III. Model documentation and write-up

1. **Who are you and what do you do professionally?**

My name is Benjamin (Ben) Carrión. I hold a bachelor degree in hydraulic engineering and a Master of Science on coastal engineering.

I work mostly as numerical modeller of coastal hydrodynamics and morphodynamics for a port and coastal consulting company (PRDW).

As a result, I code and deal with data regularly. I'm quite new to the data science world, formally, but I wanted to see how far could I make it (in the competition) with my current background.

2. **High level summary of your approach: what did you do and why?**

The key element to realize is that, at least in this particular case of seasonal breeding in a rather extreme ecosystem, what you really want to estimate is the annual rate of change of populations rather than the actual populations.

My thought was that populations are in a delicate equilibrium between competitors, preys, and predators, which can be written down using differential equations, on which the absolute rate of change between breeding seasons (birth rate minus death rate) depends in some complex way on the other species' populations.

This differential system can be solved and the levels of different populations will change until an equilibrium is reached. However, given the seasonal character of Antarctica, the "time step for the integration", sort of speak, of these equations is fixed, and could be too large to reach smooth equilibrium. Instead, the population levels oscillate seemingly at random, but around a certain equilibrium value.

This might seem far-fetched, but it is (somewhat) what happens with lemming populations in Canada. The idea of "suicidal lemmings" arises from the large variations in their population from season to season. In the particular case of lemmings what happens is that they food (plants) grows at a slower rate than the lemmings are born. So, high natality increases the lemming population, which depletes their food supply, which doesn't grow as fast, so many lemmings die. This is a very rough sketch, but illustrates the idea. Specifically, it shows that the overall rate of change seems to depend largely on the previous' year population.

Now, for the case of the penguins it is hard to see any trend at all per nest, given the patchiness of the dataset. Some sites might have only a few observations, or even only one. Moreover, fewer observations were available for previous years too deep in the past. The dataset is far from homogeneous, and any method to fill it up might include large errors. Also, it wasn't clear that certain penguin types favoured larger or smaller groups: you could find some nest of a certain type with a few dozen individuals, or other nest of the same penguin type with a few thousand ones, even for close locations. In that regard, location didn't seem very relevant either.

However, if the **mean count per penguin type** are considered, trends are more visible: large variations around a rather stable average value. This is my main hypothesis.

Now, the actual rates are still quite noisy, but correlate well with the average mean population per site per penguin type of the previous year. Correlation with other environmental variables, such as ice extent or water/air temperature was not clearly seen, since these parameters also

present a rather noisy behaviour (over a small mean trend: rising both temperatures and ice extent).

The train of thought was to forecast a mean rate per penguin type, and apply it at each site, using the last valid measurement. In fact, if you consider a rate of change 1.0, i.e. no change from the last valid value, you already get a good estimate for the future year populations. This indicates that (mean) rates seem to be rather close to 1.0, and what you should aim for is a method to estimate the deviation from that value, for each forecasted year.

3. **What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?**

First I tried a more probabilistic approach, trying to simulate a new time series, using the past data to estimate the kernel for the rates of change of populations. This approach failed miserably.

Later on, I tried to correlate the rates of change to several environmental values, both local for each nest and averaged over the whole Antarctica or even the southern hemisphere. However, the forecasted rates varied way too much when compared with the observed rates. The best predictor always remained the last year population (per penguin type).

4. **Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?**

Yes. I used jupyter notebooks to explore and plot the data provided, and other environmental variables such as time series ice extent and sea/air temperature.

5. **How did you evaluate performance of the model other than the provided metric, if at all?**

I used the AMAPE metric, over a test set composed of the last available years (2010-2013).

6. **Were there other fields or features you felt would have been very helpful to have? That is, what might the organization want to collect in the future that was not in the given data?**

I feel they provided lots of additional information of where to look for data, and ecosystem descriptions. Probably some other useful ones would include estimation of fishes and krill, since I thing that actual measurements would be hard to come by.

7. **Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?**

Unfortunately, I did not produce a model, but rather I just made a good guess of the mean population change rate for the 3 years in the test set. My scrip will only produce my submission.

You can, nonetheless, change the estimated rates of change, and feed them into the model. This could be done with some other model or fit.

8. Do you have any useful charts, graphs, or visualizations from the process?

Yes, I do. You can find them in the notebook "explore_data". Here a brief description of the plots, and the insight I got from them.

