War of the Wilds: Future Invaders and Invasions

Nicholas Lu
The Haverford School, Haverford, PA 19041
nichlu@haverford.org

Abstract

Invasive plant species can change an ecosystem's food web by destroying or replacing native food sources, providing little to no food value for wildlife. They are also able to alter the abundance of diversity of species that are important habitats for native wildlife. Some aggressive species are even capable of changing ecosystem conditions, from soil chemistry to the intensity of wildfires. This project uses several algorithms to understand the factors that contribute to the spread of invasive species. Our model uses three different algorithms: One-Way Analysis of Variance (ANOVA), Linear Regression (LR) as well as Nonlinear Regression (NR) to examine the characteristics of the ecological profile of current invasive species, specifically its spread and length in time it has had to spread. Based on statistical profiling of 77,075 records of 75 invasive species, over 10% of the species have matched the profile of known invasive plants, which are likely to become the next global invaders. These results present an opportunity to implement timely and proactive management strategies against biological invasions.

Keywords:
Machine Learning, Invasive Species, Linear Regression, ANOVA, Nonlinear Regression

1. Introduction
The regions under study were Eastern and Southern Africa, whose ecosystems, while officially protected, are currently exposed to invasion by invasive plant species (Obiri 2011). In Eastern and Southern Africa, the spread of invasive species is impacting the livelihoods of the rural poor who are dependent on natural resources for income and survival, as well as undermining foreign development investment in the area. Given that their dependency on natural resources has increased more significantly over the last decade (Shackleton et al. 2000, and Witt et al. 2010), efforts to understand the root causes of the growth and spread of invasive species are needed to be proactive about their current and future survival. However, there has been a limited study on what promotes or prevents the growth of these invasive species in those regions. Therefore, using a set of algorithms, this study investigates how time and the type of the invasive species control their behavior.

Three different machine learning algorithms, one-way ANOVA, Linear Regression (LR), and Nonlinear Regression (NR) will be performed in order to uncover the nature of the impact that the types of the invasive species and time have on the size of their spread.

Minimal studies have attempted to precisely model and predict the growth or spread of invasive plants, specifically in regards to using methods such as Machine Learning (Schneider et al. 2021). However, the vast majority of studies seem to attempt to predict events or scenarios that may come as a result of invasive plants (Faccenda et al. 2021) or seek to predict the distribution of invasive plants in either a highly general or specific area (Ahmed et al. 2020, and Jones et al. 2010). As a result, this paper seeks to understand (1) whether the size of spread of invasive plants varies by the types of species, and (2) whether time factor predicts their growth differently. More specifically, this study explores whether the relationship between time and the size of the spread of invasive species is curvilinear, such that spread increases precipitously early on and starts to slow down after some time. The hypothesized nonlinear relationship is further probed by using the types of the species as a moderator, such that the nonlinear relationship between the time and the size of spread varies by the types of the species.

2. Literature Review
Previous studies on predicting the spread of invasive plants or determining key factors, either to determine the relationship between temperature on growth of invasive plants (Skálová et al. 2015) or to predict future locations of invasive plants through analyzing the ecological temperature (Ebeling et al. 2008).

Plants spread by dispersing their seeds using everything from fire to animals. Some studies have already measured the quantity and quality of seed dispersal by animals in general (Schupp 1993), and others focusing on measuring more specific animals such as birds (Gosper et al. 2005). Similarly, studies have been written surrounding the spread of invasive plants in relation to fire, specifically fire's ability to create new habitats (Maringer et al. 2012), which further presents another way plants can spread. Invasive species are known to spread in different ways. More specifically, some invasive species have particularly aggressive root systems that spread distances from a single plant and release chemicals to harm the seeds of other plants (Ens et al. 2008), whereas some other invasive species may grow so densely that they smother the species around them (Burke et al. 2011).

Thus, this study aims to support the hypotheses that not only were the variance on the growth of the spread differed by the types of the invasive species, but that the time factor was significantly related to the size of the spread as evidenced by both linear and nonlinear regression model. While the findings of the linear model initially presented in the current paper confirmed the role of time as a significant factor, findings further indicated that a nonlinear association significantly better predicts the growth pattern, which is a key contribution of this study to the literature.

