III. Model documentation and write-up

You can respond to these questions either in an e-mail or as an attached file (any common document format is acceptable such as plain text, PDF, DOCX, etc.) **Please number your responses.**

1. **Who are you (mini-bio) and what do you do professionally?**

   I am Aharon Ravia, 32 years old from Israel. Currently, a PhD candidate in the neurobiology department at the Weizmann Institute of Science, studying computational models for human olfaction (the sense of smell).
   Throughout my academic and professional career, I have studied math & neuroscience and was employed as a data scientist, and as a data science team manager.
   I am also a huge fan of penguins, watched them in Australia and New Zealand, but had not yet been to Antarctica.)

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

   My first entry to the competition was late, less than 2 weeks before the deadline. As I have seen the amount of the data, which was quite sparse, I thought that there is not much time to go beyond simple methods. My experience with time series data had thought me that the best estimation for next step change is 0. So I predicted the last known result for each location as the estimation for the years not revealed. Astonishingly this approach alone did quite well and resulted is 16-20 place in the leaderboard.
   This thing has thought me few things, the first, that it was also significantly better than the linear fit solution, and many other submissions, so significant that I ruled out (almost completely) the possibility that it won't be the case on the private leaderboard set.
   In addition, the AMAPE estimation I had on my training set was so far away from the result on the leaderboard set, that I realized I have to relate carefully to my validation result.
   I wanted to further combine in the model 3 different type of models. The first is estimation of the mean, which is done by last estimation (could be done as an average of last N observations). The second is a recent trend, changes in recent years. The third is a the whole time series trend. This approach was partially adapted from facebook's prophet https://research.fb.com/prophet-forecasting-at-scale/ blog post.
   I also wanted to study more from the public leaderboard set (with a risk of overfit) whether there was an additional general trend of increase / decrease that years.
   Eventually I combined the constant estimation, with a minor general trend and a linear fit that was computed differently for each year.

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

   I thought about taking geographical data into account, but couldn't use it efficiently.

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

   Yes, I used Matlab for exploration simple manipulation and most submissions. I also used it to explore geographical data, but without any significant findings.

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

   Did not use other methods. For sanity checks, I checked the submissions against each other, to see that the change is reasonable.

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?**

   Not a lot to tell here. As most counts (75%) were ground counts, maybe collecting the team identity could play a role here.

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.)?**

   No, the Jupyter notebook should work.

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

   I have submitted a notebook (matlab live script in pdf) but later it was evident that the main geographic point was absent there. I had a nice geographic visualization of the different populations and the trend in population change (it was a bit difficult to export for now).

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?**

   The next step will be to try and model each type of penguins population differently. There also seems to be a need to differently use the different assessments according to their type / grade, and use differently reliable counts versus less reliable ones. Weather can also play part in the population change, and has interesting implications.

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

    Sure, no problems.