Análisis de las llamadas de A. pertinax (variable Tcall dur)

José R. Ferrer Paris

2019-11-15

## Exploración de los datos

Leemos el archivo de datos de Aratinga pertinax, versión de 2019, y reclassificamos la variable Region para facilitar interpretación:

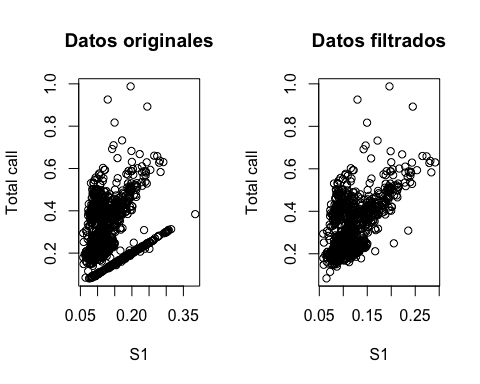
dts <- read.csv(sprintf("%s/data/mdf\_JR\_15viii19.csv",script.dir))  
str(dts)

## 'data.frame': 1351 obs. of 9 variables:  
## $ IndivGroup: Factor w/ 97 levels "AUA01","AUA02",..: 1 1 1 1 1 1 1 1 2 2 ...  
## $ soundfile : Factor w/ 1351 levels "0211327a","0211344a",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ S1\_dur : num 0.156 0.133 0.14 0.136 0.149 ...  
## $ Tcall\_dur : num 0.154 0.137 0.144 0.137 0.144 ...  
## $ RecSite : Factor w/ 37 levels "A6","A7","B1",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Lat : num 12.5 12.5 12.5 12.5 12.5 ...  
## $ Long : num -69.9 -69.9 -69.9 -69.9 -69.9 ...  
## $ LocCode : Factor w/ 14 levels "AUA","BON","CUR",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Region : Factor w/ 2 levels "isl","main": 1 1 1 1 1 1 1 1 1 1 ...

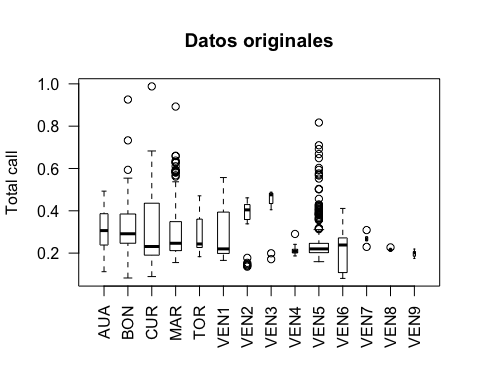
dts$Region <- factor(dts$Region,levels=c("main","isl"))

Verificamos las dos variables (duración de la primera sílaba y duración total). Existe una correlación natural entre ambas, pero hay un grupo que tiene un valor casi idéntico y se aprecia como un línea recta. No se si esto es un error, o si se desea incluir este tipo de observaciones en el análisis, pero me parecen que deberían excluirse. [Esas son llamadas de una sola sílaba, y por ende, S1=Tcall. Esas llamadas NO deben ser excluidas. Puntos que caen debajo de esa línea reflejan errores de medición; medí las duraciones manualmente (en espectrogramas), porque muchas tenían problemas con “background noise.”]

layout(matrix(1:2,ncol=2))  
plot(Tcall\_dur~S1\_dur,dts,xlab="S1",ylab="Total call",main="Datos originales")  
plot(Tcall\_dur~S1\_dur,subset(dts,Tcall\_dur>S1\_dur\*1.1),xlab="S1",ylab="Total call",main="Datos filtrados")

