DATA SCIENCE IN ECOLOGY:

BIRD MIGRATION USING CITIZEN SCIENCE DATA

Pedro Nicolau MSc

A LITTLE CONTEXT…

- Born and raised in Lisbon;
- Started birding at 12 and became active member of the Portuguese birdwatching community;
- Bachelor's Degree in **Biology** at **FCUL** – *Environmental Biology;*

- Master's Degree in **Biostatistics**

Ciências
ULisboa

Master's Thesis: *Estimating the spatio-temporal variation in bird phenology using citizen science data*

ALISON JOHNSTON

BTO
Looking out for birds

Cornell

TIAGO MARQUES

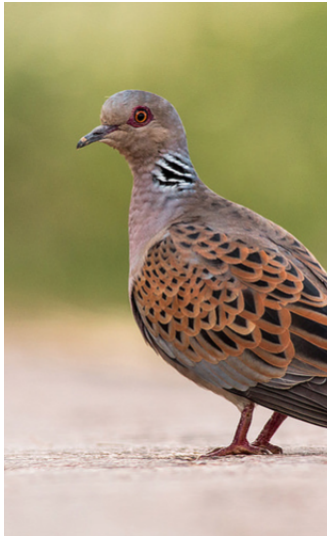
FC
Ciências
ULisboa

University of
St Andrews

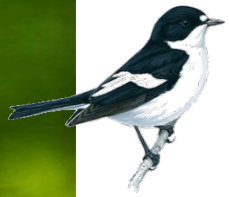
ECOLOGICAL PROBLEMS

- Migratory birds are being deeply affected by climate change.
- Spring has advanced and so they are forced to adapt: migrate early from the breeding grounds and start breeding earlier.
- Failing to adapt -> Extinction

- Not all birds are adapting. Why?
- Previous studies are limited and don’t show consensus
- One possible mechanism: **Gap between arrival and breeding**.



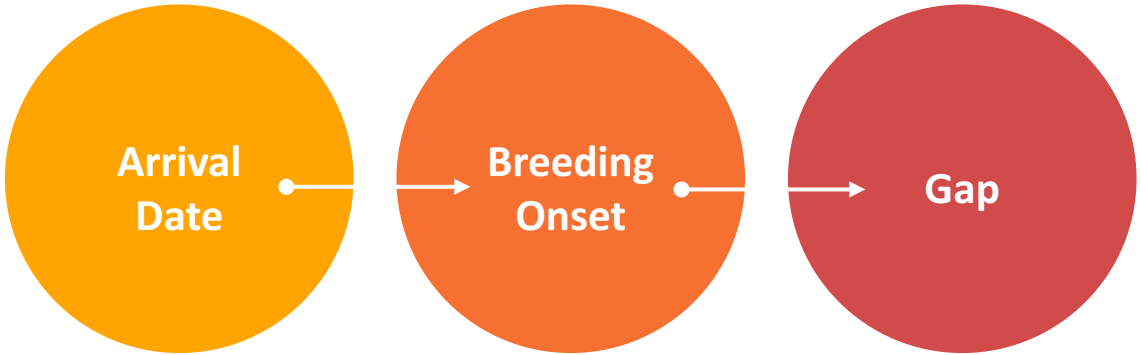
OUR PROJECT



What is the **time period** (gap) between arrival and breeding in long-distance migrants?

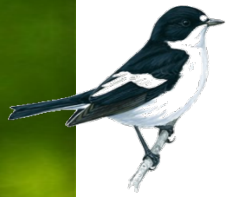
Does it **vary with latitude**? And with **year**?

Pied Flycatcher as our study species;
UK from 2013 to 2016.



