8-2019

Determining Potential Invasion of Kudzu Bug (Megacopta cribraria) in North and South America and Estimating Its Invasion Dynamics with Geopolitical-Unit Level Records

Wanwan Liang
University of Tennessee, wliang@vols.utk.edu

Recommended Citation

Liang, Wanwan, "Determining Potential Invasion of Kudzu Bug (Megacopta cribraria) in North and South America and Estimating Its Invasion Dynamics with Geopolitical-Unit Level Records. " PhD diss., University of Tennessee, 2019.
https://trace.tennessee.edu/utk_graddiss/5627

Determining Potential Invasion of Kudzu Bug (*Megacopta cribraria*) in North and South America and Estimating Its Invasion Dynamics with Geopolitical-Unit Level Records

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Wanwan Liang

August 2019

DEDICATION

This dissertation is dedicated to my family for their support and to Dr. Jerome Grant for his great kindness, support, and mentoring. I do whatever it takes to achieve my professional goals.

ACKNOWLEDGEMENTS

First of all, I would like to thank my major advisor, the best advisor in the world, Dr. Jerome F. Grant. It was my great honor to be your student. Over the last four years, I’ve learned a lot from you on how to succeed in work but also to be supportive to other people. Thank you for being supportive for me to follow my research passion and help me find out answers when I have concerns. Without your supervision and great kindness, I wouldn’t have made it this far academically. I appreciated you over the last four years and will always be grateful to you in the future. I also appreciate the Department of Entomology and Plant Pathology for giving me this chance of being a Ph.D. student in this great department, where people work innovatively and energetically to pursue Real Life Solutions.

I also would like to thank all my committee members: Dr. Monica Papeş, Dr. Liem Tran, Dr. Scott Stewart, and Dr. Gregory Wiggins. Dr. Papeş, thank you so much for introducing a bigger world of species’ distribution modeling to me, and having very detailed and critical suggestions on my research. I also appreciate you a lot for letting me work in the Spatial Analysis Lab. Dr. Tran, thank you so much for introducing me to the spatial modeling world and always guiding me on my research. Your supervision and support had always contributed substantially to my progress on my research. Dr. Stewart, I really appreciate you for always being very supportive of me, providing detailed comments, and guiding me from an integrated pest management perspective. Dr. Wiggins, I appreciate you for helping me through these years and giving me critical and constructive comments on my research. I couldn’t remember how many times I asked

you for your suggestion on everything, and you always tried to help me even when I just dropped by your office. I have most sincere gratitude for you for everything.

Except my current committee members, I also thank my former committee members, Dr. Robert Washington-Allen and Dr. James Vogt. Dr. Robert Washington-Allen, I appreciate you for introducing the remote sensing world to me. I remember how passionate you were every time you talked about remote sensing applications. Because of you, I also developed a research passion for remote sensing for its applications on natural resources. I do hope I can find the opportunity to collaborate with you in the future. Dr. Vogt, I remember the way you encouraged and helped me when I first started the project and I appreciate you for having very detailed comments on my proposal, which always reminded me of the right way to write.

Additionally, I'd like to thank Dr. Mongi Abidi for sharing good papers on image processing with remote sensing data and teaching me image processing techniques. I also like to thank Dr. Luis Carrasco Tornero and Mr. Eric Carr for sharing Lidar knowledge with me.

I'd like to thank all my lab members and friends, and they are the people who made life happy and enjoyable during my 4-year program. I hope all my lab members and friends succeed at their career, and most importantly live happy lives. Last but not the least, I would like to thank my parents, my sister, and my brother for their unconditional love and support to me and always being there for me during some difficult times. I also like to thank my little sweet boy Pudding, a goldendoodle, who always makes me laugh and doesn’t get mad at me, for most of times, when I am too lazy or too busy to accompany him.

ABSTRACT

Biological invasions have put growing threats on ecosystems and human society, exacerbate effects of climate change, and cause economic loss globally. Modeling invasion risk and processes of pest species are essential for early prediction and warning and are increasingly used for detection and control of invasion outbreaks. Kudzu bug, *Megacopta cribraria* (F.), native to Asia, has become a pest in both agricultural and urban areas since its initial discovery in the United States (U.S.) in 2009. As the establishment of kudzu bug is relatively new in the U.S., its potential invasion risk in the Americas, spread dynamics, and factors that may impact its spread, are not well understood. Thus, this research was intended to address these gaps. This project also evaluated methods that can better estimate invasion risk and dynamics. Findings of this research can guide management of kudzu bug and also provide guidance on approaches to estimate invasion risk and dynamics.

High invasion risk of kudzu bug was predicted by species distribution models (SDMs) in eastern U.S., Central America, and central South America. Kudzu bugs generally inhabit warm (annual mean temperature around 15℃) and humid (annual mean precipitation around 1300mm) regions. Due to non-adaptive niche shift, kudzu bugs occupied different environmental conditions between the native and invaded ranges. Using kudzu bug as a case study and seven SDMs, non-adaptive niche shift does not necessarily challenge transferability of SDMs. Additionally, the spatial range where PAs are extracted can significantly impact both interpolation and transferability of SDMs.

Geopolitical-unit invasion record is capable of estimating invasion dynamics. Boundary displacement methods have the best estimations for both overall rate and spread dynamics. However, for spread without a clear infestation outline, area-based regression methods can be good alternatives for estimating invasion dynamics. The spread rate of kudzu bug was 76 km/year during 2010-2016, however, the rate varied largely among different regions (45 - 144 km/year). The availability of host plants seemed to play an important role on the spread of kudzu bug in the southern region, while the low temperature in January was the most important factor in the northern region.

PREFACE

“The measure of who we are is what we do with what we have.”

-- Vince Lombardi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION AND LITERATURE REVIEW

1.1 Biological Invasions

As a major component of global change, biological invasions continue to threaten ecosystems and human societies, exacerbate effects of climate change, and cause economic loss globally (Paini et al., 2016; Pejchar & Mooney, 2009; Walsh et al., 2016). Invasive species cause public health issues, threaten endangered species, cause biodiversity loss in their invaded ranges, and cause loss to agriculture, forest, society, and other segments of economy (Pimentel et al., 2005). For example, the economic loss caused by invasive species annually has been estimated to be at least $120 billion in the United States (U.S.) (Pimentel et al., 2005) and £1.7 billion in Great Britain (Williams et al., 2010). Many endangered species are threatened by invasive species worldwide (Gurevitch & Padilla, 2004; Pimentel et al., 2005). Gurevitch and Padilla (2004) summarized that among the species listed in the International Union for Conservation of Natural Resources Red List, 882 terrestrial species, 59 freshwater species, and 87 marine species are endangered by invasive species. Management of invasive species, therefore, becomes essential to minimize their negative impacts. Management, including preventing introductions and early detection, becomes important to minimize the impacts of invasive species.

This research focused on kudzu bug, *Megacopta cribraria* (F.) (Hemiptera: Heteroptera: Plataspidae), which is a relatively new invasive species in the U.S. However, it causes economic losses in soybean, and is also an urban pest. Research on kudzu bugs in this project can facilitate early detection, monitoring, and integrated pest management.

1.2 Establishment and Assessment of Kudzu Bug in the United States

In 2009, a new invasive insect, kudzu bug, was first reported in the U.S. in Georgia (Suiter et al., 2010). Kudzu bug is named after its primary plant host, the Asian kudzu vine (*Pueraria montana* (Lour.) Merr.), which is in the pea family (Fabaceae). As a species introduced into the U.S. in the 1870s, kudzu has a wide distribution in the U.S., especially in the southeastern region (**Figure 1.1**). However, kudzu bug was reported to occur on 33 plant species and to feed on at least 9 plant species in the U.S. (Gardner et al., 2013). Kudzu bug was first described by Fabricius (1798) and is native to Asia (Eger et al., 2010). Kudzu bug is the only known species in the family Plataspidae to occur in the Americas (Ruberson et al., 2013). Since its invasion, kudzu bug has spread quickly throughout the southeastern U.S. By the end of 2018, kudzu bug had been confirmed in 13 states (EDDMapS 2019, **Figure 1.2**).

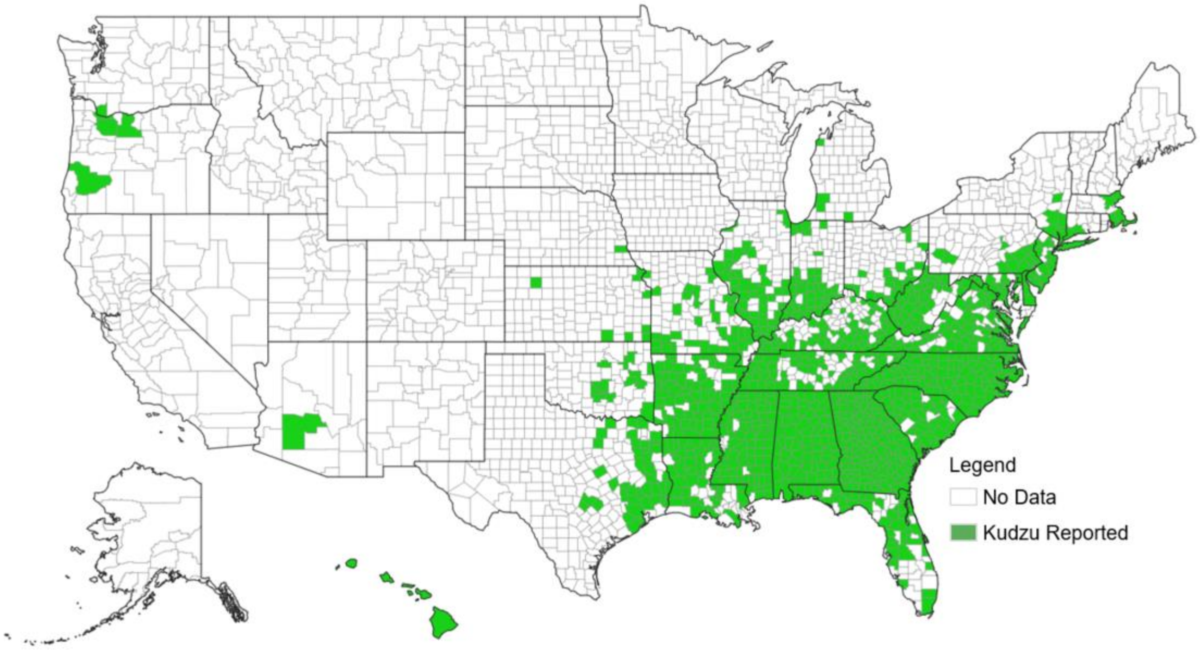

Figure 1.1 Known distribution of kudzu in the U.S. (EDDMapS, 2019).

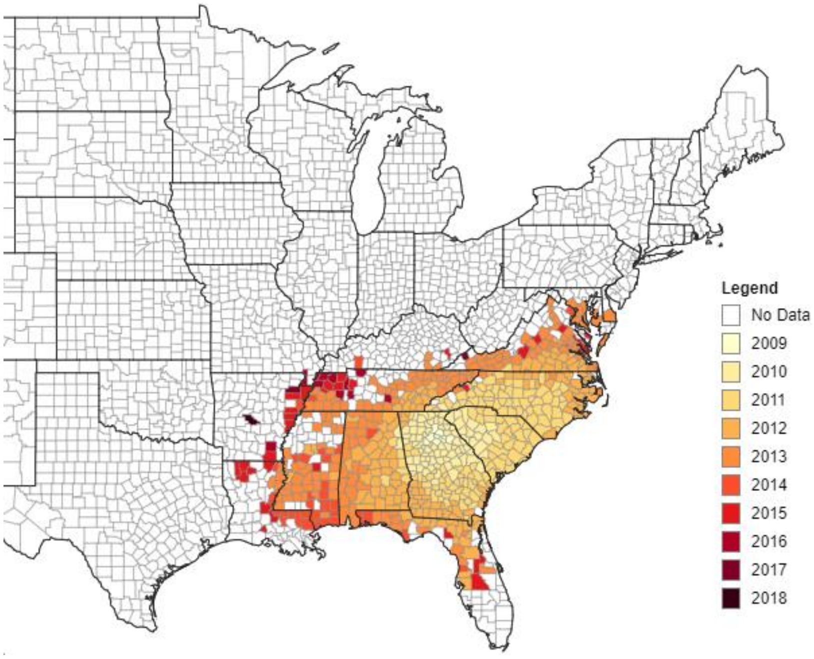

Figure 1.2 Known distribution of kudzu bug in the U.S. (County colors represent first year of detection in that County).

1.2.1 Invasion Risk Modeling

Distribution Modeling of Invasive Species

Modeling and predicting invasion processes of pest species are essential for early prediction and warning (Sutherst, 2003) and are increasingly used for detection and control of invasion outbreaks (Václavík & Meentemeyer, 2009). Invasive pest modeling is critical to manage, mitigate or prevent the infestation of invasive pest species (Gallien et al., 2010). The ecological niche-based methods have widely been used to model the potential geographic distribution of invasive species. Hutchinson (1957) described the ecological niche as a multi-dimensional environmental space where both abiotic and biotic factors permit positive growth of species. The ecological niche can be furthered described as the fundamental niche (FN) and the realized niche (RN). FN delineates the abiotic environmental conditions that allow an organism to survive, while RN describes the set of conditions actually occupied by a species after interacting with other species (Hutchinson, 1957). Compared to RN, FN only considers abiotic factors and is easily to measure, thus FN or a subset of FN has been increasingly used to model the potential geographic extent of targeted species (Guisan & Zimmermann, 2000; Soberón & Peterson, 2005). Niche-based distribution models construct relations between the occurrences of species with various environmental features (Randin et al., 2006).

As kudzu bug is a relatively new invasive insect, adequate research has not been available to predict the invasion risk of kudzu bug in North and South Americas. Although the climatic preference of kudzu bug is also not fully understood (Gardner et al., 2013;

Zhang et al., 2012), the knowledge could contribute to its early warning and detection in new regions, as well as provide information on pest ecology.

Transferability of Distribution Models for Invasive Species

Estimating potential biological invasion is an important tool to manage non-native species. To predict the potential invasion pattern of a given species in a new range, native occurrence information is generally used in species distribution modeling (SDM). Thus, successfully predicting invasion patterns closely relies on the transferability of models from native ranges to invaded ranges. However, transferability of SDMs can be challenged by niche shift of an invasive species in its new spatial range (Broennimann et al., 2007; Early & Sax, 2014; Parravicini et al., 2015). Soberón and Peterson (2011) suggested that, in some cases, the “niche shift” is more likely a result of differences in environmental conditions, or environmental space anisotropy, between two spatial ranges rather than a true adaptive niche shift of species. The difficulty in distinguishing adaptive from non-adaptive niche shift continues to cause concerns when considering the transferability of SDMs.

When estimating potential invasion patterns with presence-only data, the spatial range from which pseudo-absences (PAs, also called background data) are extracted certainly impacts the transferability of SDMs (Phillips, 2008; Barbet-Massin et al., 2012). However, a quantitative assessment of the impact of PAs on both interpolations and transferability of commonly used models is lacking. This information is important to consider, especially when niche shift is observed between the two ranges.

1.2.2 Estimating Spread Dynamics of Invasive Species

Modeling invasion dynamics is also important to the management of invasive species, as it facilitates prediction of spatial and temporal invasion risks of the species, enhances early detection, guides the construction of early warning systems, and determines important factors affecting the invasions (e.g., Liang et al., 2019; Paini et al., 2016; Stohlgren & Schnase, 2006). In practice, estimating spread rates of invasive species has been conducted on various species on all spatial scales. For research at large scales, such as regional, continental, or even global scales, researchers usually have to collect all available records from online databases, published research, surveys, or field sampling (e.g., Masciocchi & Corley, 2013; Pyšek et al., 2008; Suarez et al., 2001). Consequently, data for large-scale research usually has coarse and non-unified resolution, whereas geopolitical-unit level data are usually the most abundant and available records of invasive species at such scale (Evans & Gregoire, 2007; Liebhold et al., 1992; Tobin et al., 2007; Tobin et al., 2015).

Researchers worldwide have used geopolitical-unit records to estimate invasion rates of various species (e.g., Evans & Gregoire, 2007; Horvitz et al., 2017; Perrins et al., 1993). Multiple methods, including regression, boundary displacement, and minimum spread distance methods, have been commonly used. Several researchers compared the accuracy of these common methods to estimate spread of invasive species (e.g., Gilbert & Liebhold, 2010; Tobin et al., 2015). However, existing research only compared the overall estimated rate (Gilbert & Liebhold, 2010; Tobin et al., 2015). Spread of invasive species, especially at large scales, is commonly complex due to spatial heterogeneity and stochastic events, such as long distance jump dispersal (LDJD) (Hastings et al., 2005; Pyšek et al., 2008). Estimating spreads with geopolitical-unit data, however, further increases the uncertainties, as there can be large variations in the sizes of geopolitical units (Hastings et al., 2005; Pyšek et al., 2008). Consequently, spread rates possibly vary in different spatial regions and temporal periods. Thus, compared to one single overall spread rate, estimating spread dynamics is more informative to understanding invasions (Hastings et al., 2005).

Currently, research is lacking on a comprehensive evaluation of common methods to estimate both overall rate and spread dynamics with geopolitical-unit invasion record. Additionally, as the establishment of kudzu bug is relatively new in the U.S., the invasion dynamics and the factors that impact its spread are not well understood. Consequently, one of the objectives is to address these research gaps.

1.3 Aims and Objectives

Aims of this project include: 1) modeling the potential invasion risk of kudzu bug with niche-based SDMs, 2) assessing the transferability of SDMs with non-adaptive niche change and impact of PAs using kudzu bug as a case study, 3) evaluating performances of commonly used methods to estimate invasion dynamics with geopolitical-unit level record, and 4) estimating spread dynamics of kudzu bug with geopolitical-unit level data.

Three objectives are associated with the first research aim: 1) examining if the invaded populations of kudzu bug in the U.S. occupy the same environmental space as the native populations in Asia, 2) predicting the potential invasion of kudzu bug in North and South America with niche-based models, and 3) analyzing the climatic preference of kudzu bug.

For the second aim, my research objectives include: 1) evaluating the accuracy and transferability of commonly used SDMs with kudzu bug under non-adaptive niche shift between native and invaded ranges, and 2) analyze how model performance is affected by use of PA samples from outside the native range of the species.

To address the third aim, which was to evaluate performances of commonly used methods to estimate invasions, I conducted analysis and comparison incorporating the following aspects: 1) the accuracy of commonly used methods to estimate spread dynamics with irregularity and stochasticity in spread, 2) the impact of the size of geopolitical unit on each method, and 3) the similarities of all methods to estimate spread rate and dynamics.

Regarding the fourth aim, which was to estimate the spread dynamics of kudzu bug in the U.S., I first estimated the spread rate of kudzu bug during years 2010-2016. I then determined important spatial factors that influenced the spread of kudzu bug to provide information for early detection and management of kudzu bug.

CHAPTER II: PREDICTING THE POTENTIAL INVASION OF KUDZU BUG, *MEGACOPTA CRIBRARIA* (HETEROPTERA: PLATASPIDAE), IN NORTH AND SOUTH AMERICA AND DETERMINING ITS CLIMATIC PREFERENCE

This chapter is a reformatted version of a paper by the same name published in *Biological Invasions* by Liang, W., Tran, L., Washington-Allen, R., Wiggins, G., Stewart, S., Vogt, J., & Grant, J.

Liang, W., Tran, L., Washington-Allen, R., Wiggins, G., Stewart, S., Vogt, J., & Grant, J. (2018). Predicting the potential invasion of kudzu bug, *Megacopta cribraria* (Heteroptera: Plataspidae), in North and South America and determining its climatic preference. *Biological Invasions*, *20*, 2899-2913.

2.1 Abstract

Modeling potential invasion of an introduced organism is a critical tool for early management of damaging species, such as kudzu bug, *Megacopta cribraria* (F.) (Hemiptera: Heteroptera: Plataspidae). To predict the potential invasion of kudzu bug in North and South America, the species distribution models Genetic Algorithm for Rule-set Production (GARP) and Maximum Entropy (Maxent) were used. The D metric was used to test the niche equivalency and similarity between native and invaded populations of kudzu bug. Results suggested that kudzu bugs currently occupied unequal environmental spaces between the two ranges. Therefore, distribution models using GARP and Maxent were constructed using occurrences in both native and invaded ranges. Area under the curve (AUC), true skill statistics (TSS), and omission rate (OR) were used to evaluate and compare the models. Results indicated both models had good performance, but Maxent (AUC=0.971, TSS=0.946, OR=0.019) performed better than GARP (AUC=0.922, TSS = 0.860, OR=0.037). This research confirmed the effectiveness of using occurrence data in both ranges to predict potential invasions. Kudzu bugs prefer warm (annual mean

temperature around 15℃) and humid (annual mean precipitation around 1300mm) regions. Distribution models generated by both methods indicated similar regions with high invasion risk. Management programs that include quarantine and prevention measures are suggested for these regions to avoid outbreaks of kudzu bug.

2.2 Introduction

Species distribution models (SDMs) or ecological niche models have been commonly used to predict the potential distribution of species for various purposes, including biological conservation, invasion prediction, paleobiology, spatial epidemiology, and impacts of climate change on biodiversity (Franklin, 2013; Guisan & Thuiller, 2005; Mainali et al., 2015; Svenning et al., 2011). SDMs identify relationships between observed occurrences and environmental variables by using statistical models or theoretically derived response curves (Guisan & Thuiller, 2005; Elith & Franklin, 2013). The ecological niche, more specifically the fundamental niche (FN), or a subset of it, has been increasingly used to estimate species' geographic extent (Guisan & Zimmermann, 2000; Peterson & Soberón, 2012). The ecological niche is described as a multi-dimensional space, including both abiotic and biotic factors, that permits positive growth of a given species, whereas the FN delineates only the abiotic environmental conditions (Hutchinson, 1957; Pearson & Dawson, 2003).

To model the potential invasion of a species in the invaded range, the classical approach uses the distribution data from the native range and then projects the model into the invaded range (Peterson, 2003). One underlying assumption of this approach is that the species will conserve its niche when it invades a new geospatial range (Wiens & Graham, 2005). Niche conservatism refers to the tendency of species to retain their ancestral ecological traits (Wiens & Graham, 2005). However, this assumption can sometimes be incorrect (Broennimann et al., 2007), and makes the classical approach insufficient to predict potential biological invasion. For some species, niche shift may occur as with the invasions resulting in the change of habitat preference in the new range (Broennimann et al., 2007; Randin et al., 2006). This change is called evolutionary niche shift and had been evidenced in published research (Broennimann et al., 2007; Guisan et al., 2014; Randin et al., 2006). However, Soberón and Peterson (2011) suggested that in some cases the "niche shift" is more likely a result of environmental heterogeneity between two geospatial ranges, namely non-evolutionary niche shift, than a real niche shift of the given species. Here the term environmental space shift is used instead of niche shift to represent both types of niche shift. Once the environmental space shift exists, the distribution models trained with native occurrence data may have limited ability to predict the potential invasion in the new range as the native environmental conditions may fail to express the true habitat requirement in the new range. Broennimann and Guisan (2008) and Jiménez-Valverde et al. (2011) suggested that the potential distribution model for invasive species could be developed using occurrence data from both the native and the invaded ranges when the occupied environmental space differs between the two ranges. Thus, testing the conservatism of environmental niche is crucial for determining whether occurrence data in native range or in both native and invaded ranges should be used to predict the potential invasion.

It is not known if kudzu bug conserves its environmental space in the U.S., leading to lack of adequate research to predict the invasion risk of kudzu bug in North and South Americas. Whereas the climatic preference of kudzu bug is also not fully understood (Gardner et al., 2013; Zhang et al., 2012), knowledge on the climatic preference of this pest species could contribute to its early warning and detection as well as provide information on pest ecology. To address these gaps, I examined if the invaded populations of kudzu bug in the U.S. occupied the same environmental space as the native populations in Asia. The stability, expansion, and unfilling of occupied environmental space in the U.S. were calculated to better understand the niche dynamics of kudzu bug in the new range. Niche stability and expansion (1- niche stability) measure the proportion of occupied environmental space in the invaded range that is overlapping and non-overlapping, respectively, with that in the native range. The niche unfilling is the proportion of occupied environmental space in the native range, which is available but has not been occupied in the invaded range. Two commonly used niche-based methods to model the potential invasion of kudzu bug in North and South America. Finally, the variable importance and climatic preference of kudzu bug were analyzed for its value of providing information for management of kudzu bug, as well as its value to increase knowledge on kudzu bug ecology.

2.3 Methodology

2.3.1 Species Point Occurrence Data

Occurrence data of kudzu bug were collected in both the native range in Asia and the invaded range in the U.S. In total, 319 points of occurrence data were included. All the occurrence data are available on request. Readers are referred to Zhu et al. (2012) for detailed information on data acquisition in the native range. After deleting redundant records, 164 occurrences were retained in Asia. Occurrence data of kudzu bug in the invaded range in the U.S. were acquired from the Early Detection and Distribution Mapping System (EDDMapS, 2019). The EDDMapS dataset contained the date and location of confirmed occurrences of kudzu bug since 2009 (Gardner et al., 2013).

Excluding repeated records, 155 occurrences were retained in the invaded range. These occurrence data were distributed throughout the southeastern U.S.

2.3.2 Environmental Variables

The ‘bioclimatic’ variables from WorldClim (Hijmans et al., 2005) and elevation from the Hydro-1K digital elevation model dataset were used as the environmental features (USGS, 1996). The spatial range of variables includes Asia, North and South America. Among all the 19 bioclimatic variables, 12 bioclimatic variables were selected, 6 temperature-related variables and 6 precipitation-related variables (**Table 2.1**).

Table 2.1 Full name and brief name of 12 bioclimatic variables selected for use in the models; T = temperature, P= precipitation

Abbreviate Name	Full Name	Abbreviate Name	Full Name	Abbreviate Name	Full Name
BIO1	annual mean T	BIO8	mean T of the wettest quarter	BIO16	P of the wettest quarter
BIO2	mean diurnal T range	BIO11	Mean T of the coldest quarter	BIO17	P of the driest quarter
BIO5	maximum T of the warmest month	BIO12	annual mean P	BIO18	P of the warmest quarter
BIO7	mean annual T range	BIO15	P seasonality	BIO19	P of the coldest quarter

These variables were selected based on two principles: 1) only variables that are important to kudzu bug were selected based on its biology, as research suggested that temperature and humidity both impact kudzu bug (Shi et al., 2014; Wang et al., 1996), and 2) variables that have strong correlation ($r>0.8$) with the remaining variables were deleted. These 12 bioclimatic variables plus elevation provided 13 variables for the models, and all variables had the same resolution of 30 arc-second (approximately 1 km).

2.3.3 Test of Environmental Space Difference

Several methods had been used to test the niche conservatism (Guisan et al., 2014), and researchers suggested that quantifying niche changes by directly comparing the environmental attributes while considering the available environmental space between two spatial ranges is a most appropriate method (Broennimann et al., 2012; Guisan et al., 2014). Therefore the method proposed by Broennimann et al. (2012) was used to test the equivalence and similarity of environmental spaces occupied by native and invaded kudzu bug populations. This method takes the available environmental space in different ranges into account by applying kernel smooth function to correct occurrence densities with available environmental densities. The niche equivalency and similarity was then tested on the niche overlap along using the D metric (Broennimann et al., 2012; Warren et al., 2008), which was first proposed by Schoener (1968). Using the method proposed by Broennimann et al. (2012), the D-metric can be expressed as:

$$D = 1 - \frac{1}{2}\left(\sum_{ij}\left|Z_{1ij} - Z_{2ij}\right|\right),$$

where Z_{1ij} and Z_{2ij} represent the occupancy of entity 1 and entity 2 at given environmental conditions, respectively. The D value varies from 0 to 1, with 0 indicating no niche overlap and 1 indicating niche identical. To test the niche equivalency, all occurrences in both spatial ranges are first pooled together, and then randomly split into two datasets, this procedure can be repeated n times (n=1000 in this research). The number of occurrences in the two new datasets stays the same as the number in the original datasets (Warren et al., 2008). For each repetition, the simulated niche overlap D is calculated. With enough repetition, the derived D values can be used to construct a robust null distribution of D value, based on which the niche equivalency can be accepted if the observed D value falls in the density of 95% of the simulated D values. Although the niche similarity test is different from the niche equivalency test, niche similarity test also used the randomization procedures to construct null distribution of D values. Interested readers are referred to Broennimann et al. (2012) for detailed description on the niche similarity test.

To test the shift of environmental space of kudzu bug between native and invaded populations, only Asia and the U.S. were used as the study area. A principal component analysis (PCA) was conducted on the occurrences and the available environmental conditions to ensure the native and invaded ranges were represented equally. Five hundred pixels from both ranges (totally 1000 pixels) were extracted randomly to represent the available environmental conditions. The first two axes of PCA were then used to calculate the overlap of environmental spaces between the native and invaded populations. The available environmental spaces and the occupied environmental space by native and invaded populations were also visualized on the first two axes of PCA. The “ecospat” package in R (Di Cola et al., 2017) was used to test the equivalency and similarity of occupied environmental space by native and invaded populations. Niche expansion, stability, and unfilling were also calculated using “ecospat” package.

2.3.4 Model Development

To predict the invasion risk of kudzu bug, both North and South America were included as study area. Two commonly used niche-based methods were considered for this purpose: Genetic Algorithm for Rule-Set Production (GARP) (Stockwell, 1999) and Maximum Entropy Model (Maxent) (Phillips et al., 2006). GARP and Maxent models were chosen because prior research demonstrated their better performances when compared with many other models (Hernandez et al., 2006; Tsoar et al., 2007). Both models only require occurrence data and use pseudo-absence data during the construction of models by extracting random points from the geospatial extent of input environmental variables. Occurrence data from both native and invaded ranges were used to train the model. Occurrence data in Asia and the U.S. were first compiled (totally 319 occurrence), and then 70% of occurrences were randomly extracted (n = 223) as training data and the remaining 30% (n = 96) were used as extrinsic testing data. This random splitting was repeated 10 times to generate 10 sets of training and testing data. Thus, GARP and Maxent models were run and evaluated ten times and with each time an individual set of training and testing data were used. The averaged model of 10 runs was used as the final model for GARP and Maxent.

GARP

The GARP modeling system is a genetic algorithm that produces sets of rules that delineate ecological niches of species (Stockwell, 1999). GARP models were developed using the “best subsets” procedure in OpenModeller (De Souza Muñoz et al., 2011). In this procedure, the GARP model was run 100 times on an individual training dataset (out of 10 data splits; see previous section) and the training dataset was further split internally into 70% intrinsic training data and 30% intrinsic testing. This internal split of the data is necessary to select 10 best models (of the 100 obtained) and generate a GARP prediction. This procedure was conducted for the 10 random splits of training and testing datasets. The output GARP predictions were then evaluated by using the remaining extrinsic testing data.

