Fitting Grade of Membership Models to Binary and Binomial data (*methClust*)

Kushal K. Dey

Report - 3rd Jan, 2018

Let x_{ij} be the binary code (1/0) for presence/absence of species j in site i.

$$x_{ij} \sim Ber(p_{ij})$$

We assume p_{ij} to have a lower dimensional representation

$$p_{ij} = \sum_{k=1}^{K} \omega_{ik} f_{kj}$$

$$\omega_{ik} > 0, \ \sum_{k=1}^{K} \omega_{ik} = 1, \ \ 0 \leq f_{kj} \leq 1$$

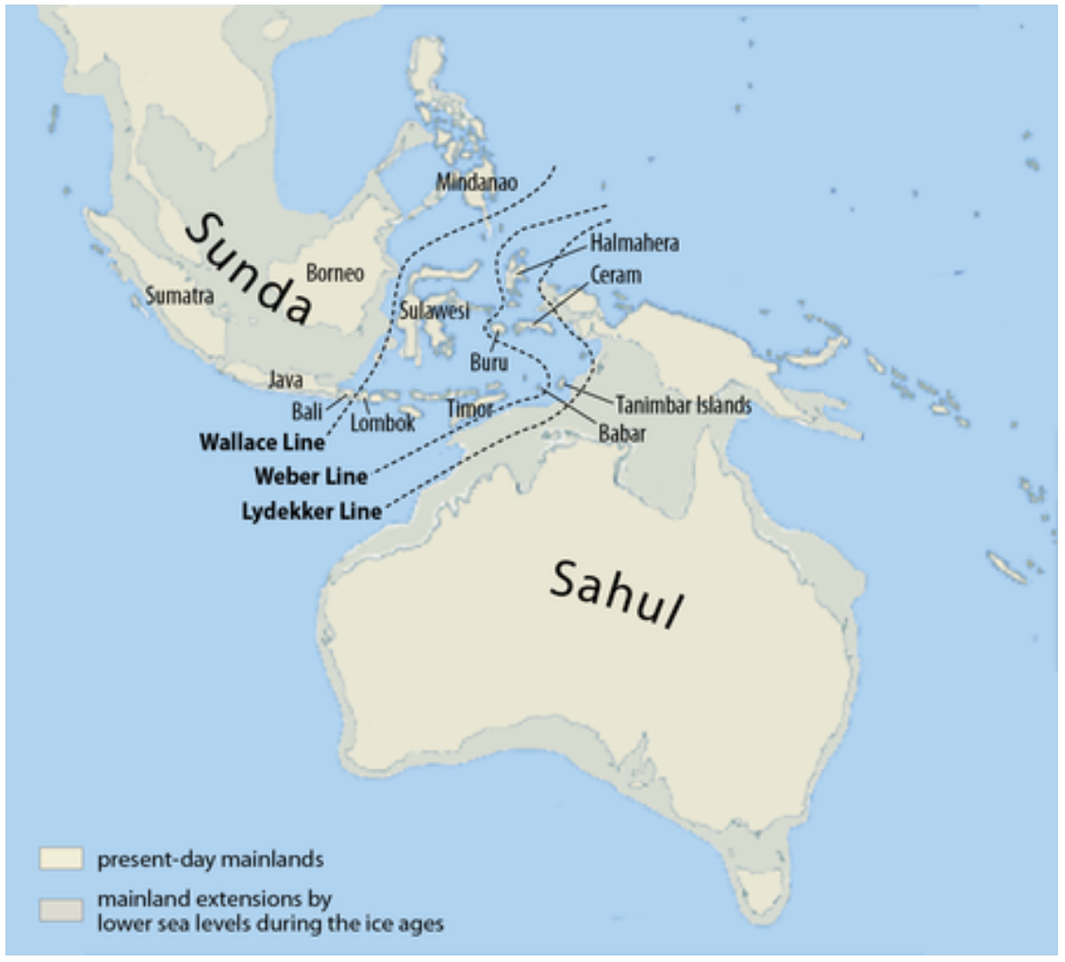
ω_{ik} : grade of membership of cluster k in sample i

f_{kj} : probability of presence of species j in cluster k

Application of Binary GoM model on ecological examples

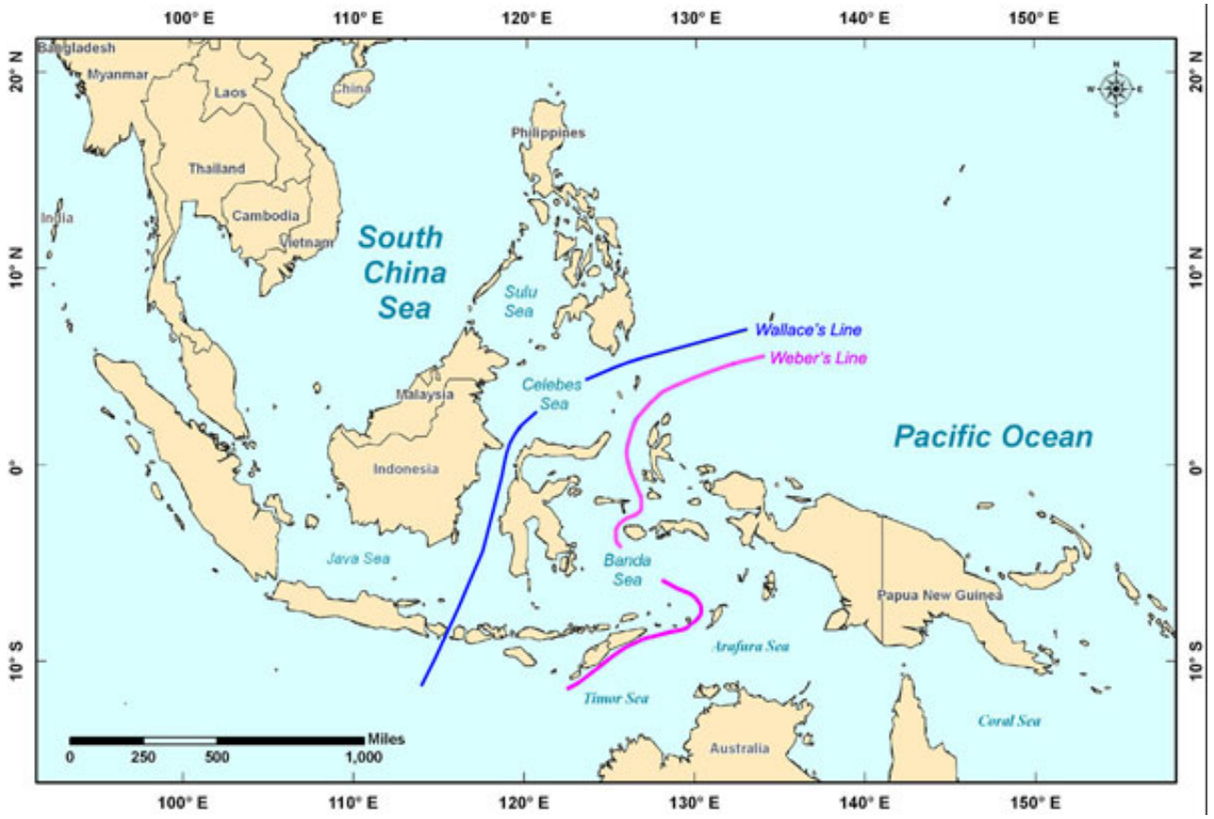
with Alex White and Trevor Price

The historical significance of the Wallace Line

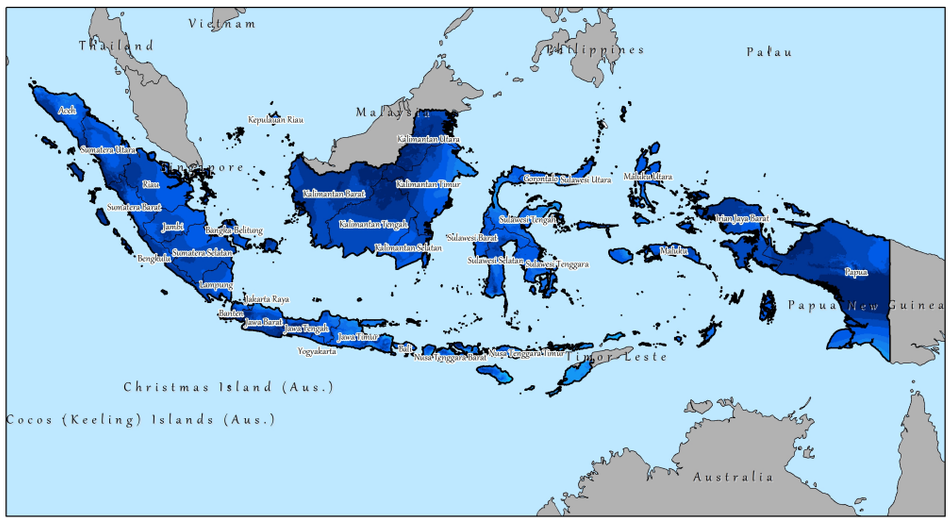


Indonesia map of Köppen climate classification

Equatorial climate (Af) Monsoon climate (Am) Tropical savanna climate (Aw)



Rainfall



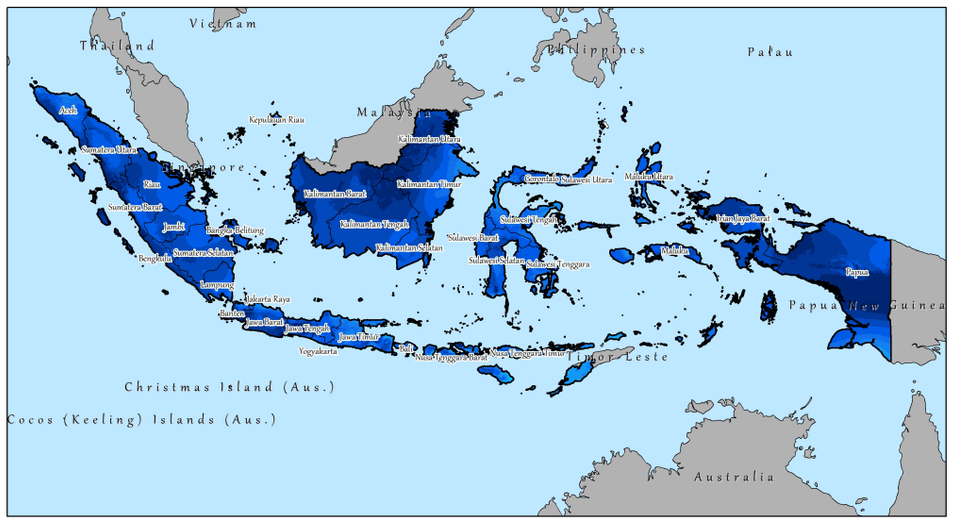
Precipitaiton (mm/year)

0 - 250	1,001 - 1,250	2,001 - 2,500
251 - 500	1,251 - 1,500	2,501 - 3,000
501 - 750	1,501 - 1,750	3,001 - 3,500
751 - 1,000	1,751 - 2,000	> 3500

Indonesia map of Köppen climate classification

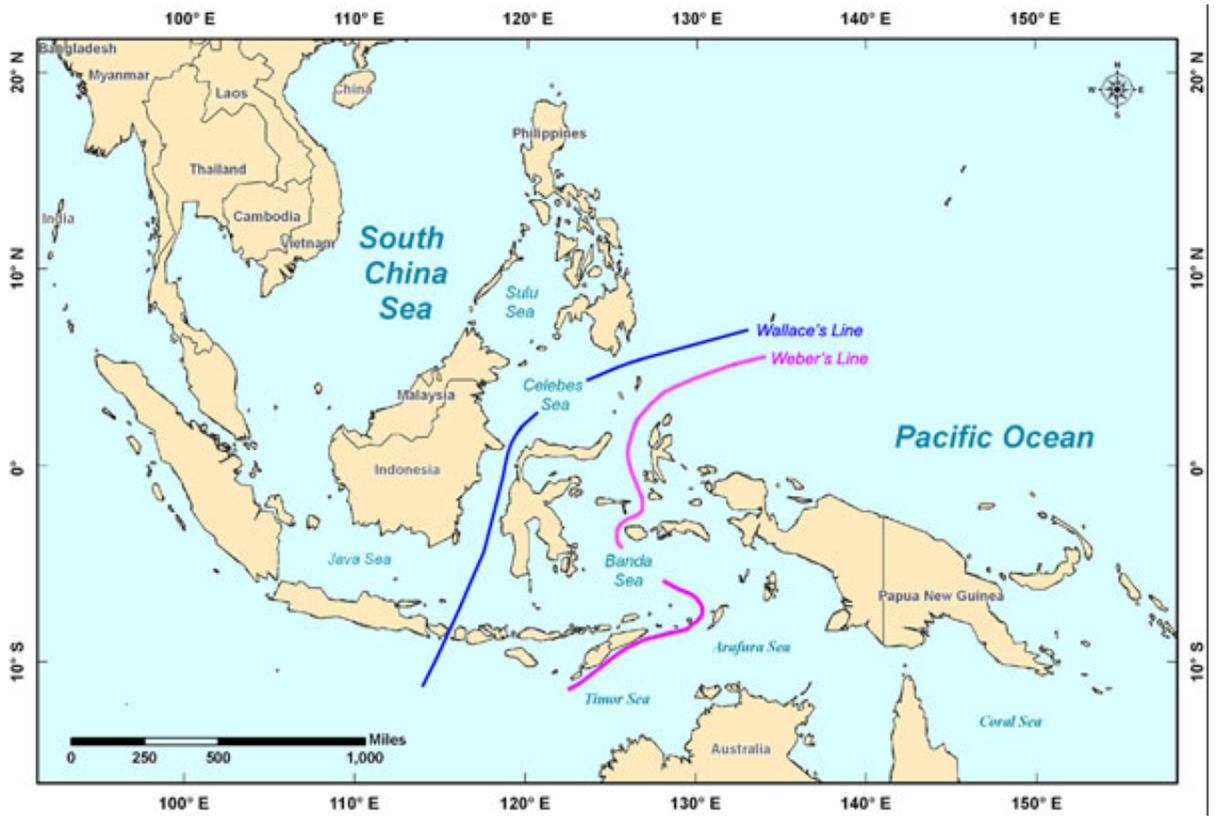
Equatorial climate (Af) Monsoon climate (Am) Tropical savanna climate (Aw)