Figure 1 shows the location of nest, identified by penguin type. It is clear that Adelie penguins live all around the Antarctic continent, while both Chinstrap and Gentoo live in the Antarctic Peninsula. From this spatial distribution, and considering the difference in diets, it is expectable that Adelie populations will show a somewhat different behaviour than the other two types. Chinstrap and Gentoo, living in the same area are expected to behave more similarly.

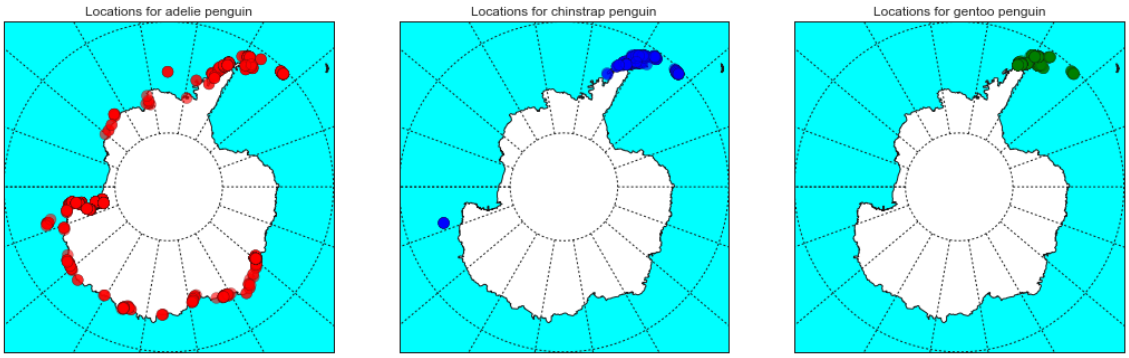

Figure 1: location distribution per penguin type.

Figure 2 shows the time series of mean nest count per penguin type. A couple of features are remarkable:

- All populations seem to oscillate around a rather fixed average.
- The averages for Chinstrap and Gentoo are similar among them, and much lower than the average for Adelie.
- Broadly, population changes are mild for Gentoo, large for Adelie, and show spikes for Chinstrap.

Figure 3 shows the relation between the rate of change of a penguin type population and the population size of the previous year. The relation seems to be restorative: negative rates for large populations (larger than the historic mean), and positive change for small population (smaller than the historic mean). The correlation ranges from $R^2 = 0.37$ for Gentoo up to 0.57 for Chinstrap penguins. Thus, this seems to be the main predictor for the total range.

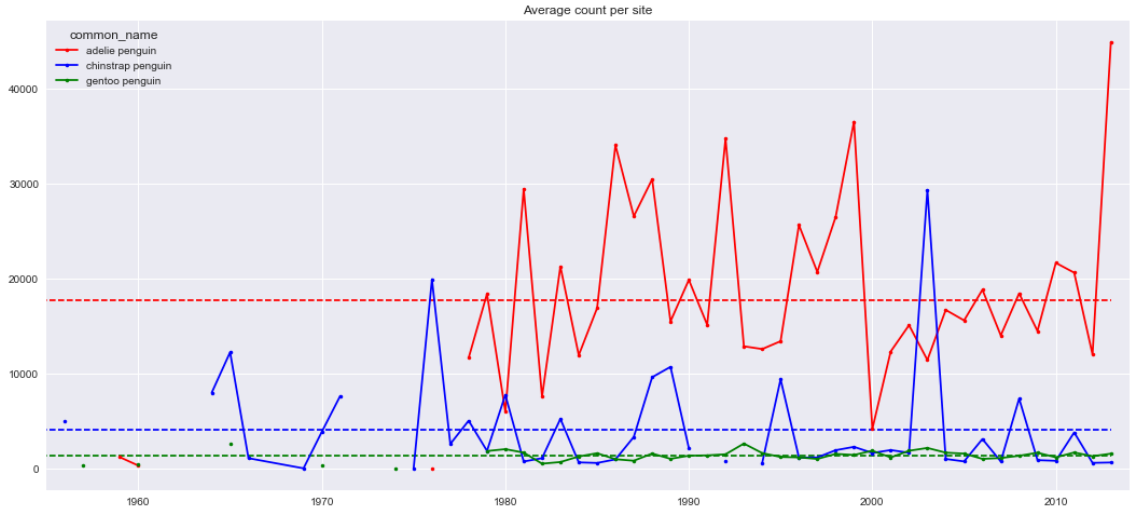

Figure 2: time series of mean nest count (solid lines), and the overall average (dashed lines) per penguin type.

