A Random Forest Regression Model for Predicting the Movement of Horseshoe Crabs in Long Island Sound

Samah Senbel[1(✉)], Jo-Marie Elisha Kasinak[2], and Jennifer Mattei[2]

[1] School of Computer Science and Engineering, Welch College of Business and Technology, Sacred Heart University, Connecticut, USA
senbels@sacredheart.edu

[2] Department of Biology, College of Arts and Science, Sacred Heart University, Connecticut, USA
{kasinakj,matteij}@sacredheart.edu

Abstract. Developing models to predict animal movement patterns is an important area of study in ecology and wildlife management. Project *Limulus,* a community research program, has been tracking the movement of tagged American horseshoe crabs (*Limulus polyphemus*) in Long Island Sound since 1998. During the spawning season, horseshoe crabs are captured by hand, tagged and then released. Recaptured horseshoe crabs give valuable information about their behavior and movement patterns. In this paper, we tested various models to find the best predictor for the movement of horseshoe crabs based on the observed movement activity in previous years. We experimented with three different models: Linear Regression, Decision Tree, and Random Forest Regression models. We used the data for 2018 as our test set and the data of all previous 19 years as our training set (19,219 records). The Random Forest Regression model proved to be the best predictive model and resulted in the smallest RMSE and MAE, as well as the smallest maximum error in prediction. The predicted horseshoe crab locations can be targeted in the next season for recapturing tagged horseshoe crabs. It also concentrates the scientists' effort and time to find the maximum number of horseshoe crabs.

Keywords: Prediction model · Random forest · Horseshoe crabs · Animal tracking · Movement prediction · Long Island Sound

1 Introduction and Literature Review

The study and understanding of the movement of animals is a long-standing area of scientific interest. Movement ecology has been practiced for hundreds of years by farmers tracking locust outbreaks, to migratory bird counts organized by non-profit groups starting at the turn of the twentieth century [1]. Animal movement studies provide in-depth knowledge of animal behavior, when and where animals move, and in some cases, why they move (e.g. for food or reproduction) [2]. The technology to track animals

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 107–119, 2021.
https://doi.org/10.1007/978-3-030-86973-1_8

has changed vastly over the years, but the goals are unchanged and remains critically important for wildlife conservation.

In the 1990s, Global Positioning Systems (GPS) and satellite technology became a popular, albeit invasive, way to perform animal tracking. A radio frequency identification (RFID) tag or a GPS devise (e.g. radio collar) is physically attached to the animal, and its location is transmitted periodically to a central monitoring location [3]. This method is the current favorite of researchers due to the continuous data locations being sent on a regular basis. However, it is an expensive option and the tags are usually heavy making them more suitable for larger animals [4]. Another invasive but much less harmful method is the use of externally applied, traditional plastic or metal tags with unique identification numbers on them that are attached to a large group of animals. They are later recaptured and their location, health, and mates noted. This methods works best with smaller animals, insects, and birds. Typically, as many animals as possible are tagged to increase the probability of recapture. This method depends heavily on citizen scientists or community volunteers [5] and researchers to report any sightings or recaptures of these animals.

The American horseshoe crab, *Limulus polyphemus*, can be found along the Atlantic coast of the United States from Maine to Florida and in parts of the Gulf Coast. This study concentrated on Long Island Sound (LIS) which is bounded by the coasts of CT and Long Island, New York. Horseshoe crabs move to find food in the coastal benthic zone and during the spawning season to find nesting sites along shorelines. The horseshoe crabs in LIS come to shore from mid-May through June for the spawning season. Horseshoe crabs play a key role in the eel and whelk fishing industry as bait and an unparalleled, integral part in ensuring the safety of vaccines and medical devices as products from their blood are used for bacterial toxin testing [6, 7, and 8].



Fig. 1. A tagged horseshoe crab

Project *Limulus* [9], a community research program, at Sacred Heart University focuses on the ecology and population dynamics of horseshoe crabs in LIS since 1998. It is run as a research and educational tool to increase public awareness of the horseshoe crab and its connection to the LIS ecosystem and human health. It is also a data-gathering and analyzing network to help state managers direct conservation programs for the horseshoe crab population in LIS, to allow for its sustainable harvest. Their movement

pattern over the seasons is of particular interest as individuals of the population frequently cross State boundaries with very different management programs [10, 11]. Figure 1 shows a tagged horseshoe crab. The tag has a unique ID number and a website to visit for reporting recaptures.

