Analisis\_generico\_baza

# Definir directorio de trabajo

Aqui probablemente el codigo te dara problemas, porque queremos definir como directorio de trabajo la carpeta de COPLAS porque los archivos estan ahi. Como este Markdown esta dentro de un Proyecto con su carpeta asociada, R entiende que el directorio es ese (".../Instar/implementaciones/instar\_baza/scripts/git\_Instar\_baza"). Estrictamente hablando deberiamos escribir el codigo de manera que cuando llamemos a un archivo le demos el path apropiado ("...\_validacion\_datos\_validacion\_PLAGAS"). Pero tambien se puede definir el directorio de manera manual (Session > Set Working Directory > Choose Directory), y asi R si acepta que sea una carpeta distinta a donde esta el proyecto. Por mi las dos opciones valen, la que te sea mas facil :)

He definido el WD de manera manual.

Definir como directorio la carpeta: Instar>implementaciones>instar\_baza>scripts>git\_Instar\_baza>git\_Instar\_baza Para poner esta carpeta he copiado los archivos de los que se alimenta este codigo, que son: - zona\_baza.csv (Datos de COPLAS en Baza)

* cortijuela.csv (Datos de COPLAS en Sierra Nevada)
* instar\_bz008 (Datos Instar)
* instar\_bz006
* instar\_sn020
* instar\_sn021

# Cargar paquetes

# DATOS INSTAR

## Importar datos INSTAR

* La tabla de datos se construye en excel y se guarda como .csv
* Se abre con procesador de texto, y se comprueba que la fecha está en formato mm-dd-yy
* Se guarda como .txt Las columnas que constituyen la tabla son: fecha, huevo, L1, L2, crisalida, radiacion, tmax, tmin, tmed y exergia.

TOMAMOS EL RODAL GR-023-008 (Sierra de Baza) COMO EJEMPLO

INSTAR<-read.table("instar\_bz008.txt", header=T, sep=",", dec=".")  
INSTAR$fecha <-as.Date(INSTAR$fecha, format="%m-%d-%y")

## Anadir biociclos al data.frame

Aparece dos chunks para este proceso, uno para Baza y otro para Sierra Nevada, ya que tienen fechas de inicio diferentes.

## Agregacion INSTAR

Media del vigor y sumas (l1, l2 y l1+l2) por biociclo

# DATOS COPLAS

## Importar datos COPLAS

COPLAS<-read.csv("zona\_baza.csv", header=TRUE, sep=",")  
names(COPLAS)[1] <- "RODAL"

Activar el rodal necesario para el analisis

## Definir biociclo COPLAS

## Union de tablas INSTAR y COPLAS

En esta tabla aparecen los datos de Instar agregados, los Ln de sumL1, sumL2 y sumL1+sumL2, el anio y el grado de infestacion de COPLAS.

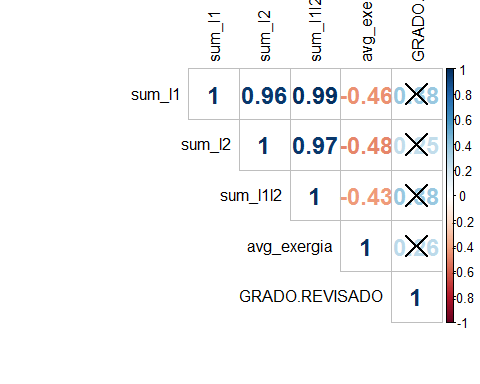
# VALIDACION EXTERNA

## Correlacion [L1, L2, L1+L2 vs. COPLAS]

Ya que es dificil explicar que una variable es dependiente de otra. La mejor forma de validar el modelo (con datos reales) es haciendo una correlacion. Con este analisis sabremos si existe relacion o no entre la variable Instar y COPLAS. Sabremos tambien la intensidad de esta relacion y si es significativa. Como nuestros datos no siguen una distribucion normal, haremos una correlacion con datos no parametricos: 'Spearman'

Funcion en R (pedir a Antonio este codigo) que plotea la relacion entre ellos con un circulo (cuanto mas relacionados, mas grande el circulo) y si es negativa o positiva. Da tambien el p valor de la correlacion. *"Puedo tener fuerte relacion pero el coef no es diferente de 0, a lo mejor con mas datos si."*"

Deberiamos esperar que haya mas relacion entre coplas y exergia, luego coplas y larvas. Cuando hay relacion negativa es que es inversa. Y como no sabemos quien es x e y, quiere decir que cuando una aumenta, otra disminuye, peor no sabemos quien aumente y quien disminuye.



