



UNIVERSITY OF  
GOTHENBURG

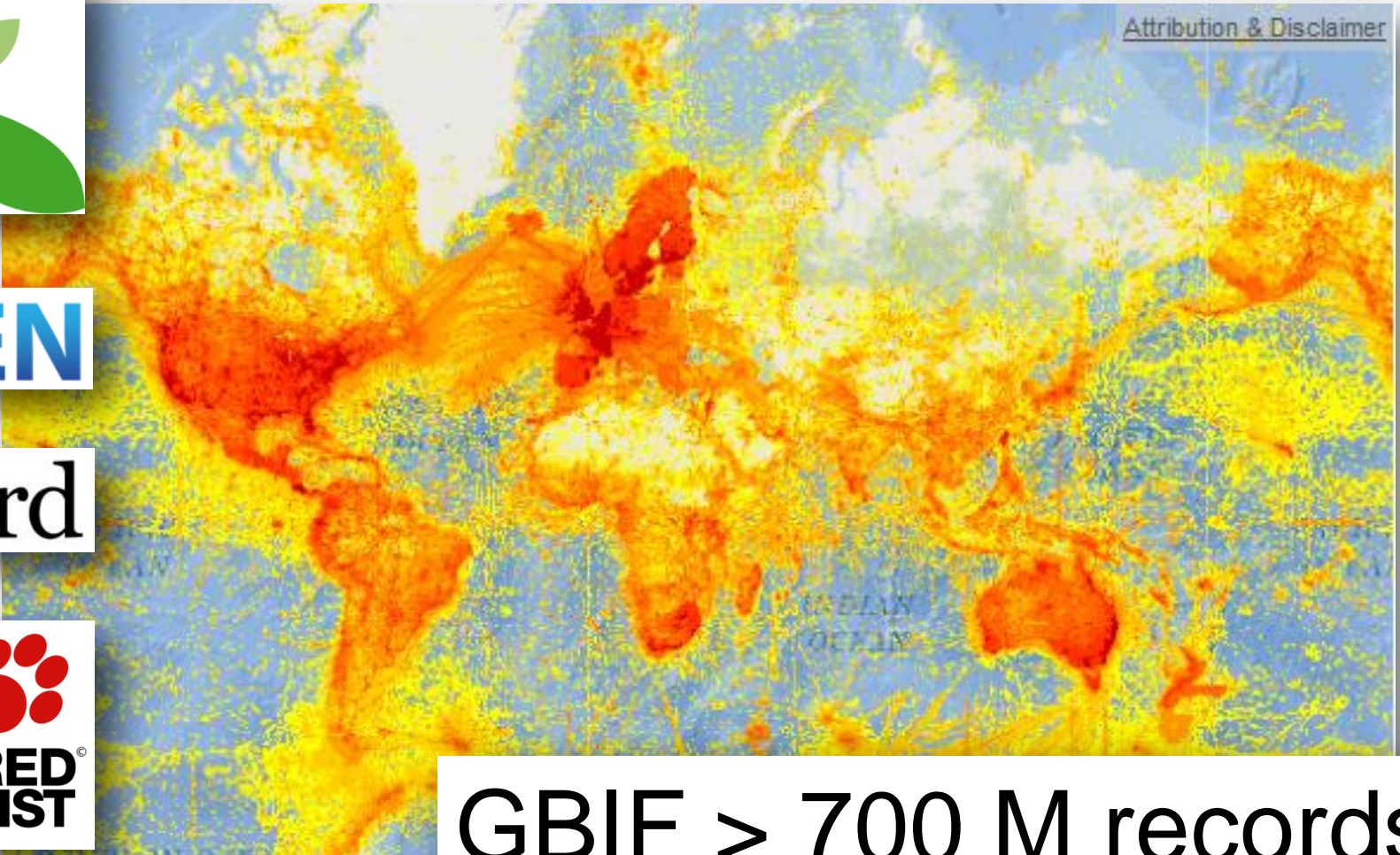
The Antonelli Lab

Evolution and biogeography with focus on the Neotropics

# Coordinate errors and uneven sampling in Biodiversity data: speciesgeocodeR and sampbias

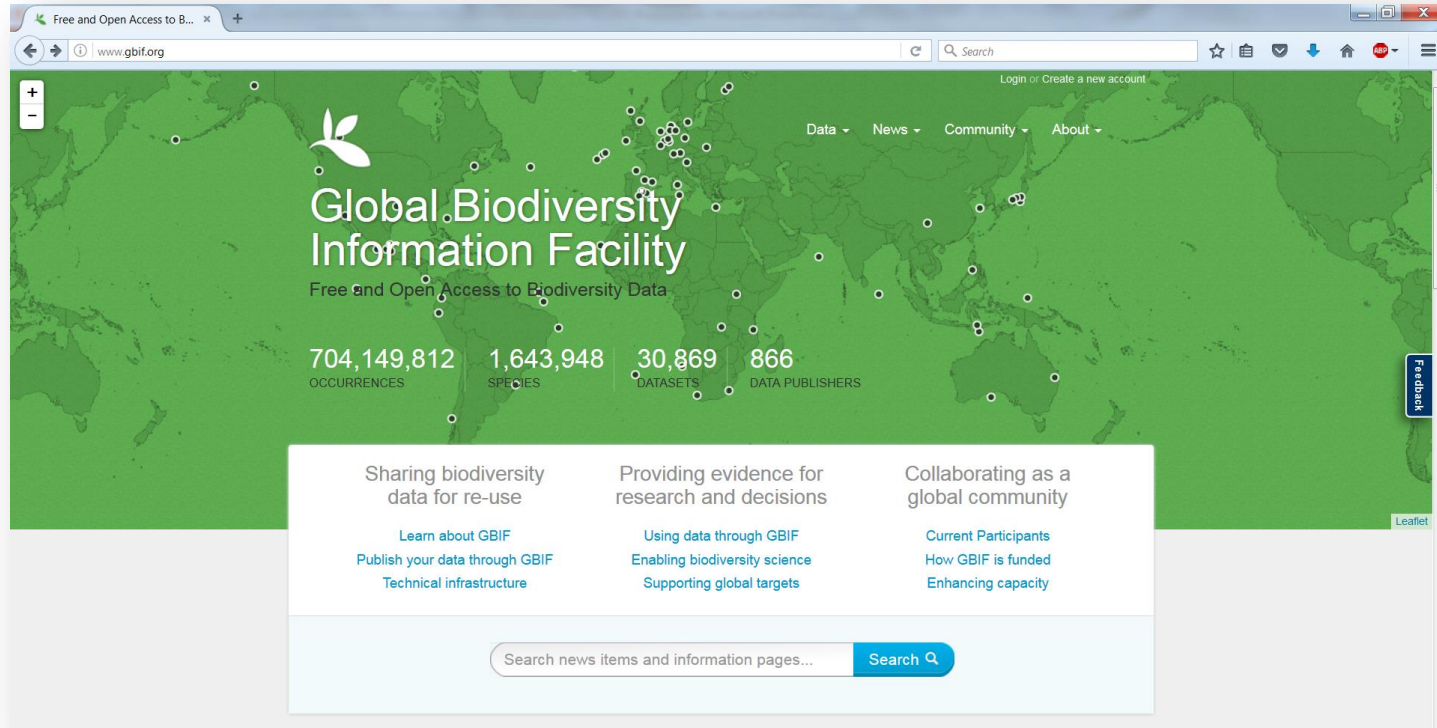
ALEXANDER ZIZKA

# Species distribution data – a revolution



GBIF > 700 M records

# GBIF



The screenshot shows the GBIF website homepage. The background is a green world map with white dots representing biodiversity data points. The GBIF logo, a stylized white leaf, is in the top left. The main heading is 'Global Biodiversity Information Facility' in white, followed by the tagline 'Free and Open Access to Biodiversity Data'. Below this, four statistics are displayed: 704,149,812 OCCURRENCES, 1,643,948 SPECIES, 30,869 DATASETS, and 866 DATA PUBLISHERS. A navigation bar at the top right includes links for Data, News, Community, and About, along with a search bar and a login link. A footer