Arrival Modelling using BirdTrack Data

Page Huang

August 24, 2023

Overview

The pipeline consists of:

- Creating a presence-absence dataset from raw csv data for a given species and year (with one row for each complete list, a 1 or 0 to indicate whether the species was spotted, and a list length indicator)
- Cleaning this dataset if needed (filtering out lists of a certain list length)
- For each grid reference square, if it passes thresholding, fitting a binomial GAM and producing an estimated arrival date based on the data. The equation used is presence of species $\sim$ s(day of list) + s(list length).
- Once first stage estimates have been produced, another GAM is fitted to spatially smooth arrival dates. The equation used is arrival date $\sim$ s(easting, northing).

The arrival date is taken as the date where ten percent of the "first peak" of the reporting rate minus minimum reporting rate (to account for increased baseline numbers in case of overwintering) is.

Three heatmaps of the UK are produced: two showing first-stage GAM modelling (one is coloured to show spread of arrival dates, the other coloured to show volume of data used for each square), and one showing the smoothed date model. There is also a separate script for analysing an individual grid reference, which produces a line graph of the fitted curve, as well as weekly reporting rate bar charts and intercepts for the first peak detected and the ten percent of that peak.

Some running notes:

- By default, the pipeline processes all the squares in the UK – restricting to an eg. region will have to be done before the pipeline is run. It is possible to manually modify the second stage smoothing GAM to smooth over a specific region (by manually changing the grid for second stage smoothing) but there are currently no functions built for this.
- The crop on the basemap is dependent on the maximum and minimum northing and easting GAMs done.
- Species code used are the two-letter species codes.
- Parameters are set at the top of the script. This includes grid square size in metres – 10000, 20000, and 50000 respectively.
- For raw CSVs with a lot of data, I have found that creating the presence-absence dataset, saving this dataset as a CSV, then reading that into the script instead (using readr) causes the pipeline to run faster.
- The basemap used is from https://gadm.org/download_country.html. The basemap function should automatically download the shapefile into a "data_in" folder if not already present, though may have to be manually downloaded if not.

Results

The initial species data used to develop this pipeline was Swallow. This is because it is abundant in occurrences and has a clearly defined arrival pattern country-wide during the summer. As a consequence, the pipeline currently only works for summer migrants. More examples of species analysed are in the appendix.

Below are the maps produced from the data (with one with 20km squares), as well as an example of the line graph and an example of a “regional” dataset with text (note that the “show text” parameter should always be set to false when displaying the second stage smoothing as there is no text label created for that):

These are done with the default thresholds (need at least 100 lists and 30 detections for a square and 4 lists per month), as well as cleaning of lists of length 1 (as there is the possibility of a user incorrectly inputting a casual list as a complete list).