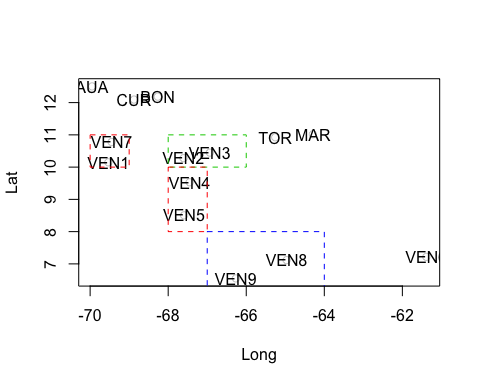
 Verificamos la duración total por localidad (LocCode). El ancho es proporcional al número de observaciones. Mi sugerencia para no sacrificar tantas muestras es combinar los grupos más pequeños según proximidad geográfica y similitud en la duración. [Como un punto de interés es de investigar/cuantificar la variabilidad de llamadas dentro de localidades (LocCode), me preocupa variar demasiado la escala geográfica de las muestras en cada LocCode. O sea, me preocupa que al combinar grupos, los LocCodes en tierra firme abarquen areas mucho más grandes que los LocCodes en islas. Por esa razón, creo que tal vez sería major eliminar/excluir los LocCodes con muestras muy pequeñas: VEN3, VEN7, VEN8, y VEN9... posiblemente VEN4 tambien. Se mantendría la escala geográfica de muestreo a través de los LocCodes, y todavía tendríamos más o menos el mismo número de LocCodes en islas y en tierra firme.]

boxplot(Tcall\_dur~LocCode,dts,las=2,varwidth=T,xlab="",ylab="Total call",main="Datos originales")



La distribución espacial de las localidades y la agrupación sugerida es la siguiente.

xys <- with(dts,aggregate(data.frame(Lat,Long),by=list(LocCode=LocCode),median))  
plot(Lat~Long,xys,col="grey81")  
with(xys,text(Long,Lat,LocCode),cex=.7)  
rect(-67,6,-64,8,border=4,lty=2)  
rect(-66,10,-68,11,border=3,lty=2)  
rect(-67,8,-68,10,border=2,lty=2)  
rect(-69,11,-70,10,border=10,lty=2)

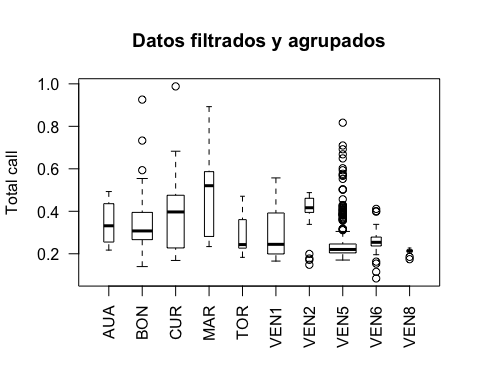


A continuación aplicamos el filtro y agrupamos las localidades según la exploración anterior.

dts <- subset(dts,Tcall\_dur>S1\_dur\*1.1)  
dts[dts$LocCode %in% c("VEN2","VEN3"),"LocCode"] <- "VEN2"  
dts[dts$LocCode %in% c("VEN7","VEN1"),"LocCode"] <- "VEN1"  
dts[dts$LocCode %in% c("VEN8","VEN9"),"LocCode"] <- "VEN8"  
dts[dts$LocCode %in% c("VEN4","VEN5"),"LocCode"] <- "VEN5"  
##dts <- subset(dts,!LocCode %in% c("VEN3","VEN7","VEN8","VEN9") & (Tcall\_dur>S1\_dur\*1.1))  
dts$LocCode <- droplevels(dts$LocCode)  
dts$IndivGroup <- droplevels(dts$IndivGroup)

El gráfico de duración por localidad queda de la siguiente manera.