Rainfall

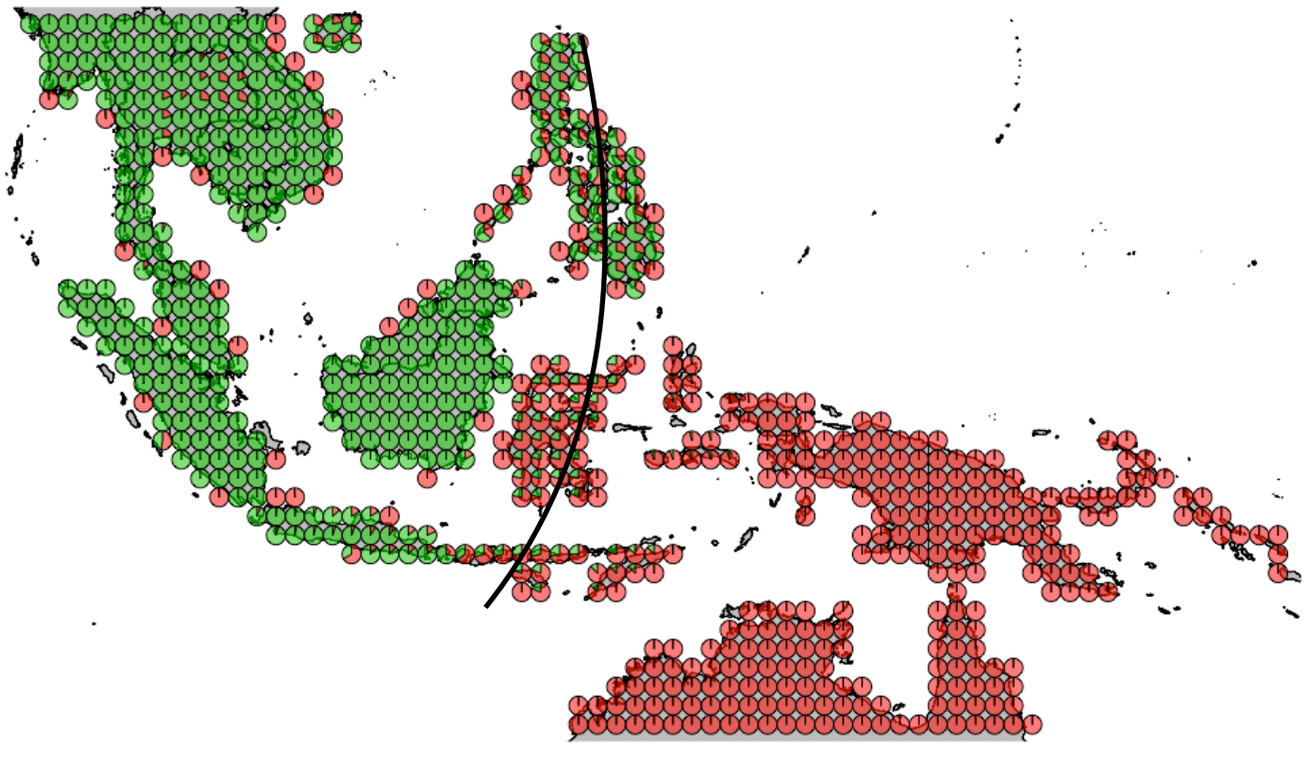


Precipitaiton (mm/year)

0 - 250	1,001 - 1,250	2,001 - 2,500
251 - 500	1,251 - 1,500	2,501 - 3,000
501 - 750	1,501 - 1,750	3,001 - 3,500
751 - 1,000	1,751 - 2,000	> 3500