3. Methods

3.1 Project Research 1

To ensure that the data meet the minimum assumptions of parametric statistics, several tests were performed. First, the histograms (Figures 1 and 2) of variables "time" and "spread" were used in order to determine whether or not the data was normalized. Then, Skewness statistic and Kurtosis statistic was run on the data to determine if it deviated significantly from normal distribution. For a regression model, a multicollinearity issue was inspected using Tolerance and VIF (variance inflation factor).

For the project, one-way ANOVA and a multiple regression model were used to understand the factors that contribute to the spread of invasive species, and to identify the factors that contribute to where the spread is most likely to occur. The programming language R was used for statistical computing and graphics. In particular, the packages caret, dplyr, rpart, e1071, tidyverse, kernlab, and splitstackshape were used.
The dependent variable, the plant spread, will measure the distance between the current plant and the first plant of its species recorded. As a result, the total distance that the plant has spread from the time it was first recorded was cataloged.

One independent variable was a categorical variable indicating a total of 75 different types of species of invasive plants. The one-way ANOVA will determine whether the size of spread varies by different types of species and where the differences may be statistically significant.
The second independent variable is length in time from when each plant was first discovered to the day the current plant was cataloged. As a result, the exact time interval since the initial plant was introduced was determined. One-way ANOVA was utilized to interpret the impact that the different species of plant would have on the distance said species spread.

To test whether the time elapsed had an impact on how far the plant spread, a moderated curvilinear regression model was performed in a hierarchical fashion. First, a linear regression (the baseline model) was performed followed by a nonlinear model in the second step to examine whether a quadratic equation better describes the relationship between time and the size of spread.

3.2 Hypotheses

The present study hypothesizes that the different species of invasive plants impact the spread of invasive plants differently.

Hypothesis 1: The types of species of invasive plants impact the spread of invasive plants differently.
Furthermore, for a plant to spread, it takes time for the parent plant to grow to completion and for the seeds to spread, and can be affected by outside factors such as temperature and the concentration of (Morison et al. 999). Given that invasive species require certain time to adjust in a new environment, it can be argued that the size of its growth and thus its spread will be more pronounced after several years before starting to make significant impact in the surrounding areas (Hastings et al. 2004, and Horgan-Kobelski et al. 2012). Thus, this paper hypothesizes that the length in time elapsed will have a nonlinear impact on the spread of invasive plants.

Hypothesis 2: The length in time elapsed will have a nonlinear impact on the spread of invasive plants.

Since the size of spread of invasive species may differ depending on the type of the species (Coutts et al. 2010), the nonlinear regression model is further probed by exploring the role of the types of the invasive species as the moderator. Thus, the following hypothesis is tested:

Hypothesis 3: The nonlinear relationship between the length in time elapsed and the spread of invasive plants will be moderated by the type of species of the invasive plants.

3.3 Data Collection 1

The CABI Africa Invasive and Alien Species dataset was used, which holds over 77,00 records of over 70 invasive alien species from Ethiopia to Uganda. The data were originally collected by a group of researchers from the University of Chicago through roadside surveys over the course of fieldwork trips over a 43-year period from 1974 to 2017, leading to a total of 77,075 observations of invasive plants which they wrote a study about (Witt et al. 2018). The dataset was first sorted into subsets based upon the species of plant, to group plants of the same species together. Afterwards, the subsets were further sorted by the date the plant was first discovered.

4. Data and Statistical Analyses

4.1 Data Presentation

Table 1. Description of Independent Variables

Variable	Destruction	Type
verbatimScientificName	The species of the plant	Character
EventDate	The date the plant was discovered and cataloged, which will be used to determine TimeElapsed	Character
decimalLatitude	The coordinate latitude of the plant in decimals, which will be used to determine DistanceSpread	Float
decimalLongitude	The coordinate longitude of the plant in decimals, which will be used to determine DistanceSpread	Float

TimeElapsed	The length in time between the current plant and the first cataloged plant of its species	Integer

Descriptive statistics of the variable are presented in Table 2.