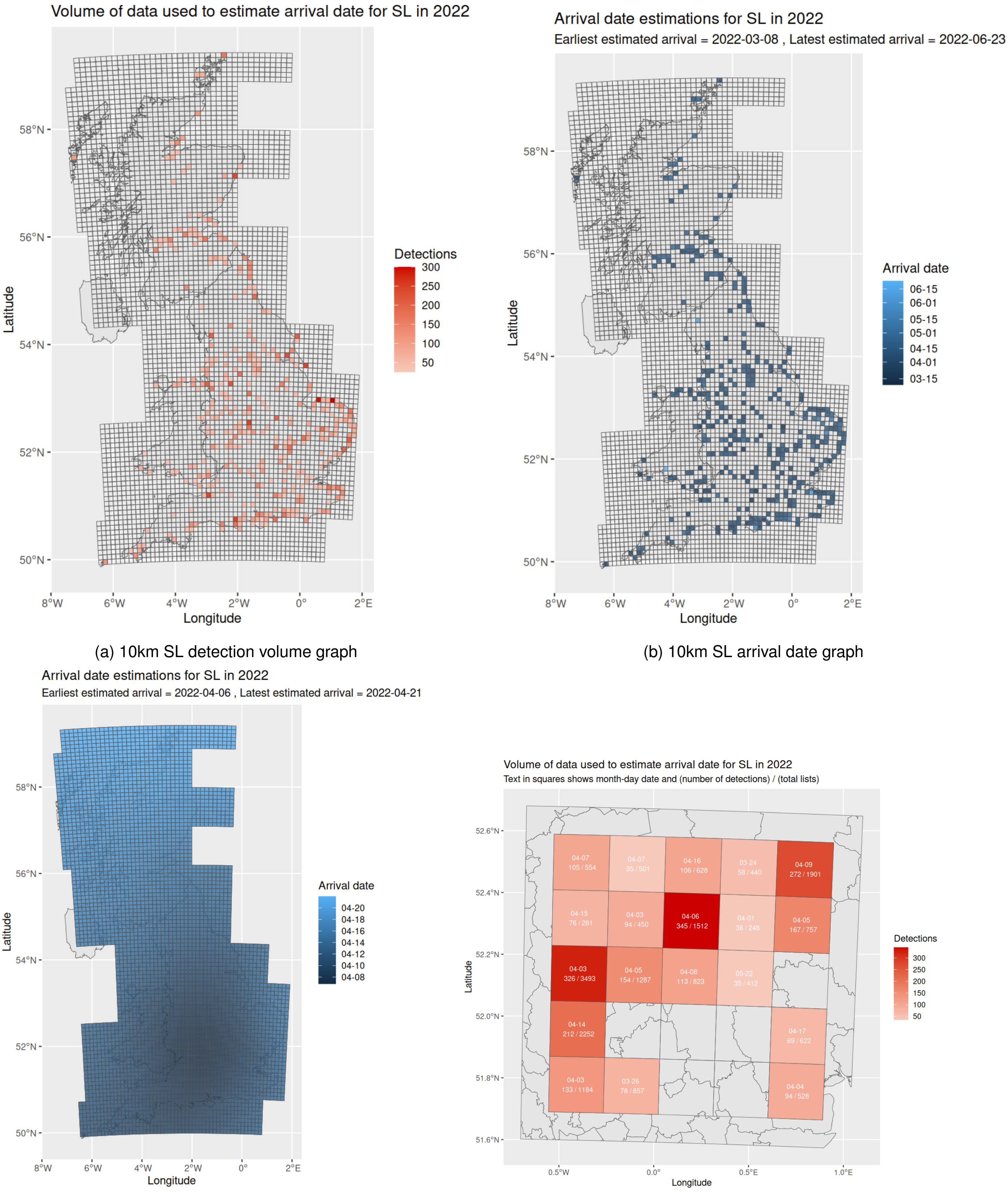
I personally found that I had to set the lists per month threshold to a lower value than I was expecting, as it would end up excluding squares with otherwise valid data for a species. Since the intended species are migrant species, if there are little/no lists in months that a species is not present in the country, it has no impact – and if it is in months when the species is present, it will be detected in the species detection threshold. Also, I did not see any drastic effect on date estimations by filtering length of one lists. However, the thresholds remain in case they may be needed for another species.

Similar species such as House Martin and Swift work well with the pipeline. However, the pipeline starts to run into difficulty with overwintering and regionally-resident species, like Chiffchaff. Some measures implemented to try and prevent incorrect arrival date prediction include:

- Taking into account an increased baseline due to overwintering [4a]
- Ignoring small peaks (currently set to ten percent of the maximum reporting rate)
- Ignoring any initial downwards reporting rate slope [4b]
- Rejecting squares that may pass initial thresholding, but produce a GAM that has no peak (which can happen if species is prevalent throughout the year eg. certain pockets of Chiffchaff) [4c]

Even despite these measures, there are some squares where overwintering Chiffchaff cause ”peaks” in the middle of January big enough to bypass the small peaks thresholding. For example, the square used in [5] has enough lists in the month of February and March to bypass the monthly list coverage threshold, albeit no detections, so the baseline level is set to zero. A similar issue would also occur if baseline level at the beginning of the year was high, but at the end of the year was low or close to zero.

Other (non-winter) migrant species do not have enough data to create the maps. For example, there were only 28 present records for nightingale for the whole of the UK for 2019, and hence no GAMs were produced. The same occured for Little Tern in 2022. In order to prevent unneccessary processing, a ”total detections” threshold for the entire dataset needs to be added, and for sparse data potentially a GAM fitted for the whole country data instead [6].