section contains three columns of links related to sharing data, using data for research, and collaborating. At the bottom, there is a search bar for news items and information pages.

Free and Open Access to Biodiversity Data

Global Biodiversity Information Facility

704,149,812 OCCURRENCES 1,643,948 SPECIES 30,869 DATASETS 866 DATA PUBLISHERS

Sharing biodiversity data for re-use  
[Learn about GBIF](#)  
[Publish your data through GBIF](#)  
[Technical infrastructure](#)

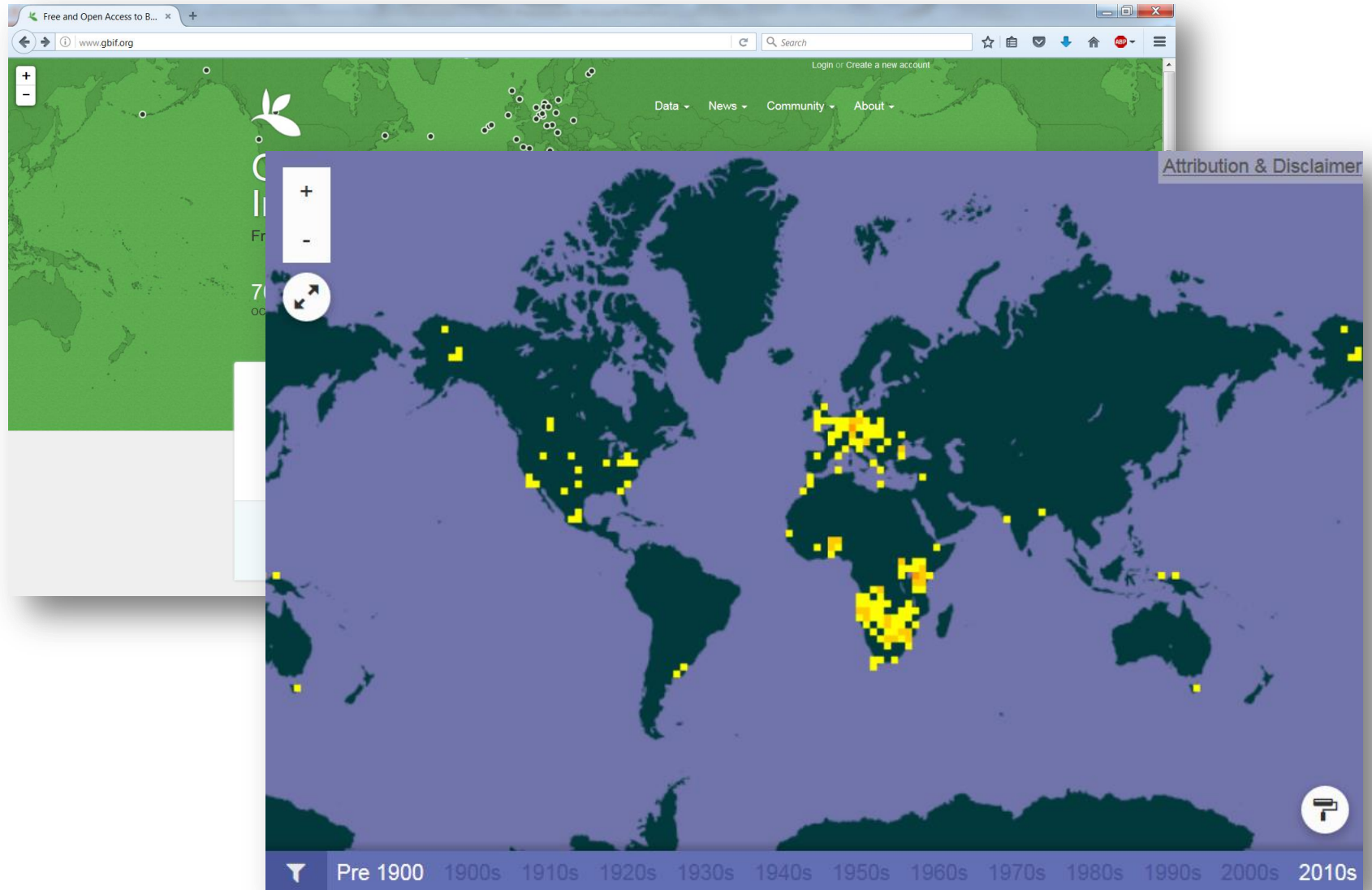
Providing evidence for research and decisions  
[Using data through GBIF](#)  
[Enabling biodiversity science](#)  
[Supporting global targets](#)

Collaborating as a global community  
[Current Participants](#)  
[How GBIF is funded](#)  
[Enhancing capacity](#)

Search news items and information pages... Search



# GBIF



# Fantastic!

---

# Fantastic!

---

But two caveats:

1. Data quality
2. Sampling bias

# Fantastic!

---

But two caveats:

