Understanding the Health of a Bird Population: A Visual Analytical Approach

Alex Wilkes. 1833720*

Data Science MSc. Kings College London

1 Introduction

We have evidence that Kasios has been undertaking industrial activity that has been harmful to the Blue-Pipit population in the Boonsong Lekagul Wildlife Preserve. Kasios deny the accusations, and have provided recordings of what they claim are Blue-Pipits that they say shows the population is in good health. To support us in investigating this counter evidence, we have a large selection of recordings of different birds throughout the park, and metadata for the recordings which includes the species being recorded, the time and date of the recording, and the location of it.

2 Research Questions

We will assess the evidence provided by Kasios in the context of the trusted data provided by previous researchers. Specific research questions I will address are:

Part A: How can we characterise the patterns of all of the bird species over the time of the collection? Does the collection indicate any trends in the populations? For example, have the size of the Blue Pipit populations increased or decreased over time? Have the populations moved within the park? Can we infer what may have caused any changes we observe? Does the data supplied by Kasios fit within the trends identified? If not, it may support a hypothesis that the data has been fabricated. Finally, can we determine from the sound files provided whether Kasios have actually recorded Blue-Pipits?

Part B: Does the set of recordings provided support the claim of Pipits being found across the Preserve? Can we see evidence in visualisations of the sound files themselves?

3 Literature review

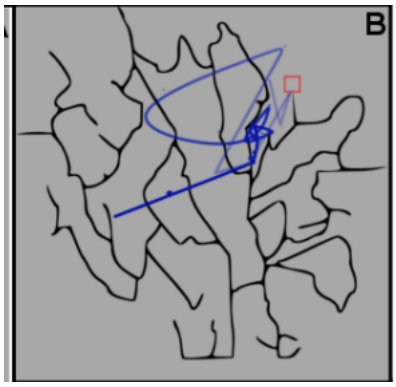
Cakmak et al. [3] take a map based approach, plotting the locations of the recordings of individual species to conclude that Rose-Crested Blue Pipits migrated southwestwards in the Wildlife Preserve away from the contaminated area.

Baeumle et al. [2] use graded opacity to show movement paths for the species over time. They calculate 'centroids' and 'centroid paths' which are the geometric centres of observations. 'Centroids are computed as a discounted sum from previous and the current year to smooth the appearance and outliers.' They are able to calculate 'a path through all the centroids of one species, where time is mapped onto the opacity' which 'makes it possible to analyze movement patterns over time.' The authors also use heatmaps to visualise where populations have seen large differences, and demonstrate an approach, 'small multiples,' with the map divided in to 16 sections which are shaded according to the presence of the species.

*e-mail: alexander.wilkes@cantab.net

Other literature relevant to mapping species includes Sullivan et al. [6] who visualise frequencies of recordings as a heatmap. They augment interactive heatmaps with the locations of oil spills, so the effect on bird populations can be explored. Their approach shows the value of the map based approach, which brings together the phenomena we are exploring (migration of a species) and the geography of potential causes e.g. oil spills or chemical dumps then inviting us to infer causality. Andrienko and Andrienko [1] argue that using visualisation in this exploratory process elevates its value beyond its traditional role to simply 'represent the final results of computations.' They describe how complex decisions can be driven by visual exploration.

Regarding Part B, Grosche, Mueller and Serra [4] demonstrate how a song can be translated in to a spectrogram, and then a constellation diagram to create a 'fingerprint' of a recording that is robust to noise and distortion. They show it is possible to visually identify when two recordings are of the same music. Bauemle et al. (2018) found that visualizing the spectrograms in combination with the results of the classifier allows us to verify the performance of the classifier in a detailed manner. their work is inspired by Hellmuth et al. [5] who set out a system for audio matching and identification in a way that is robust to typical signal disruptions (e.g. microphone distortion, compression).



4 Part A - Design

I will recreate and extend the approach of Baeumle et al. [2]. In order to create the movement paths, I use the metadata provided by researchers. This data is spatio-temporal; it contains the coordinates of observations, and the date/time they were recorded. We are told to 'assume we have a reasonable distribution of sensors and human collectors providing the recordings, so that the patterns are reasonably representative of the bird locations across the area.' I also take as given that the patterns are representative across time.