our line



Cluster 1

grey teal

pacific koel

Nankeen night heron

Pacific black duck

Willie wagtail

Cluster 2

white breast waterhen

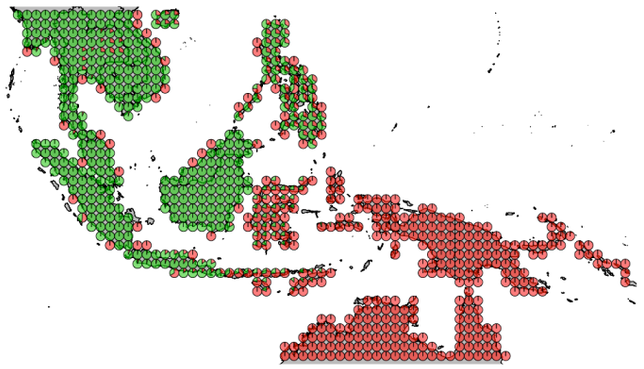
Lesser cuocal

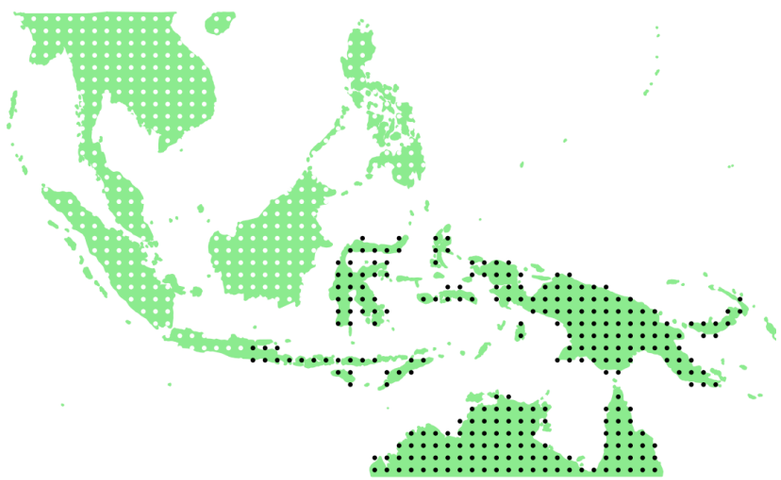
Asian palm swift

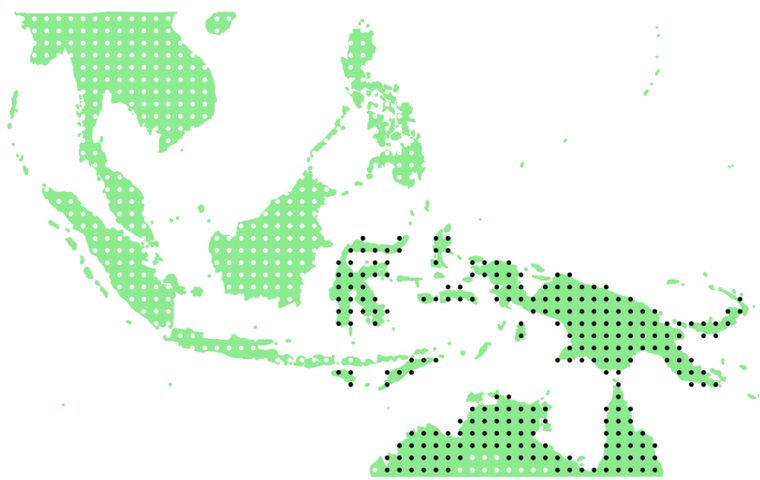
common moorhen

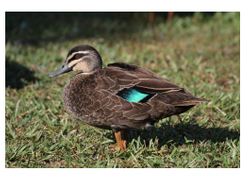
black naked monarch

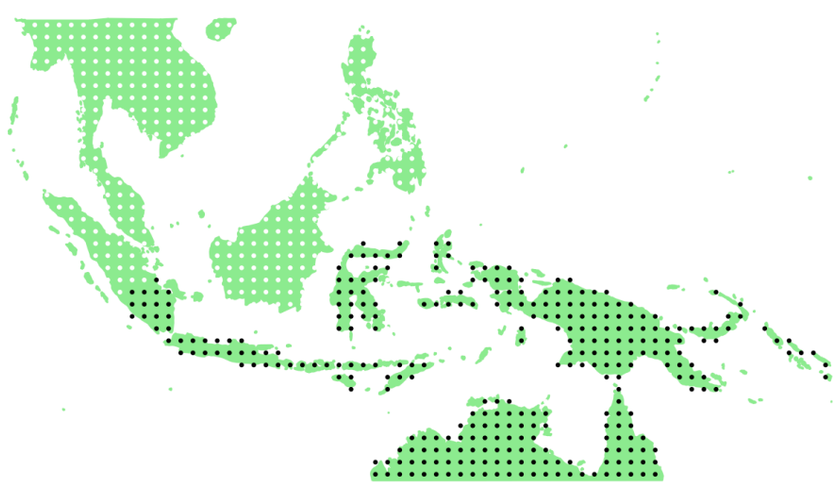
Cluster 1 Top Birds Distribution

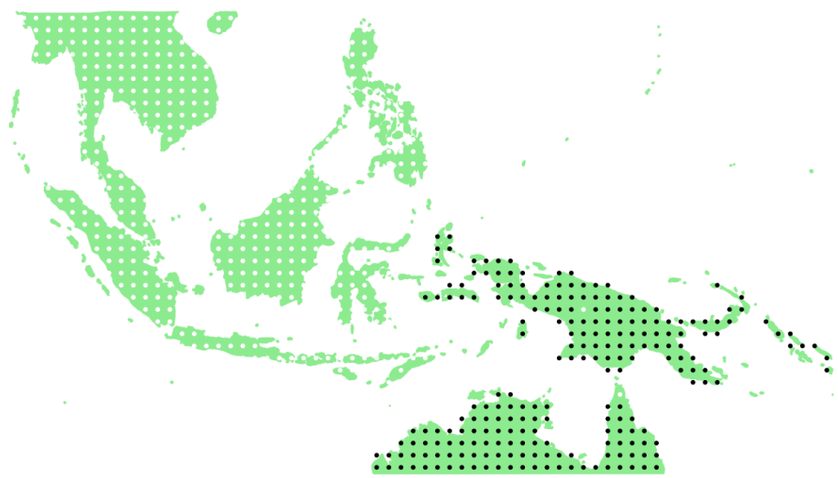




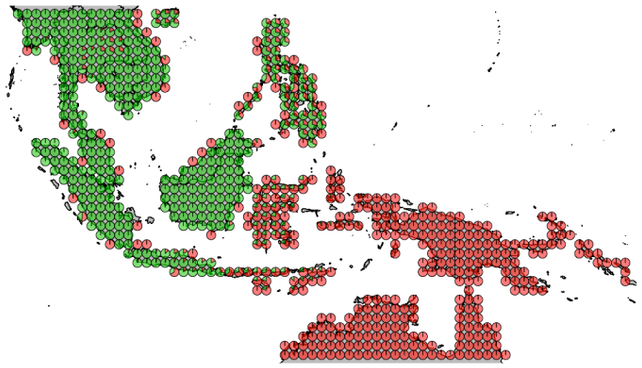


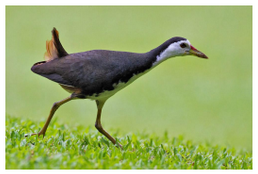


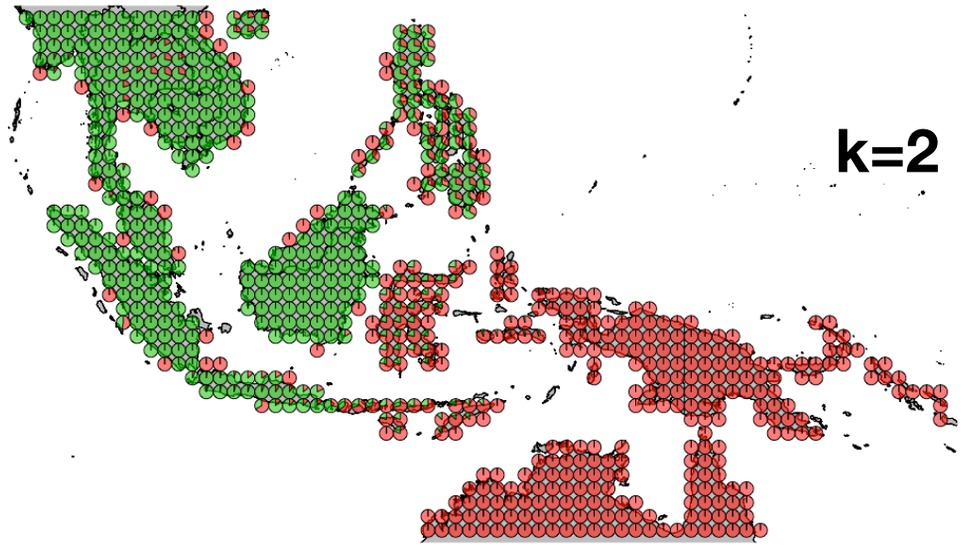


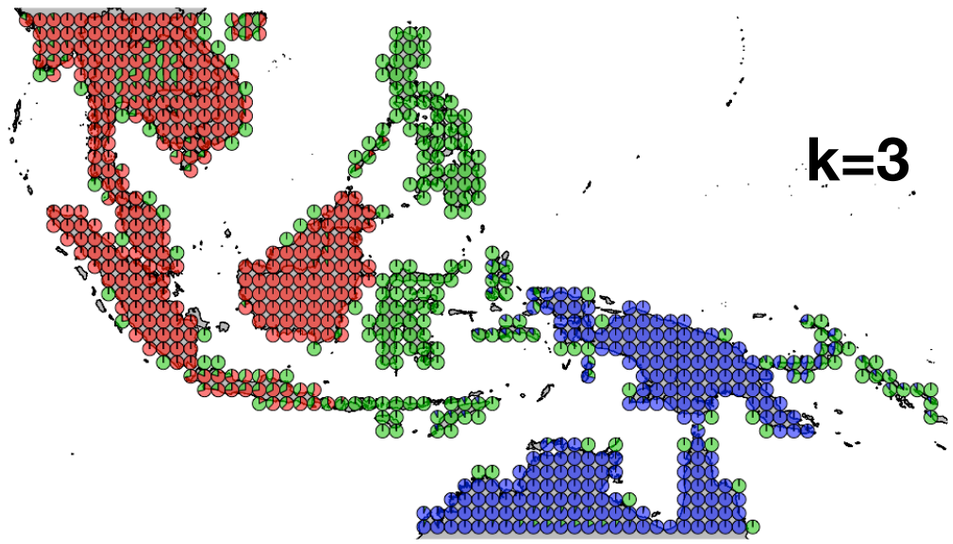


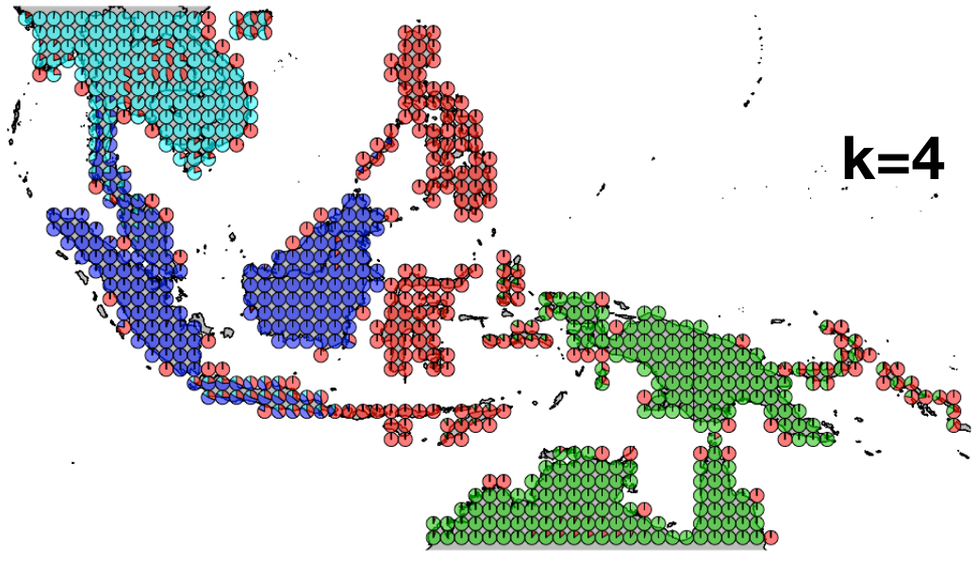
Cluster 2 Top Birds Distribution

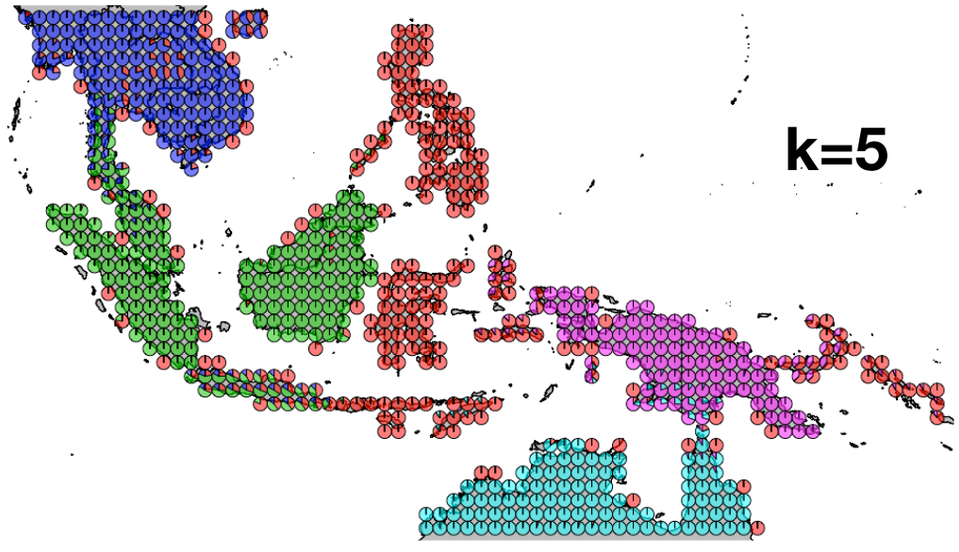


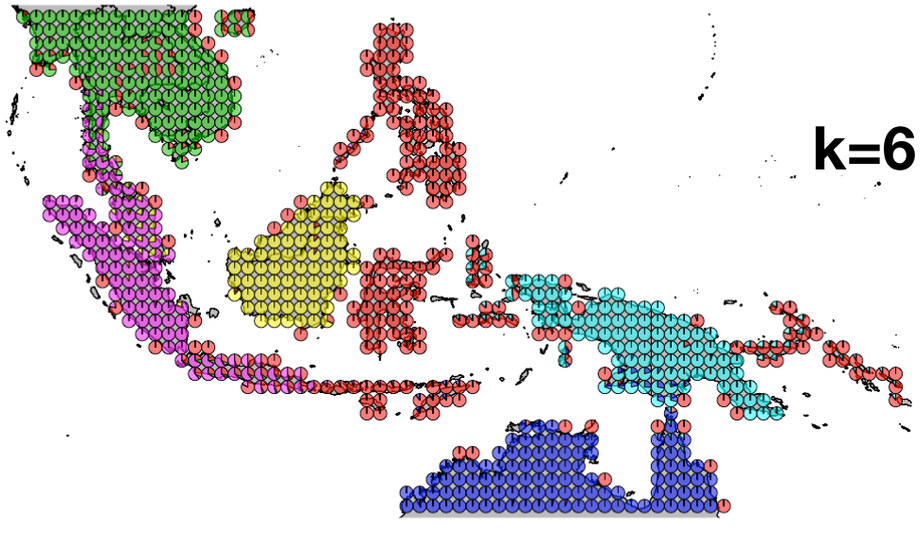


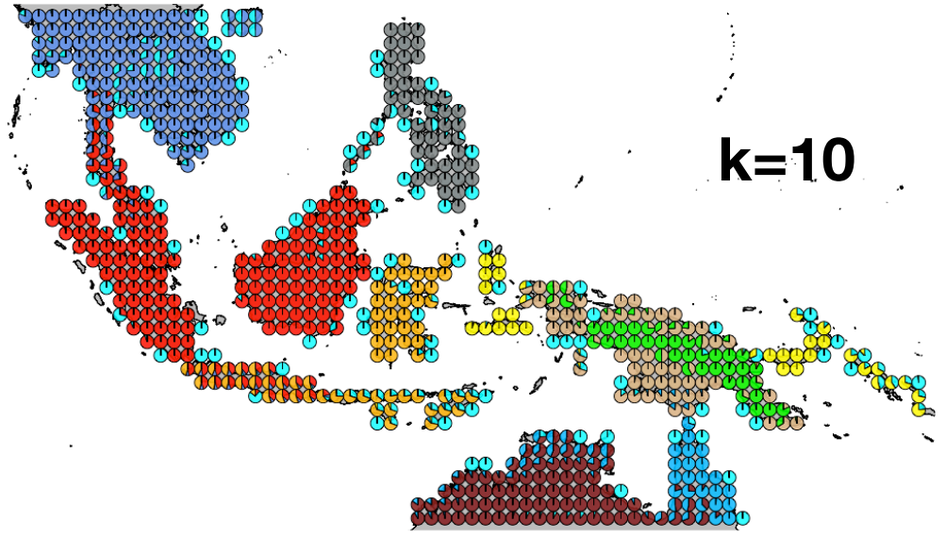

k=2


k=3


k=4

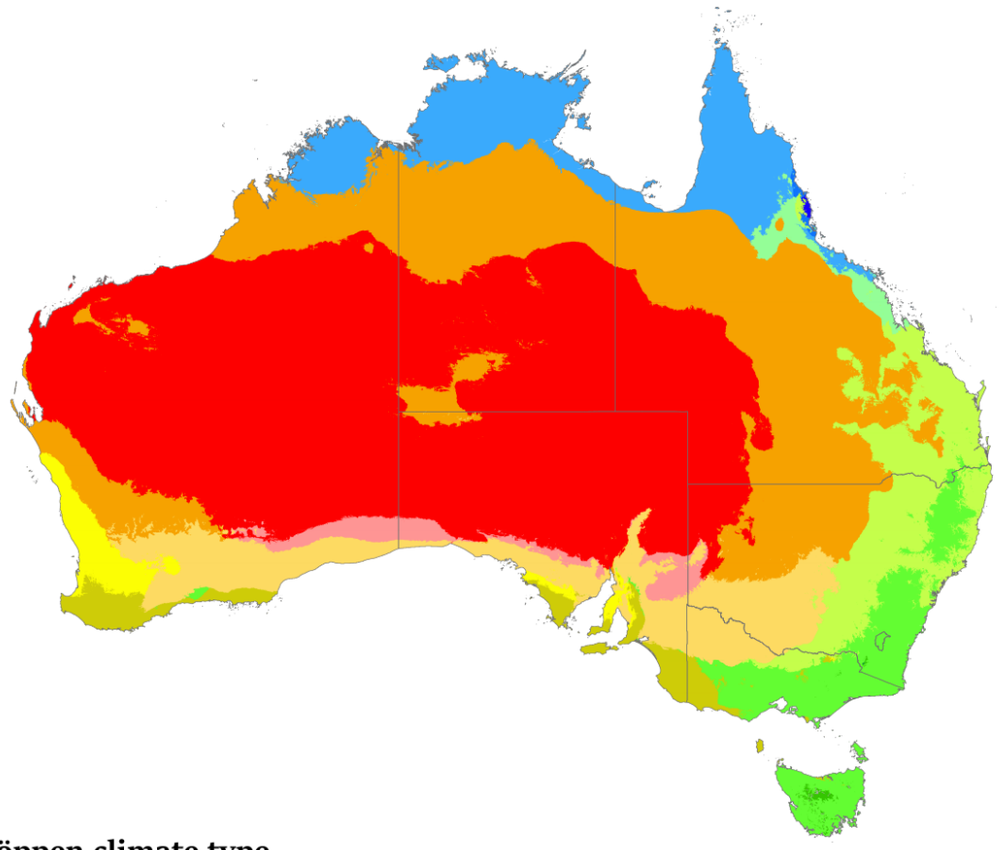

k=5


k=6


k=10

