

Diego Cueva



Biólogo da Universidad Nacional de Colombia. Mestrando em Sistemática, taxonomia animal e biodiversidade no Museo de Zoología da USP.

Meu projeto de mestrado: Filogenia molecular dos representantes do gênero *Thraupis* Boie, 1826 (Aves: Passeriformes) e revisão taxonômica do complexo *Thraupis episcopus* (Linnaeus, 1766) e *T. sayaca* (Linnaeus, 1766).

Interesses de pesquisa: Taxonomia e sistemática das aves neotropicales.

[exec](#)

Proposta A

A função irá comparar variáveis contínuas entre dois grupos. Primeiro vai testar a normalidade dos dados com uma prova Shapiro-Wilk com base no alfa selecionado pelo usuário. Dependendo do resultado a função faria uma prova t-student ou um U-test de Mann-Whitney. Finalmente o resultado na tela seria uma tabela com o nome da prova feita e o valor-p por cada coluna.

Argumentos `data.frame` = coluna1 (categorias: macho - fêmea, espécie1 - espécie2, jovem - adulto), coluna 2 em diante (variáveis contínuas: peso, resistência, comprimento da estrutura, etc)

Alfa = 0.01, 0.05, 0.1

Bicaudal= TRUE, FALSE

NA.rm= TRUE

A função foi pensada pra ajudar na comparação morfológica usando espécimes de coleção. Comparar variáveis contínuas é muito comum em estudos taxonômicos e diferenciar por sexo as espécies sem dimorfismo sexual óbvio, amplia as oportunidades de pesquisa.

Proposta B

A função B vai pegar uma tabela de dados de registros de uma espécie. A tabela deve ter data e coordenadas (sistema decimal) de cada um dos registros (observações, capturas, coletas, etc). Vai pegar os dados por mês e vai calcular a distância média entre pontos em cada mês. A função vai comparar as distâncias durante o ano com as distâncias durante o mês pra saber se há movimentos anormais (migração?). Finalmente os doze pontos das médias vão aparecer sobre a tela de um mapa além de uma linha pra ilustrar a área do 50% dos dados. Ilustrando graficamente o movimento.

Argumentos

`data.frame` = incluindo vector de data e coordenadas Y e X em sistema decimal.

Índice= permitindo o usuário eleger se quer calcular a média ou mediana.

A função foi pensada pra ilustrar o movimento de espécies, migração local, atitudinal ou de qualquer tipo. Existem grandes bases de dados (online), coleções e estudos que tem os dados pra estudar diferentes tipos de migração. Os movimentos da fauna de América do Sul é pouco conhecida e pode ser útil pra estratégias de conservação de áreas importantes, reflorestamento, etc.

Oi, Diego,

Ambas as suas propostas parecem factíveis, mas a proposta B me parece mais interessante, não só pelo tema, mas também porque a proposta A é muito similar à função “simula” que usamos em aula. De qualquer jeito, em relação às duas propostas, algumas dicas:

- Proposta A: valeria a pena pensar que o argumento de entrada de dados não precisa necessariamente ser duas colunas de um data-frame, mas sim dois vetores. Além disso, o argumento de alfa não precisa ter apenas 3 opções, a função fica mais versátil se você deixar o usuário escolher entre qualquer valor entre 0 e 1.

- Proposta B: não ficou muito claro como a comparação das distâncias durante o mês com as distâncias durante o ano revelaria se o padrão é anormal. Aliás, o que seria um padrão anormal de movimentação? Além disso, não entendi muito bem o que seria o gráfico: seriam os pontos de coletas ou as médias das distâncias que seriam plotados? Que linha de 50% seria essa? Para deixar a proposta mais clara, você poderia fazer uma lista das etapas necessárias para a função correr bem, assim eu poderia avaliar melhor se ela faz sentido mesmo.

Sugiro investir em detalhar mais essa proposta B, conforme as dicas que sugeri. Ou, se você gostar muito da proposta A, tentar tornar ela um pouco mais complexa e diferente da função simula. Pode ir modificando aqui abaixo mesmo (mas sempre deixe a original aqui na página) e eu vou comentando por aqui mesmo.

Qualquer coisa pode me mandar um email também.

