Model Suitable Habitat Distribution of Triatominae with Random Forest

L. Huang[1,3], C. Panethymitaki[1], R. Langdon[1,3], A. Sanchéz[3,9], W.H. Mobley[4], M.G. Shensky[5], T. P. Feria Arroyo[6], T. Oraby[7], E. Rebollar-Téllez[8], C. Dawson[1,2] and K.A. Brown[1,3]

[1]Oden Institute for Computational Engineering and Sciences, [2]Department of Aerospace Engineering and Engineering Mechanics, The University of Texas at Austin; [3]Cavendish Laboratory, University of Cambridge, UK; [4]Texas Advanced Computing Center, Austin, TX; [5]University of Texas Libraries, The University of Texas at Austin; [6]Department of Biology, [7]School of Mathematical and Computational Sciences, The University of Texas Rio Grande Valley; [8] Universidad Autónoma de Nuevo León, Mexico; [9]London Institute of Medical Sciences, Imperial College, London

Abstract

Chagas is caused by the protozoal parasite *Trypanosoma cruzi*. This vector-borne, neglected sub-tropical disease impacts over 8 million people across the Americas, for which the triatominaes (kissing bugs) are the primary vector carriers. The presence of triatominaes is the most crucial factor in Chagas infection. With most studies focusing on South America, Chagas is under more neglect in North America. Our study aims to fill this gap.

Understanding how triatominaes distribution varies in this century can guide us to take action to prevent the spreading of Chagas. Climate change is vital in influencing each kissing bug species' suitable habitat distribution. The Random Forest is powerful in predicting species distribution. Inspired by these, we started this project to use the random forest to predict triatominae's suitable habitat distribution under climate change scenarios. This study is part of the FloDisMod Project.

Research Questions

(1) Can we predict the triatominaes' future suitable habitat distribution based on their current distribution?

(2) How will climate change influence triatominaes' suitable habitat distribution?

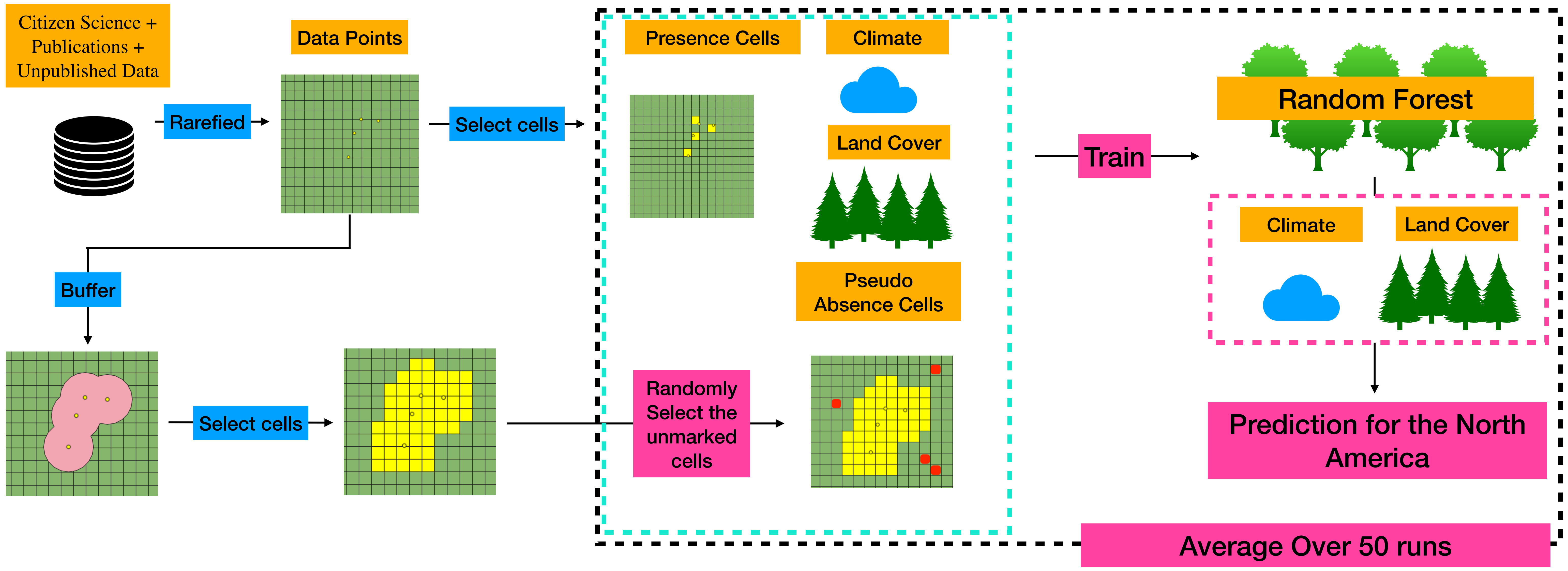
(3) What precautions can we take to stop triatominaes from infecting people?