Maxent

Maxent modeling system predicts species’ geographic distribution by fitting variable data with a probability distribution. This probability distribution is generated using training data and has the maximum entropy by subjecting to constraints representing the unknown information (Phillips et al., 2006). Maxent models were implemented using the Maxent program (Phillips et al., 2006) with the default parameters. All 10 models (based on the 10 random splits of occurrence datasets; see above) were then evaluated by using the remaining extrinsic testing data. Logistic output was used as the model output type. The Maxent logistic output is an estimate of suitability of one area to the target species (Phillips & Dudík, 2008). The logistic output of the Maxent model varies from 0 to 1, to better visualize the output model the final average model was multiplied by 100 to provide the same scale of values as GARP.

2.3.5 Model Evaluation Statistics

Three methods were used to evaluate the models: the area under the receiver operating characteristic curve (AUC) (Fielding & Bell, 1997), omission rate (OR), and true skill statistic (TSS) (Allouche et al., 2006). The overall AUC, OR, and TSS were calculated to evaluate and compare the performance of two models, and the three measures for the native and invaded ranges were also calculated. To compare the performance of the GARP and Maxent models, the one-tailed Wilcoxon signed-rank test was used to compare the measures of two models.

AUC is a generally accepted and threshold-independent measurement of model performance (Fielding & Bell, 1997). The value of AUC varies from 0 to 1, with AUC $\geq$ 0.9 considered as good performance (Swets, 1988). TSS takes both omission and commission errors into account by adding sensitivity and specificity and subtracting 1. TSS varies from -1 and +1, where 1 indicates perfect model performance and values of zero or less indicate poor performance. “ROCR” package (Sing et al., 2005) in R software was used to calculate AUC and TSS. The AUC and TSS were calculated for all the 10 models of each algorithm. OR is a threshold-dependent measurement. To convert the suitability model to binary presence/non-presence model, the threshold for GARP and Maxent model was carefully selected. As GARP usually predicts a large or even overly extensive area of potential distribution (Peterson et al., 2007), the threshold for GARP was set to 50% of the best subset models, which means only places predicted by at least half of the best models are considered as potential distribution area. To ensure that the OR of GARP and Maxent is comparable, the “equalized predicted area” method proposed by Phillips et al. (2006) was used to set the threshold of Maxent. The threshold was set for each Maxent model to give the Maxent binary model the same number of pixels predicted as potential distribution area as the GARP binary model.

2.3.6 Variable Importance and Climatic Preference

To determine the climatic preference of kudzu bug, the most important variables for the construction of both GARP and Maxent models were first selected. Then the co-occurrence of predicted invasion risk and the values of important variables were analyzed. To determine the most important variables, the jackknife test was used. The Maxent program (Phillips et al., 2006) internally conducted the jackknife test and evaluation on the variable contribution. The jackknife test on the GARP model was run externally by generating 13 models that used one variable in isolation and 13 models that excluded one variable in turn. Analysis of the response of the predicted invasion risk values to important variables was only conducted for the Maxent model.

2.4 Results

2.4.1 Environmental Space Shift

Based on the PCA on both the available and occupied environment in two ranges, the first two PCs explained 64.69% of the total variance (**Figure 2.1 A-B**). According to the first two PCs, the total overlap of occupied environmental space between the native and invaded populations is 0.196 (**Figure 2.1 A**). The statistical test suggested that the two populations did not occupy an equivalent environmental space, as the observed D value fell outside of the 95% confidence interval of simulated values (**Figure 2.1 C**). The test on the similarity of environmental space indicated that the occupied environmental space by native populations was similar with the invaded populations (*P*-value=0.069, **Figure 2.1 D**). However, the same test on invaded populations suggested a difference indicating the occupied environmental space by invaded populations was not similar with native populations (*P*-value=0.047, **Figure 2.1 E**). The niche expansion, stability, and unfilling were 0.044, 0.956, and 0.670, respectively, indicating the dissimilarities of occupied environmental spaces were mainly caused by niche unfilling of kudzu bug in the U.S. Thus, the low overlap of occupied environmental space (0.196) also mostly resulted from the niche unfilling of kudzu bug in the invaded range.

2.4.2 Model Evaluation and Comparison

Maxent had a continuous output with values varying from 0 to 100, and higher values represented a higher invasion risk. The output of GARP “best subset” procedure was categorical with values varying from 0 to 100. Ten runs of the Maxent model returned 10 AUCs varying from 0.96 to 0.97, and the AUCs of 10 GARP models varied from 0.91 to 0.93. GARP and Maxent both had mean AUCs higher than 0.9 indicating that both models had good performance. Maxent had significantly higher AUC than GARP based on the one-tailed Wilcoxon signed-rank test (*P*-value=0.001, **Table 2.2**). The TSS had the same pattern indicating that Maxent had better performance than GARP (*P*-value=0.001, **Table 2.2**). Because TSS showed the same pattern with AUC, only AUC and OR were used for model comparison hereafter.

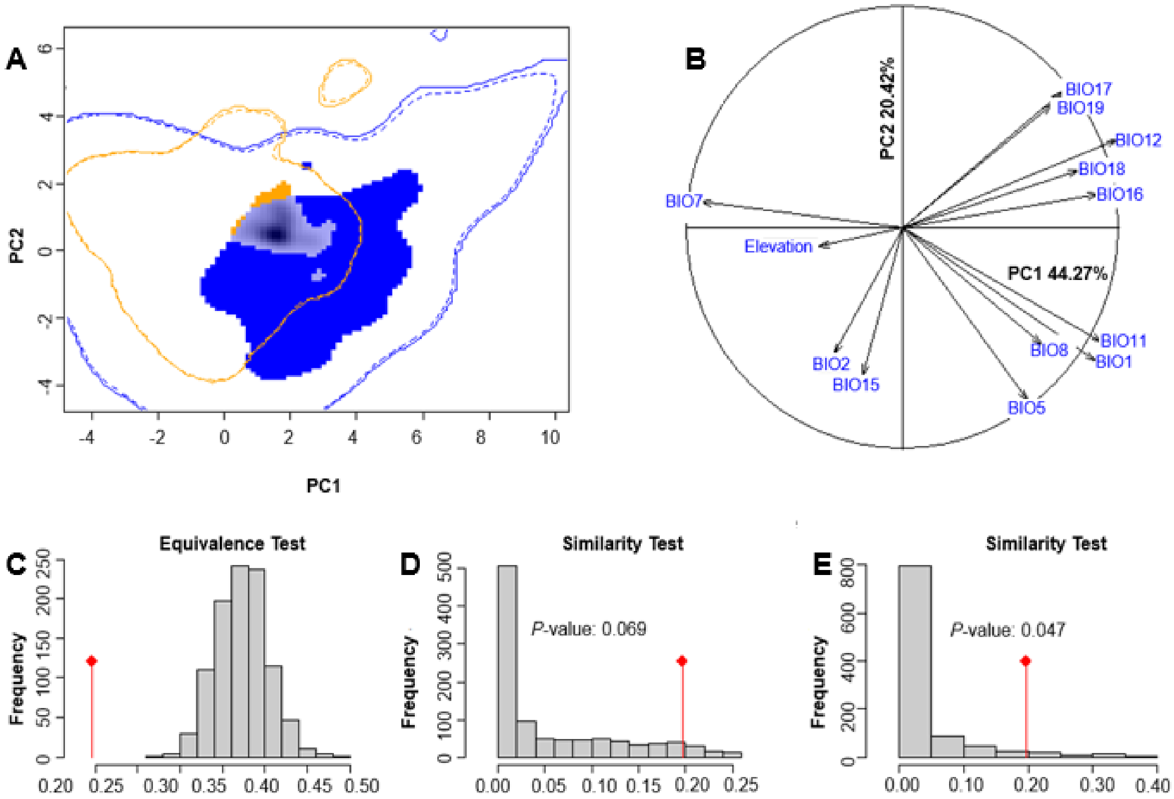

Figure 2.1 Analysis on environmental space shift of kudzu bug. **A** Environmental occupancy of kudzu bugs in native and invaded ranges, the dashed orange and blue contour lines represent the 95% of the available environmental space in the U.S. and Asia, respectively, while the solid ones represent 100% of the available environmental space. The orange and blue gradients indicate occupied environmental spaces in the U.S. and Asia, respectively, while the grey to black gradients indicate overlap of occupied environmental spaces (the darker gradients indicating higher density of overlap). **B** Correlation circle of the first two principal components (PCs), which totally explain 64.69% of the variable variance. **C-E** Observed environmental space overlap (bars with red diamond) and simulated niche overlaps (grey bars) on test of environmental space equivalency (**C**), environmental space similarity of Asia to U.S. (**D**), and environmental space similarity of U.S. to Asia (**E**) of two populations.

Table 2.2 Comparison of AUC and OR of GARP and Maxent Models

Model	Area Under the Curve (AUC)	True Skill Statistic (TSS)	Omission Rate (OR)
Maxent	0.971±0.004	0.946±0.017	0.019±0.011
GARP	0.922±0.007	0.860±0.020	0.037±0.010
P-value[a]	0.001	0.001	0.005

[a] *P*-value of the one-tailed Wilcoxon sign-ranked test on AUC and OR between GARP and Maxent model

Using 50% of best subset models as the threshold rule, the percent area predicted by 10 GARP models varied from 18.47% to 21.90%. The threshold value of 10 Maxent models varied from 0.93 to 1.83 when predicting the same area as GARP. The *P*-value for the Wilcoxon signed-rank test on OR was 0.005 (**Table 2.2**), indicating that the Maxent models had a significantly lower OR than the GARP models (*P*-value=0.005, **Table 2.2**). This outcome was consistent with the AUC result, suggesting that the Maxent models performed better than the GARP models. The final binary GARP and Maxent models predicted 20.96% of the total study area as potential distribution area for the kudzu bug.

2.4.3 Invasion Risk

The GARP and Maxent model pixels with higher values suggest regions with more suitable environment for the species and, thus, at a higher risk of invasion. GARP and Maxent models both predicted high suitability values for most of the observed occurrence points (**Figures 2.2, 2.3**). The suitability value of occurrence points from the Maxent model varied between 1.71 and 79.61 (mean = 57.165), while the suitability value from the GARP model varied between 28 and 100 (mean = 97.490). Despite the distinct difference of suitability values predicted by two models, both models had high agreement in terms of relative invasion risk (**Figure 2.3**). In the Americas, high invasion risk was predicted in the eastern U.S., most regions in the Central America, and the central region in South America (**Figure 2.3 A, B**). High agreement was also observed in binary models (**Figures 2.2, 2.3**): both models indicated similar regions that will be or already were infested by kudzu bug; however, differences in the two binary models were also observed. Part of Greenland was predicted as potential distribution area by Maxent model but not by GARP model (**Figure 2.3 C, D**). The situation was reversed in northern Brazil where the GARP model classified potential suitable area while the Maxent model classified the same type of area as unsuitable (**Figure 2.3 C, D**).

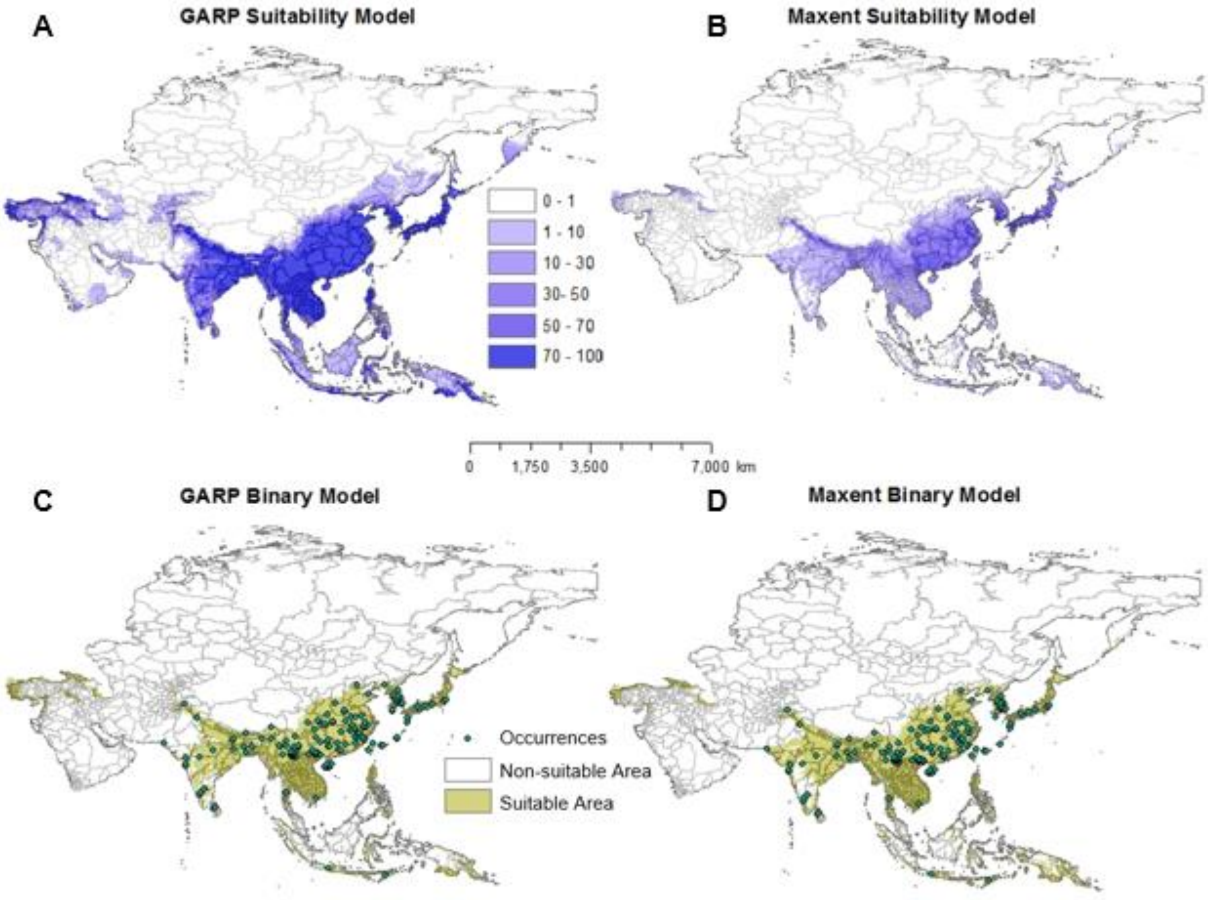

Figure 2.2 Predicted suitability maps of kudzu bug in native range**. A-B**, Predicted suitability maps of kudzu bug in native range in Asia. The higher the value is, the more suitable the area is. **C-D**, Binary predictions of the suitability maps in Asia. Occurrences, which were used to develop and evaluate models, were shown as green dots.

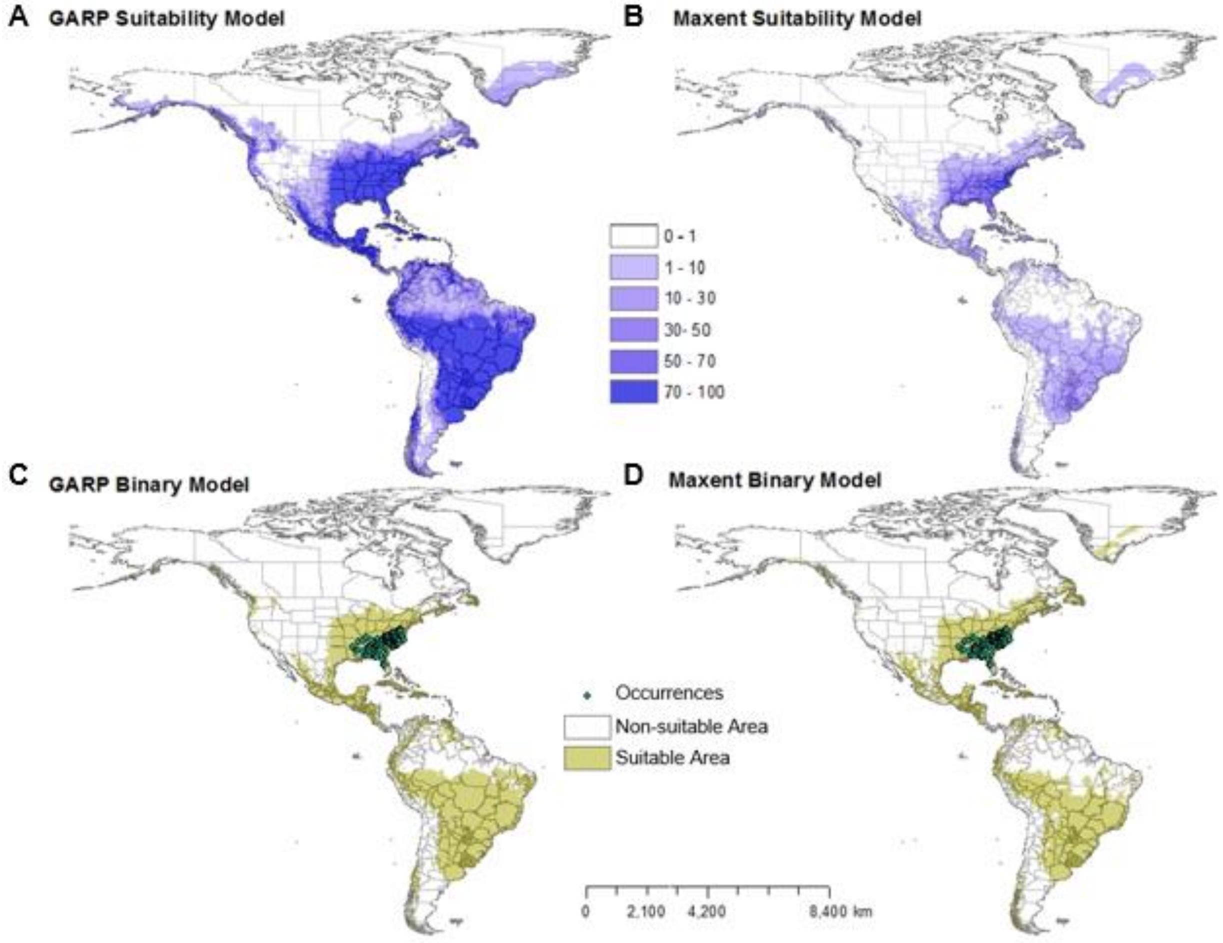

Figure 2.3 Predicted invasion risk model of kudzu bug in America**. A-B**, Predicted invasion risk model of kudzu bug in North and South America. The higher the value is, the more risky that the region would be infested by kudzu bug. **C-D**, Binary predictions of suitability maps in North and South America. Occurrence of kudzu bug in U.S., which were used to develop and evaluate models, were shown as green dots. Regions that were covered by green dots in **C-D** were all predicted as suitable places.

2.4.4 Variable Importance and Climatic Preference

GARP and Maxent yielded similar rankings of variable importance (**Figure 2.4**). Both models ranked annual mean precipitation (Annual P, BIO12), mean temperature of the coldest quarter (BIO11), precipitation of the warmest quarter (BIO18), annual mean temperature (BIO1) and mean annual temperature range (BIO7) as five of the six most important variables. The models also ranked elevation, mean diurnal temperature range (BIO2), and precipitation seasonality (BIO15) as the three least important variables. However, differences also were observed for specific rankings of these five variables. The GARP model ranked precipitation of the wettest quarter (BIO16) as the fourth important variable, while Maxent ranked it as the seventh most important variable. Analysis of variable contribution to the Maxent models generated similar results with the jackknife tests: BIO12, BIO11, BIO1, BIO8 and BIO7 are five of the most important variables, contributing 28.4%, 19.8%, 16.2%, 12.7% and 11.7% to the model, respectively. Cumulatively, the five most important variables contributed 88.8% to the Maxent models.

Response curves for the five most important variables in Maxent were generated to examine the climatic preference of kudzu bug. For annual mean precipitation (**Figure 2.5 A**), high suitability (>0.5) were limited between 1000 and 1700mm and the highest suitability was predicted at 1300mm. For mean temperature of the coldest quarter (**Figure 2.5 B**), annual mean temperature (**Figure 2.5 D**), and mean annual temperature range (**Figure 2.5 E**), the high probabilities of occurrence were predicted from 0 to 12℃, 12 to 20℃, and 28 to 35℃, respectively, and the highest probabilities occurred at 6℃, 15℃, and 32℃, respectively. The response curve of the suitability to precipitation of the warmest quarter is different from the former four variables (**Figure 2.5 C**). The high suitability occurred at the range of 330 to 375 mm and values greater than 440mm, and the highest suitability occurred at 750mm.

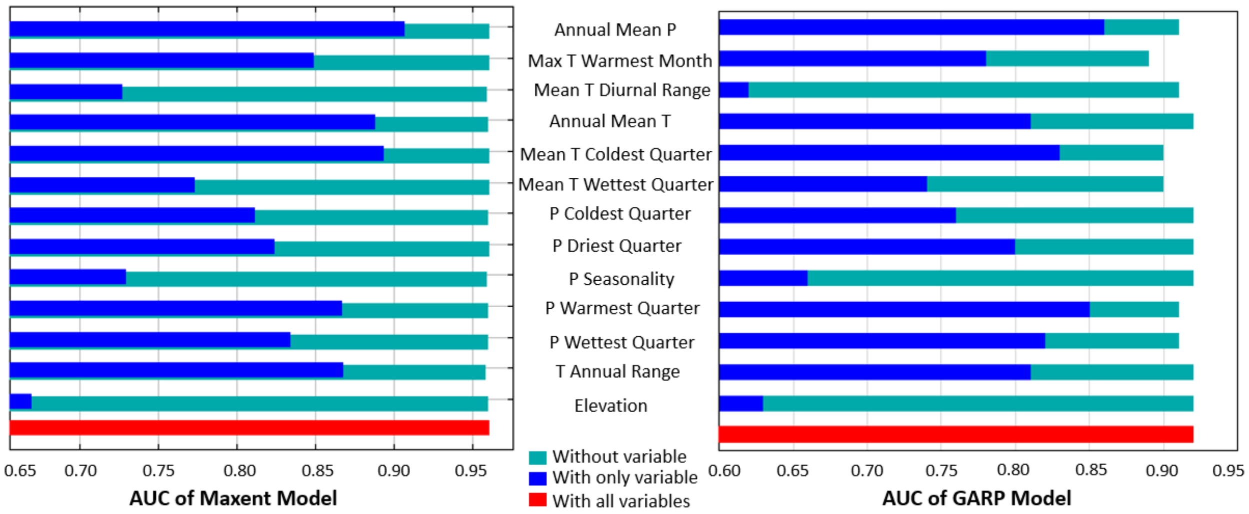

Figure 2.4 Jackknife test of AUC for the Maxent and GARP models (P, T and Max are abbreviations for Precipitation, Temperature, and Maximum, respectively).

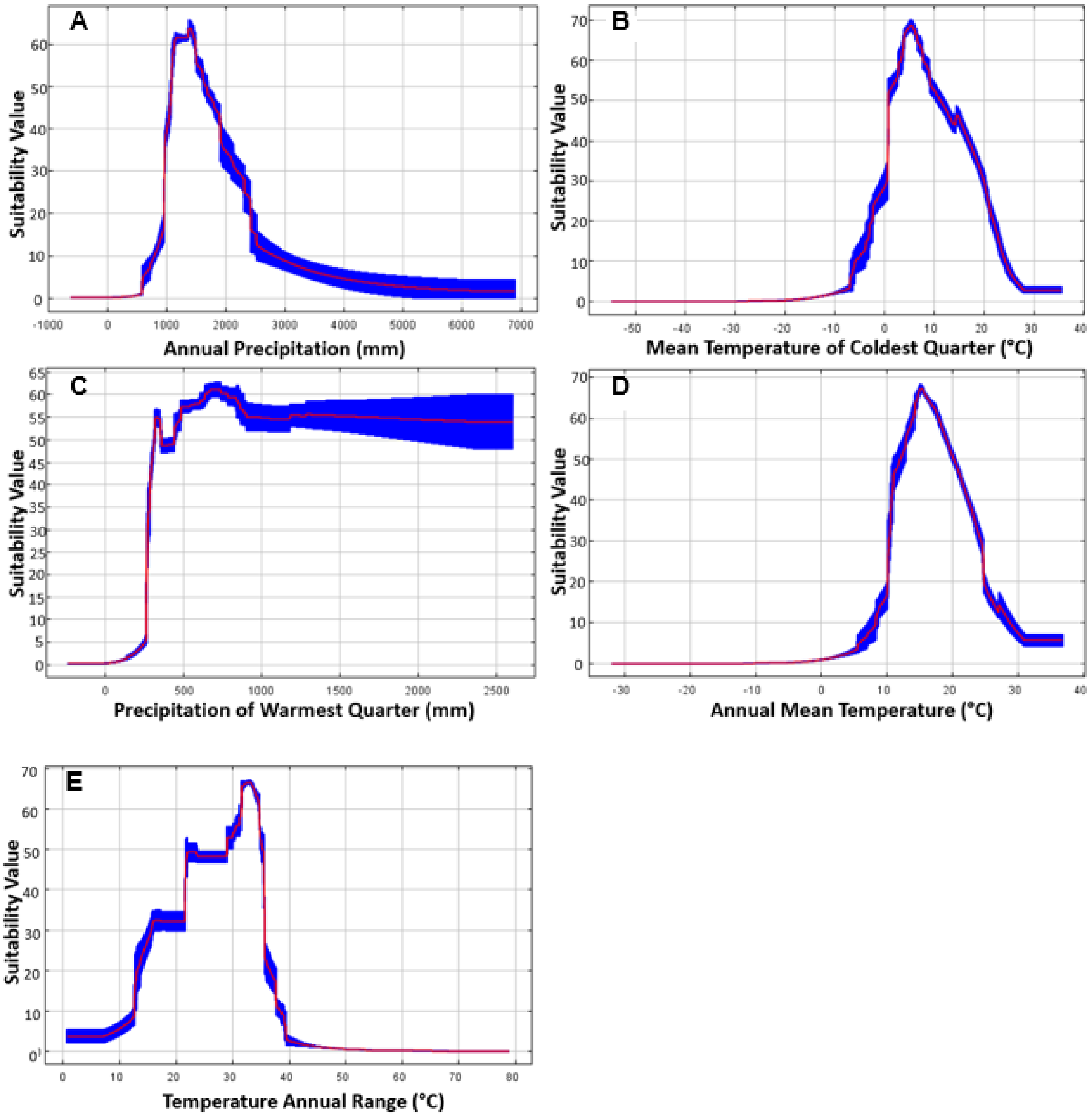

Figure 2.5 Response curve of the predicted suitability value by Maxent model to five most important variables (**A** – Annual Mean Precipitation, **B** – Mean Temperature of Coldest Quarter, **C** – Precipitation of Warmest Quarter, **D** – Annual Mean Temperature, and **E** – Temperature Annual Range) (red curves indicate mean response and the blue margins reflect ± 1 Standard Deviation over 10 runs)

2.5 Discussion

2.5.1 Environmental Space Shift

The niche equivalency test indicated that the native and invaded populations of kudzu bug do not inhabit in identical environmental conditions. However, this finding does not indicate a niche shift. Niche change had been proved rare among multiple groups of species (Petitpierre et al., 2012; Strubbe et al., 2013). Based on two independent studies with different groups of species, Petitpierre et al. (2012) and Strubbe et al. (2013) concluded that niche difference is more likely caused by niche unfilling than real niche change due to niche expansion. In this research, a small proportion of niche expansion (4.44%) and high niche stability (95.56%) were observed. However, the niche unfilling in the U.S. is as high as 67%, indicating that kudzu bugs have not been reported in 67% of the environmental space, which is occupied by native populations and available in the U.S.

Table 2.3 Comparison of AUC and OR of GARP and Maxent models obtained for native and invaded training ranges

Model	Area Under the Curve (AUC)			Omission Rate (OR)		
	Asia	U.S.	P-value[a]	Asia	U.S.	P-value[b]
Maxent	0.953+0.006	0.990+0.002	0.001	0.038+0.022	0	0.0002
GARP	0.936+0.006	0.912+0.009	0.001	0.072+0.019	0	<0.0001
P-value[c]	0.001	0.001	NA	0.009	NA	NA

[a]P-value of difference of AUC between native range and invaded range of GARP or Maxent model; [b]P-value of difference of OR between native range and invaded range of GARP or Maxent model; [c]P-value of difference between GARP and Maxent model in the same range of AUC or OR.

This niche unfilling suggests that the invasion of kudzu bug is far from equilibrium and more places could be further infested in the U.S. Based on the available environmental space in native and invaded ranges, environmental heterogeneity also exists. The study area in Asia has more available environmental conditions than the U.S. Although the niche expansion in the U.S. is only 4.44%, the inequivalent environmental spaces occupied by native and invaded species and environmental heterogeneity between two ranges both prove the effectiveness of using occurrence information from the native and invaded ranges for predicting the potential invasion.

2.5.2 Significant Difference of Model Performance between Native and Invaded Ranges

Attention should be given to the significant difference on performance of GARP and Maxent models between the invaded and native ranges. Both AUC and OR indicated that the Maxent and GARP models performed significantly different between native and invaded ranges (**Table 2.3**). The difference of model performance in native and invaded regions may result from the intrinsic characteristics of models and the features of occurrence data. It is noticed that occurrence data in the invaded range had a dense distribution in the southeastern U.S. (**Figure 2.3 C**), while occurrence data in the native range was scattered around a much larger region in Asia (**Figure 2.2 C**). As the Maxent model is based on a fitted probability distribution (Phillips et al., 2006), environmental conditions that were more frequent in the dataset are taken as the better potential environment by Maxent. Whereas the GARP model delineates the potential distribution area of species by setting a series of rules, it is less impacted by the sampling density.

As a result, GARP did not constantly show better performance in the U.S. where the sampling density is higher confirming that GARP is less sensitive to the sampling bias (Costa et al., 2010). The significant difference of model performance or sample density between the native and invaded ranges was also observed in similar research conducted by Broennimann and Guisan (2008), and Sobek-Swant et al. (2012). Therefore attention should be given to the difference in sample density between the native and invaded ranges, as the difference may introduce a sampling bias in the models.

