Fergusson College
Department of Statistics

A CRISIS LIKE NEVER BEFORE

Deccan Education Society's
Fergusson College (Autonomous), Pune
Department of Statistics
STS 3609 - Statistics Practical III

CERTIFICATE

This is to certify that, Mr./Ms. _____ with roll number _____ has satisfactorily completed the project work entitled _____ towards partial fulfilment of B.Sc Statistics, Semester VI course during the academic year 2022-2023.

Place : Pune

Date: /04/2023

(Name and signature of the guide) (Name and signature of the HOD)

Internal Examiner External Examiner

CONTENTS

GROUP MEMBERS		
	Roll Number	Name
1	213111	Harshal Mandawade
2	213121	Shreyas Yadav
3	213153	Saee Kurhade
4	213242	Dhanshree Khetre
5	213267	Sakshi Jain
6	213285	Shreyash Patil

Guided by:

Dr. Subhash Shende

Head, Department of Statistics
Fergusson College

ACKNOWLEDGEMENT

We would like to take this opportunity to thank all those who are responsible for the completion of this project. Our sincere thanks to our project guide and Head of the Department of Statistics Dr. Subhash Shende for his valuable guidance and support in all the stages of our project work, his meticulous insights helped us to improve the quality of the project we have been making. Our project was greatly benefited by his statistical knowledge and experience which were useful at the critical stages of the project. This project could not have been possible without him. We would also like to thank Mrs. Deepa Kulkarni for her valuable guidance for our project. In addition, we would also like to thank the Department of Statistics, Fergusson College for providing necessary assistance, infrastructure and facilities required. We extend our heartfelt gratitude to all the people who directly or indirectly are involved with our project.

INTRODUCTION

Biodiversity, or the variety of life on Earth, is essential for maintaining healthy ecosystems and providing the resources that sustain human societies. However, biodiversity is under threat from a range of factors, including habitat loss, climate change, pollution, and over-exploitation. To address these threats and promote the conservation of biodiversity, a number of organizations and initiatives have been established, including the International Union for Conservation of Nature (IUCN) and its Red List of Threatened Species. The IUCN Red List is the world's most comprehensive inventory of the conservation status of plant and animal species. It provides a systematic approach to evaluating the extinction risk of species based on a range of criteria, including population size, geographic range, and habitat degradation. The Red List serves as a key tool for identifying species that require conservation action, and for monitoring trends in species populations over time. Conservation efforts based on the Red List have led to the recovery of some species, but many remain under threat. As such, there is a continued need for monitoring and management of biodiversity, as well as for research into the drivers of biodiversity loss and the effectiveness of conservation measures. Through its efforts to promote sustainable development and conservation of biodiversity, the IUCN and the Red List play a critical role in ensuring the continued survival of Earth's diverse and valuable species. In this project, we aim to estimate and quantify the extinction risk faced by a particular species and forcast their upcoming conservation status. Also to investigate the public awareness and attitudes towards biodiversity conservation and threatened species through a primary survey. Through our project we assessed 'The Wildlife Protection Amendment Act (2002)' which had been enacted by the Parliament of India, was effective in the recovering the population of Bengal Tigers.

ABSTRACT

This project aims to investigate the relationship between species diversity and conservation status using data from the **International Union for Conservation of Nature(IUCN)** Red List. We analyzed the distribution of species across different Red List categories (i.e., Least Concern, Vulnerable, Endangered, Critically Endangered, etc.) and examined the factors that influence a species' likelihood of being classified as threatened. This project investigates the trends and predictions of species conservation status using time series analysis and classification models. We applied exponential smoothing to Red List data to identify patterns in the distribution of species across different conservation categories over time. Our analysis revealed that the number of species classified as threatened has been steadily increasing in recent years. To predict future trends, we used classification models to assess the likelihood of species moving between different conservation categories. Also we have estimated the conservation status and quantified the extinction risk faced by a particular species using the concept of Markov chain followed by computation of the one-step transition probability matrix. Our results suggest that conservation efforts may be most effective when targeted at species that are currently classified as vulnerable, as these are at the greatest risk of becoming endangered or extinct. We have also quantified the effects of certain laws and regulations that have been implemented for biodiversity conservation. We have assessed **The Wildlife Protection Amendment Act (2002)** for recovering the population of a particular species (here, Bengal Tiger) by employing the technique Difference-in-Difference analysis. Additionally, we conducted a primary survey of the general public to gauge awareness and attitude towards biodiversity conservation.Our findings indicate that while the majority of respondents recognize the importance of conservation. Overall, this project provides insights into the current state of species conservation and highlights the importance of proactive conservation measures to protect threatened species and preserve global biodiversity. Following are some of the terminologies:

1.IUCN Red List

A comprehensive database of the conservation status of species around the world, maintained by the International Union for Conservation of Nature (IUCN).

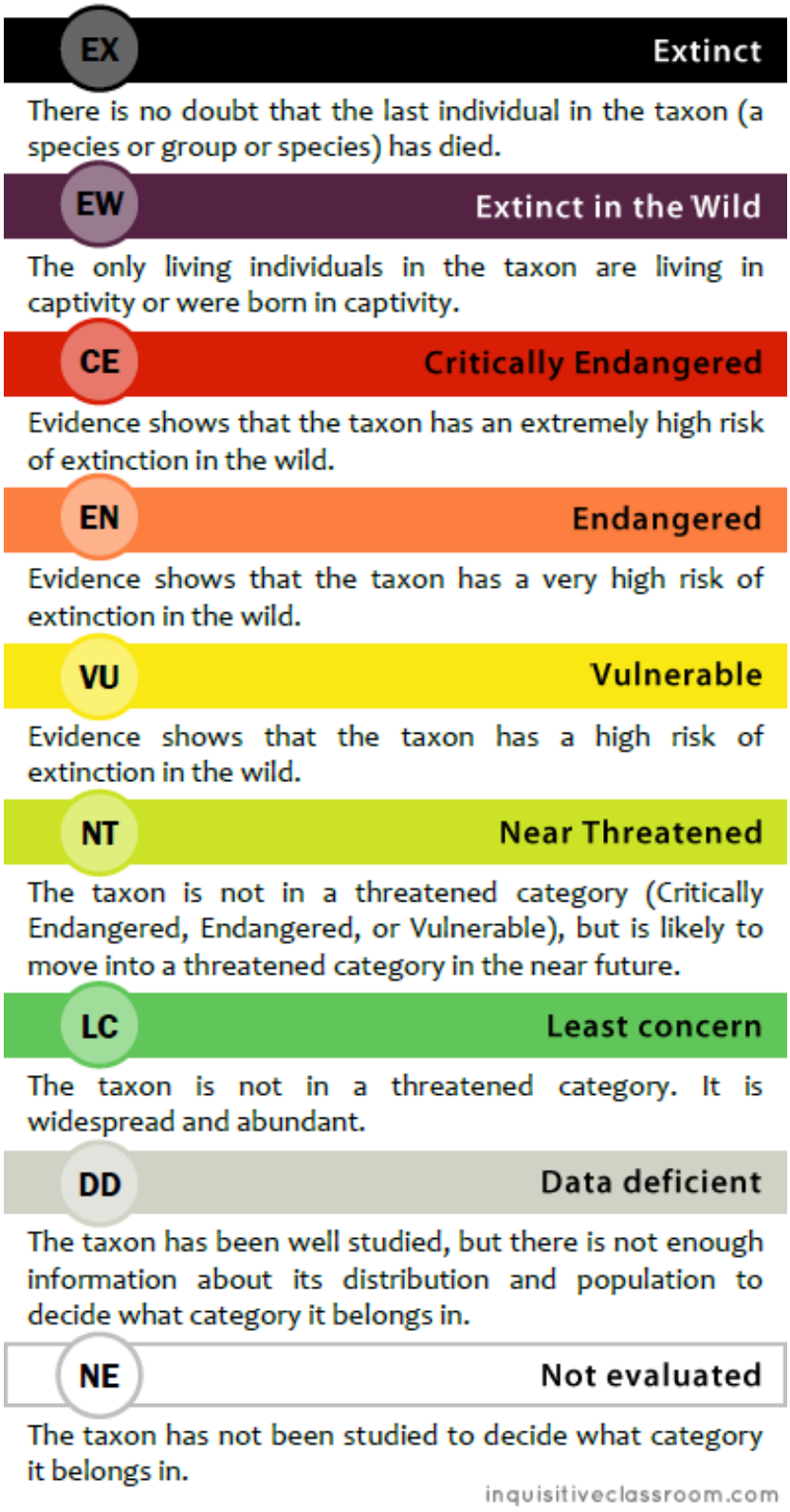
2.Species Conservation Status

Species conservation status: A designation assigned to a species by the IUCN Red List based on its population size, trends, and threats to its survival.

3.Endangered species

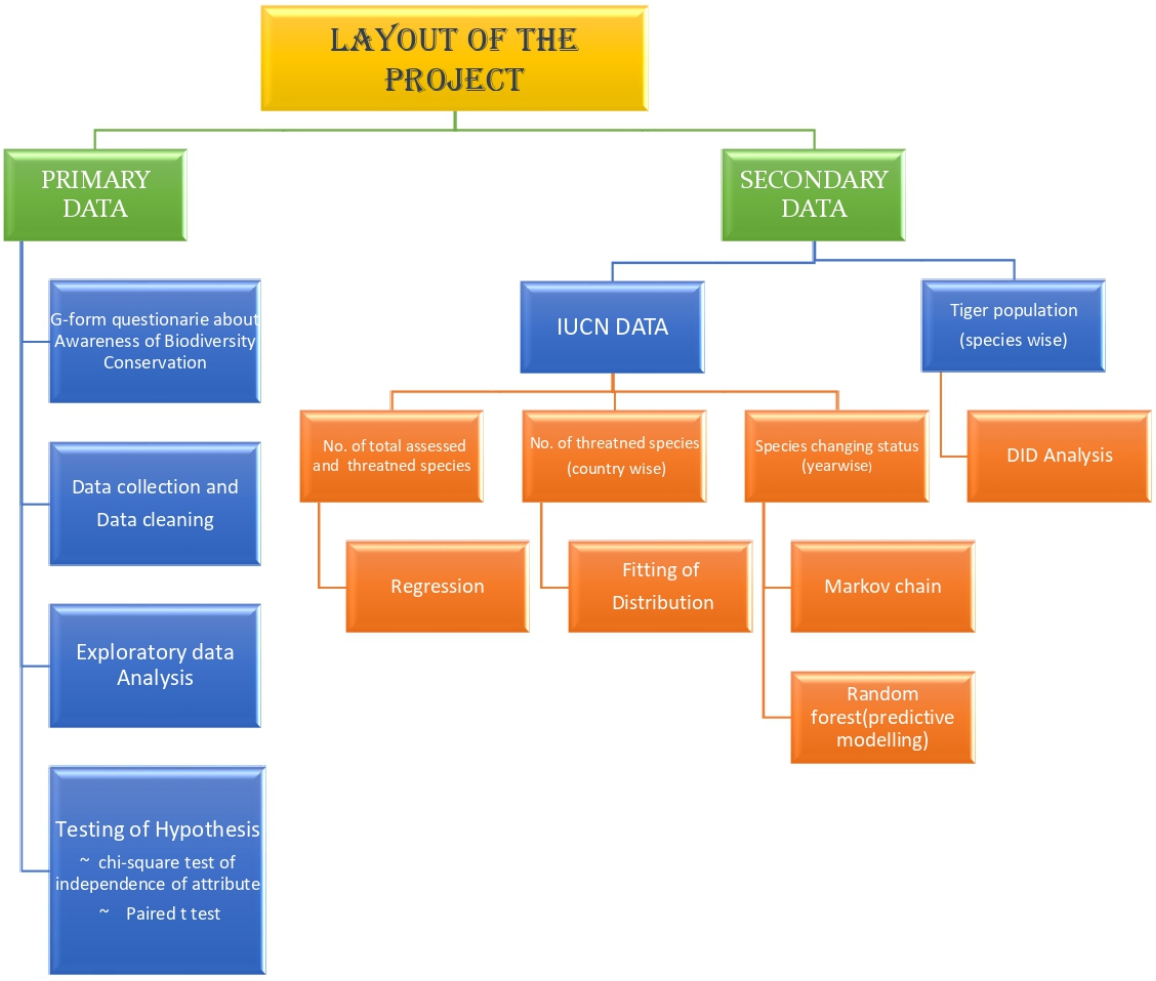
Species that are at high risk of extinction in the wild, often due to a combination of factors including habitat loss, human impact, and low population size.

IUCN Red List Categories version 3.1



MOTIVATION

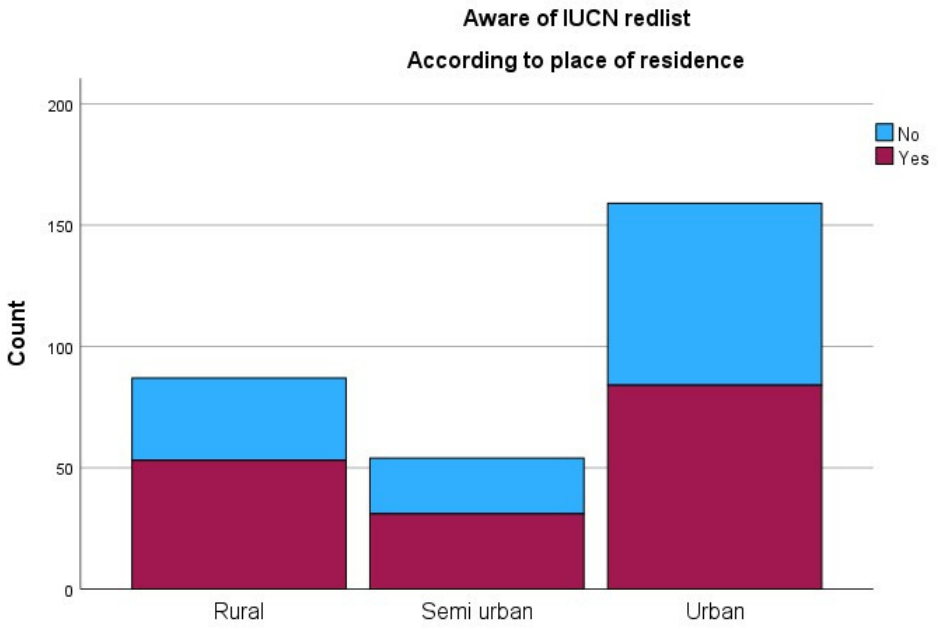
Biodiversity loss is one of the most pressing environmental challenges facing our planet today. Biodiversity provides a range of ecosystem services that are vital for human well-being, including air and water purification, carbon sequestration, and crop pollination. By studying the Red List and conservation efforts, our project can highlight the importance of protecting species and ecosystems for the benefit of both wildlife and humans. According to the IUCN Red List, thousands of species are currently threatened with extinction, and many more are likely to become endangered in the coming years. Understanding the factors that contribute to species declines and developing effective conservation strategies is critical for preventing extinctions and preserving biodiversity. We aim to explore the effectiveness of conservation strategies by examining the impacts of the Wildlife Protection Amendment Act of 2002 on the conservation status of species in India. By using a difference-in-differences analysis, we hope to shed light on the factors that contribute to successful conservation efforts and identify areas where conservation strategies can be improved. We will also use one-step transition probability matrices to estimate the conservation status of different species and build classification models. The motivation of this project to highlight the importance of biodiversity conservation and contribute to the development of effective conservation strategies for the benefit of both wildlife and human populations. Ultimately, this research could inform policy decisions and help guide conservation efforts aimed at preserving biodiversity and protecting vulnerable species.


LAYOUT OF THE PROJECT
PRIMARY DATA
SECONDARY DATA
G-form questionarie about Awareness of Biodiversity Conservation
Data collection and Data cleaning
Exploratory data Analysis
Testing of Hypothesis
~ chi-square test of independence of attribute
~ Paired t test
IUCN DATA
Tiger population (species wise)
No. of total assessed and threatned species
No. of threatned species (country wise)
Species changing status (yearwise)
DID Analysis
Regression
Fitting of Distribution
Markov chain
Random forest(predictive modelling)

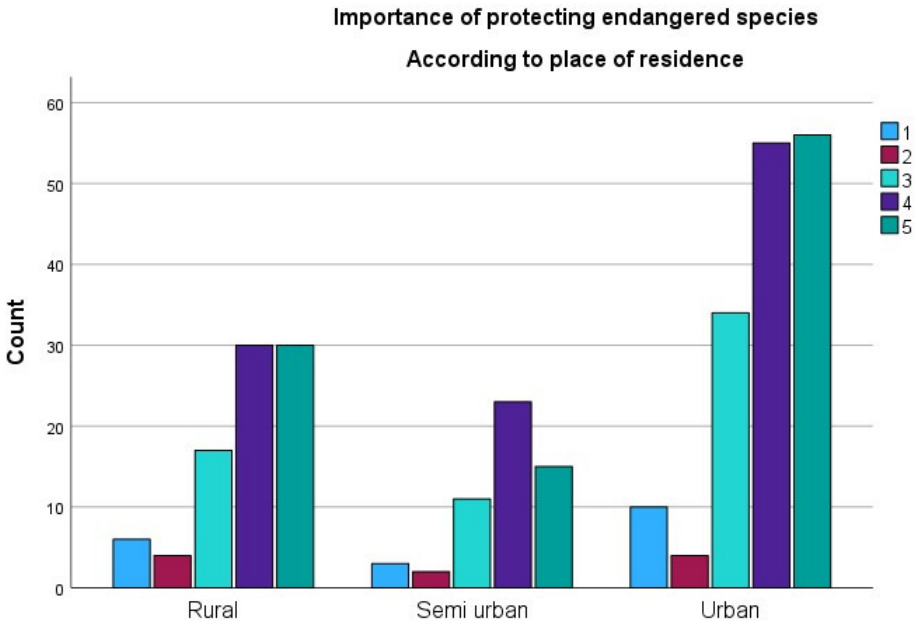
OBJECTIVES

- To analyse and understand the opinion of the general public regarding the topic
- To estimate the number of species to be threatened in the near future
- To model the number of species expected to be threatened in a particular country.
- To highlight the severity of the extinction crisis we are facing
- To predict the conservation status of different species belonging to certain taxons based on their current conservation status
- To test the effectiveness of Wildlife Amendment Act 2002 for conserving and countering the declining Bengal Tiger population in India

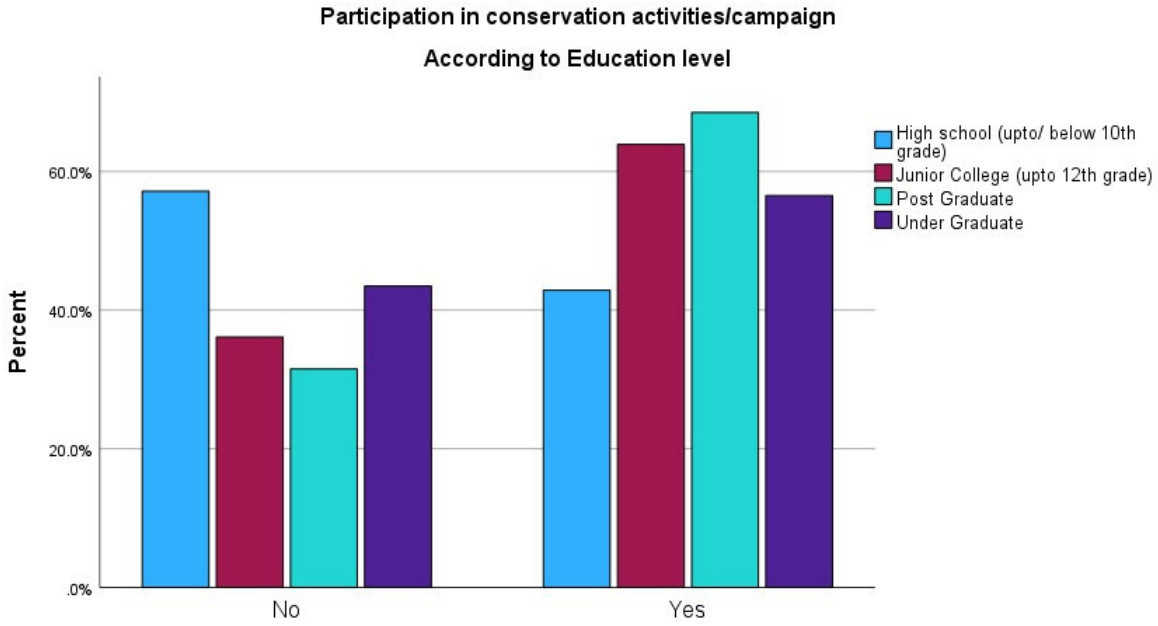
Exploratory Data Analysis on Primary Data



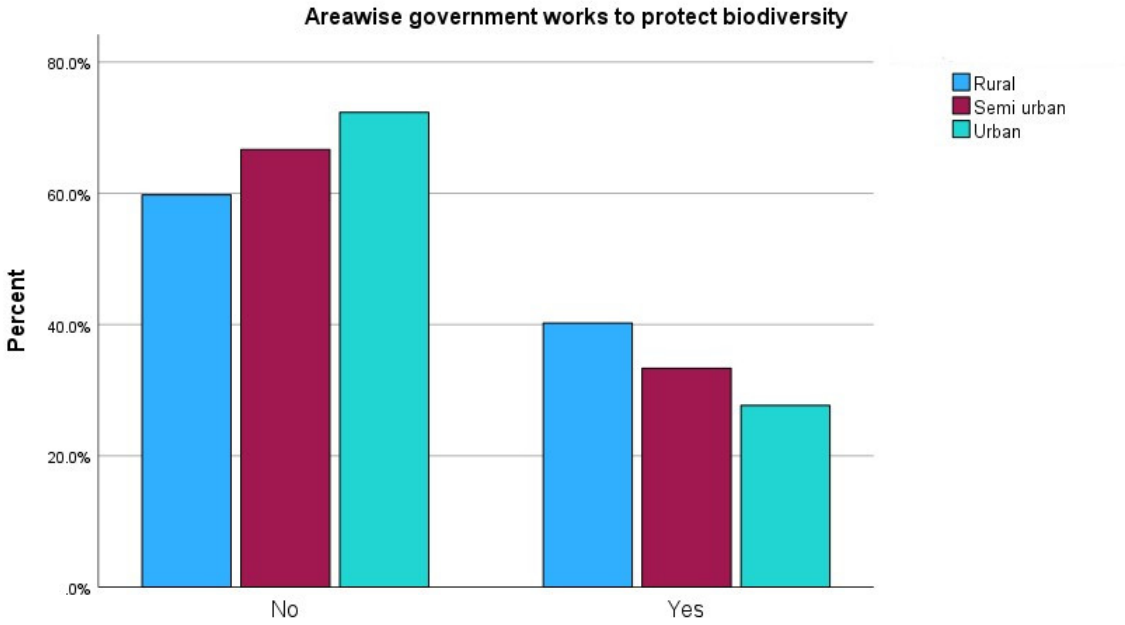
Interpretation:According to the place of residence,we can say that approximately 50% of individuals are aware about IUCN Redlist.