Three research projects also concentrate on the study of the movement of horseshoe crabs in particular. A study in New England [12] uses both acoustic telemetry and tagging to study the movement of horseshoe crabs on Pleasant bay in Cape Cod, Massachusetts, USA. They tagged and followed the movement of 70 crabs over a period of two years using acoustics and 2000 horseshoe crabs using tagging similar to our project. Their results are very interesting, but we note the limitation of the study due to its short time period and small horseshoe crab population test set. Another project was conducted on the Asian horseshoe crabs [13] in the south of Japan. A group of 20 horseshoe crabs were fitted with electronic tags and their movement was tracked for two years. Their mating habits were also studied. Although it is a useful and informative study, the small set of horseshoe crabs that were tagged and studied does not give a complete picture of the movement of the Asian horseshoe crabs. Closer to LIS, Bopp et al. [11] conducted a study on the movement pattern and survival rates of horseshoe crabs in NY State waters. They divided the NY coastal area into three parts: Jamaica bay, south shore of Long Island, and North shore of Long Island. They conclude that the survival rate is highest in the northern part of Long Island for both male and female horseshoe crabs. And, there is a trend of movement towards the north shore of Long Island as well. They limited their work to New York City shores. Their results supports the findings of our previous work, where we observed a trend for horseshoe crabs to move towards the shores of CT from the NY metropolitan region [14].

This prediction model focuses on the population dynamics of horseshoe crabs in LIS. The study of the movement of horseshoe crabs is an important part of Project *Limulus*. With 20 years of accumulated movement data, it is possible to paint a fairly accurate picture of their travel behavior both in the short term and over the years19. Our analysis showed an interesting trend in movement: horseshoe crabs are mostly static, but when they do move, there is a tendency to move towards the northern part of LIS. The data shows that horseshoe crabs from Rhode Island rarely travel west to the LIS, and that those from the northern part of the Sound rarely travel west or south. Meanwhile, there is a slight tendency for horseshoe crabs from the western New York City part of LIS to travel northeast, and horseshoe crabs from the south shores of LIS also show a tendency to move northeast [14].

Building on this collection of work on the horseshoe crab population in LIS, it was imperative to have a good estimate of the best locations to capture horseshoe crabs in the future and to predict the possible movement of crabs that have not been recaptured based on the general movement of the population. We experimented with three different models: Linear Regression, Decision Tree, and Random Forest Regression models. We used the data for 2018 as our test set and the data of all previous 19 years as our training set (19,219 records).

In this paper, we present a prediction model for finding the best horseshoe tagging locations using three different methods, and compare their performances. In Sect. 2, we describe the dataset characteristics, performance measures, and the prediction models. Section 3 presents our implementation results and conclusion.

2 Prediction Model

Our three models were coded using the Python coding language [15] on a regular Intel i5 laptop. The dataset was divided into two parts, a training set ("train") for building our models and a testing set ("test") for measuring their performance. Each set is further divided into the 5 input fields: initial longitude, initial latitude, initial date, is Male, and is Female, (used in X_train and X_test) and the two output fields, final longitude and final latitude (used in Y_train and Y_test). Figure 2 shows how the dataset was split into those 4 subsets.

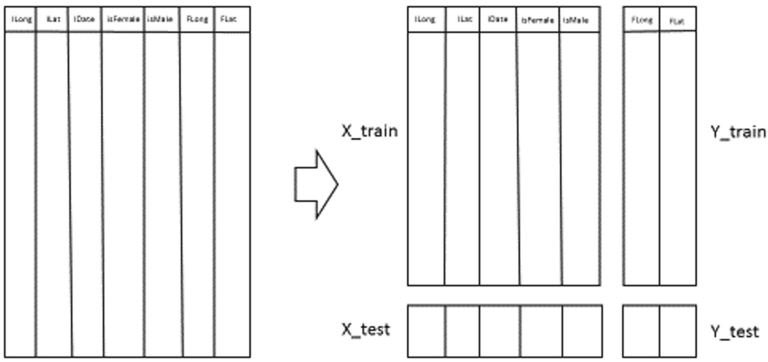

Fig. 2. An illustration of how the dataset was split into X_train, Y_train, X_test, and Y_test.