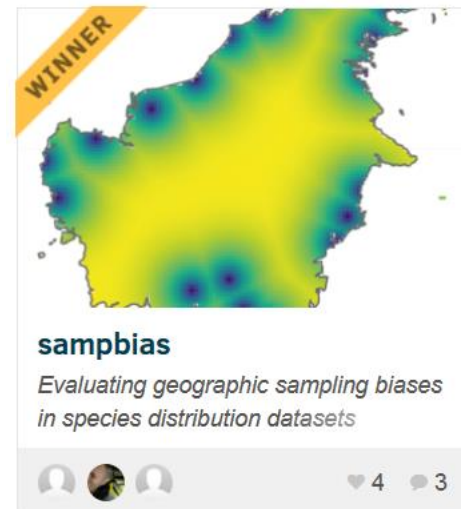
1. Data quality
2. Sampling bias



# Fantastic!

But two caveats:

1. Data quality
2. Sampling bias





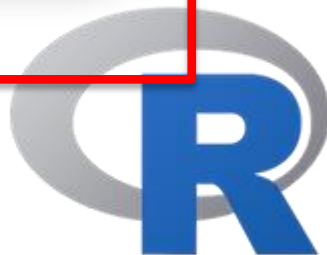
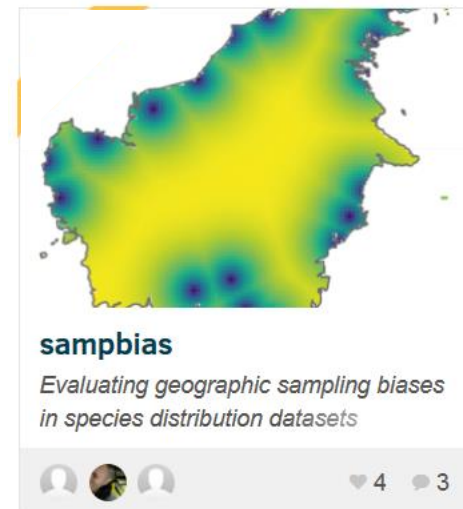
# Fantastic!

## But two caveats:

### 1. Data quality



### 2. Sampling bias



# 1. Data quality



# 1. Data quality



# Common coordinate errors



# Common coordinate errors



# Common coordinate errors



- Invalid or missing





# Common coordinate errors



- Invalid or missing
- Zeros



# Common coordinate errors



- Invalid or missing
- Zeros
- Seas



# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids



# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids
- Capitals



# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids
- Capitals
- Institutions





# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids
- Capitals
- Institutions
- Wrong country





# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids
- Capitals
- Institutions
- Wrong country
- Outliers



# Common coordinate errors



- Invalid or missing
- Zeros
- Seas
- Country centroids
- Capitals
- Institutions
- Wrong country
- Outliers
- Urban areas



# Example from the coffee family (Rubiaceae)



(a) ▲ Verified Dataset



(b) ▲ GBIF Dataset

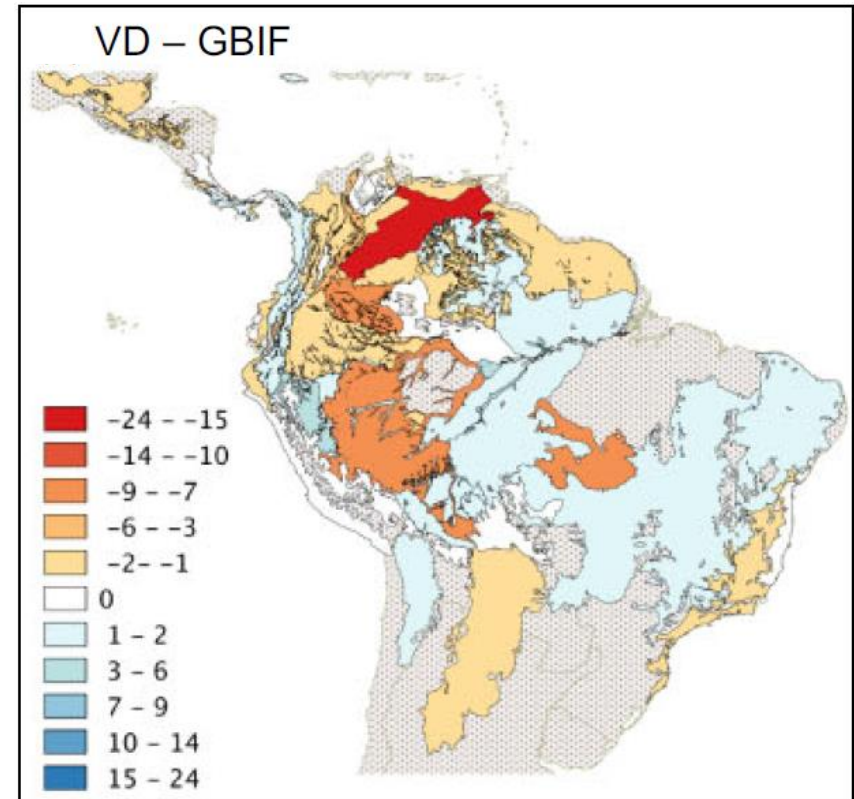
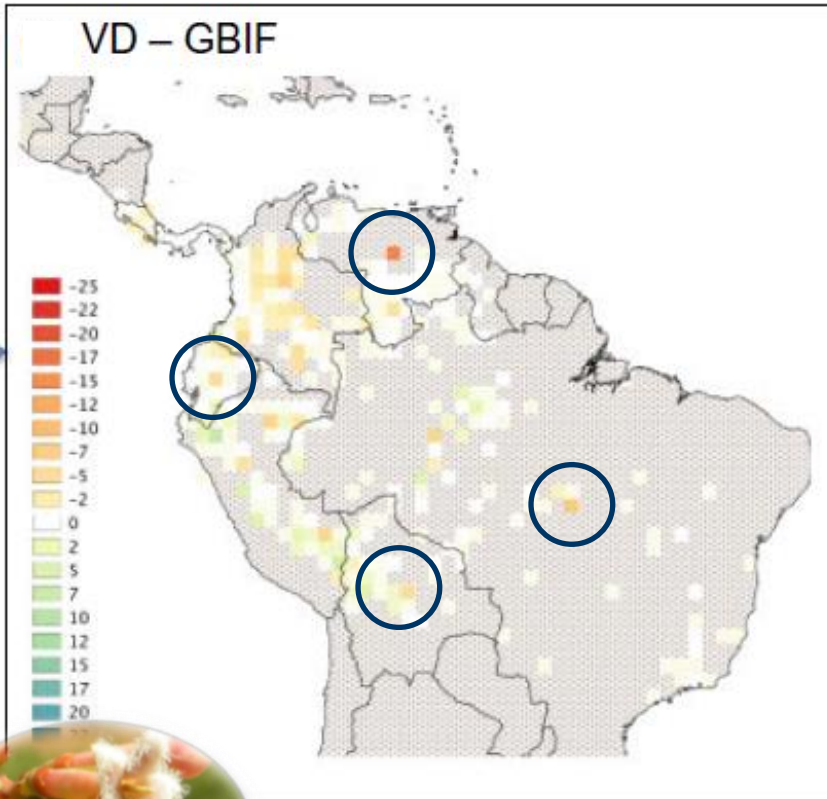




