chính nhưng chất lượng thông tin thấp. Mô hình này có thể là phương pháp thích hợp để thay thế cho phương pháp thành lập bản đồ đất truyền thống ở Việt Nam hiện nay.

- Để tránh ảnh hưởng của một loại đất chiếm phần lớn trong các điểm mẫu đầu vào đến kết quả dự báo của mô hình, cần lựa chọn tỷ lệ mẫu tương đối hợp lý giữa các nhóm đất.

- Cần tiếp tục nghiên cứu đánh giá khả năng ứng dụng mô hình MLR ở những dạng địa hình khác nhau và lựa chọn các biến phù hợp với từng dạng địa hình để có những đánh giá toàn diện hơn.

TÀI LIỆU THAM KHẢO

- A.P.D.Turetta, M. L. Mendoca Santos, L.H.C.Anjos and R.L.L.Berbara (2006). Chapter 22: Spatial-Temporal Changes in Land Cover
- Carré, F. and M. C. Girard (2002). "Quantitative mapping of soil types based on regression kriging of taxonomics distances with landform and land cover attributes". *Geoderma*, 111: 241-263.
- Fagerland M. W., D. W. Hosmer and A. M. Bonfi. 2008. 'Multinomial goodness-of-fit tests for logistic regression models.' *Statistics in Medicine*. 27(21).
- FAO (2001). Lecture notes on the Major Soils of the World.
- Finke, P. (2011). Syllabus for the course Soil prospection and classification in the Physical Land Resources program.
- Kempen, B., D. J. Brus, Heuvelink GBM, Stoorvogel JJ (2009). "Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach". *Geoderma*, 151: 311-326
- Lagacherie, P. and A. B. McBratney (2007). Chapter 1. Spatial soil information system and spatial soil inference systems: perspective for Digital Soil Mapping. *Digital Soil Mapping: An Introductory Perspective*. P. Lagacherie, A. B. McBratney and M. Voltz. Amsterdam, Elsevier. Development in Soil Science, pp. 3-24.
- Martin, M. A. and J. M. Rey (2000). "On the role of Shannon's entropy as a measure of heterogeneity". *Geoderma*, 98: 1-3.
- Olaya, V. F. (Ed.) (2004). A gentle introduction to Saga GIS Gottingen, Germany.
- WRB, I. W. G. (2006). World Reference Base for soil resources 2006. Rome, FAO.
- Yang, L., Y. Jiao, S. Fahmy, A-Xing Zhu, S. Hann, J. E. Burt, and Feng Qi. (2011). "Updating Conventional Soil Maps through Digital Soil Mapping". *Soil Science Society of America*, 75(3): 1044-1053.