Performance Measure: For continuous-valued estimation problems, two performance measures are typically used to measure the accuracy of a model's prediction: the root mean squared error (RMSE) and the mean absolute error (MAE). RMSE measures the root of the average squared error of our predictions. The smaller the MSE, the better. Equation 1 shows the MSE calculation. RMSE is simply the root of the MSE.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{1}$$

The MAE is another popular measure: the error is calculated as an average of absolute differences between the target values and the predictions. Equation 2 shows how to calculate the MAE.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2}$$

An important distinction between this metric and the RMSE is that MAE does not penalizes huge errors as much as MSE does. Thus, it is not as sensitive to outliers compared to the mean square error. We used both metrics and they gave similar results.

The Movement Dataset and Its Features: Our dataset consists of a large repository of 19,219 movements of horseshoe crabs throughout the LIS area collected between 1998 and 2019. Each record consists of the horseshoe crab tag number, sex, initial latitude and longitude, initial capture (tagging) date, final latitude and longitude, and final recapture date. In the early years of Project *Limulus* (1998–2008), the data collection was mainly in the south-central CT area (between longitudes −72.5 and −73.5 and between latitude 41.1 and 41.3 approximately) and on a relatively small-scale. 2009 to 2013 were years of large-scale data collection as well as a great expansion in the collection area to cover all of LIS. From 2014 to 2018, the data collection size was stabilized and still covers the entire LIS area, however the data collection size was smaller due to less funding for the project. Figure 3 shows the number of recaptured horseshoe crabs, respectively from 2000–2018.

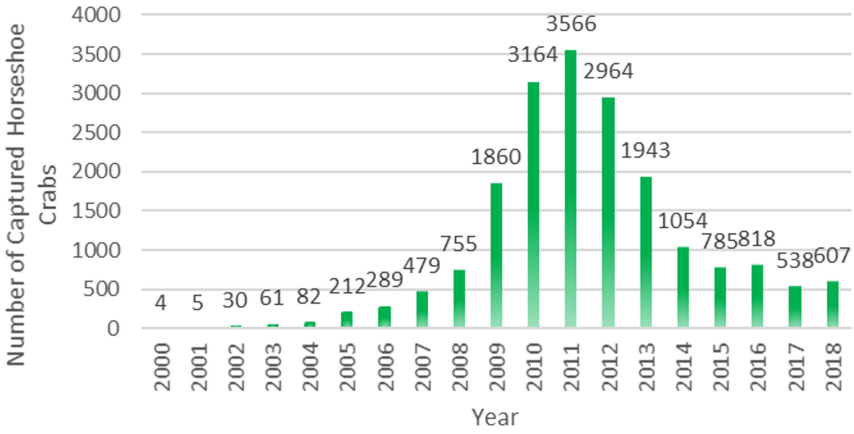

Fig. 3. Number of recaptured horseshoe crabs from 2000–2018.

The nature of capture-and-release animal tracking means that only a small percentage of the tagged animals are recaptured. In Project *Limulus*, about 17% of all tagged horseshoe crabs were ever re-captured, and the majority of crabs are only recaptured once (Table 1).

Table 1. Number of horseshoe crabs recaptures

Number of recaptures	Number of crabs
0	~37,000
1	12364
2	2544
3	564
4	131
5	43
6	6
7	1 (Horseshoe Crab ID 172143)
8	1 (Horseshoe Crab ID 192594)
9	1 (Horseshoe Crab ID 182857)

Our prediction model would generate a list of target beaches to be visited by community volunteers in the expectation of capturing previously tagged horseshoe crabs. This would hopefully lead to a rise in the number of recaptures, thereby providing more information about the movement of the horseshoe crabs. Some particular crabs have been captured multiple times over the years, making their movement, growth and health status data particularly valuable. It would be beneficial to be able to predict which beach to focus our search time on, to increase the number of recaptured individuals.

Prediction Models: Our dataset consists of a large table of 19,219 horseshoe crabs movements over 20 years. Each record contains information about one horseshoe crab movement: Initial longitude and latitude, initial date, final longitude and latitude, final date and sex. Our goal is to be able to predict the final longitude and latitude from the initial longitude and latitude, sex, and initial date tagged. Table 2 shows the correlation matrix of the final longitude and latitude in relation to the other fields.

Table 2. Correlation matrix of the final longitude and latitude.