2.5.3 Selection of Threshold for GARP and Maxent

The choice of threshold for the GARP and Maxent models is crucial for converting models into binary prediction. The GARP model tends to predict overly extensive areas of potential distribution, whereas the Maxent model tends to be too conservative. Choosing 50% of the best subset models as the threshold for GARP coincides with the observation by Phillips et al. (2006) that for GARP models, areas predicted by 5 to 10 out of the 10 best-subset models were appropriate to be classified as the species' potential distribution area. The value of Maxent threshold was selected to match the percent area predicted as potential distribution to that of binary GARP models. By using this method, the potential distribution model generated by GARP was in high agreement with that generated by Maxent. This result further confirmed that the Maxent model resembles the GARP model when lower levels of predictions were used as threshold (Peterson et al., 2007). The fixed value 50 is not the optimal method to convert suitability maps into binary prediction maps. Both Liu et al. (2005) and Jiménez-Valverde and Lobo (2007) concluded that the fixed threshold was the worst criteria to use to convert suitability maps into binary prediction maps, while the sensitivity-specificity based thresholds were most accurate. However, in this research, as the OR, which is derived from confusion matrix, was used to compare the performance of GARP and Maxent, using a threshold derived from confusion matrix would make the comparison bias towards GARP model. It is recommended that readers select the most appropriate threshold method based on research objectives.

2.5.4 Invasion Risk and Climate Preference of Kudzu Bug

Regions with high invasion risk include most states in the eastern U.S., most countries in Central America, and countries in central South America. Currently kudzu bug is only present in the southeastern U.S. in Americas. However, climate conditions (including temperature and precipitation) of Central America and a large proportion of South America are similar with the conditions of south Asia, where most of the occurrences in native range were distributed (Hijmans et al., 2005). This similarity explains why high invasion risk was predicted in those regions and suggests that kudzu bug can successfully establish there once it has been accidently introduced. Although kudzu bug is the only member found in America in the family of Plataspidae, a close relative of kudzu bug, *Coptosoma xanthogramma* (White), had been found in Hawaii in 1965, and established as a pest of legumes (Beardsley & Fluker, 1967), suggesting kudzu bug can establish in tropical climate regions, where part of Central and South America are located. However, soybean is an important crop in Central and South American countries including Mexico, Brazil, and Argentina, thus to avoid potential economic losses of soybean by kudzu bug, serious quarantine is recommended in these regions to prevent the entry of kudzu bug. In the U.S., the high proportion of niche unfilling suggests that more places could be further infested. Meanwhile, a majority of the soybean production is distributed in upper Midwest, where most of the regions have not been infested by kudzu bug. Therefore, early detection is also highly recommended in the upper Midwest of the U.S.

Kudzu bugs generally inhabit warm (annual mean temperature around 15℃) and humid (annual mean precipitation around 1300mm) regions. This finding coincides with the description of the climatic preference of kudzu bug in native range in Asia (Wang et al., 1996). It had been conjectured that low winter temperature (below -14°C) might kill some bugs in Georgia (Gardner & Olson, 2016). This finding confirmed this conjecture: kudzu bug had low possibility (< 30%) of occurrence in regions with mean temperature of the coldest quarter around 0 °C, whereas the possibility decreases to almost 0 (around 3%) when the mean winter temperature reached -10 °C. In addition to the variables considered in this research, occurrence of the main host plant (i.e. soybean and kudzu) also have played an important role in determining the invasion of kudzu bug (Liang et al., 2019).

2.6 Conclusions

Based on the current distribution of kudzu bug in the U.S., this study suggests that the environmental conditions where kudzu bugs inhabit are not the same between native and invaded ranges. The effectiveness of using occurrence data from both native and invaded ranges to predict potential biological invasion was therefore confirmed. In addition, careful examination of the difference of sample density between native range and invaded range is suggested to avoid sample bias for similar research. Kudzu bug generally inhabits warm and humid places. Management, including monitoring, detection and preventive measures, are suggested for these regions to avoid invasion and outbreak of kudzu bug. Regions with suitable climatic conditions for kudzu bug could be managed to change the microenvironment to avoid severe damage from kudzu bug.

CHAPTER III: THE EFFECT OF PSEUDO-ABSENCE SELECTION METHOD ON TRANSFERABILITY OF SPECIES DISTRIBUTION MODELS IN THE CONTEXT OF NON-ADAPTIVE NICHE SHIFT

This chapter is a reformatted version of a paper by the same name published in *Ecological Modelling* by Liang, W., Papeş, M., Tran, L., Grant, J., Washington-Allen, R., Stewart, S., & Wiggins, G.

Liang, W., Papeş, M., Tran, L., Grant, J., Washington-Allen, R., Stewart, S., & Wiggins, G. (2018). The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. *Ecological Modelling, 388*, 1-9.

3.1 Abstract

Transferability of species distribution models (SDMs) is key to predicting invasion patterns and can be challenged if niche shift occurs in the invaded range. Recently, a non-adaptive niche shift caused by environmental anisotropy was observed on kudzu bug in the invaded range in the U.S. Thus, the first research question of this research is whether transferability of SDMs would be challenged by non-adaptive niche shift using kudzu bug as a case study. When using native occurrences to estimate potential invasions with presence-only modeling methods, it is important to constrain the pseudo-absence (PA) sampling to the species' native range. However, some studies including highly cited ones, do not follow this approach to selecting PA samples. Consequently, the second research question is how model performance would be affected by use of PA samples from outside the native range of the species. To answer the first question, I further quantified the environmental space anisotropy and non-adaptive niche change, and then evaluated the performances of seven SDMs. To answer the second question, I compared the interpolation and transferability of seven SDMs trained with PAs from the

native range and from both native and invaded ranges. Results confirmed that the environmental space anisotropy (*P*=0.01) and non-adaptive niche change (*P*=0.01) are both statistically significant. Of the seven SDMs used, four models had transferability indices higher than 0.9. Boosted regression tree and random forests both had good interpolation and transferability (AUC>0.80 and kappa>0.60), whereas three other models showed good interpolation and fair transferability (AUC>0.70 and kappa>0.40). Inclusion of pseudo-absences from the invaded range significantly increased the interpolation (*P*<0.001) but decreased the transferability (*P*<0.01) of almost all models. Findings in this research suggest that SDMs can still show good transferability with non-adaptive niche shift, thus native occurrence information should be used in similar situations. This research confirmed that it is crucial to constrain the PAs to the same spatial range as presences to accurately model potential invasions.

3.2 Introduction

Estimating the potential distribution of a species across different spatial and temporal scales has become an increasing practical use of SDMs (Elith & Leathwick, 2009; Mainali et al., 2015). In this practice, SDMs are usually trained with occurrence information and environmental predictors from one spatial-temporal range, and then projected to a different range to identify potential distributional areas of a given species (Peterson, 2003; Verbruggen et al., 2013). One underlying assumption of this practice is that the given species conserves its niche across different spatial and temporal scales (Wiens & Graham, 2005). Another assumption is that the given species is in equilibrium with the environment in the spatial range from where the occurrence information is extracted for model training (Elith & Leathwick, 2009; Gallien et al., 2012).

To predict the potential invasion pattern of a given species in a new range, native occurrence information is generally used in SDMs. Thus, successfully predicting invasion patterns closely relies on the transferability of models, which is defined as the ability of SDMs to predict occurrence in a largely unsampled spatial range or time period (Heikkinen et al., 2012; Randin et al., 2006). However, several researchers have suggested that species could shift their niche in new spatial ranges (Broennimann et al., 2007; Early & Sax, 2014; Gallagher et al., 2010; Medley, 2010; Parravicini et al., 2015), which challenges the transferability of SDMs as the same species may survive under different environmental conditions in the invaded ranges (Broennimann et al., 2007; Early and Sax, 2014; Parravicini et al., 2015). As a result, the use of only native occurrence information to model potential invasion is debatable. Broennimann and Guisan (2008) and Jiménez-Valverde et al. (2011) suggested that SDMs for estimating potential invasions could be developed using occurrence information from both native and invaded ranges. However, the drawback of using occurrence information from both ranges is that the equilibrium assumption is likely violated if the invasion is still ongoing (Early & Sax, 2014; Elith & Leathwick, 2009; Jiménez-Valverde et al., 2011).

A study of 50 terrestrial plant invaders, however, suggested that substantial niche shifts are rare (Petitpierre et al., 2012), while the same conclusion was also made for birds and other taxa (Peterson, 2011; Strubbe et al., 2013). Soberón and Peterson (2011) suggested that, in some cases, the "niche shift" is more likely a result of differences in environmental conditions, or environmental space anisotropy, between two spatial ranges than a true, adaptive niche shift of species. The concern that the non-adaptive niche shift would affect the transferability of SDMs has not been fully addressed. Specifically, it is unclear whether both native and invaded range occurrence datasets are needed to estimate invasion patterns when a non-adaptive niche shift exists, given the drawback of invaded range data violating the equilibrium assumption underlying SDMs. When estimating potential invasion patterns with presence-only data, the spatial range from which pseudo-absences (PAs, also called background data) are extracted certainly impacts the transferability of SDMs (Phillips, 2008; Barbet-Massin et al., 2012).

For studies that use only occurrences from the native range of a species to train models, the PAs also should be restricted to the native range (Peterson, 2003; Phillips, 2008). However, in several highly cited studies, the PAs were extracted from both the native and invaded ranges when only native occurrences were used to construct the

SDMs (Broennimann & Guisan, 2008; De Meyer et al., 2010; Gallardo et al., 2013; Verbruggen et al., 2013). Additionally, in some studies, the methods used to extract the PAs were not mentioned (Fitzpatrick et al., 2007; Loo et al., 2007; Mau-Crimmins et al., 2006). Evaluation of the impact of PA data on model performance is not rare (Anderson & Raza, 2010; Barbet-Massin et al., 2012; Zhu et al., 2014). However, what is missing is a systematic and quantitative assessment of the impact of selecting PAs from both native and invaded ranges, while only native occurrences are used, on both interpolations and transferability of multiple commonly used models. This information is important to consider, especially when niche shift is observed between the two ranges.

This study had two objectives: 1) evaluate the interpolative accuracy and, more importantly, transferability (also called extrapolative accuracy) of SDMs obtained with seven commonly used techniques, under non-adaptive niche shift between native and invaded ranges, and 2) examine how inclusion of PAs from the invaded range, when only native occurrences are used for model training, impacts model interpolation and transferability. This research was conducted with kudzu bug (*Megacopta cribraria*; Hemiptera: Plataspidae) for several reasons. First, kudzu bug is a newly invasive species in the United States (U.S.), where it was first found in 2009 (Suiter et al., 2010); however, it is a well-studied species in Asia, its native range. Kudzu bug spread fast, such that it had been reported in more than 650 counties in the U.S. by the end of 2017, ensuring enough occurrence data in the invaded range for SDMs. Additionally, a non-adaptive niche shift was observed between native and invaded populations of kudzu bug in the U.S. (Liang et al., 2018a). Modeling the potential distribution of kudzu bug is also important for practical applications. Results of this research can provide valuable information for management of kudzu bug, as it has become a pest in both agricultural and urban areas (Eger et al., 2010). To fulfill the first objective, the differences of both available and occupied environmental spaces between native and invaded ranges of kudzu bug was quantified. As empirical research, these findings can be used to guide selection of SDMs when a non-adaptive niche shift is observed and to select PAs for accurate modeling of species' distribution.

3.3 Methodology

3.3.1 Presence and Pseudo-absence Data

Presence To compare the interpolation and transferability of different models, point presence data of kudzu bug were collected in both native and invaded ranges. Readers are referred to Liang et al. (2018a) and Zhu et al. (2012) for detailed information on data acquisition. Redundant observations were deleted to ensure only one observation per 1000-m grid. In total, three datasets were generated: Dataset I included 164 presence records in the native range (Asia), Dataset II contained 152 presence records in invaded range of kudzu bug (U.S.), and Dataset III combined occurrences from both the native and invaded ranges and therefore had 316 presence records (**Figure 3.1**). All occurrence data are available to readers upon request.

Pseudo-absence (PA) To address the lack of absence data, 10,000 PAs were extracted for model training and intrinsic evaluation. To quantify and compare the accuracy of model interpolation and transferability, PAs were extracted in the same spatial range as the presences used for model training. However, to fulfill the second objective, the Dataset IV was generated, which included the presences from Asia but 10,000 PAs extracted from both Asian and U.S. ranges.

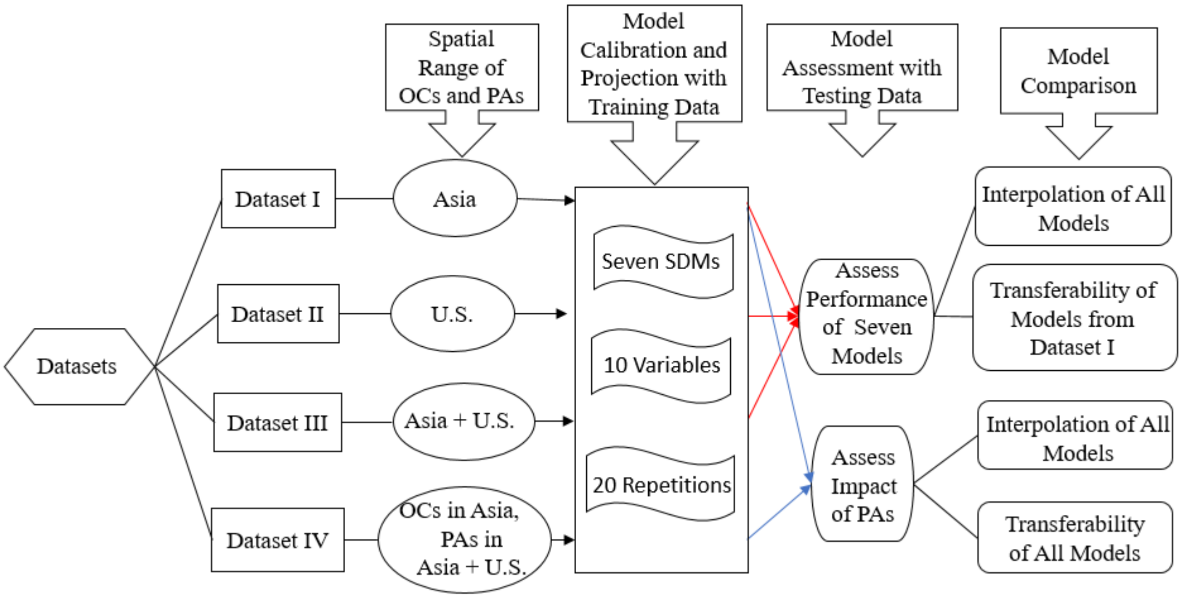

Figure 3.1 Flow chart of species distribution modellings (SDMs) for comparing performance of seven models and assessing impact of pseudo-absences (PAs) on model interpolation and transferability. OCs is short for occurrences.

3.3.2 Environmental Variables

To characterize the environmental space in both ranges (native and invaded), an elevation variable from the Hydro-1k digital elevation model dataset (USGS, 1996) and nine climatic variables from the “WorldClim” dataset (Hijmans et al., 2005) were used. Instead of using all 19 climatic variables available in WorldClim, only nine variables that were not highly correlated ($r<0.8$) were selected based on the climatic conditions of 10,000 randomly selected points. The nine climatic variables are mean diurnal temperature range (BIO2), maximum temperature of the warmest month (BIO5), mean temperature of the wettest quarter (BIO8), mean temperature of the coldest quarter (BIO11), annual mean precipitation (BIO12), precipitation seasonality (BIO15), precipitation of the driest quarter (BIO17), precipitation of the warmest quarter (BIO18), and precipitation of the coldest quarter (BIO19). All 10 variables (elevation and climate) used had a resolution of 30 arc-seconds (approximately 1 km).

3.3.3 Quantifying Available and Occupied Environmental Space Change

To represent the available environmental space in both native and invaded ranges, the values of the 10 environmental variables at 1,000 randomly selected points from each range (2,000 points total) were extracted. The 10 environmental features of the presence datasets in the two ranges were considered to represent the occupied environmental space. Principal component analysis (PCA) was run separately on the available and occupied environmental spaces of the two ranges. To quantify the change in both occupied and available environmental space between two ranges, a between-class analysis was conducted on the PCA results of occurrences and background data

(Broennimann et al., 2007; Soberón & Peterson, 2011). Monte Carlo simulation was then used to test the significance of the between-class inertia. Between-class analysis (Dolédec & Chessel, 1994) measures the difference between groups of individuals, while the inertia quantifies how much variance of the data is caused by group difference. However, to visualize the occupied and available environmental space on the principal components (PCs), PCA on the environmental variables of all background and occurrence data was conducted in the two ranges.

3.3.4 Model Development

Training and Testing Dataset

The flow chart of distribution modeling is shown in **Figure 3.1**. To generate the models, the data were randomly split into 70% for training and 30% for extrinsic testing of the models. This step was replicated 20 times. For each model run, the 70% training data was further split into 70% for model construction and 30% for intrinsic model evaluation.

Modeling Techniques

Seven widely used algorithms for species distribution modeling were included: artificial neural network (hereafter ANN; Hecht-Nielsen, 1988), boosted regression tree (hereafter BRT; Friedman, 2001), genetic algorithm for rule-set production (hereafter GARP; Stockwell, 1999), generalized additive model (hereafter GAM; Hastie & Tibshirani, 1990), multiple adaptive regression splines (hereafter MARS; Friedman, 1991), maximum entropy (hereafter MAXENT; Phillips et al., 2006), and random forest (hereafter RF;

Breiman, 2001). Although only GARP and MAXENT were developed for presence-only species data, other studies have successfully used the other five presence-absence models on presence-only data (Broennimann & Guisan, 2008; Elith et al., 2006) and thus were included here. Among the seven algorithms, ANN, BRT, MAXENT, and RF are four machine-learning methods. GAM and MARS are two additive models, and BRT and RF are decision tree based methods. MAXENT is based on maximum entropy principle, while GARP is a genetic algorithm. Details of these seven models were well documented by Elith et al. (2006), Lek and Guégan (1999), Phillips et al. (2006), and Stockwell (1999), thus they are not described here. To run the models, MAXENT (Version 3.3.3) and GARP programs were used, and the "biomod2" package in R for the remaining five modeling methods (Thuiller et al., 2016).

3.3.5 Model Evaluation and Comparison

Two metrics, area under the curve (AUC) of the receiver operating characteristic and maximum kappa, were used to evaluate the models. AUC is a threshold-independent measurement of model performance (Fielding & Bell, 1997). The value of AUC varies from 0 to 1, representing the possibility that a randomly selected prediction from the presence group has a higher probability of presence than a randomly selected prediction from the absence group (Fielding & Bell, 1997). A value of 0.5 implies the model prediction is no better than a random guess, while a value in the range of 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, and 0.9-1.0 suggests fail, poor, fair, good, and excellent performance of models, respectively (Swets, 1988). The kappa statistic (Cohen, 1960) is a threshold-dependent measure of model performance, and the maximum kappa was used as the model measure in this research. The value of kappa statistic varies from -1 to 1. The kappa (K) values were interpreted as representing: poor models K< 0.4; good 0.4<K<0.75; and excellent models 0.75<K<1 (Landis & Koch, 1977). AUC and kappa were calculated using "ROCR" package (Sing et al., 2005) and "ecospat" package (Broennimann et al., 2016), respectively, in the R program. For the ensemble model, which will be described in Section 2.6, omission rate was used to assess their performance in both Asia and the U.S.

To evaluate model interpolation, the same number of PAs were randomly extracted as the corresponding extrinsic testing presences. For the first three datasets, the extrinsic PAs were extracted in the same spatial range as the presences, while for Dataset IV the extrinsic PAs were extracted from both ranges (native and invaded). To assess the transferability of algorithms, only models trained with Dataset I and Dataset IV were used (presences from the native range; PAs from native and native plus invaded range, respectively) and evaluated with all invaded range presences and the extrinsic PAs extracted from invaded range. To decrease the uncertainty, 100 sets of extrinsic PA data were randomly extracted for each model evaluation. Each model output was evaluated 100 times with the corresponding testing data and 100 sets of extrinsic PAs, and the mean value was calculated as the final measurement of each model. When extracting extrinsic PAs, the pixels that had been chosen for model construction were avoided. For each algorithm and dataset, the evaluation was conducted on all 20 replicate runs. To compare the transferability of each algorithm, a transferability index (TI) for both AUC and kappa was calculated as follow:

$$TI_{AUC} = \frac{AUC_{extrapolation}}{AUC_{interpolation}} \text{ and } TI_{Kappa} = \frac{Kappa_{extrapolation}}{Kappa_{interpolation}}$$

Tukey's HSD (honest significant difference) test was used to conduct multiple comparisons on AUC and kappa of the seven algorithms. The paired one-tail Wilcoxon signed-rank test was used to conduct pairwise comparisons.

3.3.6 Model Visualization

To visualize the potential distributions estimated by the models constructed with the first three datasets (Datasets I, II, and III), an ensemble model was derived by averaging the best four models selected from 20 replicate runs. The ensemble model was further evaluated with all occurrences from the same range as the occurrences used for model construction. Although this evaluation approach was based on data that had been used during model construction, this analysis was applied only to find a possible threshold, which maximizes the kappa, to better visualize the potential distribution as estimated by the ensemble model.

3.4 Results

3.4.1 Quantifying Changes in Occupied and Available Environmental Spaces

The PCA on the environmental features of occurrence data indicated that the first two PCs explained 64.3% of the total variance. The between-class inertia ratio was 0.335, indicating that 33.5% of the variable inertia can be explained by the differences in spatial location of occurrence points. The Monte Carlo simulation on between-class inertia ratio (*P*=0.01) suggested a substantial difference of occupied environmental conditions between native and invaded ranges. The results of between-class inertia analysis on the available environmental space are consistent with the observations by Liang et al. (2018a) that environmental space anisotropy exists between native and invaded ranges. The first and second PCs explained 41.4% and 21.5% of the total variance of variables, respectively (**Figure 3.2 A**). The between-class inertia on the available environmental space was 6.07%, and the difference of environmental space between the two ranges was significant (*P*=0.01).

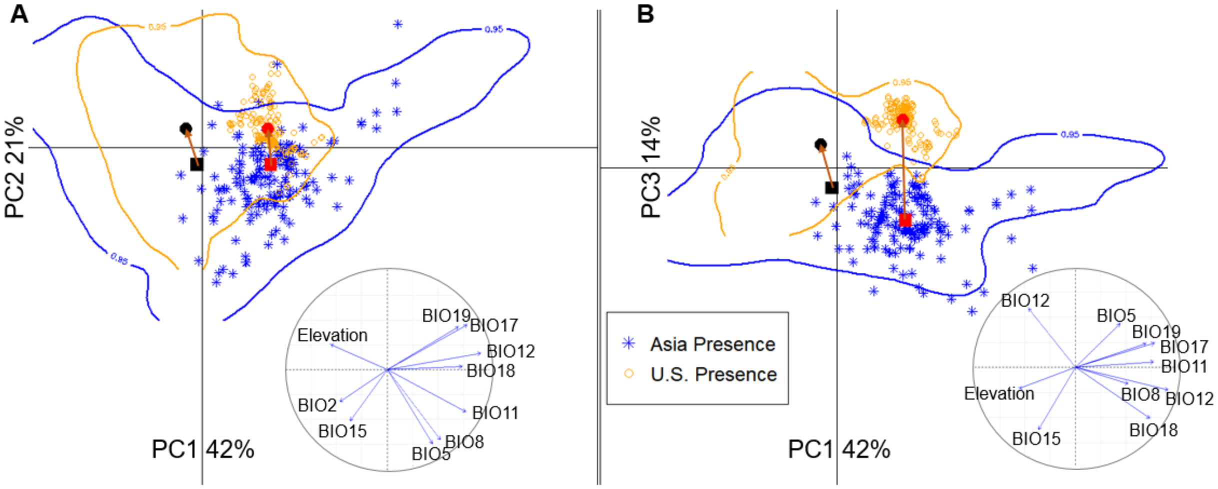

Figure 3.2 Plots of available and occupied environmental spaces in the area, where kudzu bug is found (native Asia and invaded U.S.). **A** and **B** show available and occupied environmental space along the first two PCs (principal components) and the first and third PCs, respectively. Cumulatively, the first three PCs explained 77% of the total variance. The orange and blue lines indicate 95% of the available environmental space in the U.S. and Asia, respectively. The blue stars and orange circles indicate occupied environmental space by kudzu bug in Asia and U.S., respectively. The black dot and square represent the centroid of the available environmental space in U.S. and Asia, respectively, while the red dot and square represent the centroid of the occupied environmental space in U.S. and Asia, respectively. The brown arrows indicate the shift of the centroid of available and occupied environmental spaces between native and invaded ranges.

Based on the first two PCs (**Figure 3.2 A**), the majority of presences in both native and invaded ranges were distributed in the environmental space that was available in both ranges. However, based on PC1 and PC3, presences in Asia and the U.S. were mostly distributed in environmental spaces which are not available in the other ranges (**Figure 3.2 B**). The environmental space anisotropy mainly occurred along the PC2 and PC3 (explaining 35% of the variance). The shift of centroids of available and occupied environmental spaces between two ranges mainly occurred in the same dimension (brown arrows in **Figure 3.2**), whereas along PC1, which explained the most variance of variables (42%), no discrepancy was observed between two ranges. However, it is clear that along PC1 both the available and occupied environmental spaces in Asia included wider ranges than those in the U.S. (**Figure 3.2**).

3.4.2 Model Interpolation and Transferability

Model Interpolation

The mean AUC and kappa of all models were higher than 0.8 and 0.45, respectively, except for the GARP model trained with the native presences (AUC=0.78, kappa=0.57, **Figure 3.3**). This result suggested nearly all models had good interpolations. Models trained with Dataset II had constantly higher AUC and kappa than the ones trained with Dataset I and Dataset III, while models trained with Dataset I constantly showed the lowest performances among the three datasets. Both measurements indicated that MAXENT, BRT, MARS, and RF were the highest performing four models for all datasets (AUC > 0.89, kappa>0.65), whereas the first three models were ranked in the highest level of interpolation five out of six times (**Figure 3.3**).

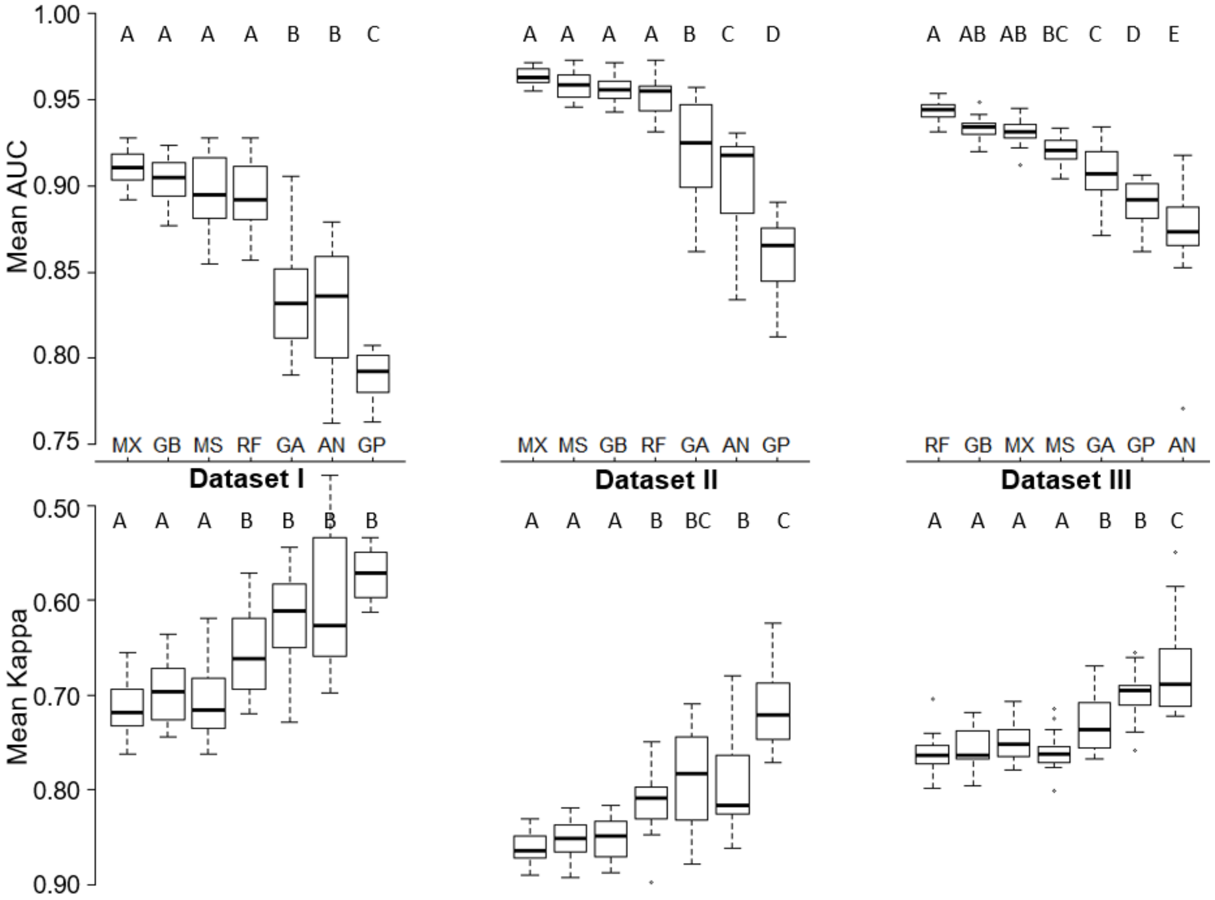

Figure 3.3 Mean AUC and kappa for all models trained with three datasets. Dataset I, II, and III include presences and pseudo-absences in native Asia, invaded U.S., and both native and invaded ranges, respectively. Models followed by same letters within a dataset are not significantly different (Tukey's HSD test at α=0.05). MX, GB, MS, AN, GA, GP are short for MAXENT, BRT, MARS, ANN, GAM, and GARP, respectively. Dots in the plot indicate potential outliers. MAXENT, BRT, MARS, and RF were constantly ranked as the best four models.

Model Transferability

The ranking based on model transferability suggested RF, BRT, MAXENT, and ANN were the highest performing models (**Figure 3.4**). RF and BRT both had good transferability (AUC>0.8), while MAXENT, ANN, and GARP all had fair transferability (AUC>0.7). Based on kappa, the first five models had good transferability (kappa>0.4). However, from the highest to lowest, the two evaluation measurements had the same rank of transferability index: RF, BRT, ANN, GARP, MAXENT, MARS, and GAM. The two models with highest transferability (BRT and RF) were both tree-based algorithms, while the two models with lowest transferability (MARS and GAM) were both regression-based algorithm. Although the AUC and kappa on model transferability were significantly lower than the ones on model interpolation, four models had transferability indices higher than 0.9 based on AUC.