Data

We combined North America kissing bug observations from various sources, including citizen science (iNaturalist) and published and unpublished observations. This formed a database containing 12 species: *T. sanguisuga*, *T. rubida*, *T. recurva*, *T. protracta*, *T. mexicana*, *T. mazzottii*,*T. neotomae*, *T. longipennis*, *P. lecticularia*, *T. indictiva*, *T. gerstaeckeri and T. dimidiata*. The historical and future projected climate (33 variables) and land cover (7 variables) data (1km resolution) under CMIP6 scenarios came from AdaptWest[1] and the study of Guangzhao Chen et al.[2].

Method

We proposed one solution to training Random Forest with presence-only data by randomly selecting pseudo-absence points outside buffers around presence points. With the presence records from 1991-2020 (in total 2282 records), we trained the Random Forest Classifier with climate and land-cover data to predict the presence-absence of each species at a given location in North America for 11 species (not enough data points for *T. neotomae*). The models were evaluated with 5-fold cross-validation TSS, ROC and AUC scores, and the variable importance was accessed with Boruta, Shapely Values and mean-decrease in mini-impurity importance. We further generated their suitable habitat spatial distribution map for 1991-2020 and the climate suitability projections for 2011-2040, 2041-2070, and 2071-2100 under four CMIP6 climate change scenarios: SSP1 RCP2.6, SSP2 RCP4.5, SSP3 RCP7.0, and SSP5 RCP5.8.