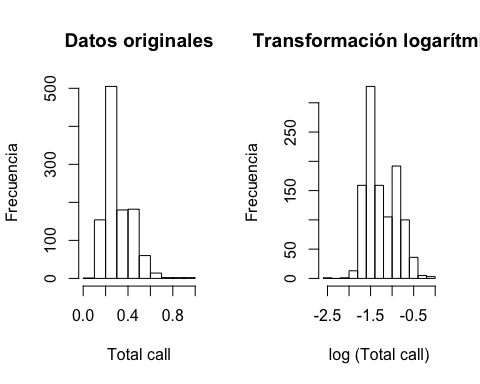
boxplot(Tcall\_dur~LocCode,dts,las=2,varwidth=T,xlab="",ylab="Total call",main="Datos filtrados y agrupados")



Vemos que la localidad VEN2 (combinación de VEN2 y VEN3) tiene una duración más parecida a la de las islas, y la de TOR tiene una duración más parecida a la de tierra firme. Por otro lado VEN5 tiene mucha variabilidad. ¿Alguna sugerencia sobre como explicar esto? [Por el momento no...]

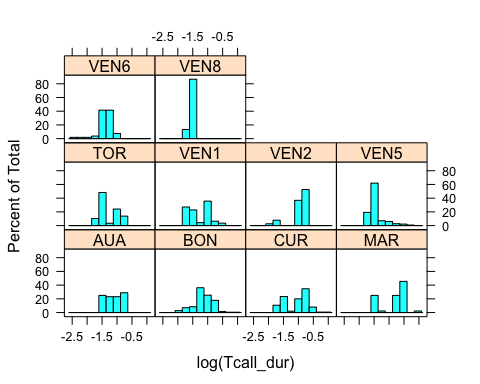
Otro aspecto clave es que la distribución de valores de la variable duración total no es normal, de hecho es ligeramente asimétrica y esto genera una distribución asimétrica de residuales en los modelos. Después de varias pruebas, parece que la transformación logarítmica mejora la distribución de los residuales, sin embargo, al aplicarla resalta una ligera bimodalidad en los datos. [Nota: estas distribuciones no incluyen las llamadas de una sola sílaba, ya que éstas fueron filtradas... habría que ver como cambian las distribuciones si la llamadas de una sílaba son incluídas. Supongo que la bimodalidad va a ser peor al incluir llamadas de una sola sílaba...]

layout(matrix(1:2,ncol=2))  
hist(dts$Tcall\_dur,ylab="Frecuencia",xlab="Total call",main="Datos originales")  
hist(log(dts$Tcall\_dur),ylab="Frecuencia",xlab="log (Total call)",main="Transformación logarítmica")

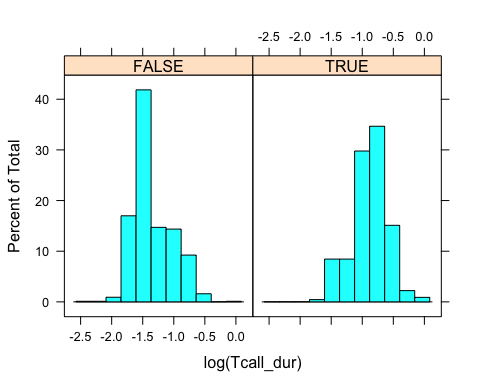


Al revisar el histograma por Localidad, pareciera que la bimodalidad es más pronunciada en algunas localidades que en otras

histogram(~log(Tcall\_dur)|LocCode,dts)

 Pero parece estar más relacionada con la duración de la primera sílaba. Por ejemplo si se compara las que tienen una primera silaba corta vs larga (umbral arbitrario de 0.135ms), los histogramas se ven mejor. ¿Puede que se trate de varios tipos de llamadas/comportamientos diferentes? [Ver nota anterior sobre filtración de las llamadas con una sola sílaba. En cuanto a tu pregunta sobre tipos de llamadas... es muy posible que la categoría de “loud contact calls” podría incluir varios sub-tipos de llamada, pero no es obvio como definirlos. Son llamadas fuertes y frecuentes, que son producidas mientras vuelan los pericos, y cuando están posados y llamando de uno a otro. Bien podrían estar diciendo “aqui estoy!” o “ven para acá!” o “ya voy!” . Yo no he podido reconocer sub-categorías (y otros investigadores trabajando con psitácidos tampoco sub-dividen la categoría de “loud contact call”).]