Using this data we explore research question 1. We can see trends in the total population over time, using the frequency and location of recordings as a proxy for the birds themselves. A histogram showing the frequency of recordings over time will convey

overall population numbers within the park. The location of birds will give us snapshot locations of the birds, but these will need both aggregating and smoothing.

Aggregation is required because recordings are not spaced at regular temporal intervals; sometimes a number of recordings occur in close succession, at other times there are large gaps between them. Instead I bin the data in to regular periods and take aggregate statistics from those bins. Through trial and error, it appears blocks of three days are optimal for the task at hand. Aggregating is achieved using existing methods available in the pandas library.

The point that we plot is the centroid of the observations within that period. This is simply the mean of the relevant observations in each attribute. This is demonstrated in figure 1, with triangles representing individual observations and the circle being calculated 'centroid' of the species. This entirely presumes there is a single population of any given species within the park, which is a strong assumption. I proceed with this assumption because it is made by Baeumle et al. [2]. However, further research in this area may look to use a clustering technique to try and detect multiple distinct populations of the same species. for this spatial data.

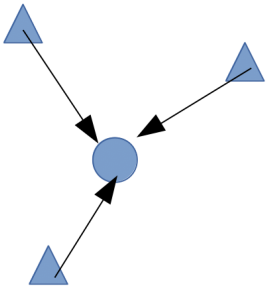


Figure 1: Identifying the centroid of a set of observations

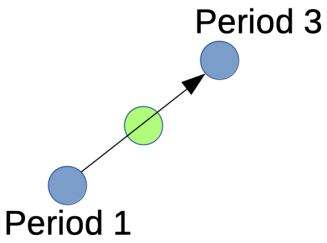


Figure 2: Example of interpolation

The plot includes information about how old an observation is. The plot uses opacity as a visual channel to convey this magnitude, despite being one of the less effective visual channels for conveying magnitude. However, I have selected it initially anyway because it conveys a sense in which older data is less prominent than newer data, which lends to the overall salience of the visual channel. I later experiment with colour instead as a more effective visual channel.

Smoothing is achieved firstly by linearly interpolating results between dates in which data isn't available. For example, in figure 2, an observation occurs in one location in period 1, period 2 contains no data, and another observation occurs in period 3. The period 2 estimate for the location of the species is assumed to be the midpoint of the vector between them (the green circle). This is again implemented using methods available in the pandas library.

Smoothing is also implemented by taking moving averages of the locations of the data. This reduces the impact of any individual outliers/noise in the data and stops them from sharply distorting the paths that we plot whilst preserving the overall trend.

5 Part A - Results and Discussion

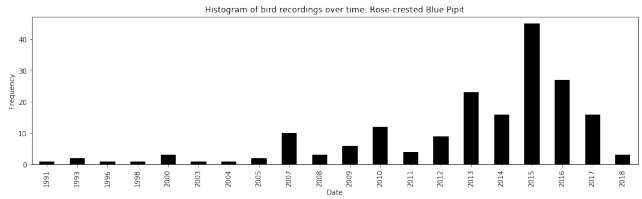


Figure 3: Distribution of recordings for the Blue-Pipit over time

The histogram in figure 3 shows that the population of Rose-Crested Blue-Pipits has changed significantly over time. We can see significant explosions in the population occurring in 2007, 2013 and 2015 in which it peaked at its maximum. Since 2015 the population has decreased significantly each year, which timing suggests may relate to activity by Kasios, but the overall number is still similar to that of the long term pre-2007 levels which suggests that if anything, it is the large increases in the intervening years which may be a notable feature and in fact recent decreases are a regression to this long run equillibrium.

We may imagine that, for example, a predator of the species has died off, or their prey has become abundant and this is why the population of Blue Pipits increased so sharply, and that is no longer the case. Of course, these narratives are not mutually exclusive. It may be that something changed which resulted in the population growing, and it may be that the activities of Kasios have since reversed that. The strength of this approach is that we can clearly see the trend in the size of the populations, the limitation is that its very difficult to infer why these changes occurred. Such explanations without further analysis are pure conjecture.