Continental Analysis of Birds presence absence using Binomial GoM models

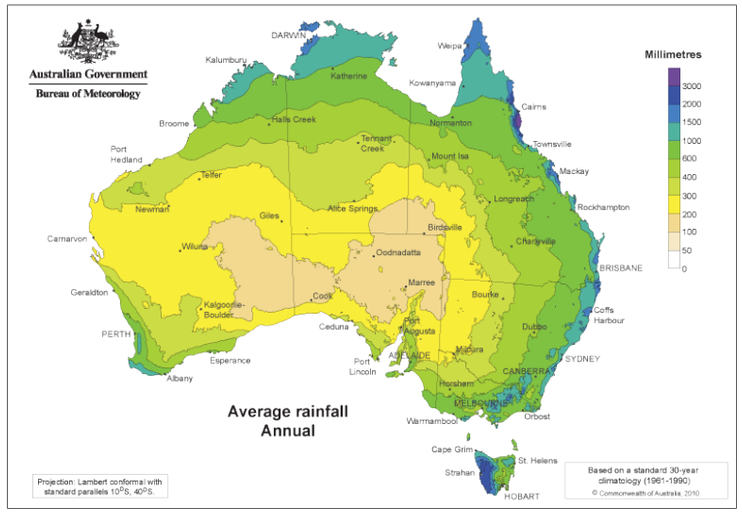
Köppen climate types of Australia

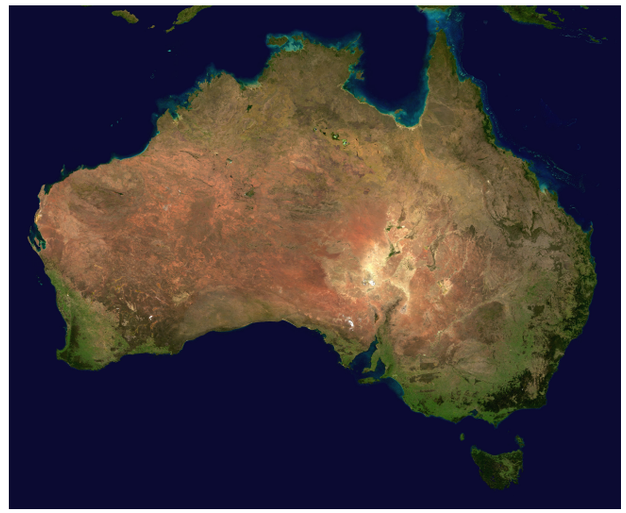


Köppen climate type

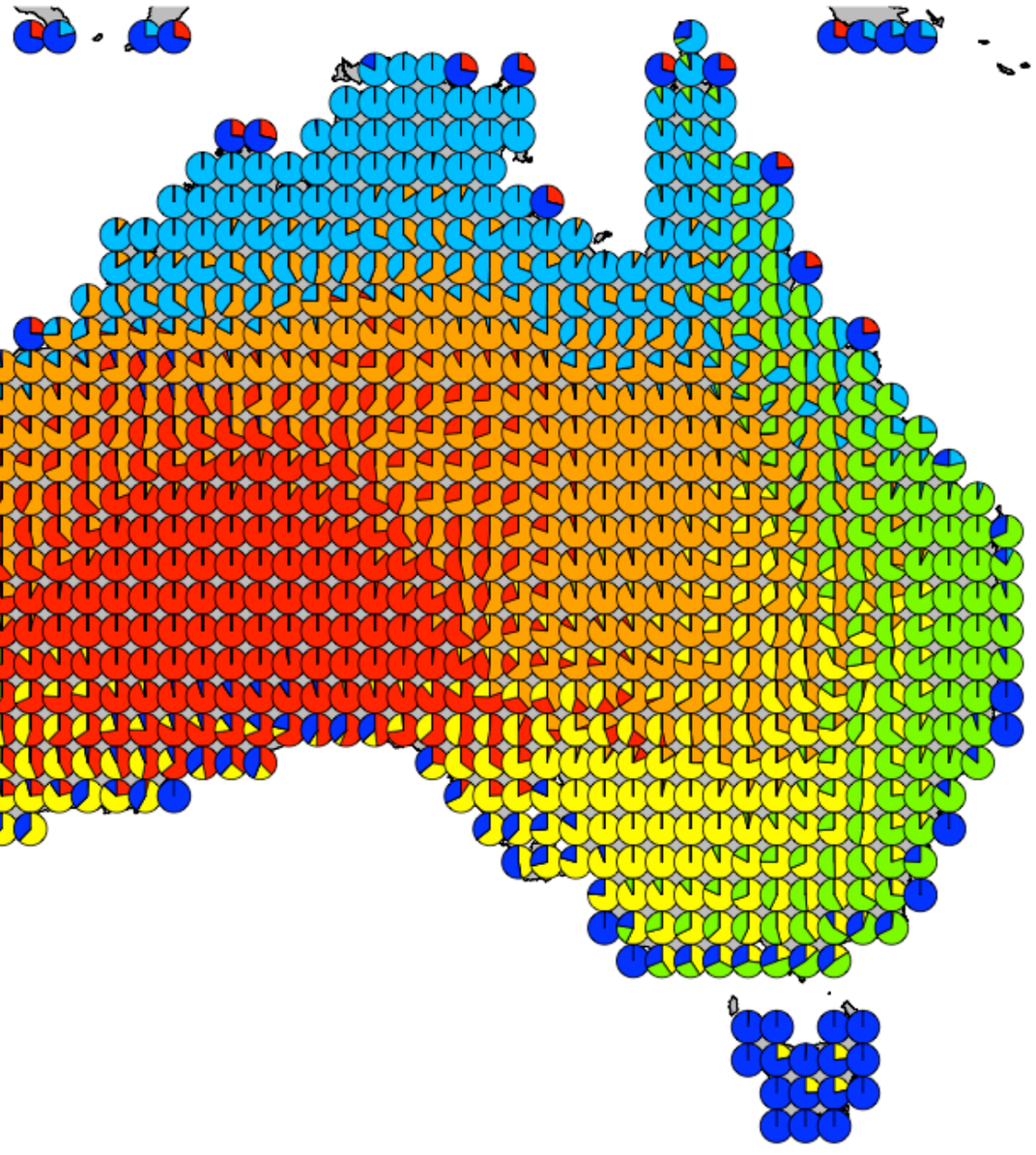
ET (Tundra)	*Cwa (Humid subtropical)*	*BWk (Cold desert)*
Dfc (Subarctic)	*Csb (Warm-summer mediterranean)*	*BWh (Hot desert)*
Cfc (Subpolar oceanic)	*Csa (Hot-summer mediterranean)*	*Aw (Savanna)*
Cfb (Oceanic)	*BSk (Cold semi-arid)*	*Am (Monsoon)*
Cfa (Humid subtropical)	*BSh (Hot semi-arid)*	*Af (Rainforest)*

**Isotherm used to separate temperate (C) and continental (D) climates is -3°C*
Data source: Climate types calculated from data from WorldClim.org





Binomial GoM model (K=6)



Top 5 distinguishing birds for each cluster

cluster 1 (orange)





cluster 2 (red)





cluster 3 (yellow)




cluster 4 (sky blue)





cluster 5 (green)

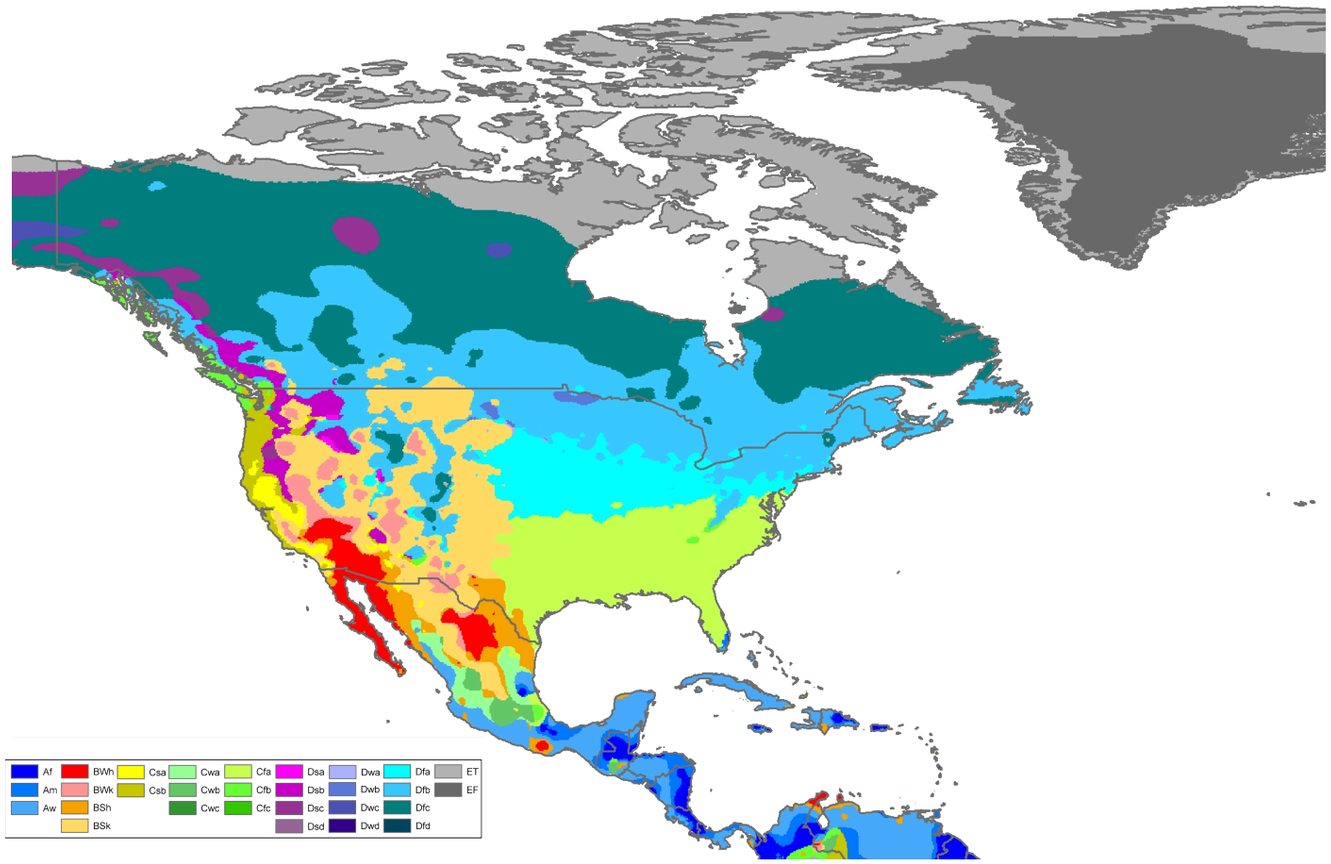


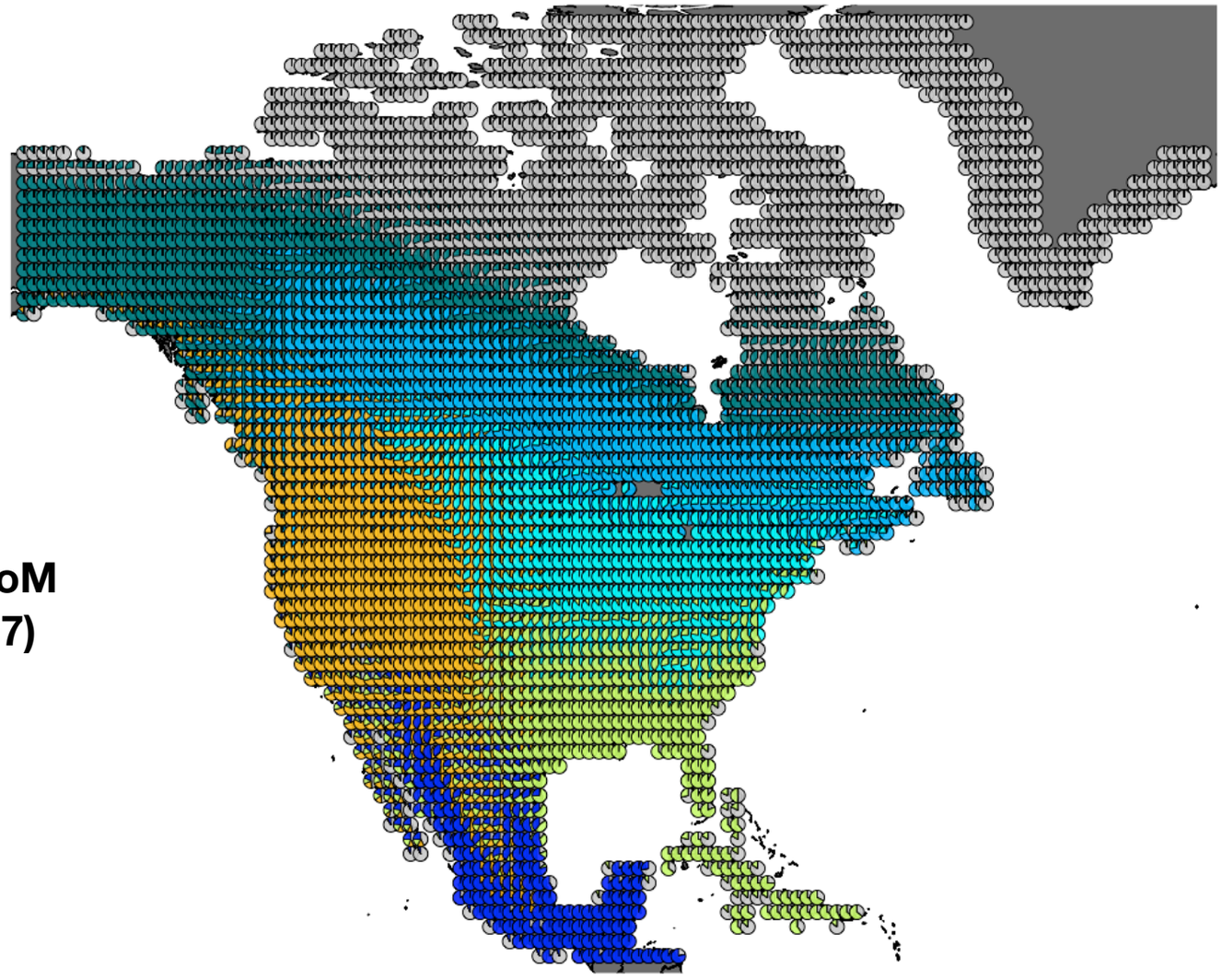


cluster 6 (blue)