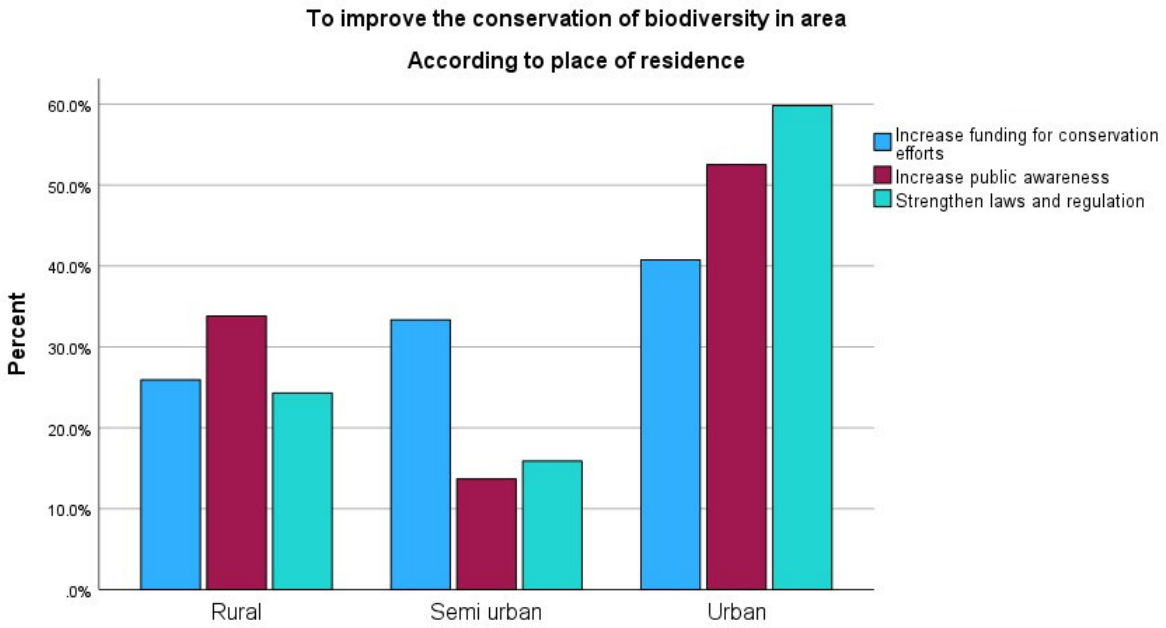
Interpretation:According to our data set people residing from urban area more importance should be given to protect endangered species.



Interpretation:Here,we can observe that Post Graduates and Under Graduates highly participates in biodiversity conservation campaigns.



Interpretation:According to people residing in rural areas government is taking enough efforts to protect biodiversity.



Interpretation:We can observe that, in rural areas to improve the biodiversity conservation public awareness should be increased while semi urban residents thinks that funding for conservation efforts should be increased and according to people residing in urban area laws and regulations should be strengthened.

Testing of Hypothesis

In this section, we perform the following tests on the primary data:

1. Chi Square Test of Independence of Attributes
2. The Wilcoxon sign ranked test

1. Chi Square Test of Independence of Attributes

i) Association Between Area of Residence and Concerned about Biodiversity

H_0 : The attributes are independent
H_1 : The attributes are not independent

Level of concerned	Rural	Urban
Not concerned	9	10
Concerned	5	5
Neutral	28	34
Very concerned	53	55
Extremely concerned	45	56

```
Pearson's Chi-squared test
    data:contingency table
    X-squared = 0.53739, df = 4, p-value = 0.9698
```

We observe that, $\chi^2{}_{(calc)} < \chi^2{}_{(table)}$ hence we accept H_0 at 5% LOS.

ii) Association between info about IUCN and residence

H_0 : The attributes are independent
H_1 : The attributes are not independent

Known about IUCN/Area of Residence	Rural	Semi-urban	Urban
Yes	34	23	75
No	53	31	84

```
Pearson's Chi-squared test\\
data: contingency table\\
X-squared = 1.5463, df = 2, p-value = 0.4616\\
```

We observe that, $\chi^2{}_{(calc)} < \chi^2{}_{(table)}$ hence we accept H_0 at 5% LOS.

iii) Association between level of education and participation in conservation activities

H_0 : The attributes are independent
H_1 : The attributes are not independent

Participation in conservation activities		
Level of Education/	Yes	No
High School	6	5
Junior college	13	23
UG	23	50
PG	80	101

```
Pearson's Chi-squared test\\
data: contingency table\\
X-squared = 4.6647, df = 3, p-value = 0.1981\\
```

We observe that, $\chi^2{}_{(calc)} < \chi^2{}_{(table)}$ hence we accept H_0 at 5% LOS.

iv) Association between Area of Residence and Enough measures to protect Biodiversity by Govt

H_0 : The attributes are independent
H_1 : The attributes are not independent

Enough to protect Biodiversity by Govt		
Area of Residence	Yes	No
Rural	52	35
Semi-urban	36	18
Urban	115	44

```
 Pearson's Chi-squared test\\
 data:  contingency table\\
X-squared = 4.0826, df = 2, p-value = 0.1299}\\
```

We observe that, $\chi^2{}_{(calc)} < \chi^2{}_{(table)}$ hence we accept H_0 at 5%LOS.

v) Association between Area of Residence and Importance of protecting endangered species

H_0 : The attributes are independent
H_1 : The attributes are not independent

Area of Residence		
Importance of protecting endangered species	Rural	Urban
Not important	5	7
Somewhat important	7	3
Neutral	18	7
Very important	23	24
Extremely important	88	118

```
Pearson's Chi-squared test\\
 data:contingency table\\
 X-squared = 10.12, df = 4, p-value = 0.03845}\\
```

We observe that, $\chi^2{}_{(calc)} > \chi^2{}_{(table)}$ hence we reject H_0 at 5% LOS.

The Wilcoxon sign ranked test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to compare two related samples, matched pairs, or repeated measurements on a single sample. It is used to determine whether the median difference between two groups is statistically significant or not. It is a useful alternative to the paired t-test when the data do not meet the assumptions of normality and equal variances.

Before using this test we have to perform **Shapiro -Wilk Test of Normality** for checking non-normality of this data for both the varibles.

H_0 : The variable is Normally distributed.

H_1 : The variable is not Normally distributed.

```
# shapiro.test(a)

# Shapiro-Wilk normality test
#  data:  a
#  W = 0.91103, p-value = 2.491e-12

# shapiro.test(b)

# Shapiro-Wilk normality test
# data:  b
# W = 0.83043, p-value < 2.2e-16
```

Decision: We observe that,$W_{(calc)} > W_{(table)}$, Hence, we REJECT H_0 at 5% LOS.
The variables are not normally distributted. Therefore we can use non-parametric test.

In our priamry survey, we have asked people regarding their engagement in conservation activites or any other awareness campaign and take their responses about level of concenrn before and after participating in such activites. By conducting the Wilcoxon signed-rank test, we can determine whether participation in biodiversity and conservation campaigns leads to a significant increase in individuals' concern levels regarding conservation activities. This information can be useful for designing future campaigns and interventions aimed at promoting biodiversity conservation.

H_0:There is no significant difference in the concern levels of individuals before and after participating in biodiversity and conservation campaigns.

H_1: There is a significant difference with a higher median concern level after participation.

```
# Wilcoxon rank sum test with continuity correction

# data:  u and v
# W = 26652, p-value < 2.2e-16
```

Regression analysis of yearly total assessed and total threatened species

Overview

To carefully assess and analyse the figures presented by the IUCN Red List so as to quantify the risk of extinction currently faced by all taxons globally, it is imperative to also understand the assessment efforts undertaken by the organisation.

To get an insight into the assessment efforts of the organisation, a comparison needs be done between the total number of species assessed every year and the total number of species that turn out be threatened. As of now, the IUCN has assessed a total of at least 150,388 species globally with 28% (42,100) of all assessed species being classified as threatened.

However, it is important to note that this percentage figure was very high when the assessments just began (for example in the year 2002, it was as high as 66.92%). The main reason for this decrease in the percentage is solely because of the assessment efforts of the IUCN. As more number of species are being assessed, more we are geting a clearer picture of the extinction crisis we facing now.

We thus attempt to model the number of species that will be assessed in the following years and how many out of the total assessed will be deemed as threatened by using the technique of regression analysis.

Regression analysis

Introduction

Regression analysis is a statistical modelling field of study which makes use of mathematical functions to predict the outcome of a certain dependent/response variable influenced by other regressor/explanatory variables.

The term 'regression' was first coined by Sir Francis Galton, while he was studying the relation between the height of son and their father's. In simple terms 'regression' means 'regressing towards mean'.

Depending on the nature of relation between the regressor and the response variable(s), a mathematical function is known as the best model if it minimizes the least square error. That is, the method of least square estimation is most widely used for fitting a regression model to a data.

Some different types of curves usually fitted are:-

- Single or Multiple linear regression
- Polynomial and Non-linear regression
- Logistic regression

General Idea

Consider fitting a simple linear regression model. It requires estimation of the regression coefficients β_0 and β_1 where the regression coefficient is given by :-

$$Y = \beta_0 + \beta_1 * X + \epsilon$$

where,
Y = Outcome variable
β_0 = Interpreted as mean of distribution of Y when $X = 0$
β_1 = Average change in Y produced by unit change in X
ϵ = Random error component

The equation of best fit changes as the correlation between the response and regressor variables change.

Assumptions:-

The following assumptions are made for all regression models:-
1.) Errors are normally distributed with mean=0 and a constant varience=σ^2 (that also means error terms are homoscedastic).
2.)Regressor variables should be uncorrelated to one another.
3.)There should NOT be autocorrelation between the residuals.

Thus, in our case, the response variables are total number of species assessed by IUCN and total number of species deemed threatened by IUCN , while the regressor variable is time (two separate models have been made)

Model I:-
Response Variable = Y : Number of species assessed
Regressor Variable = X : Time(in years)

Model II:-
Response Variable = Y : Number of species threatened
Regressor Variable = X : Time(in years)

Regression analysis for the given data and problem under consideration

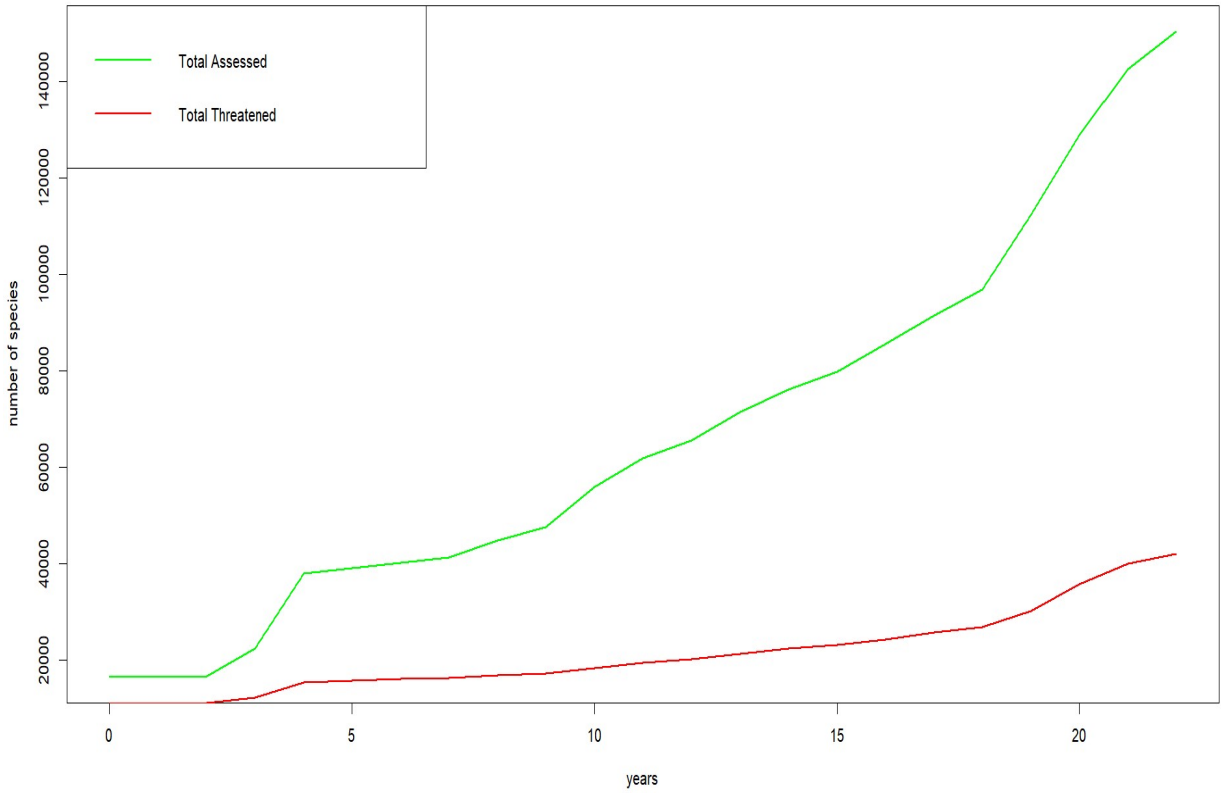
The first step in deciding the type of regression analysis to be used, is to observe the scatter plot and look for a general trend. This initial step helps us in understanding the general relationship between the response and the regressor variable and accordingly choosing a

model to be fitted initially.
The data was made available by the IUCN and it gives the total number of species assessed and total number of species threatened for all taxonomies, every year. A glimpse of the data is given for reference.

Year[3]		VERTEBRATES						INVERTEBRATES								PLANTS[4]							FUNGI & PROTISTS				TOTAL
		Mammals	Birds	Reptiles	Amphibians	Fishes	Subtotal (Vertebrates)	Insects	Molluscs	Crustaceans	Corals	Arachnids	Velvet worms	Horseshoe crabs	Other invertebrates	Subtotal (Invertebrates)	Mosses	Ferns & allies	Gymnosperms	Flowering plants	Green algae	Red algae	Subtotal (Plants)	Lichens	Mushrooms	Brown algae	Subtotal (Fungi & protists)	
2022 (version 2022-2)	Total assessed	5,973	11,188	10,222	7,486	25,351	**60,220**	12,441	9,032	3,197	831	441	11	4	905	**26,862**	329	747	1,046	60,470	16	58	**62,666**	86	539	15	**640**	**150,388**
	Total threatened	1,340	1,400	1,842	2,606	3,551	**10,739**	2,345	2,399	745	253	251	9	2	157	**6,161**	181	288	436	24,000	0	9	**24,914**	62	226	6	**294**	**42,108**
2021 (version 2021-3)	Total assessed	5,968	11,162	10,148	7,296	22,581	**57,155**	12,100	9,019	3,189	848	441	11	4	902	**26,514**	282	739	1,016	56,232	16	58	**58,343**	76	474	15	**565**	**142,577**
	Total threatened	1,333	1,445	1,839	2,488	3,332	**10,437**	2,270	2,385	743	232	251	9	2	150	**6,042**	165	281	403	22,477	0	9	**23,335**	56	208	6	**270**	**40,084**
2020 (version 2020-3)	Total assessed	5,940	11,158	8,492	7,212	22,005	**54,807**	10,865	8,847	3,188	864	358	11	4	842	**24,219**	282	674	1,016	48,323	16	58	**50,369**	55	353	15	**423**	**128,918**
	Total threatened	1,323	1,481	1,458	2,442	3,210	**9,914**	1,926	2,300	742	237	203	9	2	148	**5,489**	165	265	403	19,518	0	9	**20,360**	48	185	6	**239**	**35,765**
2019 (version 2019-3)	Total assessed	5,850	11,147	7,829	6,794	19,199	**50,819**	8,696	8,749	3,181	864	344	11	4	839	**22,688**	281	641	1,014	36,623	13	58	**38,630**	27	253	15	**295**	**112,432**
	Total threatened	1,244	1,486	1,409	2,200	2,674	**9,013**	1,647	2,250	733	237	197	9	2	146	**5,221**	164	261	402	14,938	0	9	**15,774**	24	140	6	**170**	**30,178**
2018 (version 2018-2)	Total assessed	5,692	11,126	7,127	6,722	16,803	**47,470**	8,037	8,627	3,180	864	324	11	4	839	**21,886**	102	558	1,012	25,771	13	58	**27,514**	23	43	15	**81**	**96,951**
	Total threatened	1,219	1,492	1,307	2,092	2,332	**8,442**	1,537	2,195	733	237	182	9	1	146	**5,040**	76	249	401	12,564	0	9	**13,299**	20	33	6	**59**	**26,840**
2017 (version 2017-3)	Total assessed	5,674	11,122	6,278	6,609	16,409	**46,092**	7,639	8,413	3,177	864	249	11	4	773	**21,130**	102	479	1,012	22,566	13	58	**24,230**	13	43	15	**71**	**91,523**
	Total threatened	1,204	1,469	1,215	2,100	2,386	**8,374**	1,414	2,187	732	237	170	9	1	143	**4,893**	76	246	401	11,773	0	9	**12,505**	10	33	6	**49**	**25,821**
2016 (version 2016-3)	Total assessed	5,567	11,121	5,338	6,534	16,134	**44,694**	6,587	7,276	3,177	862	212	11	4	480	**18,609**	102	417	1,011	20,652	13	58	**22,253**	8	25	15	**48**	**85,604**
	Total threatened	1,194	1,460	1,079	2,068	2,359	**8,160**	1,268	1,984	732	237	166	9	1	73	**4,470**	76	217	400	10,941	0	9	**11,643**	7	21	6	**34**	**24,307**
2015 (version 2015.4)	Total assessed	5,502	10,424	4,669	6,460	14,462	**41,517**	5,573	7,216	3,168	862	210	11	4	472	**17,516**	102	365	1,011	19,206	13	58	**20,755**	9	25	15	**49**	**79,837**
	Total threatened	1,197	1,375	944	1,994	2,271	**7,781**	1,046	1,950	728	237	164	9	0	67	**4,201**	76	197	400	10,551	0	9	**11,233**	7	22	6	**35**	**23,250**
2014 (version 2014-3)	Total assessed	5,513	10,425	4,414	6,414	12,457	**39,223**	5,304	7,217	3,164	856	209	11	4	453	**17,218**	102	360	1,010	18,195	13	58	**19,738**	4	1	15	**20**	**76,199**
	Total threatened	1,199	1,373	927	1,957	2,222	**7,678**	993	1,950	725	235	163	9	0	65	**4,140**	76	194	400	9,905	0	9	**10,584**	4	1	6	**11**	**22,413**
2013 (version 2013.2)	Total assessed	5,506	10,065	4,204	6,409	11,172	**37,356**	4,610	6,809	3,163	856	35	11	4	423	**15,911**	102	342	1,010	16,766	13	58	**18,291**	2	1	15	**18**	**71,576**
	Total threatened	1,143	1,308	879	1,950	2,110	**7,390**	896	1,898	723	235	21	9	0	40	**3,822**	76	187	399	9,394	0	9	**10,065**	2	1	6	**9**	**21,286**
2012 (version 2012.2)	Total assessed	5,501	10,064	3,755	6,374	10,590	**36,284**	4,003	6,183	2,399	858	34	11	4	50	**13,542**	102	311	1,012	14,178	13	58	**15,674**	2	1	15	**18**	**65,518**
	Total threatened	1,139	1,313	807	1,933	2,058	**7,250**	829	1,857	596	236	20	9	0	23	**3,570**	76	167	374	8,764	0	9	**9,390**	2	1	6	**9**	**20,219**
2011 (version 2011.2)	Total assessed	5,499	10,052	3,336	6,338	9,554	**34,779**	3,844	5,422	2,399	856	33	11	4	52	**12,621**	101	310	1,020	12,994	13	58	**14,496**	2	1	15	**18**	**61,914**
	Total threatened	1,138	1,253	772	1,917	2,028	**7,108**	741	1,673	596	235	19	9	0	24	**3,297**	80	163	377	8,527	0	9	**9,156**	2	1	6	**9**	**19,570**

The last column indicates the total species assessed as well as the total number of species threatened. The data was imported in R and further analysis was performed.