We consider whether this visualisation is effective against the following five tests.

- **Does it exhibit graphical integrity?** The visualisation does not mislead. Axes begin at zero so the relative sizes of the bars in the histogram are true and the scale is uniform. There is little risk of misinterpreting it.
- **Is it simple?** The visualisation is simple and clutter free. The only elements included are those conveying a key part of the story it tells. The data ink ratio is very high.
- **Does it use the right display?** The histogram is effectvely a bar chart measuring frequency which is appropriate for the story we are trying to tell. A line chart may also be appropriate.
- **Does it use colour strategically?** Colour is deliberately not used in this visualisation as it is not required as a visual channel.
- **Does it tell a story?** We can clearly see the trends in the population size over the time period our data covers.

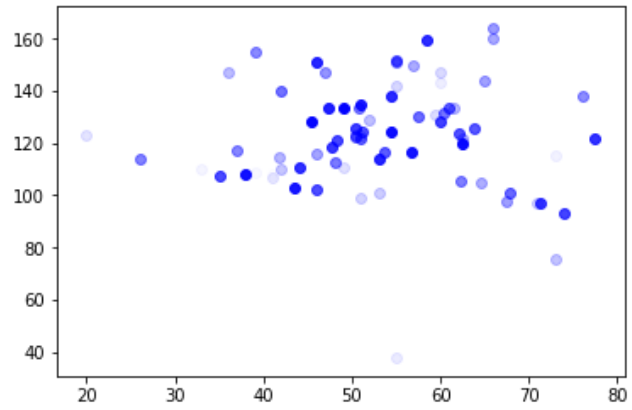


Figure 4: Simple plot of centroids with no interpolation or smoothing

In order to analyse the movement of the species within the preserve over time, I plot the calculated centroids, initially with no interpolation of values.

Figure 4 is successful in using opacity as a visual channel to show how old a set of observations is. However, it is difficult to glean how the species was moving over time. The opacity measure is also misleading, as we have simply ordered the data from oldest to newest and opacity reflects the observation's position in this order. Given the high variance in the gap between observations within the data, this is problematic. This confirms our requirement for both aggregation and smoothing. Even without these things, the plot is useful and allows us to infer that the species existing primarily in the north of wildlife preserve, and that the species does indeed move around.

Figure 5: Plot showing centroids with interpolation, smoothing and graded opacity

In figure 5 it is much easier to infer the movement of the species over time. Interpolation and smoothing of the values has created a much more useful visualisation. However, because the species has effectively returned to an area it has begun in, it is very difficult to see the lower opacity 'path' behind the higher opacity path. This is a key failing of the technique.

For the next plot, instead of using opacity as a visual channel, I experiment with using colour instead and show the most successful of these attempts.

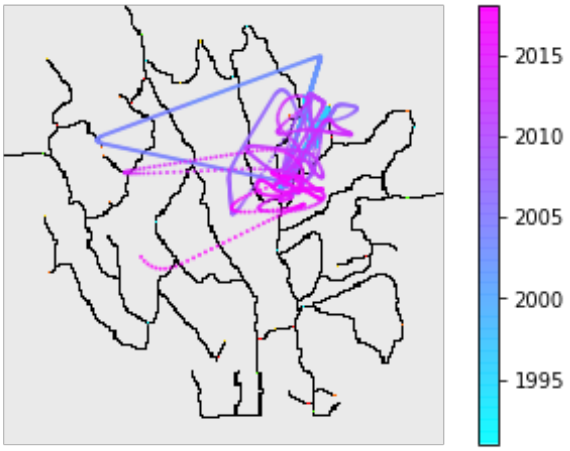


Figure 6: Plot showing centroids with interpolation, smoothing and graded colour

This plot makes it easier to identify where the species was in the same space and its movements within those periods because they are displayed with the same opacity. The addition of the colour bar allows us to identify the period associated with a segment of the path.

Finally, we plot the locations of the recordings given to us by Kasios on the same graph in order to understand whether they fit in to the pattern we can observe.