Visual Interpretation

The ensemble model used to estimate the potential distribution of the kudzu bug was derived from BRT, MARS, MAXENT and RF (**Figure 3.5 A-F**). For Dataset I, the ensemble model derived from the best four models (Dataset I-derived ensemble model) successfully predicted high probability of presence (>50) in southern Asia where occurrence data were distributed (**Figure 3.5 A**). However, in the U.S. the high values were only predicted narrowly in the south (**Figure 3.5 B**) rather than the southeastern region, where most of the occurrences were observed and reported. Similarly, the Dataset II-derived ensemble model successfully predicted high values for the southeastern U.S. (**Figure 3.5 D**), but low values for the vast range of Asia (**Figure 3.5 C**). Compared with

Datasets I and II, Dataset III-derived ensemble model predicted high values for both south Asia and southeastern U.S (**Figure 3.5 E, F, Table 3.1**).

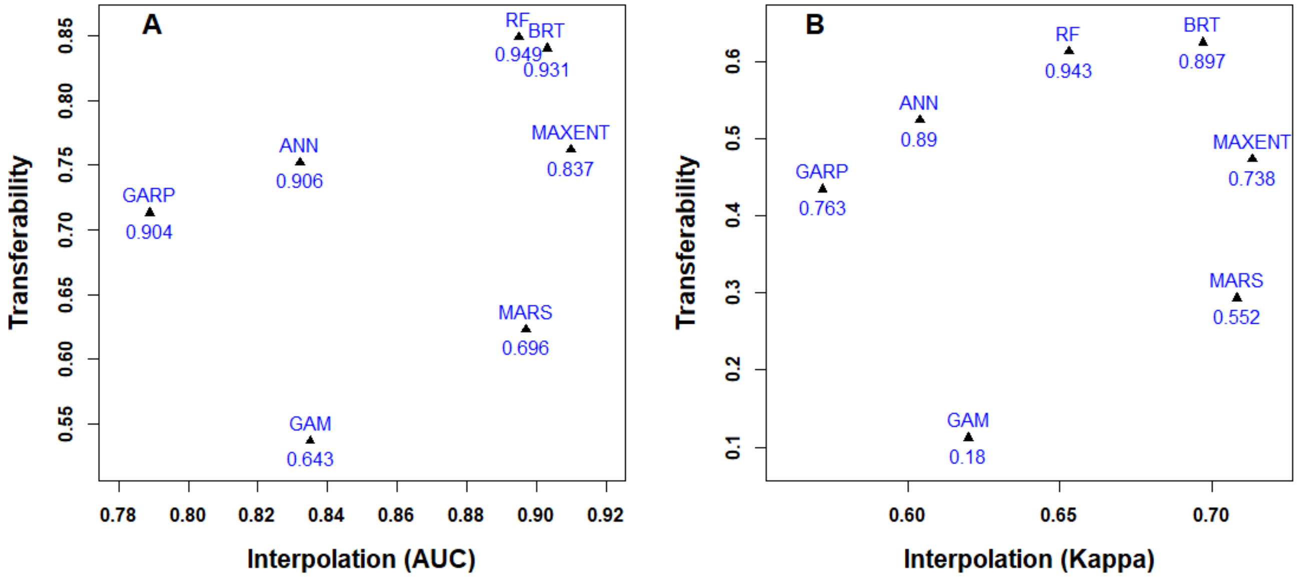

Figure 3.4 Evaluation, with AUC (**A**) and kappa metrics (**B**), of interpolation and transferability abilities of seven algorithms trained with Dataset I (presences and pseudo-absences from native range, Asia). The values of transferability indices are shown in blue in the two plots. Both AUC and kappa indicate that RF, BRT, and ANN have the highest transferability.

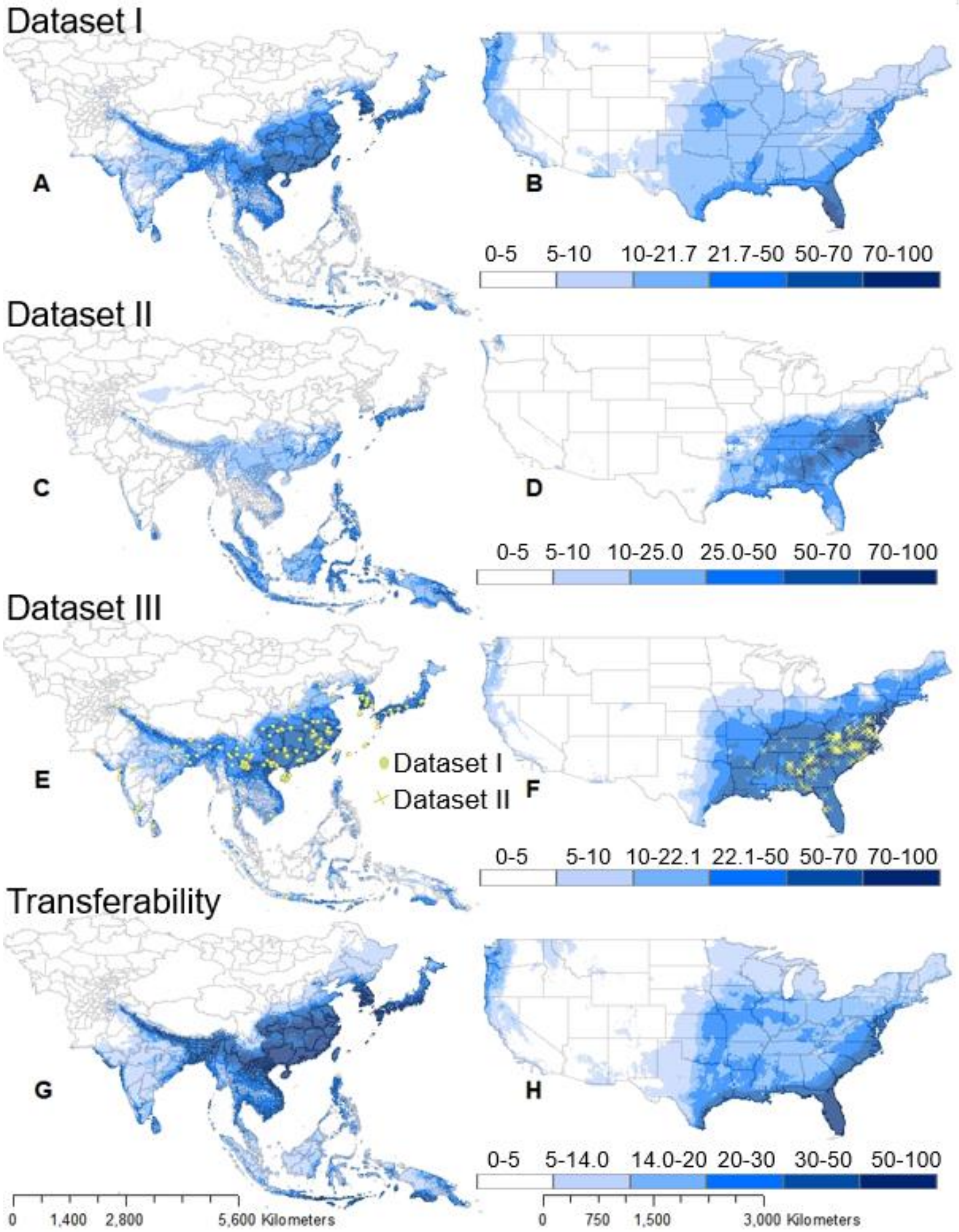

Figure 3.5 Ensemble models of presence probability of kudzu bug derived from three datasets. Dataset I, II, and III include presences and pseudo-absences in native Asia, invaded U.S., and both native and invaded ranges, respectively. The transferability model was trained with Dataset I. The thresholds for the corresponding ensemble models at which the maximum kappa of models was reached were 21.7, 25.0, 22.1, and 14.0.

Table 3.1 Mean predicted value, omission rate, and percent area of potential distribution obtained with the three ensemble models for interpolation and the ensemble model for transferability

Performance	Dataset	Mean predicted value		Omission rate (%)		Percent area (%)	
		Asia	U.S.	Asia	U.S.	Asia	U.S.
Interpolation	I	64.55	21.27	0.61	68.42	18.07	9.51
	II	12.94	68.05	86.59	0	6.55	15.05
	III	58.90	79.19	4.27	0	16.78	26.16
Transferability	I	66.53	25.95	0	2.63	30.29	38.79

The ensemble model of the four models with highest transferability was derived from RF, BRT, MAXENT and ANN (**Figure 3.5 G, H**). The ensemble models obtained using the maximum kappa as the cut-off threshold selection predicted the largest areas of potential distribution for both native and invaded ranges (**Table 3.1, Figure 3.5 G, H**), while the Dataset III-derived ensemble model predicted moderate potential distribution areas for both ranges (**Table 3.1, Figure 3.5 E, F**).

3.4.3 Impact of Spatial Range of Pseudo-absences on Model Performances

As AUC and kappa indicated similar results, only AUC was presented in this section (**Figure 3.6**). All models, which were trained with native occurrences but with PAs from both native and invaded ranges (hereafter Asia/All), had significantly higher AUCs on interpolation than models (P <0.001 for all), which were trained with native occurrences and PAs from native range (hereafter Asia/Asia) (**Figure 3.6**). However, except for GARP (P= 0.999), the AUCs on transferability of all models trained with Asia/All were much lower than models trained with Asia/Asia (P<0.01 for all, **Figure 3.6**). Meanwhile, all transferability indices (TIs) of models trained with Asia/Asia were significantly higher than those trained with Asia/All except GARP (**Table 3.2**).

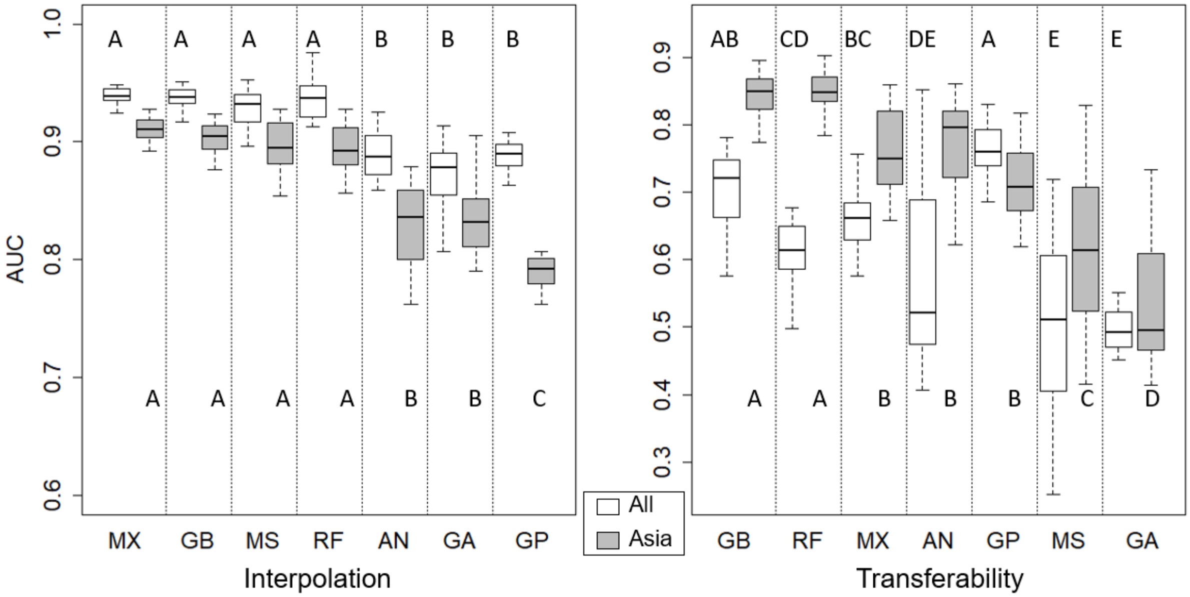

Figure 3.6 Mean AUC of interpolation and transferability for all seven models trained with pseudo-absences (PAs) from native range (Asia) only and all ranges (Asia + U.S.) of kudzu bug. MX, GB, MS, AN, GA, and GP are short for MAXENT, BRT, MARS, ANN, GAM, and GARP, respectively. Models followed by same letters within a group are not significantly different (Tukey's HSD test at α=0.05). Letters in the upper and lower spaces indicate differences among models trained with background data from all the ranges and only from the native range, respectively. *P*-values of the pairwise comparisons on interpolation AUC of models using native PAs and all-range PAs are smaller than 0.001 for models. *P*-values of the pairwise comparisons on transferability AUC are smaller than 0.01 for all models, except GARP model (*P*=0.999).

Table 3.2 Transferability index values of seven algorithms trained with background data from both native (Asia) and invaded (U.S.) ranges and only native range of kudzu bug

Background Range	RF	BRT	ANN	GARP	MAXENT	MARS	GAM
Asia	0.949A	0.931A	0.906AB	0.904AB	0.837B	0.696C	0.643C
Asia + U.S.	0.662CD	0.754B	0.653CD	0.869A	0.708BC	0.553E	0.577DE
P	0.0000	0.0000	0.0001	0.0768	0.0000	0.0077	0.0007

P represents the P-value of the Wilcoxon Signed-Rank test on the difference of transferability index of the models trained with all background data and the Asian background data. Values followed by the same letters within a row are not significantly different (Tukey's HSD test at α=0.05)

3.5 Discussion

3.5.1 Model Interpolation and Transferability

Predictive SDMs are important to provide information on management of invasive species and conservation of endangered species. This case study identified that BRT, MARS, MAXENT, and RF were the highest ranked models of interpolation for all datasets used. Although the performance of a given SDM varies according to the species studied (Elith & Graham, 2009), MAXENT, BRT, and MARS frequently perform better than most methods as shown from research on multiple species (Elith & Graham, 2009; Elith et al., 2006; Guisan et al., 2007; Heikkinen et al., 2012; Wisz et al., 2008). This good performance indicates that these three models generally have better capability of modeling the relationships between species and selected environmental features. The low performance of GARP is also frequently observed (Guisan et al., 2007; Elith &

Graham, 2009; Heikkinen et al., 2012; Wisz et al., 2008). Elith and Graham (2009) found that the GARP model failed to model the relations of a simulated species to all environmental features considered, while other four models all correctly modeled the relationships to some extent. This failure of recognizing the relationship of species to environment may lead to the low performance of GARP models.

Interpolation and transferability are equally important when the SMDs are aimed to predict the potential distribution of species in largely unsampled regions or time periods. Thus, models that have both high interpolation and transferability are desirable. This case study identified that BRT and RF were among the highest ranked models of both interpolation and transferability. Although in most cases there is a trade-off between model precision (here as interpolation) and generality (here as transferability) (Guisan and Zimmermann, 2000), this case study confirmed that the achievement of both precision and generality of SDMs is possible. Mi et al. (2017) also found that RF showed the best performance in terms of interpolation and transferability with a research on three crane species. The result that tree-based algorithms (BRT and RF) showed good transferability is consistent with the research by Prasad et al. (2006) and Mi et al. (2017). The low transferability of regression-based algorithms (MARS and GAM) is consistent with the result of Prasad et al. (2006) on both models and Heikkinen et al. (2012) on MARS, but contrary to the observation on GAM by Heikkinen et al. (2012). However, the high transferability of RF in this research is contrary to the result from Breiner et al. (2018) where they observed low transferability of RF comparing to models like MARS, MAXENT, GAM, and GBM. The inconsistency of model rankings on transferability in different research suggest that there is not a single best transferable model for all cases (Breiner et al., 2018; Mainali et al., 2015).

Although a significant shift was observed between both the available and occupied environmental spaces in the native (Asia) and invaded (U.S.) ranges of kudzu bug, five of the seven models showed fair transferability, while at least two of the models showed good transferability. Findings of this research are comparable with the study by Heikkinen et al. (2012) in which ANN, GARP, BRT, RF and MAXENT were classified as complex models and the transferability index (mean=0.900, based on AUC) was similar to results in this study (mean=0.905). However, with the non-adaptive niche shift, high transferability of models was not expected. The good and fair transferability of the SDMs in this research could result from the fact that the shift of occupied environmental space is caused by a non-adaptive niche shift rather than an adaptive niche shift. Although researchers found that the environmental difference between two ranges decreased transferability of SDMs (Qiao et al., 2018; Yates et al., 2018), this study suggests that even with significant changes of environment between two ranges, SDMs trained with native-only occurrences can still show good transferability. Thus, unlike suggestions by Broennimann and Guisan (2008) and Jiménez-Valverde et al. (2011), when niche shift is observed, using occurrence information from both native and invaded ranges would not always be preferred, especially when the equilibrium in the invaded range is not reached. Before deciding what occurrence information should be used for predicting potential invasion patterns, one could first determine the type of niche shift (adaptive or non-adaptive) as well as the environmental space anisotropy between native and invaded ranges. In some cases, non-adaptive niche shift does not challenge the transferability of

SDMs, thus, the use of native-only occurrence information is suggested. For the niche shift caused by niche expansion or environmental space anisotropy, which would seriously challenge the transferability of SDMs, one could run the SDMs including occurrence data in new ranges where the invasive species have already established (Gallagher et al., 2010).

3.5.2 Impact of Spatial Range of Pseudo-absences on Model Interpolation and Transferability

Research on the impact of the spatial range of PAs on model performances is not rare (Barbet-Massin et al., 2012; Stokland et al., 2011; VanDerWal et al., 2009), however, such research only considered model interpolation. This systematic evaluation of the impact of the spatial range of PAs on model transferability has general implications on the selection of PAs for estimating potential distributions of non-native species. Model transferability can also be significantly impacted by the spatial range of PAs, thus confirming the importance of constraining PAs to the same spatial range as occurrences. The inclusion of PAs from the invaded range when only native occurrences are used can significantly increase model interpolation but decrease model transferability, if environmental space anisotropy exists between native and invaded ranges. As the available environmental features in the invaded range are significantly different from those in the native range, the PAs from the invaded range increased the ability of models to discriminate the presence from the pseudo-absence. Thus, including PAs from the invaded range significantly improved model interpolation for all seven algorithms used here. However, as some PAs from the invaded range were actually extracted from the suitable range of the invasive species, this inclusion decreased the ability of models to recognize the potential distribution area in the new range, resulting in underestimation of model transferability. Anderson and Raza (2010) and Zhu et al. (2014) also found that constraining the PAs into a properly small region improved model transferability, and can be explained as the overfitting of models with PAs from larger spatial extent. Factors other than PAs, such as environmental features used for model construction, also can impact model transferability (Verbruggen et al., 2013; Yates et al., 2018).

Using only native occurrences but PAs from both ranges increases model accuracy at the expense of decreasing model generality. Broennimann and Guisan (2008) concluded that occurrence information from both native and invaded ranges should be used by comparing the interpolation and transferability of SDMs trained with occurrence information in the native range only and in both native and invaded ranges. However, in their research, Broennimann and Guisan (2008) used the same PAs extracted from the study area for all models, including models trained only with native occurrences (the original description of their method of extracting PAs is “a corresponding number of pseudo-absences were sampled randomly across both native and invaded ranges”; Broennimann and Guisan, 2008). Several other researchers also used PAs from spatial ranges where presence data were not used to develop distribution models (De Meyer et al., 2010; Gallardo et al., 2013; Verbruggen et al., 2013). As suggested by this study, PAs from the spatial range where presences were not used for model training can significantly decrease model transferability, especially when the environmental space anisotropy exists between two ranges. Therefore it is important to constrain the PAs to the same spatial range as the presence data to determine the best dataset used for development of SDMs.

3.6 Conclusions

This study addressed two questions on distribution modeling of invasive species with non-adaptive niche shifts between native and invaded ranges. I first evaluated the interpolation and transferability of seven commonly used models, and then systematically assessed the impact of spatial range where the PAs are extracted on models' interpolation and transferability. I found that five models (RF, BRT, MAXENT, ANN, and GARP) showed good or fair transferability (AUC > 0.7, transferability index >0.8, **Figure 3.4**), indicating the non-adaptive niche shift does necessarily challenge the transferability of SDMs. Although Broennimann and Guisan (2008) suggested that both native and invaded occurrences should be used for invasion modeling when a niche shift is observed, one should first determine the type of niche shift (adaptive or non-adaptive). Non-adaptive niche shift, in some cases, does not challenge model transferability, thus occurrence data in native range should be used, especially when the equilibrium in the invaded range is not reached. I also found that the spatial range where PAs are extracted can significantly impact both interpolation and transferability of SDMs, when environmental space anisotropy exists between two ranges. Using kudzu bug as a case, extracting PAs from both the native and invaded ranges, while only native presences were used for model construction, significantly overestimates interpolations of all models ($P<0.001$ for all, **Figure 3.5**), but underestimates transferability of all models (**Table 3.2**). The challenge of the environmental space anisotropy on model transferability can be largely avoided or reduced by constraining the PAs to the same spatial range as the presences. Therefore, constraining the PAs to the same spatial range as presences used for model construction to retain interpolation and transferability. For the classical approach of modeling potential invasions with native occurrences, the PAs should only be extracted from the native range to avoid misestimates of model performances. Based on potential distribution maps, kudzu bug may further spread north and west of the U.S. into major soybean growing areas, including Kentucky, Indiana, Ohio, Illinois, Missouri, Iowa, Kansas, Oklahoma, and Texas (**Figure 3.5 H**). Given the serious economic damage of kudzu bug on soybean (Seiter et al., 2012), more stringent detection programs are suggested for these states to reduce the infestation risk and potentially economic loss.

CHAPTER IV: MODELING INVASIONS WITH GEOPOLITICAL-UNIT LEVEL RECORDS: COMPARING PERFORMANCES OF COMMON METHODS TO ESTIMATE OVERALL SPREAD RATE AND DYNAMICS

4.1 Abstract

To determine best method to estimate invasion dynamics, I evaluated the abilities of common methods to estimate expansion pattern, overall rate, and spread dynamic of invasive species, assessed impacts of variations in geopolitical-units on each method, and finally analyzed similarity patterns of all methods. A new boundary displacement method (i.e., CtdBD) and an alternative method (i.e., NctyArea) of using the cumulative number of infested geopolitical units (Ncty) to estimate spread with geopolitical-unit data was proposed. Simulated spread using counties in the U.S. was used, and eight methods to estimate the spread rates were applied. Three regions with different mean sizes of counties were selected to conduct simulations independently, and three spread scenarios were simulated. R^2 and root mean square error were used to evaluate the performance of all methods. Pearson correlation coefficient was used to assess the similarity of estimated spread by all methods. Finally, kudzu bug, *Megacopta cribraria*, was used as a case study to test the generality of results concluded from the simulated research. CtdBD, Ncty and NctyArea methods correctly estimated the expansion patterns for all scenarios and regions. Two boundary displacement and two area-based regression methods estimated highly correlated spread rates and were the best four methods, among which CtdBD had the best estimation. Distance-based regression methods are sensitive to the irregularity and stochasticity in spread, and the minimum spread distance method had a low ability to estimate spread. NctyArea had better performances than Ncty when geopolitical units vary largely in size. The case study showed consistent results with the simulated research. Both regression and boundary displacement methods can

estimate expansion pattern, overall rate, and spread dynamics of invasive species. Generally, the abilities of all methods to estimate spread decrease with the increase of irregularity and stochasticity in spread. Boundary displacement methods best estimate overall rate and spread dynamics, however, for spread without clear infestation outlines, area-based regression methods can be good alternatives.

4.2 Introduction

Modeling invasion dynamics is important to the management of invasive species, as it can facilitate prediction of spatial and temporal invasion risks of the species, enhance early detection, guide the construction of early warning system, and determine important factors affecting the invasions (e.g., Liang et al., 2019; Paini et al., 2016; Stohlgren & Schnase, 2006). In practice, estimating spread rates of invasive species had been conducted on various species, such as aquatic animal, disease, insect, and plant, on all spatial scales including landscape (e.g., Haregeweyn et al., 2013), local (e.g., Sharov et al., 1999), regional (e.g., Evans & Gregoire, 2007), continental (e.g., Pyšek et al., 2008), and global scales (e.g., Suarez et al., 2001). Selection of invasion records for estimating spread is closely related to the spatial scales of research. For research conducted at local or smaller scales, spread data collected through field sampling or censuses are usually used (e.g., Pratt et al., 2003; Sharov et al., 1999). For plant species, time-series satellite or aerial images are also constantly used (e.g., Haregeweyn et al., 2013; Pyšek et al., 2008). However, for research with large scale, such as regional, continental, or even global scales, researchers usually have to collect all available records from online database, published research, survey, or field sampling (e.g., Evans & Gregoire, 2007; Masciocchi & Corley, 2013; Pyšek et al., 2008; Suarez et al., 2001). Consequently, data for large-scale research usually has coarse and non-unified resolution.

The collected data are often converted to geopolitical-unit level data recording presence/absence of invasive species for each geopolitical unit (e.g., Evans & Gregoire, 2007; Horvitz et al., 2017; Lantschner et al., 2014). Additionally, quarantines initiated by governments or institutions are usually conducted at geopolitical-unit levels, such as county or township (e.g., Perrins et al., 1993; Tobin et al., 2007 & 2015). Therefore, geopolitical-unit level data are usually the most abundant and available records of invasive species (Evans & Gregoire, 2007; Liebhold et al., 1992; Tobin et al., 2007; Tobin et al., 2015). During the past decades, researchers worldwide had used geopolitical-unit record to estimate invasion rates of various species (e.g., Evans & Gregoire, 2007; Horvitz et al., 2017; Perrins et al., 1993). Such research was mostly conducted at regional scale (e.g., Horvitz et al., 2017; Morin et al., 2007; Perrins et al., 1993), but could also be conducted at continental (e.g., Lantschner et al., 2014; Pyšek et al., 2008), or global scale (e.g., Liu et al., 2014; Suarez et al., 2001).

Researchers have developed multiple mathematical models, such as partial differential equation (Fisher, 1937) and simple reaction-diffusion model (Skellam, 1951), to estimate spread rates of invading organisms in relation to their population density and other life characteristic-related parameters (reviewed in Hastings et al., 2005). These mathematical models may be capable to estimate spread at small scale for some species (Hastings et al., 2005; Suarez et al., 2001). However, for spread at large spatial scales, external factors, such as spatial heterogeneity, human transport processes, and air current, are more important to the spread of invading species than species-intrinsic characteristics (Hastings et al., 2005; Liang et al., 2019; Nathan, 2006). Additionally, researchers found that long distance jump dispersal (LDJD) caused by random events are much more influential than local dispersal for many species (Kot et al., 1996; Mineur et al., 2010; Nathan, 2006). Consequently, invasive species at large spatial scale often show highly asymmetric spread pattern with large stochasticity, resulting in inabilities of mathematical models on estimating spread rates.

Estimating spread of invasive species over large spatial scale, therefore, uses alternative approaches. Multiple methods, including regression, boundary displacement, and minimum spread distance methods, had been commonly used. Several researchers had compared the accuracy of these common methods on estimating spread of invasive species (e.g., Gilbert & Liebhold, 2010; Tobin et al., 2015). However, existing research only compared the overall estimated rate (Gilbert & Liebhold, 2010; Tobin et al., 2015). Spread of invasive species, especially at large scales, is commonly complex due to spatial heterogeneity and stochastic events, such as LDJD (Hastings et al., 2005; Pyšek et al., 2008). Estimating spreads of invasive species with geopolitical-unit data, however, further increases the uncertainties, as there can be large variations on the sizes of geopolitical units (Hastings et al., 2005; Pyšek et al., 2008). Consequently, the spread rates possibly vary at different spatial regions and temporal periods. Thus, compared to one single overall spread rate, estimating spread dynamics is more informative to understand the invasion dynamics (Hastings et al., 2005). In addition to the lack of a comprehensive evaluation of common methods on estimating both overall rate and spread dynamics, a systematic evaluation of these methods with consideration of variation of geopolitical unit and anisotropy of spread due to spatial heterogeneity and LDJD is also lacking.

Therefore, in this research, I aimed to evaluate the performances of common methods to estimate spread dynamics and overall rate using geopolitical-unit record with consideration of real world irregularity and stochasticity. Specifically, the evaluations were conducted from following aspects: 1) the accuracy of commonly used methods on estimating spread dynamics and overall rate with irregularity and stochasticity in spread, 2) the impact of size of geopolitical-unit on each method, and 3) the intrinsic similarities of all methods on estimating spread rate and dynamics. Simulated spread data were used as spread rate and dynamics can be accurately derived from the simulations. A new boundary displacement method and a regression method were proposed for estimating spread of invasive species with geopolitical-unit data.

The commonly used methods were first reviewed with a particular focus on using geopolitical-unit invasion data in Section 4.3. Section 4.4 included details of eight methods used in this research, methods of simulating spread data, and evaluation criteria. Section 4.5 and 4.6 described the results and discussions of this research, respectively.

4.3 Overview of Commonly Used Methods to Estimate Spread

The most commonly used methods to estimate spread are regression and boundary displacement methods, which can be used with all types of invasion record (summarized in Gilbert & Liebhold, 2010; Tobin et al., 2015). The general idea of the regression method is to regress the spread measurement against the time when the infestation is first observed. Boundary displacement methods estimate spread rates as the distance between consecutive infestation outlines of different periods. The minimum spread distance method also had been used (Aikio et al., 2010; Horvitz et al., 2017; Suarez et al., 2001). Here the details and use of these methods were summarized for estimating spreads with geopolitical-unit data.

4.3.1 Regression Method

Spread Distance With geopolitical-unit records, the methods of calculating spread distance are multiple. One way is to derive the minimum distance between the spread origin and the polygon of each geopolitical unit (Tobin et al., 2007 & 2015), and then regress the mean distance between spread origin and all geopolitical units infested in a same period against the invasion time. Another method is to use the distance between the spread origin and the centroid of each infested geopolitical unit (Evans & Gregoire, 2007; Tobin et al., 2007).

Square Root Area Square root area method is often used with population information of invasive species (Gilbert & Liebhold, 2010), but can also be used with geopolitical-unit records. The square root of cumulative area of all infested geopolitical units for each period is used as measurement (Tobin et al., 2015). This method, first suggested by Skellam (1951), assumes an invasive species is spreading by approximately concentric circles, for which the total spread distance (D) can be estimated as $= \sqrt{A}/\sqrt{\pi}$, where A represents the cumulative area of spread regions (Shigesada & Kawasaki, 1997).