The total number of assessed species and the total number of threatened species were plotted against time(from the year 2002 to 2022). The observed plot was:-



Thus the graphs suggest that linearity is absent for both the response variables. We shall verify this with the diagnostic plots.

Using R, the linear regression lines were fitted to the response variables.(Figure 1)

The diagnostic plots were plotted for both the models too. For demonstration, consider the diagnostic plots for the number of total assessed species.(Figure 2)

The diagnostic plot 1 suggests that there is non-linear relationship between the response and regressor variable. Additionally, the scale-location plot indicates presence of heteroscedasticity. Thus, the fitted is not good and we should use non-linear models to best describe our data.

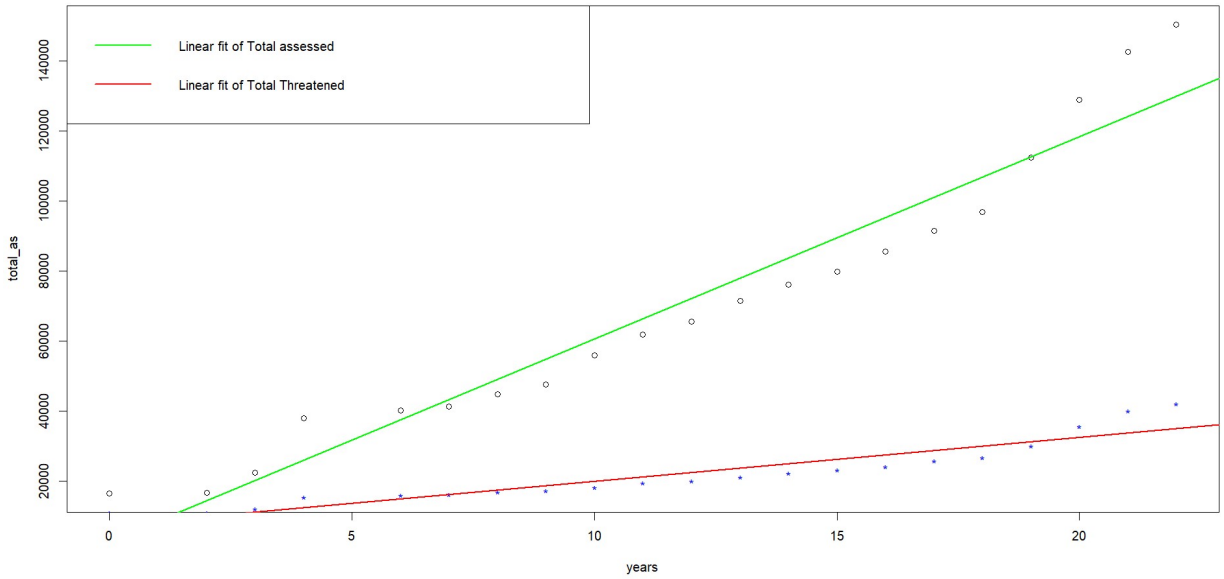


Figure 1: Best fitted line for the response variables

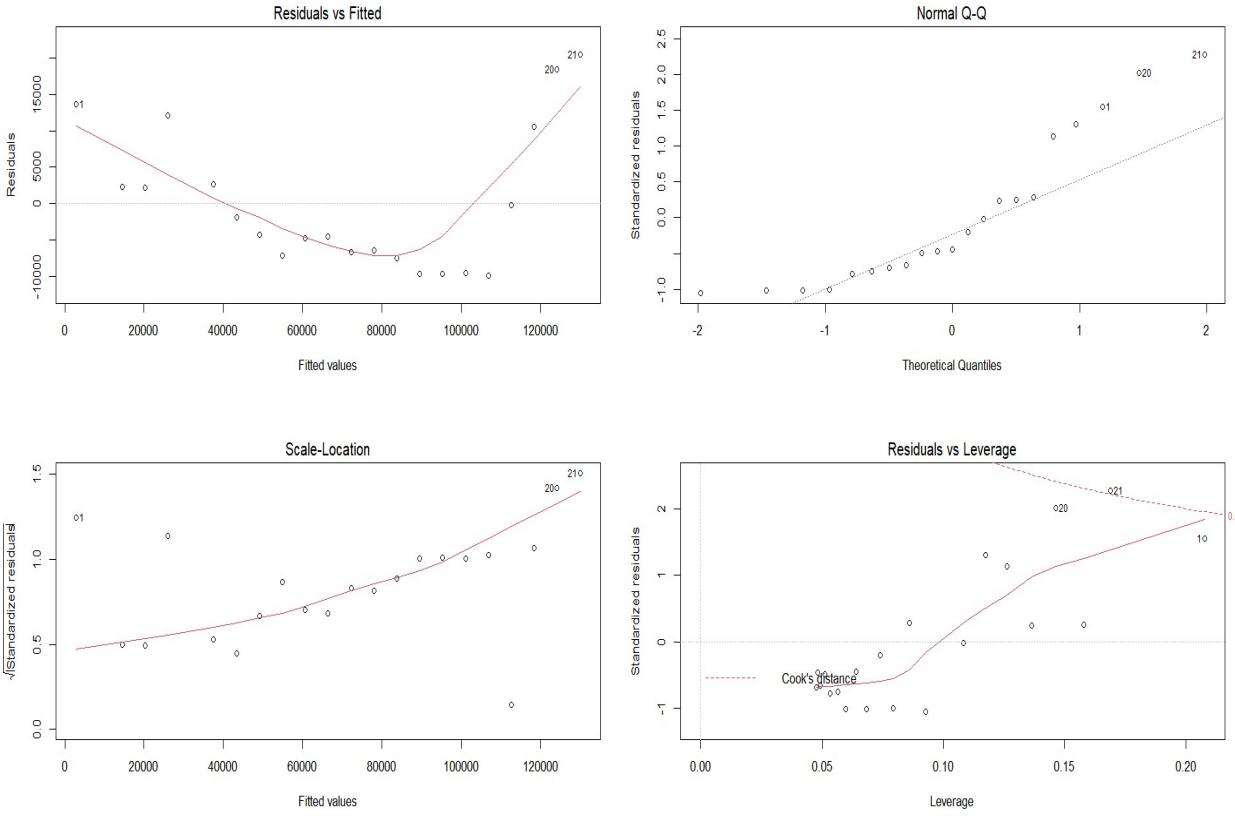


Figure 2: Diagnostic plots for model of total assessed species

Fitting of non linear models

Now since the diagnostic plots suggest a non-linear model, a number of non-linear models were fitted and the best possible model was taken into account.

For both the response variables certain models were fitted and the results have been summarised. The table indicates which assumptions have been violated leading to rejection of the model.

Model I:-

For modelling total species assessed:-

Model	Residuals vs Fitted values	Standardized residuals	Autocorrelation	Normality
$Y = a \cdot e^{bX}$	Yes	No	Yes	No
$Y = a \cdot b^{X}$	No	No	Yes	No
$Y = a \cdot X^{b}$	Yes	No	Yes	Yes
$Y = a + b \cdot X + c \cdot X^2$	Yes	No	Yes	No
$Y = a + b \cdot X + c \cdot X^2 + d \cdot X^3$	No	No	No	No

Thus, the cubic equation was overall best fit, where no assumptions were violated.

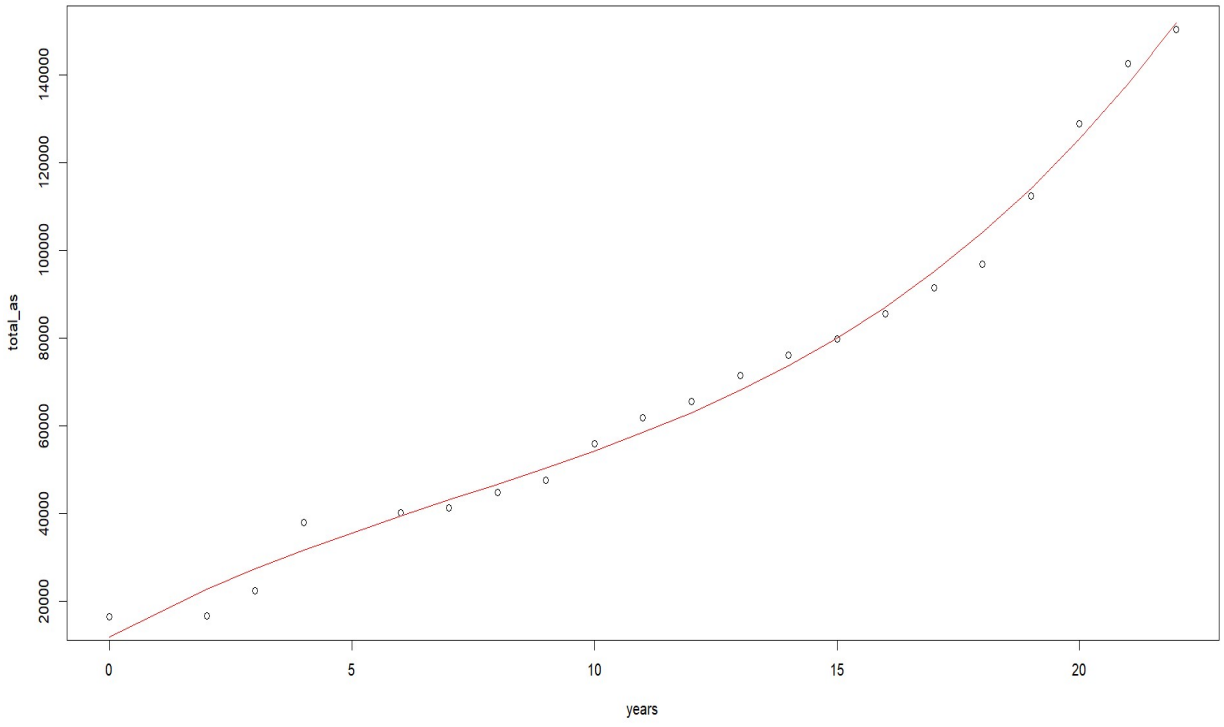


Figure 3: Regression curve on actual values

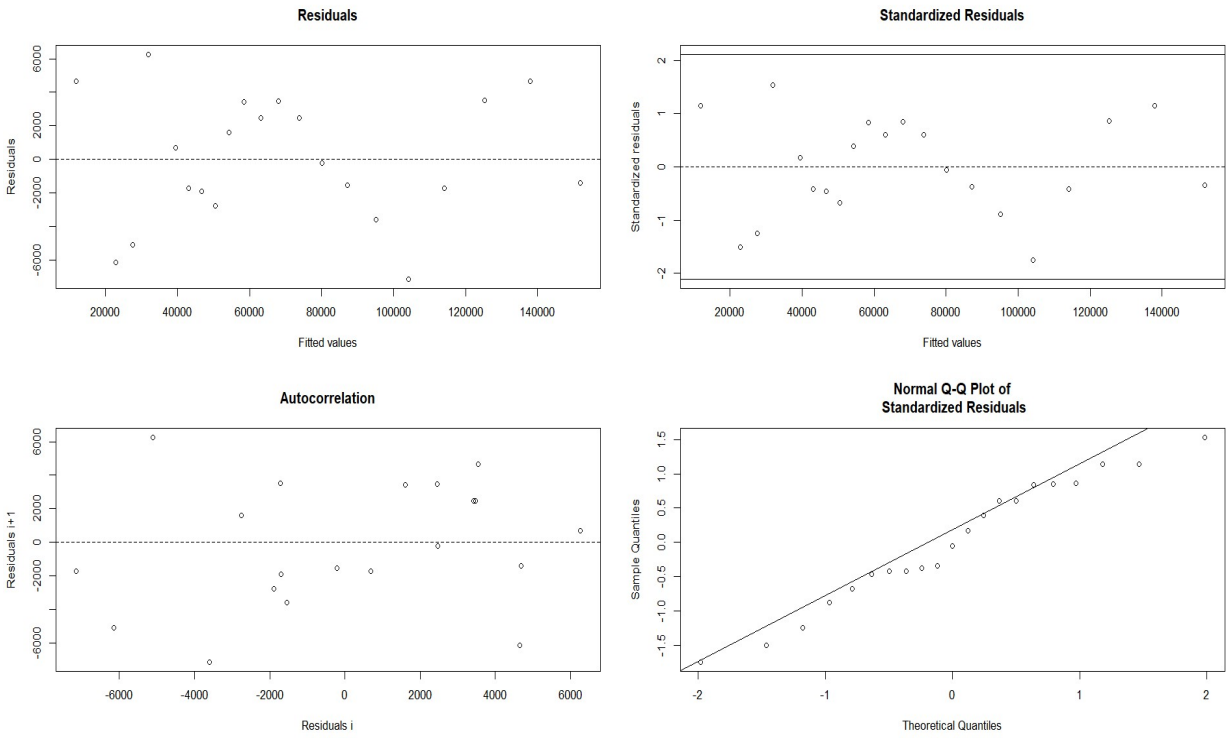


Figure 4: Diagnostic plots for the model

Concise summary of the model and its interpretation is given as:-

```
> overview(model)

------
Formula: total_as ~ a + b * years + c * years^2 + d * years^3

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 11847.815   3376.729   3.509 0.002693 **
b  6142.485   1282.478   4.790 0.000171 ***
c  -355.686    134.181  -2.651 0.016813 *
d    16.620      3.953   4.204 0.000597 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4083 on 17 degrees of freedom

Number of iterations to convergence: 1
Achieved convergence tolerance: 2.791e-08

------
Residual sum of squares: 2.83e+08

------
t-based confidence interval:
         2.5%       97.5%
a 4723.538988 18972.09190
b 3436.692166  8848.27768
c -638.783504   -72.58897
d    8.278848    24.96081

------
Correlation matrix:
           a          b          c          d
a  1.0000000 -0.8274881  0.6831721 -0.5913462
b -0.8274881  1.0000000 -0.9638657  0.9081558
c  0.6831721 -0.9638657  1.0000000 -0.9853145
d -0.5913462  0.9081558 -0.9853145  1.0000000
```

Interpretation:-

The fitted equation is,

$$Y = 11847.815 + 6142.485X - 355.686X^2 + 16.62X^3$$

Further,the t-values indicate that all the regression coefficients are statistically significant,that is all the coefficients used in the model are important for predicting the number of species that will be assessed in the coming years.

Thus the very same procedure used for the fitting of the above model,was employed while fitting the curve for total threatened species. It was overall found out that the best fitted curve is once again a cubic one,just like the one above. The summary tables have been drawn, along with the plots for better visualisation.

Model II:-

For modelling total species threatened:-

Model	Residuals vs Fitted values	Standardized residuals	Autocorrelation	Normality
$Y = a \cdot e^{bX}$	Yes	No	Yes	No
$Y = a \cdot b^{X}$	Yes	No	Yes	No
$Y = a \cdot X^{b}$	Yes	No	Yes	Yes
$Y = a + b \cdot X + c \cdot X^2$	Yes	No	Yes	No
$Y = a + b \cdot X + c \cdot X^2 + d \cdot X^3$	No	No	No	No

Thus,here too the cubic equation was overall best fit, where no assumptions were violated.

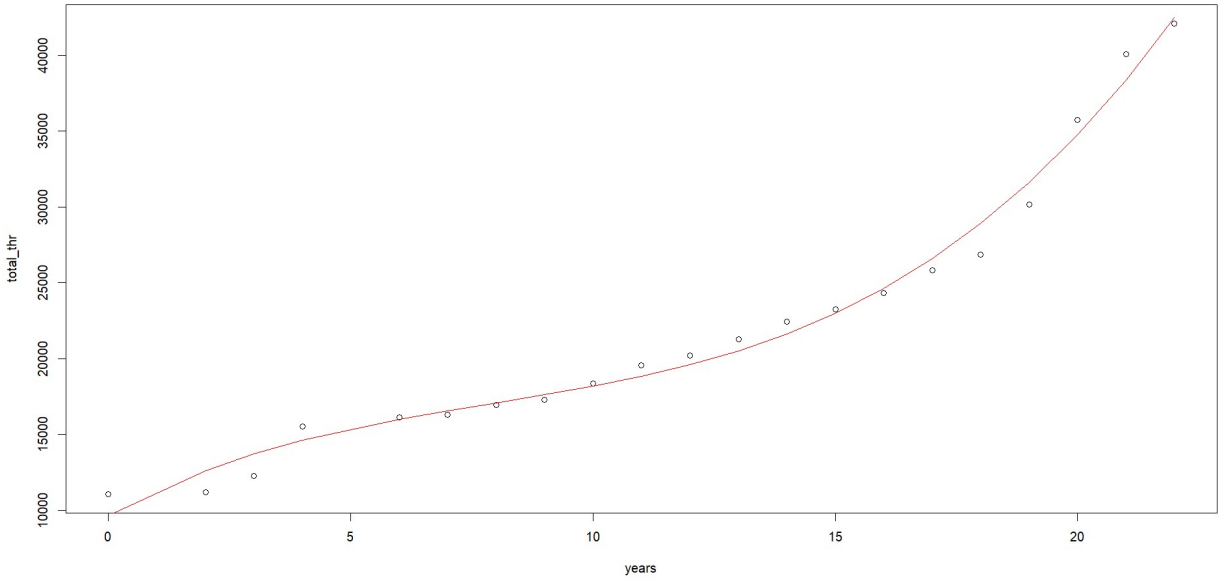


Figure 5: Regression curve on actual values

Thus the diagnostic plots for the model, which validate the model are:-

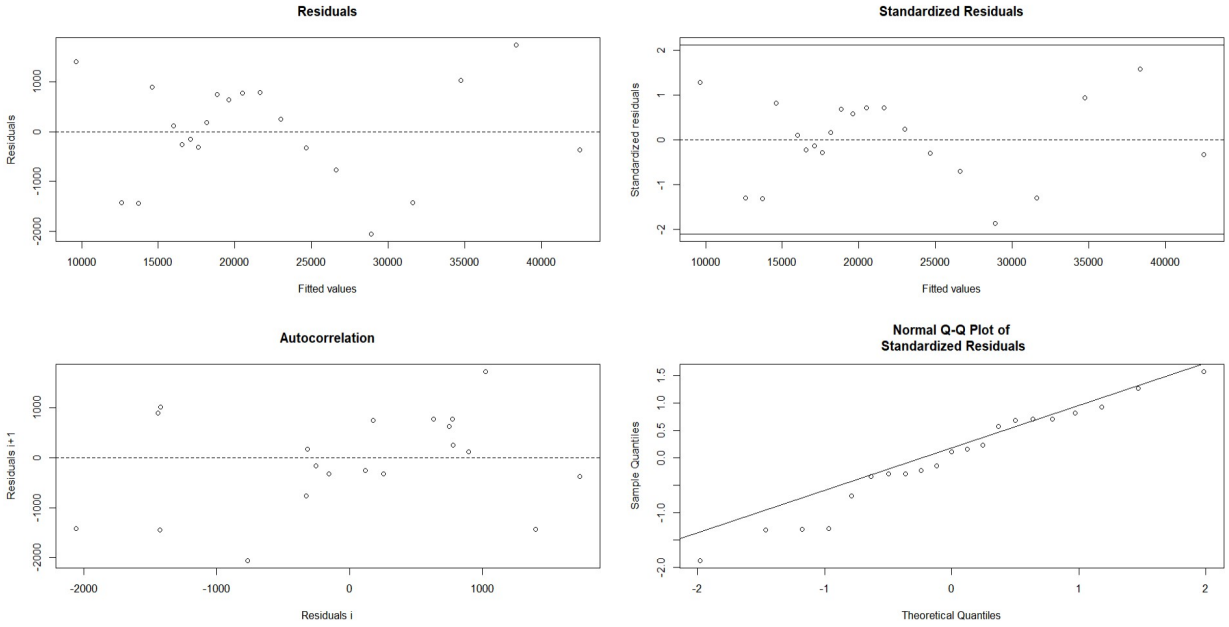


Figure 6: Diagnostic plots for model

The final model and the interpretation are as follows

```
> overview(model)

------
Formula: total_thr ~ a + b * years + c * years^2 + d * years^3

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 9645.266    907.670  10.626 6.31e-09 ***
b 1761.458    344.732   5.110 8.72e-05 ***
c -156.398     36.068  -4.336 0.000449 ***
d    6.553      1.063   6.167 1.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1098 on 17 degrees of freedom

Number of iterations to convergence: 1
Achieved convergence tolerance: 2.475e-08

------
Residual sum of squares: 20500000

------
t-based confidence interval:
        2.5%        97.5%
a 7730.25035 11560.282640
b 1034.13699  2488.779243
c -232.49493   -80.300997
d    4.31118     8.795317

------
Correlation matrix:
           a          b          c          d
a  1.0000000 -0.8274881  0.6831721 -0.5913462
b -0.8274881  1.0000000 -0.9638657  0.9081558
c  0.6831721 -0.9638657  1.0000000 -0.9853145
d -0.5913462  0.9081558 -0.9853145  1.0000000
```

Interpretation:-

The fitted equation is,

$$Y = 965.266 + 1761.458X - 156.398X^2 + 6.553X^3$$

Further, the t-values indicate that all the regression coefficients are statistically significant,that is all the coefficients used in the model are important for predicting the number of species that will be labeled threatened out of the total species assessed in the coming years.