—*Diana Garcia*

Proposta B reformulada

A função B vai pegar um data.frame com registros de uma espécie. O data.frame deve ter a data e coordenadas (sistema decimal) de cada um dos registros (observações, capturas, coletas, etc).

Eu quero que a função primeiro faz grupos por mês. Exemplo: todos os dados de janeiro, fevereiro, ..., dezembro (doze grupos).

A função vai calcular a mediana das coordenadas X, Y e os quartis 25% e 75% por mês.

(decidi que a media é melhor que a média porque a gente quer representar o movimento da maioria

das aves, e não quero muita influência dos pontos extremos)

Eu gostaria que a função mostra-se na tela duas cosas.

1- Uma matriz de distâncias entre pontos

2- Um mapa com as doze medianas (pontos), e cada ponto com duas linhas pra mostrar os quartis em X e Y. Cada ponto pareceria uma cruz, o ponto medio a mediana e cada linha 25% dos dados.

Parecido de plotar dois boxplot um em X e outro em Y.

Argumentos

data.frame = incluindo vector de data e coordenadas Y e X em sistema decimal.

A função foi pensada pra ilustrar o movimento de espécies, migração local, atitudinal ou de qualquer tipo. Existem grandes bases de dados (online), coleções o estudos que tem os dados pra estudar diferentes tipos de migração. Os movimentos da fauna de América do sul é pouco conhecida e pode ser útil pra estratégias de conservação de áreas importantes, reflorestamento, etc.

Oi, Diego,

Legal, agora consegui entender o que seria esse gráfico, acho que a proposta dá para fazer sim. Mas fiquei com algumas dúvidas ainda:

- Na verdade, primeiro, gostaria de saber se isso é comum de se fazer nessa área. Pode ser total ignorância minha, mas fazer uma mediana de coordenadas me apareceu estranho. Na minha cabeça, a união dos pontos externos em uma área de distribuição para cada mês faria mais sentido... É o padrão se fazer a mediana? Se for, não tem problema nenhum, só queria ter certeza.
- A matriz de distância entre pontos seria uma matriz de distância entre quais pontos? Todos? Ou uma matriz por mês?
- Aliás, o melhor seria essa matriz não ser apenas impressa na tela, mas retornar como o objeto de saída da função, não? Ou foi isso que você quis dizer?
- A **mediana** é melhor que é média, certo? Acho que rolou um erro de digitação ali... Mas na verdade eu achei legal a ideia anterior do usuário poder escolher se quer mediana ou média, isso torna a função mais flexível e é mais um desafio de código que pode ser interessante como exercício no R. Que tal colocar isso?
- Por último, você desancou da opção de comparar distâncias por mês e por ano para ver se há indícios de migração? Por quê? Eu não entendi como você poderia calcular isso, mas não me pareceu algo a ser descartado, só melhor explicado. Você pode falar para a função retornar um aviso na tela a depender do resultado de uma tarefa. Teria que definir exatamente que tarefa é essa, mas acho que pode ser um incremento interessante à função.

Bom, acho que é isso, estamos chegando lá, até o fim da semana você já deve ter a proposta pronta para começar a trabalhar na função.

—[Diana Garcia](#)

Olá Diego, acho que a Diana já indicou caminhos para fechar a proposta B. Veja se consegue reformulá-la a partir das sugestões. — [Alexandre Adalardo de Oliveira](#) 2016/04/29 12:43

Oi Diana, a verdade a proposta B não minha área principal de pesquisa e só até este ano, olhe algo parecido. Acho que desenvolver esta função vai ser útil pra ilustrar movimento de animais (movimentos a grande ou pequena escala). No tropico tem muito bicho que faz movimentos relativamente curtos procurando chuvas, disponibilidade de alimento, etc. e são categorizados como não migratórios. Acho que a mediana permitiria perceber aqueles pequenos movimentos melhor que a média (pela influência dos pontos extremos) e o polígono. A união dos pontos extremos vai fazer um polígono que pode não mudar ao longo de um ano porque alguns poucos indivíduos extremos fiquem na sua área (exemplo: *Pandion haliaetus* é uma águia migratória e os indivíduos de primer ano ficam em América do Sul, mas a maior parte da espécie e todos os indivíduos sexualmente maduros voam de volta para Estados Unidos e Canadá. Um polígono não ilustraria muito bem isso, mas a mediana sim). Além disso as barras (quartis) vão permitir olhar se os bichos acumulam ou não (na época seca as aves aquáticas tendem acumular nos corpos de agua que ficam, eu esperaria ter linhas mais cortas).