histogram(~log(Tcall\_dur)|(S1\_dur>.135),dts)



Mi recomendación para los análisis es: \* filtrar llamadas consistentes de una sola silaba (Tcall\_dur<S1\_dur*1.1)*  agrupar las localidades de tierra firme por proximidad y similitud para homogeneizar un poco el tamaño muestral \* usar la transformación logarítmica para la duración \* Es necesario considerar como aminorar la bimodalidad en el modelo, después de considerar varias opciones estoy usando un ‘offset’ de S1\_dur (suponiendo que la duración de S1 tiene un efecto directo en Tcall), pero tal vez haga falta introducir alguna variable adicional o restringir el análisis a un subconjunto de grabaciones más homogeneo [Ver comentarios anteriores sobre filtración y agrupación de localidades. En cuando al ‘offset’ de S1... creo que sería mejor hacer el ajuste de modelos de dos formas (como hicieron para la tesis de Irene): en un caso se consideran modelos para la variable Tcall\_dur, y en otro caso se consideran modelos para la variable S1\_dur. O sea, no mezclar Tcall\_dur y S1\_dur en un solo modelo, pues no tenemos ninguna hipótesis acerca de efectos de S1\_dur sobre Tcall\_dur. Simplemente son dos medidas diferentes de duración para caracterizar las llamadas.]

## Ajuste de modelos

Vamos a considerar varios modelos para la variable Tcall\_dur y utilizar el valor de AIC calculado por la función para escoger el mejor modelo. Debemos obviar el valor de las pruebas de t y F en las tablas, y basar la selección en las diferencias en AIC.

A partir de la exploración de los datos usamos un modelo con efecto fijo constante (offset(S1\_dur)) y dos alternativas para representar el efecto geográfico : offset(S1\_dur) + Region o offset(S1\_dur) + Long + Lat

La nomenclatura de los modelos es la siguiente

|  |  |  |  |
| --- | --- | --- | --- |
| MODELO | componente | valor |  |
| f0\_\_ | efecto fijo | offset(S1\_dur) |  |
| f1\_\_ | efecto fijo | offset(S1\_dur) + Region |  |
| f2\_\_ | efecto fijo | offset(S1\_dur) + Long + Lat |  |
| f\_0\_ | Efecto aleatorio | LocCode/IndivGroup |  |
| f\_1\_ | Efecto aleatorio | Region + LocCode/IndivGroup |  |
| f\_\_0 | Heterocedasticidad | No |  |
| f\_\_1 | Heterocedasticidad | Si |  |

[Me parece que el incluir Lat podría introducir una confusión con el efecto de Region, ya que todas las islas tienen valores altos de latitud. No veo la ventaja de incluir Lat y Region... creo que sería mejor usar solamente Region, porque un objetivo central del estudio es de entender cómo cambian las vocalizaciones de poblaciones aisladas en islas.]

El modelo nulo con efecto fijo de la isla

f000 <- lme(log(Tcall\_dur)~offset(log(S1\_dur)),dts,random=~1|LocCode/IndivGroup, method="ML")  
f100 <- lme(log(Tcall\_dur)~offset(log(S1\_dur))+Region,dts,random=~1|LocCode/IndivGroup, method="ML")  
f200 <- lme(log(Tcall\_dur)~offset(log(S1\_dur))+Lat+Long,dts,random=~1|LocCode/IndivGroup, method="ML")  
  
f010 <- update(f000,random=~1|Region/LocCode/IndivGroup)  
f110 <- update(f100,random=~1|Region/LocCode/IndivGroup)  
f210 <- update(f200,random=~1|Region/LocCode/IndivGroup)  
  