Estimating the number of species expected to be assessed and threatened in the next six years:-

Year	Number of species expected to be assessed	Number of species expected to be threatened
2023	1,67,183	38,475
2024	1,84,147	43,744
2025	2,02,793	49,644
2026	2,23,221	56,214
2027	2,45,531	63,493
2028	2,69,821	71,522

Conclusion:-

The best possible curves for predicting the number of species that will be assessed in the coming years and out of which how many will be threatened have been thus obtained .

With these models,we have estimated the number of species which will become more prone to extinction for the coming years. This hence becomes an elementary step towards really estimating the extinction risk that a species can face.

Probabilistic Model of number of species threatened (in each major taxonomic group) by country

Overview

For any given year, to assess the severity of extinction risk worldwide it is necessary to estimate how many species on an average are expected to be threatened in a particular year per country. This threat can be quantified by and estimated by fitting an appropriate probability distribution model.

By fitting a probability distribution, we aim to describe the given data and gain more insights into it, as well as predicting outcomes or estimating figures. Additionally, based on the model obtained simulations can be performed and hence appropriate inferences can be drawn.

Introduction

Statistically, fitting of a distribution to a given data set means estimating the parameters of the distribution which are to be used for modelling the given data.

It involves choosing and selecting a probability distribution/mathematical function which best describes and most closely approximates the observed data.

Fitting of a distribution usually involves the following process:-

1. Plot the histogram of the data to get an idea of the curve of the distribution of the data set
2. Carry out summary statistic analysis and observe the relationships between measures of central tendency(for example mean,median), measures of dispersion, skewness and kurtosis; if any and look for any theoretical properties best describing the data.
3. Estimate the parameters of all possible distributions.
4. Perform the goodness of test to get an idea of how ideal the fit was.
5. Repeat the above steps till the best possible fit (model) has been found and applied.

Note:- There is no rigorous or a well defined approach towards fitting a distribution. Many a times, the variable under consideration is an important factor in determining the best probability model. Along with that, earlier experience and intuition also prove to be helpful in such scenarios.

Problem under consideration

The data to which a probability distribution is to be fitted was available on the official website of IUCN Red List. A sample of the data points is attached for reference.

Name	Mammals	Birds	Reptiles*	Amphibians	Fishes*	Molluscs*	Other Inverts*	Plants*	Fungi*	Chromists*	Total
Antarctic											
Antarctica	2	5	0	0	1	0	0	0	0	0	8
Bouvet Island	1	1	0	0	0	0	0	0	0	0	2
French Southern Territories	3	14	7	0	10	0	0	2	0	0	36
Heard Island and McDonald Islands	1	10	0	0	2	0	0	0	0	0	13
South Georgia and the South Sandwich Islands	3	6	0	0	0	0	0	0	0	0	9
Caribbean Islands											
Anguilla	1	0	8	0	48	0	25	3	0	0	85
Antigua and Barbuda	2	3	8	0	48	0	25	2	0	0	88
Aruba	2	1	2	0	50	1	25	5	0	0	86
Bahamas	5	9	13	0	60	1	27	28	0	0	143
Barbados	3	3	7	0	47	0	26	4	0	0	90
Bermuda	4	3	4	0	35	2	35	8	0	0	91
Bonaire, Sint Eustatius and Saba	3	1	7	0	51	0	27	6	0	0	95
Cayman Islands	1	1	11	0	43	2	28	23	0	0	109

Figure 7: Number of threatened species, taxonomy wise per country for the year 2022

The figures indicate the number of species threatened in every taxon for every country in the year 2022. In our case, we want to model the **total** number of threatened species for any particular country, chosen at random
Thus, for our situation the random variable can be given as:-

X : Total number of species threatened in a particular country

Now, after the random variable was defined, the dataset was imported in Excel, and then on R software for analysis. To understand the general nature of the data, histogram was plotted and was obtained as:-

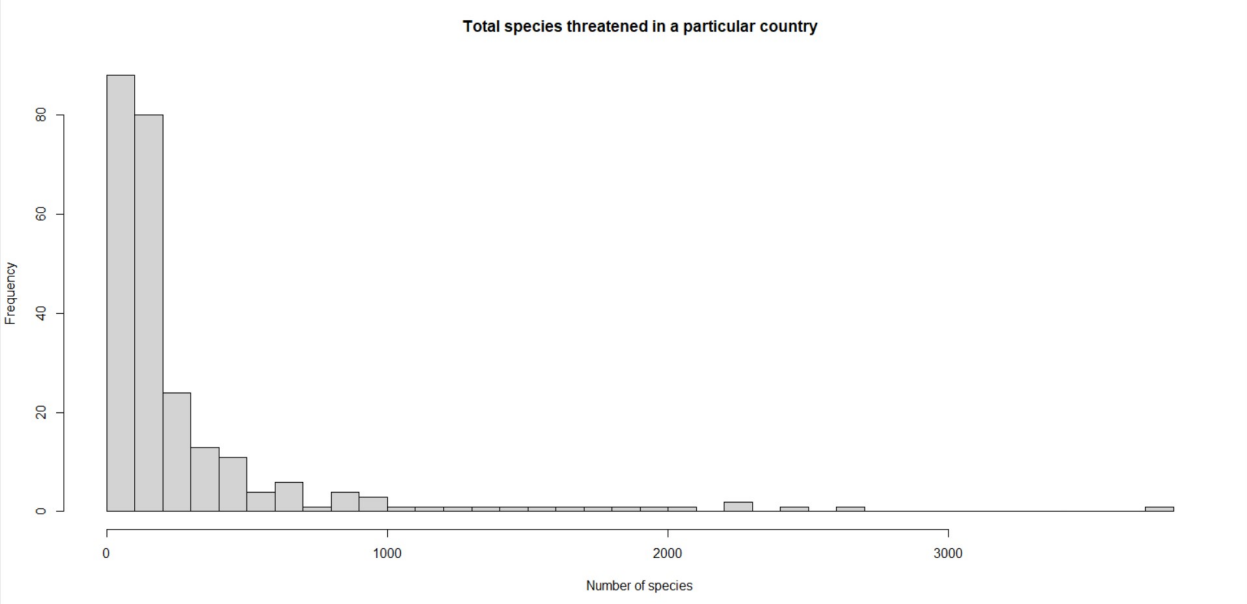


Figure 8: Histogram for given data

Using the packages 'MASS' and 'fitdistrplus' in R software, fitting of a distribution was carried out.(Refer Appendix)

The density curve and the corresponding cumulative distribution for the given data were plotted to get a general idea of the distribution.

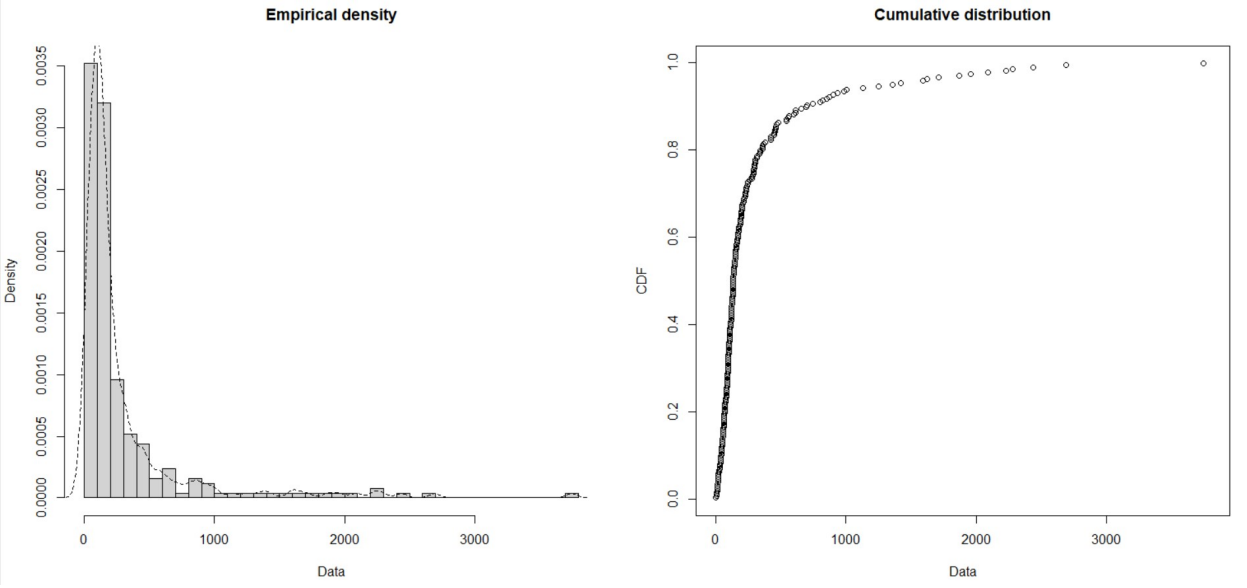
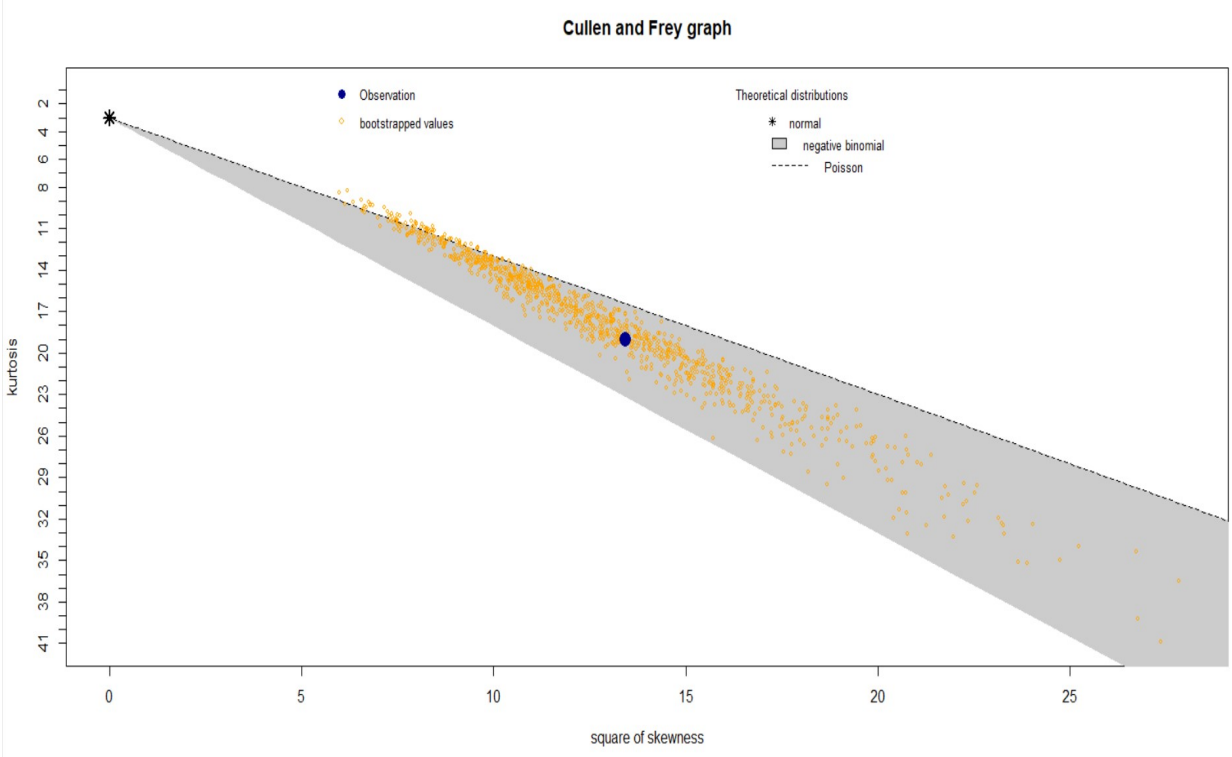


Figure 9: Density and PMF plot

To estimate best distribution to be fitted, we make use of the Cullen-Frey graph.



Thus it can be observed that, the observations(blue dot) and a majority of the bootstrapped values lie in the grey shaded region, indicating that our data might be following a negative binomial distribution.

```
> gofstat(list(fnbin,fgeom),fitnames=c('negative binomial','geometric'))
Chi-squared statistic:  89.78903 85.04634
Degree of freedom of the Chi-squared distribution:  12 13
Chi-squared p-value:  5.424182e-14 1.227876e-12
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
       obscounts theo negative binomial theo geometric
<= 21         16               26.394635      17.612306
<= 47         16               21.796856      19.221800
<= 63         17               11.520717      11.029915
<= 85         18               14.235725      14.240323
<= 95         16                5.966788       6.136863
<= 111        17                8.988342       9.404806
<= 127        16                8.376850       8.918170
<= 137        17                4.956155       5.337980
<= 165        16               12.862447      14.036892
<= 197        16               13.103034      14.522683
<= 244        16               16.680585      18.718323
<= 316        16               20.840172      23.566114
<= 461        17               29.668234      33.343062
<= 854        16               37.112585      39.291751
> 854         20               17.496874      14.619010

Goodness-of-fit criteria
                                 negative binomial geometric
Akaike's Information Criterion          3349.886  3355.801
Bayesian Information Criterion          3356.929  3359.322
> |
```

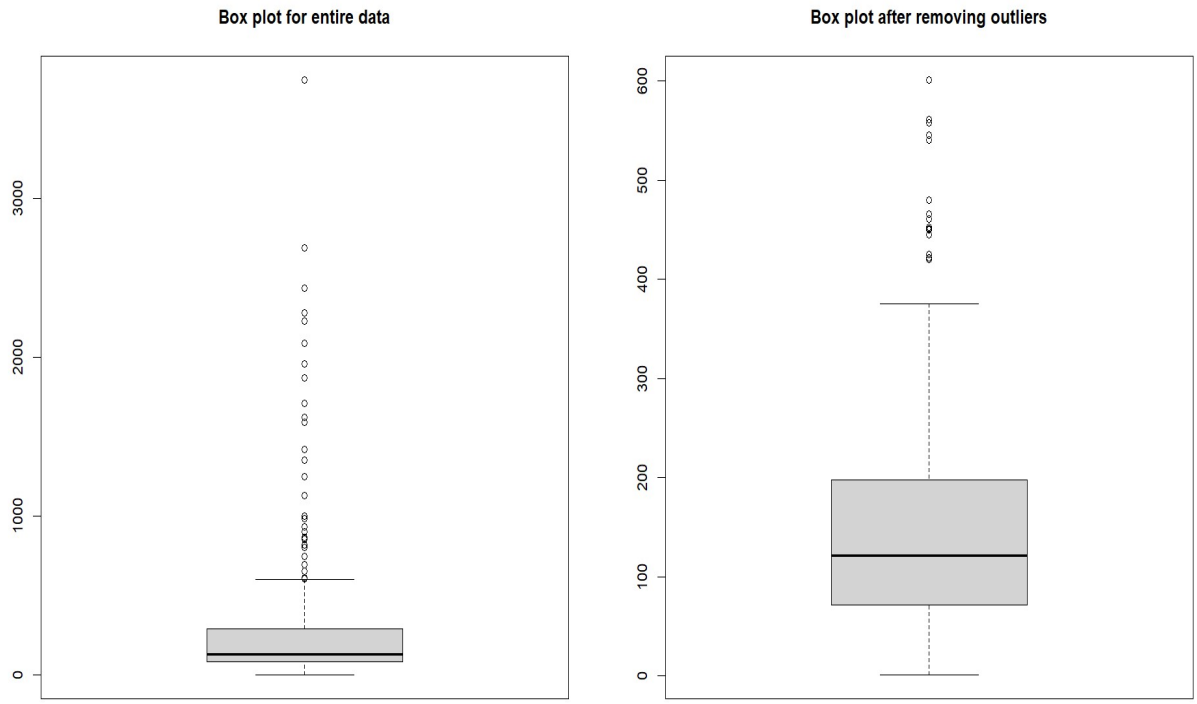
However, here the p-value is around zero, which suggests neither fit is good

Reason:- The main reason for this negligible p-value is the presence of too many outliers. Referring to the histogram for the data, it is evident that there are many values which heavily distort the measures of central tendency and dispersion.

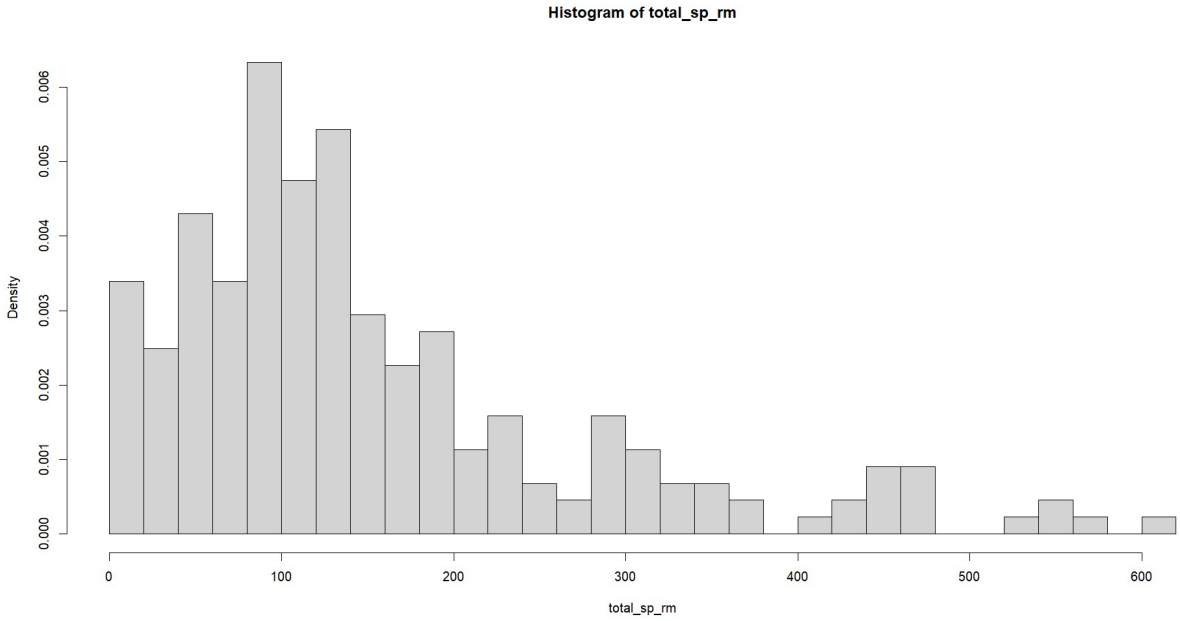
To cater to this problem, we shall fit a model after removing its outliers. The procedure is very similar to that used above.

Detection of outliers:- The easiest way of detecting outliers, is by drawing a box plot of the data. Here we compare the box plots of the data before and after removing the outliers.(Refer Appendix for R program).