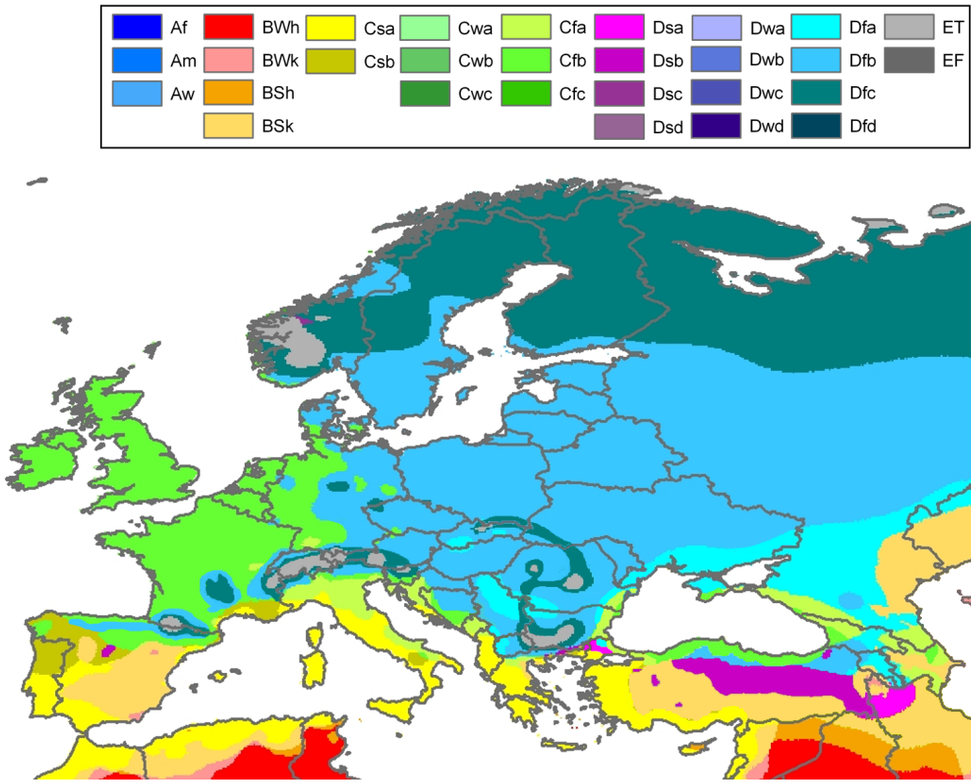


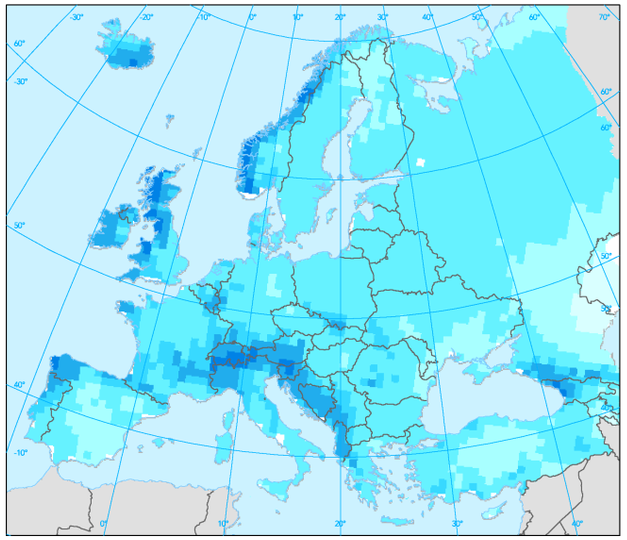




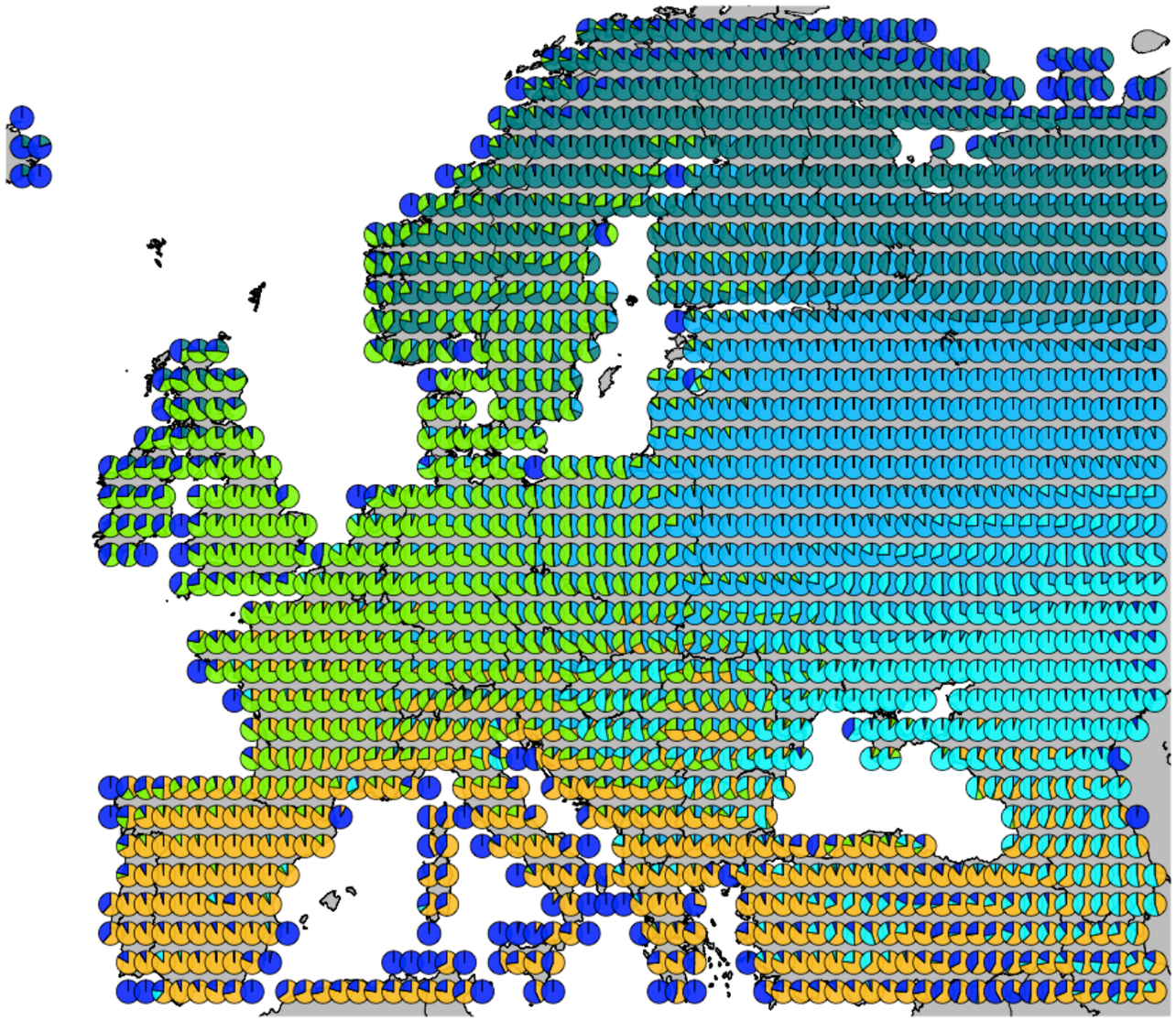


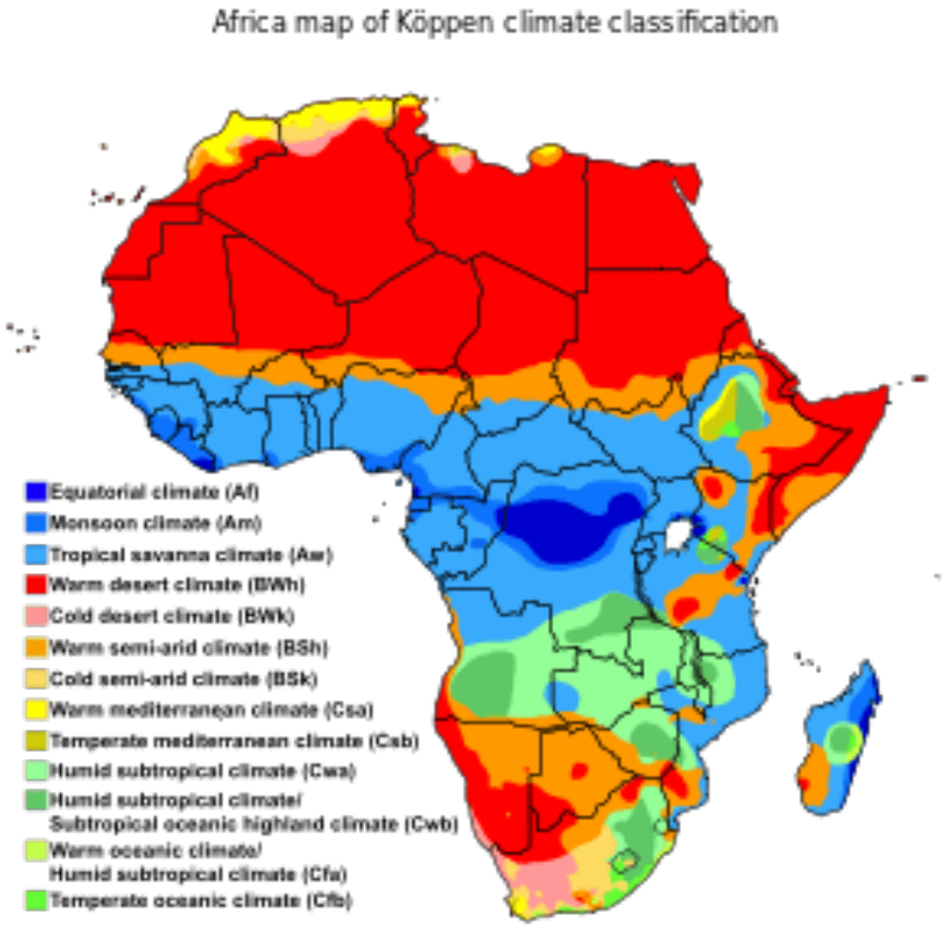
Binomial GoM model (K=7)

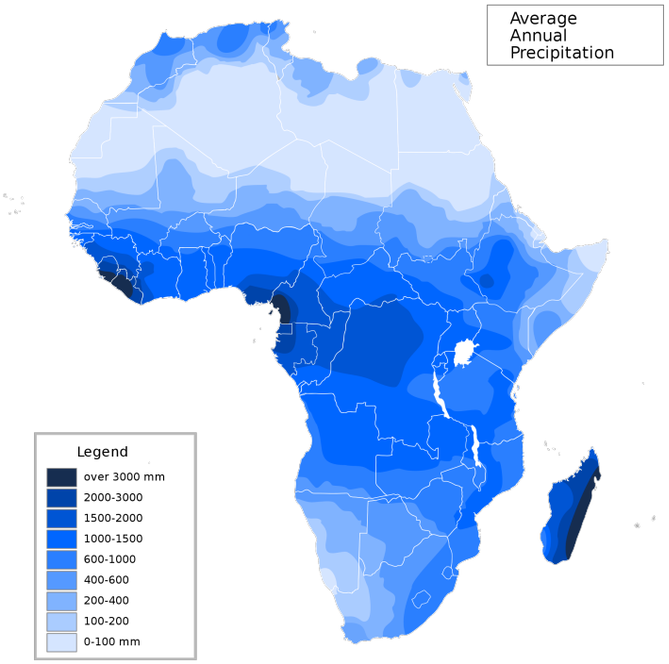




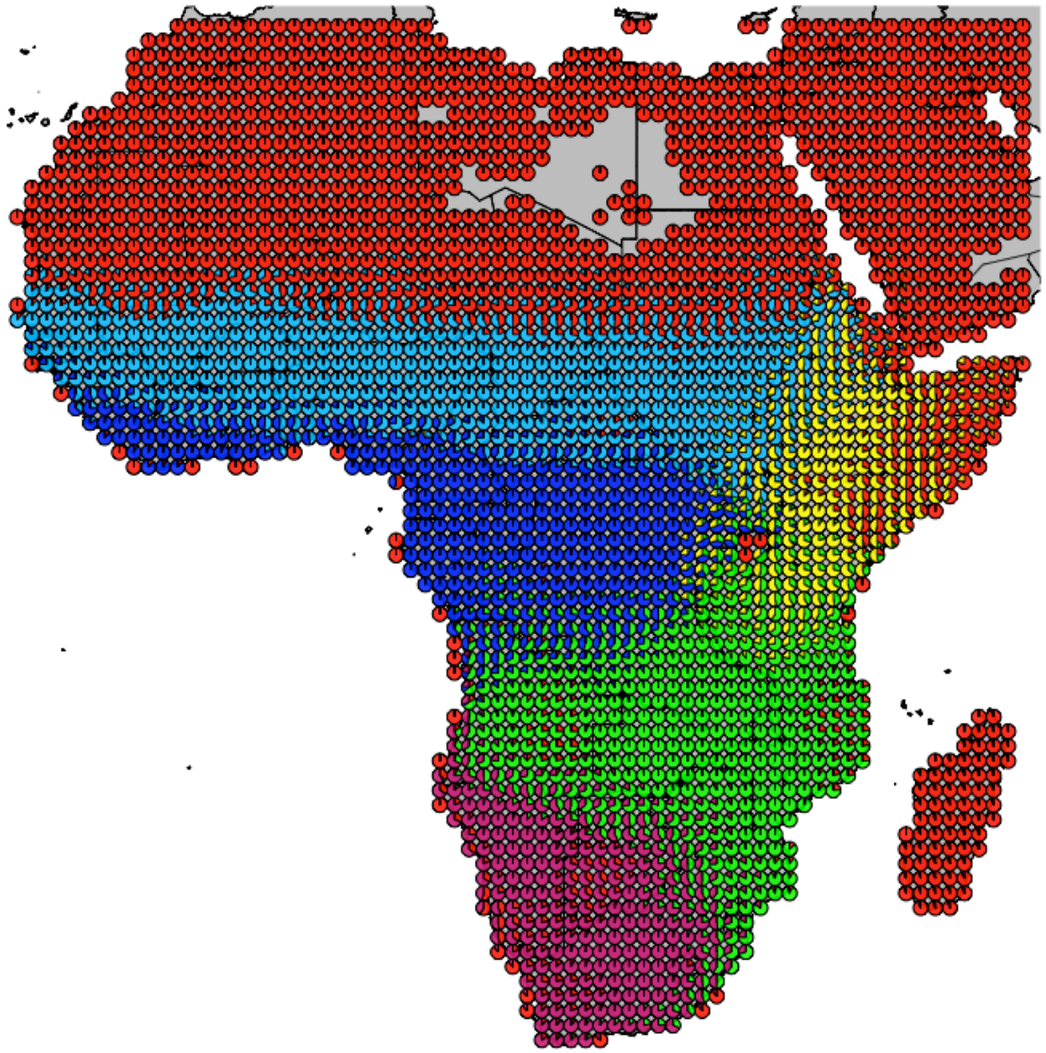
Binomial GoM model (K=6)