# Example from the coffee family (Rubiaceae)



~10,000 records & 121 species



# What to do?



- Filtering
- Expert opinion
- Automated cleaning

# speciesgeocodeR



- Automatically flag suspicious records + easy visualization
- Biodiversity institutions (6,000 soon 11,000)

Töpel, Zizka et al. 2016,  
<https://github.com/azizka/speciesgeocodeR>



# Workflow



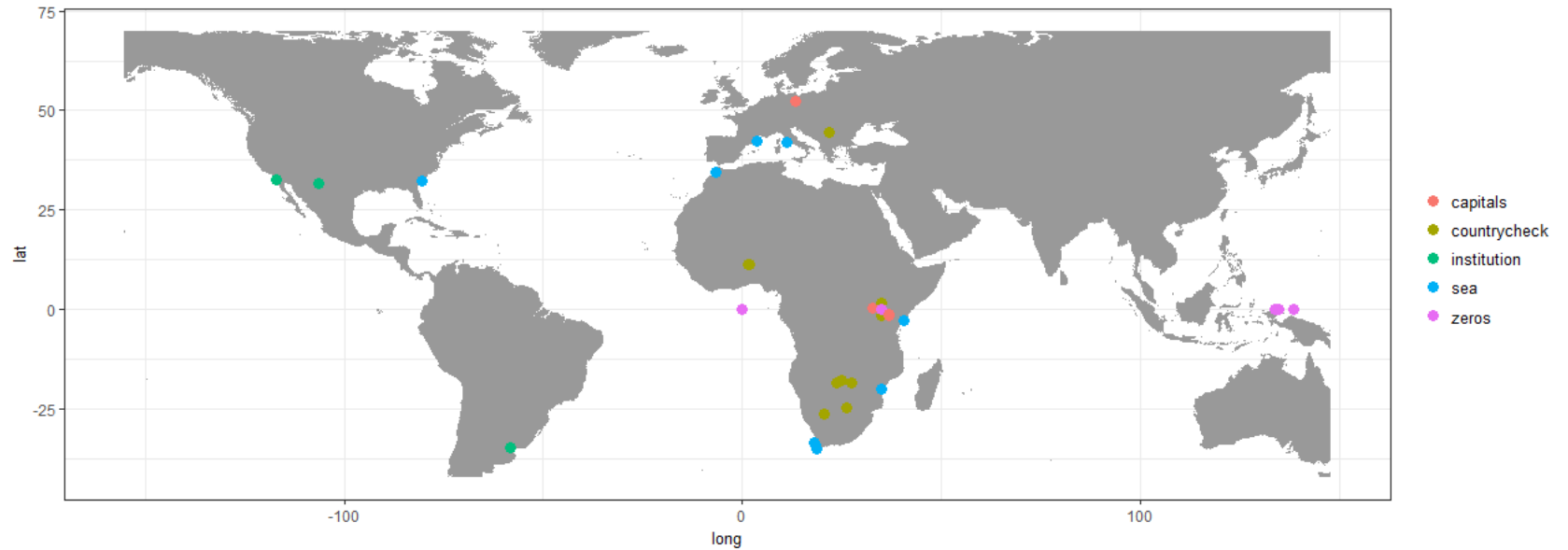
```
dat <-read_csv("input/0053170-160910150852091.csv")
dat %>%
  dplyr::select(decimallongitude, decimallatitude)%>%
  CleanCoordinates(countries = unlist(dat[, "countrycode"]), institutions = T, inst.rad = 0.01) %>%
  plot(clean = F)
```

```
## running validity test
## running zero coordinate test
## flagged 27 records
## running capitals test
## flagged 5 records
## running centroids test
## flagged 22 records
## running seas test
## flagged 14 records
## running countrycheck test
## flagged 16 records
## running GBIF test
## flagged 0 records
## running institutions test
## flagged 9 records
## flagged 67 of 1332 records, EQ = 0.05
```

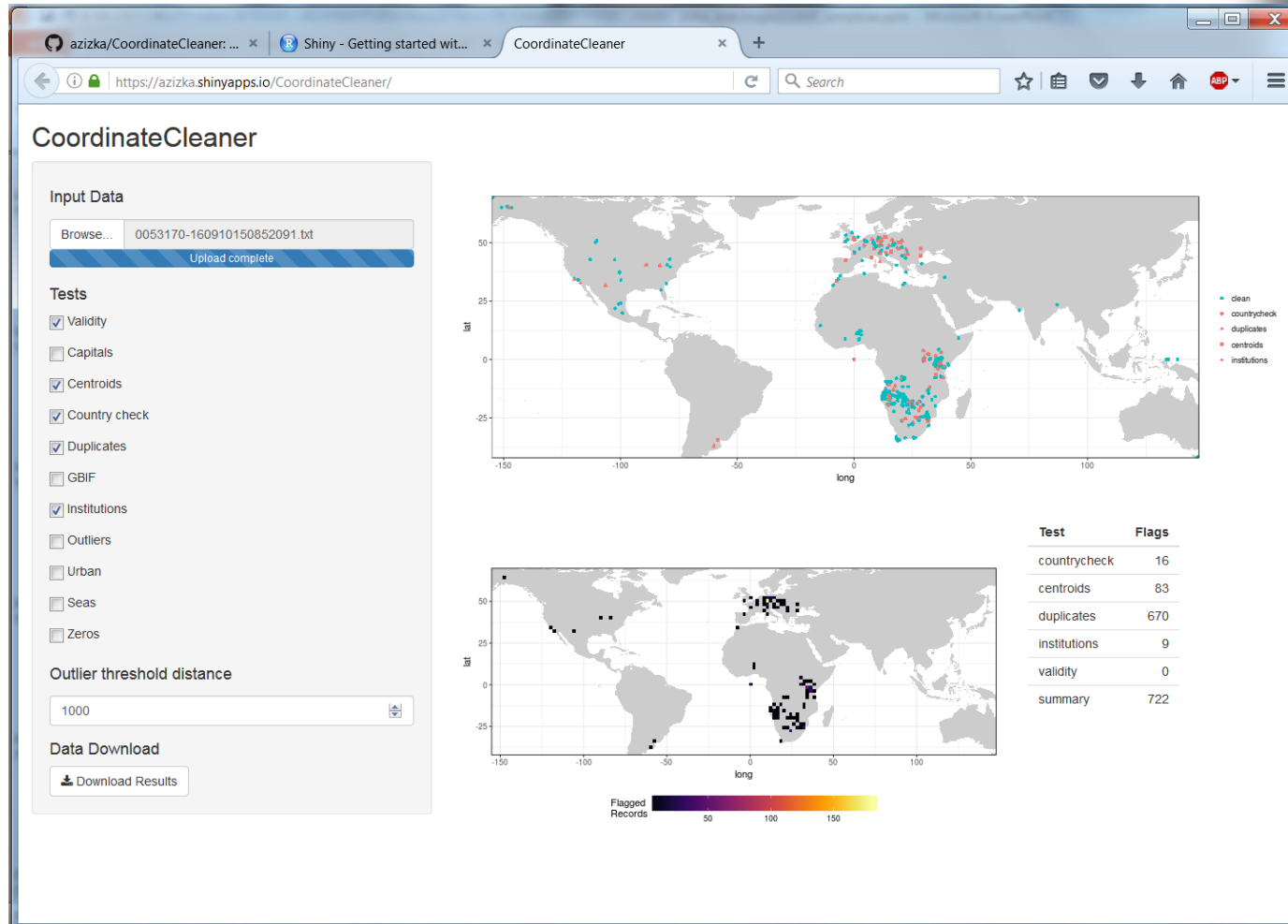
##	decimallongitude	decimallatitude	validity	zeros	capitals	centroids	sea
## 1	1.46994	11.45179	TRUE	TRUE	TRUE	TRUE	TRUE
## 2	1.47035	11.46455	TRUE	TRUE	TRUE	TRUE	TRUE
## 3	1.49126	11.44658	TRUE	TRUE	TRUE	TRUE	TRUE
## 4	1.58874	11.40803	TRUE	TRUE	TRUE	TRUE	TRUE
## 5	1.47939	11.44816	TRUE	TRUE	TRUE	TRUE	TRUE
## 6	1.35556	11.24228	TRUE	TRUE	TRUE	TRUE	TRUE
##	countrycheck	gbif	institution	summary			
## 1	TRUE	TRUE	TRUE	TRUE			
## 2	FALSE	TRUE	TRUE	FALSE			
## 3	TRUE	TRUE	TRUE	TRUE			
## 4	FALSE	TRUE	TRUE	FALSE			
## 5	TRUE	TRUE	TRUE	TRUE			
## 6	TRUE	TRUE	TRUE	TRUE			