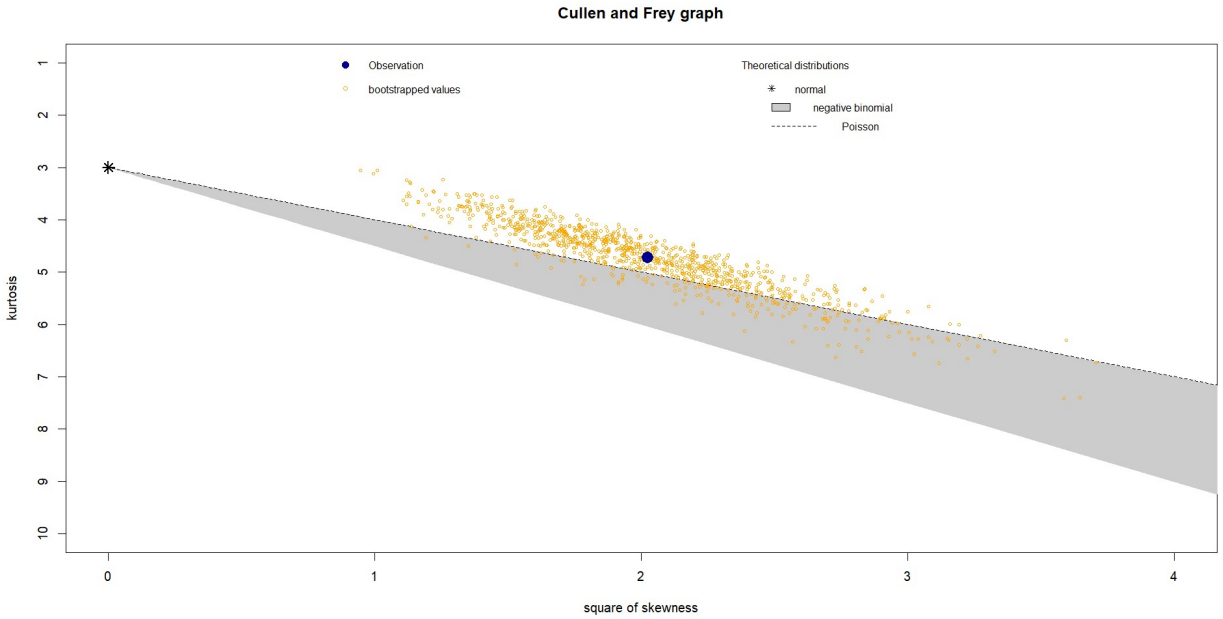
A total of 29 observations have been outliers and they have been removed to increase the goodness/reliability of the model.



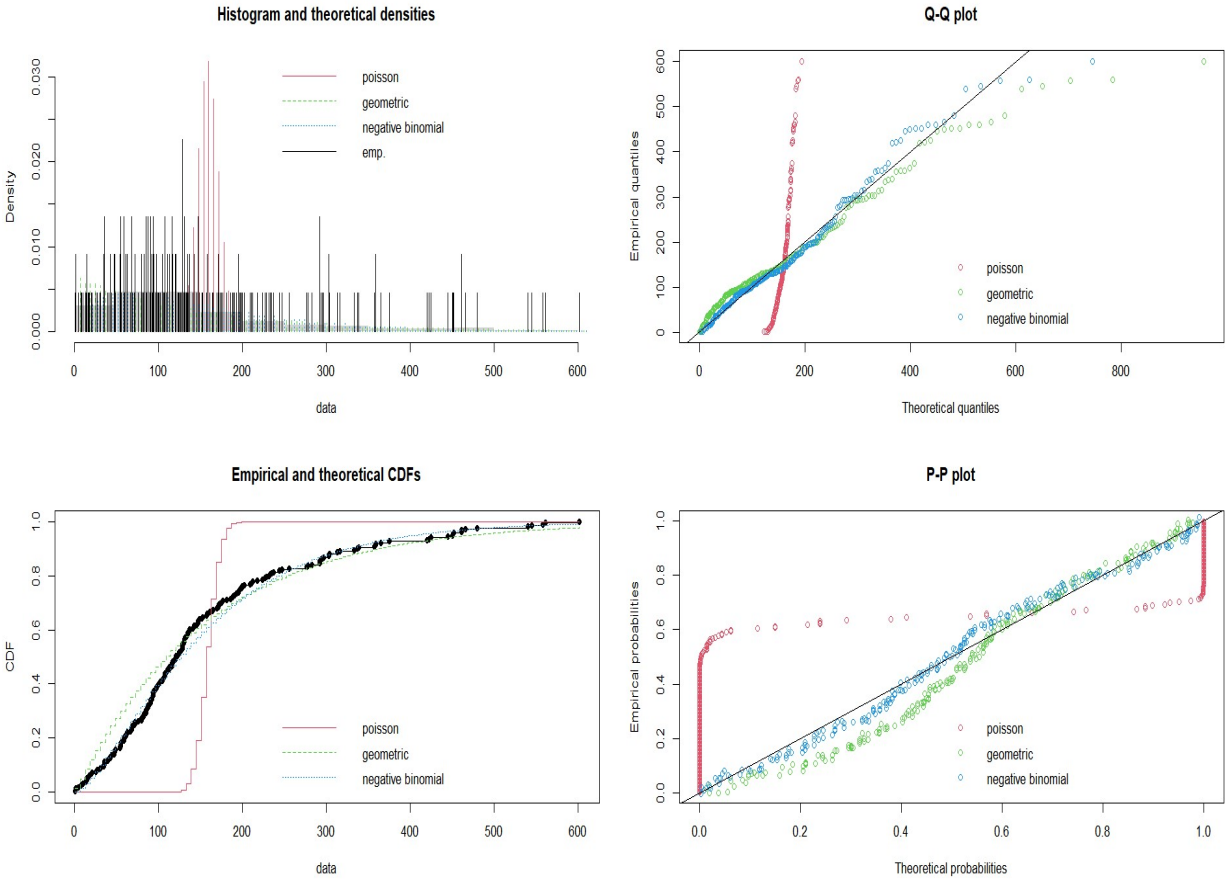
Thus, the corresponding histogram after removal of outliers was obtained.



Thus, the Cullen Frey graph for best fit(given below) suggests negative binomial and poisson distributions to be approximately good.



To get a better idea, we make use of the diagnostic plots as done earlier.



Thus it can be observed from the Q-Q,P-P plots and cumulative distribution plot that the negative binomial very accurately describes the data.

Hence it is important to now check the goodness of fit for fitted models(namely, negative binomial and geometric). The output has been attached for reference.

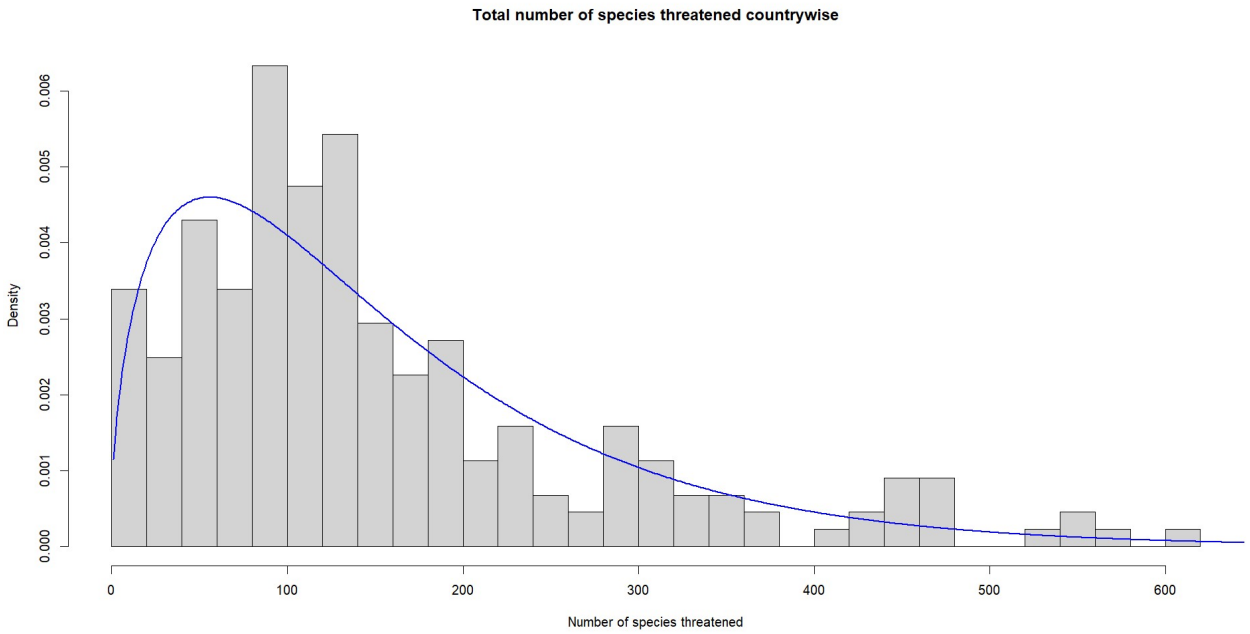
```
> gofstat(list(fnbin,fgeom),fitnames=c('negative binomial','geometric'))
Chi-squared statistic:  17.04175 39.04627
Degree of freedom of the Chi-squared distribution:  12 13
Chi-squared p-value:  0.1480336 0.0001965714
Chi-squared table:
       obscounts theo negative binomial theo geometric
<= 20         15                12.39665      27.670759
<= 44         15                22.60158      27.407395
<= 60         15                16.22553      16.077306
<= 80         15                19.97298      17.923942
<= 91         15                10.53160       8.927132
<= 102        15                10.09645       8.323029
<= 116        15                12.14274       9.783440
<= 129        16                10.52247       8.335670
<= 143        15                10.50421       8.237556
<= 171        16                18.48423      14.426642
<= 199        15                15.34654      12.069954
<= 247        15                20.08944      16.284706
<= 316        15                18.53529      16.193928
<= 452        15                16.37163      17.001643
> 452          9                 7.17868      12.336899

Goodness-of-fit criteria
                               negative binomial geometric
Akaike's Information Criterion          2658.472  2678.828
Bayesian Information Criterion          2665.268  2682.226
>
```

The observed p-value for the model of negative binomial distribution is 0.1480336, while that for geometric distribution is 0.0001965714, clearly suggesting that the negative binomial distribution is a good fit for the given data, as per the chi-square goodness of fit test.
The estimated parameters for the distribution are:-

```
> fnbin$estimate
      size         mu
  1.559133 156.518749
```

That is, the parameters of the negative binomial distribution are:- k= 1.559133 ≈ 2 and p=0.0098631 (or mean = 156.518749), that is

$$X \sim NB(k = 2, p = 0.0098631)$$



Why a negative binomial distribution?

A major indication that the distribution could be following a negative binomial distribution (and not a poisson) was the relation between mean and variance. Here, variance was larger than the mean clearly indicating that the distribution was either geometric or negative binomial(poisson distribution was fitted to the data for demonstration purpose).

Additionally, negative binomial distribution has a wide range of applications in the fields of biology and ecology, further making this a model that can be implemented in our case.

This model can help identifying particular countries that have a higher count of threatened species and those which need targeted conservation efforts further assisting in policy making and decision implementing for conserving a particular species.

Markov Chain Analysis of conservation status of species (taxonomy wise)

Overview

As described in the earlier sections, the IUCN allocates conservation statuses (DD,LC,NT, VU, EN,CR,EW and EX) to all living organisms on this planet. To estimate and quantify the extinction risk faced by a particular species(belonging to a particular taxonomy group), it is important to compute the chance of transitioning from one conservation status to other.

This can be done best by the concept of Markov chain, since for the determination of the upcoming conservation status it is enough to take into consideration the most recent one. Thus, we shall first begin by elaborating on the concept of Markov chain followed by computation of the one-step probability matrix (and the procedure of obtaining the same) so as to draw conclusions and inferences and predict the upcoming conservation status for a particular species. We can learn a lot about the patterns of the transition from one conservation status to other by making use of Markov chains.

Markov Chain

Introduction:-

Markov chain was first introduced by the Russian mathematician Andrey Markov, and it is a special case of Stochastic Processes(Discrete time and Discrete state space). It is a mathematical and a statistical model which helps in explaining the chance of transitioning from one state to other; where the probability of going in the next state is only influenced by the current state (recent past) of the random variable.

Definition:-

A stochastic process $\{X_n : n = 1, 2, 3, ...\}$ with a finite state space/countably infinite state space(S) is said to be a Markov Chain(M.C) if $\forall i, j, i_0, j_0, ..., i_{n-1} \in S$ and n=1,2,3,....∞

$$P[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, ..., X_2 = i_2, X_1 = i_1] = P[X_{n+1} = j | X_n = i]$$

Applications of Markov chain:-

Markov chains are highly applicable in various fields of human endeavours such as:-

- Animal population mapping
- Stock market analysis
- Rigourously used in statistical mechanics, astronomy and thermodynamics
- Weather forecasting
- Machine learning and artificial intelligence

Basic terminologies:-

Persistent state:-
A persistent,also known as recurrent state, of a Markov chain is a state which possesses a positive(non-zero) probability of re-entering to itself after any number of transitions, irrespective of the initial state.
Steady state:-
It describes the long term behavior of the Markov chain, after having being ran for a (sufficiently) long time. Thus in short, it describes the probability of X_n being in each state on an average after many transitions.

Formulation of the one step transition probability matrix for the given IUCN Red List dataset:-

For the computation of the One Step TPM (Transition Probability Matrix), we imported the dataset made available by IUCN Red List which displayed the conservation status of a particular species in the previous year and in the current year. A sample of the dataset has been attached on the following page for reference. (Figure 16)

Scientific name	Common name	IUCN Red List (2021) Category	IUCN Red List (2022) Category	Reason for change	Red List version
MAMMALS (Mammalia)					
Hypogeomys antimena	Malagasy Giant Jumping Rat	EN	CR	G	2022-1
Macaca fascicularis	Long-tailed Macaque	VU	EN	N	2022-1
Macaca nemestrina	Southern Pig-tailed Macaque	VU	EN	N	2022-1
Mesocricetus auratus	Golden Hamster	VU	EN	N	2022-1
Microcebus lehilahytsara	Goodman's Mouse Lemur	VU	NT	N	2022-1
Plecturocebus vieirai	Vieira's Titi Monkey	DD	CR	N	2022-1
Presbytis comata	Javan Surili	EN	VU	N	2022-1
Pteropus vampyrus	Large Flying-fox	NT	EN	N	2022-2
Pygeretmus zhitkovi	Greater Fat-tailed Jerboa	NT	LC	N	2022-1
Rhinolophus hillorum	Upland Horseshoe Bat	NT	VU	N	2022-2
Rhinopithecus brelichi	Grey Snub-nosed Monkey	EN	CR	N	2022-1
Sapajus cay	Azara's Capuchin	LC	VU	G	2022-1
BIRDS (Aves)					
Acanthiza katherina	Mountain Thornbill	LC	VU	G	2022-1
Accipiter imitator	Imitator Goshawk	VU	NT	N	2022-1
Accipiter nanus	Dwarf Sparrowhawk	NT	LC	N	2022-1
Acrocephalus orinus	Large-billed Reed-warbler	DD	LC	N	2022-1
Acrocephalus sorghophilus	Streaked Reed-warbler	EN	CR	G	2022-1
Actenoides bougainvillei	Moustached Kingfisher	EN	LC	N	2022-2
Actenoides princeps	Scaly-breasted Kingfisher	NT	LC	N	2022-2
Afropavo congensis	Congo Peafowl	VU	NT	N	2022-1
Amazona bodini	Northern Festive Amazon	NT	LC	N	2022-1
Amazona farinosa	Southern Mealy Amazon	NT	LC	N	2022-1
Amazona festiva	Southern Festive Amazon	NT	LC	N	2022-1
Amazona kawalli	White-faced Amazon	NT	LC	N	2022-1
Amytornis ballarae	Kalkadoon Grasswren	LC	VU	G	2022-1
Amytornis housei	Black Grasswren	NT	LC	N	2022-1
Amytornis woodwardi	White-throated Grasswren	VU	EN	G	2022-1
Apalis chirindensis	Chirinda Apalis	LC	VU	G	2022-1
Apalis sharpii	Sharpe's Apalis	LC	NT	G	2022-1
Aphelocephala leucopsis	Southern Whiteface	LC	VU	G	2022-1
Aratinga auricapillus	Golden-capped Parakeet	NT	LC	N	2022-1
Arborophila mandellii	Chestnut-breasted Partridge	VU	NT	N	2022-1
Archboldia papuensis	Archbold's Bowerbird	NT	LC	N	2022-2
Argya rufescens	Orange-billed Babbler	NT	LC	N	2022-1
Asthenes helleri	Puna Thistletail	VU	LC	N	2022-1
Astrapia mayeri	Ribbon-tailed Astrapia	NT	LC	N	2022-2
Atlapetes nigrifrons	Black-fronted Brush-finch	NT	LC	N	2022-2
Atlapetes rufigenis	Rufous-eared Brush-finch	NT	LC	N	2022-1
Bolemoreus hindwoodi	Eungella Honeyeater	LC	NT	G	2022-1
Botaurus poiciloptilus	Australasian Bittern	EN	VU	G	2022-2
Bubo sumatranus	Barred Eagle-owl	LC	NT	G	2022-1
Calidris acuminata	Sharp-tailed Sandpiper	LC	VU	G	2022-1
Callaeas wilsoni	North Island Kokako	NT	LC	N	2022-2
Callocephalon fimbriatum	Gang-gang Cockatoo	LC	VU	G	2022-1
Calyptorhynchus lathami	Glossy Black-cockatoo	LC	VU	G	2022-1
Campephilus gayaquilensis	Guayaquil Woodpecker	NT	LC	N	2022-1
Caprimulgus ruficollis	Red-necked Nightjar	LC	NT	G	2022-1
Carpococcyx radiceus	Bornean Ground-cuckoo	NT	VU	G	2022-1

Figure 10: Conservation statuses, for every species, taxonomy wise in the years 2021 and 2022

Remark:- The dataset only considers the species which have changed their conservation status and does not take into account those which have not.

REPTILES									
	DD	LC	NT	VU	EN	CR	EW	EX	
DD	0	140	17	19	39	14	0	0	**229**
LC	20	0	41	18	10	2	0	0	**91**
NT	3	64	0	21	20	2	0	0	**110**
VU	12	36	22	0	50	23	1	2	**146**
EN	10	13	9	21	0	39	0	1	**93**
CR	6	1	2	4	21	0	0	1	**35**
EW	0	0	0	1	0	1	0	1	**3**
EX	0	0	0	0	0	3	1	0	**4**

Figure 11: Total count of transitions taking place between the conservation statuses

Such datasets were also available for the previous years starting from 2007 to present(2022). These datsets were then imported in an excel file,cleaned and sorted out taxonomy wise (that is, species of mammalia, reptilia,amphibians,birds and plantae taxons were grouped separately) for further analysis.
Further, a count of the transitions taking place between the conservation statuses was taken taxonomy wise, as explained below.
For instance, the number of transitions that took place from the conservation status DD to EN is 39 (simply put, it indicates the number of species ,which were in the previous year under consideration, DD and in the following year they changed their status to EN)

The last column indicates the total number of transitions that took place **given some initial** conservation status. The individual probabilities were thus obtained by taking the proportion, that is,by dividing individual row values by the total value.
Thus, in the given scenario, the data can be expressed as a Markov chain as:-

$\{X_n : n \geq 1\}$ is a Markov chain with state space S={DD,LC,NT,VU,EN,CR,EW, EX} and the random variable X_n can be described as :
X_n : conservation status of a particular species belonging to a particular taxonomy group

The one step transition probability matrix for taxon reptilia was thus obtained as:-

REPTILES								
	DD	LC	NT	VU	EN	CR	EW	EX
DD	0	0.61135	0.07424	0.08297	0.17031	0.06114	0	0
LC	0.21978	0	0.45055	0.1978	0.10989	0.02198	0	0
NT	0.02727	0.58182	0	0.19091	0.18182	0.01818	0	0
VU	0.08219	0.24658	0.15068	0	0.34247	0.15753	0.00685	0.0137
EN	0.10753	0.13978	0.09677	0.22581	0	0.41935	0	0.01075
CR	0.17143	0.02857	0.05714	0.11429	0.6	0	0	0.02857
EW	0	0	0	0.33333	0	0.33333	0	0.33333
EX	0	0	0	0	0	0.75	0.25	0

Figure 12: One step transition probability matrix for reptiles

Similarly, repeating the above procedure, the one step transition probability matrix for different taxonomies(mammals,birds,amphibians and plants) was obtained.

