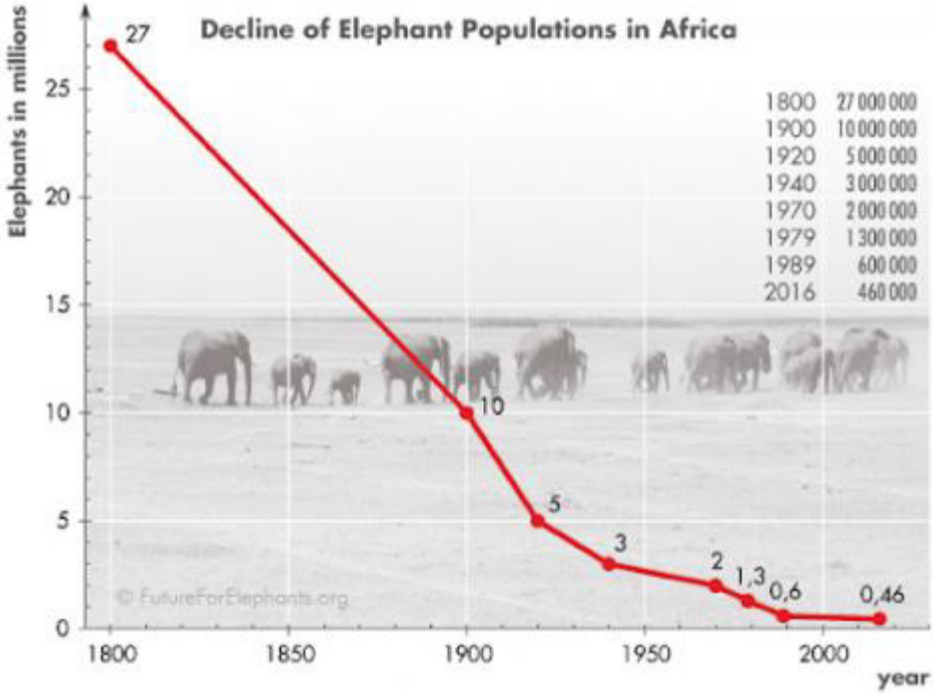
Jacob Harkins

CM3203 - Predicting animal movements using collar data

Project Description

In the current environmental climate, more species of animals are getting driven to extinction than ever before, whether it be deforestation, climate change, or the main driver for this project – poaching. Elephant numbers across the globe have dropped by 62% within the past decade, from poachers seeking profit from ivory, meat, and other body parts – projections show that they be mostly extinct within the next decade. With wildlife crime bringing in between $7 billion to $23 billion each year, it is no surprise that wildlife reserves are struggling to maintain a defence against poachers killing endangered animals for 'medicinal purposes' and trivial vanity items.



For one, patrols are often difficult to organise – whilst a group of animals could be spotted in one location on one day, that same pack could be anywhere within the compound sometime later, making it hard for rangers to make sure that poachers do not get close to the protected animals. Furthermore, if rangers get information that poachers have been spotted within a certain part of the compound, it is hard to narrow down where the poachers may have travelled from the last sighting – the animals could have settled down anywhere, and the rangers must race against time to check all possible locations.

For these reasons, with the datasets provided by researchers at the "Danau Girang Field Centre" in Malaysia (which comprises of one dataset for pythons, the rest being elephants) the goal for this project is explore the data with the intention of helping law enforcement to better understand how animals move so animals can be located easier.

Before being able to develop a model however, the different datasets will have to be processed and cleaned if needed, as with real world data there is often large discrepancies with some pieces of data, or other inconsistencies in how the data was collected.

After that has been done, multiple techniques will have to be developed, tested, and compared with each other to see which is the most effective at predicting animal movements. Once the most effective technique has been chosen, it will be developed within Jupyter Notebook with the dataset to allow a clear visualisation to be developed.

Once a sufficient model has been developed, then the aim is for the model to be able to take an input date from the user and predict where animals will be within the reserve at that given date.

Project Aims and Objectives

Aims

- To be able to explore the data with the intention of helping law enforcement to better understand how animals move so animals can be located easier.
- To be able to make it easier for rangers to organise patrols.
- To make it harder for poachers to get closer to animals.