	Final longitude	Final latitude
Initial longitude	0.9408	0.7755
Initial latitude	0.8008	0.8722
Initial date	0.1146	0.0973
Is male	−0.0907	−0.0617
Is female	0.0963	0.0667

The table shows a high correlation between the final location and the initial location data and a weak correlation with the initial date and gender (~0.1). This implies that our prediction model will heavily rely on these two fields (initial longitude and initial latitude) and we can ignore the other fields with little change in results.

Three different regression models were experimented with for our dataset: Linear Regression, Decision Trees and Random Forests. Ordinary Least Squares Linear Regression [17] is the simplest regression model as it assumes a simple direct linear relationship between the output data (Final longitude and latitude) and the input variables, and aims to minimize the square error between the prediction line and the input data.

A decision tree predictor [18] is another popular technique based on recursively building a decision tree by selecting the best attribute to split the data into two parts. Then it recursively repeats the process on each part until either the data size is zero or all tuples belong to one group, or there are no more attributes to split on. This tree is then used to predict the new data instance values. Figure 4 shows a 3-level decision tree for prediction the final longitude, and it shows clearly that the preferred splitting attribute is the Initial Longitude (ILong) as expected from the correlation matrix (Table 2). A typical decision tree would have tens of levels for good accuracy. We show a 3-level decision tree just for clarity. Although simple and clear to trace decision trees are not optimal due to several factors: decision trees are sensitive to the specific data on which they are trained. So, they tend to over-fit the training data and a small change in the training data would result in a different decision tree and a different prediction. Also, decision trees are computationally expensive to train, carry a big risk of overfitting, and tend to find local optima because they cannot go back after they have made a split.

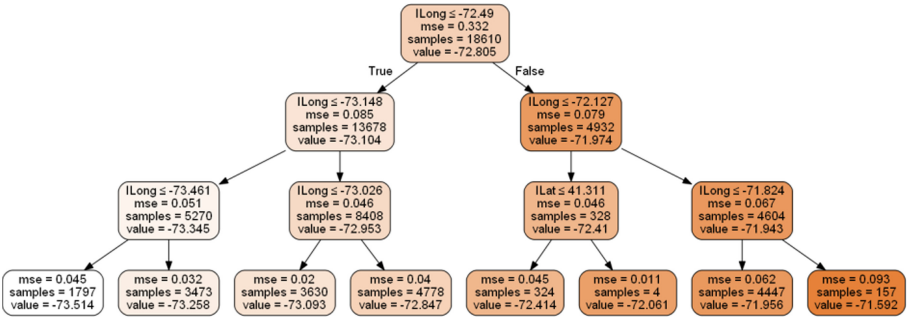

Fig. 4. A small 3-level decision tree regression model.

The third technique we used is the Random Forest Ensemble Regression model [19]. An Ensemble method aggregates the output of several similar or different machine learning techniques to produce an output that is better than using any individual method. A random forest regression ensemble is a meta-estimator which aggregates many decision trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees, as shown in Fig. 5.

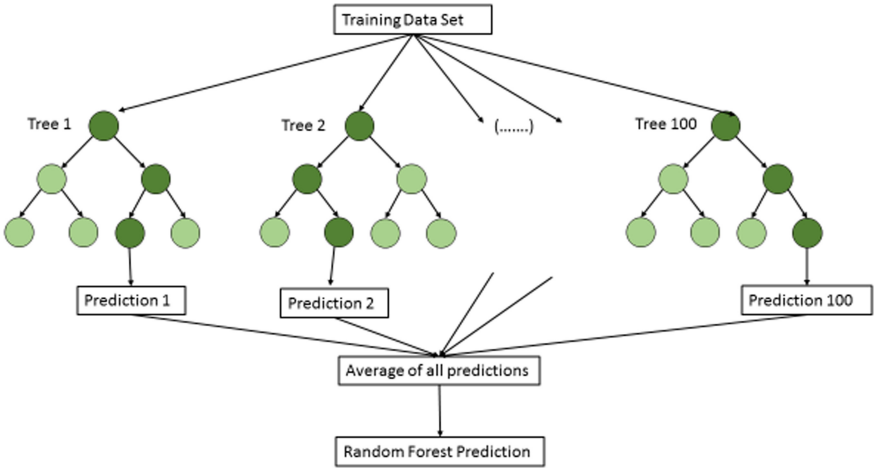

Fig. 5. A simplified version of the random forest regression model.