<https://github.com/azizka/speciesgeocodeR>

# Visualization



# Shiny app



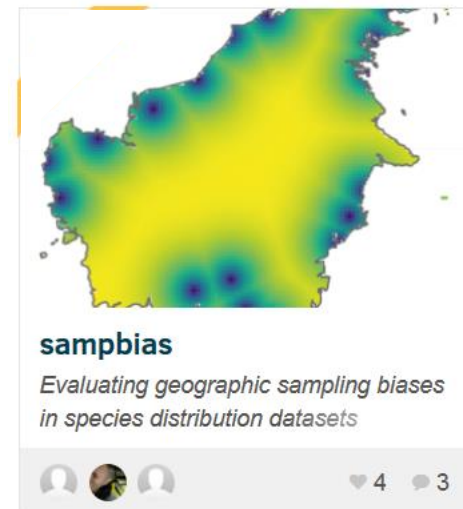
<https://azizka.shinyapps.io/CoordinateCleaner/>

# Fantastic!

But two caveats:

1. Data quality

2. Sampling bias



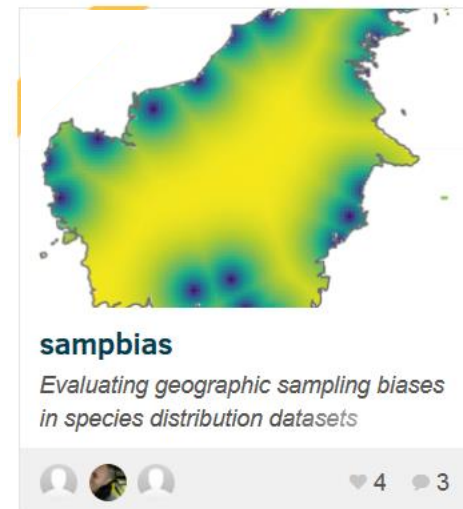
# Fantastic!

## But two caveats:

1. Data quality



2. Sampling bias



# Geographic features and sampling

SampBias





# Geographic features and sampling

SampBias



# Geographic features and sampling

SampBias

- Roads



# Geographic features and sampling

SampBias

- Roads
- Cities



# Geographic features and sampling

SampBias

- Roads
- Cities
- Airports





# Geographic features and sampling

SampBias

- Roads
- Cities
- Airports
- Rivers

➔ Sampling increases with accessibility



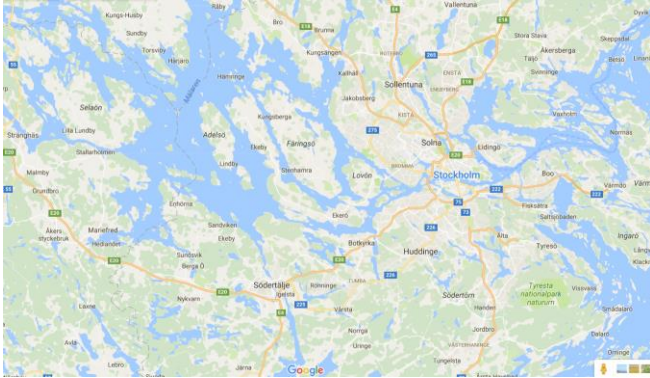
- Relation between anthropogenic geographic features and occurrence records
- Based on gazetteers and frequency distributions of distances



# How does it work?

**SampBias**

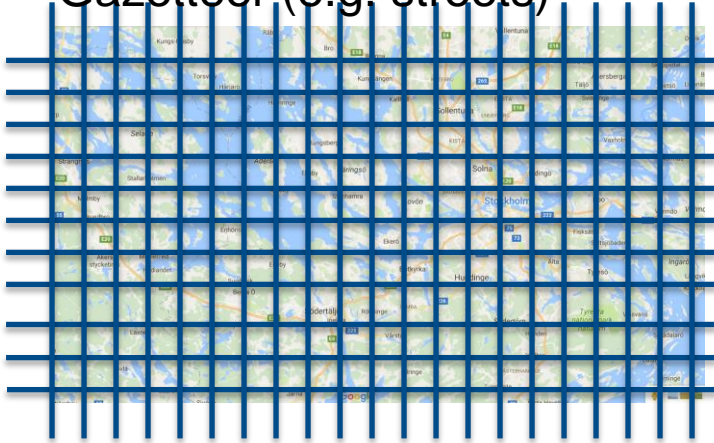
Gazetteer (e.g. streets)