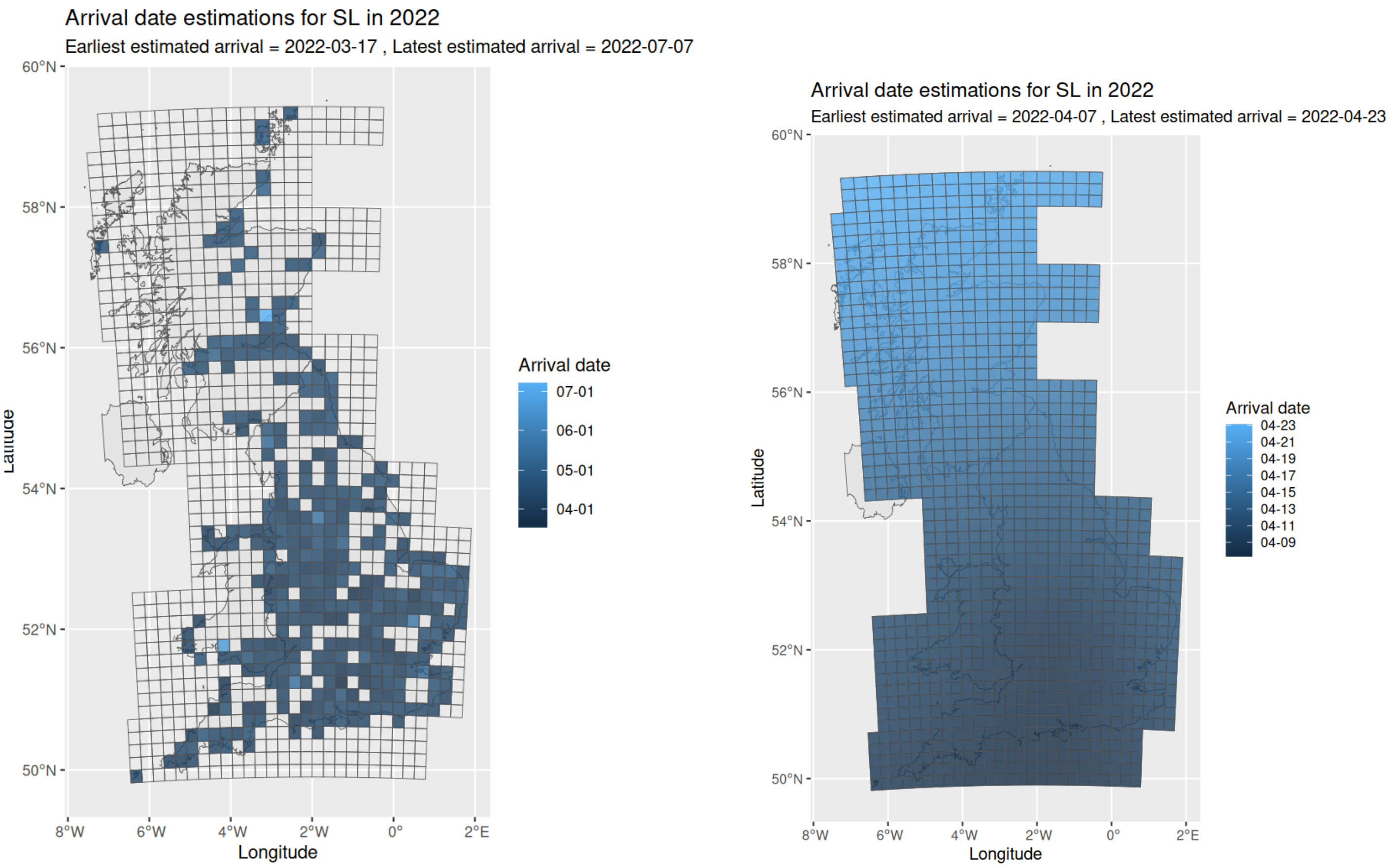
(a) 10km SL detection volume graph

(b) 10km SL arrival date graph

(c) 10km SL smoothed arrival date

(d) 10km SL for Thetford (TL)

Figure 1: Graphs generated for 2022 Swallow data, for the UK and for Thetford.


Arrival date estimations for SL in 2022
Earliest estimated arrival = 2022-03-17 , Latest estimated arrival = 2022-07-07
60°N
58°N
56°N
54°N
52°N
50°N
Latitude
8°W
6°W
4°W
2°W
0°
2°E
Longitude
Arrival date
07-01
06-01
05-01
04-01
Arrival date estimations for SL in 2022
Earliest estimated arrival = 2022-04-07 , Latest estimated arrival = 2022-04-23
Arrival date
04-23
04-21
04-19
04-17
04-15
04-13
04-11
04-09

(a) SL 20km squares

(b) SL smoothed 20km squares

Figure 2: 2022 Swallow data but using 20km grid references. Whilst the first stage date spread becomes skewed towards later dates compared to the 10km grid references, the second stage smoothing is more similar to the 10km version.