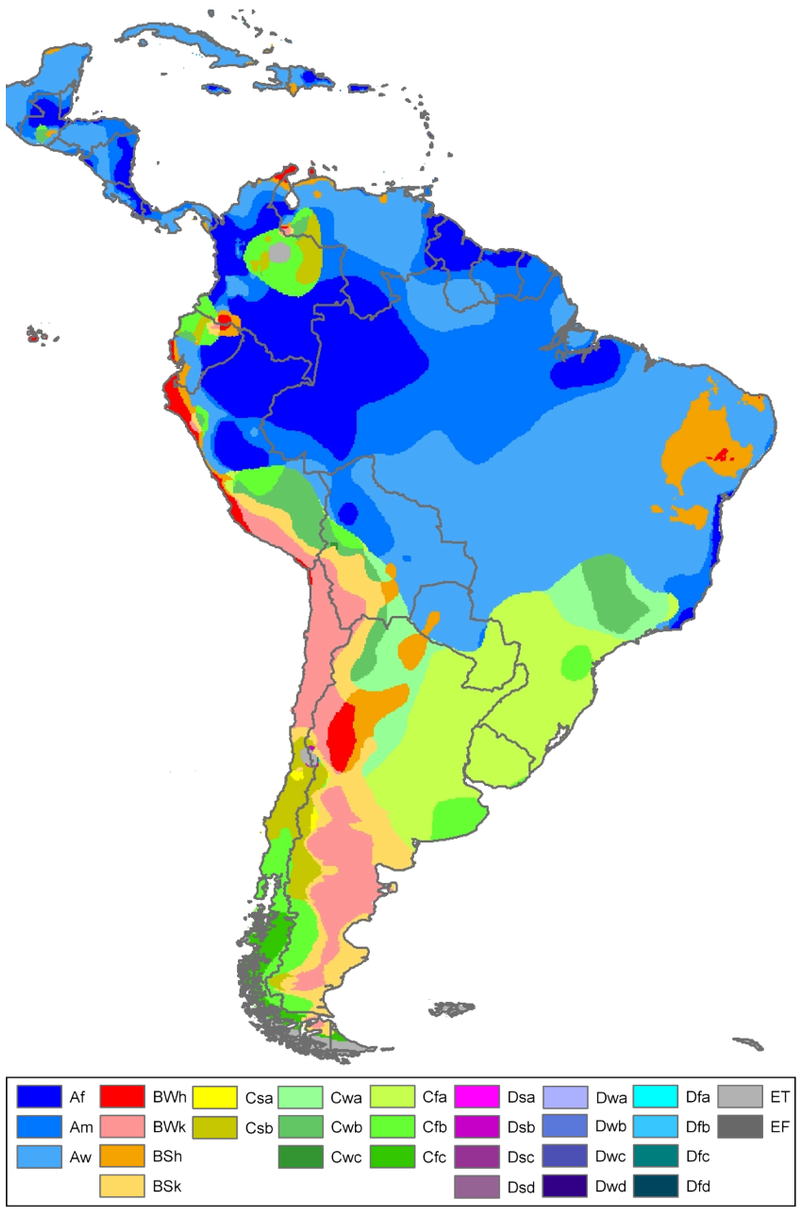




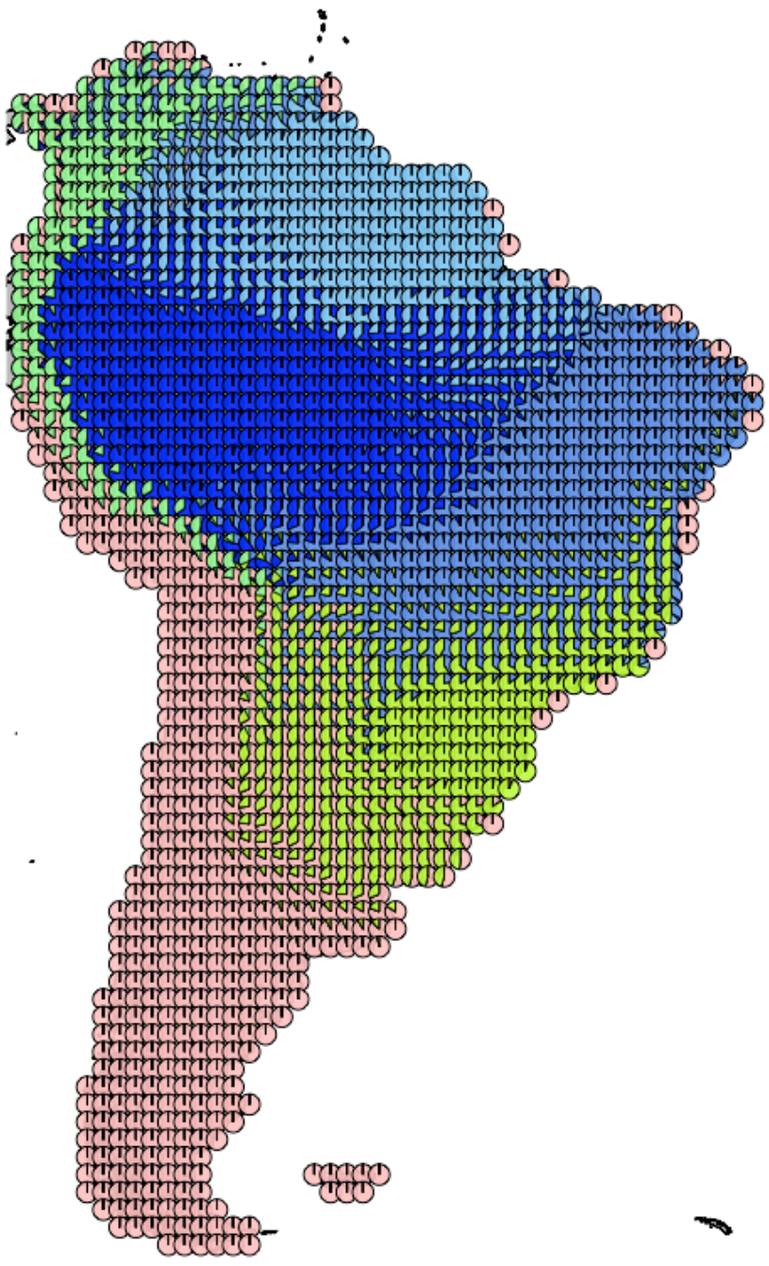


Binomial GoM model (K=6)



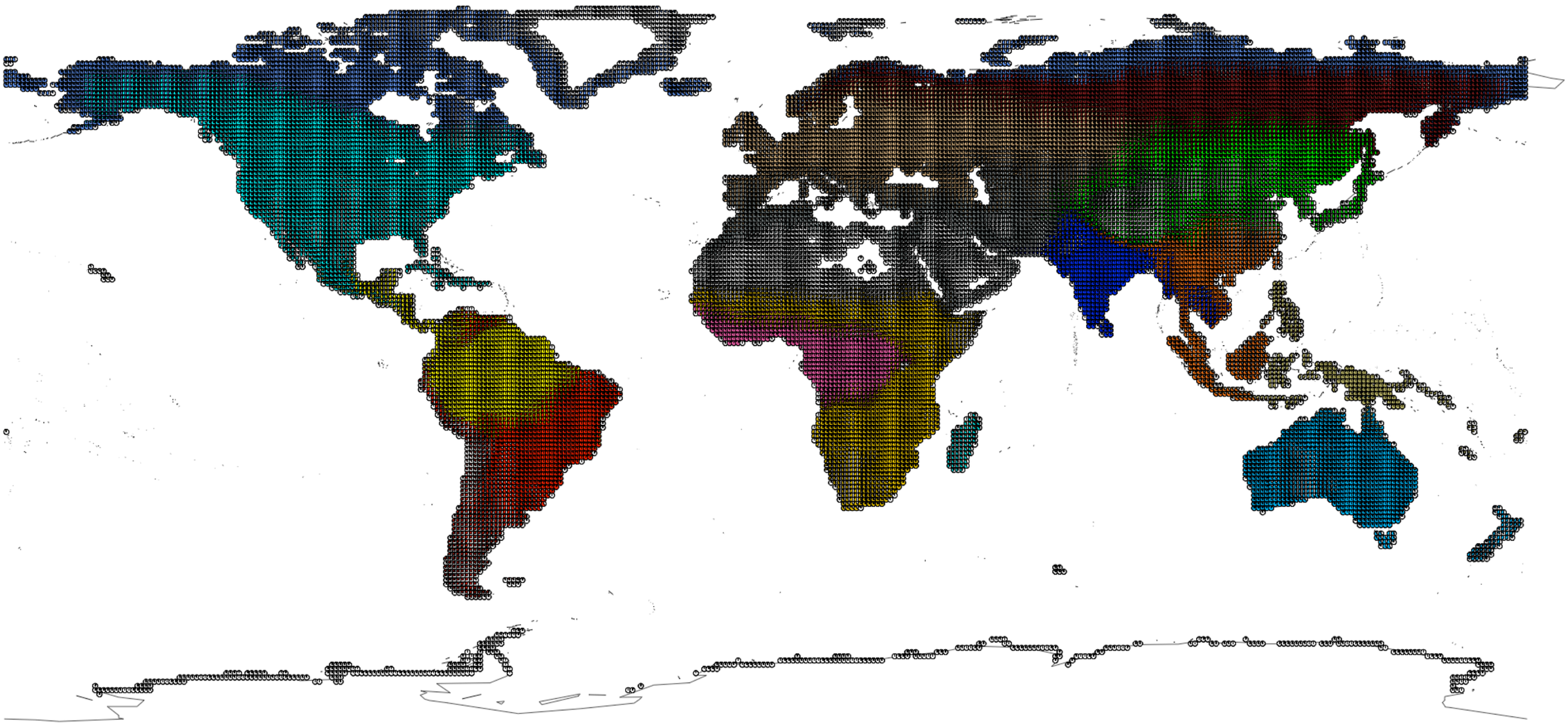


Binomial GoM model (K=6)



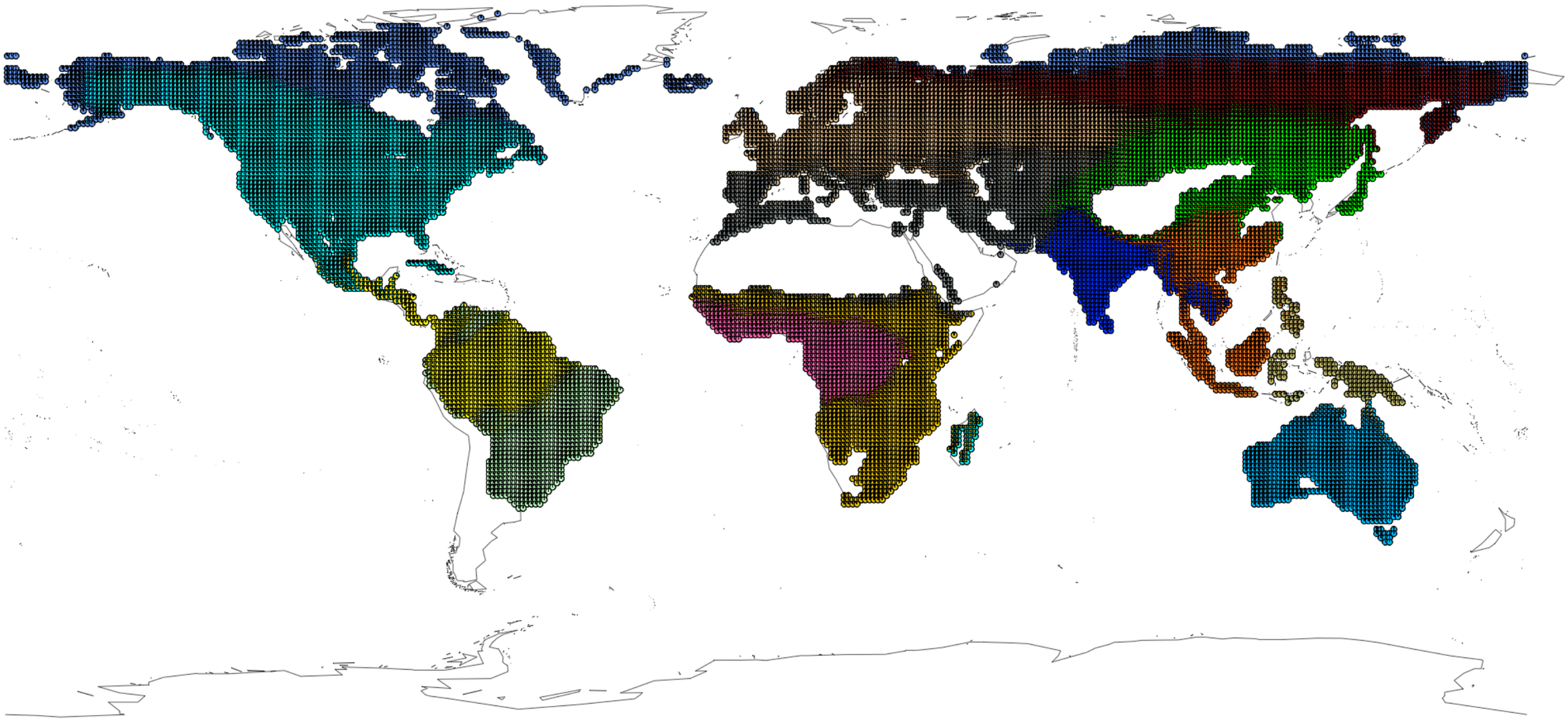
Global Analysis of Birds presence absence using Binomial GoM models

Binomial GoM model k= 15



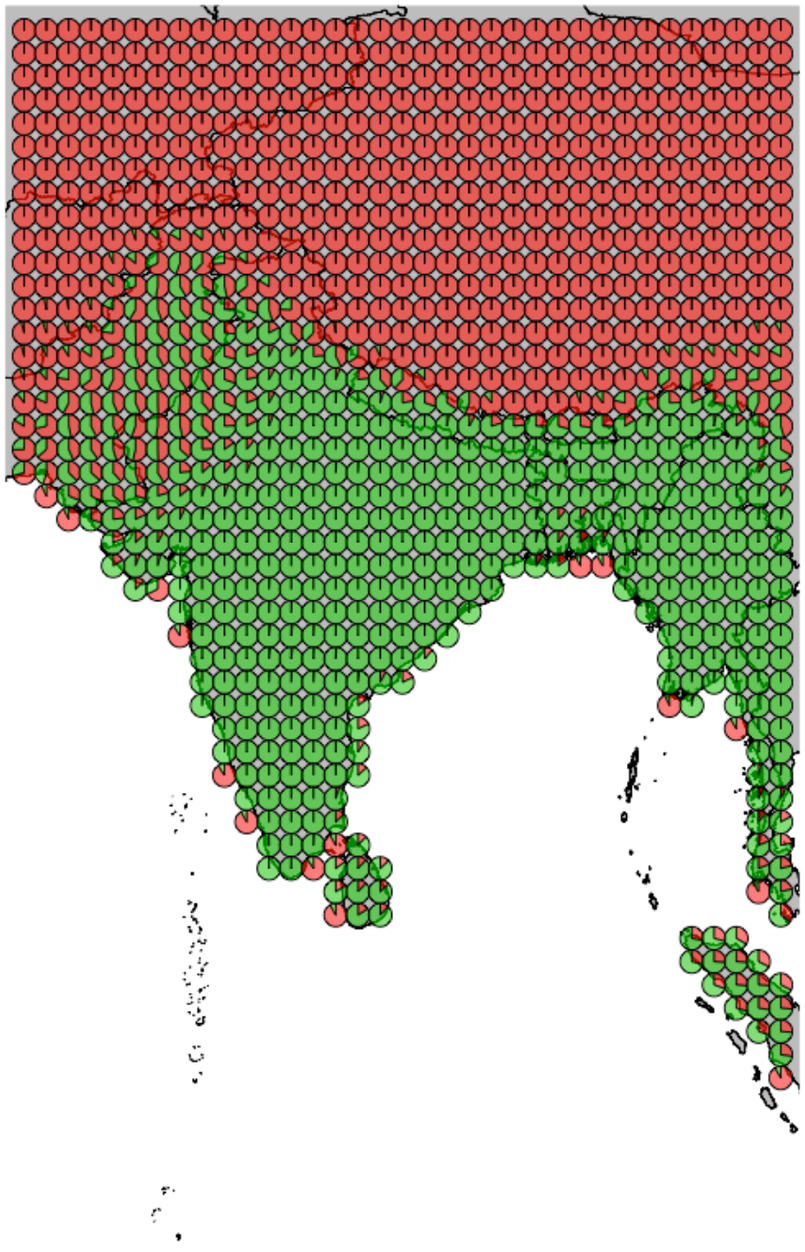
Binomial GoM model k= 15

(*removing pies with high grades in the cluster*
for desert and snow birds)