* La correlacion de exergia y coplas resulta ser menos fiable de los esperado. Posiblemente porque su modelizacion es simple.
* La correlacion frente a COPLAS de L1+L2 y L1 son iguales. Esto quiere decir que L2 no aporta mucho (o nada) cuando se agregan. Esto probablemente sea porque la agregacion de las larvas indica la carga defoliadora en el sistema y al hacer esta agregacion (pobre) se le da la misma capacidad defoliadora a L1 que a L2, cuando en realidad L2 tiene mas.

# VALIDACION INTERNA

## ACP

Autocorrelation plot. Nos ayudara a saber si existen ciclos dentro de toda la serie temporal. ESTE ANALISIS SOLO SE HACE CON LOS DATOS DE BAZA

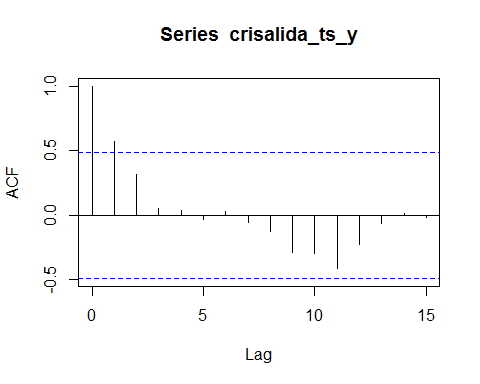
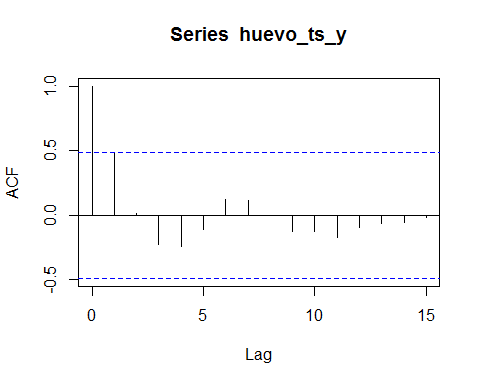
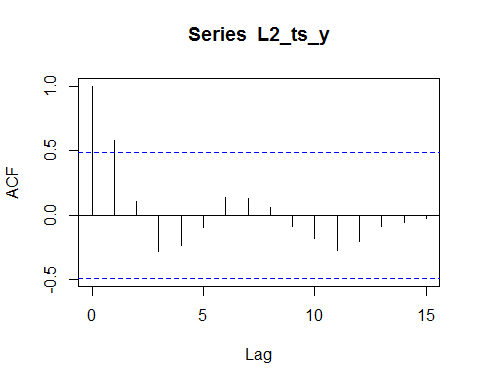
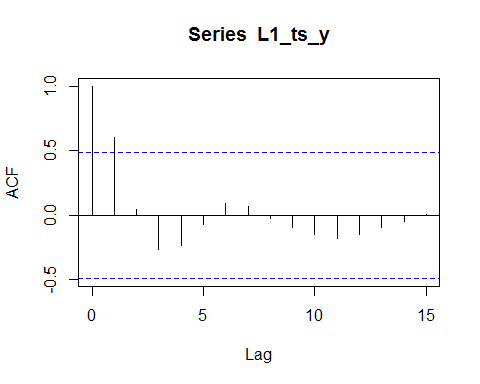
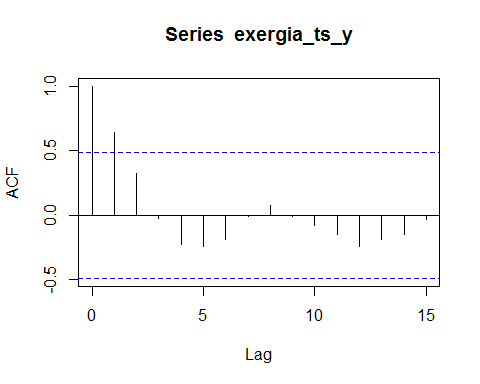
"*Those plots are showing you the correlation of the series with itself, lagged by x time unitscorrelation of the series with itself, lagged by x time units. So imagine taking your time series of length TT, copying it, and deleting the first observation of copy#1 and the last observation of copy#2. Now you have two series of length T for which you calculate a correlation coefficient. This is the value of of the vertical axis at x=1 in your plots. It represents the correlation of the series lagged by one time unit. You go on and do this for all possible time lags xx and this defines the plot*.