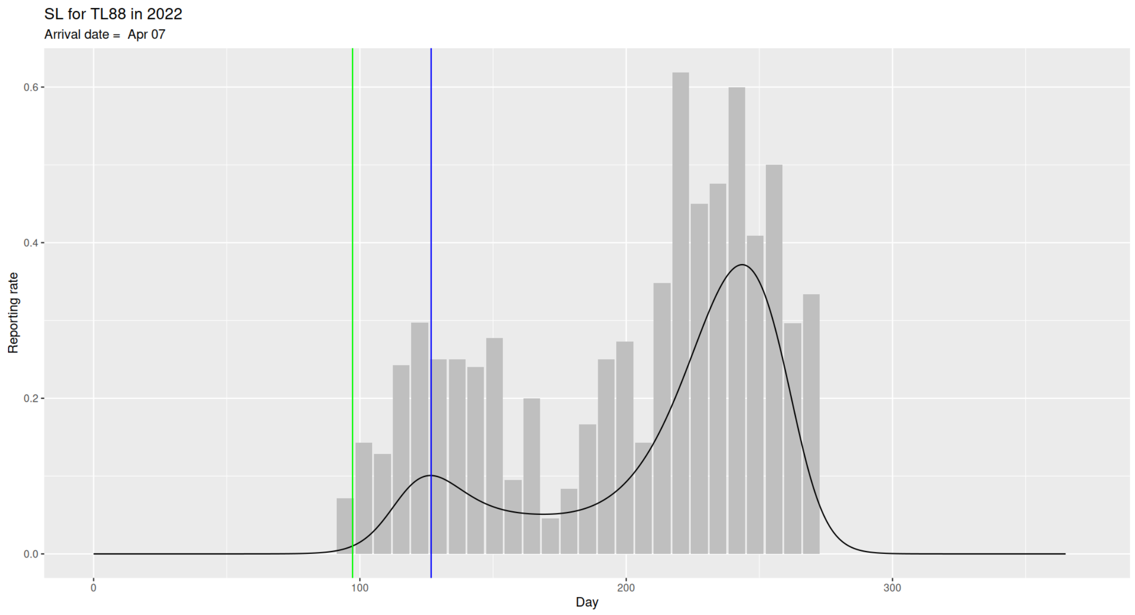
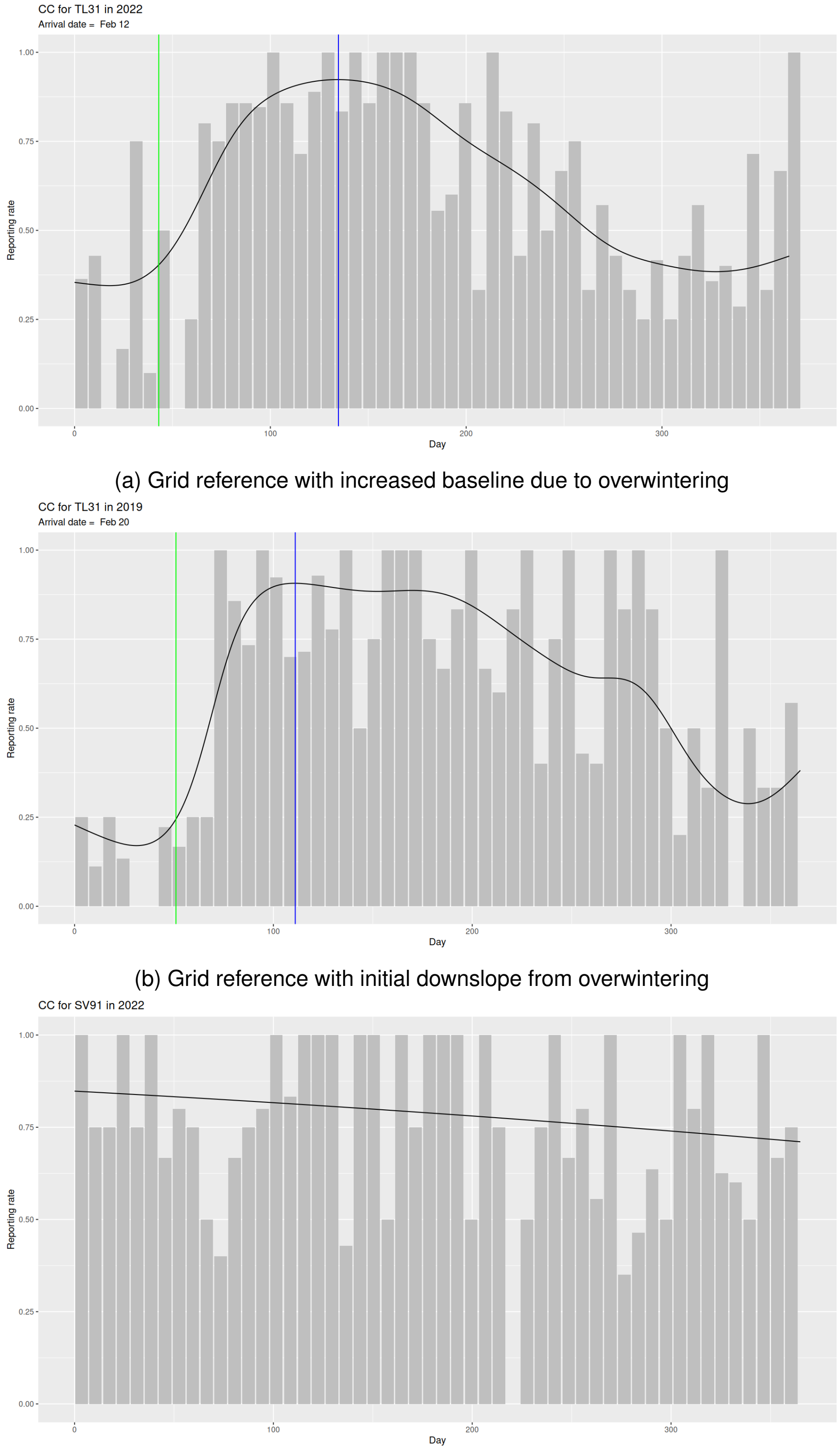