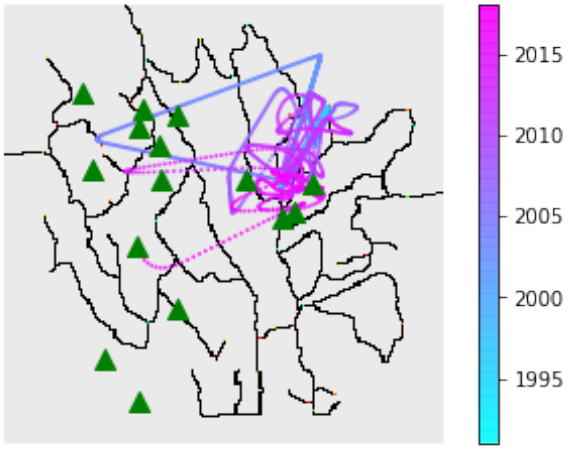


Figure 7: Plot showing centroids with Kasios' test sightings overlayed

The technique allows us to see that although a small proportion of the recordings are in locations in which we expect to see Blue-Pipits, most of them are not. Using colour as a visual channel for the year attribute allows us to infer that even where a recording is in a location in which Pipits have been seen (i.e. the top left quadrant of the map), this was not during a period of time in which we know the recordings from Kasios were made (2017-2018). This is strong evidence that the recordings provided by Kasios are not Blue-Pipits but are perhaps another bird. It may also be evidence that they have been fabricated entirely.

If we do suppose that another species has been mistaken for the Blue-Pipit, we may repeat the visualisation with the same set of recordings from Kasios but different underlying movement paths plotted for different species. Doing this for the Blue-Collared Zipper shows the recordings for this species as a better fit for the data from Kasios as illustrated in figure 8.

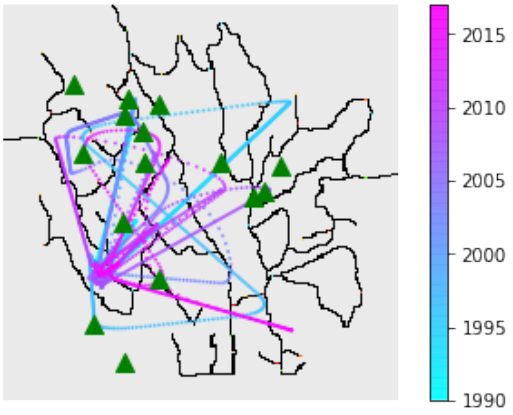


Figure 8: Plot showing centroids for Blue-Collared Zipper species with Kasios' test sightings overlayed

The main criticism of this approach is that in computing the centroids, we lose information about the 'spread' of observations around these centroids. We risk inferring a false sense of how compact the distribution of the birds is by plotting only its mean at a point in time. Ironically the original full dataset of individual observations may give us a better visual sense of both centre and spread, and can also point us towards the same conclusion that the Kasios recordings are not taken in locations we would expect to find Pipits, as show in figure 9.

Again, we consider whether this visualisation is effective against the following five tests.

- **Does it exhibit graphical integrity?** Generally the visualisation does not mislead in that it does show where the centroids are on the map at a point in time. However, there is a risk of misleading the audience as to how compact the species distribution is. There is also a risk that the interpolation suggests the movement of the species is smoother than it really is.

- **Is it simple?** The visualisation is simple and clutter free. The map is required so we can understand the positions marked in absolute terms. The plot indicates the location of the species and the colour is required so we can infer the time attribute. Unnecessary axes have been removed. However, this isn't the simplest visualisation that tells the same story (as shown in fig 9) and so is in some sense, not as simple as it could be.

- **Does it use the right display?** The display is technically a scatter plot overlayed on a map. This display allows us to relate the position of the centroids we calculate to absolute points in the park.

- **Does it use colour strategically?** Colour is used strategically as a visual channel (time) and it's very effective in doing this when accompanied with the colour bar. We saw that this was in fact more effective than using opacity. Colour is not used unnecessarily anywhere.

- **Does it tell a story?** We can clearly see which parts of the parks the species has inhabitated at a specific point in time.