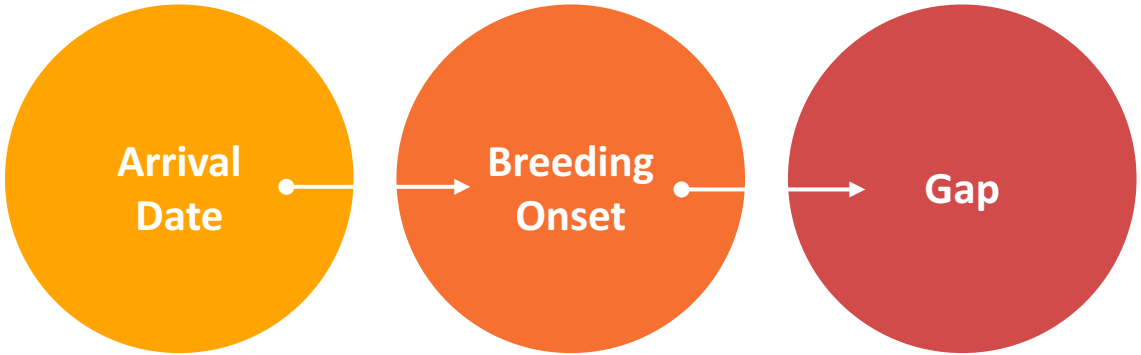
Individual data would be very costly, so we must estimate these dates separately at the population level

OUR PROJECT



Derived estimate from modelling presence/absence from large-scale databases collected by regular volunteers

Specialized volunteers that collect specific nest data annually (NEST RECORD SCHEME)



Breeding date – Arrival date

Looking at 50% of the population, instead of average of individuals

SOURCE OF DATA: **CITIZEN SCIENCE**

Research conducted using data collected by regular volunteers, with specific knowledge in certain scientific areas

- **Large volumes of data with low investment;**
- **Extensive representability in space and time;**

- Subject to a number of **biases and confounding variables**: it is important knowing what it can be used for, and especially what it cannot.

CITIZEN SCIENCE *in ornithology*

- Birdwatching has a very big number of people across the globe, **two million birdwatchers in the UK**
- Several online platforms, such as **BirdTrack or eBird**, allow users to submit their bird observations on a daily basis

- Thousands of complete checklists per day, providing presence/absence

eBird

iGoTerra
THE WILDLIFE NETWORK TO LEARN, EXPERIENCE AND SHARE

METHODS
Pre-processing

Data processing

- Raw data contained over 10 million observations (> 5GB);
- All analysis conducted in **R**, mostly with package ***dplyr****;*

- *Geographical coordinates like latitude, longitude and altitude + habitat variables were cross-referenced with external databases;*

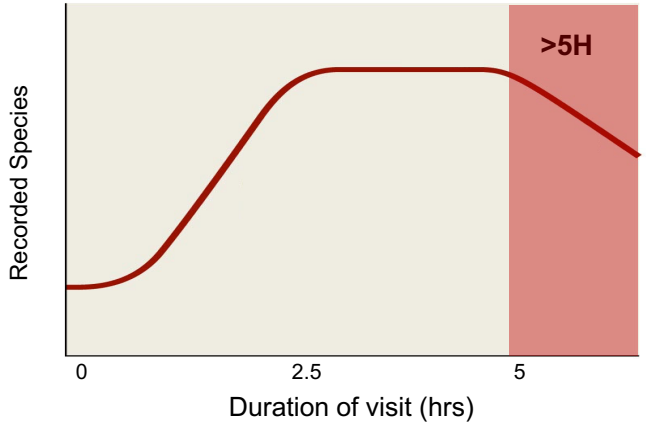
- Extensive filtering involving removing checklists that were:
  - **Duplicated;**
  - **Too long;**
  - **Incomplete**;
  - **Outside of the breeding range** or at migration hotspots...

Less than 90% of the initial observations made it

METHODS
Pre-processing

Checklist Duration

- Number of species is expected to increase with duration of visit;
- What if longer visits actually harmed the detection of certain species?
- **Number of recorded species started decreasing when over 5 hours...**