Two modifications are made to keep the individual decision trees from being too similar: First, the number of features that can be split on at each node is limited to some percentage of the total. This ensures that the ensemble model does not rely too heavily on any individual feature, and makes fair use of all potentially predictive features. Second, each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness that prevents overfitting.

3 Results and Discussion

We tested all three models using the data for the 2018 season (607 movements) as our test value and all the seasons before that as the training value (18,612 movements). A sample output is displayed in Fig. 6. The solid black arrow shows the actual movement of a particular horseshoe crab. The red dot is the location estimate produced by linear regression model, the green one is the prediction of the decision tree regression model, and the blue dot represents the prediction of the random forest predictor. In general, the random forest predictor gave the best performance for most horseshoe crabs.



Fig. 6. Sample prediction output. (Color figure online)

A model would be very useful to predict the approximate area where horseshoe crabs would be found, but would not accurately predict the location of a particular horseshoe crab. Figure 7 shows the predictor results for all 607 horseshoe crabs in the test set. The black dots represents the actual final locations of the horseshoe crabs. The Red dots represent the linear regression, the green dots represent the decision tree predictor, and the blue dots represent the results of the random forest predictor. All three predictors did not perform well with the data points on the south side of the Sound (i.e. Long Island's North shore). This is due to a general lack of data points there in both the training and test sets.

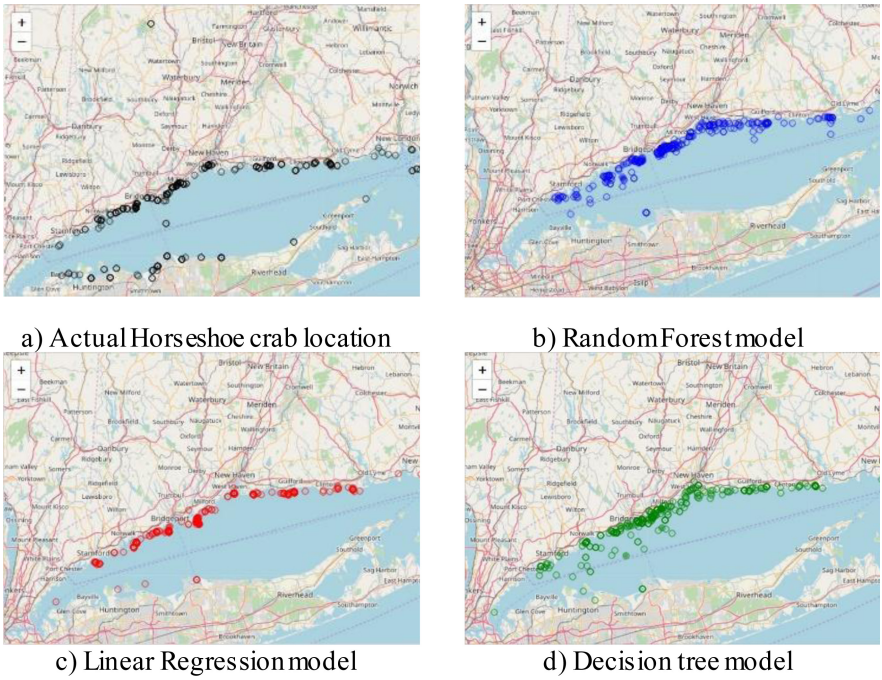

Fig. 7. Collective result of prediction for all three models. (Color figure online)

Table 3 shows the results for the 2018 season prediction based on all previous seasons. We tried two different methods: predict both final longitude and latitude together, or predict each one alone and average the error. An interesting observation is that all techniques performed better in predicting the latitude, compared to the longitude. This is due to the fact that the latitude range is much smaller than the longitude range for the LIS region. Also, we observe that we get better results when we predict the longitude and latitude separately. The least error was obtained when using the Random forest regression model and when predicting the longitude and latitude separately. Based on these results, we conclude that the random forest regression model is the most suitable for our prediction model, and predicting the longitude and latitude individually is slightly better than running the model on them both together.

Table 3. Performance results for the three prediction models.