## con esta parametrización no hace falta usar pdDiag  
##f010 <- update(f000,random=list(LocCode=pdDiag(~Region),IndivGroup=pdDiag(~Region)))  
##f110 <- update(f100,random=list(LocCode=pdDiag(~Region),IndivGroup=pdDiag(~Region)))  
##f210 <- update(f200,random=list(LocCode=pdDiag(~Region),IndivGroup=pdDiag(~Region)))  
  
f001 <- update(f000,weights=varIdent(form=~1|Region))  
f101 <- update(f100,weights=varIdent(form=~1|Region))  
f201 <- update(f200,weights=varIdent(form=~1|Region))  
f011 <- update(f010,weights=varIdent(form=~1|Region))  
f111 <- update(f110,weights=varIdent(form=~1|Region))  
f211 <- update(f210,weights=varIdent(form=~1|Region))

Comparamos el valor de AIC

(mis.aics <- AIC(f000,f100,f200,f010,f110,f210,  
 f001,f101,f201,f011,f111,f211))

## df AIC  
## f000 4 145.5782  
## f100 5 142.9303  
## f200 6 138.5812  
## f010 5 146.9251  
## f110 6 144.9303  
## f210 7 140.5812  
## f001 5 144.9056  
## f101 6 142.2596  
## f201 7 137.8655  
## f011 6 146.2524  
## f111 7 144.2596  
## f211 8 139.8655

Reordenamos los modelos según el AIC

aic.tab <- cbind(mis.aics,delta.AIC=mis.aics[,2]-min(mis.aics[,2]))  
aic.tab[order(aic.tab$AIC),]

## df AIC delta.AIC  
## f201 7 137.8655 0.0000000  
## f200 6 138.5812 0.7157406  
## f211 8 139.8655 2.0000005  
## f210 7 140.5812 2.7157412  
## f101 6 142.2596 4.3941496  
## f100 5 142.9303 5.0648610  
## f111 7 144.2596 6.3941501  
## f001 5 144.9056 7.0401163  
## f110 6 144.9303 7.0648615  
## f000 4 145.5782 7.7126924  
## f011 6 146.2524 8.3869201  
## f010 5 146.9251 9.0596738

Los cuatro mejores modelos () incluyen efectos fijos de Latitud y Longitud que puede interpretarse como fuerte evidencia de la existencia de un gradiente espacial en la longitud de la llamada, pero difieren en los componentes de varianza (efectos aleatorios y heterocedasticidad).

El mejor modelo incluye efectos aleatorios en dos niveles (localidades y grupos) y heterocedasticidad entre regiones (continente vs isla), mientras que la inclusión de un tercer nivel de efectos aleatorios o la exclusión de la heterocedasticidad tiene menos soporte.

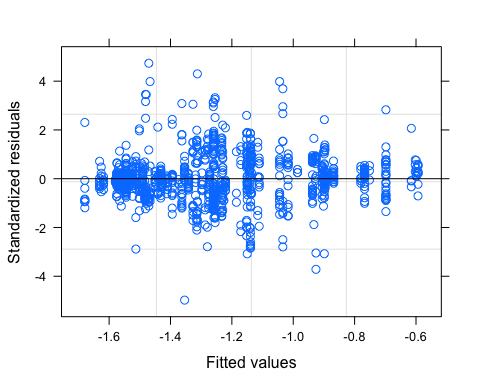
A continuación el resumen del mejor modelo. Los efectos fijos sugieren que hay efectos positivos de la latitud y longitud en la duración de la llamada, pero el efecto de la latitud es de mayor magnitud. La mayor porción de varianza está en los residuales, seguida por los grupos, y la menor porción en las localidades, y la función de varianza residual es ligeramente menor en tierra firme que en las islas.

summary(f201)