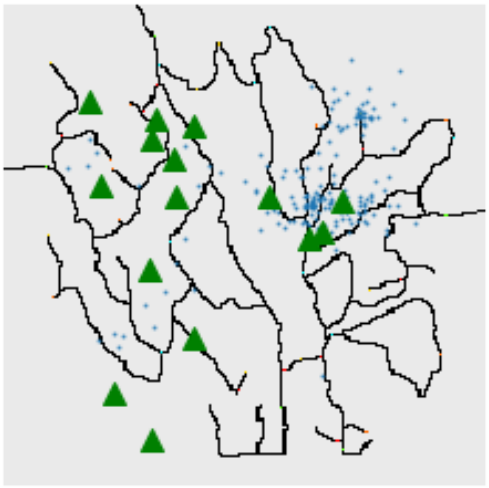


Figure 9: Simple plot of individual observations of Blue Pipits with Kasios' test sightings overlayed

6 Part B Design

In Part B intend to create constellation diagrams for the recordings provided by the original researchers and further constellation diagrams for the recordings provided by Kasios. This will allow us to visually compare the recordings to see if they are sufficiently similar to be considered as from the same species.

The data we have is a digital sound recording which can be interpreted as tabular data, or a set of time series (one for each frequency). The first step to creating a constellation plot is to extract a spectrogram from the recording. The spectrogram plots the frequency of the sound against time, with amplitude represented through the visual channel of colour. I will use the librosa python library to read the mp3 files that we are given, and generate a spectrogram.

Once a spectrogram has been extracted, we have a visual representation of the sound but we must reduce it to the 'peaks.' That is to say we find the points which represent local maxima of amplitude within the image. The original authors do not detail how these peaks are found, so I tried a number of algorithms. The first three algorithms I tried were those included in the SKImage python library for 'blob detection,' specifically 'Difference of Gaussians,' 'Difference of Hessians' and 'Laplacian of Gaussians.' These are all able to identify 'blobs' within an image. They are all somewhat able to identify the peaks within the spectrogram, but differ in their ability to do this effectively, and also the time taken to process.

I also implement my own very simple algorithm to identify the peaks; a thresholding approach. I convert the image to grayscale, and identify all pixels in the spectrogram that are beyond than a parameterised threshold. The advantage of this approach is it's very simple, the disadvantage is that it tends to find multiple points in areas of high amplitude rather than single peaks. We also have to select an appropriate threshold.

Based on experimenting with all these approaches, I have selected the 'Difference of Gaussians' as the most effective method for identifying the peaks within the image.

Once we are able to explicitly extract a constellation diagram from a sound file, we need a way to compare these. This could be as simple as using the matplotlib libary to plot them on the same axes. This would be of limited use as we may be comparing files of different length or starting at different points in the bird's call/song. In those cases, the two constellation plots are unlikely to immediately line up.

Instead I use the Bokeh interactive plotting library. This library is written and configured in Python, but returns HTML/Javascript for an interactive plot that can be zoomed and translated. I will build a very basic user interface for translating the plots. This is required to allow the user to explore the two plots together to see if they are indeed a match.

In all of the constellation diagrams I produce, we use the visual channels of position to show where the peaks in amplitude for a recording are, and the visual channels of colour to show categorical information (i.e. which recording a peak is from). All of these are identity channels. There are effectively no magnitude channels in this visualisation.