Figure 3: Resulting line graph when running the script to analyse an individual grid reference



(a) Grid reference with increased baseline due to overwintering

(b) Grid reference with initial downslope from overwintering

(c) Grid reference that due to frequent detections does not form a peak

Figure 4: Examples of grid references with various modelling challenges.

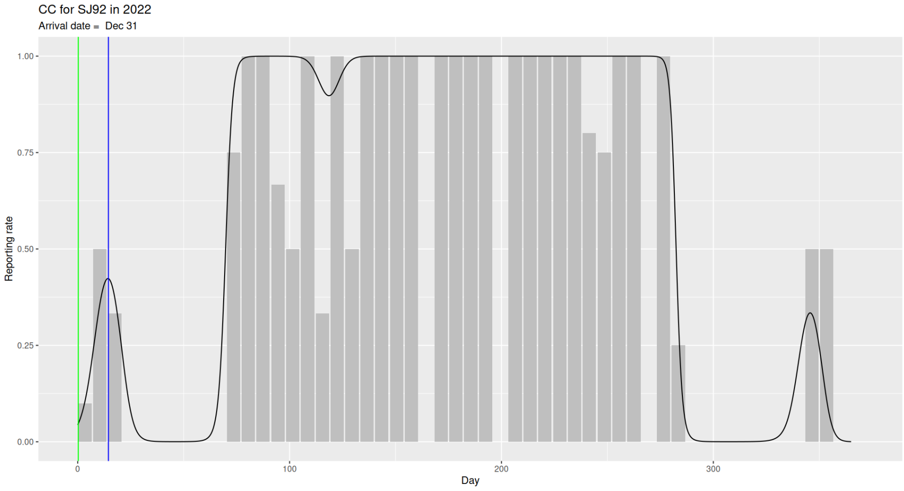


Figure 5: Grid reference with an early ”peak”.