The further Markov chain analysis was performed on R-Software.(Refer Appendix for code)

```
MC1
 A  8 - dimensional discrete Markov Chain defined by the following states:
 DD, LC, NT, VU, EN, CR, EW, EX
 The transition matrix  (by rows)  is defined as follows:
           DD         LC         NT         VU         EN         CR          EW         EX
DD 0.00000000 0.61135371 0.07423581 0.08296943 0.1703057 0.06113537 0.000000000 0.00000000
LC 0.21978022 0.00000000 0.45054945 0.19780220 0.1098901 0.02197802 0.000000000 0.00000000
NT 0.02727273 0.58181818 0.00000000 0.19090909 0.1818182 0.01818182 0.000000000 0.00000000
VU 0.08219178 0.24657534 0.15068493 0.00000000 0.3424658 0.15753425 0.006849315 0.01369863
EN 0.10752688 0.13978495 0.09677419 0.22580645 0.0000000 0.41935484 0.000000000 0.01075269
CR 0.17142857 0.02857143 0.05714286 0.11428571 0.6000000 0.00000000 0.000000000 0.02857143
EW 0.00000000 0.00000000 0.00000000 0.33333333 0.0000000 0.33333333 0.000000000 0.33333333
EX 0.00000000 0.00000000 0.00000000 0.00000000 0.0000000 0.75000000 0.250000000 0.00000000
```

Figure 13: One step TPM in R

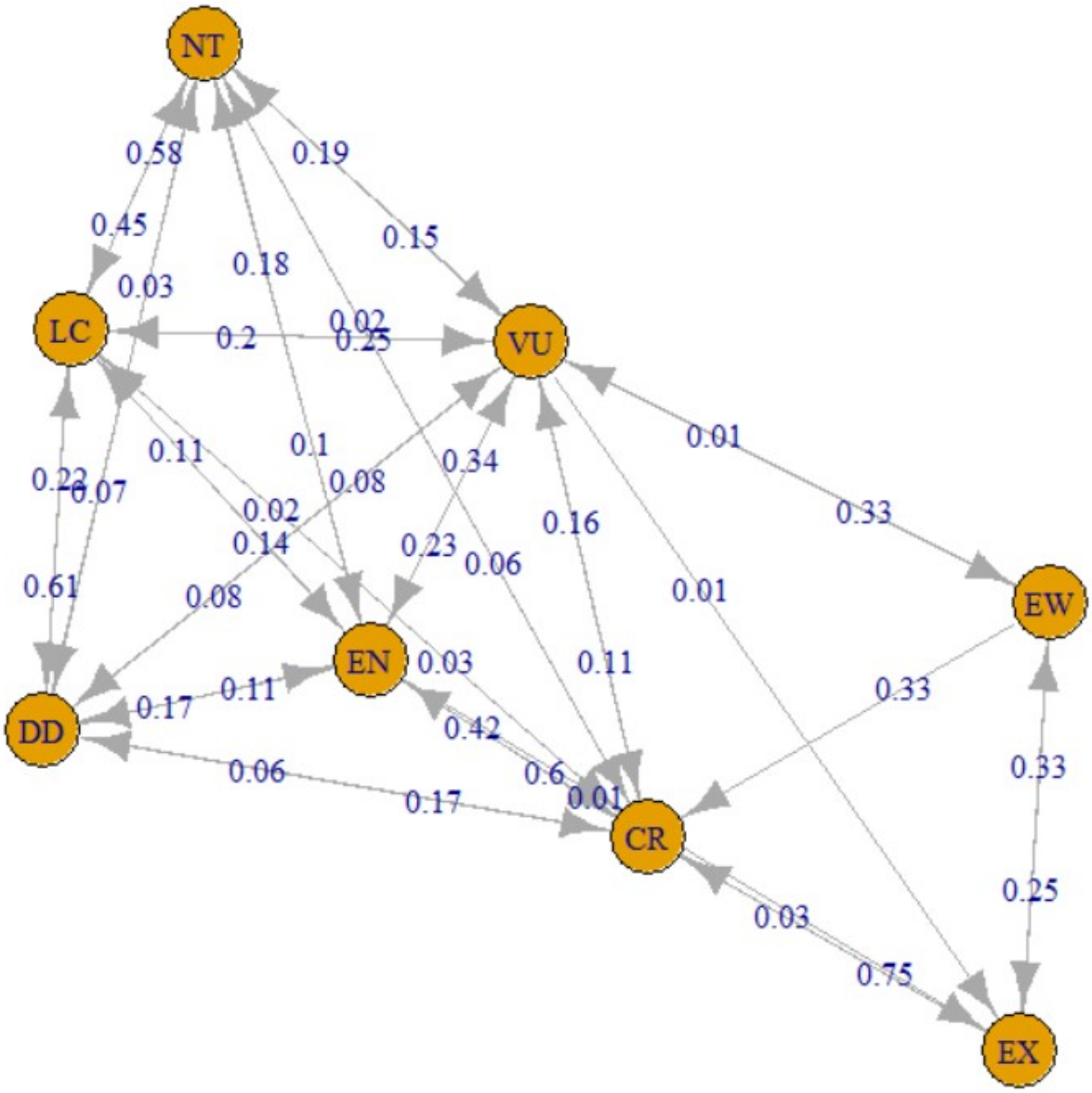


Figure 14: Sample Markov Chain plot

To begin with the analysis of the above Markov chain, it imperative to understand the nature of the individual states(that is, to check which states are persistent,transient,absorbing etc.) so as to get some initial insights.

Using the R code given in the appendix, it was inferred that all the states are **persistent**

```
> summary(ml_trans_r)
MCl  Markov chain that is composed by:
Closed classes:
DD LC NT VU EN CR EW EX
Recurrent classes:
{DD,LC,NT,VU,EN,CR,EW,EX}
Transient classes:
NONE
The Markov chain is irreducible
The absorbing states are: NONE
```

Figure 15: Corresponding output,summarising nature of all states

Interpretation:- It indicates that all states are persistent, that is regardless of the initial/current conservation status of a particular species, it has positive chances of re-entering other states; which means although any species's current conservation status may not indicate that it is threatened, in the near future, it is susceptible towards being threatened and facing increasing risks of extinction.

Thus based on the above interpretation, it can be said that, irrespective of whether a species is threatened or not, it cannot be directly inferred that a species classified as LC(Least Concern) or NT(Near Threatened) need not require any initial care and precautions for their conservation. Hence, when it comes to conservation of bio-diversity, neglecting the biological importance of species in the LC or NT categories is neither advisable nor correct.

After having understood the nature of the individual states, it is also important to comprehend what would possibly be the conservation status of a particular species after a certain number of years. As time increases (or as $t \to \infty$), however it is observed that the probabilities in the one step TPM converge to a particular value. This principle is called as steady states.

For demonstration, a plot was made by plotting the transition probabilities against the probabilities of TPM at the nth year (or nth step) with 'DD' as a conservation status being the initial one.(for simplicity and ease, graphs for transitions to states 'EW' and 'EX' have not been plotted)

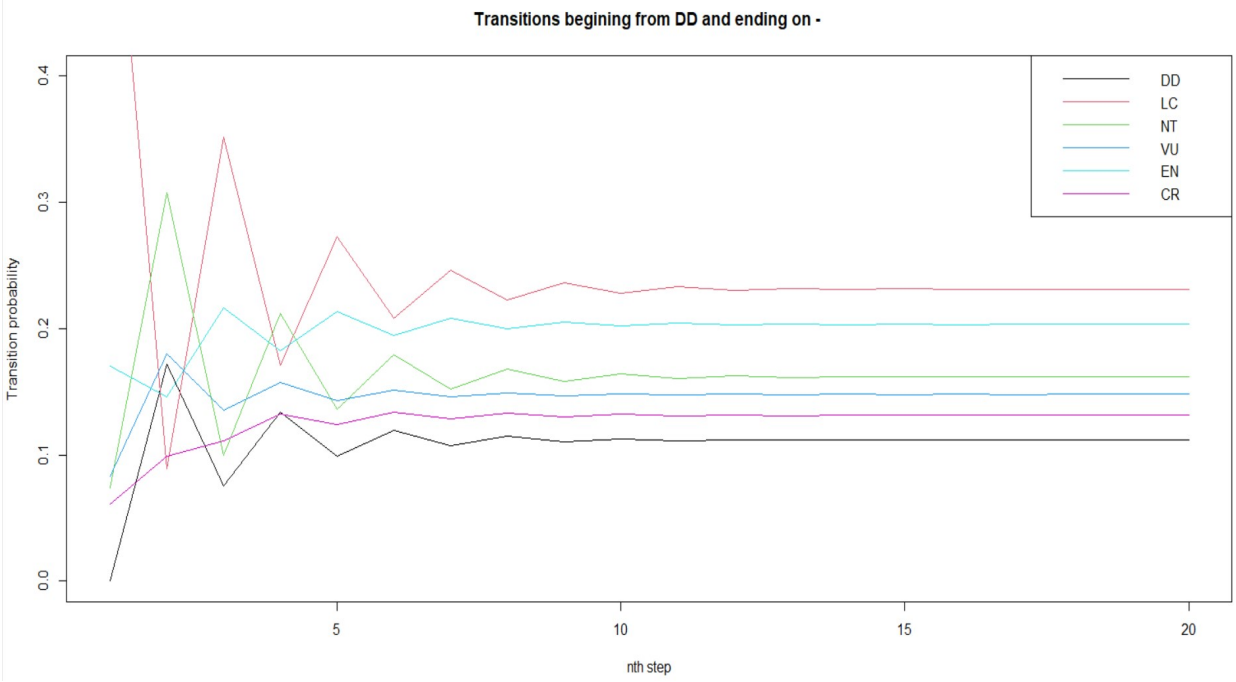


Figure 16: Plot demonstrating steady states

The steady states for the taxonomy of reptiles was thus obtained as:-

```
> ss=steadyStates(ml_trans_r);ss
           DD       LC        NT        VU        EN        CR          EW          EX
[1,] 0.1117838 0.231238 0.1619645 0.1479608 0.2033732 0.1313424 0.003278189 0.009059037
>
```

Figure 17: Steady state vector of animals belonging to taxonomy reptiles

For visualisation,the plot of the steady state is :-

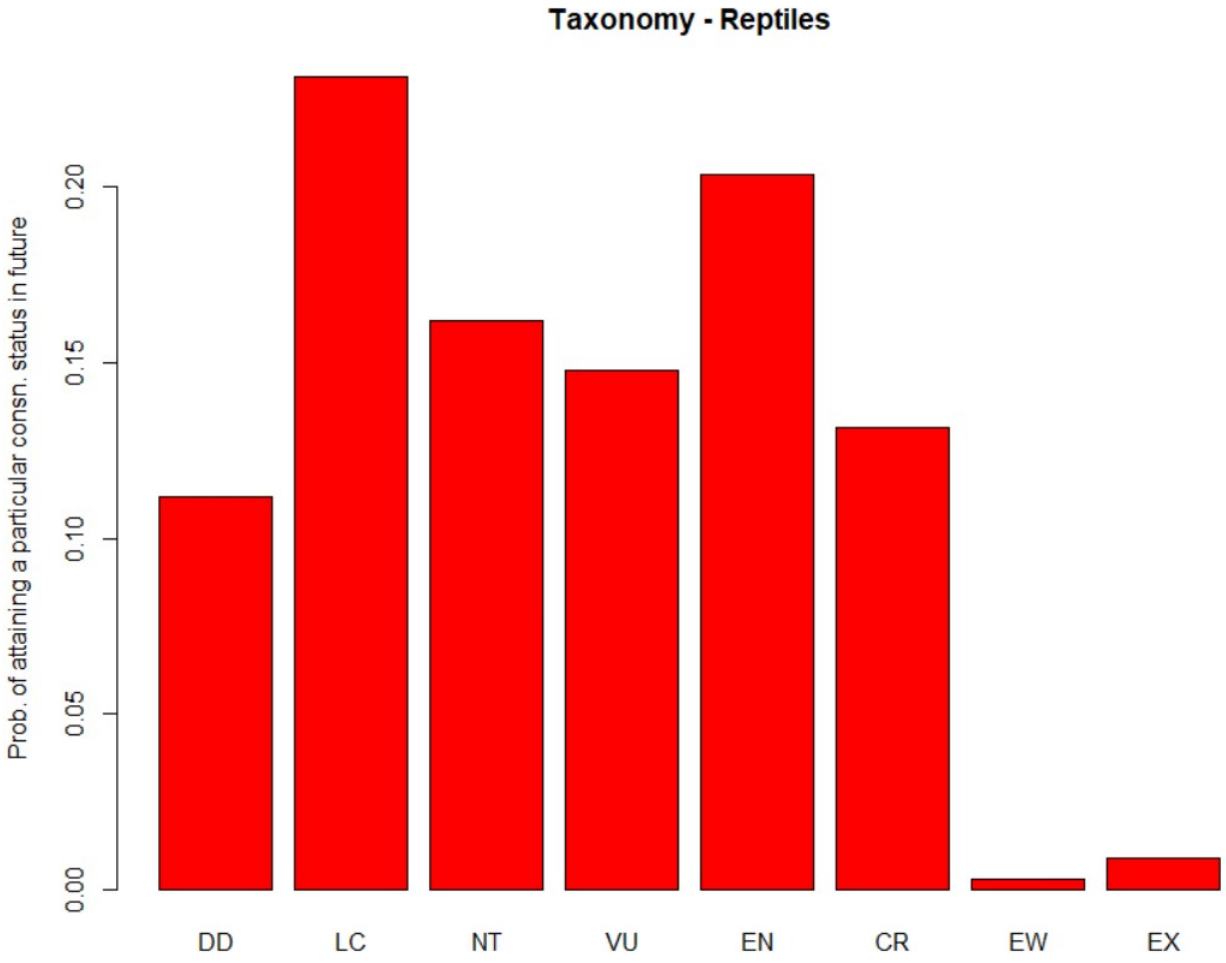


Figure 18: Bar plot of the steady states

Interpretation:- The above vector gives us the long term probabilities of attaining a given conservation status. For instance, we can infer that the probability of a particular species(from taxonomy reptilia) will become endangered is approximately 0.2033732.

Simulations using R

```
> sim_r1=rmarkovchain(n=20,ml_trans_r,t0='LC');sim_r1
 [1] "EN" "LC" "NT" "LC" "VU" "LC" "NT" "VU" "CR" "EN" "LC"
[12] "DD" "LC" "NT" "VU" "EN" "DD" "LC" "VU" "LC"
> sim_r1=rmarkovchain(n=20,ml_trans_r,t0='NT');sim_r1
 [1] "LC" "VU" "NT" "VU" "EN" "VU" "LC" "DD" "EN" "CR" "EN"
[12] "VU" "LC" "DD" "LC" "NT" "VU" "NT" "LC" "EN"
> sim_r1=rmarkovchain(n=20,ml_trans_r,t0='VU');sim_r1
 [1] "NT" "VU" "EN" "VU" "CR" "EN" "CR" "DD" "LC" "DD" "CR"
[12] "EN" "VU" "DD" "LC" "NT" "LC" "DD" "LC" "NT"
> sim_r1=rmarkovchain(n=20,ml_trans_r,t0='EN');sim_r1
 [1] "DD" "LC" "NT" "EN" "CR" "EN" "VU" "EN" "CR" "EX" "EW"
[12] "VU" "LC" "NT" "LC" "NT" "VU" "EN" "LC" "NT"
> sim_r1=rmarkovchain(n=20,ml_trans_r,t0='CR');sim_r1
 [1] "EN" "CR" "VU" "EN" "CR" "EN" "DD" "LC" "NT" "LC" "NT"
[12] "VU" "NT" "VU" "NT" "LC" "VU" "LC" "NT" "LC"
>
```

Figure 19: Simulation for different conservation status of taxon reptilia for a period of 20 years

In this very similar manner, the nature of the states and steady states for the other taxons were computed and their bar plots were plotted for better visualisation.

The nature of the states of the Markov chains of different taxons are:-

Taxon name	Recurrent states	Nature of Markov chain
Mammals	All	Irreducible
Birds	All	Irreducible
Amphibians	All	Irreducible
Plants	All	Irreducible

The steady states for the remaining taxons are:-

```
> ssm                    # Steady states for mammals taxon
            DD        LC        NT        VU        EN        CR         EX
[1,] 0.03393349 0.1472259 0.1655445 0.2576513 0.2612462 0.1235653 0.01083333
> ssb                    # Steady states for birds taxon
             DD        LC       NT        VU        EN        CR          EX
[1,] 0.004723902 0.2383209 0.281048 0.2318464 0.1654953 0.0708089 0.007756653
> ssa                    # Steady states for amphibians taxon
            DD        LC        NT        VU        EN        CR         EW          EX
[1,] 0.07251541 0.2333064 0.1411952 0.1997426 0.2029122 0.1179508 0.02537913 0.006998365
> ssp                    # Steady states for plants taxon
           DD        LC        NT        VU        EN        CR          EW          EX
[1,] 0.070683 0.2244427 0.1362042 0.2185463 0.2285773 0.1085713 0.008620126 0.004355124
>
```

Figure 20: Steady states of Mammals,Birds, Amphibians and Plants taxon respectively

The corresponding bar plots for the steady state of each taxon are:-

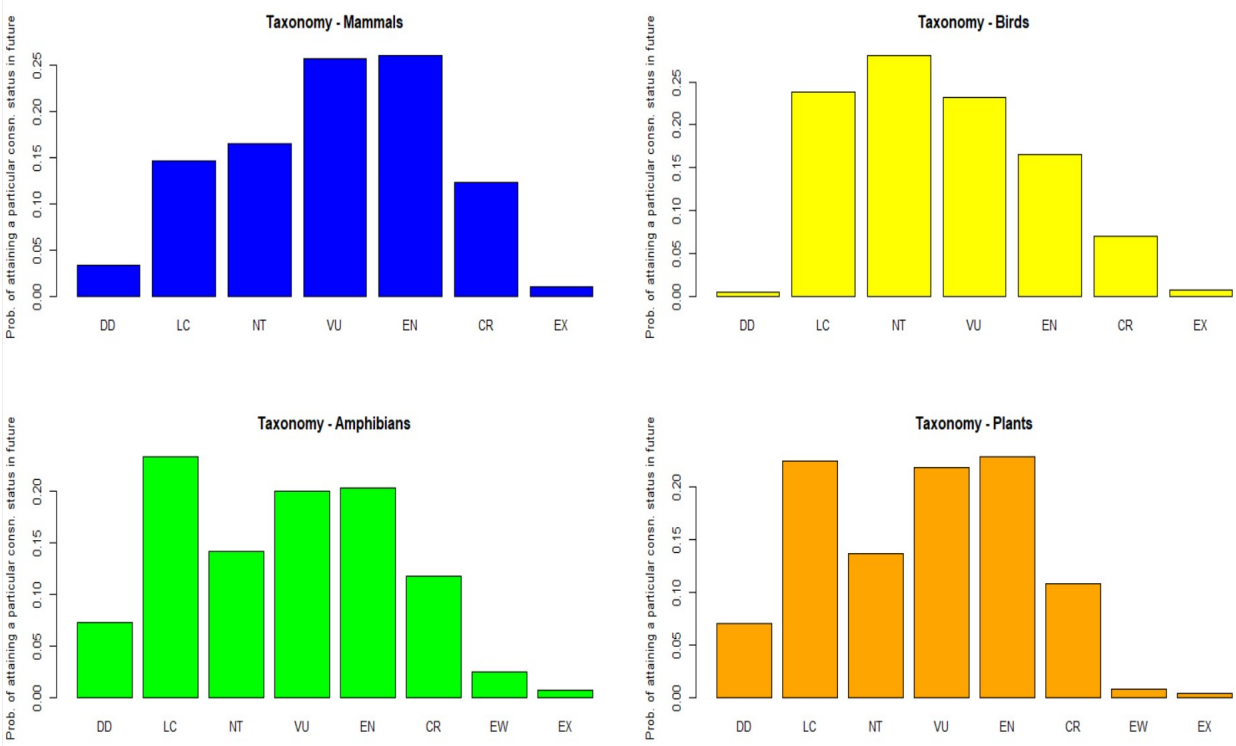


Figure 21: Visualisations of steady states

Interpretation:- All the states for the Markov chains of all the taxonomies are persistent, which clearly implies (as stated in the interpretation of persistent states for taxon reptilia) that, although a species might be classified currently as NT or LC, it does not indicate that it will not face any risks of being threatened or extinct in the near future.

Thus, every floral and faunal species, irrespective of its current conservation status can need special care, measures and human interventions for their conservation and protection when necessary.

Additionally, it is also evident from the plots of steady states that, for all the taxonomy groups,in the long run there is a higher probability that a species will be more prone towards being threatened in the far future, further increasing their risks of extinction; simply put, every taxon is in trouble (with amphibians being the most threatened and susceptible to immediate extinction). This is an **extremely** serious issue and should be addressed to immediately if the extinction of the majority of the extant species is to be prevented, or at least warded off.

Conclusion

We have now estimated the extinction probabilities of all species which belong to a particular taxonomy group. Additionally, we have also inferred that a massive number of species are under increasing risks of being threatened and extinction. According to the United Nations, 'Around 1 million animal and plant species are now threatened with extinction, many within decades, more than ever before in human history'(IPBES Report 2019). Thus we are in an **immensely huge** extinction crisis, one never witnessed by mankind, which entirely is of our making.

What can be done?

It will take many generations to undo what has been done so far. Whatever actions and steps that are to be taken for curbing the increasing extinction risk faced by all flora and fauna, should be taken now, at all levels from individual to global.

To begin,it is initially imperative to make people aware of the ongoing extinction crisis and the adverse effects on human life associated with it. It is also of utmost importance to let the general public know about the fragility of our ecosystem and how it works and the role of all living organisms in it.

The next step is to take action! It all begins with individuals taking the necessary steps and actions like switching to sustainable products, carefully observing and scrutinizing how subtle actions adversely affect the environment, avoiding products made from animal parts like leather, crocodile skins, ivory show pieces, etc, reducing personal carbon footprint to reduce the impact of climate change and so on.