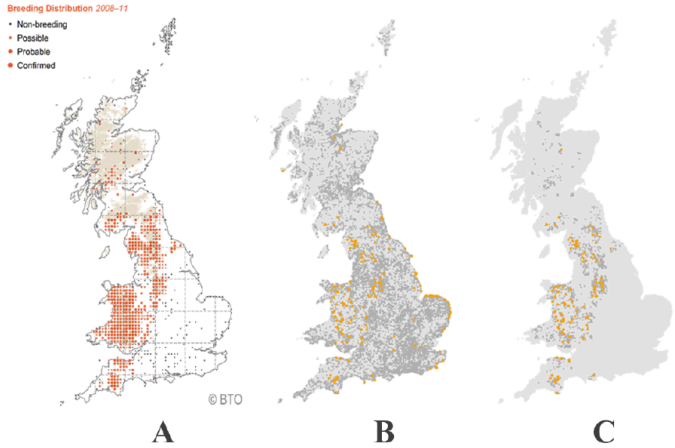


METHODS
Pre-processing

Filtering by breeding grounds

- Excluding passage (migratory) birds to make sure our observations referred to **individuals at the breeding grounds;**

- **Breeding Bird Atlas (2007-11)** was used as the database to provide the 10-km squares where pied flycatcher was detected breeding



METHODS
Overview

R package ‘mgcv'
Generalized Additive Models (GAMs)

- Modelling the response variable (probability of detection) as a function of a spatio-temporal joint smooth and other covariates describing habitat and effort: *computation-efficient functions like **bam***

- Ten-folded Cross-Validation to test the predictive abilities;

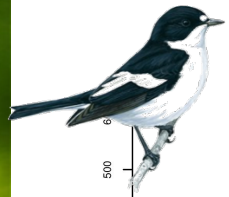
- Non-parametric **bootstrapping** (lengthy running periods)
    - *Multi-node cluster to run models individually*

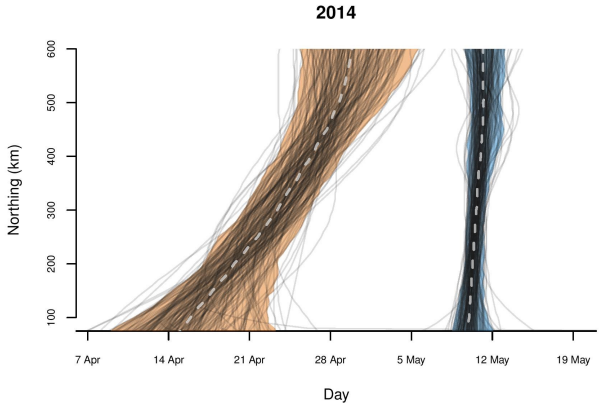
- **To obtain the gap, subtract the median bootstrapped values between each set of estimated dates.**

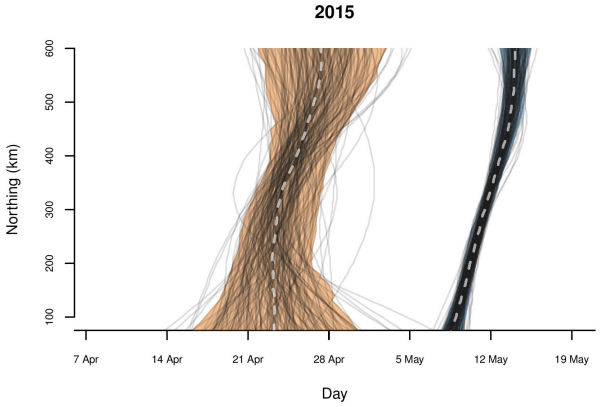
RESULTS

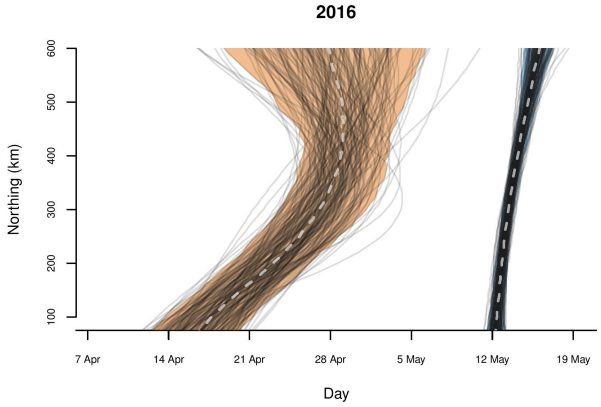
Arrival date vs Breeding onset

(2013-2016)