Objectives

- Process and clean the data to a sufficient level – spend time understanding what each variable represents, get an idea of which aspects of the data may be more relevant than others by drawing simple visualisations (heat maps, correlation graphs etc.) for each of the datasets with different variables through to get a better feel for it before developing techniques. Make sure I fully understand how the different variables were measured (e.g., distance measured from a certain point in the compound).
  Look for any outliers that may have to be excluded/corrected in the dataset, as well as inconsistencies in the collection methods. This could be done for each dataset in

many ways e.g., by displaying all the points on a map and seeing where there are large jumps in distances.
Work out how data should be manipulated to better fit the objectives/aim.

- To develop different prediction techniques – research which techniques would be most applicable to this scenario by looking through various respectable sources, and then implement each of them into Jupyter Notebook.
- To test the different prediction techniques – test how accurate the techniques were at predicting the location of the animals at certain times of the year, based on the input.
- To compare the different prediction techniques – compare each of the techniques with one another and perform various tests to see which one would be most appropriate for the aims I am trying to achieve. The tests can involve which technique had the highest percentage accuracy in predicting the location of the animals, run-time in how long it takes
- Develop a model with the chosen technique – implement the chosen technique into Jupyter Notebook and integrate with the datasets, allowing the user to input a chosen date, returning a position/area on the map of the reserve for the animals.

(Gantt chart is attached as excel spreadsheet)

Work Plan

Familiarise (with), process, clean data

I intend to get information from those that collected the data on what the variables are in their own words, as well as how they are measured to ensure I am not making incorrect assumptions. Furthermore, different visualisations of the data will be made to help understand the trends within the data and what correlates with one another e.g., heat maps, correlation graphs etc.

I will visualise all the different data points for each dataset on the research reserve area, with number labels to see if there are any outliers with regards to large jumps in distance that do not make sense in the context e.g., the distance an elephant has travelled between two data points is much higher than average.

The data will have to be processed depending on what will be most applicable for my task – it may be too difficult to use the raw data that has multiple points each day to attempt to predict animal movements, the data may be much easier to work with if instead, average data points were made for each week/fortnight/month.

Develop prediction techniques

The prediction techniques that I will develop into Python will have to be researched to see if they are suited to the data set that I have been given. Seeing how the goal of this project is to help law enforcement to better understand how animals move, then it perhaps a regression model that would be most suitable to predict two values: longitude and latitude. Perhaps a better way to do so initially would be to have two separate models: one that

predicts a latitude value and one that predicts the longitude value, to lower the risk of errors at the beginning or getting myself too confused.

Each of the prediction techniques that I have chosen will have to be implemented into Jupyter Notebook with Python so that I can start using it with the datasets. I will have to familiarise myself with the workings of each of the techniques and then think about how to adapt it into Python. After doing so, it is likely best to test the implementation with data taken online that already have known expected values that I can compare with to make sure implementation went smoothly.

Test prediction techniques

To test the various techniques, the datasets will have to be split into training set, validation set, and test set. After training the machine repeatedly on the training set with one of the regression models, it would be tested against the validation set whilst it is still in development to make sure that the model has not only got good at predicting with familiar data. Following this, the machine will be using the test data set (which only contains data that the machine has not got any experience in handling) to predict the latitude and longitude values, using its previous training to do so. If the model fits to the test data as well as the training data, then minimal overfitting has taken place – making it a desirable choice.

There are multiple ways to compare which regression model is more suited for this data set, each of which may be more applicable depending on the type of regression model used including:

- Average error: the numerical difference between a predicted value and the actual value present in the data.
- R-squared: the proportion of variance for a dependent variable (is only applicable for linear regression, so how useful it is depending on the regression model chosen).
- Mean square error: the average squared distance between an estimated point and the actual value

And many others.

It could also be a good idea to cross-validate to reduce the chances further of overfitting.

Compare prediction techniques

The techniques will have to be compared with each other, their pros and cons weighed up against one another, the primary concern is accuracy so that will likely be the deciding factor. Depending on which of the past comparison methods is most applicable, the value collected for each different regression model will be compared with one another, the one which has the highest ‘accuracy’ will potentially be picked.