Number of Infested Geopolitical-Unit Directly regressing the cumulative number of infested geopolitical units on invasion times had been commonly used (e.g., Perrins et al., 1993; Pyšek et al., 2008; Suarez et al., 2001). However, I argue that the square root of the cumulative number of infested units should be used instead (Williamson et al., 2005). If it is assumed that the total number of infested geopolitical units is n and the mean county size is $\bar{A}$, then the total infested area is $A = n\bar{A}$. As presented above, the spread distance would be $D = \sqrt{A}/\sqrt{\pi}$, by replacing A with $n\bar{\mathrm{A}}$, the spread distance can be estimated as $D = \sqrt{n\bar{\mathrm{A}}}/\sqrt{\pi}$. Thus, the total spread distance is linearly associated with $\sqrt{n}$, and the spread rate estimated from $\sqrt{n}$ would be linearly correlated with the ones by spread distance regression and square root area regression. Additionally, I proposed that $\sqrt{n\bar{\mathrm{A}}}$ can be an alternative to both the square root area regression method and square root number of infested geopolitical unit method.

4.3.2 Boundary Displacement Method

For geopolitical-unit data, the outer boundary of geopolitical units infested within the same period is commonly used as the infestation outline (e.g., Liebhold et al., 1992; Tobin et al., 2015). Slight changes on the infestation boundaries are common to avoid folds, islands, and gaps on the boundary (Sharov et al., 1999). I proposed a new method of deriving the infestation boundary by using polylines connecting all the outer centroids of newly infested geopolitical units as the infested boundary. This new method avoids the needs of changing infestation boundaries, and the infestation boundaries can be delineated automatically using programs like R.

4.3.3 Minimum Spread Distance Method

This method takes the minimum distance between a newly infested geopolitical unit and all units infested in earlier periods as the distance that a species has to spread to invade the new unit (Aikio et al., 2010; Horvitz et al., 2017). The mean of all minimum distances of geopolitical units infested in the same period is taken as the spread rate in that period. Similar with boundary displacement methods, MinD also directly estimates temporal spread dynamics.

4.4 Methodology

4.4.1 Spatial Area of Simulated Spread

To represent the real-world variety of geopolitical units for the generality of the simulated research, the counties in the U.S. were used for the simulated study (**Figure 4.1 A**). The county distribution in the U.S. is representative of many countries where size of geopolitical unit varies largely among different regions. To evaluate the performance of common methods to estimate spread with different geopolitical-unit sizes, the simulation research was conducted individually in three regions, Region 1 (R1), Region 2 (R2), and Region 3 (R3) (**Figure 4.1 A**). R1, R2, and R3 represent three different types of regions: 1) region with counties defined by small areas, 2) region with counties defined by small and large areas, and 3) region with counties defined by large areas. The county size and its variance in these three regions were listed in **Table 4.1**. In each region, an origin was selected, from which the spread was simulated in all directions, to make sure each spread direction has available ground area to conduct the simulations.

4.4.2 Simulation of Three Expansion Types and Three Spread Scenarios

To evaluate the ability of common methods to estimate expansion pattern of invasive species, three expansion types summarized by Shigesada et al. (1995) were simulated: 1) linear expansion, 2) biphasic expansion resulting from two linear-expansion phases, and 3) logistic growth function expansion. These three expansion types were commonly seen patterns of invading organism in research (e.g. Lantschner et al., 2014; Mineur et al., 2010; Perrins et al., 1993). Additionally, three spread scenarios were simulated to compare performances of all methods under different scenarios. A symmetric spread (S1) was first simulated to evaluate accuracies of all methods under this ideal scenario (**Figure 4.1 A-B**). To evaluate capability of all methods on dealing with anisotropy and stochasticity in spread, an asymmetric spread (S2) caused by spatial heterogeneity (**Figure 4.1 C-D**) and a LDJD (S3) caused by stochasticity (**Figure 4.1 E-F**) were simulated. The three expansion types were simulated for all spread scenarios in three regions.

Table 4.1 Statistics of county size in each region for different types of expansion

Spread scenarios	Expansion type	No. of infested county			Mean area (km^2)			Coefficient of variation		
		R1	R2	R3	R1	R2	R3	R1	R2	R3
S1	Type 1	637	240	103	1188.21	3173.15	7564.29	0.12	0.07	0.29
	Type 2	919	371	172	1235.23	3133.17	7243.13	0.15	0.07	0.31
	Type 3	930	377	172	1236.82	3175.75	7229.18	0.16	0.10	0.32
S2	Type 1	516	198	82	1125.04	3133.54	8472.74	0.02	0.09	0.29
	Type 2	760	209	95	1176.03	3231.86	8062.71	0.06	0.07	0.33
	Type 3	758	209	95	1175.65	3231.86	8062.71	0.07	0.13	0.26
S3	Type 1	914	381	146	1271.61	3273.84	8858.33	0.12	0.04	0.35
	Type 2	1147	471	200	1290.96	3275.94	8169.05	0.13	0.04	0.35
	Type 3	1158	479	200	1288.32	3274.22	7881.00	0.14	0.08	0.34

S1, S2, S3, and No. are short for symmetric spread, asymmetric spread, long distance jump dispersal, and number, respectively. R1, R2, and R3 represent small, small and large, and large sized counties.

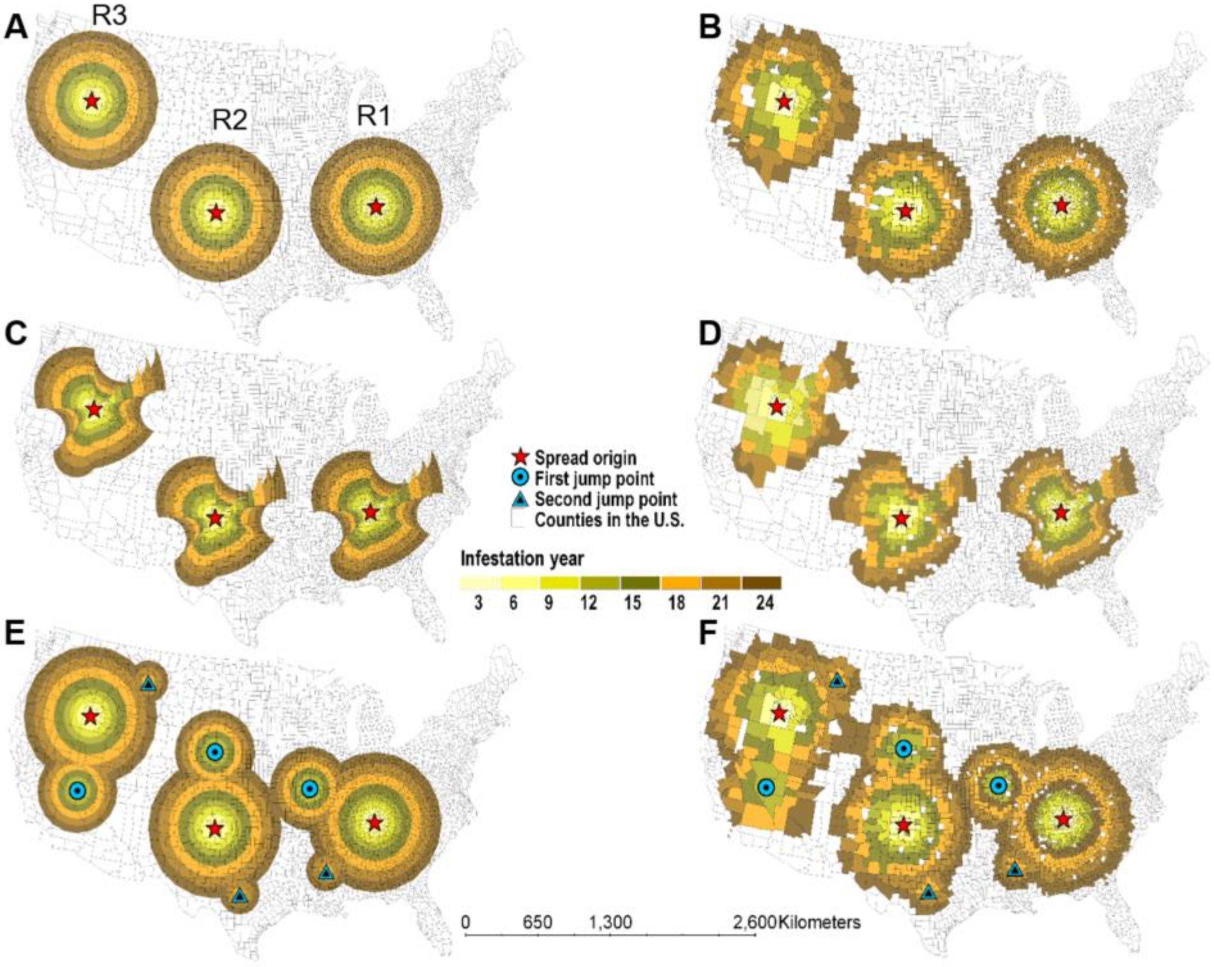

Figure 4.1 Examples of simulated invasion dynamics and conversion to county-level data using Type 2 expansion pattern (biphasic expansion). **A**, **C**, and **E** Simulated biphasic expansion during year 3-24 for scenario 1, 2, and 3, respectively; **B**, **D**, and **F** county-level spread record converted from **A**, **C**, and **E**, respectively. The distances between the spread origin and first and second jump points in **E** and **F** are set the same in all three regions.

Simulation of Symmetric Spread For the simulation of S1, the Type 1 expansion has a constant spread rate for all periods. The spread rate for Type 1 expansion was set to 20 km/year, as this is about the mean spread rate of invasive species based on multiple research (e.g., Horvitz et al., 2017; Suarez et al., 2001; Tobin et al., 2009). For Type 2 expansion the simulated rate was set to 20km/year for the first 12 invasion years and 30km/year for the following 12 years, thus the mean spread rate is 25km/year for the whole period. The Type 3 expansion follows a logistic growth function:

$$y = 826/(1 + e^{-0.43*(x-10)}).$$

Simulation of Asymmetric Spread Simulation of the three expansion types for S2 is similar with that for S1, except that the spread rates varied among different directions (**Figure 4.1 C-D**). The mean simulated rate at all directions varied between 10-24 km/year for Type 1 expansion, and 12-31 km/year for Type 2 and Type 3 expansions.

Simulation of Long-Distance Jump Dispersal To simulate the S3, two jump dispersal events were added for the symmetric spread (S1) with one occurring in year 9 and another occurring in year 18 (**Figure 4.1 E-F**). To make the S3 in three regions comparable, the distances among the two jump points and the spread origin are set the same for all regions (**Figure 4.1 E-F**). The jump point would become a new spread origin, from which further spreads will occur in all directions at some rates, and this rate was set to 20km/year for all jump points and expansion types for clarity and simplicity.

Estimation of Simulated Rates Unlike S1, the spread rates for S2 and S3 can not be directly derived from an expansion algorithm. Two methods, square root area and spread distance (Dis) methods, were used to calculate the simulated overall rate and spread dynamics. Calculation of spreads by Dis method is similar with boundary displacement method, except that instead of using mean distances between two consecutive boundaries, the Dis method calculates the rate dynamics by subtracting the mean distances between two consecutive boundaries and the spread origin. These two methods were chosen as they can directly be used with simulated data without converting to county-level data, and also both methods use objective measures without the need of any manual changes.

Converting Simulated Spread to Geopolitical-Unit Data For all simulations, the spread was measured once every three years. To convert the simulated spread (shown in **Figure 4.1 A, C, E**) to geopolitical-unit spread record (shown in **Figure 4.1 B, D, F**), counties that were infested in the same periods (i.e., every three years) were first selected. To eliminate margin effect, a county is only defined as first infested when more than 10% area of that county get infested. To reflect stochasticity occurred often in real situations, 5% of the counties first infested in each period were randomly set as non-infested counties.

4.4.3 Methods to Estimate Overall Rate and Spread Dynamics

The technique flow of the simulation research is summarized in **Figure 4.2.** Four regression methods, two boundary displacement methods and a minimum spread distance method (MinD) were used. Details of all methods were listed in **Table 4.2**. For CtyBD and CtdBD methods, distances between two consecutive boundaries were measured as the mean length of all transects radiating from the spread origin at 2°

(Liebhold et al., 1992; Liang et al., 2019). For the MinD method, we first calculates the minimum distance between the centroid of a newly infested county and centroids of all counties infested in earlier periods, and then take the mean of the minimum distances of all newly infested counties as the spread rate in the new period.

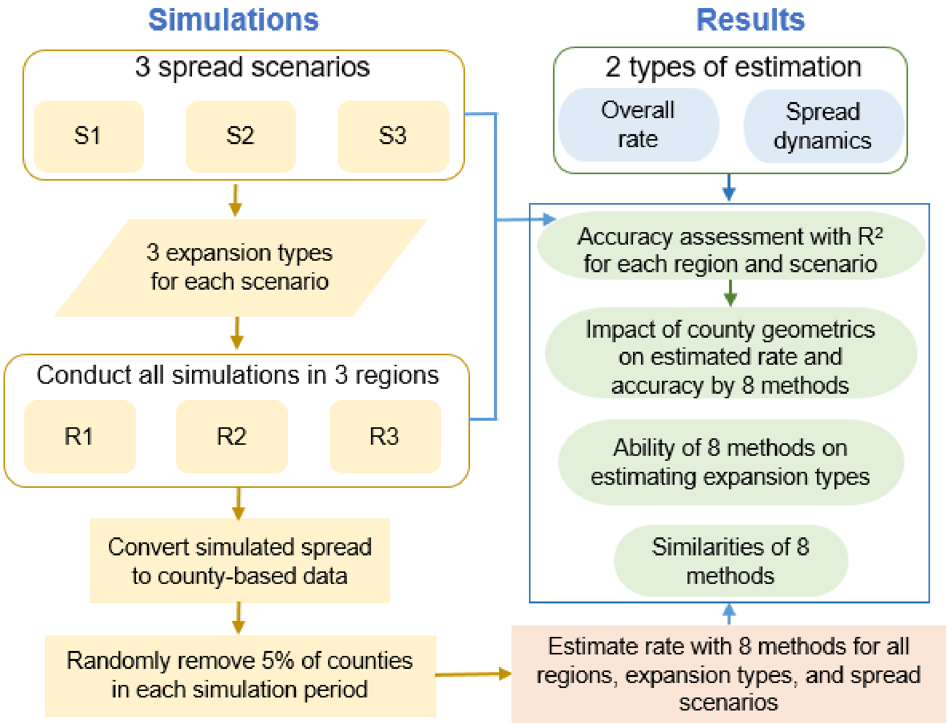

Figure 4.2 Technique flow of the simulation study. CV is short for coefficient of variation of county size. S1, S2, and S3 represent symmetric spread, asymmetric spread, and long distance jump dispersal, respectively.

]

Table 4.2 Full and abbreviation names of all methods and the measurements used by all methods to estimate spread dynamics

Method	Name of method		Measurement
	Full name	Abbreviation	
Regression method	Centroid distance between spread origin	CtdDis	Mean distance between county centroids and spread origin
	Minimum distance between county and origin	MinD Reg	Mean of the minimum distance between counties and spread origin
	Square root of infested area	Area	$\sqrt{A/\pi}$
	Number of infested county	Ncty	$\sqrt{n}$
	Square root area estimated from number of infested county	NctyArea	Ncty*$\sqrt{\bar{A}/\pi}$
Boundary displacement	Centroid boundary	CtdBD	Mean distance between two consecutive boundaries
	County boundary	CtyBD	
Minimum spread distance		MinD	Mean of the minimum distance between centroids of newly and earlier infested counties

$A, n, and\ \bar{A}$ represent the cumulative infested area, all number and mean area of infested counties, respectively.

CtdBD, CtyBD, and MinD methods directly estimate spread dynamics in different periods, and the overall rate was calculated as the mean rate overall all periods. For regression method, spread dynamics were estimated as the difference of measurements between two consecutive periods. For Type 1 expansion, the overall rate was estimated as the slope of a linear regression for regression methods, whereas for Type 2 expansion, the overall rate was estimated as the mean of two slopes of a segmented linear regression with break point at year 12. Finally, for Type 3 expansion pattern, instead of estimating spread rate using derivative of logistic growth function, the below function was used to make the derived overall rate more representative of the practice.

$$Overall\ rate = \frac{cumulative\ spread\ distance}{total\ spread\ time}$$

4.4.4 Evaluation Statistics

Ability of All Methods to Estimate Expansion Patterns

For CtdBD, CtyBD, and MinD methods, the cumulative values of spread measurement for each period were derived. To test whether all methods can accurately estimate the right expansion patterns, three regression models were fitted, i.e., linear, biphasic linear, and non-linear with logistic function, to the estimated spread measure or cumulative measure for non-regression methods for each expansion type, region, and spread scenario. The AICs of three regression models were derived and the model with lowest AIC was assigned as the estimated expansion pattern by the method.

Accuracy and Similarity of All Methods

To evaluate the accuracy on overall rate and spread dynamics estimated by each method, the R^2 and the root mean square error (RMSE) between the estimated and simulated spread were derived by each region and spread scenario. R^2, which varies from 0 to 1, is a scale-independent criteria and can measure ability of all method on estimating the real spread patterns, with 1 indicating perfect estimation. The RMSE measures the absolute deviation of estimated spread from the simulated spread, but can not be used on Ncty method due to its scale-sensitivity.

To assess the similarity of overall rate and spread dynamics estimated by all methods, the Pearson correlation coefficient (r) was calculated. Hierarchical clustering (HC) with complete linkage was used to group all methods based on their similarity of estimating overall rate and spread dynamics by different regions and spread scenarios.

Impact of County Size on Estimation Accuracy

To determine whether the county size and its variation affect the values of estimated overall rate and spread dynamics by each method, the significance of correlations between the estimated spread with the mean and coefficient of variation (CV) of county size was tested. To assess the impact of county size and its variation on accuracies of estimated overall rate and spread dynamics, the significance of correlations between the mean and CV of county size and R^2 of the estimation was tested for each region and spread scenario.

4.5 Results

4.5.1 Ability of All Methods on Estimating Expansion Patterns

Not only regression methods but also boundary displacement methods can estimate the expansion patterns (**Figure 4.3**, for clarity only Type 1 and 3 patterns in R1 and R3 were shown). CtdBD, Ncty and NctyArea methods correctly estimated the expansion patterns for all scenarios and regions. For S1, except MinD, all other methods correctly estimated the right expansion patterns in all regions. For S2 and S3, Area and CtyBD correctly estimated all expansion patterns for R1 and R2. MinD Reg, and MinD methods often misclassified the expansion patterns for S2 and S3, and CtdDis often misclassified the expansion patterns for S3.

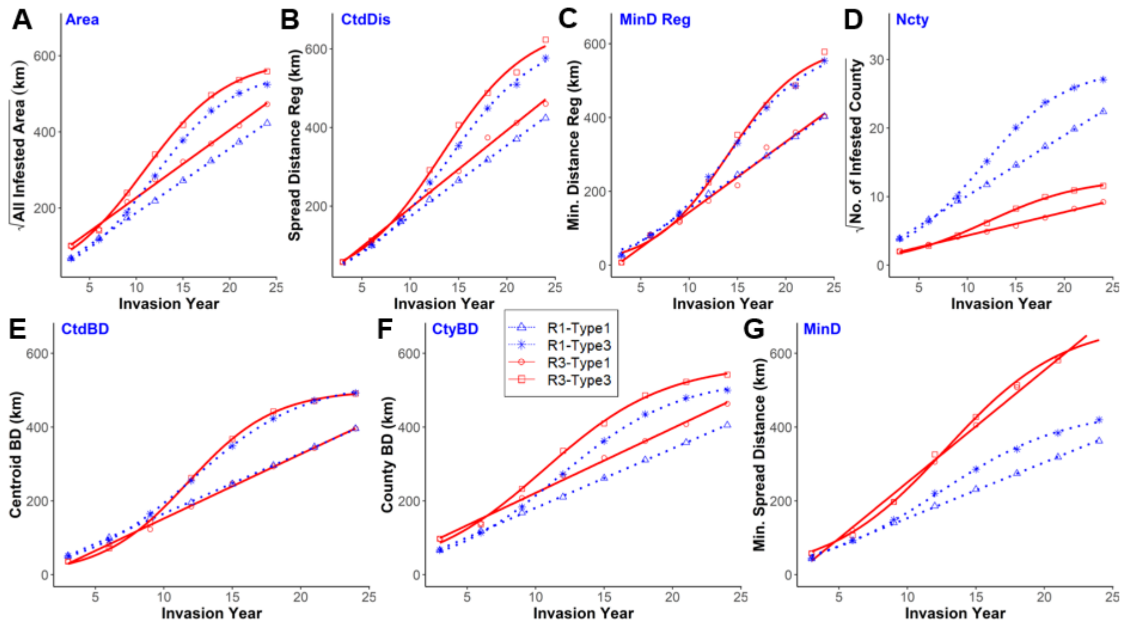

Figure 4.3 A-G Estimated spread patterns for asymmetric spread scenario in Region 1 (R1) and Region 3 (R3) by four regression methods (**A-D**), two boundary displacement methods (**E-F**), and minimum spread

distance method (**G**). For clarity, only type 1 and type 3 expansion patterns are shown in the plots. NctyArea method estimated exactly the same pattern with Ncty method, thus is not shown.

4.5.2 Accuracy of All Methods

For both scenario 2 and 3, the spread becomes asymmetric (**Figure 4.1**). Consequently, the simulated spread rates measured by spread distance and square root area differ with each other. This difference of simulated rate suggests that the spread rate of invasive species can be multiple based on the estimation methods, whereas different methods capturing different aspect of the spreads. Therefore, estimating spread dynamics is more informative and meaningful than estimating an overall spread rate. The spread rate estimated from spread distance method was used hereafter.

Accuracy on Estimating Overall Spread Rate

For all scenarios, the MinD method consistently estimated higher spread rate in regions with larger county size leading to underestimation of rate in R1 but overestimation in R3. For S1, all other methods estimated similar spread rates with simulated rates (**Table 4.3**). However, the MinD Reg method estimated a significantly lower spread rate (*P*=0.003), whereas the CtyBD estimated a higher spread rate than the simulations (*P*=0.002). For S2 and S3, all regression methods tended to estimate significantly higher spread rates than the simulations (**Table 4.3**), whereas the boundary displacement methods estimated significantly higher spread rate when LDJD occurred.

Table 4.3 Simulated and estimated overall spread rates and the paired T-test between simulated and estimated rates for each spread scenario

Spread scenario	Type	Simulated rate (km/y)	Region	CtdDis	Area	MinD Reg	Ncty	NctyArea	CtdBD	CtyBD	MinD
S1	I	20.00	R1	19.81	19.45	19.53	1.02	20.23	20.13	20.11	15.91
			R2	19.86	19.91	19.61	0.61	19.26	19.93	20.55	20.54
			R3	20.20	19.80	19.82	0.38	18.25	20.12	21.57	27.99
	II	25.00	R1	24.90	24.34	24.38	1.16	22.56	25.21	25.87	18.46
			R2	24.90	24.71	24.48	0.78	24.79	24.97	25.43	22.55
			R3	24.95	24.25	24.60	0.52	25.52	25.16	26.28	29.07
	III	25.20	R1	25.15	25.17	24.01	1.27	25.20	25.06	25.23	19.07
			R2	25.24	25.70	23.70	0.81	25.75	25.12	25.73	23.15
			R3	25.60	26.42	23.48	0.55	26.38	25.38	26.56	28.78
		P		0.987	0.661	0.003	NA	0.468	0.260	0.002	0.742
S2	I	16.68	R1	18.00	16.89	17.84	0.88	16.70	16.42	16.13	15.15
		16.59	R2	19.03	17.41	18.82	0.54	17.01	16.85	17.18	20.19
		16.59	R3	19.72	17.75	19.09	0.34	17.73	17.16	17.42	29.09
	II	21.44	R1	23.54	21.14	23.09	1.06	20.49	21.08	20.81	17.56
		21.33	R2	24.47	22.61	24.41	0.70	22.50	21.90	22.23	22.21
		21.32	R3	23.95	21.37	23.57	0.45	22.64	20.36	21.14	29.34

Table 4.3 Continued

Spread scenario	Type	Simulated rate (km/y)	Region	CtdDis	Area	MinD Reg	Ncty	NctyArea	CtdBD	CtyBD	MinD
	III	21.42	R1	24.04	21.86	23.08	1.13	21.83	21.04	20.65	17.88
		21.31	R2	24.63	22.92	23.11	0.73	23.50	22.05	21.87	22.82
		21.30	R3	25.97	23.30	24.09	0.48	24.54	21.67	21.17	28.22
		P		0.000	0.005	0.000	NA	0.016	0.748	0.744	0.175
S3	I	24.66	R1	26.31	25.99	25.94	1.28	25.72	25.66	25.45	20.79
			R2	26.13	26.71	25.90	0.81	26.16	26.57	26.84	26.03
			R3	25.13	26.53	24.70	0.48	25.44	25.40	25.79	33.51
	II	28.07	R1	31.32	28.32	30.73	1.33	26.88	29.37	30.05	22.32
			R2	32.76	29.06	32.46	0.89	28.77	29.71	30.32	26.96
			R3	29.12	29.80	28.32	0.58	29.42	28.70	30.18	33.61
	III	28.21	R1	29.70	28.71	28.53	1.42	28.71	27.88	28.51	21.76
			R2	30.04	29.44	28.48	0.91	29.44	28.77	30.19	25.76
			R3	28.36	29.51	25.98	0.59	29.51	28.87	30.54	31.56
		P		*0.003*	*0.000*	*0.087*	*NA*	*0.009*	*0.002*	*0.000*	*0.976*

S1, S2, and S3 represent symmetric spread, asymmetric spread, and long distance jump dispersal, respectively. *P*: *P* value of paired T-test between estimated and simulated rates by each scenario (one-tale test for *P*<0.10, two-tale test for *P*>=0.10)

Based on R^2, all methods had best estimates for symmetric spread and least estimates for LDJD (**Table 4.4**). The MinD and Ncty methods had low ability to estimate spread when spread in all three regions was considered. For S1 and S2, the CtdBD, CtdDis and MinD Reg were the best three methods. However, the distance-based regression methods had low ability to estimate spread rate for S3, whereas CtyBD, CtdBD, and Area were the best three methods for this scenario. Except Ncty and NctyArea, all other methods had best estimates for R1 (**Table 4.4**). Despite the large size and variations of county in R3, all methods had higher R^2 than that for S3, suggesting the LDJD challenges the estimation of spreads more than irregularities caused by county sizes. CtdBD method had the best estimate, followed by CtyBD, Area, and NctyArea.

Accuracy on Estimating Spread Dynamics

Compared to the overall rate, the ability to estimate spread dynamics decreased for all methods (**Table 4.5**). Except Ncty, NctyArea, and MinD methods, all other methods had highest performances for S1 and R1 and lowest performances for S3 and for R3, suggesting the increasing challenge of estimating spread dynamics with the increase of stochasticity and irregularity in spread and county size. The MinD method had low performances for all scenarios and regions. CtdDis and MinD Reg methods only had good estimation for S1 (**Table 4.5**). CtdBD, NctyArea, Area, and CtyBD were the best four methods for all scenarios and regions based on both R^2 and RMSE (**Table 4.5**), whereas only CtdBD showed consistent satisfying estimates for all regions and scenarios ($R^2 > 0.75$).

Table 4.4 R^2 of estimated overall spread rate by each scenario and region, and R^2 and root mean square error (RMSE) of estimated spread rates in all regions for all scenarios

Spread		CtdDis	Area	MinD Reg	Ncty	NctyArea	CtdBD	CtyBD	MinD	*P*
R^2 by scenario	S1	0.99	0.95	0.96	0.09	0.87	1.00	0.95	0.04	*NA*
	S2	0.93	0.92	0.96	0.09	0.86	0.94	0.93	0.03	*0.007*
	S3	0.72	0.89	0.47	0.02	0.79	0.85	0.89	0.00	*0.021*
R^2 by region	R1	0.91	0.98	0.89	0.93	0.92	0.98	0.98	0.89	*NA*
	R2	0.84	0.96	0.80	0.96	0.95	0.97	0.98	0.76	*0.033*
	R3	0.82	0.95	0.81	0.92	0.91	0.98	0.97	0.41	*0.035*
All overall rates	R^2	0.85	0.95	0.81	0.14	0.91	0.98	0.97	0.13	
	RMSE	2.13	1.06	1.78	NA	1.25	0.71	1.22	5.19	*NA*
Tukey's range test		a	a	ab	bc	a	a	a	c	

S1, S2, and S3 represent symmetric spread, asymmetric spread, and long distance jump dispersal, respectively. *P: P* value of one-tale paired T-test of R^2 between S1/R1 and other scenarios/ regions.

4.5.3 Impact of County Size and its Variation on Estimation of Spread

In regions with larger county size, the MinD method always estimated higher spread rates whereas the Ncty method always had lower values (**Table 4.3**). Consequently, strong correlations existed between the county size and the estimated spread for these two methods (**Table 4.6**). Additionally, the larger county size also lead to higher estimated spread dynamics for Area, CtdBD, and CtyBD methods. For the accuracy of estimated overall rate, significant correlation of R^2 with mean and CV was only observed on the Area method (**Table 4.6**). However, for the estimated spread dynamics, significantly negative correlations were observed on Area, CtdBD, and CtyBD methods (**Table 4.6**), indicating the accuracy of these methods are negatively impacted by the county size and its variations. Whereas for Ncty and NctyArea methods, variations in mean county size is more influential than the size of county on the estimation accuracy.

Table 4.5 R^2 of estimated spread dynamics by each scenario and region, and R^2 and root mean square error (RMSE) of estimated spread dynamics for all regions and scenarios

Spread		Regression Method					Boundary Displacement			*P*
		CtdDis	Area	MinD Reg	Ncty	NctyArea	CtdBD	CtyBD	MinD	
R^2 by scenario	S1	0.83	0.84	0.81	0.36	0.84	0.90	0.83	0.24	*NA*
	S2	0.60	0.79	0.45	0.35	0.89	0.80	0.77	0.22	*0.036*
	S3	0.05	0.57	0.06	0.24	0.63	0.77	0.58	0.07	*0.004*
R^2 by region	R1	0.37	0.96	0.39	0.46	0.89	0.97	0.96	0.27	*NA*
	R2	0.31	0.87	0.32	0.48	0.92	0.90	0.88	0.24	*0.013*
	R3	0.31	0.57	0.31	0.38	0.65	0.77	0.55	0.30	*0.009*
All spread dynamics	R^2	0.33	0.76	0.34	0.24	0.80	0.86	0.76	0.19	
	RMSE	7.44	3.86	7.25	NA	3.50	3.30	4.13	8.75	*NA*
Tukey's range test		b	a	b	b	a	a	a	b	

S1, S2, and S3 represent symmetric spread, asymmetric spread, and long distance jump dispersal, respectively. *P: P* value of one-tale paired T-test of R^2 between S1/R1 and other scenarios/ regions.