# How does it work?

**SampBias**

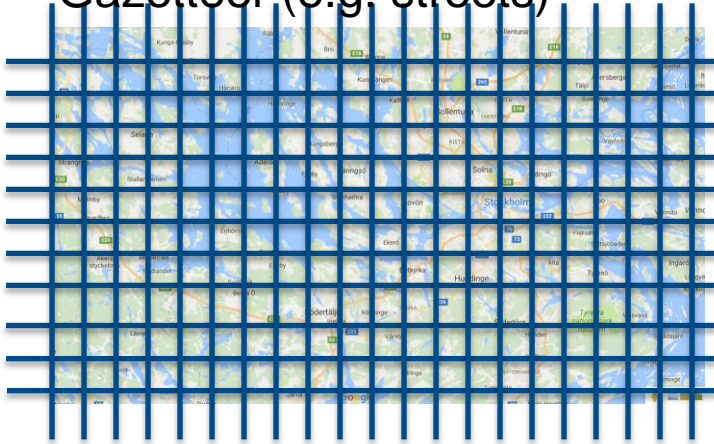
Gazetteer (e.g. streets)



# How does it work?

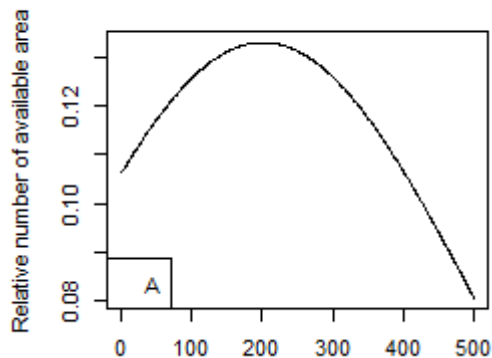
SampBias

Gazetteer (e.g. streets)



Minimum distance

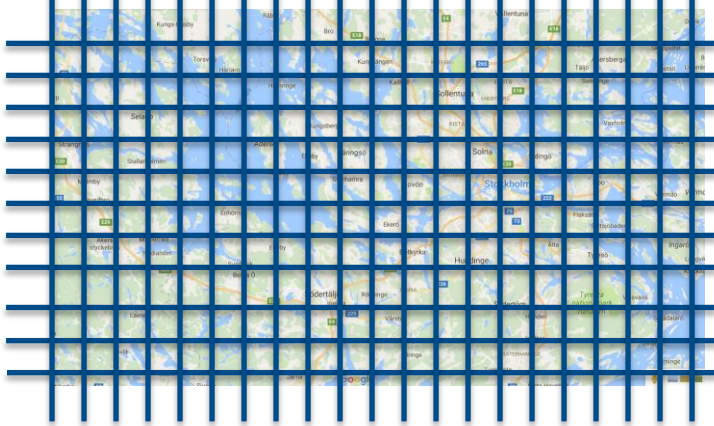
Expected distribution



# How does it work?

SampBias

Gazetteer (e.g. streets)

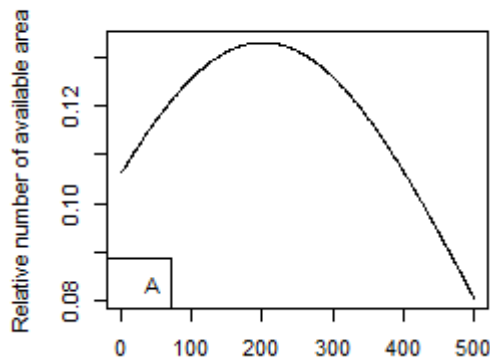


Occurrences



Minimum distance

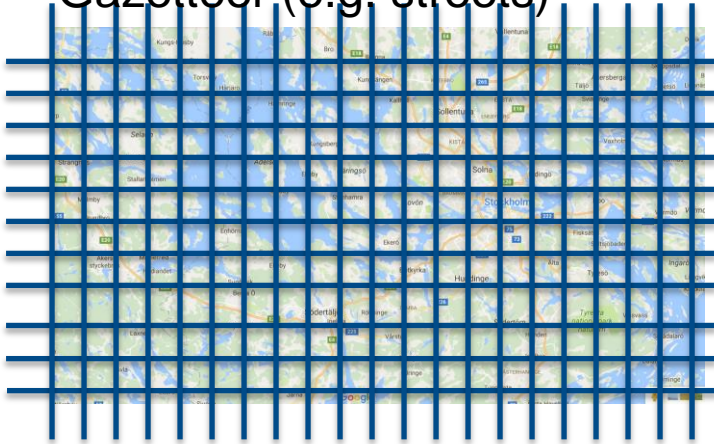
Expected distribution



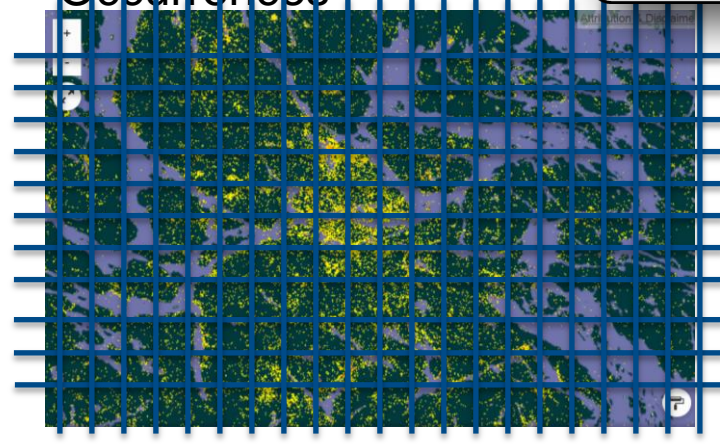
# How does it work?