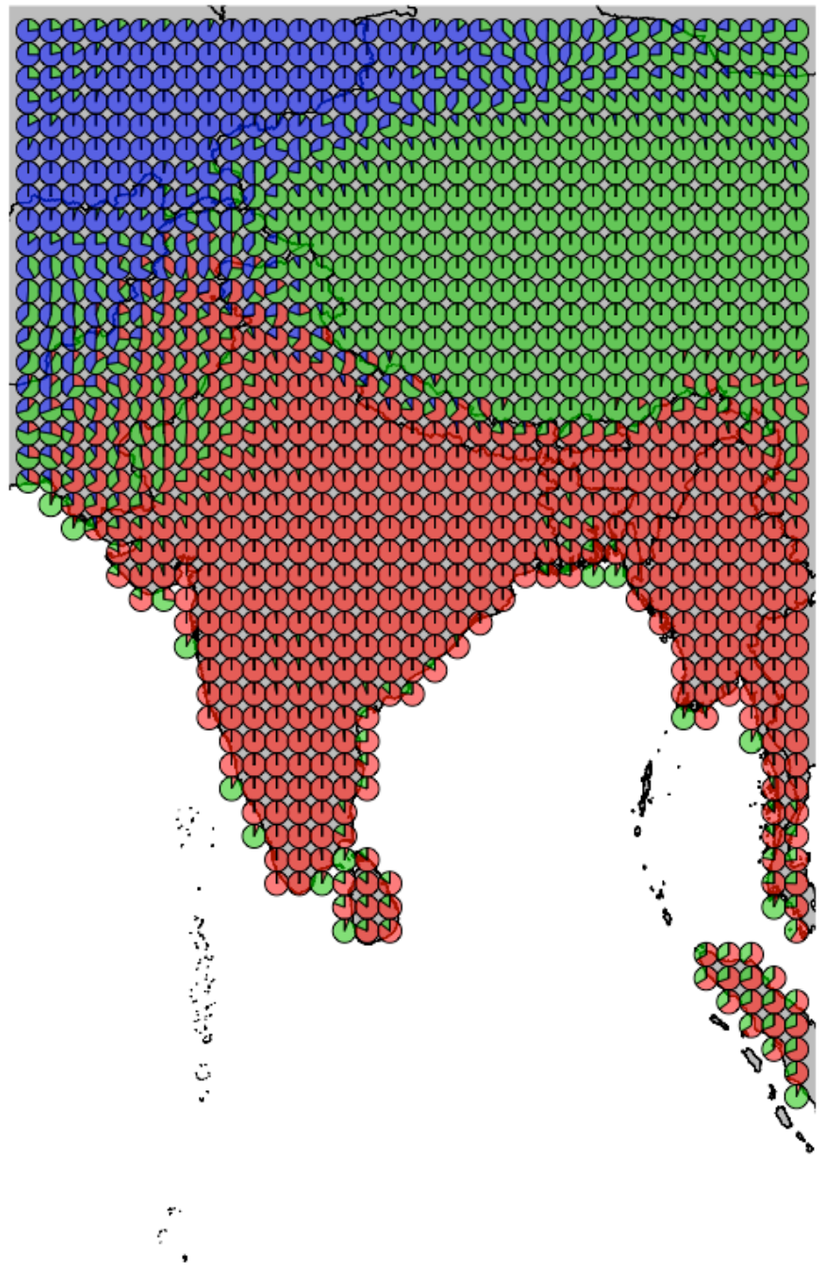


Zooming on the Indian subcontinent

k=2

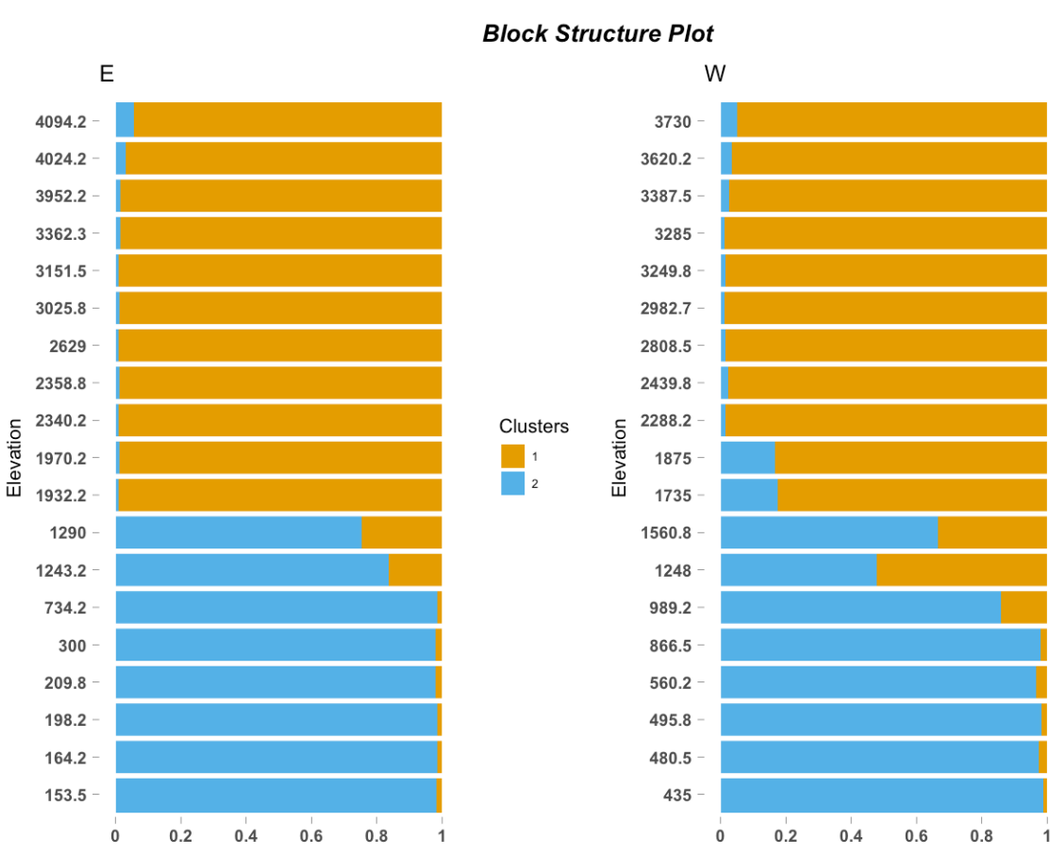


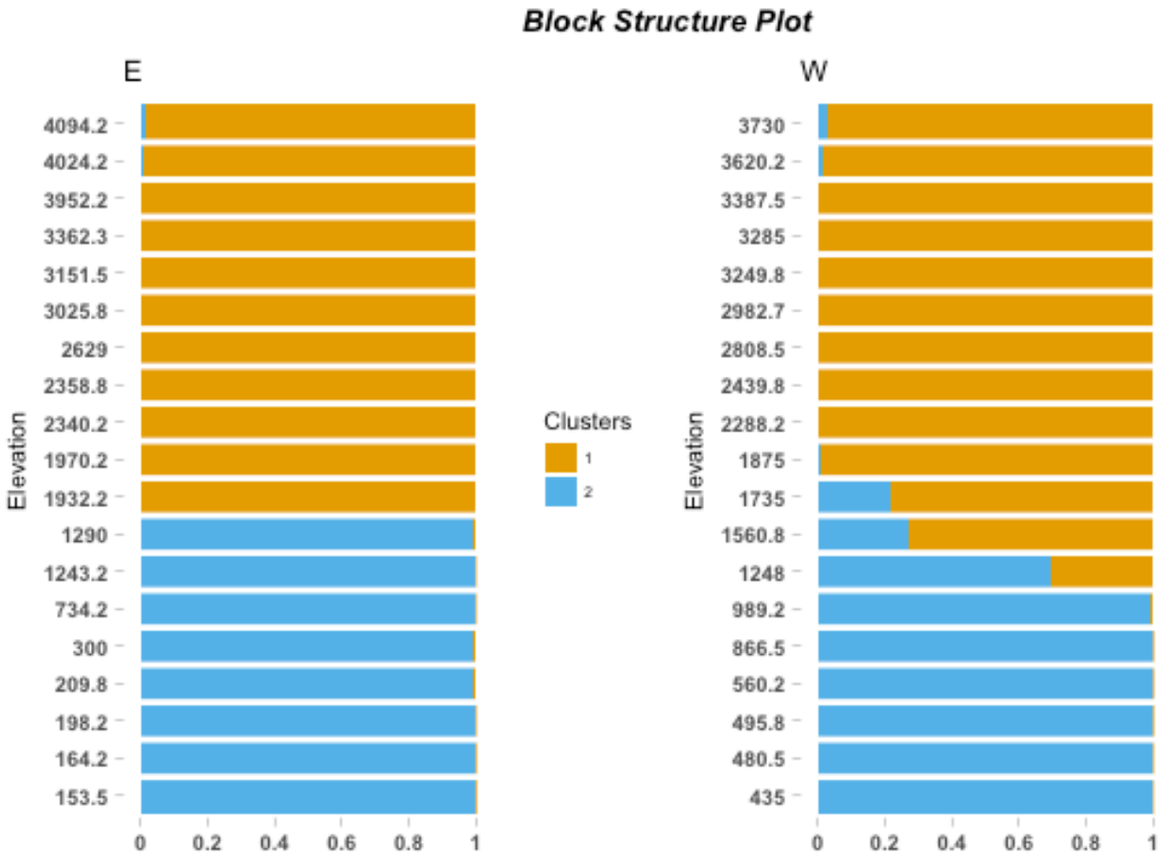
k=3



Both at the global scale analysis and the sub-analysis looking at the regional presence absence patterns of the Indian species, we find that the Himalayan bird communities are a mix of two types - one from North of Himalayas and other from Southern and SE Asia.

Local Bird Abundance Distributional Patterns in the Himalayas





Presence Absence Data (Binomial GoM model K=2)

Local Abundance Data (Multinomial GoM model K=2)

General methClust model

Applications to methylation examples

with Sebastian Post and Kevin Luo

Let M_{ij} (U_{ij}) be the number of methylated (unmethylated) sites in genomic bin j and sample i

$$M_{ij} \sim Bin(M_{ij} + U_{ij}, p_{ij})$$

where we assume a lower dimensional representation of p_{ij}

$$p_{ij} = \sum_{k=1}^{K} \omega_{ik} f_{kj}$$

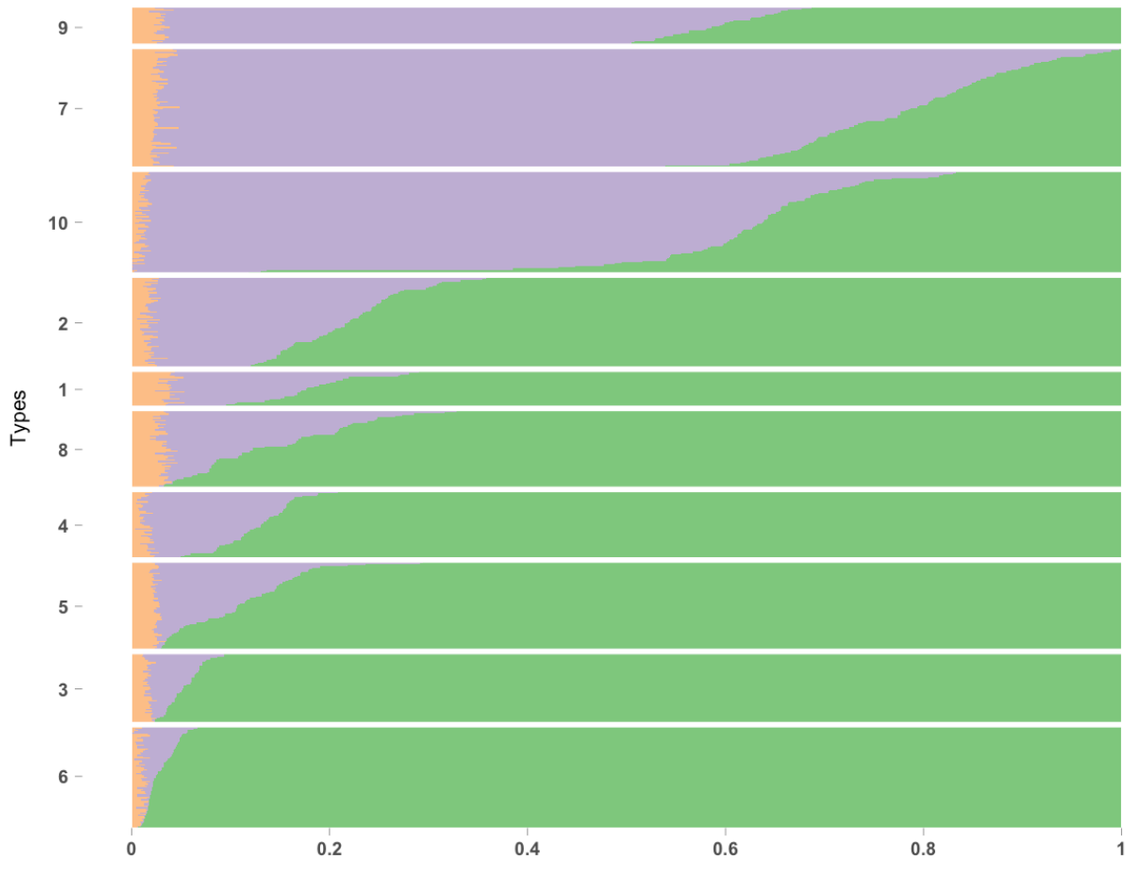
$$\omega_{ik} > 0, \ \sum_{k=1}^{K} \omega_{ik} = 1, \ \ 0 \leq f_{kj} \leq 1$$