Table 4.6 Correlation coefficient between mean of county size and estimated overall rate and spread dynamics, and between mean/coefficient of variation of county size and R^2 for each region, expansion type, and spread scenario

Correlation of A and B		Regression Method					Boundary Displacement		MinD
A	B	CtdDis	Area	MinD	Ncty	NctyArea	CtdBD	CtyBD	
Mean	Overall rate	0.00	0.01	-0.04	-0.85***	0.13	0.00	0.07	0.91***
	Dynamics	0.08	0.15*	-0.02	-0.55***	-0.04	0.15*	0.15*	0.67***
Mean	R^2 of overall rate	-0.06	-0.68*	-0.23	0.51	0.51	0.18	0.42	-0.58
CV		0.07	**-0.62**	-0.19	0.18	0.18	0.13	0.14	-0.38
Mean	R^2 of dynamics	-0.08	-0.85***	-0.22	-0.32	-0.32	-0.69**	-0.86***	-0.38
CV		-0.05	-0.81***	-0.16	-0.53*	-0.53*	-0.62**	-0.77***	-0.34

*, **, and ***Significant at level α=0.05, 0.01, and 0.001, respectively. Bold: Significant at level α=0.10.

4.5.4 Similarity of All Methods

Similarity of Overall Rate For the overall rate, all methods estimated highly correlated overall rate in R1 (r>0.90, **Figure 4.4 A**) and R2 (r>=0.85, **Figure 4.4 B**), whereas the large county size in R3 only dramatically decrease the similarities of MinD with the remaining methods (r<= 0.80, **Figure 4.4 C**). For both S1 and S2, except Ncty and MinD methods, all other methods estimated highly correlated overall rate among each other and with the simulation (r>=0.90, **Figure 4.4 E, F**), confirming comparably good performance of these methods. For S3 similarities of CtdDis and MinD Reg with other methods also decreased, however high positive correlations were still observed among Area, NctyArea, CtdBD, and CtyBD methods (**Figure 4.5 G**).

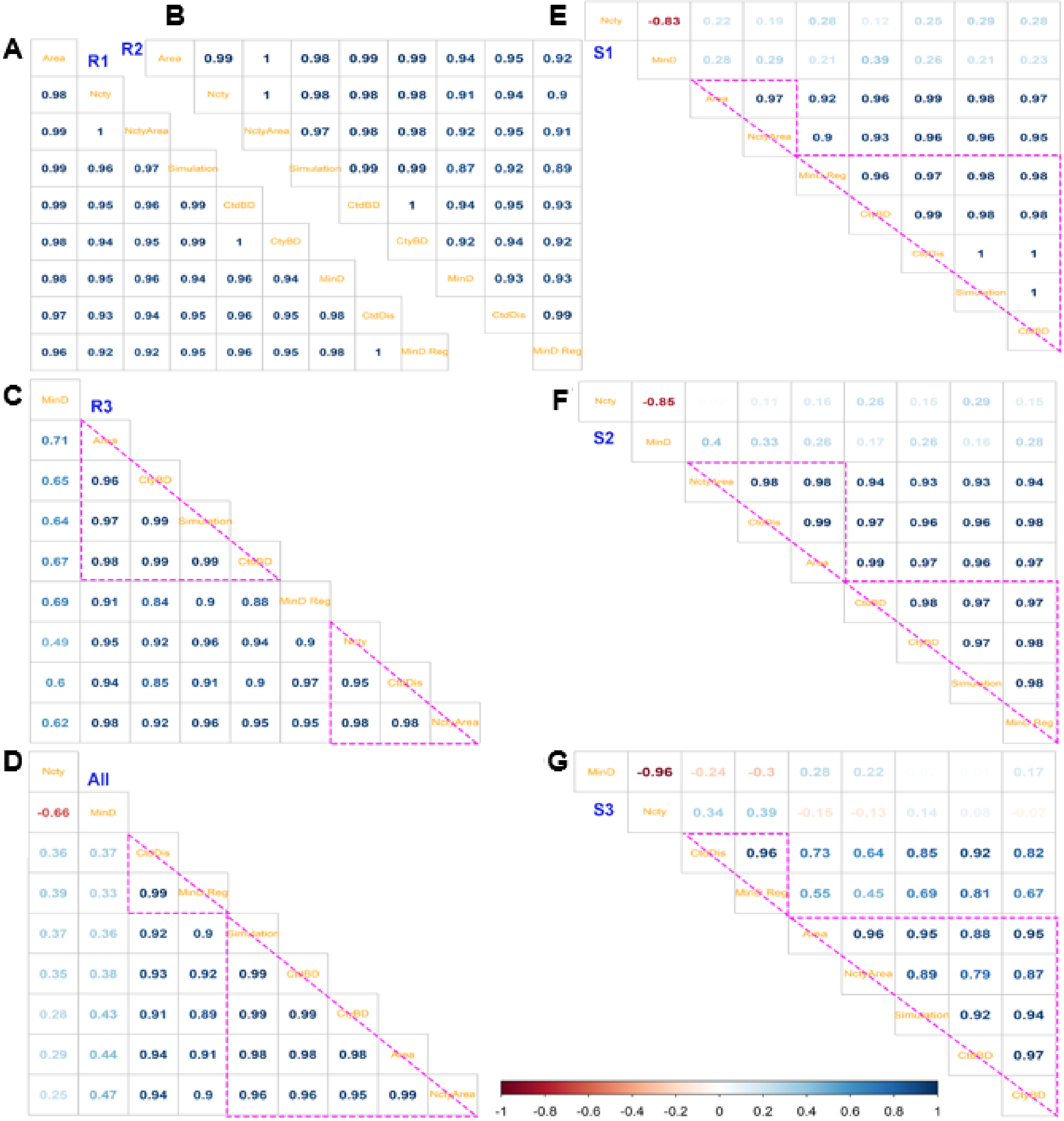

Figure 4.4 Correlation matrix (CM) and similarity pattern (SP) of estimated overall spread rate by all methods. **A-D** CM and SP in region 1 (R1), region 2 (R2), region 3 (R3), and all regions for three spread scenarios, respectively; **E-G** CM and SP of symmetric spread (S1), asymmetric spread (S2), and long distance jump dispersal (S3) for all regions, respectively. Methods that are enclosed in the same triangle are classified in the same group by hierarchical clustering based on similarity of estimated rate.

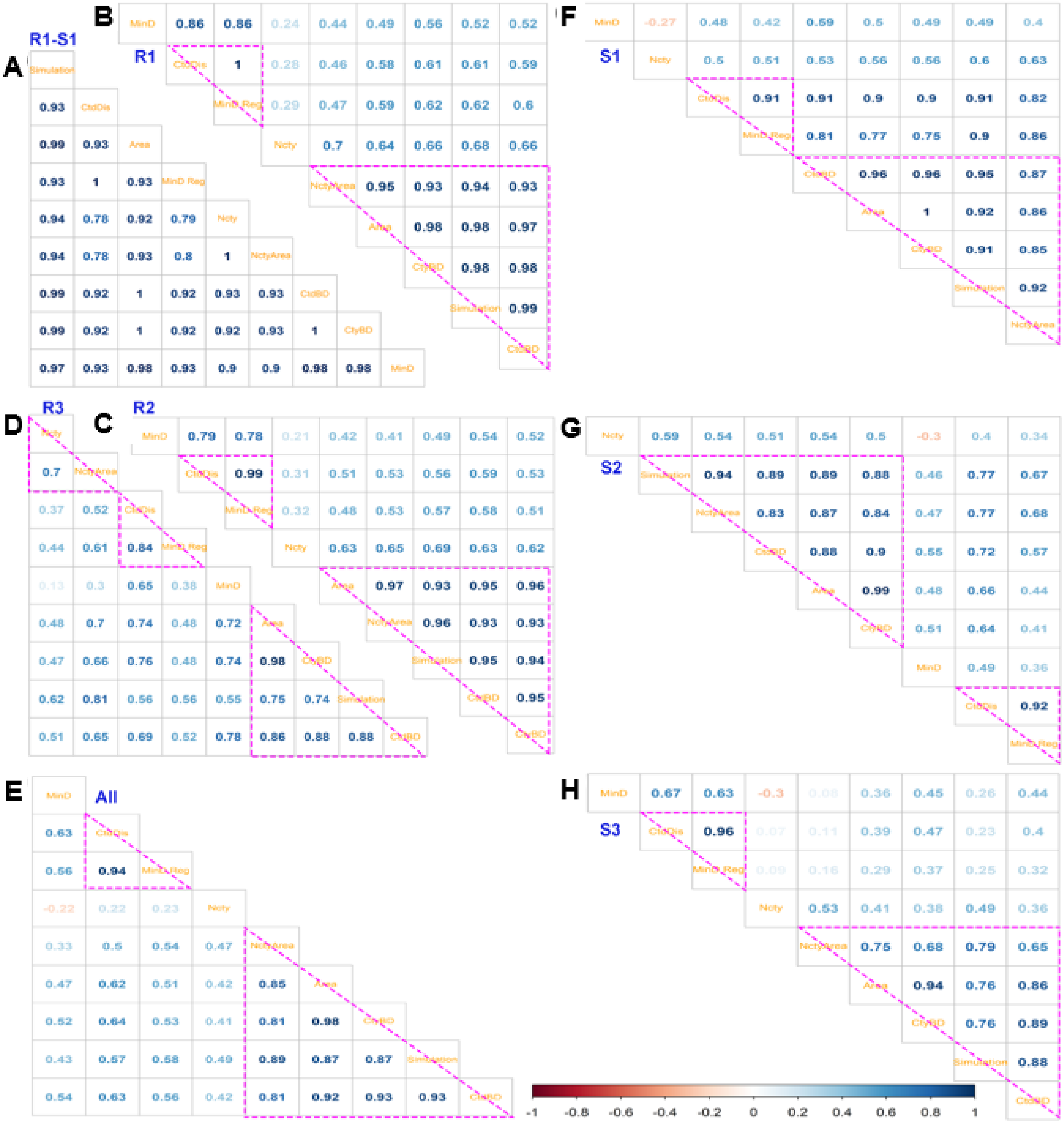

Figure 4.5 Correlation matrix (CM) and similarity pattern (SP) for estimated spread dynamics by all methods. **A** CM in region 1 (R1) for symmetric spread (S1)**; B-E** CM and SP in R1, region 2 (R2), region 3 (R3), and all regions for three spread scenarios, respectively; **F-G** CM and SP of S1, asymmetric spread (S2), and long distance jump dispersal (S3) for all regions, respectively. Methods that are enclosed in the same triangle are classified in the same group by hierarchical clustering based on similarity of estimated rate.

Similarity of Spread Dynamics Compared to the overall rate, the spread dynamics are more sensitive to the irregularities and stochastic events. Consequently, high correlations among all methods was only observed for S1 in R1 (**Figure 4.5 A**, r>0.70). With the increase of anisotropy and stochasticity in spread in R1, the similarities of the CtdDis, MinD Reg, MinD, and Ncty methods decreased with other methods (**Figure 4.5 B**). Nevertheless, estimated dynamics by NctyArea, Area, CtyBD, and CtdBD methods still had high correlations among each other and with the simulated dynamics (r>0.90, **Figure 4.5 B-C**). The correlation of all methods in R2 showed the same pattern with that in R1 (**Figure 4.5 C**). Similar with the overall rate, the similarity of all methods was further weakened in R3, and high correlations was only observed among Area, CtyBD, and CtdBD methods (r>0.80, **Figure 4.5 H**).

Overall Similarity The CtdDis and MinD Reg methods always estimated highly correlated overall rate and spread dynamics (r>0.95, **Figures 4.4 & 4.5**), whereas MinD and Ncty methods always estimated low similarity with other methods when spreads in all regions were analyzed together. For the similarity patterns on all estimations, except Ncty and MinD, all other methods estimated highly correlated overall rate among each other and with the simulations (**Figure 4.4 D**), whereas Area, NctyArea, CtDBD, and CtyBD methods estimated highly correlated spread dynamics among each other and with the simulations (**Figure 4.5 E**). These strong correlations indicated that the estimated spread patterns were essentially similar. Based on similarities of estimated overall rate and spread dynamics, Area, CtDBD, CtyBD and NctyArea methods were constantly classified into one group (**Figures 4.4 & 4.5**).

4.6 Discussion

Estimating spread is critical for prediction and early management of invasive species and understanding important factors affecting the spread (Liang et al., 2019; Paini et al., 2016; Stohlgren & Schnase, 2006). For non-native species, invasion histories recorded at geopolitical unit are often more available than fine resolution records. Researchers worldwide had used geopolitical-unit level data to estimate spread of invasive species without a full evaluation on ability of common methods on estimating spread with geopolitical-unit level data (e.g., Evans & Gregoire, 2007; Liang et al., 2019; Tobin et al., 2009). This research evaluated the ability and similarity of commonly used methods to estimate spread while considering the variation of geopolitical-unit size and heterogeneity and stochasticity in spread of invasive species across large spatial area. A new boundary displacement method (i.e., CtdBD) and regression method (i.e., NctyArea) was proposed to estimate invasions with geopolitical-unit level data. Findings of this research can guide selection of optimal methods for estimating invasions with both geopolitical-unit level data and other types of invasion record, based on irregularities in data and anisotropy and stochasticity in spread.

4.6.1 Ability of All Methods to Estimate Spread of Invasive Species

Regression methods had long been used to determine expansion patterns of invasive species (e.g., Liang et al., 2019; Liebhold et al., 1992; Mineur et al., 2010; Perrins et al., 1993). The cumulative value of boundary displacement methods can also be used to estimate the expansion patterns. The CtdBD always estimated the right expansion patterns as well Ncty and NctyArea method regardless of variations in county size and anisotropy and stochasticity in spread. Additionally, CtyBD and Area methods can also be used to estimate expansion patterns.

Compared to the overall spread rate, spread dynamics over different invasion periods can provide more complete knowledge on the spread of invasive species and better facilitate further analysis or management of invasive species (Hastings et al., 2010; Liang et al., 2019). However, the estimation of spread dynamics is much more challenging than the estimation of overall rate. Boundary displacement and minimum spread distance methods had been commonly used to estimate spread dynamics of invasive species (e.g. Horvitz et al., 2017; Sharov et al., 1999; Wang & Wang, 2006). Tobin et al. (2007) concluded that the boundary displacement method is better than regression method as the former method can describe dynamics of spread rate. However, this research suggests that regression methods can also accurately estimate spread dynamics by using the difference of measurements between two consecutive periods. CtdBD, CtyBD, Area, and NctyArea methods were the top four methods to estimate both overall rate and spread dynamics. Thus, for spread without clear infestation outline, Area and NctyArea methods can be used to estimate the spread dynamics. For spread with clear infestation outlines, CtdBD can be a top choice as it constantly had the best estimation.

Generally, the abilities of all methods to estimate spread rate decrease with the increase of anisotropy and stochasticity in spread. However, the distance-based regression methods (i.e., CtdDis and MinD Reg) are more sensitive to these irregularities and stochasticity than the area-based regression methods (i.e., Area and NctyArea) and boundary displacement methods, as they can be more easily skewed by unrepresentative long distances caused by stochasticity or heterogeneity. Consequently, CtdDis and MinD

Reg methods had low ability to estimate overall rate for LDJD, and can only estimate spread dynamics well without high asymmetry or LDJD in spread.

Researchers had used MinD method to estimate spread dynamics (e.g., Aikio et al., 2010; Horvitz et al., 2017), and Ncty method to estimate the overall rate and expansion patterns of non-native species (Perrins et al., 1993; Pyšek et al., 2008). However, when evaluating the accuracy of estimated overall rate across different regions, Ncty and MinD methods constantly showed low ability to estimate both overall rate and spread dynamics for all spread scenarios due to their sensitiveness to size of geopolitical unit. Additionally, MinD constantly showed lowest ability to estimate both overall rate and spread dynamics, and Ncty had low ability to estimate spread dynamics. Thus, MinD method is not recommended to estimate spread of invasive species, whereas Ncty is only suggested to estimate overall spread rate when the mean sizes of geopolitical unit across different periods or regions are relatively uniform. NctyArea can be a good alternative to the Ncty method as it rectifies the Ncty method by the mean area of geopolitical unit.

4.6.2 Estimation of Spread with Asymmetric Spread and Long-Distance Jump Dispersal

Compared to LDJD, asymmetric spread caused by spatial heterogeneity does not seriously challenge the ability of all methods to estimate overall spread rate, meanwhile all boundary displacement methods and area-based regression methods can also show good estimation of spread dynamics under this scenario. However, when the spread is highly asymmetry and study area is large, such as regional or bigger scales, estimating spread of invasive species by taking the whole infested region as one area can not reveal the spatial dynamics of spread caused by heterogeneities (Andow et al., 1990; Liang et al., 2019). To better reveal localized spread dynamics for asymmetric spread, Andow et al. (1990) first proposed to divide the large infested area into multiple neighborhoods to increase homogeneity within each neighborhood. Fraser et al. (2015) divided the whole infested region into three segments based on the introduction origins and spatial heterogeneity, whereas Morin et al. (2009) divided the whole infested area into uniform size of sectors to reflect the spatial anisotropy in the spread of hemlock wooly adelgid. However, these methods do not classify neighborhoods from a quantitative perspective. Liang et al. (2019) proposed a quantitative method, spatial constrained clustering, to classify large heterogeneous region into environmentally homogeneous sub-regions. The use of neighborhood measurement is suggested for highly asymmetric spread, especially when assessing impact of spatial factors contributing to the asymmetrical expansions is of interest. Additionally, dividing a highly asymmetric spread into several relatively symmetric spread within each sub-region, the estimation accuracy can also be improved as suggested by this research.

Asymmetric spread caused by LDJD challenges the ability of all methods to estimate both overall rate and spread dynamics. LDJD of invasive species is often caused by random event (Nathan, 2006; Suarez et al., 2001). However, despite its rarity and stochasticity, LDJD events greatly facilitate the expansion of invasive species and can be more influential than local dispersal for some species (Kot et al., 1996; Mineur et al., 2010; Nathan, 2006; Shigesada et al., 1995; Suarez et al., 2001). For LDJD, estimation of spread rate could not tell the real dispersal ability of the species, but a rate impacted by stochastic events. LDJD events cause new spread origins and obscure the spread patterns (Mineur et al., 2010; Nathan, 2006). Thus, to better estimate local spread rates of invasive species, researchers could set multiple spread origins including the ones caused by LDJD, if the expansions originated from the LDJD point are easily recognized (Muirhead et al., 2006; Suarez et al., 2001). However, for the LDJD-point originated spread that are not easily recognized, the boundary displacement method, CtdBD and CtyBD, and area-based regression methods, Area and NctyArea, could be preferred as they are less sensitive to the LDJD than distance-based regression methods.

4.6.3 Similarity of All Methods to Estimate Spread of Invasive Species

Comparison of common methods to estimate spread of invasive species had been conducted by researchers (Gilbert & Liebhold, 2010; Tobin et al., 2007; Tobin et al., 2015), whereas assessment of the similarity of these methods is rare. A variety of measurements, such as spread distance or area and number of infested geopolitical unit, are used by different methods, thus the estimated spread rates may vary upon the measurements. However, regardless of the differences in estimated rates, the spread patterns, i.e. the spatial or temporal spread dynamics, revealed by different methods could be essentially similar (Liang et al., 2019; Weber, 1998). In some practice, such as determining important factors affecting spread of invasive species (Lantschner et al., 2014; Liang et al., 2019; Sharov et al., 1999), the accuracy of estimated spread patterns matter more than the values of spread rates.

When the spread is symmetric and the county size is relatively uniform, all eight methods estimate similar spread dynamics among each other and with real spread dynamics. However, with all the irregularities and stoachasticity, Area, CtdBD, CtyBD, and NctyArea methods estimated similar spread dynamics and can be classified into one group, whereas the Area, CtdBD, and CtyBD have higher similarities among each other than with NctyArea method. When the overall spread rate is the question of interest and no LDJD occurred, except Ncty and MinD methods, all other six methods estimated similar spread patterns among each other and with real spread dynamics, suggesting all these methods could have robust estimation of overall rate regardless of the variations in geopolitical unit. However, for LDJD only Area, CtdBD, CtdDis, and NctyArea methods estimated similar overall rates with the simulation. Unsurprisingly, the two distance-based regression methods, CtdDis and MinD Reg, always estimate highly correlated overall rates and spread dynamics. No significant difference of accuracy on estimated overall rate (P=0.49) and spread dynamics (P=0.48) was observed between the two methods. The two boundary displacement method always estimated highly correlated overall rate and spread dynamics ($r>0.85$), whereas CtdBD had higher accuracy to estimate spread dynamics than the CtyBD method (P=0.014).

4.7 Conclusions

Using simulated spread data, I found geopolitical-unit invasion record is capable of accurately estimating spread of invasive species. Both regression methods and boundary displacement methods can be used to estimate the overall invasion rate, expansion pattern, and spread dynamics of non-native species. Selection of an optimal method depends on the question of interest, anisotropy and stochasticity in spread, and variations and sizes of geopolitical unit. Among the eight methods considered, except MinD method, all other methods can be used for spread without LDJD when the question of interest is the overall rate and sizes of geopolitical unit are relatively uniform. However, if the question of interest includes estimating overall rate in several sub-regions, among which mean sizes of geopolitical units vary largely, both MinD and Ncty method could be avoided, whereas NctyArea method is a good alternative for Ncty method under this scenario. The spread distance-based regression methods (i.e., CtdDis and MinD Reg) are more sensitive to the irregularities and stochasticity in spread than the area-based regression methods and boundary displacement methods. For LDJD, boundary displacement methods and area-based regression method, i.e., Area and NctyArea, can be used to estimate overall spread rates. Estimating spread dynamics is more informative but more challenging than overall spread rates. Boundary displacement methods and area-based regression methods estimate the most reliable spread dynamics for all scenarios. For both overall spread rate and spread dynamics, boundary displacement methods have the best estimations. However, for spread rate without clear infestation outline, area-based regression methods can be good alternatives to estimate spread of invasive species. Additionally, because research using geopolitical-unit level records to estimate spread usually has a large spatial scale, the use of neighborhood measurement is suggested for highly asymmetric spread to better estimate invasion dynamics.

CHAPTER V: DETERMINING SPREAD RATE OF KUDZU BUG (HEMIPTERA: PLATASPIDAE) AND ITS ASSOCIATIONS WITH ENVIRONMENTAL FACTORS IN A HETEROGENEOUS LANDSCAPE

This chapter is a reformatted version of a paper by the same name published in *Environmental Entomology* by Liang, W., Tran, L., Wiggins, G., Grant, J., Stewart, S., & Washington-Allen, R.

Liang, W., Tran, L., Wiggins, G.J., Grant, J.F., Stewart, S.D., & Washington-Allen, R. (2019). Determining Spread Rate of Kudzu Bug (Hemiptera: Plataspidae) and Its Associations with Environmental Factors in a Heterogeneous Landscape. *Environmental Entomology, 48*, 309-317.

5.1 Abstract

Modeling invasion dynamics is valuable to guide management through early detection and prevention of further invasion. Herein, I first estimated the spread rate of kudzu bug with county-level invasion records and then determined important spatial factors impacting its spread during years 2010-2016. As kudzu bug infests a large heterogeneous area and shows asymmetric spread, I first utilized spatially constrained clustering (SCC), an unsupervised machine learning method, to divide the infested area into eight spatially-contiguous and environmentally homogenous neighborhoods. Distance regression and boundary displacement methods were then used to estimate the spread rates in all neighborhoods. Finally, multiple regression was applied to determine spatial factors influencing the spread of kudzu bug. The average spread rate reached 76 km/year by boundary displacement method, however, the rate varied largely among eight neighborhoods (45 - 144 km/year). In the southern region of the infested area, host plant density and wind speed were positively associated with the spread rate, whereas mean annual temperature, precipitation in the fall, and elevation had inverse relationships. In

the northern region, January minimum temperature, wind speed, and human population density showed positive relationships. This study increases the knowledge on the spread dynamics of kudzu bug. This research highlights the utility of SCC to determine natural clustering in a large heterogeneous region for better modeling local spread patterns and determining important factors impacting the invasions.

5.2 Introduction

Modeling invasion dynamics is a valuable tool to guide management, enhance early detection, and facilitate control of kudzu bug in its invaded range, and possibly prevent further invasion (Epanchin-Niell & Hastings, 2010). Various types of data have been used and compared to assess their usefulness in predicting spread dynamics of invasive species. Although the most desirable and accurate invasion records come from time-series field mapping (Hastings et al., 2005), increasingly, invasion histories that are recorded at geopolitical unit levels, such as county, vice-county, or township, have been used to estimate invasion dynamics (Evans & Gregoire, 2007; Hudgins et al., 2017; Liebhold et al., 1992; Perrins et al., 1993; Tobin et al., 2007). Although county-level quarantine records have coarse resolution, Tobin et al. (2007 & 2015) demonstrated that spread dynamics derived from these spatially crude records do not differ substantially from estimates derived from more extensive and costly trapping records.

The spread of invasive species in real-world situations is rarely uniform. In contrast, it tends to be faster in some regions/directions but slower in the others. The asymmetry of spread is usually attributed to a combined effect from various factors, such as geographic barriers, human-assisted movement, and environment heterogeneity (Andow et al., 1990; O’Reilly-Nugent et al., 2016; Fraser et al., 2015). For the asymmetric spread, the boundary displacement method is able to estimate regional spread dynamics by deriving the change of boundary displacement at specific temporal and local spatial scales (Morin et al., 2009). On the other hand, Andow et al. (1990) proposed to divide the infested areas into multiple neighborhoods within which the environmental conditions are similar. Estimating spread of invasive species for different homogeneous neighborhoods in a large heterogeneous area would provide better information on the spread dynamics and environmental factors contributing to the asymmetrical expansions. For example, Fraser et al. (2015) calculated the spread rate of American mink (*Neovison vison*) in three different regions in north-east Scotland and were able to get better estimates of its spread patterns. Lantschner et al. (2014) divided the infested region of woodwasp (*Sirex noctilio)* in the southern hemisphere into eight regions to analyze impact of environmental factors on its spread rate. However, currently a method, which quantitatively classifies the large heterogeneous area into several homogeneous neighborhoods, is lacking. In this study, the spatial constrained clustering (SCC) was used to quantitatively classify our large heterogeneous infested area into multiple homogeneous neighborhoods. SCC is an unsupervised machine learning method to determine natural clustering regarding environmental conditions within a large heterogeneous region, while assuring spatial adjacency of each cluster.

As the establishment of kudzu bug is relatively new in the U.S., the invasion dynamics and the factors that impact its spread are not well understood. Consequently, this research aims to address this research gap. The incidence of kudzu bug in the U.S. has been recorded at the county level as presence/absence since 2009. Therefore, this county-level invasion record was used to estimate the spread rate of kudzu bug. Two methods were used to estimate spread rates. The first method estimates the spread rate of kudzu bug across the whole infested area to get the overall spread rate in the U.S., whereas the second method estimates the spread rate in different neighborhoods to better determine its spread dynamics across heterogeneous environment. The SCC method was used to divide the large infested area into relatively-homogeneous neighborhoods. Finally, important environmental factors that influenced the spread of kudzu bug, during 2010-2016, were determined to provide information for early detection and management of kudzu bug.

5.3 Methodology

5.3.1 Environmental Factors and Neighborhood Classification

To classify the study area into environmentally homogenous neighborhoods, the potentially important factors related to the spread of kudzu bug were identified based upon studies on its biology. The spread of kudzu bug is thought to be facilitated by human-assisted movement of life stages, host plants and prevailing wind currents (Gardner et al., 2013). Research also indicates that both high (33℃) (Shi et al., 2014) and low temperatures (<5.1℃) (Grant & Lamp, 2017) inversely affected the establishment of kudzu bug populations. Furthermore, Liang et al. (2018a) suggested precipitation was an important factor for habitat suitability of kudzu bug. Therefore, environmental factors including natural, sociological, and host plant availability were included to determine their relationship with the spread rate of kudzu bug. Detailed information on environmental factors and data sources is listed in **Table 5.1**. Except for kudzu density and road density (**Table 5.1**), all other variables were obtained as spatial grid files. The kudzu density variable was obtained as spatial points where kudzu was observed, while the road density variable was obtained as spatial lines delineating the road networks in the U.S. For consistency with other variables, kudzu and road density variables were converted to spatial grid files indicating the mean number of reported kudzu and road length in each grid, respectively. The resolution of all variables was scaled to 1km.

Based upon reports compiled in EDDMapS, a repository of invasive species tracking records (EDDMapS, 2019), only two additional counties had documented occurrences of kudzu bug after 2017. Therefore all the counties that were infested by kudzu bug during 2010-2016 were included in the study area (i.e., 650 counties were finally used as our study area) (**Figure 5.1 A**, EDDMapS, 2019). To generate neighborhoods for analysis, the infested area was first uniformly divided into 180 sectors using lines radiating from the invasion origin in R with the “rgdal” package (**Figure 5.1 A**) (Bivand et al., 2014, R Core Team, 2013). SCC provided in the Spatial Statistics Toolbox in ArcGIS 10.5 (ESRI, 2011) was then used to classify all the 180 sectors into several homogeneous areas or clusters. This method classifies the spatial regions into given number of clusters based on input environmental factors by maximizing the environmental similarity within each cluster and minimizing the similarity among different clusters. Meanwhile, the SCC method also ensures the spatial connectivity of each cluster. To classify the 180 sectors into environmentally homogenous clusters, the mean values of the factors listed in **Table 5.1** in all sectors were extracted. The SCC was then conducted on the 180 sectors with the extracted factor values and classified all sectors into eight spatially contiguous clusters (**Figure 5.1 A**). Eight clusters were used (i.e., neighborhoods, **Figure 5.1 A**) as the neighborhoods for subsequent analysis, as this classification ensures the homogeneity within each neighborhood and also roughly corresponds to topographic variation in the landscape.

Table 5.1 Environmental factors used to classify neighborhoods and determine their importance related to spread rate of kudzu bug, and their original resolution and data sources.