It could also be a good idea to test multiple methods instead of just one of the methods shown above, to make sure a technique is not chosen that may appear to be suitable based on one of the values above, but may not be when put into practise.

Jacob Harkins

Develop model

Fully develop the final chosen model within python, allowing the model to take an input date, and if the previous steps have been completed as expected, it should return a prediction. A raw latitude and longitude value could be returned, or perhaps the point could be shown over an image of the preservation reserve. Whilst this should be the primary aim of the task, if there is time then perhaps features that allow the user to select different packs of elephants or select the animals by name could potentially be implemented.

Risk number #	Risk cause	Risk effect	Likelihood (almost certain, likely, possible, unlikely or rare)	Consequences (negligible, minor, moderate, major, or crucial)	Risk ranking (low, moderate, high, or very high)	Mitigation strategies
Risk 1	Prediction techniques are incorrectly implemented.	Incorrect understanding on how the techniques do not work; Incorrect implementation into Python.	Possible	Crucial	**High**	1. Spent a lot of time getting a deeper understanding of the techniques before implementing them, read up through various respected sources. 2. Test implemented techniques with online data sets that already have confirmed true values - compare the
Risk 2	Outliers in the data are left unchecked	Not enough testing on the data before it is implemented into the different models and machine learning.	Possible	Moderate	**High**	1. Spend time visualising the data in different ways to try and spot these visualisations (can plot points chronologically on the reserve map and look for any large jumps between two points).
Risk 3	Model is unable to accurately predict animal movements.	Unforeseen issues make the task more difficult, incorrect implementation of model.	Possible	Crucial	**High**	1. Take a step back when a brick wall is hit - talk to supervisors and spend more time researching the concepts behind what I am trying to achieve. 2. Do vigorous testing before developing a final model.
Risk 4	Incorrect prediction technique is chosen	Only testing the model with one calculated variable that it may be skewed towards.	Possible	Moderate	**High**	1. Test each model/prediction technique for multiple variables.

References

Butler, R., 2022. *62% of all Africa's forest elephants killed in 10 years (warning:* graphic images). [online] Mongabay Environmental News. Available at: <https://news.mongabay.com/2013/03/62-of-all-africas-forest-elephants-killed-in-10-years-warning-graphic-images/> [Accessed 7 February 2022].

En.wikipedia.org. 2022. Mean squared error - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Mean_squared_error> [Accessed 7 February 2022].

En.wikipedia.org. 2022. Training, validation, and test sets - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets> [Accessed 7 February 2022].

Globalpartnership.org. 2019. Risks and mitigation strategies template. [online] Available at: <https://www.globalpartnership.org/content/risks-and-mitigation-strategies-template> [Accessed 7 February 2022].

Google.com. 2022. Before you continue to Google Maps. [online] Available at: <https://www.google.com/maps/place/Danau+Girang+Field+Centre/@5.4315253,118.010532,13z/data=!4m12!1m6!3m5!1s0x323f23c100a4214b:0x47f4b4f6fc957d2f!2sDanau+Girang+Field+Centre!8m2!3d5.4137278!4d118.0376297!3m4!1s0x323f23c100a4214b:0x47f4b4f6fc957d2f!8m2!3d5.4137278!4d118.0376297> [Accessed 7 February 2022].

Hennrich, T., 2019. The Fall of the Gentle Giants. [online] The Sting. Available at: <https://chssting.com/3311/news/the-fall-of-the-gentle-giants/> [Accessed 7 February 2022].

InDataLabs. 2021. What is Predictive Performance Models and Why Their Performance Evaluation is Important. [online] Available at: <https://indatalabs.com/blog/predictive-models-performance-evaluation-important> [Accessed 7 February 2022].

Lehmacher, W., 2016. Wildlife crime: a $23 billion trade t*hat's destroying our* planet. [online] World Economic Forum. Available at: <https://www.weforum.org/agenda/2016/09/fighting-illegal-wildlife-and-forest-trade/> [Accessed 7 February 2022].