7 Part B - Results and Discussion

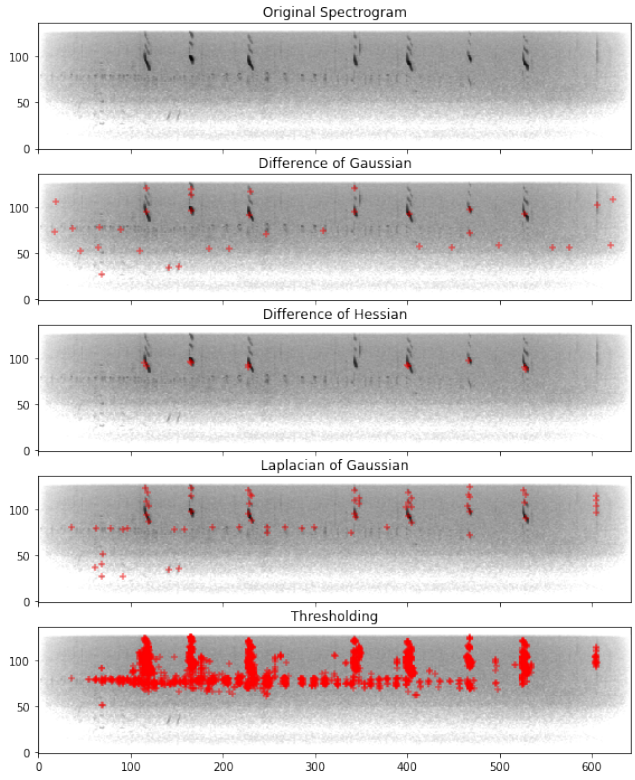


Figure 10: Plot 5

Figure 10 shows the spectrogram and the results of the difference approaches used to find the peaks within it. Plotting the results of each approach on top of the image of the spectrogram allows us to quickly see how effective each method is. The Difference of Gaussian approach tends to find most peaks, but misses some of those close to others, which may suggest a calibration error. The Difference of Hessian approach finds very few of the peaks, regardless of how it is calibrated. The Laplacian of Gaussian approach finds most peaks. The thresholding approach I designed certainly finds all peaks, but as predicted, is unable to select a single peak in dark areas of the spectrogram. I select the Difference of Gaussian approach to go forward.

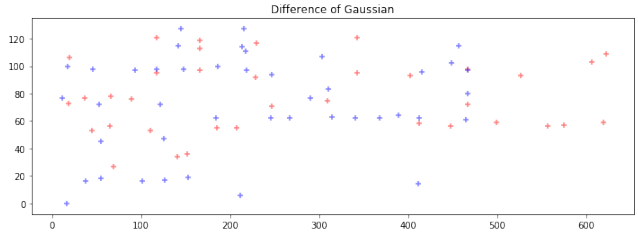


Figure 11: Plot 5

Figure 11 shows these constellation plots generated using the Difference of Gaussian approach for two different files overlayed over each other. In this case they're two different recordings of the same species, and so we might well expect to see the constellation plots lining up. The visual channel of colour conveys which of the two files the point in the constellation plot belongs to.

In order to see if the constellation plots line up, we need the ability to translate one of the plots over the other. The interactive Boken plot allows the user to compare the constellation plots we generated in an interactive framework. It starts with them selecting the two sounds they would like to compare. One of these may be the unknown files given to us by Kasios. The user can translate one of the two plots left, right, up and down, as well as zooming and panning across both plots. I use the visual mark of a cross as this is the closest to what the original authors used, and it allows the user to see at a very granular detail whether the constellation plots line up.

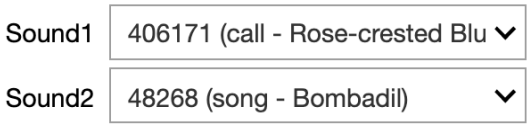


Figure 12: File selection widget

Figure 13: Interactive Bokeh constellation plot

At this point we have everything we need to make a visual analytics informed decision on whether the data supplied by Kasios is indeed recordings of Blue-Pipits. We can test the accuracy of the approach by comparing species of the same bird. This should, for an accuracte classification approach, find that the recordings are the same. However, when we do exactly this, it is not conclusive that the approach works. There may be a number of reasons for this, but first and foremost, we make the strong assumption that birds of the same species always sound the same. In fact, it may well be the case that birds vary their calls over time and over space. It may even be the case that bird calls are entirely unique to the individual bird.

When comparing the recordings provided by Kasios to samples of Blue-Pipts from the researchers, we do not see the constellation plots lining up which could be evidence they are not Blue-Pipits, but unfortunately the approach does not allow us to draw such a strong conclusion.

For the final time, we consider whether the visualisation is effective against the following five tests.