	Least square linear regression model		Decision tree regression model		Random forest regression model with 100 estimators	
	Max error	MAE	Max error	MAE	Max error	MAE
Predict longitude and latitude together	1.555	0.098	1.616	0.099	1.433	0.097
Predict longitude only	2.399	0.143	2.487	0.163	2.270	0.141
Predict latitude only	0.712	0.053	0.746	0.055	0.662	0.050
Average of both	1.555	0.098	1.616	0.109	**1.466**	**0.096**

Performance Analysis: The dataset size and composition greatly affects the performance results. An interesting observation is the varied performance of the random forest prediction model results over the years, as the dataset grows in both size and geographical area. Table 4 shows the random forest prediction model's RMSE results over the past ten years. For each year, we used that year's data as the test set and the data of all the previous years as the training data. The RMSE is the average RMSE for predicting the longitude and latitude separately. Similar results were obtained for the MAE.

Table 4. Random forest prediction model results for the years 2008–2018.

Test year	Train set size	Test set size	Percentage of testing set	RMSE	MAE
2008	1163	755	39%	0.146	0.068
2009	1918	1860	49%	0.121	0.049
2010	3778	3164	46%	0.097	0.043
2011	6942	3566	34%	0.103	0.045
2012	10508	2964	22%	0.112	0.055
2013	13472	1943	13%	0.156	0.077
2014	15415	1054	6%	0.194	0.113
2015	16469	785	5%	0.204	0.116
2016	17254	818	5%	0.187	0.110
2017	18071	538	3%	0.173	0.103
2018	18609	607	3%	0.183	0.096

Interestingly, the table could be divided into three areas with different prediction results: 2008–2010 the prediction error decreases steadily, 2011–2015 the prediction error increases, and then in 2016–2018 the prediction error showed a gradual descent over time.

The prediction error was decreasing from 2008 to 2010 as expected, as the tagging/recapture activity was mainly concentrated on the north central part of LIS. However, as the dataset changed over time by having a larger geographical area to cover all of the LIS region after 2011, the prediction results had a larger error, as the training set mainly contained data from the northern shores of LIS area and the testing sets started to have more data from the western and southern shores. Those new data points had no similar movement history in the training set and therefore most of them were predicted with a huge error (2011–2015). Then, as more and more data points in the western and southern parts of LIS were added, we observed a decrease in the prediction error (2015 - present), as the training set gets more varied year after year to cover the entire LIS area. We expect their prediction error to decrease in future seasons as more data is added. Also, we recommend more capture activity in the southern and western parts of the Sound to be able to gather more movement data there.

Based upon our experimental results, we conclude that the best prediction model for horseshoe crab movement in LIS is the random forest prediction model with 100 estimators. Also, the results have shown that performance is expected to improve over the years as more data points are added over the seasons.

4 Discussion

The prediction model would be used for two purposes: First, to determine the best locations for the tagging/recapture teams to search for horseshoe crabs, thus saving their time and effort. Also, particular horseshoe crabs are of interest, those who have been tracked for the longest time, or have traveled long distances, or have a particular interesting feature. The model can predict the best location to find that individual horseshoe crab as shown in Figs. 6 and 7.

A weakness of the prediction model is its dependency on the dataset being used, as is common with all machine learning techniques. The more data that is available, the better the prediction. There has been higher tagging/recapture activity on the CT or north shore of LIS than the NY or south shore. This easily explains why the model prediction rate is better in CT. We recommend an increased tagging/recapture effort in the south on Long Island to get a better picture of the movement of horseshoe crabs there and improve our predictions for all of LIS.

While movement inside LIS is a useful ecological study for the management of horseshoe crabs, a wider look at the movement of horseshoe crabs in New England would be extremely useful in the study of the mortality and movement of the American horseshoe crab. Currently, research into the movement of horseshoe crabs in this area is plentiful but highly localized [7, 11, 12, and 16].

5 Conclusion

Animal movement prediction is a vital aspect of animal tracking using tags. It enables researchers to predict an approximate location they are expected to be in, thus enabling a better chance of re-capture of the tagged animals, and locating more members of the general population. Project *Limulus* has been monitoring and studying the population

of horseshoe crabs in Long Island Sound since 1998. Predicting the optimal beaches to target for tagging horseshoe crabs would be useful for the project to enable the capture of more horseshoe crabs per season. Also, some horseshoe crabs have been captured multiple times over several years and are of particular interest for revealing movement patterns over time. Therefore, it is of great value to recapture those particular crabs.