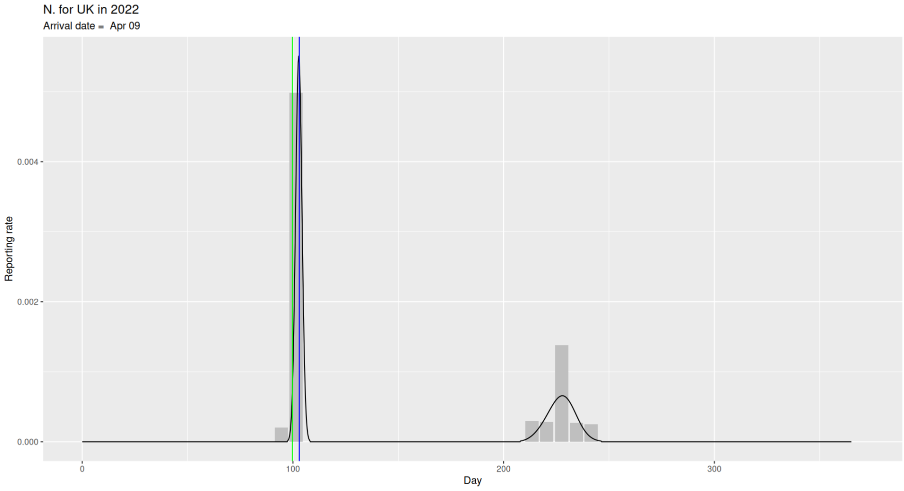


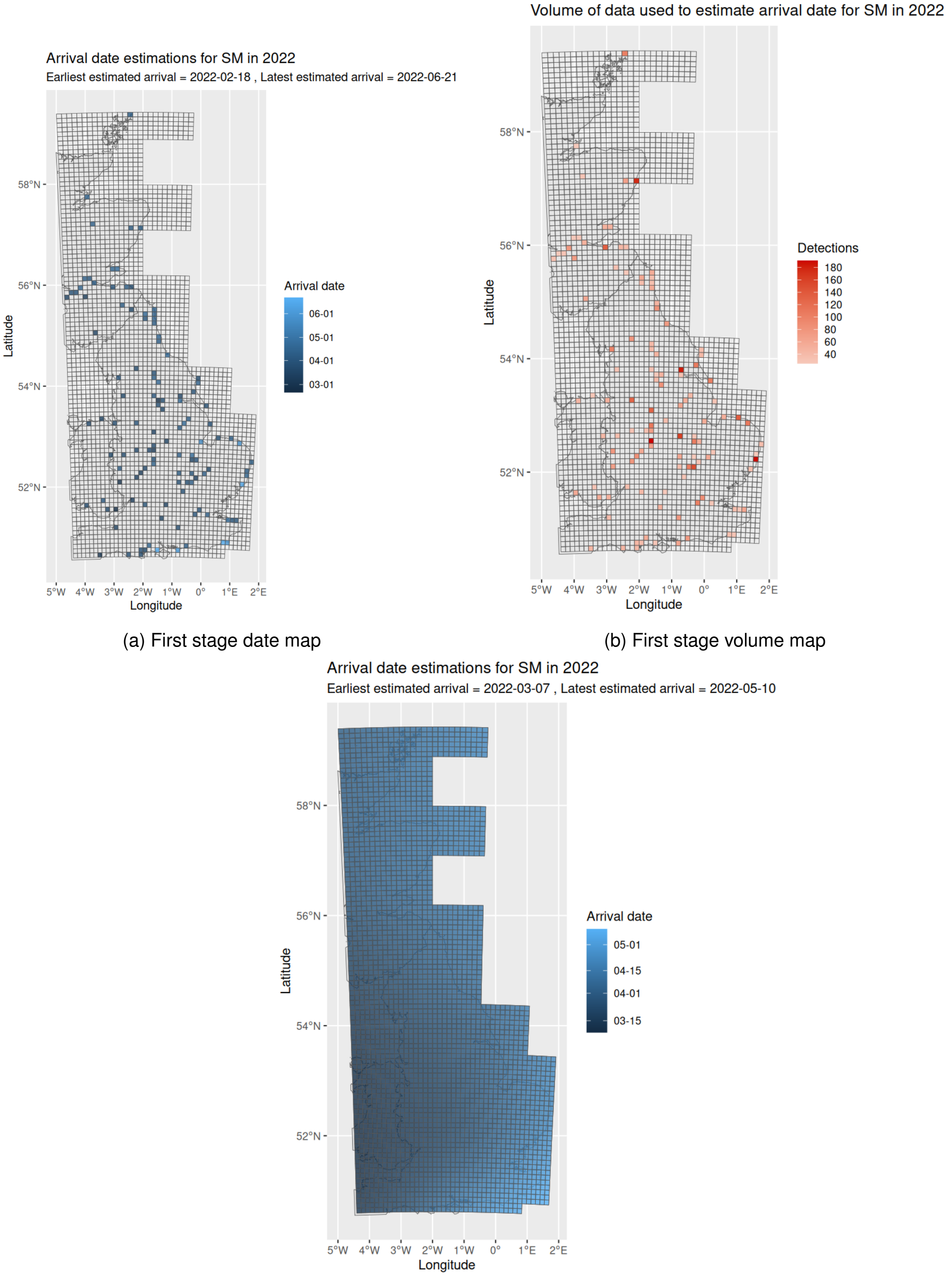
Figure 6: The single GAM fitted for Nightingale data.

Next Steps

There are three potential next steps for this pipeline, aside from adjusting thresholding/peak detection to account for more edge cases in data. One involves increasing the flexibility in what locations are mapped - for example, ignoring areas where species does not appear (inland grid references for coastal species, Scotland for a species appearing only in England, or even smaller pockets of localised species eg. Osprey [9]) or ignoring areas where migrant species have resident populations. This may involve more preliminary analysis - creating a distribution map, for instance.

Another next step involves increasing the amount of migrant species that the pipeline can apply to - namely wintering species. The simplest solution would be instead of gathering the dataset from January - January, to gather the dataset from June - June. Because the pipeline currently assumes the dataset is from January-January for a single year (in collecting the presence absence data as well as translating the day number to a date), this would require some tweaks in some functions - changing from filtering by a single year and potentially adding an half-year offset to the day number.

The last next step is to account for uncertainty in the model. This would involve using weightings in the second stage smoothing to carry forth the uncertainty in the arrival date (which could be calculated by running all potential GAM curves for a grid reference through the pipeline and calculating dates for each).