Category	Environmental factors	Resolution	Data source
Climate	January minimum temperature (T)	4-km	PRISM Climate Data (Daly et al. 2008)
	Mean annual T		
	Maximum T of hottest quarter (Q)		
	Minimum T of coldest Q		
	Mean annual precipitation (P)		
	Mean P of fall season		
	Mean P of warmest Q		
	Annual horizontal wind speed	5-km	Kalnay et al. (1996)
	Annual vertical wind speed		
Host plant	Soybean density	30-m	CropScape (Han et al., 2012)
	Kudzu density	NA	EDDMaps (EDDMaps, 2019)
Sociology	Human population density	1-km	SILVIS Lab (http://silvis.forest.wisc.edu)
	Road density	NA	(U.S. Census Bureau TIGER/Line, https://www2.census.gov/geo/tiger/TGRGDB13/)
Geography	Elevation	1-km	Hydro-1K digital elevation model (USGS, 1996)

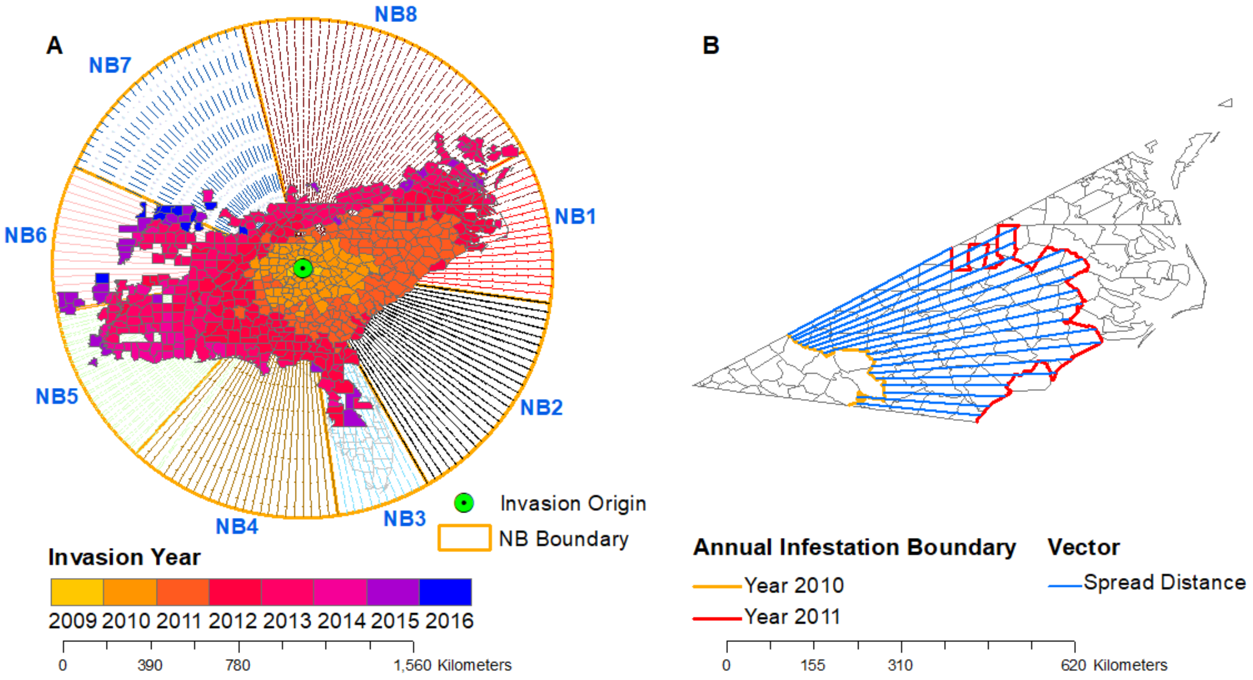

Figure 5.1 **A** Origin and spread history of kudzu bug in the southeastern U.S. and classification of eight neighborhoods (NB = neighborhood). Sectors are represented by the radial lines, and lines of the same color and type have similar environment. **B** Example of calculation for annual spread distance: mean length of blue vectors is used to indicate the spread distance of kudzu bug from 2010 to 2011 in NB 1.

5.3.2 Spread Rate of Kudzu Bug

To estimate the overall expansion pattern of kudzu bug in the southeastern U.S., the entire infested area was first taken as one region to derive the spread rate. Spread rates within each of the eight neighborhoods were also determined to better understand the localized spread dynamics of kudzu bug. The county-level records of first incidence of kudzu bug from 2010 to 2016 were used to calculate the spread distances for estimating invasion rates (**Figure 5.1 A**). The centroid of the first nine infested counties in Georgia in 2009 was used as the invasion origin of kudzu bug in the U.S. (**Figure 5.1 A**). Two commonly-used methods, boundary displacement and regression method (Gilbert & Liebhold, 2010; Tobin et al., 2015), were used to estimate the spread rates of kudzu bug.

For the boundary displacement method, the outer line of newly infested counties in each year was used as the annual infestation outline. The annual spread rate was calculated as the distance between two consecutive boundaries, which was measured as the mean length of all transects radiating from the invasion origin (**Figure 5.1 B**). For the regression method, the distances between the invasion origin and centroids of infested counties were regressed on the invasion time (i.e., when the counties were first infested) to determine the spread rates (Evans & Gregoire, 2007; Tobin et al., 2007). Two functions were used for the regression method to estimate the spread rate, a linear function and a logistic growth function as follows:

Linear function: $y = a * t + b$

Logistic growth function: $y = \frac{L}{\left(1 + e^{-k(t - t_0)}\right)}$

Where *y* is the total spread distance from invasion origin measured in each invasion period; *a* and *b* are the slope (spread rate in our case) and intercept of the linear function, respectively; *t* is the invasion time; *L* is a calculated asymptote (i.e. the furthest spread distance of kudzu bug in our case); *e* is the natural logarithm; and *k* is an instantaneous spread rate constant at t_0, the inflection point where y = *L*/2. The time unit in this research is 1 year. The Akaike information criterion (AIC) was used to compare the two models (Akaike, 1973). For the non-linear regression method with the logistic growth function, the annual spread rate was estimated as the derivative of the logistic function at each year, and the mean value of all annual rates was taken as the spread rate in a given region. To determine the similarities of spread rates estimated from the boundary displacement and regression methods, Pearson correlation coefficient, r, was calculated in R (R Core Team, 2013).

5.3.3 Impact of Environmental Factors on Spread Rate of Kudzu Bug

The annual spread rates calculated from the boundary displacement method in eight neighborhoods were used to determine their relationships with environmental factors. The mean values of all environmental factors were extracted for the newly infested areas in a given neighborhood and invasion year. In total, there were 47 observations with each observation representing the spread rate of kudzu bug in a given neighborhood in a particular year. As the study area is large, the impact of the same factor may vary across different neighborhoods. Therefore the eight neighborhoods were classified into two categories: those in the northern region (**Figure 5.1 A**) (NB1, NB6, NB7, and NB8) and those in the southern region (**Figure 5.1 A**) (NB2, NB3, NB4, and

NB5). Relationships between annual spread rates and environmental factors were analyzed for both regions. Various regression methods (e.g., Mortensen et al., 2009; Muirhead et al., 2006; Evans & Gregoire, 2007; Lantschner et al., 2014) have been commonly used to model relationships between spread rates and environmental variables. For regular regression methods an underlying assumption is that the residuals from regression model are random and independent. This assumption can be violated due to spatial autocorrelation (SA) within dependent and independent variables, in which case a spatial regression model should be used. Moran's I statistic is a widely used method to test the significance of SA within residuals of a regression model (Moran, 1950). Therefore the Moran's I statistic test was conducted on the residuals of the multiple regression of spread rates (i.e., dependent variable) on environmental variables (i.e., independent variables). The *P*-value of Moran's I statistic (0.525) indicated that the SA was not significant. Therefore, the regular regression model was used to determine environmental factors impacting spread rates of kudzu bug in the northern and southern regions. The Bonferroni outlier test provided in "car" package in R was used to detect outliers of the multiple regression model (Fox et al., 2012) and extremely significant outliers (e.g., 1 and 2 outliers in the southern and northern regions, respectively) were removed for further analyses ($P<0.01$). Outliers in this study were mainly caused by extraordinarily high spread rates of kudzu bug before year 2014 (rate > 240km/year). The best-subset selection method in R was used to select variables that resulted in a final model with highest adjusted R-square. Linear regression of individual variables was also performed to illustrate the significance of the relationship between variables and spread rate. All the data analysis in this study was conducted in R (R Core Team, 2013).

5.4 Results

5.4.1 Spread Rate of Kudzu Bug

When estimating the spread rate of kudzu bug for the entire infested area, the AIC of the non-linear regression (AIC = 8144) is lower than the linear model (AIC = 8185) indicating the former model is better. These values indicated that the spread rates of kudzu bug varied largely among invasion years and had an exponential spread during the early years. The average spread rates derived from the linear and non-linear regression are 101 km/year and 75 km/year, respectively (**Figure 5.2 A**). The annual spread distance derived from boundary displacement method differed greatly between 8-132 km/year among different invasion years (**Figure 5.2 B**), and the mean spread rate is 76 km/year, which is similar to the one derived from non-linear regression (75 km/year). The spread rate of kudzu bug decreased dramatically from 2013 (109 km/year) to 2014 (20 km/year) (**Figure 5.2 A-B**).

As the non-linear regression is better than the linear regression to estimate the spread rate of kudzu bug, therefore the non-linear regression and boundary displacement methods were included to estimate spread rates in all neighborhoods (**Figure 5.2 C**). Both methods indicated that the spread rates of kudzu bug varied largely among eight neighborhoods (**Figure 5.2 C**, **Table 5.2**). Spread rates varied from 44.6 to 143.9 for boundary displacement and 58.1 to 113.3 for non-linear regression. These varying spread rates across the study area suggests that the use of neighborhood measurements can better explain invasion dynamics of kudzu bug across heterogeneous environments than estimating one spread rate for the entire infested region. Directionally, kudzu bug had high spread rates in the northeastern (NB1) and southwestern (NB5) regions but low spread rates in the northern regions (NB7 and NB8) of the study area. Note that the estimated spread rates of kudzu bug in eight neighborhoods from non-linear regression and boundary displacement methods were similar (correlation coefficient r= 0.95) (**Table 5.2**).

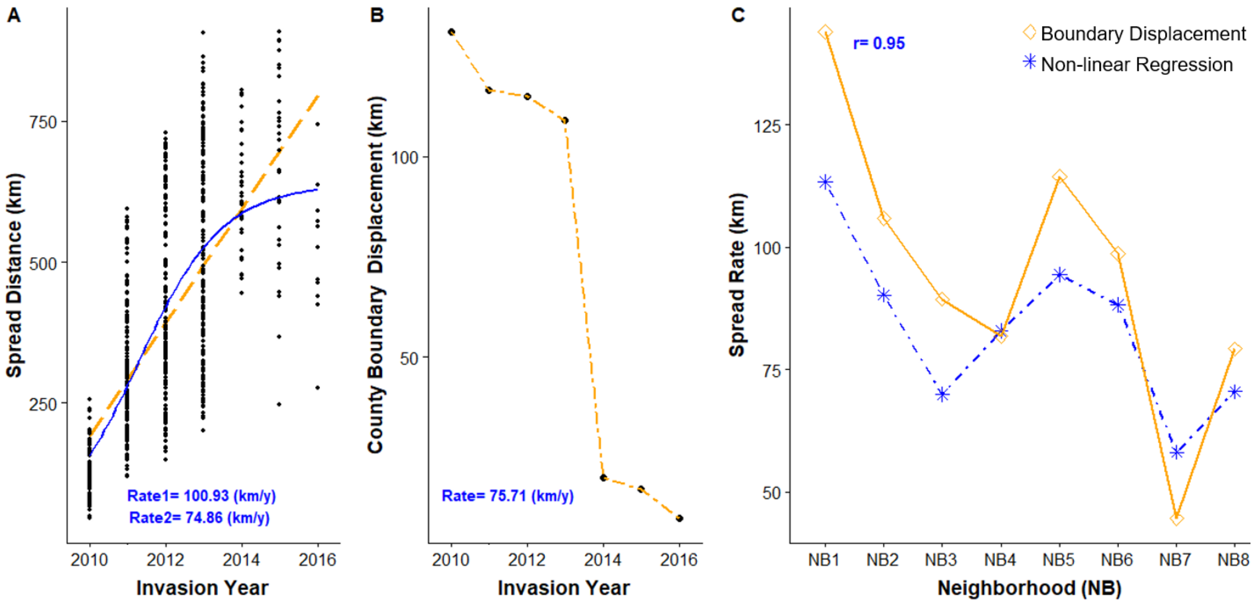

Figure 5.2 Spread rate estimated by two methods. **A** Linear (orange dashed line: Rate 1) regression and non-linear regression with logistic growth function (blue solid line: Rate 2) fit of spread distance against invasion time from 2010 to 2016 for the entire infested area. **B** Spread rate of kudzu bug using county boundary displacement methods for the entire infested area. **C** Spread rates of kudzu bug in eight neighborhoods derived from boundary displacement and non-linear regression and the correlations coefficient (r) of spread rates between two methods.

Table 5.2 Estimates of spread rate of kudzu bug in eight neighborhoods fitted by boundary displacement method and non-linear regression with logistic growth function ($R^2 = 0.95$)[1].

Neighborhood	NB1	NB2	NB3	NB4	NB5	NB6	NB7	NB8
Boundary displacement	143.9	105.9	89.4	81.8	114.5	98.7	44.6	79.2
Non-linear regression	113.3	90.3	69.9	82.9	94.4	88.3	58.1	70.5

[1] Correlation coefficient of estimated spread rates between boundary displacement and non-linear regression with logistic growth function.

5.4.2 Environmental Factors Impacting Spread Rate of Kudzu Bug

In the southern region of the infested area, all variables in total explained 99.6% of the variance. Based on the best subset selection, four variables were selected: soybean density, mean annual temperature, precipitation in fall season, and elevation (**Table 5.3**, **Figure 3 A-D**). The multiple regression model containing these four variables was highly significant ($P < 0.001$), and the four variables explained 81.4% of the variance (**Table 5.3**). Additionally, the individual fittings of these four variables with the spread rate were all significant ($P < 0.05$, **Figure 5.3 A-D**), and the fitted linear models were listed in **Figure 5.3**. Among the four variables, soybean density showed a significantly positive association with spread rate of kudzu bug (**Figure 5.3 A**), whereas the other three had significantly inverse relationships (**Figure 5.3 B-D**). Horizontal wind speed and kudzu density also had significantly positive associations with the spread rates but were not selected as part of the best subset (**Figure 5.3 E-F**).

Table 5.3 Parameters of the multiple regression model constructed with four best-subset variables for the southern regions of the infested area of kudzu bug ($R^2 = 0.814$)

Factors	Estimate	Standard error	Probability
Intercept	407.37	88.08	0.000
Soybean density	3883.27	922.96	0.001
Mean annual temperature	-13.44	4.70	0.012
Precipitation in Fall	-0.58	0.27	0.052
Elevation	-0.08	0.03	0.022

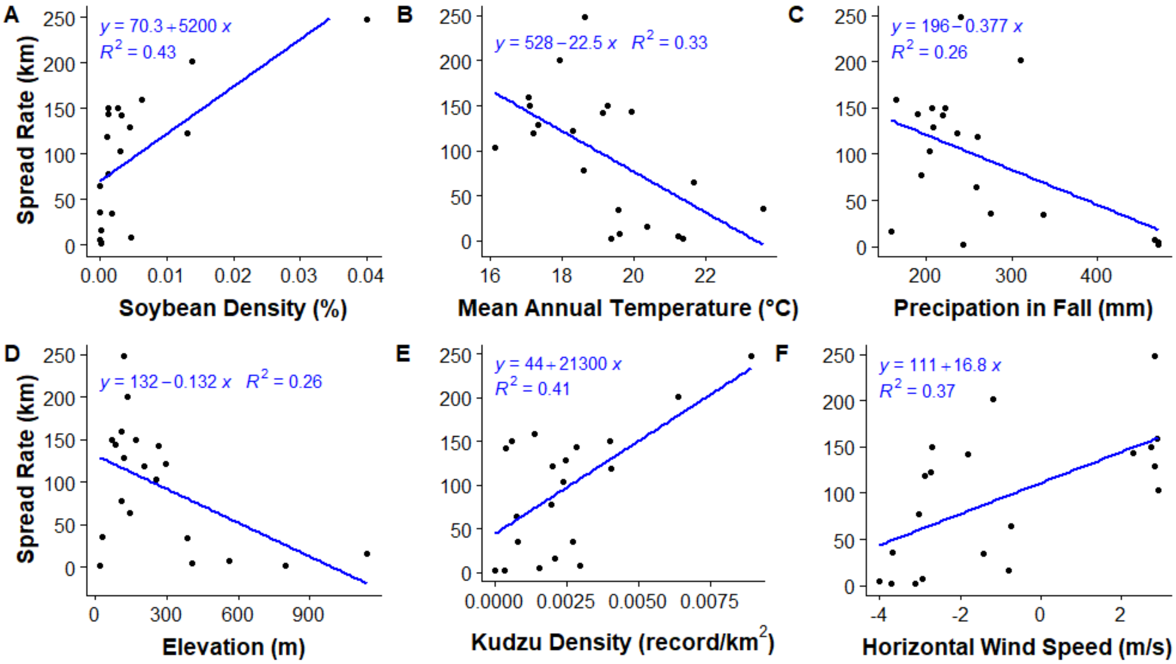

Figure 5.3 Important variables in southern regions. **A-D** Individual fitting of four best-subset variables with annual spread rates of kudzu bug from 2010 to 2016 in southern regions. **E-F** Variables that are significantly associated with spread rate but not selected by multiple regression. The *P*-values of model in **A-F** are 0.002, 0.008, 0.021, 0.021, 0.003 and 0.004, respectively. **F** Positive and negative values indicate west and east wind, respectively.

In the northern region, all the variables explained 87.6% of the variance. Using the same methods for best subset selection, three variables were selected: January minimum temperature, horizontal wind speed, and human population density. These three variables explained 59.2% of the variance (**Table 5.4**, **Figure 5.4 A-C**), and the multiple regression model containing these three variables was highly significant (P = 0.002, **Table 5.4**). For the individual fitting of each factor with spread rate, the impact of January minimum temperature and horizontal wind speed were both significant at the 0.05 level (**Figure 5.4 A-B**), whereas the impact of human population density was significant at 0.1 level (**Figure 5.4 C**). Although no significant association was observed between vertical wind and spread rates for all regions, extremely significant association was observed in the northeastern region (NB1, **Figure 5.4 D**). The vertical wind strength almost explained all the spread variance in NB1 using a quadratic function (**Figure 5.4 D**, R^2= 1), suggesting wind might be the determinant factor in this region.

Table 5.4 Parameters of the multiple regression model constructed with three best-subset variables for the northern regions of the infested area of kudzu bug (R^2 = 0.592)

Factors	Estimate	Standard error	Probability
Intercept	51.38	19.77	0.019
January minimum temperature	9.96	3.57	0.013
Horizontal wind speed	10.84	4.84	0.040
Human population density	0.48	0.20	0.029

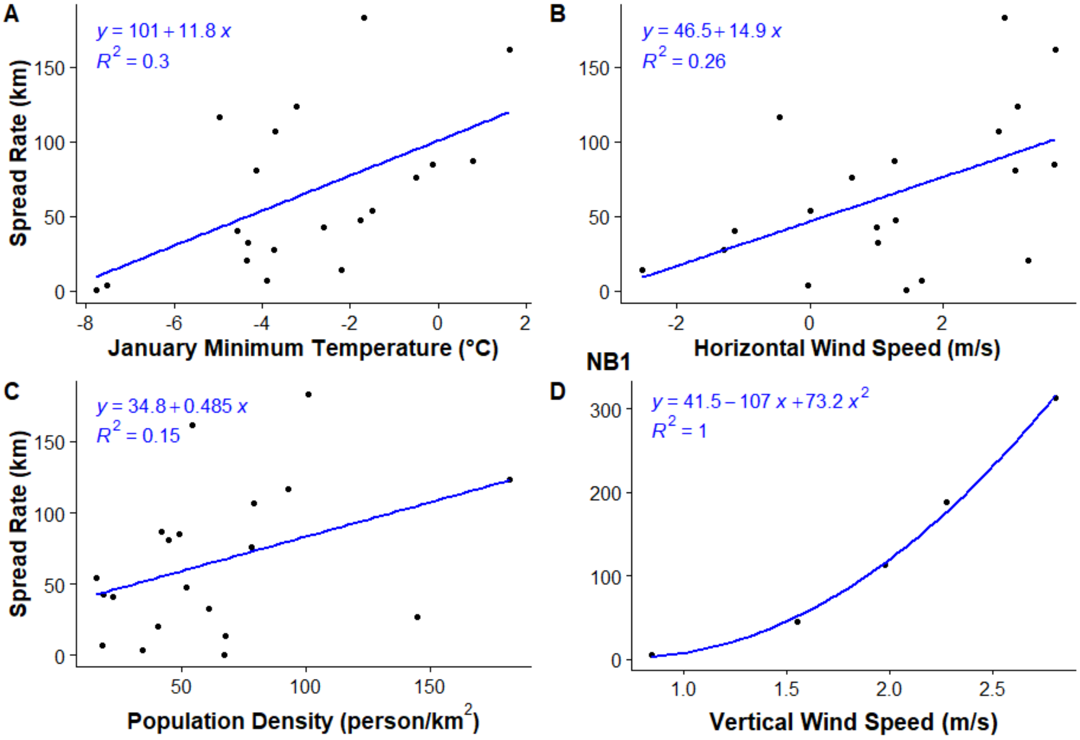

Figure 5.4 Important variables in northern regions. **A-C** Individual fitting of the three best-subset variables with annual spread rates of kudzu bug from 2010 to 2016 in northern region. *P*-values of model in **A-C** are 0.013, 0.021, and 0.089, respectively. **B** Positive and negative values indicate west and east wind, respectively. **D** Quadratic fitting between spread rate in neighborhood 1 (NB1) and vertical wind speed.

5.5 Discussion

The regression and boundary displacement methods estimated similar spread patterns across different invasion regions. With county-level invasion records, both methods captured similar spread distances at an equally coarse spatial level, thus delineating similar spread patterns. In our study, the non-linear regression with logistic growth function performed better than the linear regression on estimating the spread patterns, suggesting that kudzu bug showed an exponential expansion in the early years of invasion. Shigesada et al. (1995) classified expansion patterns of invasive species into three general types: linear, biphasic, and continually increasing patterns, and summarized the stratified diffusion that explains the non-linear expansions. Empirical research also has indicated that the non-linear relationship between expansion and invasion time is common (Weber, 1998; Williamson et al., 2005). I therefore suggest that, with county-level invasion records, multiple algorithms, including linear and non-linear should be explored and compared to determine the best model of describing the spread dynamics and expansion patterns of invasive species. Given the phenomenal spread rate of kudzu bug in northeastern and southwestern regions (more than 100 km/year), long-distance dispersal possibly occurred in the early invasion years of kudzu bug, which can greatly increase the overall invasion rate (Suarez et al., 2001).

Once introduced to a new spatial range, if an invasive species is able to establish, it starts to spread to adjacent areas. Given that most invasions are started by a few individuals and that kudzu bug was first found in nine counties in 2009 in Georgia, where it has two generations annually (Shigesada & Kawasaki, 1997; Zhang et al., 2012), it is likely that kudzu bug came into the U.S. and had already started to spread before 2009. Although the range expansion of kudzu bug has slowed dramatically, it may continue to invade the northern and western ranges of the U.S. (Liang et al., 2018a; Liang et al., 2018b). Liang et al. (2018a) used GARP and Maxent models to predict the potential invasion of kudzu bug in the Americas, and the Maxent model was more conservative than the GARP model. The slower spread rate of kudzu bug north and northwest of its invasion origin (**Figure 5.1A**) more-closely fits the prediction of the Maxent model (Liang et al., 2018a).

Ecological processes are closely affected by environmental patterns, which vary across spatial regions (Addicott et al., 1987). Consequently, the limiting factors of an invasive species might vary significantly across different regions (Morin et al., 2007). Due to the environmental heterogeneity over the large study area, different sets of environmental factors were found to significantly impact the spread of kudzu bug in the northern and southern regions of infested area. The availability of host plants played a more important role in the southern region, while the low temperature in January was the most important factor in the northern region. Horizontal winds towards the east were positively associated with spread rate of kudzu bugs in both regions. Precipitation in the fall of the year and elevation both showed inverse association with spread rate of kudzu bug only in the southern region, whereas the human population density showed a positive association with the spread rate only in the northern region. Temperature, wind current, and host plant had been found playing important roles on spread of invasive species (e.g., Evans & Gregoire, 2007; Lantschner et al., 2014; Sakai et al., 2001). The decreasing availability of host plants, increasing temperature, and high humidity in the fall may contribute to the slower spread of kudzu bug in the southern region after 2014. In the northern region, low temperature is the limiting factor of further spread of the kudzu bug. In the upper Midwestern U.S. where kudzu bug is not yet found, but high acreages of soybean are grown, the spread is influenced more by temperature than host plant. The dominant role of vertical wind strength in NB1 (northeastern region) highlights the possibility that the significant expansion of kudzu bug in northeast of its invasion origin might be attributed to dominant air currents (Gardner et al., 2013).

In addition to the environmental factors included in this research, the dramatic decrease of spread rate since year 2014 may also result from the presence of natural predators. The decreased spread rate of kudzu bug coincided with the discovery of egg parasitoids, such as *Paratelenomus saccharalis*, which has been found in Georgia, Florida, Alabama, Mississippi, and Tennessee since 2013 (Gardner & Olson, 2016; Medal et al., 2015). The entomopathogenic fungus *Beauveria bassiana* has been observed to devastate populations of kudzu bug since 2013 in Georgia, South Carolina, and Tennessee (Britt et al., 2016; Gardner & Olson, 2016; Seiter et al., 2014). Investigations on associations between kudzu bug and other natural predators deserves more attention in future studies.

Studies on estimating invasion dynamics with county-level invasion records (e.g., Evans & Gregoire, 2007; Hudgins et al., 2017; Liebhold et al., 1992; Tobin et al., 2015) often cover large areas (e.g., multiple states covering several million squared kilometers). Consequently, environmental homogeneity is less likely to exist across the whole infested area. Environmental heterogeneity can invoke complicated invasion dynamics over a large spatial area (Fraser et al., 2015) and cause significant variations of spread rates in different regions. Therefore, asymmetric spread of invasive species is more likely the case than symmetric spread (Evans & Gregoire, 2007; Suarez et al., 2001; Tobin et al., 2007). Thus, with county-level records, using the neighborhood measurement to divide a large study area into sub-areas is suggested to reflect environmental homogeneity in each sub-region for a better understanding of the spread dynamics. Additionally, as seen in this research, environmental factors may influence the invasion of the same species in different ways and magnitudes across a large heterogenous region (Hastings et al., 2005). The neighborhood measurement, therefore, can also facilitate better analysis of determinant environmental factors over heterogeneous environments (Dewhirst & Lutscher, 2009).

To divide the entire infested area into smaller neighborhoods to better estimate spread dynamics, multiple methods have been implemented. Andow et al. (1990) used boundaries reflecting major irregularities of infested regions (e.g., rivers, mountain ranges, etc.), while Fraser et al. (2015) divided the whole infested region based on the introduction origins and geographical boundaries. Morin et al. (2009) divided the infested area of hemlock woolly adelgid (HWA) into uniform size of sectors based on the vectors radiating from invasion origin to reflect the spatial anisotropy in the spread rate of HWA. Additionally, in several studies, infested regions were divided into sub-regions based on the geopolitical area (Evans & Gregoire, 2007; Lantschner et al., 2014). These methods are valuable in practice, however, they do not ensure the environmental homogeneity within each neighborhood from a quantitative perspective. The SCC method, on the other hand, can quantitatively classify large heterogenous region into spatially-contiguous and environmentally homogeneous sub-regions. This advantage of SCC should be considered in future studies focusing on estimating localized spread dynamics of invasive species and potentially important factors impacting the invasions.

5.6 Conclusions

This research increases the knowledge of spread dynamics of kudzu bug in the southeastern U.S. The influence of determinant factors (e.g., temperature, precipitation, host plant, etc.) on kudzu bug's spread rate varied across different invasion regions due to heterogeneity of the landscape. Methodologically, the study also shows the suitability of the neighborhood measurement method to estimate invasion dynamics and to determine important factors impacting biological invasions over a large heterogeneous environment. Furthermore, the SCC method was able to quantitatively classify a large and heterogeneous infested region into spatially-connected and environmentally-homogeneous neighborhoods. Thus, the use of SCC method is highlighted to classify neighborhoods for large heterogeneous region to better estimate localized spread dynamics of invasive species. In terms of invasive species-related policies, strict regulatory measures, including quarantines and international commodity and transport inspections, are recommended to reduce the potential invasion of kudzu bug, especially for regions with warm temperatures and moderate to high populations of host plants in South America. Focusing on the U.S., for the purpose of early detection and management of kudzu bug, monitoring soybean fields and kudzu sentinel plots should be conducted regularly in regions where no kudzu bugs are reported but are potential distribution areas (Liang et al., 2018a & 2018b). Furthermore, decreasing the availability of host plants, especially kudzu, would be a valid method, but impractical, to control the population of kudzu bug and slow its spread in the U.S.

CHAPTER VI: CONCLUSIONS

Based on the current distribution of kudzu bug in the U.S., the environmental conditions in the areas inhabited by kudzu bugs are not the same between native and invaded ranges. In the U.S., kudzu bugs have not been reported in 67% of the available environmental space, compared to space occupied by populations in their native ranges. These data suggest that the invasion of kudzu bug is far from equilibrium and more places could be further infested in the U.S. Kudzu bugs generally inhabit warm (annual mean temperature around 15℃) and humid (annual mean precipitation around 1300mm) regions. Based on the invasion risk model, most states in eastern U.S., most countries in Central America, and countries in central South America have high invasion risk of kudzu bug. Soybean is an important crop in Central and South American countries, including Mexico, Brazil, and Argentina. Thus, to avoid potential economic losses of soybean by kudzu bug, quarantines, as well as monitoring and detection programs, in these regions are recommended to prevent the entry and establishment of kudzu bug. Meanwhile, a majority of the soybean production in the U.S. is distributed in the upper Midwest, where most of the regions have not been infested by kudzu bug. Therefore, monitoring and early detection is also highly recommended in the upper Midwest of the U.S. to avoid invasion and outbreak of kudzu bug.

A significant shift of both available environmental spaces and occupied environmental space by kudzu bug was observed between the native (Asia) and invaded (U.S.) ranges of kudzu bug. However, five of the seven species distribution models (SDMs) included (i.e., RF, BRT, MAXENT, ANN, and GARP) still showed fair transferability while at least two of the models (i.e., BRT and RF) showed good transferability, indicating the non-adaptive niche shift does not necessarily challenge the transferability of SDMs. Thus, different from the suggestion from Broennimann and Guisan (2008) that both native and invaded occurrences should be used for invasion modeling when niche shift is observed, one should first determine the type of niche shift (adaptive or non-adaptive). Non-adaptive niche shift, in some cases, does not challenge model transferability, thus occurrence data in native range should be used, especially when the equilibrium in the invaded range is not reached. The spatial range where pseudo-absences (Pas) are extracted can also significantly impact both interpolation and transferability of SDMs. In the case of kudzu bug, extracting PAs from both the native and invaded ranges, while only native presences were used for model construction, significantly overestimates interpolations of all models, but underestimates transferability of all models. Thus, the challenge of the environmental space anisotropy on model transferability can be largely avoided or reduced by constraining the PAs to the same spatial range as the presences. For the classical approach of modeling potential invasions with native occurrences, the PAs should only be extracted from the native range to avoid misestimates of model performances.