## Linear mixed-effects model fit by maximum likelihood  
## Data: dts   
## AIC BIC logLik  
## 137.8655 172.8996 -61.93273  
##   
## Random effects:  
## Formula: ~1 | LocCode  
## (Intercept)  
## StdDev: 0.08208923  
##   
## Formula: ~1 | IndivGroup %in% LocCode  
## (Intercept) Residual  
## StdDev: 0.20886 0.2426937  
##   
## Variance function:  
## Structure: Different standard deviations per stratum  
## Formula: ~1 | Region   
## Parameter estimates:  
## isl main   
## 1.0000000 0.9289595   
## Fixed effects: log(Tcall\_dur) ~ offset(log(S1\_dur)) + Lat + Long   
## Value Std.Error DF t-value p-value  
## (Intercept) 1.1019388 1.0848752 1008 1.015729 0.3100  
## Lat 0.1169593 0.0251532 82 4.649878 0.0000  
## Long 0.0526380 0.0180340 82 2.918818 0.0045  
## Correlation:   
## (Intr) Lat   
## Lat 0.385   
## Long 0.980 0.559  
##   
## Standardized Within-Group Residuals:  
## Min Q1 Med Q3 Max   
## -4.97928301 -0.47730485 -0.05176557 0.44342414 4.73457250   
##   
## Number of Observations: 1102  
## Number of Groups:   
## LocCode IndivGroup %in% LocCode   
## 10 94

Los residuales del modelo tienen una distribución ligeramente asimétrica y muestran varios valores extremos en ambas colas (< -3 o >3).

plot(f201)



Los intervalos de confianza aproximados sugieren efectos significativos en los efectos fijos. Los efectos fijos de los modelos alternativos son muy similares.

intervals(f201,which="fixed")

## Approximate 95% confidence intervals  
##   
## Fixed effects:  
## lower est. upper  
## (Intercept) -1.02403394 1.10193884 3.22791162  
## Lat 0.06698971 0.11695926 0.16692880  
## Long 0.01681146 0.05263799 0.08846452  
## attr(,"label")  
## [1] "Fixed effects:"

Los intervalos de confianza aproximados para los componentes de varianza muestran: \* componentes de similar magnitud entre grupos y residual y menor valor para localidad, \* el componente de heterocedasticidad no es completamente diferente a uno (valor de referencia para las islas).

intervals(f201,which="var-cov")

## Approximate 95% confidence intervals  
##   
## Random Effects:  
## Level: LocCode   
## lower est. upper  
## sd((Intercept)) 0.03017115 0.08208923 0.2233472  
## Level: IndivGroup   
## lower est. upper  
## sd((Intercept)) 0.1750121 0.20886 0.2492541  
##   
## Variance function:  
## lower est. upper  
## main 0.8512431 0.9289595 1.013771  
## attr(,"label")  
## [1] "Variance function:"  
##   
## Within-group standard error:  
## lower est. upper   
## 0.2272778 0.2426937 0.2591552

Aproximadamente 50% de la varianza es residual (dentro de grupos o “no explicada”), 40 % se debe a diferencia entre grupos dentro de las localidades y solo 6% se puede atribuir a diferencias entre localidades

vs <- varcomp(f201)  
round(vs\*100/sum(vs),1)

## LocCode IndivGroup Within   
## 6.2 39.9 53.9   
## attr(,"class")  
## [1] "varcomp"

Los resultados con modelos alternativos son similares (la proporción de varianza atribuible a la region en los modelos f211 y f210 es equivalente a cero).

vs <- varcomp(f200)  
round(vs\*100/sum(vs),1)

## LocCode IndivGroup Within   
## 6.4 41.9 51.7   
## attr(,"class")  
## [1] "varcomp"

vs <- varcomp(f211)  
round(vs\*100/sum(vs),1)

## Region LocCode IndivGroup Within   
## 0.0 6.7 43.1 50.2   
## attr(,"class")  
## [1] "varcomp"

vs <- varcomp(f210)  
round(vs\*100/sum(vs),1)

## Region LocCode IndivGroup Within   
## 0.0 6.4 41.9 51.7   
## attr(,"class")  
## [1] "varcomp"