Preferi jogar fora a opção da media pela influência dos dados extremos.

Não desencanei comparar distancias, só ache que é melhor comparar em uma matriz, fica mais organizado. A comparação é a distância entre a mediana dos meses ao longo do ano. Janeiro vs janeiro (vai dar 0), Janeiro vs fevereiro, janeiro vs março e assim com todos os meses. Acho legal que a matriz seja um objeto de saída.

Mãos à obra.

Oi, Diego,

Ah, agora entendi porque a área de distribuição não seria boa para ilustrar os movimentos, vamos ficar com essa ideia gráfica que você falou mesmo então!

A ideia de incorporar a opção de média era para ser um desafio de programação para incrementar seu trabalho final, mas se você realmente acha que não será

útil de modo algum, ok... Mas pense que no mínimo seria interessante para você ter a opção de comparar o gráfico com a média e o gráfico com a mediana para ver se a mediana é realmente o melhor jeito!

E, por último, agora fiquei confusa com as matrizes. Uma matriz de distância entre pontos (como você colocou na proposta reformulada) é diferente de uma matriz de distância entre a mediana dos pontos para cada mês (como você explicou na resposta)... Para sua proposta ficar fechadinha, seria legal definir melhor que objeto de saída será esse antes de entregar a função.

De resto, mãos à obra mesmo! Boa sorte!

—*Diana Garcia*

Oi Diana,

Então, prefiro tentar primeiro só com a mediana se consigo terminar mais rápido, procuro incluir a media mas acho que já estou bem longe da minha área de confort jajaja.

A matriz seria assim: `Dist = matrix(NA, nrow = 12, ncol= 12)`

então na coluna 1 linha 1 o valor é 0 porque é a distancia entre a mediana de janeiro e janeiro (o mesmo ponto). Na coluna 10 linha 5 quero a distancia entre a mediana de outubro (10) e maio (5). etc.

Obrigado pela ajuda, já tenho minhas primeiras duvidas com as datas. Vou procurar.

Oi, Diego,

Ok, se a média é pouco útil realmente não tem porque usar. Minha preocupação com isso era realmente para incrementar a função, que agora que consegui entender mesmo (ufa, agora deu!), me pareceu um pouco simples. Mas conversei com outros monitores e tivemos outras ideias de como dar uma incrementada sem muito mais trabalho:

- Que tal colocar uma opção para o usuário escolher o que fazer com NAs?
- E se houver registros de mais de uma espécie? Daí a função calcularia uma matriz de distância mês a mês para cada espécie e poderia plotar cada espécie com uma cor diferente no gráfico! Ou mesmo dar a opção para o usuário escolher se quer tudo no mesmo gráfico ou um gráfico por espécie. Não seria útil poder comparar espécies?
- Por último, mas não sei se é viável para o tipo de dado, e se a função também conseguir separar por indivíduos jovens e adultos também? De novo, a função poderia fazer as mesmas coisas que sugeri para as

espécies.

Enfim, o ideal mesmo seria que você escolhesse alguma dessas coisas para dar uma incrementada na função se conseguir, ok? Sei que parece difícil quando você lê, mas quando você quebrar a cabeça para resolver vai perceber que não é tão complicado assim. Confie em você! E o fórum está aí para tirar qualquer dúvida também, lembre-se disso!

—*Diana Garcia*

Oi Diana,

Tá, vou incluir a opção de trabalhar com contagem de indivíduos, se o usuário quer. Não é igual um registro de 1 indivíduo que um registro de 100 (o registro de 100 tem que pular os dados mais pra seu lado, eu acho)

Também vou incluir a opção de eliminar os dados com NA (o usuário esqueceu de anotar o número de indivíduos) ou trocá-lo por 1, assim o usuário pode usar as coordenadas daquele registro.

e seu da certo o tempo incluir a opção de selecionar a espécie que quer plotar em caso que tenha mais de uma.