The process is extremely slow but it will be extremely fruitful in the long run.

Remarks

For the taxons mammals and birds, very few observations were belonging to the category EW (Extinct in the Wild) and to avoid inflated probabilities, they have been put in the status EX(Extinct).

Predictive modelling using Random forest for finding conservation status of species

Intro

If we continue our chain of thought from the previous section on Markov chain which states that the current conservation status of a species is essentially only dependent on its previous year's status. In this section we use the same data as the previous one, but take a look at it from a Data Science perspective.

We have made a Random Forest classifier and trained on our data where it models the current conservation status of a species using some explanatory variables.

Terminology

- **Feature:** Generally also referred to as Independent variable, explanatory variable, or attribute.
- **Label:** Generally also referred to as Dependent variable, Response variable.
- **Greedy algorithm:** A greedy algorithm is an algorithm makes the locally optimal choice at each step without looking back to it in the future.
- **Bagging:** In ensemble learning, it is like voting for the class with same classifier trained on different datasets. It is also called bootstrap aggregation.

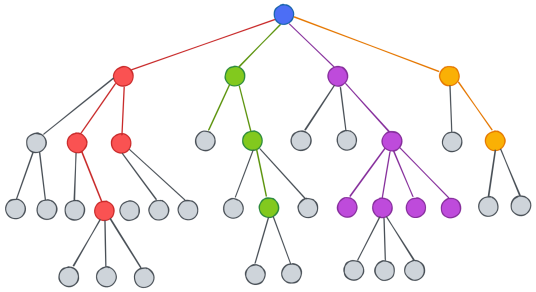
Decision Trees

Let us say every data point in our dataset consists of a feature vector x and a corresponding label y.Then, we denote the entire dataset by

$$D = \{(X, Y)\} = \{(x^{(i)}, y^{(i)}\}_{i=1}^{n}$$

Where $y^{(i)}$ is a categorical variable for a classification problem. In our case, it is the conservation status of the species for a year, and the feature vector consists of its taxonomy and its conservation state for the previous year.

Decision Trees are greedy non-parametric supervised learning algorithms which can be used for classification as well as regression. There are a few algorithms used to make decision trees, most popular being Iterative dichotomizer 3 (ID3) and Classification and Regression Tree (CART). The structure of a descision tree goes as follows:



A decision tree consists of one root node, decision nodes and leaf nodes.

- **Root node:** We begin with asking "which attribute should be tested at the root?" This attribute will become our root node. Here, the entire population is being analysed.
- **Decision nodes:** Here, we make 'splits' in our data and check conditions to further divide our dataset for classification. They are based on attributes.
- **Leaf node or terminal node:** They are nodes at the end of the tree which do not split into more nodes. they each consist of a class.

Measuring Impurity in a node

A node's purity is measured on the basis of how many datapoints of different categories are present in its region.
Let $p_{i,j}$ be the proportion of data points in node i belonging to class j. Some of the common meaures of impurity are:

- **Gini Index:** We classify observations to class j with probability $p_{i,j}$. Formula of Gini Index is given by

$$G_i = \sum_{j=1}^{J} p_{i,j}(1 - p_{i.j})$$

- **Entropy:**It is the disorder among a node. It is given by the formula:

$$H_i = -\sum_{j=1}^{n} p_{i,j} \log_2 p_{i,j}$$

- **Information Gain:** It is the expected reduction in entropy.

$$\text{IG(attr)} = \text{entropy of dataset - entropy of attribute}$$

Training

Step 1: Select the root node, or the best attribute for splitting the data. This is done using feature selection which utilizes impurity measures.
Step 2: Divide the root node into subsets that contains possible values for the best attributes.
Step 3: Generate the decision tree node, which contains the best attribute.
Step 4: Repeat from step two but using the current most leaf nodes until we reach a stopping criterion.

When do we stop splitting?

There are a few stopping points, and we stop if we reach any one of them.

- Every observation in the node belong to the same class.
- There are no more attributes on the basis of which we can make a split. Here, we label all the observations with the majority class.
- We are left with no more observations in the node.

One of the biggest disadvantages of Decision trees is that they have high variance due to being non parametric. To rectify this, we use ensembles.

Random Forest

Ensembles use multiple models and combine their predictions to get better performace as compared to individual models.
Random forest is an Ensemble learning method that uses multiple decision trees to reduce variance and get better performace. A random forest algorithm uses bagging with decision trees.

Step 1: Take k samples from our dataset, $D_1, D_2, ..., D_k$ with replacement.
Step 2: On each D_i, we train a Decision tree, but only taking some of the features for splitting in each one. **Step 3:** Our model is:

$$model = \frac{1}{k} \sum_{j=1}^{k} h_j(x)$$

Random Forest Using Sci-kitlearn

Sklearn implements the CART algorithm by default for training decision trees as well as random forest.

e Our dataset:

	category	IUCN Red List (2021)	IUCN Red List (2022)
0	MAMMALS (Mammalia)	EN	CR
1	MAMMALS (Mammalia)	VU	EN
2	MAMMALS (Mammalia)	VU	EN
3	MAMMALS (Mammalia)	VU	EN
4	MAMMALS (Mammalia)	VU	NT
...	...	...	...
2019	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	CR	EN
2020	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	EN	CR
2021	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	VU	EN
2022	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	EX	CR
2023	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	NT	VU

We can see that both our features and labels are categorical variables. sklearn used an optimized version of CART by default. There are many advantages of CART over ID3 or other algorithms like, being faster and more efficient as it is simpler and only uses binary trees. But since CART produces only binary trees, i.e., non leaf nodes will only be splitted into two categories. Thus, our two features need to be One Hot encoded.

One Hot Encoding: Say we have a categorical feature with k classes. Then One Hot encoding it replaces it with k binary features. after doing that, our data looks like:

	AMPHIBIANS (Amphibia)	BIRDS (Aves)	CYADS (Cycadopsida)	FISHES AND AQUATIC SPECIES	FISHES AND OTHER AQUATIC SPECIES	FLOWERING PLANTS (Liliopsida and Magnoliopsida)	HYDROZOANS (Hydrozoa)	INSECTS	MAMMALS (Mammalia)	MOLLUSCS
0	0	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...
2019	0	0	0	0	0	1	0	0	0	0
2020	0	0	0	0	0	1	0	0	0	0
2021	0	0	0	0	0	1	0	0	0	0
2022	0	0	0	0	0	1	0	0	0	0
2023	0	0	0	0	0	1	0	0	0	0

Now, We split the dataset into training and test set so that we can evaluate it later.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size = 0.25)
```

Now, We a train random forest classifier on our training set. We use the Gini index criterion for feature selection in the decision nodes.

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 100,
                                    criterion = 'gini')
classifier.fit(X_train, y_train)
```

Predicting the test set results:

```
y_pred = classifier.predict(X_test)
```

Evaluating the results

Now, we evaluate our predictions using confusion matrix and classification metrics like precision recall and f1 score.

```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)
```

```
                LC   NT  VU  EN  CR  EX
LC      [[205  10  10  12   3   0]
NT       [ 35  18   6   4   6   1]
VU       [ 41   4  11  10   5   0]
EN       [ 53   5   0  22   0   0]
CR       [ 27   2   0   3   9   1]
EX       [  1   0   0   2   0   0]]
```

	precision	recall	f1-score	support
0	0.57	0.85	0.68	240
1	0.46	0.26	0.33	70
2	0.41	0.15	0.22	71
3	0.42	0.28	0.33	80
4	0.39	0.21	0.28	42
5	0.00	0.00	0.00	3
accuracy			0.52	506
macro avg	0.37	0.29	0.31	506
weighted avg	0.49	0.52	0.48	506

Metrics

Here we can see that the accuracy of our model on unseen data is approximately 52%. It means that the model is predicting the status of the species correctly 52% of the time. But, we have to keep in mind that there is a limitation of accuracy as a metric. Suppose we make the given model,

$$\text{model} = \text{argmax}(\text{Response Variable of data set})$$

Then, this model will classify any species into the class with the maximum number of species regardless of what the input parameters were, thus making it a not very ideal metric in our data due to the high difference in number of species in the red list categories.

Precision: Also called positive predictive value, it is measuring the proportion of the species in each red list category that actually belonged in it.

Recall/Sensitivity: It is measuring the proportion of species in each red list category whose status was predicted correctly.

F1- Score: It is the harmonic mean of precision and recall.

Interpretation

From looking at the metrics we can conclude that the previous year's conservation status of a species does affect it for the consecutive year but there are several other factors for example climate change, habitat loss and degradation, environmental contamination and many more that can affect the future conservation status of a species and all of them should be taken into consideration to avoid any species facing a serious extinction crisis.

Difference in Difference Analysis for testing the effectiveness of Wildlife Amendment Act 2002

Overview

Now that we have analysed the data that has been collected, compiled and made available by the IUCN and have assessed certain factors which influence biodiversity risks as well as the conservation status of innumerable living organisms, we will also quantify the effects of certain laws and regulations that have been implemented to counter biodiversity loss.

The Government of India, for instance, implemented a conservation program in 1973, which we know as 'Project Tiger' for countering the declining population of the Bengal Tiger, a tiger species which resides in India. But how effective have many such conservation programmes and policies been? Is the government doing enough for protecting this majestic animal?

We will now be assessing whether 'The Wildlife Protection Amendment Act (2002)', which had been enacted by the Parliament of India, was effective in the recovering the population of a particular species (here, Bengal Tiger) whose population was rapidly declining.

For this, we shall be employing the technique 'Difference-in-Difference' analysis. Before we begin let us briefly describe The Wildlife Protection Amendment Act (2002)

The Wildlife Protection Amendment Act (2002)

According to the Constitution of India, it is defined as :- "An Act to provide for the protection of wild animals, birds and plants and for matters connected therewith or ancillary or incidental thereto with a view to ensuring the ecological and environmental security of the country.".

The Wildlife Protection Act was first implemented in the year 1972, a year before 'Project Tiger', for protecting plants and animals by adopting various measures like :-

- Declaring and protecting areas of land of zoological and botanical significance as natural reserves,sanctuaries and/or national parks
- Prohibiting hunting and poaching of animals
- Prohibition of Cutting/Uprooting specified plants
- Constitution of various bodies (e.g National Tiger Conservation Authority(NTCA)) for protecting specific flora and fauna

Moreover, this act has divided the species of animals and plants according to their need of conservation, population trends and generation lengths into six schedules/lists that provide varying degrees of protection, with animals and plants belonging to the Schedule I (first schedule) being endangered species.

Animals like the Black Buck, **Bengal Tiger**, Kashmiri Stag, Blue Whales belong to the first schedule and are entitled to rigorous protection. Additionally, poachers or hunters are severely punished for harming,hunting or poaching species belonging to this category.
Thus, The Wildlife Protection Amendment Act (2002), just like other laws and regulations which require amendments, was enacted with the aim of decreasing the risk of extinction faced by the threatened species of India and is also known for aiding in countering the decreasing tiger populations and decreasing the cases of hunting of leopards nation wide.

The Difference-in-Difference(DiD) analysis

Introduction

The DiD analysis is a technique which is widely and extensively used in the field of economics, econometrics and social sciences for checking,testing and validating the effect of a policy,rule or regulation which has been implemented.The DiD model is a powerful and flexible **regression technique**(whose model is stated in the 'Implementation of DiD for testing effectiveness of Wildlife Amendment Act 2002' subsection later in the project) that can be used to estimate the differential impact of a 'treatment' on the treated group of individuals or things. Alternatively, DiD is typically used to estimate the effect of a specific intervention or treatment (such as a passage of law, enactment of policy, or large-scale program implementation) by comparing the changes in outcomes over time between a population that is enrolled in a program (the intervention group) and a population that is not (the control group).

Definition

DiD is a quasi-experimental approach that makes use of longitudinal data from treatment and control groups(defined below) to obtain an appropriate counterfactual to estimate a causal effect i.e the difference-in-difference method is a quasi experimental method that compares changes in outcomes over time between a population undergoing a certain treatment(a law, a program) and a population that is not.

The difference-in-difference (DiD) technique is often called as the 'controlled before-and-after study'.

Basic Terminologies

- **Treatment Group** :-
  Also known as experimental group, it is the group to which a particular treatment(for example, in our case; laws, programs, regulations, acts) are applied and the effects of the same are to studied by researchers, analysts etc.
- **Control Group** :-
  It is the group which does not receive any treatment or the treatment that is being studied and has been implemented in the corresponding treatment group.

- **Counterfactual**:-
  It is the estimated value a treatment group can take, had the treatment not taken place. Simply put, it is the value taken the treatment group if the intervention/treatment never happened or was not implemented.

Graphically, the difference-in-difference analysis can be depicted as shown in the following figure.

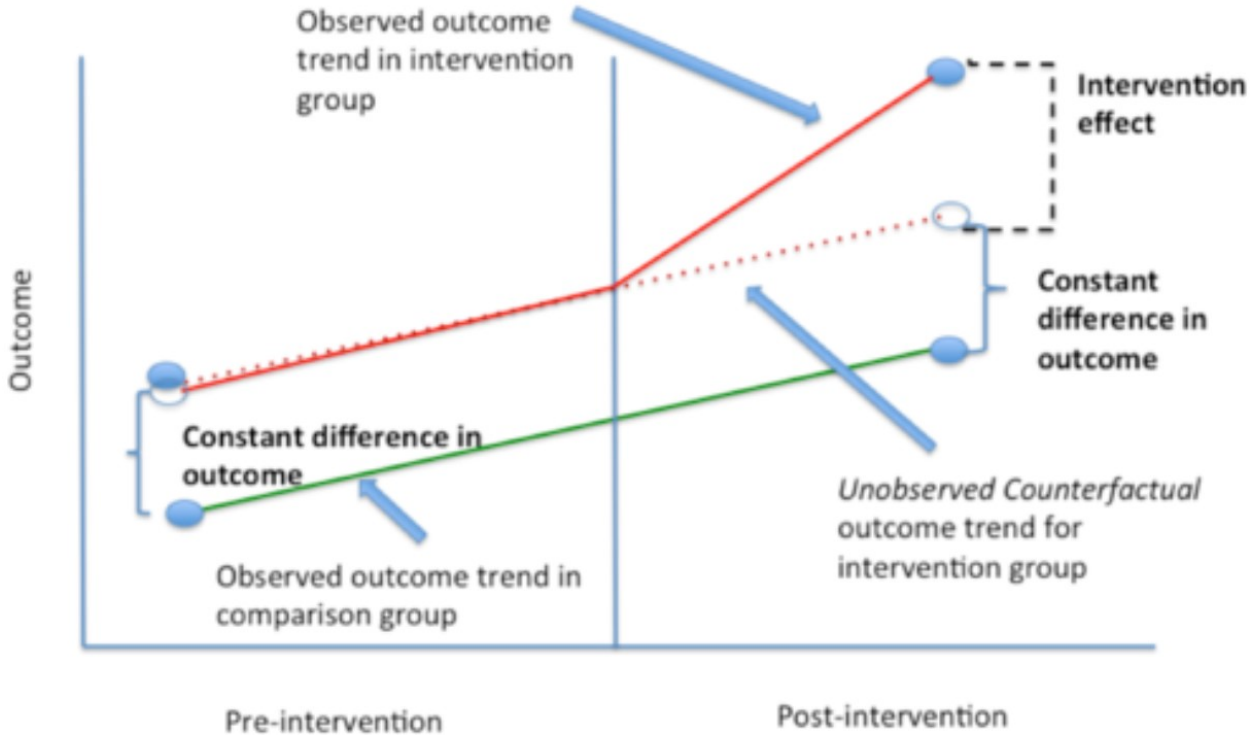


Figure 22: DiD Graphical representation

Assumptions

The DiD model is a powerful tool, provided it's assumptions hold true. Some of the assumptions are :-

- **Parallel trends**:-
  This assumption is the most important of all assumptions and also the most difficult to meet. Parallel trends assumptions states that the trends for the treatment group and the control group, in the period before the treatment/intervention was applied, should be parallel. This is an indication that there are no time varying factors and that the treatment was mainly responsible for the effects observed in the treatment group.
- **Treatment and control groups are comparable**:-
  The treatment and control groups should be similar in all relevant aspects, except for the treatment under study.

- **No spill over effects**:-
  There are no factors that affect or influence the outcome of either groups.

Implementation of DiD for testing effectiveness of Wildlife Amendment Act 2002

Year	Bengal	Malayan	Siberian	Sumatran
1900	44600	8900	9900	8400
1905	39600	7300	8400	7600
1910	38800	6900	8280	7480
1915	35600	5360	7830	6938
1920	31300	4200	7110	6300
1925	25100	3650	6415	5850
1930	24290	3200	6050	5382
1935	21400	2600	5260	4758
1940	18430	2210	4750	4429
1945	14250	1570	3750	3820
1950	11270	1150	3160	3415
1955	7200	910	2425	2950
1960	4230	690	1970	2570
1965	3140	610	1740	2230
1970	2480	550	1480	1943
1975	2368	480	1360	1630
1980	2260	430	1220	1380
1985	2210	390	1062	1150
1990	2300	341	880	920
1995	2280	270	720	660
2000	2220	245	690	580
2005	2380	230	580	515
2010	2540	300	450	440
2015	2980	380	470	400
2020	3010	383	490	400

Figure 23: Count of different tiger species worldwide (extant)

To begin with, the data for the extant tiger species worldwide has been collected (Figure 2). In our case,
Treatment Group:- Bengal tiger
Control Group:- Extant species of tigers (Sumatran, Siberian and Malayan)
We now check the assumption of parallel trends, the validity of the main assumption of the analysis (Figure 3).

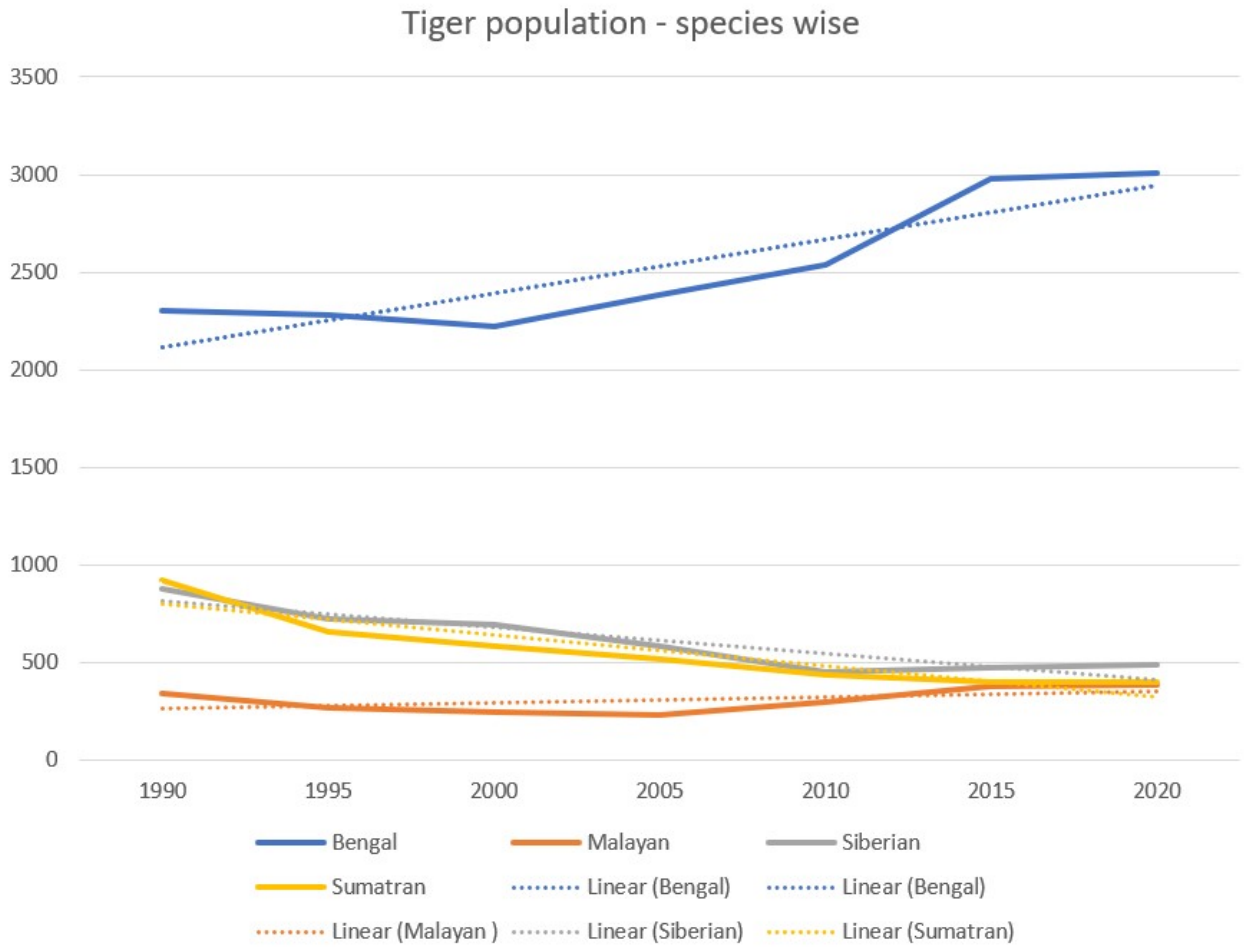


Figure 24: Graph for checking the validity of parallel trends assumption

Thus, from the graph it can be inferred that the trend of the treatment group (bengal tiger) is nearly/ approximately parallel to all the trends of all the control groups(Sumatran, Siberian and Malayan tiger). Since the effectiveness of treatment(Wildlife Amendment Act 2002) was implemented in 2002, we shall consider the count only for the years 1990 to 2015, since they will be sufficient for the evaluation. Furthermore,from the graph (approximately for the year 2002), it is evident that implementation of the Act (2002)caused a change in the trend of the count of the bengal tiger species.