Table 2. Description of Independent Variables

Variable	Mean	Median	Range	Standard Deviation
DistanceSpread	13.312	13.360	34.820	7.037
TimeElapsed	2090	1974	7764	1184.808

Due to its nature as a categorical variable, the frequency of verbatim Scientific Name is shown in Table 3.

Table 3. Frequency Reporting of Invasive Plants by Type

Name	Number	Percent (%)
Agave sisalana	N = 1608	6.169%
Azadirachta indica	N = 1393	5.344%
Calotropis procera	N = 1035	3.971%
Cascabela thevetia	N = 1821	6.986%
Datura stramonium	N = 1001	3.840%
Euphorbia tirucalli	N = 1139	4.370%
Grevillea robusta	N = 1460	5.601%
Jacaranda mimosifolia	N = 1004	3.852%
Lantana camara	N = 3330	12.775%

Leucaena leucocephala	N = 1559	5.981%
Parthenium hysterophorus	N = 1178	4.519%
Psidium guajava	N = 927	3.556%
Ricinus communis	N = 1962	7.527%
Senna didymobotrya	N = 902	3.460%
Senna occidentalis	N = 1100	4.220%
Senna siamea	N = 2250	8.632%
Senna spectabilis	N = 953	3.656%
Tithonia diversifolia	N = 1443	5.536%
Agave sisalana	N = 1608	6.169%
Azadirachta indica	N = 1393	5.344%
Calotropis procera	N = 1035	3.971%
Cascabela thevetia	N = 1821	6.986%
Datura stramonium	N = 1001	3.840%
Euphorbia tirucalli	N = 1139	4.370%
Grevillea robusta	N = 1460	5.601%
Jacaranda mimosifolia	N = 1004	3.852%
Lantana camara	N = 3330	12.775%
Leucaena leucocephala	N = 1559	5.981%

Parthenium hysterophorus	N = 1178	4.519%
Psidium guajava	N = 927	3.556%
Ricinus communis	N = 1962	7.527%
Senna didymobotrya	N = 902	3.460%
Senna occidentalis	N = 1100	4.220%
Senna siamea	N = 2250	8.632%
Senna spectabilis	N = 953	3.656%
Tithonia diversifolia	N = 1443	5.536%

4.2 First Statistical Tool

One-way ANOVA is most often used to determine if scores of the outcome variable differ significantly between three or more groups.

Table 4. One Way ANOVA Results

Variable	DF	Sum Sq	Mean Sq	F Value	Pr(>F)
verbatimScientificName	18	527406	29300	999.9	p<0.001
Residuals	26048	763275	29		

The one-way ANOVA result showed that the size of spread significantly differs by the types of invasive plants ($p < .001$). Thus, there is significant evidence that different species of invasive plants have different methods of spread, with some spreading faster than others. Post hoc analysis was performed to identify where the differences in the size of spread were statistically significant. The size of spread appears to differ significantly between most of the comparisons. There were several exceptions. For example, The size of spread was not significantly different between Azadirachta indica and Calotropis procera ($p = .699$) and between Cascabela thevetia and Senna siamea ($p = .997$).

4.3 Second Statistical Tool (see section 3.2)

To test whether the time elapsed had an impact on how far the plant spread, a moderated curvilinear regression model was performed in a hierarchical fashion. First, a linear regression (the baseline model) was performed followed by a nonlinear model in the second step to examine whether a quadratic equation better describes the

relationship between time and the size of spread. The result of the baseline model (see below) showed that Time Elapsed was a significant predictor.

However, after running the nonlinear regression and moderated nonlinear regression, it was determined that the quadratic regression model was more accurate. Thus, it appears that invasive plants spread much faster over time, possibly due in part to having more plants to spread.

Table 5. Results of Hierarchical Regression Analysis

	Dependent Variable		Adjusted
Step 1:	Distance spread	0.1555	0.1555
Step 2:	Distance spread squared	0.1689	0.1688

Due to the significant positive correlation between TimeElapsed and DistanceSpread, it can be interpreted that the distance an invasive plant species spreads increases as the length in time increases. Table 6 shows the size of the coefficient (b = .0023), which indicates that one-day growth is associated with .0023 kilometers.