In this paper, we present a horseshoe crab location predictor for estimating the locations of horseshoe crabs for the next season, based on previously observed movements over the years. We tried three different regression models: Ordinary Least Square Linear regression, Decision Tree regression, and Random Forest Ensemble regression. Based on our results, we recommend the Random Forest regression model with 100 estimators as it gave the best predictions compared to the others. The measure of accuracy used were the root mean squared error (RMSE) and the mean absolute error (MAE). The prediction model proved accurate when the 2018 summer capture season was used as a test set. The expectation is that the predictions will get more accurate as the data set grows over the seasons.

Acknowledgment. We thank the US Fish & Wildlife Service for their support with the tagging program, CT Sea Grant, Disney Conservation Fund, The Nature Conservancy, Audubon CT, CT Audubon Society, Sacred Heart University undergraduate Biology Majors, and the College of Arts and Sciences for funding and time supporting this project.

References

1. Allen, A., Singh, N.: Linking movement ecology with wildlife management and conservation. Front. Ecol. Evol. **3**, 155 (2016). https://doi.org/10.3389/fevo.2015.00155
2. Fraser, K., Davies, K., Davy, C., Ford, A., Flockhart, D., Martins, E.: Tracking the conservation promise of movement ecology. Front. Ecol. Evol. **6**(150), 1–8 (2018)
3. Benson, E.: Wired Wilderness: Technologies of Tracking and the Making of Modern Wildlife. The Johns Hopkins University Press, Baltimore (2010)
4. Peniche, G., Vaughan-Higgins, R., Carter, I., Pocknell, A., Simpson, D., Sainsbury, A.: Long-term health effects of harness-mounted radio transmitters in red kites (Milvus milvus) in England. Vet. Rec. **169**, 311 (2011)
5. Li, Z.: MoveMine: mining moving object data for discovery of animal movement patterns. ACM Trans. Intell. Syst. Technol. (TIST) **2**(4), 1–32 (2011). Article no. 37
6. Krisfalusi-Gannon, J., et al.: The role of horseshoe crabs in the biomedical industry and recent trends impacting species sustainability. J. Front. Mar. Sci. **5**, 185 (2018)
7. Mattei, J., Kasinak, J., Senbel, S., Bartholomew, K.: The power of citizen science: 20 years of horseshoe crab community research merging conservation, education and management. In: Tanacredi, J.T., et al. (eds.) 4th Proceedings of Horseshoe crab Conservation and Biology, Springer-NATURE (2021, in press)
8. Beekey, M., Mattei, J., Pierce, B.: Horseshoe crab eggs: a rare resource for predators in Long Island Sound. J. Exp. Mar. Biol. Ecol. **439**, 152–159 (2013)
9. Project *Limulus*. www.projectlimulus.org. Accessed 30th Aug 2020
10. Bakibayev, T., Kulzhanova, A.: Common movement prediction using polynomial regression. In: 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan, pp. 1–4 (2018)

11. Bopp, J., Sclafani, M., Smith, D., et al.: Geographic-specific capture-recapture models reveal contrasting migration and survival rates of adult horseshoe crabs (*Limulus polyphemus*). Estuaries Coasts **42**, 1570–1585 (2019)
12. James-Pirri, M.: Seasonal movement of the American horseshoe crab Limulus polyphemus in a semi-enclosed bay on Cape Cod, Massachusetts (USA) as determined by acoustic telemetry. Curr. Zool. **56**(5), 575–586 (2010)
13. Wada, T., Mitsushio, T., Inoue, S., Kioke, H., Kawabe, R.: Movement patterns and residency of the critically endangered horseshoe crab Tachypleus tridentatus in a semi-enclosed bay determined using acoustic telemetry. PLoS ONE **11**(2), e0147429 (2016)
14. Youssef, I., Senbel, S., Kasinak, J.E., Mattei, J.: RCrab: an R analytics tool to visualize and analyze the movement of horseshoe crabs in long island sound. In: 2019 IEEE Long Island Systems, Applications and Technology Conference, NY, USA, pp. 1–6 (2019)
15. Van Rossum, G.: Python tutorial, Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995. http://www.python.org. Accessed 5th Sep 2020
16. Swan, B.: Migrations of adult horseshoe crabs, Limulus polyphemus, in the middle Atlantic bight: a 17-year tagging study. Estuaries **28**, 28–40 (2005)
17. Goldstein, H.: Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika **73**(1), 43–56 (1986)
18. Shalev-Shwartz, S., Ben-David, S.: Chapter 18. Decision trees. In: Understanding Machine Learning. Cambridge University Press (2014)
19. Ho, T.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp. 278–282, August 1995