Obrigado, abçs.

Função

```
## data.frame with the columns sp (specie), y (latitud), x (longitud),  
date(y/m/d), ind (individuals count)*  
## *optional  
  
install.packages("rworldmap") # It is necessary to install the package  
"rworldmap" to use the function  
  
migration = function(x, specie="", na.rm=FALSE, na.one=TRUE,  
ind.count=FALSE)  
{  
  cc=x # Transform the data set  
  if (specie=="") ## User didn't pick up a specie from the data set  
  {  
    cc$sp[1] # Select the specie at the first position  
  }  
  else  
  {
```

```

    cc=cc[cc$sp==specie, ] # Pick up all the rows with the specie name the
user pick up
}
if(na.rm==TRUE) # The users want to delete the NAs
{
    cc=na.omit(cc) # Delate the rows with NAs in the column ind The other
columns shouldn't have NAs, doesn't make sense to have rows without
coordanates or withour date
}
if(na.one==TRUE) # In case the user didn't want to delete the rows with
NAs and want to use the coordanates at least once, It will turn the NAs as
1, so the record will appear as 1 individual, better than nothing
{
    cc[is.na(cc)]=1 # Turns NAs into 1
}
if(ind.count==TRUE) # If the user wants to use records with individuals
counts
{
    y=c() # New vector for the loop
    x=c() # New vectortor for the loop
    date=c() # New vector for the loop
    for(i in 1:length(cc$ind)) # This loop is going to multiple each row by
the number of individuals recorded, so locations with more individuals will
attract the median more than those with a single individual
    {
        Y=rep(cc[i,"y"], cc[i,"ind"]) # multiply the latitude
        y=c(y,Y) # Put the data in the vector
        X=rep(cc[i,"x"], cc[i,"ind"]) # multiply the longitude
        x=c(x,X) # Put the data in the vector
        DATE=rep(cc[i,"date"], cc[i,"ind"]) # multiply the date
        date=c(date,DATE) # Put the data in the vector
    }
    cc=data.frame(y, x, date) # Create a data.frame over the first one
}
cc$date=gsub("-", "/", cc$date) # In case the date have "-" it replaces
them for "/" not sure why, but the function works better this way
cc$date=as.Date(cc$date) # Turn the date column into date format
cc$month=as.numeric(format(cc$date, format="%m")) # creates a new column
with only the month
ymed=tapply(cc$y, cc$month, FUN=median) # Calculated the median latitude
for month
yinf=tapply(cc$y, cc$month, FUN=quantile, probs = c(0.25), names=F) #
Calculated the 0.25 quantil of the latitude for month
ysup=tapply(cc$y, cc$month, FUN=quantile, probs = c(0.75), names=F) #
Calculated the 0.75 quantil of the latitude for month
xmed=tapply(cc$x, cc$month, FUN=median) # calculated the median longitude
for month
xinf=tapply(cc$x, cc$month, FUN=quantile, probs = c(0.25), names=F) #
calculated the 0.25 quantil of the longitude for month
xsup=tapply(cc$x, cc$month, FUN=quantile, probs = c(0.75), names=F)
#Calculated the 0.75 quantil of the longitude for month