Consider, control group as Sumatran tiger. The required figures for the analysis are:-

	1990	1995	2000	2005	2010	2015
Bengal	2300	2280	2220	2380	2540	2980
Sumatran	920	660	580	515	440	400

Figure 25: Tiger counts of bengal and sumatran tigers from 1990 to 2015

Now, in our case, the pre-treatment period are the years 1990,1995 and 2000, while the post treatment period are the years 2005,2010 and 2015. We want to check the **average effectiveness** of the Act. Thus, we take the average values of the pre-treatment and the post-treatment period.
The notation that is needed for the analysis are as follows:-

$$D_t = \begin{cases} 0 & \text{if observation is from control group} \\ 1 & \text{if observation is from treatment group.} \end{cases}$$

$$D_p = \begin{cases} 0 & \text{if observation is from pre-treatment period} \\ 1 & \text{if observation is from post-treatment group.} \end{cases}$$

Model of the Difference-in-Difference estimates :

$$Y = \beta_0 + \beta_1 * D_p + \beta_2 * D_t + \beta_3 * D_t * D_p$$

where:
Y = Outcome variable (here, tiger population of a particular species)
β_0 = Original regression intercept
β_1 = Average change in Y from 1st to 2nd time period that is common to both groups
β_2 = Average difference in Y between the 2 groups that is common in both periods
β_3 = Average differential change in Y from 1st to 2nd time period of treatment group relative to control group

Computation of the regression coefficients:-

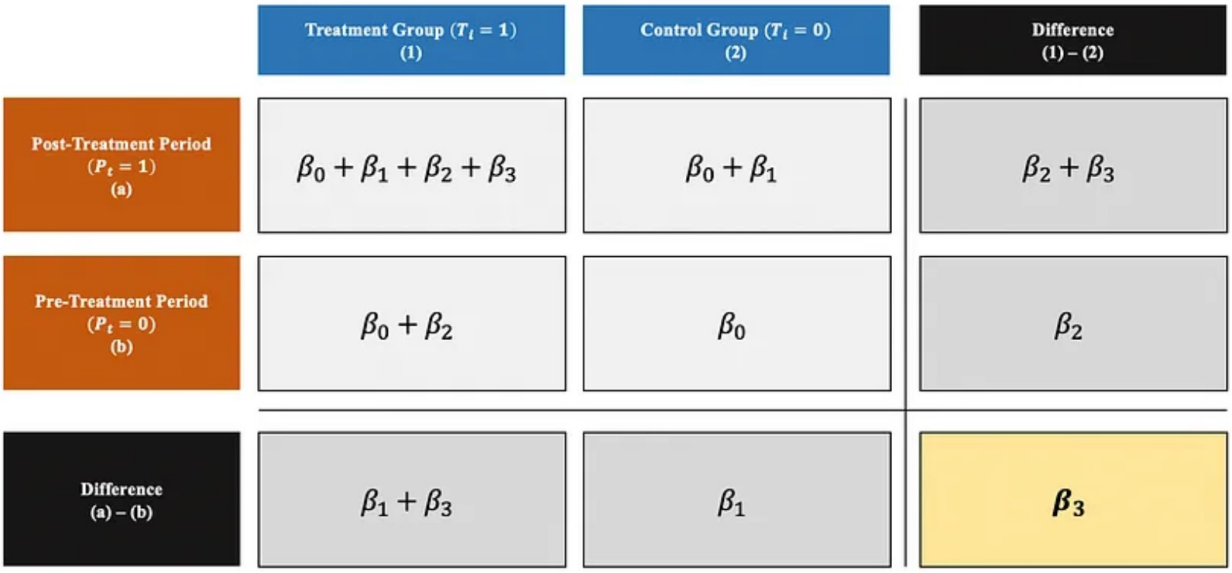


Figure 26: Estimated regression coefficient values

The data in the Figure 4 displayed previously can be expressed thus in terms of indicator variables D_t and D_p as:-

		OUTCOME (Y)	TREATMENT	
Species	Time	Tiger population(estimate)	W.A.A 2002	Post treatment indicator
1	0	2300	1	0
1	0	2280	1	0
1	0	2220	1	0
0	0	920	0	0
0	0	660	0	0
0	0	580	0	0
1	1	2380	1	1
1	1	2540	1	1
1	1	2980	1	1
0	1	515	0	1
0	1	440	0	1
0	1	400	0	1
where				
1=Bengal tiger	1=Time period after 2002		1=Treatment group	
0=Sumatran tiger	0=Time period before 2002		0=Control group	

Figure 27: Datapoints in terms of indicator variables D_t and D_p

The above table can be interpreted as follows, for instance, the first three cells denote the observations for bengal tigers before the years 2002; while, similarly, the last three cells denote

observations for sumatran tigers before the years 2002(the indicators have been attached for reference).

We shall now calculate the average values for both the tiger species in pre-treatment as well as post-treatment periods and tabulate them as :-

	T=1 (I) (Treated)	T=0 (II) (Control)	Diff. (I)-(II)
P=1 (a)	2633	452	2181
P=0 (b)	2267	720	1547
Diff. (a)-(b)	366	-268	**634**
Difference pre-treatment	1547		
Difference post-treatment	2181		
Difference in Differences	**634**		
beta0	720		
beta1	-268		
Beta2	1547		
beta3	634		

Figure 28: Average values (colour coded) for each case

Thus from the above calculated values and figure 5 as reference, the values of the regression coefficients have been calculated and have displayed accordingly. Above all, the difference-in difference can be calculated as:-

$$DiD = Differenceposttreatment - Differencepretreatment$$

which comes out to be 634 tigers. The final step is to estimate the counterfactual, that is, how much would have been the population of bengal tigers had the treatment(Wildlife Amendment Act 2002)not taken place.
Estimating counterfactual is fairly intuitive and can be calculated as :-

$$C = \beta_0 + \beta_1 + \beta_2$$

Thus,substituting the values of the respective coefficients, the counterfactual comes out to be 1999.

Interpretation of the output

The counterfactual value thus is interpreted as :- given the trend, had the treatment not happened ,there would have been on an average 1999 tigers i.e (T=1,P=1) would have been 1999 in contrast to the total count of 2633 tigers!

Moreover, it has been observed that, pre-treatment difference was 1547 tigers and the post-treatment difference was 2181 tiger. The difference-in-difference is 634 tigers, which indicates that the population of bengal tiger has significantly increased.

Conclusion

Based on the figures obtained after conducting the analysis, we can conclude that the Wildlife Protection Amendment Act (2002) was helpful in replenishing the bengal tiger population in India.

Remark

Although bengal tigers also reside in the neighbouring countries of India, it is important to note that these populations are scattered and nearly 80-90 % of all the bengal tigers reside in India. Thus, India as a country being home to the largest tiger population on the planet, the total count of the bengal tigers in India greatly determines and influences the total counts of the species. Simply put, India plays a major role in determining whether this majestic animal will dwindle or thrive in the near future.

Limitations of the Project

Limitations of the probabilistic model fitted for estimating number of species threatened in a country:-

1. The model is applicable only for ONE year (2022).Due to lack of such data for previous years, a very robust model which could have helped us in similar such modeling for the future years(using machine learning algorithms) could not be done.
2. It is always not a very good practice to directly remove the outliers or influential points. However, to get a more precise model, it had to be done here.
3. Due to no access to professional softwares like EasyFit and limited knowledge about probability distributions, the best possible fit was found out only for some known standard discrete distributions in R.

Limitations of the Markov chain analysis performed on the conservation status of a species belonging to certain taxon :-

1. The analysis has been performed based on the data provided by the IUCN and does not take into account biological factors influencing the changing conservation status of a particular species.

   For instance, after a species is declared extinct, it is of no debate that the last individual of the species has died. However, in our case we can still observe specific species changing their conservation status from EX to say CR. This happens because of the regular data collection and updating by the IUCN (thus making EX a non-absorbing state).
2. Due to limited knowledge of the concept of Markov Chains, an in depth analysis by making use of additional concepts like **Monte Carlo Markov Chains(MCMC)**, which are very extensively and regularly used by IUCN for analysing population of a particular species, could not be done.
3. Due to time constraints, a more concise and in depth version of the IUCN Red List data (for various particular species) could not be obtained.

Limitations of the random forest model fit for predicting conservation status of a species

1. We can notice that for the Extinct class, both the precision and recall are zero. We notice that it is because the support for EX class is only 3, i.e., there were only 3 species in the test set that were extinct consequently leading to these metrics.
2. Decision trees and Random forest model on a categorical but gives equal weight to all the classes. Our Response vector was ordinal as it is in the order of increasing threat to a species, but this ordinality was not taken into consideration.

3. Precision, Sensitivity and consequently F1-score do not consider ordinality as well. For example, classifying a Critically Endangered species as Endangered is less erroneous than say classifying it as Near Threatened.

Limitations of the Difference-in-Difference analysis performed on checking the effectiveness of WAA2002 :-

1. The figures of population of all the tiger species for years before 2000 are rough estimates of the populations owning to the fact that the censuses conducted earlier were not accurate.
2. The data sources are scattered and hence the reliability of the analysis depends upon the authenticity of these sources.
3. No data was explicitly available for the bengal tiger populations in the neighbouring countries of India, hence the fact stated in the remark was used for the analysis.

Scope of the topic and looking into the future

Statistical tools are being extensively used in predicting and estimating extinction risks that some species will face so as to aid in implementation of conservation programs and laws. An in depth statistical analysis of biological as well as man made factors responsible for exacerbating risks of extinctions faced by particular species, can help identify the root cause of the driving factors behind the same, consequently assisting in decision making for conservation of the species. Such analyses conducted world wide and spread among the common public can help spread general awareness about the extinction crisis we are going through. 0 In fact the severity of the crisis we are facing is not known to many. Our planet has been through Five Mass Extinction events and now,the ***Sixth Mass Extinction*** has began! However, unlike these previous extinction events, it is **very important** to note that the previous mass extinction events were caused by natural phenomena, while the sixth one has been initiated by us.

As the saying goes, better late than never. Inferences and conclusions drawn from statistical analysis can at least serve as an instrument towards conveying the world the severity of this crisis, the importance of ecological balance and coexistence and above all,the importance of the steps that need to be taken by us.

A note on packages *rredlist* and *iucn_sim*

There are packages in R-software which are **solely** dedicated towards analyzing the data presented by IUCN. The package *rredlist* has a number of commands like rl_history(), rl_country() and rl_habitats() just to name a few, which perform different tasks. Detailed information about each command can be found on the R software website or by using help() command in R software. Another package called as *iucn_sim* package is primarily used for simulating the data gathered from the *rredlist* package. The main aim of this package is to give improved predictions for extinction rates of different species. Details of this package can be found on Github.

However, if we are to get access to the data provided by IUCN through *rredlist*, we need to request API keys from the organisations for the same, which is not a very straight forward procedure and the IUCN later also demands a report about how the data has been used by us. Thus, due to many such restrictions, we were unable to fully gather in depth and concise data so as to strengthen our analysis and effectively demonstrate the scope of the topic and how big of an oragnisation the IUCN is.

Software used

R Software

R is a programming language and software environment for statistical computing and graphics. It is widely used in the scientific community for data analysis, visualization, and modeling. In our project, we have used R pacakages such as nlstools, fitdistrplus, markovchain, MASS, diagram.

Excel

Excel is a spreadsheet software developed by Microsoft. It is commonly used for tasks such as data entry, organization, analysis and for visualizing data. Excel is widely used in business and academic settings.

Python

Python is a high-level programming language that is popular in a variety of fields, including data science, machine learning and web development. It has a large library of packages for data analysis, some of which we have used are numpy (for numerical analysis), sklearn (for machine learning) and seaborn for visualization.

SPSS

SPSS (Statistical Package for the Social Sciences) is a software package for statistical analysis. SPSS has a user-friendly interface and a wide range of statistical procedures and techniques.

LaTeX

LaTeX is a document preparation system used for creating scientific and technical documents. It uses a markup language to define the structure and formatting of documents, and is known for its high-quality typesetting and support for mathematical equations and symbols.

Appendix

R Code for fitting Regression model:-

```
install.packages('nlstools')
library(nlstools)

years=c(0,2,3,4,6:22)
total_as=c(16507,16697,22424,38046,40174,41415,44838,47677,55926,61914,
    65518,71576,76199,79837,85604,91523,96951,112432,128918,142577,
    150388)
total_thr=c(11046,11167,12259,15503,16116,16306,16928,17291,18351,19570,
    20219,21286,22413,23250,24307,25821,26840,30178,35765,40084,
    42108)
plot(years,total_as,type='l ',col='green',xlab=c('years'),
   ylab=c('number of species')); points(years,total_thr,type='l',col='red')
legend('topleft',legend=c('Total Assessed','Total Threatened'),
   col=c('green','red'),lty=c(1,1))
plot(years,total_as); points(years,total_thr,pch='*',col='blue')
a=lm(total_as~years); b=lm(total_thr~years)
abline(a,col='green'); abline(b,col='red')
legend('topleft',legend=c('Linear fit of Total assessed',
      'linear fit of Total Threatened'),col=c('green','red'),lty=c(1,1))
par(mfrow=c(2,2)); plot(a); plot(b)

model=nls(total_as~a+b*years+c*years^2+d*years^3,data=data,
     start=list(a=100,b=100,c=10,d=10))
total_as_fit=fitted(model)
plot(years,total_as)
lines(years,total_as_fit,col='red')
plot(nlsResiduals(model)); coefficients(model)

model=nls(total_thr~a+b*years+c*years^2+d*years^3,data=data,
     start=list(a=100,b=100,c=10,d=10))
total_as_fit=fitted(model)
plot(years,total_thr)
lines(years,total_as_fit,col='red')
plot(nlsResiduals(model))
summary(model); coefficients(model)
```

R Code for fitting Probability model:-

```
library(MASS); library(fitdistrplus)
data=read.csv("C:\\Users\\Dell\\Downloads\\
            Table 5  Threatened species country new.csv",header=TRUE)
head(data)
total_sp=as.numeric(gsub(",","",data$Total))
plotdist(total_sp,histo=T,demp=T,breaks=32)
descdist(total_sp,discrete=T,boot=1000)
par(mfrow=c(2,2))

plot.legend=c('poisson','geometric','negative binomial')
fpoi=fitdist(total_sp,'pois')
```

```
fgeom=fitdist(total_sp,'geom')
fnbin=fitdist(total_sp,'nbinom')
denscomp(list(fpoi,fgeom,fnbin),legendtext=plot.legend)
qqcomp(list(fpoi,fgeom,fnbin),legendtext=plot.legend)
cdfcomp(list(fpoi,fgeom,fnbin),legendtext=plot.legend)
ppcomp(list(fpoi,fgeom,fnbin),legendtext=plot.legend)

boxplot(total_sp,main='Box plot for entire data')
total_sp_rm=total_sp[!total_sp %in% boxplot.stats(total_sp)$out]
boxplot(total_sp_rm,main='Box plot after removing outliers')
length(total_sp)-length(total_sp_rm)

hist(total_sp_rm,breaks=32,prob=T)
descdist(total_sp_rm,discrete=T,boot=1000)

par(mfrow=c(1,2))
plot(1:700,dnbinom(1:700,size=1.559133,prob=0.009863068636),type = 'l')

fnbin
par(mfrow=c(1,1))
hist(total_sp_rm,breaks = 32,prob=T,main='Total number of species
    threatened countrywise',xlab='Number of species threatened')
points(dnbinom(1:700,size=fnbin$estimate[1],mu=fnbin$estimate[2]),
    col='blue',type='l',lwd=2)
```

R Code for Markov Chain analysis:-

```
install.packages('markovchain');library(markovchain)
install.packages('diagram');library(diagram)
install.packages('expm');library(expm)

DD=c(0,0.611353712,0.074235808,0.082969432,0.170305677,0.061135371,0,0)
LC=c(0.21978022,0,0.450549451,0.197802198,0.10989011,0.021978022,0,0)
NT=c(0.027272727,0.581818182,0,0.190909091,0.181818182,0.018181818,0,0)
VU=c(0.082191781,0.246575342,0.150684932,0,0.342465753,0.157534247,
    0.006849315,0.01369863)
EN=c(0.107526882,0.139784946,0.096774194,0.225806452,0,0.419354839,0,
    0.010752688)
CR=c(0.171428571,0.028571429,0.057142857,0.114285714,0.6,0,0,0.028571429)
EW=c(0,0,0,0.333333333,0,0.333333333,0,0.333333333)
EX=c(0,0,0,0,0,0.75,0.25,0)

reptiles=c(DD,LC,NT,VU,EN,CR,EW,EX)
reptiles_matrix=matrix(reptiles,nrow=8,byrow=T);reptiles_matrix
m1=reptiles_matrix
rownames(m1)=c('DD','LC','NT','VU','EN','CR','EW','EX')
colnames(m1)=c('DD','LC','NT','VU','EN','CR','EW','EX')
m1_trans_r=new("markovchain",transitionMatrix=m1,states=c('DD','LC','NT',
        'VU','EN','CR','EW','EX'),name='MC1')
m1_trans_r;plot(m1_trans_r)
ss=steadyStates(m1_trans_r)
barplot(ss,main='Taxonomy - Reptiles',ylab='Prob. of
        attaining a particular consn. status in future',col='red')
summary(m1_trans_r)
simm=rmarkovchain(n=100,m1_trans_r,t0='NT');simm
```

Python code for predicting the conservation status using Random Forest:-

```
import numpy as np
import seaborn as sns
import pandas as pd
#preprocessing
dataset = pd.read_csv('factored_table.csv')
ds2 = dataset.replace(['LR/lc','LR/cd','LR/nt','CR(PE)',
    'CR(PEW)','EW','NR','DD'],['LC','LC','NT','CR','CR','EX','LC','LC'])
ds2 = ds2.drop(columns= ['Scientific name','Common name'])
# using ordinal encoder on response vector
from sklearn.preprocessing import OrdinalEncoder
enc = OrdinalEncoder(categories= [['LC','NT','VU','EN','CR','EX']])
print(ds2['IUCN Red List (2021)'].value_counts())
print(ds2['IUCN Red List (2022)'].value_counts())
f1 = enc.fit_transform(ds2.iloc[:,[1]])
f1 = f1.reshape(2024,)
# One Hot encoding for CART
ds4 = pd.get_dummies(ds3[0])
ds5 = pd.get_dummies(ds3[1])
ds3[ds4.columns]=ds4
ds4[ds5.columns]=ds5
ds4[res.columns] = res
ds3.drop(0,axis=1,inplace=True)


y = ds4['IUCN Red List (2022)']
ds4.drop('IUCN Red List (2022)',axis=1,inplace=True)
X = ds4
# splitting the data into training and test se
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size = 0.25, random_state = 0)
#fitting the model
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 100,
    criterion = 'gini', random_state = 0)
classifier.fit(X_train, y_train)
#predicting test set results
y_pred = classifier.predict(X_test)
pred = classifier.predict(X_train)
cm = confusion_matrix(y_test, pred)
```

References

https://www.traffic.org/publications/reports/skin-and-bones-report-2022
https://ntca.gov.in
fsi.nic.in
https://www.iucnredlist.org/resources/summary-statistics
https://www.wpsi-india.org/tiger/index.php
www.iucnredlist.org
https://github.com/tandermann/iucnsim
www.gbif.org
www.nationalredlist.org
https://www.researchgate.net/figure/Status-transitions-counted-in-the-IUCN-history-of-birds-class-Aves-between-2011-and_tbl1_346869022
https://www.statisticshowto.com/difference-in-differences
https://youtu.be/CT4jnOZGv_A
https://youtu.be/eiffOVbYvNc
https://youtu.be/J7q2H8aB8bQ
https://scikit-learn.org/stable/modules/tree.html
https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
http://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote1.html

Books Referred:

Thesaurus of univariate discrete probability distributions
by G. Wimmer and G. Altmann
Man–Eaters of Kumaon by Jim Corbett
The Royal Tiger of Bengal: His Life and Death by Sir Joseph Fayrer