SampBias

Gazetteer (e.g. streets)

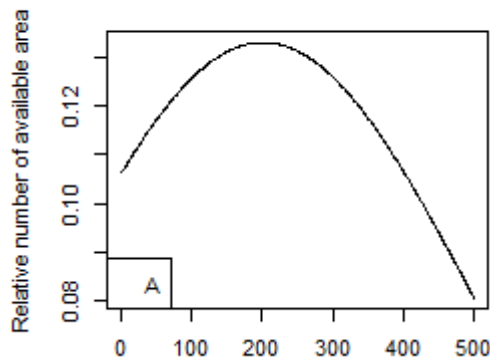


Occurrences

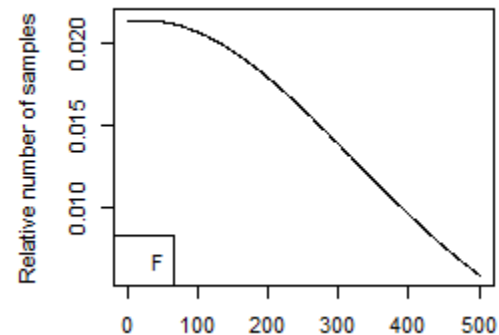


Minimum distance

Expected distribution



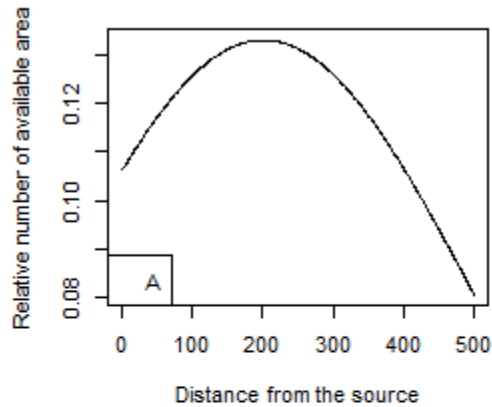
Observed distribution



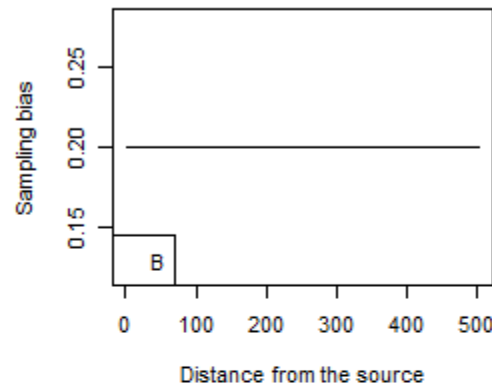
# How does it work?

SampBias

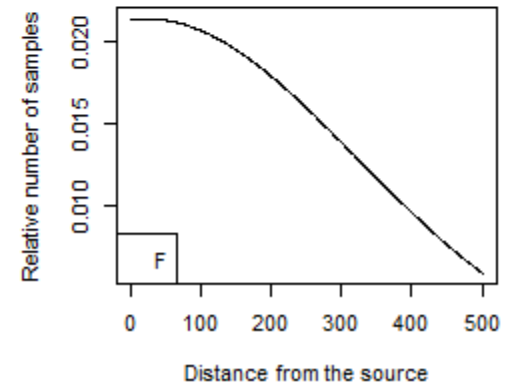
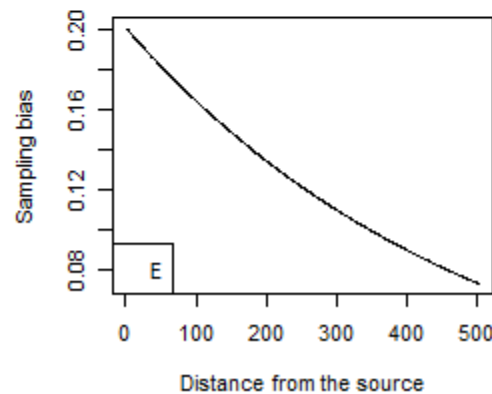
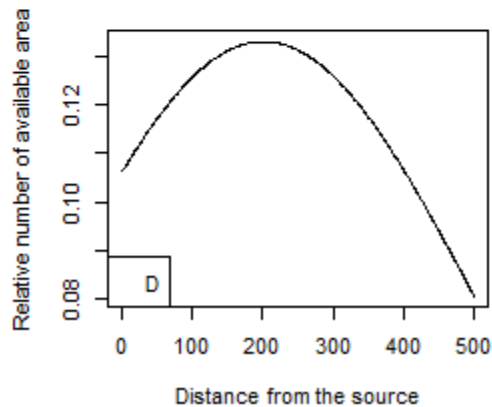
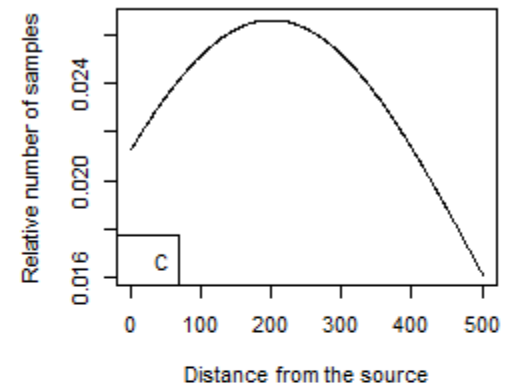
Expected distribution