Due to the significant positive correlation between TimeElapsed and DistanceSpread, it can be interpreted that the distance an invasive plant species spreads increases as the length in time increases. Table 6 shows the size of the coefficient (b = .0023), which indicates that one-day growth is associated with .0023 kilometers.

After analyzing the data, both the one-way ANOVA test and the quadratic regression model showed significant relationship between the variables. The type of invasive plant and the time the invasive plant's species has had to spread were both determined to be significant predictors on how far it will spread. Additionally, the quadratic regression model proved to be far more accurate in determining the relationship between the time elapsed and the spread of the invasive plant than the linear model.

The significance of both of the independent variables makes sense and supports the hypothesis. For the first, as different plants spread in different ways, different invasive plants would spread differently. Also, the more time a plant species has had to spread, the more plants would be able to spread further seeds of their species. While the quadratic regression model proved to be more accurate in determining the relationship between the time elapsed and the spread than the linear regression model, it lent credence to the theory that more invasive plants would continue to spread more seeds, increasing their spread exponentially.

5. Conclusions

This paper examined whether the type of the invasive plants and time factor controlled their growth behavior, using one-way ANOVA, linear regression, and moderated nonlinear regression. Via each technique, the paper was able to determine the relationship and the strength between the type of plant and time their species has had to spread to the distance they have spread. The paper determined that there is strong proof that the type of plant and the time elapsed factors into how far the invasive plant has spread, along with the fact that a nonlinear quadratic regression model more accurately represents the relationship between time elapsed and the distance the invasive plant spreads.

6. Future Research

Through the findings, the paper not only determined whether or not different types of invasive plants and the time they have had to spread impact the distance spread, but it also determined the extent to which each time factor impacts the distance spread. Using this information, the most invasive plants, or the ones that spread the farthest

fastest, are able to be determined, creating profiles for highly invasive plants which may become potential threats in the future. Future work could, knowing the independent variables and their relation to the spread of invasive plants, seek to predict location at risk of invasive plants. Additionally, broader datasets from across the world could also be used to determine where invasive plants are likely to spread.

7. References

Ahmed, N., Atzberger, C., & Zewdie, W., Integration of remote sensing and bioclimatic data for prediction of invasive species distribution in data-poor regions: a review on challenges and opportunities. *Environmental Systems Research*, *9*(1),2020. https://doi.org/10.1186/s40068-020-00195-0

Burke, J. M., & DiTommaso, A., Corallita (Antigonon leptopus): Intentional Introduction of a Plant with Documented Invasive Capability. *Invasive Plant Science and Management*, *4*(3), 265–273, 2011. https://doi.org/10.1614/ipsm-d-10-00088.1

Coutts, S. R., van Klinken, R. D., Yokomizo, H., & Buckley, Y. M., What are the key drivers of spread in invasive plants: dispersal, demography or landscape: and how can we use this knowledge to aid management? *Biological Invasions*, *13*(7), 1649–1661,2010. https://doi.org/10.1007/s10530-010-9922-5.

Ebeling, S. K., Welk, E., Auge, H., & Bruelheide, H., Predicting the spread of an invasive plant: combining experiments and ecological niche model. *Ecography*, *31*(6), 709–719,2008. https://doi.org/10.1111/j.1600-0587.2008.05470.x

Ens, E. J., Bremner, J. B., French, K., & Korth, J., Identification of volatile compounds released by roots of an invasive plant, bitou bush (Chrysanthemoides monilifera spp. rotundata), and their inhibition of native seedling growth. *Biological Invasions*, *11*(2), 275–287, 2008. https://doi.org/10.1007/s10530-008-9232-3

Faccenda, K., & Daehler, C. C., A screening system to predict wildfire risk of invasive plants. *Biological Invasions*, *24*(2), 575–589, 2021. https://doi.org/10.1007/s10530-021-02661-x