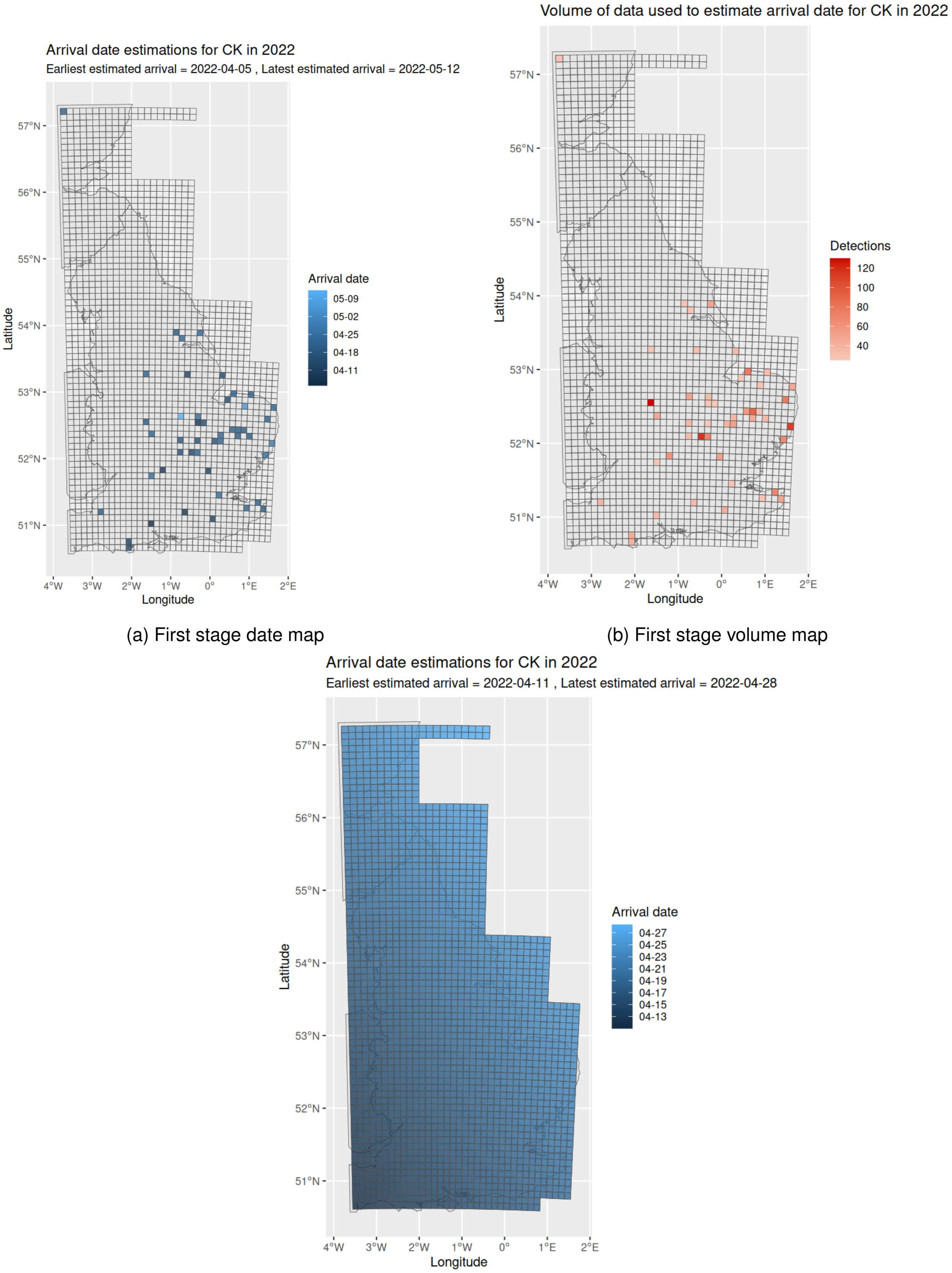

Arrival date estimations for SM in 2022
Earliest estimated arrival = 2022-02-18 , Latest estimated arrival = 2022-06-21
Latitude
58°N
56°N
54°N
52°N
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E
Longitude
Arrival date
06-01
05-01
04-01
03-01
Volume of data used to estimate arrival date for SM in 2022
Latitude
58°N
56°N
54°N
52°N
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E
Longitude
Detections
180
160
140
120
100
80
60
40
Arrival date estimations for SM in 2022
Earliest estimated arrival = 2022-03-07 , Latest estimated arrival = 2022-05-10
Latitude
58°N
56°N
54°N
52°N
5°W 4°W 3°W 2°W 1°W 0° 1°E 2°E
Longitude
Arrival date
05-01
04-15
04-01
03-15