Citizen Science + Publications + Unpublished Data

Rarefied

Data Points

Select cells

Buffer

Select cells

Presence Cells

Climate

Land Cover

Pseudo Absence Cells

Randomly Select the unmarked cells

Train

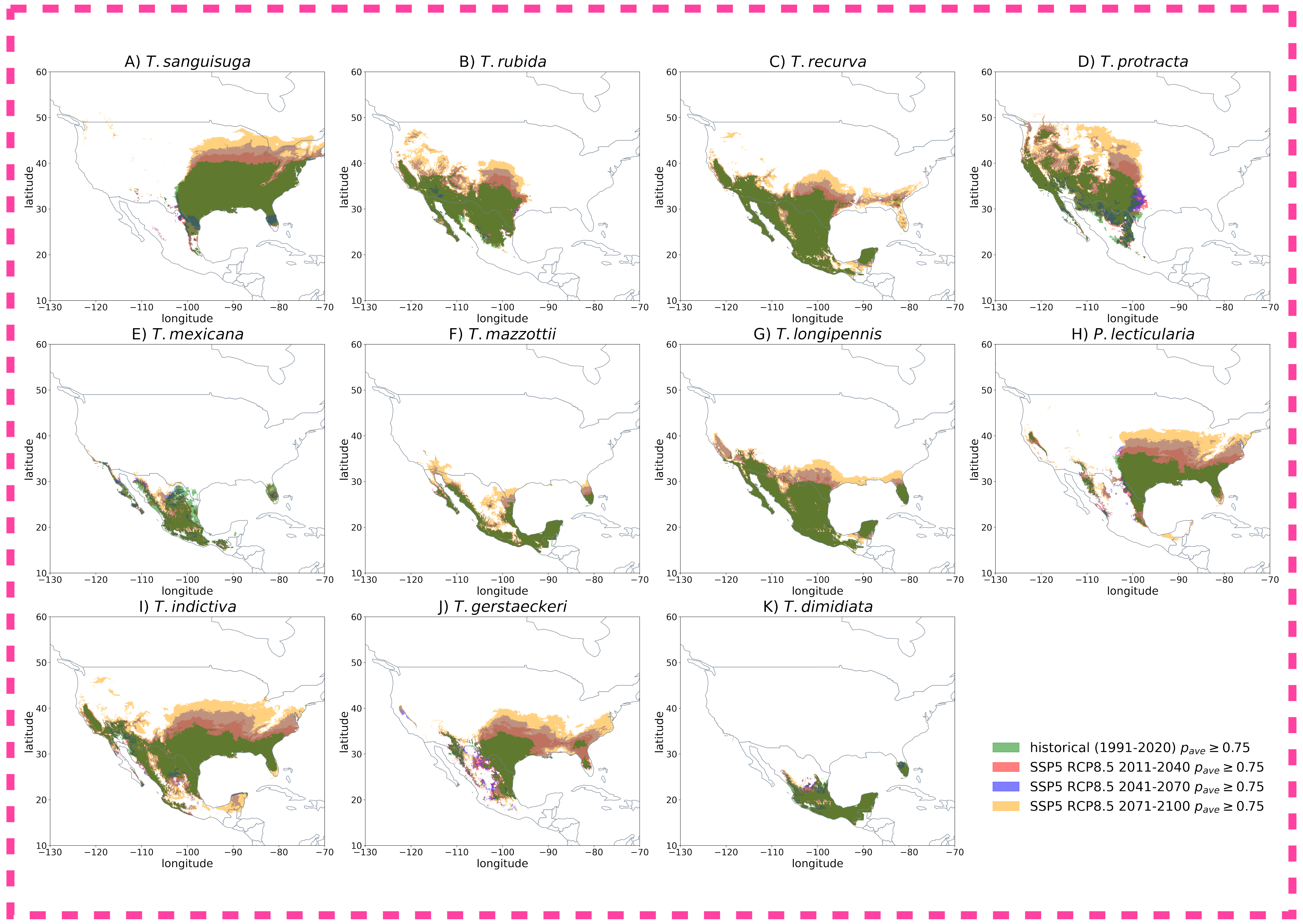
Random Forest

Climate

Land Cover

Prediction for the North America

Average Over 50 runs



A) *T. sanguisuga*

B) *T. rubida*

C) *T. recurva*

D) *T. protracta*

E) *T. mexicana*

F) *T. mazzottii*

G) *T. longipennis*

H) *P. lecticularia*

I) *T. indictiva*

J) *T. gerstaeckeri*

K) *T. dimidiata*

latitude
longitude
60
50
40
30
20
10
−130
−120
−110
−100
−90
−80
−70
historical (1991-2020) $p_{ave} \geq 0.75$
SSP5 RCP8.5 2011-2040 $p_{ave} \geq 0.75$
SSP5 RCP8.5 2041-2070 $p_{ave} \geq 0.75$
SSP5 RCP8.5 2071-2100 $p_{ave} \geq 0.75$

Result

We can answer the first two research questions:

(1) These 11 Random Forest models have 5-fold cross-validation TSS and AUC scores larger or equal to 87.2% and 97.8%. We can have a high accuracy prediction on the triatominaes' suitable habitat.

(2) We expect to see the suitable habitat of some triatominae species expanding northward. In the worst emission scenario SSP5 RCP8.5, even south of Canada expect to be celibately suitable for some of these bugs.

Discussion

To better answer the third research question, further studies need to be carried out. We need to understand how people's lifestyle contributes to the infection and acts in the ecological niche of the triatominaes. The FloDisMod project aims to introduce alternative statistical machine-learning approaches to improve the current infectious disease models. We will compare this Random Forest model with MaxEnt and Bayesian in the next step to identify the most important variables.

[1] AdaptWest Project. 2022. Gridded current and projected climate data for North America at 1km resolution, generated using the *ClimateNA v7.30* software (T. Wang et al., 2022). Available at adaptwest.databasin.org.

[2] Chen, G., Li, X. & Liu, X. Global land projection based on plant functional types with a 1-km resolution under socio-climatic scenarios. *Sci Data* **9**, 125 (2022). https://doi.org/10.1038/s41597-022-01208-6

Discuss your thoughts:
Liting Huang
liting@utexas.edu