- **Does it exhibit graphical integrity?** The visualisation does not mislead. It accurately reflects peaks in the spectrogram of the underlying sound file. The constellation plots are fundamentally quite abstract, and not something the audience will have a strong intuition for. The only inference we can expect the audience to make is whether they're the same underlying sound or not.

- **Is it simple?** The visualisation is simple and has done away with significant clutter of the underlying spectrogram. It takes a very complex question and reduces it to a very simple form.

- **Does it use the right display?** The display is tehnically a scatter plot (frequency against time). It does not represent a relationship between these two variables and we must be careful to not infer that.

- **Does it use colour strategically?** Colour is used effectively as a categorical visual channel to represent simply which of the two underlying sound file the plot belongs to.

- **Does it tell a story?** Setting aside the issues discussed above, it allows us to tell the story of whether the test data received from Kasios is credible or not and to see visually whether the test data 'sounds like' Blue Pipits.

8 Conclusion

Mapping the movement of the species over time gives us insights in to the long term trends of the species, combined with the histogram showing the size of the populations gives us a clear sense of its overall health. We are able to conclude that actually the Blue-Pipit population does move around significantly, and the timing of large movements does not seem to relate to the known dumping events around 2016. Whilst we can also conclude that the size of the population has decreased in recent years since that event, we can't say whether this is a natural regression to a long term mean or as a result of Kasios' activities. Perhaps with a dataset that stretches further back, we could determine what the long run average population size is and draw a stronger conclusion.

What we can more strongly infer is that the recordings by Kasios do not fit with our understanding of where the the Blue-Pipit population was in the park at any point in time and it is likely that Kasios have either recorded a different species, or fabricated the recordings in some other way.

The visualisation approach here passes the tests set out for an effective visualisation, telling a complex story through a couple of relatively simple visualisations. To improve upon this approach, I would introduce a method to explore the hypotheses that there were multiple distinct populations of the same species of bird within the wildlife preserve, perhaps using a k-means approach. I would also look for an extra visual channel that could convey the spread of a species around its centroid as I see the risk of misleading the audience on the spread of populations as the biggest issue with the visualisation currently.

The approach of using constellation plots as a visual analytical tool has not been as successful. This is because the hypothesis that different recordings of the same species of bird will produce very similar constellation plots has not held. This could be because of some unrevealed feature of the bird populations, e.g. bird songs for the same species are different in different parts of the park, or they change over time. It could also be because of some failing of the algorithm I've used to identify peaks within the spectrogram.

These issues aside, the visualisation approach of creating constellation plots is successful. The plots show in a much simpler way than the spectrogram where the peaks in amplitude are, and the interactive Bokeh visualisations make it easy for the user to explore and quickly establish whether they match or not. To improve these I would develop a way to explore recordings that we particularly may expect to match, e.g. from the same part of the preserve, or the same point in time.

References

[1] N. Andrienko and G. Andrienko. Informed spatial decisions through coordinated views. *Information Visualization*, 2(4):270–285, Dec. 2003. doi: 10.1057/palgrave.ivs.9500058

[2] B. Bäumle, I. Boesecke, E. Cakmak, R. Buchmüller, W. Jentner, Y. Metz, D. A. Keim, and J. Buchmuller. Interactive webtool for tempospatial data and visual audio analysis. IEEE Conference on Visual Analytics Science and Technology (VAST), Berlin, Germany, 2018.

[3] E. Cakmak, U. Schlegel, M. Miller, J. Buchmüller, W. Jentner, and D. A. Keim. Interactive classification using spectrograms and audio glyphs. IEEE Conference on Visual Analytics Science and Technology (VAST), Berlin, Germany, 2018.

[4] M. Grosche and Serra. Audio content-based music retrieval. Conference on Multimodal Music Processing, 2012. doi: 10.4230/DFU.Vol3.11041.157

[5] O. Hellmuth, J. Herre, E. Allamanche, M. Cremer, T. Kastner, and W. Hirsch. Advanced audio identification using mpeg-7 content description. 11 2001.

[6] B. Sullivan, S. Kelling, C. Wood, M. Iliff, D. Fink, M. Herzog, D. Moody, and G. Ballard. Data exploration through visualization tools. 02 2008. doi: 10.13140/2.1.4892.5126