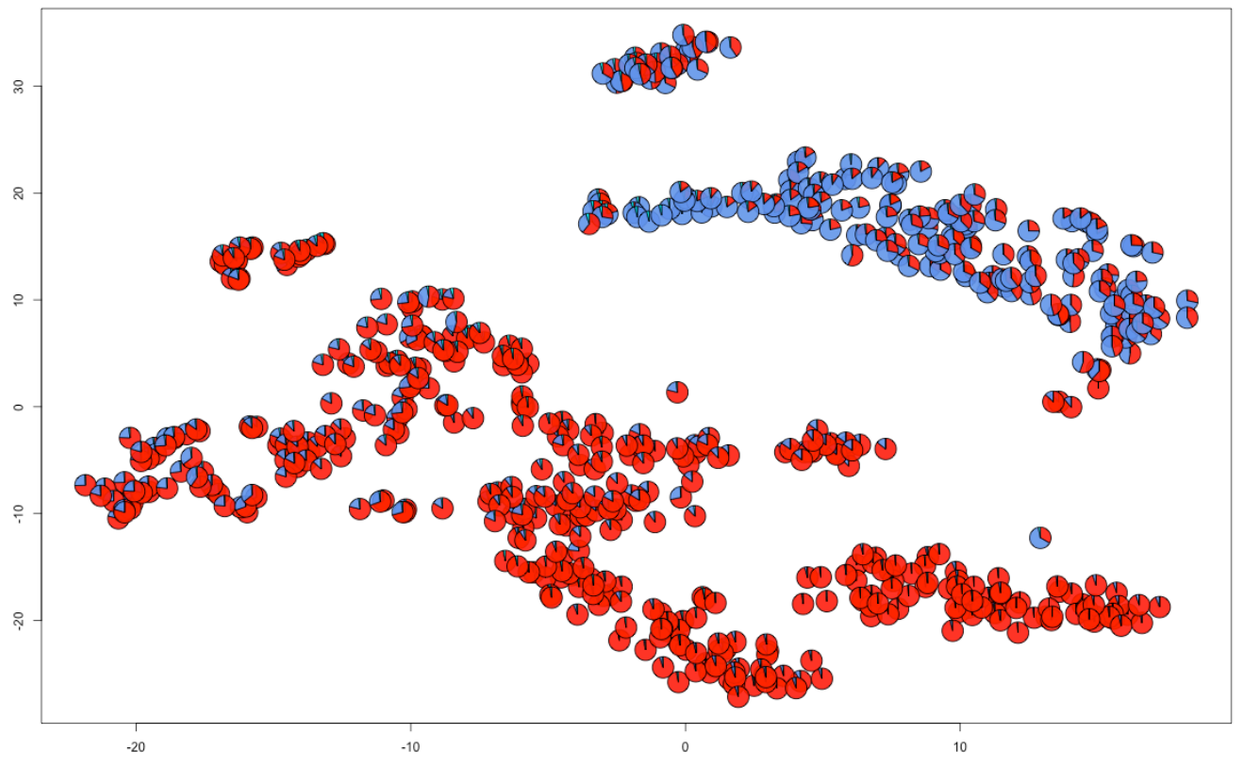
ω_{ik} : grade of membership of cluster k in sample i

f_{kj} : probability of presence of species j in cluster k

hCh data from Sebastian Pott

K=3

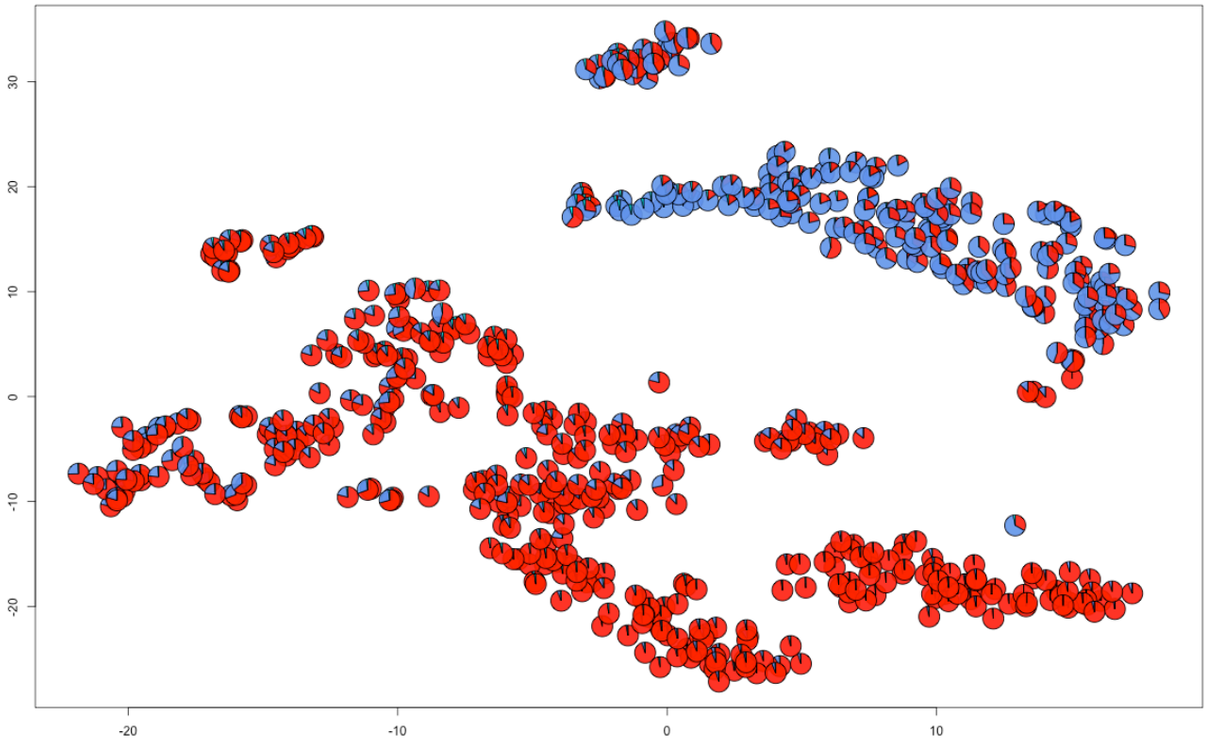




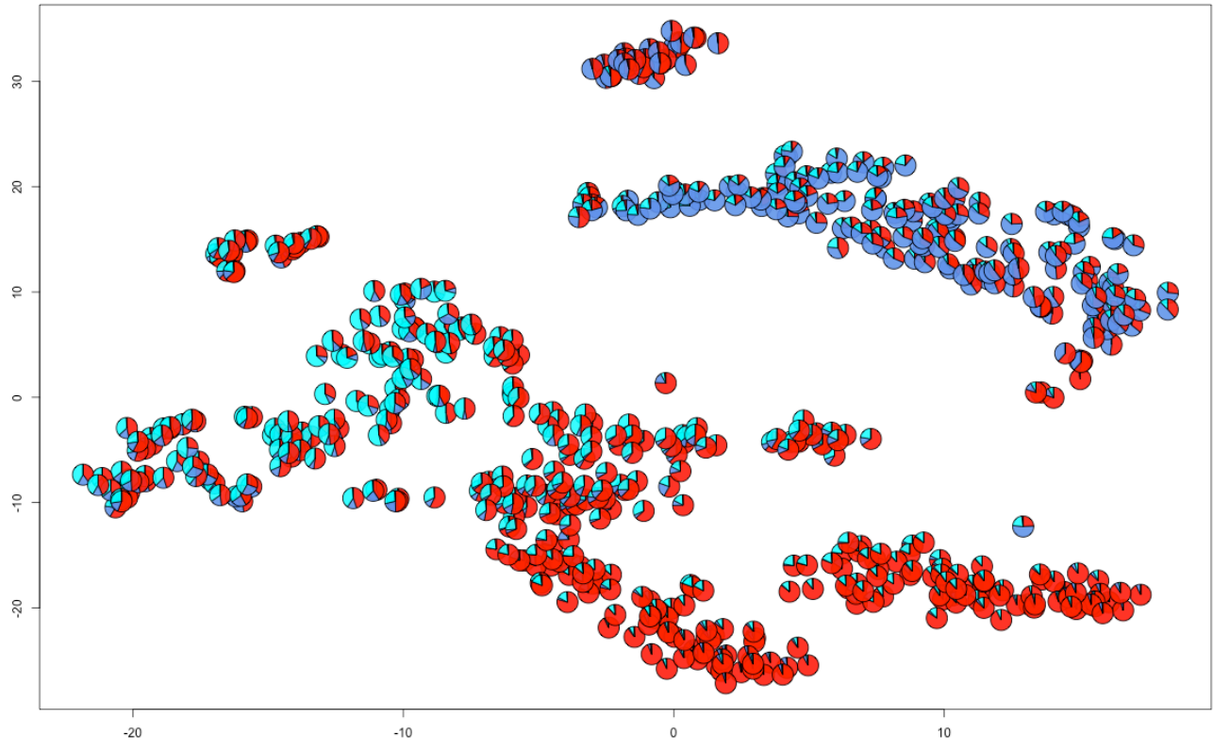
STRUCTURE representation

tSNE + STRUCTURE representation

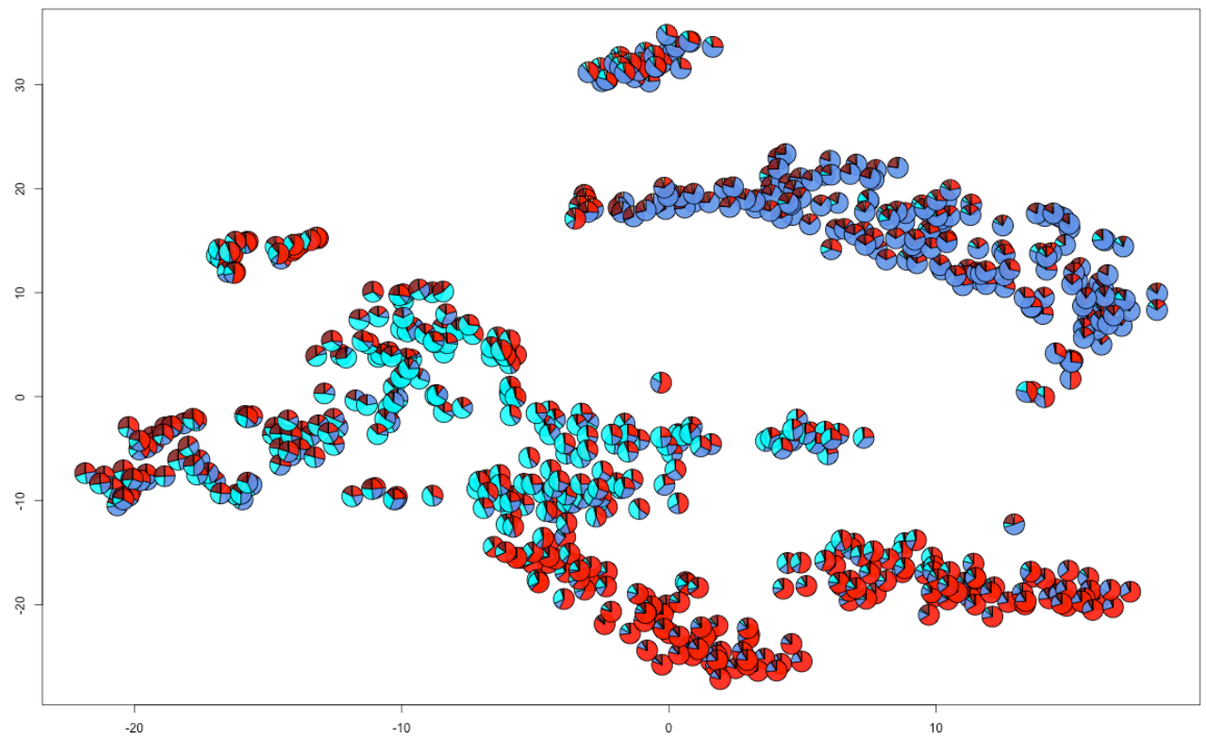
k=3


30
20
10
0
-10
-20
-20
-10
0
10

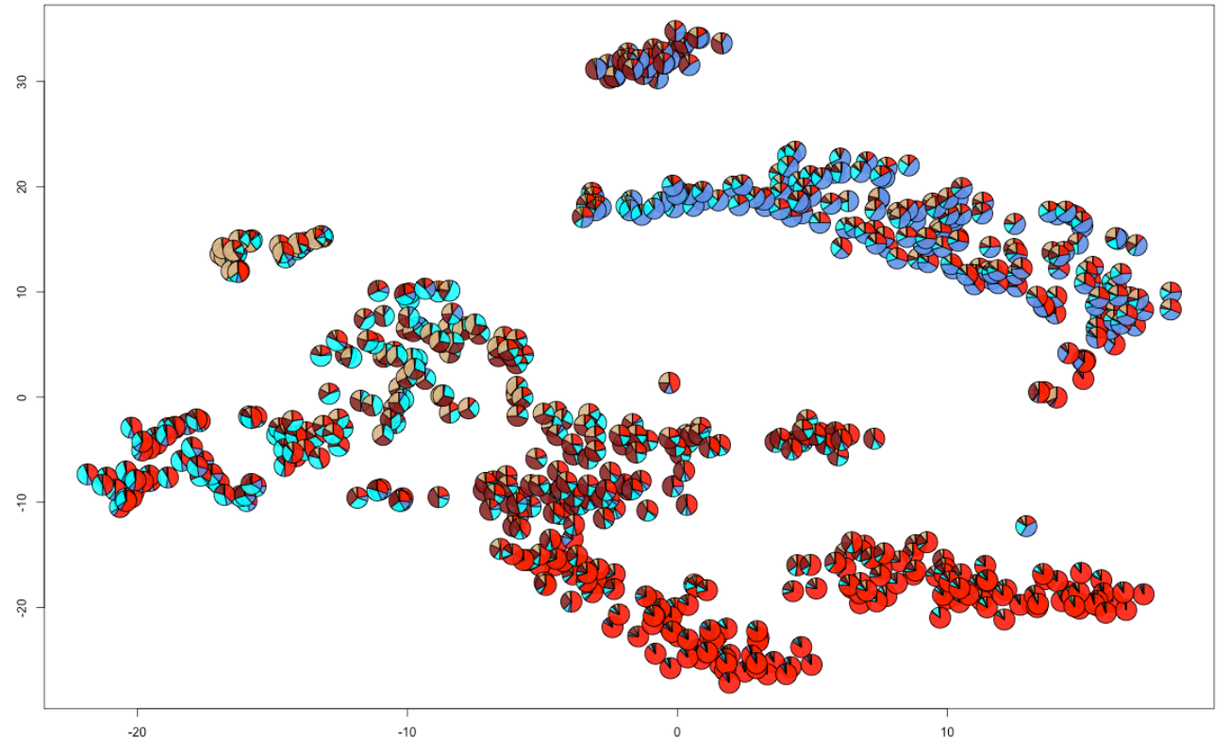
k=4


30
20
10
0
-10
-20
-20
-10
0
10

k=5


30
20
10
0
-10
-20
-20
-10
0
10

k=6


30
20
10
0
-10
-20
-20
-10
0
10