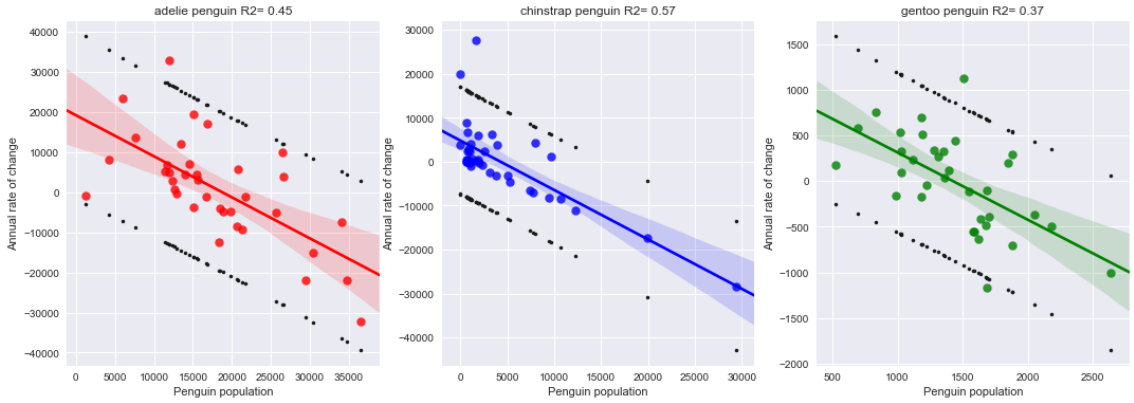

Figure 3: correlation between rate of change and population size last year penguin type.

Finally, Figure 4 shows the distribution of counts for penguin types, during all the recorded years. All of them appear to be rather exponential, which might indicate some stochastic variations might explain the different observed values.

It might be that, on average, the mean population changes with a certain rate of change, but in each particular place there will be some deviation from that trend. Moreover, some places might have some geographical constraints that limit the number of nests that can be accommodated there. This constrains appear to be random, or stochastic, variables given the exponential distribution of nests counts.

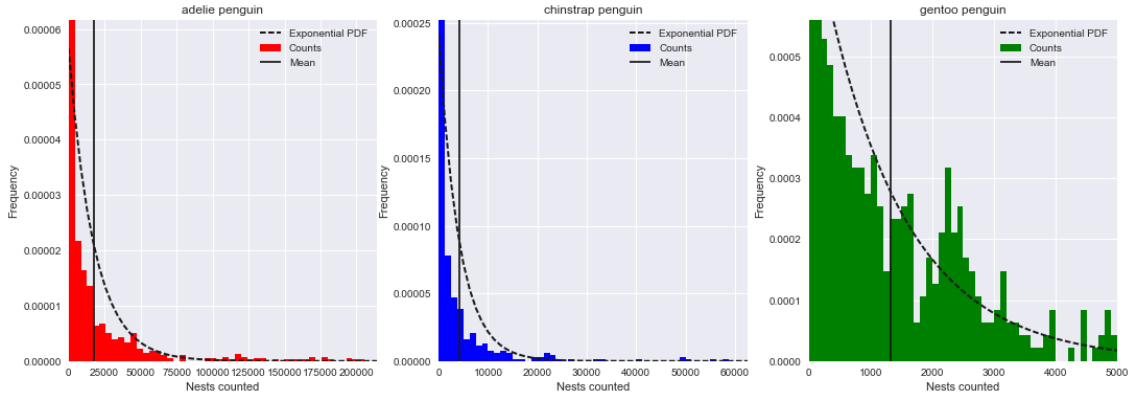

Figure 4: distribution of nest count per penguin type.

9. **If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far?**

I'd try the following:

a. The first thing is to try to actually predict the mean rate of change per penguin type, as a function of the previous year's population and some other environmental variables.
b. Later on, small variations from the penguin type rate of change could be computed for each location, depending on historic data, and local variations of environmental variables compared to global ones (e.g. local temperature vs Antarctica-averaged temperature)

10. **Are you willing to be interviewed for a blog post?**

Yes. Although I would feel a bit as an impostor since I didn't really produce a model.

I do believe, however, that the insight of focusing on the rate of change per penguin type would be of value for some actual data scientists. It reduced the problem to just a few parameters that showed to be key for the forecast. The results are nonetheless quite sensitive to these parameters, and here some more hard-core data science might be in order (which is something I can't really produce).