Effect of distance from bias source



Observed distribution





# Workflow

## SampBias

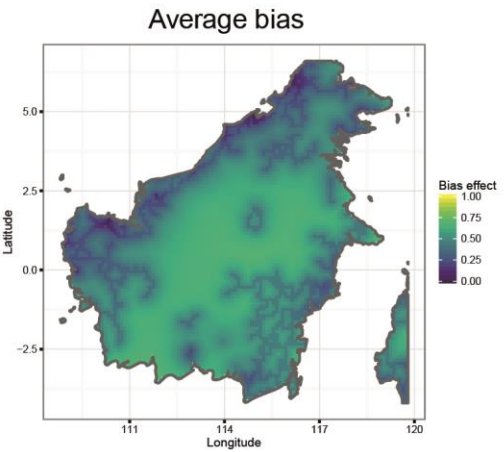
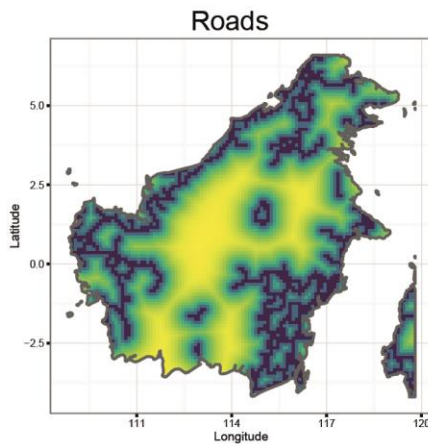
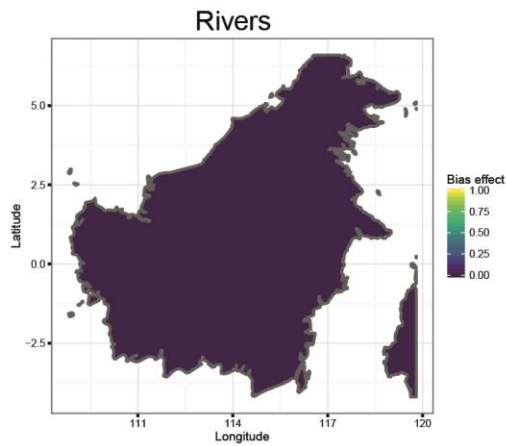
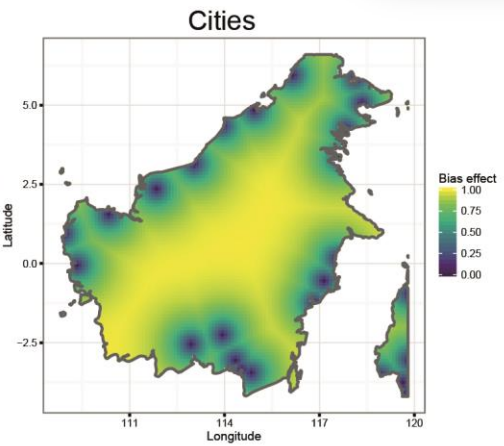
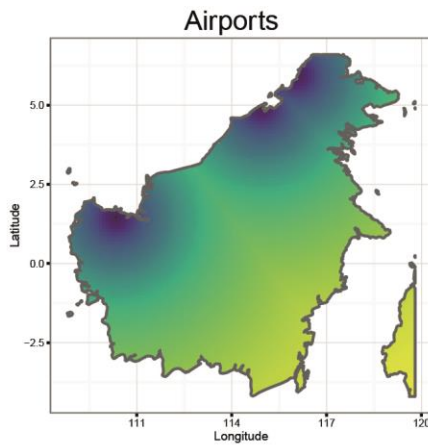
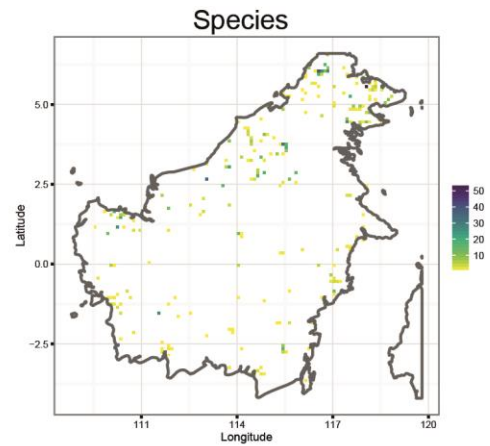
```
occ <- read.csv("input/borneo_mammals.csv", sep = "\t")
bias.out <- SamplingBias(x = occ)

summary(bias.out)
plot(bias.out)
```

```
## Number of occurrences: 5111
## Number of species: 204
## Raster resolution: 1
## Distance binsize: 1e+05
## Convexhull: FALSE
## Geographic extent:
## class      : Extent
## xmin       : 108
## xmax       : 119
## ymin       : -3
## ymax       : 7
## Bias effect at distance:
##           0      10000
## airports 0 8.994056e-05
## cities    0 5.712649e-04
## rivers    0 2.124920e-04
## roads     0 1.332444e-03
```

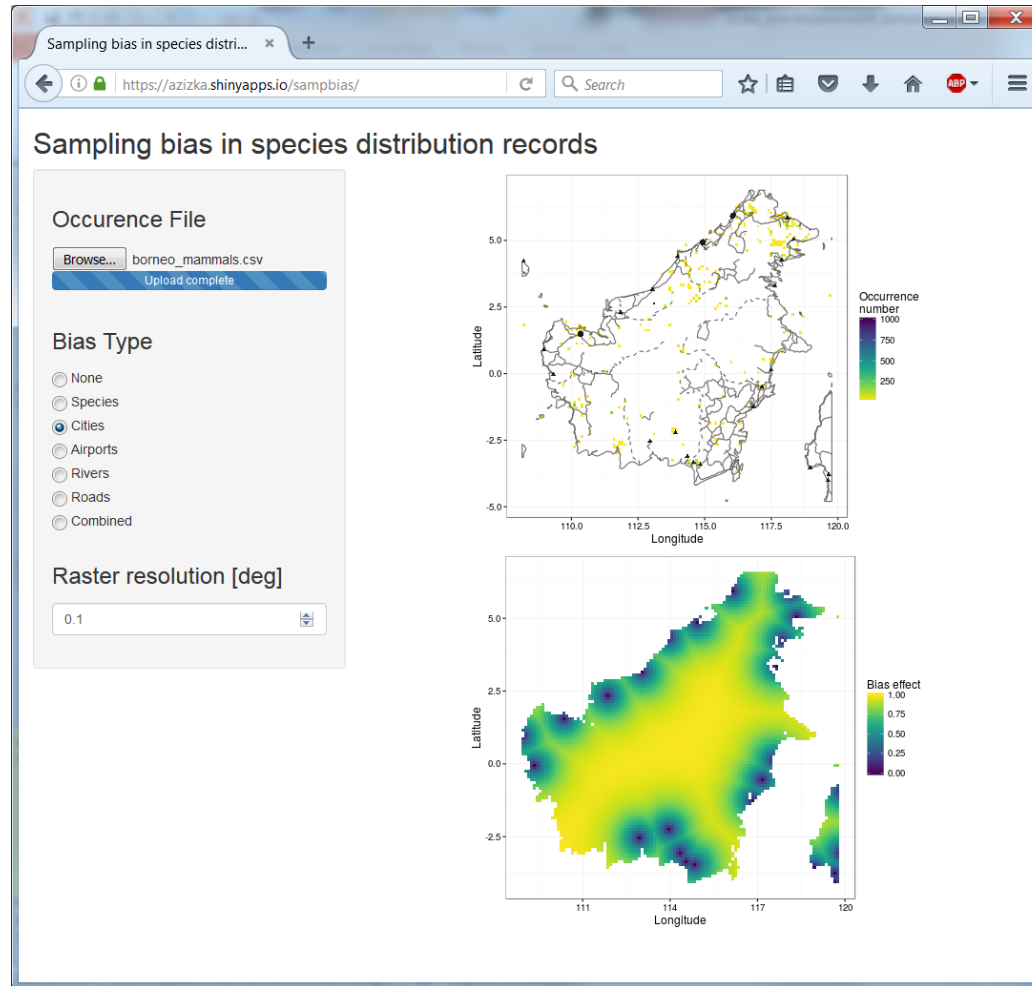
# Visualization

SampBias



# Shiny app

SampBias



<https://azizka.shinyapps.io/sampbias/>

# Summary

---

+

- Dataset specific
- Compare effect of geographical features
- Mathematically sound
- Easy to use

—

- Only exponential bias
- Multi-species datasets
- Dependent on gazetteers
- Computationally expensive

Tack så mycket!

*“Late again! ...  
This better be  
good!”*



# Tack så mycket!



**SampBias**

(BETA)

<http://antonelli-lab.net/resources.php>

<https://cran.rstudio.com/web/packages/speciesgeocodeR/>

<https://github.com/azizka/speciesgeocodeR>

<https://github.com/azizka/sampbias>