(a) First stage date map

(b) First stage volume map

(c) Second stage smoothing

Figure 7: Sand Martin

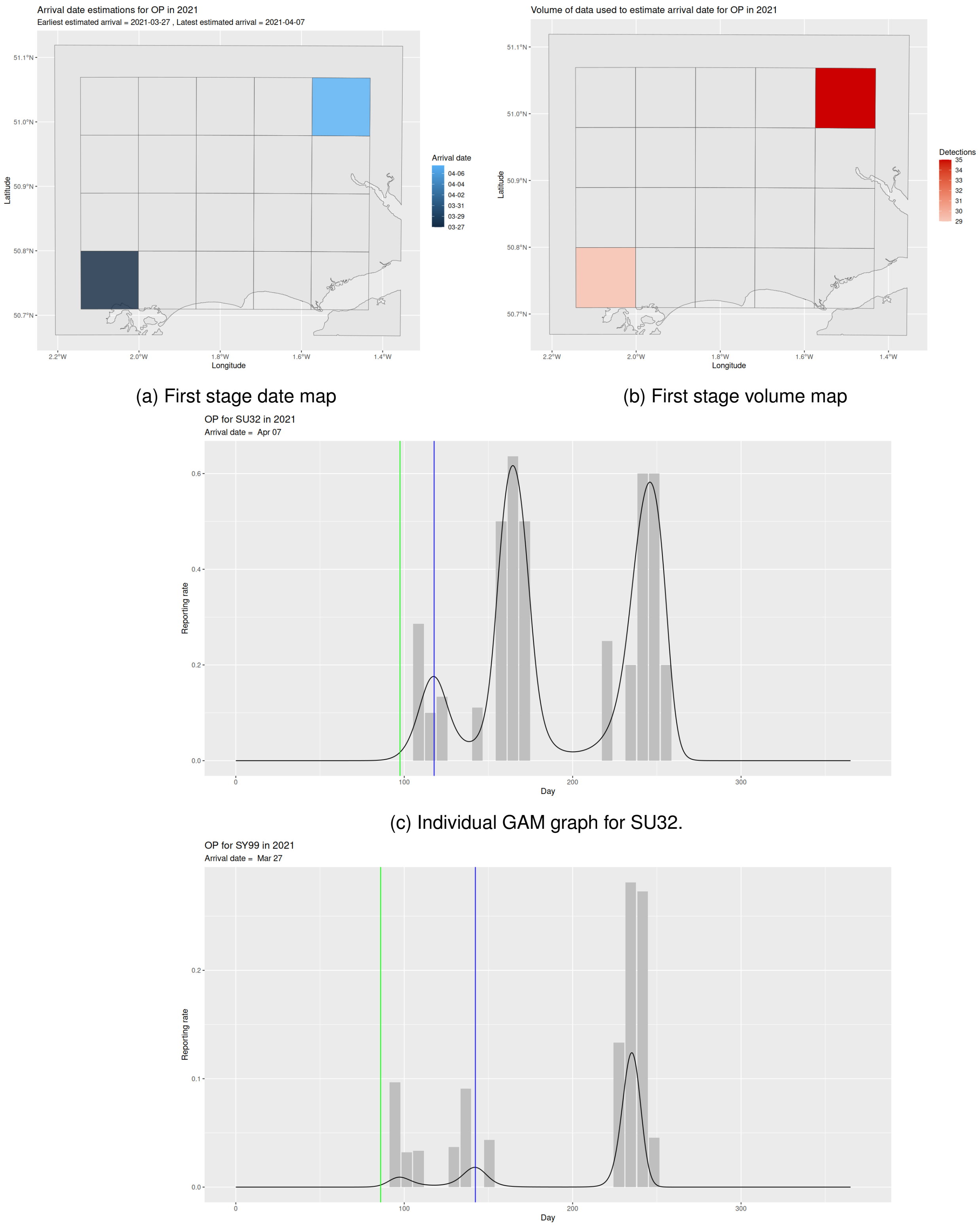

Arrival date estimations for CK in 2022
Earliest estimated arrival = 2022-04-05 , Latest estimated arrival = 2022-05-12
Arrival date
05-09
05-02
04-25
04-18
04-11
Volume of data used to estimate arrival date for CK in 2022
Detections
120
100
80
60
40
Arrival date estimations for CK in 2022
Earliest estimated arrival = 2022-04-11 , Latest estimated arrival = 2022-04-28
Arrival date
04-27
04-25
04-23
04-21
04-19
04-17
04-15
04-13
Latitude
57°N
56°N
55°N
54°N
53°N
52°N
51°N
Longitude
4°W
3°W
2°W
1°W
0°
1°E
2°E

(a) First stage date map

(b) First stage volume map

(c) Second stage smoothing

Figure 8: Cuckoo



(a) First stage date map

(b) First stage volume map

(c) Individual GAM graph for SU32.

(d) Individual GAM graph for SY99.

Figure 9: Osprey. There was not enough GAM models to do second stage smoothing, so individual graphs have been done instead.