```

```
    coor = data.frame(yinf, ymed, ysup, xinf, xmed, xsup) # New object with
the coordinates of each point for month
    R = 6371 # Radius of earth in Km (aprox)
    coor$yrad=coor$ymed*pi/180 # Turning longitude into radians to compere
distances because earth isn't plain
    coor$xrad=coor$xmed*pi/180 # Turning latitude into radians to compere
distances because earth isn't plain
    Dist = matrix(NA,12,12) # Create a matrix 12 x 12
    for(i in 1:12) # Compering the month i with..
    {
        for(j in 1:12) # ... the month j
        {
            dist = round((acos(sin(coor[i,"yrad"])*sin(coor[j,"yrad"])) +
cos(coor[i,"yrad"])*cos(coor[j,"yrad"]) * cos(coor[j,"xrad"]-
coor[i,"xrad"]))) * R),0)
            Dist[i,j]=dist # Compering distances between points
        }
    }
    Dist[is.nan(Dist)]=0 # The function may create NaNs when comparing between
the same month, here we are fixing that. Distances between the same month is
0, because is the same point
    rownames(Dist)=c("J","F","M","A","M","J","J","A","S","O","N","D") #
Changing the row names for the initial of each month
    colnames(Dist)=c("J","F","M","A","M","J","J","A","S","O","N","D") #
Changing the columns names for the initial of each month
    Mig=Dist<quantile(Dist,0.75, na.rm=T) # Assuming that the fardest points
are the arrival points, this matrix shows the probability of birds being
migrating or not. Each TRUE for month increase the change of active
migration, those are the months where researchers should look for migration.
    library("rworldmap") # Opening the package
    map <- getMap(resolution = "low") # Creating a map
    plot(map, xlim = c((round(min(coor$xinf),0)-1),
(round(max(coor$xsup),0)+1)), ylim = c((round(min(coor$yinf),0)-1),
(round(max(coor$ysup),0)+1)), asp = 1) # Ploting the map
    for(i in 1:12) # Marking the lines from quantiles 0.25 to 0.75 in X and Y
    {
        yy= c(coor$ymed[i], coor$ymed[i]) # Starting point
        xx= c(coor$xinf[i], coor$xsup[i]) # Final point
        lines(xx, yy, col= "green2") # Ploting quantiles lines from east to west
        yyy= c(coor$yinf[i], coor$ysup[i]) # Starting point
        xxx= c(coor$xmed[i], coor$xmed[i]) # Final point
        lines(xxx, yyy, col= "green2") # Plotting quantiles lines from south to
north
    }
    points(coor$xmed, coor$ymed, col = "steelblue1", cex= 2, pch=19) #
Plotting median points
    text(coor$xmed, coor$ymed, row.names(coor),cex=0.8) # Number of each month
    A=list(Dist,Mig) # Objects to recover
    return(A) # Returning objects
```

```
}
```

Arquivo da Função

[migration.r](#)

Exemplo

[cardellina_canadensis.csv](#) [setophaga_striata.csv](#)

Help

```
data.frame {colnames=c(sp,ind*,y,x,date)}  
Documentation
```

Package:unknown

R

```
migration()
```

Description

Calculated the median coordinates and the interquartile range, base it on the monthly records. It plots this information on a map.

Creates a matrix (12x12) comparing the distances between the median coordinates of each month.

Creates a second matrix (12x12) that shows the probable months that presents migration movements or not.

Usage

```
migration (x, specie = "", na.rm = FALSE, na.one = TRUE, ind.count =  
FALSE)
```

Arguments

x

A data.frame including the columns sp, ind, y, x, date.

specie

In case the data.frame have more than one specie, user must write down the name of the specie he wants to plot.

na.rm

It will eliminate the NAs from the analysis.

na.one

It will swap the NAs in the column "ind" for a 1, so those coordinates become useful when the analysis includes individuals count.

ind.count

The data analysis will be made using the individual count for each

record.

Value

comp1: Distances comparison matrix. It calculates the distance (Km) between the median coordinates per month.

comp2: Migration probability matrix. It shows the months whit higher likely to occur migration (more TRUE per month, higher chance of migration movements) or not (more FALSE per month, less chance of migration movements).

Warning

Sometimes when estimating the distance between the same month, the function will generate small numbers that will be replaced it with NaNs. The function by it self edits the Distances comparison matrix replacing the NaNs with 0.

Author

Diego Cueva

Examples

```
sp = c(rep("example", 100))
ind = c(round(runif(100, 1, 20),0))
y = c(runif(100, -60, 60))
x = c(runif(100, -100,-10))
date = seq(as.Date("2015/1/1"), as.Date("2015/12/31"), length.out=100)
e = data.frame(sp,ind,y,x,date)
migration(e, specie="example", na.rm=FALSE, na.one=FALSE,
ind.count=FALSE)
```

From:

<http://ecor.ib.usp.br/> - **ecoR**

Permanent link:

http://ecor.ib.usp.br/doku.php?id=05_curso_antigo:r2016:alunos:trabalho_final:dacuevac:start

Last update: **2020/08/12 06:04**