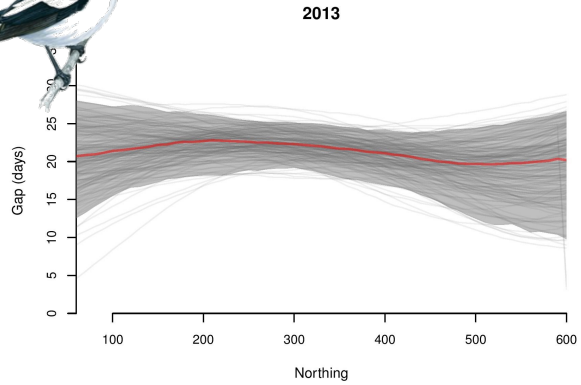


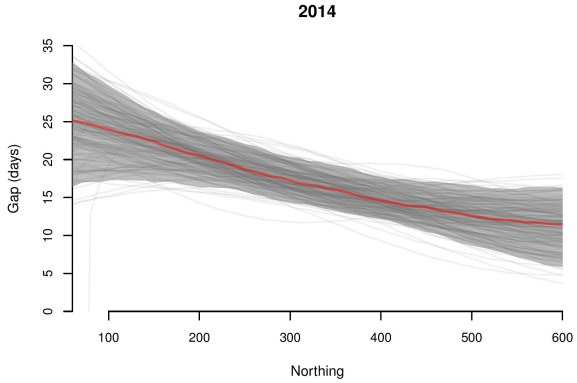


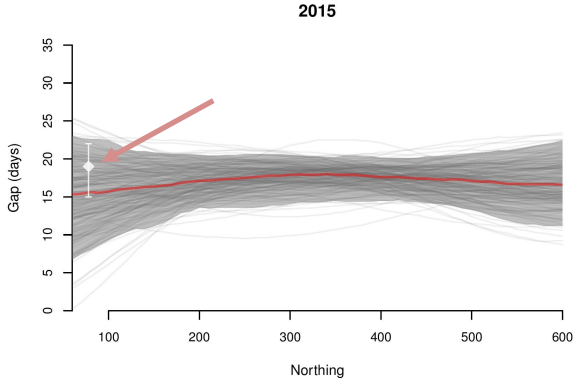
RESULTS

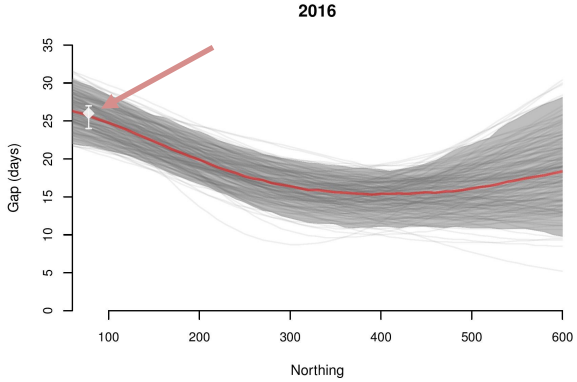
Arrival-breeding Gap

(2013-2016)









ECOLOGICAL CONCLUSIONS

- **Arrival is always later in the north**, with variation up to 15 days of difference; overall a **flexible** process
- **Breeding onset** tends to be later in the north, but with **little variation** within the same year and between years
- The gap from the end of migration to breeding onset varies from **less than 10 days to just under 30 days.**
- Birds can take less time to initiate breeding if required (**adaptable**).

DISCUSSION & REMARKS

In wanting to solve a “simple” ecological question, how much time do birds spend from migration arrival to breeding onset, we had to:

- Use massive amounts of data;
- do extensive data filtering and processing;
- complex statistical modelling and validation;

FINAL REMARKS

- Data interpretation requires ecological knowledge and both processing and modelling do not have a fixed protocol.

- Understanding the basis of the processes you are analyzing is essential, as well as working with an interdisciplinary team;

- Data are not just numbers, they reflect real processes and care must be taken! Beware of overinterpreting results.

Data Science and Ecology do go hand in hand!


Next up: Tromsø, Norway


Obrigado!