Geopolitical-unit invasion records are capable of accurately estimating spread of invasive species. Both regression methods and boundary displacement methods can be used to estimate the overall invasion rate, expansion pattern, and spread dynamics of non-native species. However, one should carefully select the optimal method to estimate spread based on the question of interest, anisotropy and stochasticity in spread, and variations and sizes of geopolitical unit. When the question of interest is the overall spread rate and sizes of geopolitical units are relatively uniform, among the eight methods considered, except MinD method, all other methods can be used for spread without LDJD. If the question of interest includes estimating overall spread rate in several sub-regions, among which mean sizes of geopolitical units vary largely, both MinD and Ncty method should be avoided, whereas NctyArea method is a good alternative for Ncty method under this scenario. The spread distance-based regression methods (i.e., CtdDis and MinD Reg) are more sensitive to the irregularities and stochasticity in spread than the area-based regression methods and boundary displacement methods. For LDJD, boundary displacement methods and area-based regression method, i.e., Area and NctyArea, can be used to estimate overall rates. Estimating spread dynamics is more informative but more challenging than overall rates. Boundary displacement methods and area-based regression methods estimate the most reliable spread dynamics for all scenarios. For both overall spread rate and spread dynamics, boundary displacement methods have the best estimations. However, for spread without a clear infestation outline, area-based regression methods can be good alternatives for estimating spread of invasive species.

The average spread rate of kudzu bug reached 76 km/year during 2010-2016, however, the rate varied largely among different invasion years (8-132 km/year) and eight neighborhoods (45 - 144 km/year). The spread rate of kudzu bug decreased dramatically from 2013 (109 km/year) to 2014 (20 km/year); meanwhile kudzu bug had high spread rates in the northeastern and southwestern regions but low spread rates in the northern regions of the study area. The influence of determinant factors (e.g., temperature, precipitation, host plant, etc.) on kudzu bug’s spread rate varied across different invasion regions due to heterogeneity of the landscape. The availability of host plants seemed to play a more important role in the southern region, while the low temperature in January was the most important factor in the northern region. Horizontal winds towards the east were positively associated with spread rate of kudzu bugs in both regions. Precipitation in the fall of the year and elevation both showed inverse association with spread rate of kudzu bug only in the southern region, whereas the human population density showed a positive association with the spread rate only in the northern region. In the northern region, low temperature is the limiting factor of further spread of the kudzu bug. Methodologically, the study also shows the suitability of the neighborhood measurement method to estimate invasion dynamics and to determine important factors impacting biological invasions over a large heterogeneous environment. Furthermore, the spatially constrained clustering (SCC) method is able to quantitatively classify a large and heterogeneous infested region into spatially-connected and environmentally-homogeneous neighborhoods. Thus, the use of SCC method to classify neighborhoods for large heterogeneous regions is recommended to better estimate localized spread dynamics of invasive species.

REFERENCES

Addicott, J.F., Aho, J.M., Antolin, M.F., Padilla, D.K., Richardson, J. S., & Soluk, D.A. (1987). Ecological neighborhoods: scaling environmental patterns. *Oikos, 49,* 340-346.

Aikio, S., Duncan, R. P., & Hulme, P. E. (2010). Herbarium records identify the role of long-distance spread in the spatial distribution of alien plants in New Zealand. *Journal of Biogeography, 37*, 1740-1751.

Akaike, H. (1998). Information Theory and an extension of the maximum likelihood principle. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, pp. 267-281. In: Petrov, B. N. & Csaki, F. (eds.), Proceedings of the 2nd International Symposium on Information Theory. Budapest: Akademiai Kiado.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology, 43,* 1223-1232.

Anderson, R.P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography, 37,* 1378-1393.

Andow, D.A., Kareiva P. M., Levin, S. A., & Okubo, A. (1990). Spread of invading organisms. *Landscape Ecology, 4,* 177-188.

Barbet-Massin, M., Jiguet, F., Albert, C.H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution, 3*, 327-338.

Beardsley, J.W., & Fluker, S. (1967). *Coptosoma xanthogramma* (White),(Hemiptera: Plataspidae) a new pest of legumes in Hawaii. *Proceedings of the Hawaiian Entomological Society, 19,* 367-372.

Bivand, R., Keitt, T., & Rowlingson, B. (2014). rgdal: Bindings for the geospatial data abstraction library. R package version 0.8-16.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Breiner, F.T., Nobis, M.P., Bergamini, A., & Guisan, A. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution, 9*, 802-808.

Britt, K., Grant, J., Wiggins, G., & Stewart, S. (2016). Prevalence and localized impact of the entomopathogenic fungus *Beauveria bassiana* on kudzu bug (*Megacopta cribraria*) in Eastern Tennessee. *Journal of Entomological Society, 51,* 321-324.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., & Graham, C.H. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography, 21*, 481-497.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T., & Guisan, A. (2007). Evidence of climatic niche shift during biological invasion. *Ecology Letters, 10*, 701-709.

Broennimann, O., & Guisan, A. (2008). Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters, 4*, 585-589.

Broennimann, O., Di Cola, V., Petitpierre, B., Breiner, F., Manuela, D., Randin, C., Engler, R., Hordijk, W., Pottier, J., Di Febbraro, M., & Pellissier, L. (2016). Package ‘ecospat’.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Costa, G.C., Nogueira, C., Machado, R.B., & Colli, G.R. (2010) Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. *Biodiversity Conservation, 19,* 883-899.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., & Pasteris, P.P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology, 28,* 2031-2064.

De Meyer, M., Robertson, M.P., Mansell, M.W., Ekesi, S., Tsuruta, K., Mwaiko, W., Vayssières, J.F., & Peterson, A.T. (2010). Ecological niche and potential geographic distribution of the invasive fruit fly *Bactrocera invadens* (Diptera, Tephritidae). *Bulletin of Entomological Research, 100,* 35-48.

De Souza Muñoz, M.E., De Giovanni, R., de Siqueira, M.F., Sutton, T., Brewer, P., Pereira, R.S., Canhos, D.A.L. & Canhos, V.P. (2011) OpenModeller: a generic approach to species' potential distribution modelling. *Geoinformatica, 15*, 111-135.

Dewhirst, S., & Lutscher, F. (2009). Dispersal in heterogeneous habitats: thresholds, spatial scales, and approximate rates of spread. *Ecology, 90,* 1338-1345.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F.T., D'amen, M., Randin, C., Engler, R., Pottier, J., Pio, D., Dubuis, A., & Pellissier, L. (2017). Ecospat: an R package to support spatial analyses and modeling of species niches and distributions. *Ecography, 40,* 774-87.

Dolédec, S., & Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology, 31*, 277-294.

Early, R., & Sax, D.F. (2014). Climatic niche shifts between species' native and naturalized ranges raise concern for ecological forecasts during invasions and climate change. *Global Ecology and Biogeography, 23,* 1356-1365.

EDDMapS (2019). Early detection and distribution mapping system. In the Univ. Georg. Cent. Invasive Species Ecosyst. Heal. Retrieved from https://www.eddmaps.org on February, 18, 2018.

Eger Jr, J. E., Ames, L. M., Suiter, D. R., Jenkins, T. M., Rider, D. A., & Halbert, S. E. (2010). Occurrence of the Old World bug *Megacopta cribraria* (Fabricius) (Heteroptera: Plataspidae) in Georgia: a serious home invader and potential legume pest. *Insecta Mundi, 121*, 1-11.

Elith, J., & Franklin, J. (2013). Species distribution modeling. *Encyclopedia Biodiversity, 6,* 692-705.

Elith, J., & Graham, C.H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography, 32*, 66-77.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics, 40,* 677-697.

Elith, J., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., & Loiselle, B. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29,* 129-151.

Epanchin-Niell, R.S., & Hastings, A. (2010). Controlling established invaders: integrating economics and spread dynamics to determine optimal management. *Ecology Letter, 13,* 528-541.

ESRI, R. (2011). ArcGIS desktop: release 10. Environmental Systems Research Institute, CA.

Evans, A.M., & Gregoire, T.G. (2007). A geographically variable model of hemlock woolly adelgid spread. *Biological Invasions, 9*, 369-382.

Fabricius, J.C. (1798). Entomologia systematica emendata et auct, secundum classes, ordines, genera, species, adjectis synonymis, locis, observationibus. Supplementum. Proft et Storch, Gopenhagan. ii: 572.

Fielding, A.H., & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environment Conservation, 24*, 38-49.

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of Human Genetics, 7*, 355-369.

Fitzpatrick, M.C., Weltzin, J.F., Sanders, N.J., & Dunn, R.R. (2007). The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography, 16*, 24-33.

Fox J., Weisberg, S., Bates, D., & Fox, M.J. (2012). R package “car v3.0-2”.

Franklin, J. (2013). Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions, 19*, 1217-1223.

Fraser, E. J., Lambin, X., Travis, J. M., Harrington, L. A., Palmer, S. C., Bocedi, G., & Macdonald, D. W. (2015). Range expansion of an invasive species through a heterogeneous landscape–the case of American mink in Scotland. *Diversity and Distributions, 21,* 888-900.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19,* 1-67.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics, 29,* 1189-1232.

Gallagher, R.V., Beaumont, L.J., Hughes, L., & Leishman, M.R. (2010). Evidence for climatic niche and biome shifts between native and novel ranges in plant species introduced to Australia. *Journal of Ecology, 98,* 790-799.

Gallardo, B., Zu Ermgassen, P.S., & Aldridge, D.C. (2013). Invasion ratcheting in the zebra mussel (*Dreissena polymorpha*) and the ability of native and invaded ranges to predict its global distribution. *Journal of Biogeography, 40*, 2274-2284.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I., & Thuiller, W. (2010). Predicting potential distributions of invasive species: where to go from here? *Diversity and Distribution, 16,* 331-342.

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N.E., & Thuiller, W. (2012). Invasive species distribution models–how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography, 21*, 1126-1136.

Gardner, W., & Olson, D. M. (2016). Population census of *Megacopta cribraria* (Hemiptera: Plataspidae) in kudzu in Georgia, USA, 2013–2016. *Journal of Entomology Society, 51*, 325-328.

Gardner, W.A., Peeler, H. B., LaForest, J., Roberts, P. M., Sparks Jr, A. N., Greene, J. K., Reisig, D., Suiter, D. R. , Bacheler, J. S., Kidd, K., & Ray, C. H. (2013). Confirmed distribution and occurrence of *Megacopta cribraria* (F.) (Hemiptera: Heteroptera: Plataspidae) in the southeastern United States. *Journal of Entomology Society, 48*, 118-127.

Gilbert, M., & Liebhold, A. M. (2010). Comparing methods for measuring the rate of spread of invading populations. *Ecography, 33*, 809-817.

Grant, J.I., & Lamp, W.O. (2017). Cold tolerance of *Megacopta cribraria* (Hemiptera: Plataspidae): an invasive pest of soybeans. *Environmental Entomology, 46*, 1406-1414.

Guisan, A., Petitpierre, B., & Broennimann, O. (2014) Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution, 29*, 260-269.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters, 8*, 993-1009.

Guisan, A., & Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135,* 147-186.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., & Peterson, A.T. (2007). What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecological Monographs, 77*, 615–630.

Gurevitch, J., & Padilla, D. K. (2004). Are invasive species a major cause of extinctions? *Trends in Ecology & Evolution, 19*, 470-474.

Han, W., Yang, Z., Di, L., & Mueller, R. (2012). CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture, 84,* 111-123.

Haregeweyn, N., Tsunekawa, A., Tsubo, M., Meshesha, D., & Melkie, A. (2013). Analysis of the invasion rate, impacts and control measures of *Prosopis juliflora*: a case study of Amibara District, Eastern Ethiopia. *Environmental Monitoring and Assessment, 185*, 7527-7542.

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models. London: Chapman and Hall, *137-173*.

Hastings, A., Cuddington, K., Davies, K.F., Dugaw, C.J., Elmendorf, S., Freestone, A., Harrison, S., Holland, M., Lambrinos, J., Malvadkar, U., & Melbourne, B.A. (2005). The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters, 8,* 91-101.

Hecht-Nielsen, R. (1988). Neurocomputing: picking the human brain. *IEEE Spectrum, 25,* 36-41.

Heikkinen, R. K., Marmion, M., & Luoto, M. (2012). Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography, 35,* 276-288.

Hernandez, P.A., Graham, C.H., Master, L.L., & Albert, D.L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography, 29,* 773-785.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology, 25*, 1965-1978.

Horvitz, N., Wang, R., Wan, F. H., & Nathan, R. (2017). Pervasive human-mediated large-scale invasion: analysis of spread patterns and their underlying mechanisms in 17 of China's worst invasive plants. *Journal of Ecology, 105*, 85-94.

Hudgins, E. J., Liebhold, A. M., & Leung, B. (2017). Predicting the spread of all invasive forest pests in the United States. *Ecology Letters, 20,* 426-435.

Hutchinson, G.E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding Remarks, 415-427.*

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., & Lobo, J.M. (2011). Use of niche models in invasive species risk assessments. *Biological Invasions, 13,* 2785-2797.

Jiménez-Valverde, A., & Lobo, J.M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica, 31*, 361–369.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., & Zhu, Y. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*, 437-472.

Kot, M., Lewis, M.A., & Van Den Driessche, P. (1996). Dispersal data and the spread of invading organisms. *Ecology, 77,* 2027-2042.

Landis, J.R., & Koch, G.C. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lantschner, M.V., Villacide, J.M., Garnas, J.R., Croft, P., Carnegie, A.J., Liebhold, A.M., & Corley, J.C. (2014). Temperature explains variable spread rates of the invasive woodwasp *Sirex noctilio* in the Southern Hemisphere. *Biological Invasions, 16,* 329-339.

Lek, S., & Guégan, J.F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling, 120,* 65-73.

Li, Y.H., Pan, Z.S., Zhang, J.P., & Li, W.S. (2001). Observation of biology and behavior of *Megacopta cribraria* (Fabricius). *Plant Protection Technology, 21*, 11-12.

Liang, W., Tran, L., Washington-Allen, R., Wiggins, G., Stewart, S., Vogt, J., & Grant, J. (2018a). Predicting the potential invasion of kudzu bug, *Megacopta cribraria* (Heteroptera: Plataspidae), in North and South America and determining its climatic preference. *Biological Invasions, 20*, 2899-2913.

Liang, W., Papes, M., Tran, L., Grant, J., Washington-Allen, R., Stewart, S., & Wiggins, G. (2018b). The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. *Ecological Modelling, 388*, 1-9.

Liang, W., Tran, L., Wiggins, G., Grant, J., Stewart, S., & Washington-Allen, R. (2019). Determining spread rate of kudzu bug (Hemiptera: Plataspidae) and its associations with environmental factors in a heterogeneous landscape. *Environmental Entomology, 48*, 309-317.

Liebhold, A.M., Halverson, J.A., & Elmes, G.A. (1992). Gypsy moth invasion in North America: a quantitative analysis. *Journal of Biogeography, 19,* 513-520.

Liu, C., Berry, P.M., Dawson, T.P., & Pearson, R.G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography, 28,* 385-393.

Liu, X., Li, X., Liu, Z., Tingley, R., Kraus, F., Guo, Z., & Li, Y. (2014). Congener diversity, topographic heterogeneity and human-assisted dispersal predict spread rates of alien herpetofauna at a global scale. *Ecology Letters, 17*, 821-829.

Loo, S.E., Nally, R.M., & Lake, P.S. (2007). Forecasting New Zealand mudsnail invasion range: model comparisons using native and invaded ranges. *Ecological Applications, 17,* 181-189.

Mainali, K.P., Warren, D.L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., Karki, D., Shrestha, B.B., & Parmesan, C. (2015). Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. *Global Change Biology, 21*, 4464-4480.

Masciocchi, M., & Corley, J. (2013). Distribution, dispersal and spread of the invasive social wasp (*Vespula germanica*) in Argentina. *Austral Ecology, 38,* 162-168.

Mau-Crimmins, T.M., Schussman, H.R., & Geiger, E.L. (2006). Can the invaded range of a species be predicted sufficiently using only native-range data? Lehmann lovegrass

(*Eragrostis lehmanniana*) in the southwestern United States. *Ecological Modelling, 193*, 736-746.

Medal, J., Santa Cruz, A., Williams, K., Fraser, S., Wolaver, D., Smith, T., & Davis, B. J. (2015). First record of *Paratelenomus saccharalis* (Hymenoptera: Platygastridae) on kudzu bug *Megacopta cribraria* (Heteroptera: Plataspidae) in Florida. *Florida Entomology, 98,* 1250-1251.

Medley, K.A. (2010). Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus Skuse* (Culicidae), revealed by reciprocal distribution models. *Global Ecology Biogeography, 19,* 122-133.

Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ, 5*, e2849.

Mineur, F., Davies, A. J., Maggs, C. A., Verlaque, M., & Johnson, M. P. (2010). Fronts, jumps and secondary introductions suggested as different invasion patterns in marine species, with an increase in spread rates over time. *Proceedings of the Royal Society of London B: Biological Sciences, 277,* 2693–2701.

Moran, P.A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37,* 17-23.

Morin, R. S., Liebhold, A. M., & Gottschalk, K. W. (2009). Anisotropic spread of hemlock woolly adelgid in the eastern United States. *Biological Invasions, 11,* 2341-2350.

Morin, R. S., Liebhold, A. M., Tobin, P. C., Gottschalk, K. W., & Luzader, E. (2007). Spread of beech bark disease in the eastern United States and its relationship to regional forest composition. *Canadian Journal of Forest Research, 37,* 726-736.

Mortensen, D. A., Rauschert, E. S., Nord, A. N., & Jones, B. P. (2009). Forest roads facilitate the spread of invasive plants. *Invasive Plant Science and Management, 2,* 191-199.

Muirhead, J.R., Leung, B., van Overdijk, C., Kelly, D.W., Nandakumar, K., Marchant, K.R., & MacIsaac, H.J. (2006). Modelling local and long-distance dispersal of invasive emerald ash borer *Agrilus planipennis* (Coleoptera) in North America. *Diversity and Distributions, 12,* 71-79.

Nathan, R. (2006). Long-distance dispersal of plants. *Science, 313,* 786-788.

O'Reilly-Nugent, A., Palit, R., Lopez-Aldana, A., Medina-Romero, M., Wandrag, E., & Duncan, R.P. (2016). Landscape effects on the spread of invasive species. *Current Landscape Ecology Reports, 1*, 107-114.

Paini, D.R., Sheppard, A.W., Cook, D.C., De Barro, P.J., Worner, S.P., & Thomas, M.B. (2016). Global threat to agriculture from invasive species. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 7575-7579.

Parravicini, V., Azzurro, E., Kulbicki, M., & Belmaker, J. (2015). Niche shift can impair the ability to predict invasion risk in the marine realm: an illustration using Mediterranean fish invaders. *Ecology Letters, 18*, 246-253.

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography, 12*, 361-371.

Pejchar, L., & Mooney, H. A. (2009). Invasive species, ecosystem services and human well-being. *Trends in Ecology & Evolution, 24,* 497-504.

Perrins, J., Fitter, A., & Williamson, M. (1993). Population biology and rates of invasion of three introduced Impatiens species in the British Isles. *Journal of Biogeography, 20,* 33-44.

Peterson, A.T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology, 78*, 419-433.

Peterson, A.T., Papes, M., & Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography, 30,* 550-560.

Peterson, A.T., & Soberón, J. (2012). Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza Conservação, 10*, 102–107.

Peterson, A.T. (2011). Ecological niche conservatism: a time-structured review of evidence. *Journal of Biogeography, 33*, 817-827.

Petitpierre, B., Kueffer, C., Broennimann, O., Randin, C., Daehler, C., & Guisan, A. (2012). Climatic niche shifts are rare among terrestrial plant invaders. *Science, 335*, 1344-1348.

Phillips, S.J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography, 31*,161-175.

Phillips, S.J. (2008). Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). *Ecography, 31,* 272-278.

Phillips, S.J., Anderson, R.P., & Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*, 231-259.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics, 52*, 273-288.

Prasad, A.M., Iverson, L.R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*, 181–199.

Pratt, P. D., Slone, D. H., Rayamajhi, M. B., Van, T. K., & Center, T. D. (2003). Geographic distribution and dispersal rate of *Oxyops vitiosa* (Coleoptera: Curculionidae), a biological control agent of the invasive tree *Melaleuca quinquenervia* in south Florida. *Environmental Entomology, 32,* 397-406.

Pyšek, P., Jarošík, V., Müllerová, J., Pergl, J., & Wild, J. (2008). Comparing the rate of invasion by *Heracleum mantegazzianum* at continental, regional, and local scales. *Diversity and Distributions, 14*, 355-363.

Qiao, H., Feng, X., Escobar, L.E., Peterson, A.T., Soberón, J., Zhu, G., & Papes, M. (2018). An evaluation of transferability of ecological niche models. *Ecography, 42,* 521-534.

R Development Core Team. (2013). R: a language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria http://www.R-project.org.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography, 33,* 1689-703.

Ruberson, J.R., Takasu, K., Buntin, G.D., Eger, J.E., Gardner, W.A., Greene, J.K., Jenkins, T.M., Jones, W.A., Olson, D.M., Roberts, P.M., & Suiter, D.R. (2013). From Asian curiosity to eruptive American pest: *Megacopta cribraria* (Hemiptera: Plataspidae) and prospects for its biological control. *Applied Entomology and Zoology*, *48*, 3-13.

Sakai, A.K., Allendorf, F.W., Holt, J.S., Lodge, D.M., Molofsky, J., With, K.A. & McCauley, D.E. (2001). The population biology of invasive species. *Annual Review of Ecology and Systematics*, *32*, 305-332.

Schoener, T.W. (1968). The Anolis lizards of Bimini: resource partitioning in a complex fauna. *Ecology, 49,* 704-726.

Seiter, N.J., Greene, J.K., & Reay-Jones, F.P. (2012). Reduction of soybean yield components by *Megacopta cribraria* (Hemiptera: Plataspidae). *Journal of Economic Entomology, 106,* 1676-1683.

Seiter, N.J., Grabke, A., Greene, J.K., Kerrigan, J.L., & Reay-Jones, F.P. (2014). *Beauveria bassiana* is a pathogen of *Megacopta cribraria* (Hemiptera: Plataspidae) in South Carolina. *Journal of Entomological Science, 49,* 326-330.

Sharov, A. A., Pijanowski, B.C., Liebhold, A. M., & Gage, S. H. (1999). What affects the rate of gypsy moth (Lepidoptera: Lymantriidae) spread: winter temperature or forest susceptibility? *Agricultural and Forest Entomology, 1*, 37-45.

Shi, S., Cui, J., & Zang, L. (2014). Development, survival, and reproduction of *Megacopta cribraria* (Heteroptera: Plataspidae) at different constant temperatures. *Journal of Economic Entomology, 107*, 2061-2066.

Shigesada, N., & Kawasaki, K. (1997). Biological invasions: theory and practice. Oxford University Press, UK.

Shigesada, N., Kawasaki, K., & Takeda, Y. (1995). Modeling stratified diffusion in biological invasions. *The American Naturalist, 146,* 229-251.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics, 21*, 3940-3941.

Skellam, J. G. (1951). Random spread in theoretical populations. *Biometrika, 38*, 196-218.

Sobek-Swant, S., Kluza, D.A., Cuddington, K., & Lyons, D.B. (2012). Potential distribution of emerald ash borer: What can we learn from ecological niche models using Maxent and GARP? *Forest Ecology and Management, 281,* 23-31.

Soberón, J., & Peterson, A.T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics, 2*, 1–10.

Soberón, J., & Peterson, A.T. (2011). Ecological niche shifts and environmental space anisotropy: a cautionary note. *Revista Mexicana de Biodiversidad, 82*, 1384-1355.

Stockwell, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science, 13,* 143-158.

Stohlgren, T.J., & Schnase, J.L. (2006). Risk analysis for biological hazards: what we need to know about invasive species. *Risk Analysis: An International Journal, 26,* 163-173.

Stokland, J.N., Halvorsen, R., & Støa, B. (2011). Species distribution modelling—Effect of design and sample size of pseudo-absence observations. *Ecological Modelling, 222,* 1800-1809.

Strubbe, D., Broennimann, O., Chiron, F., & Matthysen, E. (2013). Niche conservatism in non-native birds in Europe: niche unfilling rather than niche expansion. *Global Ecology and Biogeography, 22,* 962-970.

Suarez, A.V., Holway, D. A., & Case, T. J. (2001). Patterns of spread in biological invasions dominated by long-distance jump spread: insights from Argentine ants. *Proceedings of the National Academy of Sciences, 98,* 1095-1100.

Suiter, D.R., Eger, J.E., Gardner, W.A., Kemerait, R.C., All, J.N., Roberts, P.M., Greene, J.K., Ames, L.M., Buntin, G.D., Jenkins, T.M., & Douce, G.K. (2010). Discovery and distribution of *Megacopta cribraria* (Hemiptera: Heteroptera: Plataspidae) in northeast Georgia. *Journal of Integrated Pest Management*, *1,* 1-4.

Sutherst, R.W. (2003). Prediction of species geographical ranges. *Journal of Biogeography, 30,* 805-816.

Svenning, J.C., Fløjgaard, C., Marske, K.A., Nógues-Bravo, D., & Normand, S. (2011). Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews, 30*, 2930-2947.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Thuiller, W., Georges, D., Engler, R., Breiner, F., Georges, M.D., & Thuiller, C.W. (2016). Package ‘biomod2’.

Tobin, P.C., Liebhold, A. M., Roberts, E. A., & Blackburn, L. M. (2015). Estimating spread rates of non-native species: The gypsy moth as a case study. In Venette, R. C. (Ed.) Pest Risk Modelling and Mapping for Invasive Alien Species. *CABI International and USDA, Wallingford,* 131-145.

Tobin, P.C., Liebhold, A.M., & Roberts, E.A. (2007). Comparison of methods for estimating the spread of a non-indigenous species. *Journal of Biogeography, 34,* 305-312.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007) A comparative evaluation of presence only methods for modelling species distribution. *Diversity and Distributions, 13*, 397-405.

USGS. (1996). HYDRO1k elevation derivative database. Retrieved from http://lta.cr.usgs.gov/ hydro1k on 22 August 2016.

Václavík, T., & Meentemeyer, R.K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling, 220,* 3248-3258.

VanDerWal, J., Shoo, L.P., Graham, C., & Williams, S.E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling, 220*, 589-594.

Verbruggen, H., Tyberghein, L., Belton, G.S., Mineur, F., Jueterbock, A., Hoarau, G., Gurgel, C.F.D., & De Clerck, O. (2013). Improving transferability of introduced species' distribution models: new tools to forecast the spread of a highly invasive seaweed*. PLoS One, 8*, e68337.

Walsh, J. R., Carpenter, S.R., & Vander Zanden, M.J. (2016). Invasive species triggers a massive loss of ecosystem services through a trophic cascade. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 4081-4085.

Wang, Z., Wang, H., Chen, G., Zhuge, Z., & Tong, C. (1996). Occurrence and control of *Megacopta cribraria* (Fabricius) on soybean. *Plant Protection, 3,* 7-9.

Wang, R., & Wang, Y.Z. (2006). Invasion dynamics and potential spread of the invasive alien plant species *Ageratina adenophora* (Asteraceae) in China. *Diversity and Distributions, 12,* 397-408.

Warren, D.L., Glor, R.E., & Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution, 62,* 2868-2883.

Weber, E. (1998). The dynamics of plant invasions: a case study of three exotic goldenrod species (*Solidago* L.) in Europe. *Journal of Biogeography, 25*, 147-154.

Wiens, J.J., & Graham, C.H. (2005). Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics, 36,* 519-539.

Williams, F., Eschen, R., Harris, A., Djeddour, D., Pratt, C., Shaw, R.H., Varia, S., Lamontagne-Godwin, J., Thomas, S.E., & Murphy, S.T. (2010). The Economic Cost of Invasive Non-Native Species on Great Britain. CABI Publishing, Wallingford.

Williamson, M., Pyšek, P., Jarošík, V., & Prach, K. (2005). On the rates and patterns of spread of alien plants in the Czech Republic, Britain, and Ireland*. Ecoscience, 12,* 424-433.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions, 14*, 763-773.

Wu, M., Wu, Z., & Hua, S. (2006). A preliminary study on some biological characters of globular stink bug, *Megacopta cribraria* and its two egg parasitoids. *Journal of Fujian Agriculture for University (Natural Science Edition), 35,* 147-150.

Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford, A.J., Ban, S., Barbosa, A.M., & Dormann, C.F. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution, 33,* 790-802.

Zhang, Y., Hanula, J.L., & Horn, S. (2012). The biology and preliminary host range *of Megacopta cribraria* (Heteroptera: Plataspidae) and its impact on kudzu growth. *Environmental Entomology, 41*, 40-50.

Zhu, G., Petersen, M.J., & Bu, W. (2012). Selecting biological meaningful environmental dimensions of low discrepancy among ranges to predict potential distribution of bean plataspid invasion. *PloS One, 7,* e46247.

Zhu, G.P., Rédei, D., Kment, P., & Bu, W.J. (2014). Effect of geographic background and equilibrium state on niche model transferability: predicting areas of invasion of *Leptoglossus occidentalis*. *Biological Invasions, 16,* 1069-1081.

VITA

Wanwan grew up in Luoyang, a city that is in central China and was the capital city for 13 ruling dynasties in Chinese history. Wanwan likes running, hiking, and playing badminton with friends. Her beloved parents are always supportive of her decisions. As a generation born in 1990s, Wanwan witnessed the transition of Agriculture in China from traditional to industrial mode but also the pollution and degradation of the environment, which lead Wanwan into the broad fields of agriculture and environment. Wanwan received her Bachelor in Horticulture from Henan Agricultural University in 2013, and her Masters in Plant Protection in China Agricultural University in 2015. In fall 2015, Wanwan came to the University of Tennessee, Knoxville, and started her Ph.D. program in the Department of Entomology and Plant Pathology. At the same time, Wanwan pursued a Master's degree in Statistics. During a 4-year Ph.D. program, Wanwan conducted multidisciplinary research aiming to use statistics/machine learning, GIS, and remote sensing to deal with environment or agriculture related issues. Following graduate school, Wanwan dedicates her career to the broad field of agriculture and environment sustainability at large spatial scales either in academics or in the industrial field.