Gosper, C. R., Stansbury, C. D., & Vivian-Smith, G., Seed dispersal of fleshy-fruited invasive plants by birds: contributing factors and management options. *Diversity <Html ent Glyph="@amp;" Ascii="&Amp;"/> Distributions*, *11*(6), 549–558, 2005. https://doi.org/10.1111/j.1366-9516.2005.00195.x

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., Moore, K., Taylor, C., & Thomson, D., The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters*, *8*(1), 91–101, 2004. https://doi.org/10.1111/j.1461-0248.2004.00687.x

Jones, C. C., Acker, S. A., & Halpern, C. B., Combining local- and large-scale models to predict the distributions of invasive plant species. *Ecological Applications*, *20*(2), 311–326, 2010. https://doi.org/10.1890/08-2261.1

Maringer, J., Wohlgemuth, T., Neff, C., Pezzatti, G. B., & Conedera, M., Post-fire spread of alien plant species in a mixed broad-leaved forest of the Insubric region. *Flora - Morphology, Distribution, Functional Ecology of Plants*, *207*(1), 19–29, 2012. https://doi.org/10.1016/j.flora.2011.07.016

MORISON, J. I. L., & LAWLOR, D. W., Interactions between increasing CO2 concentration and temperature on plant growth. *Plant, Cell and Environment*, *22*(6), 659–682, 1999. https://doi.org/10.1046/j.1365-3040.1999.00443.x

Obiri, J. F., Invasive plant species and their disaster-effects in dry tropical forests and rangelands of Kenya and Tanzania. *Jàmbá: Journal of Disaster Risk Studies*, *3*(2), 2011. https://doi.org/10.4102/jamba.v3i2.39

Schneider, K., Makowski, D., & van der Werf, W. (2021). Predicting hotspots for invasive species introduction in Europe. *Environmental Research Letters*, *16*(11), 2021. 114026. https://doi.org/10.1088/1748-9326/ac2f19

Schupp, E. W., Quantity, quality and the effectiveness of seed dispersal by animals. *Vegetatio*, *107–108*(1), 15–29, 1993. https://doi.org/10.1007/bf00052209

Shackleton, C. M., McGarry, D., Fourie, S., Gambiza, J., Shackleton, S. E., & Fabricius, C. (2006). Assessing the Effects of Invasive Alien Species on Rural Livelihoods: Case Examples and a Framework from South Africa. *Human Ecology*, *35*(1), 113–127. https://doi.org/10.1007/s10745-006-9095-0

Skálová, H., Moravcová, L., Dixon, A. F. G., Kindlmann, P., & Pyšek, P. (2015). Effect of temperature and nutrients on the growth and development of seedlings of an invasive plant. *AoB PLANTS*, *7*https://doi.org/10.1093/aobpla/plv044

Sultan, S. E., Horgan-Kobelski, T., Nichols, L. M., Riggs, C. E., & Waples, R. K., A resurrection study reveals rapid adaptive evolution within populations of an invasive plant. *Evolutionary Applications*, *6*(2), 266–278, 2012. https://doi.org/10.1111/j.1752-4571.2012.00287.x

Witt, A. B. R. (2010). Biofuels and invasive species from an African perspective - a review. *GCB Bioenergy*, *2*(6), 321–329. https://doi.org/10.1111/j.1757-1707.2010.01063.x

Witt, A., Beale, T., & van Wilgen, B. W. (2018). An assessment of the distribution and potential ecological impacts of invasive alien plant species in eastern Africa. *Transactions of the Royal Society of South Africa*, *73*(3), 217–236,2018. https://doi.org/10.1080/0035919x.2018.1529003

8. Biography

Nicholas Lu is a rising sophomore at The Haverford School. His research interests encompass the application of artificial intelligence and machine learning methods to address the needs of new climate solutions by developing methods for land management practices, water security, environmental health and justice, preventing extinction and optimizing nature for human health and wellbeing. His other interests include environmental philosophy, where he ponders issues of the moral relationships between human beings and nature, and what we owe future generations when it comes to the natural world. He enjoys using philosophical frameworks including ethical and political theories to engage with environmental challenges of our time.