*The answer to your question of what is needed to report a pattern is dependent on what pattern you would like to report. But quantitatively speaking, you have exactly what I just described: the correlation coefficient at different lags of the series. You can extract these numerical values by issuing the command acf(x.ts,100)$acf*.

*In terms of what lag to use, this is again a matter of context. It is often the case that there will be specific lags of interest. Say, for example, you may believe the fish species migrates to and from an area every ~30 days. This may lead you to hypothesize a correlation in the time series at lags of 30. In this case, you would have support for your hypothesis*"

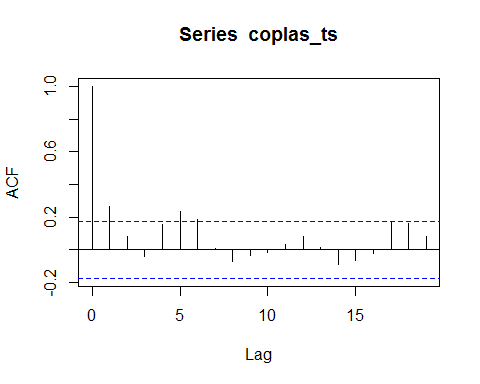
"*The blue lines give the values beyond which the autocorrelations are (statistically) significantly different from zero*"

Hay que definir el LAG. lag.max= Lo hemos definido como el numero de biociclos.

**CORRELOGRAMAS INSTAR** 

El dato de un anio esta significativamente correlacionado positivamente con el anio anterior. Es decir, el dato de la variable X del anio 't' depende del anio 'anterior't-1' de formar que cuanto mas alto el valor de la variable en 't-1', mayor es en 'este't' anio.

En exergia, por ejemplo, los Lags 3, 4, 5 indican correlacion negativa (no significativo). Quiere decir que lo que pase en 't' esta negativamente correlacionado con't-3', cuanto mayor la variable en 't-3', menos sera en 't'. Esto puede indicar densodependencia, es decir, regulacion por densidad. Una pena que no sea significativo...

**CORRELOGRAMA COPLAS** 

Los Lags 1, 5, 6 y 17 son significativos. ?? La de 5 y 6 podria referirse al ciclo del que hablan en la literatura.

## Analisis de tendencia

Se eligen indicadores de cada variable para calcular su tendencia.

En las variables huevo, L1, L2 y crisalida:

* Fecha del maximo. *HECHO*
* Fecha del primer dia distinto a cero. *EN PROCESO*
* Fecha del ultimo dia distinto a cero. *EN PROCESO*

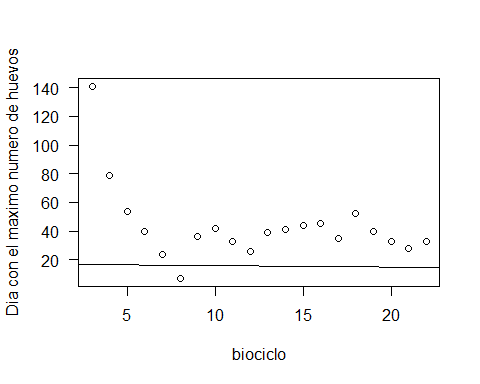
En la variable exergia:

* Fecha del maximo.*HECHO*
* Fecha del minimo. *HECHO*

**TENDENCIA EN HUEVOS**

*DIA DE PICO MAXIMO*

##   
## Call:  
## lm(formula = max\_huevo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.9192 -4.4557 -0.1653 4.3134 8.5097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.57316 2.38968 6.935 1.76e-06 \*\*\*  
## dia\_max\_huevo -0.09342 0.04698 -1.988 0.0622 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.504 on 18 degrees of freedom  
## Multiple R-squared: 0.1801, Adjusted R-squared: 0.1345   
## F-statistic: 3.954 on 1 and 18 DF, p-value: 0.06219



Tendencia negativa pero muy muy cercana a cero en el dia del valos maximo (conforme pasa el tiempo, se adelanta la fecha del pico maximo) con una significancia de 0.06. No es del todo significativo pero casi. Esto podria deberse a los altos valores de los primeros años (biociclo 3 y 4= años 1994 y 1995), que se podrian considerar dentro del periodo 'calentamiento del modelo'. Tal vez, podriamos decir que el modelo en el tiempo se va 'adecuando' a las fechas reales, si es que la fecha real del pico max de huevo es sobre el dia 40 (mediados de febrero).

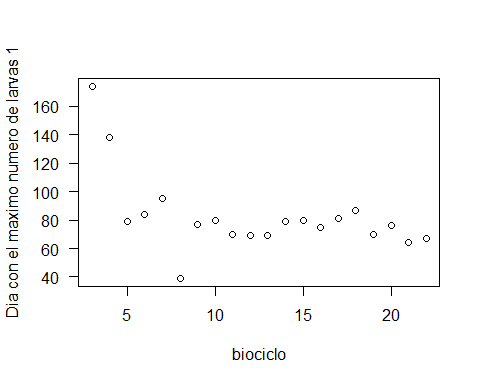
*PRIMER DIA*

*ULTIMO DIA*

**TENDENCIA EN L1**

*DIA DE PICO MAXIMO*

##   
## Call:  
## lm(formula = max\_l1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.1933 -3.1720 -0.3228 4.5269 7.8173   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.38662 3.72290 5.745 1.91e-05 \*\*\*  
## yday -0.10752 0.04276 -2.514 0.0217 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.229 on 18 degrees of freedom  
## Multiple R-squared: 0.2599, Adjusted R-squared: 0.2188   
## F-statistic: 6.321 on 1 and 18 DF, p-value: 0.02166



Pasa lo mismo que con el pico max de huevos pero con mayor significancia

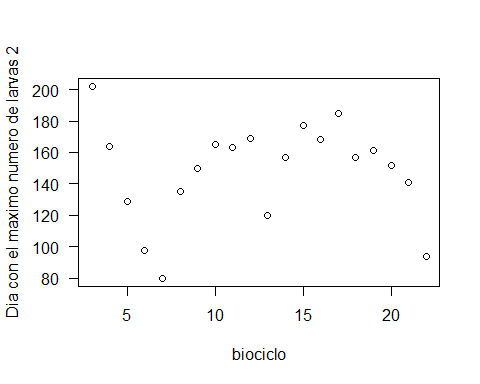
*PRIMER DIA*

*ULTIMO DIA*

**TENDENCIA EN L2**

*DIA DE PICO MAXIMO*

##   
## Call:  
## lm(formula = max\_l2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.6988 -4.6496 0.0143 4.6401 9.7014   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.950364 6.768036 1.766 0.0944 .  
## yday 0.003705 0.044693 0.083 0.9348   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.077 on 18 degrees of freedom  
## Multiple R-squared: 0.0003816, Adjusted R-squared: -0.05515   
## F-statistic: 0.006872 on 1 and 18 DF, p-value: 0.9348



La tendencia que marca es muy cercana a 0, pero esto no es significativo.

*PRIMER DIA*

*ULTIMO DIA*

**TENDENCIA EN CRISALIDA**

*DIA DE PICO MAXIMO*

DUDA: Hay varios dias en cada biociclo con el mismo numero de crisalidas. He intentado a quitar duplicados, pero claro, con que dia me quedo?

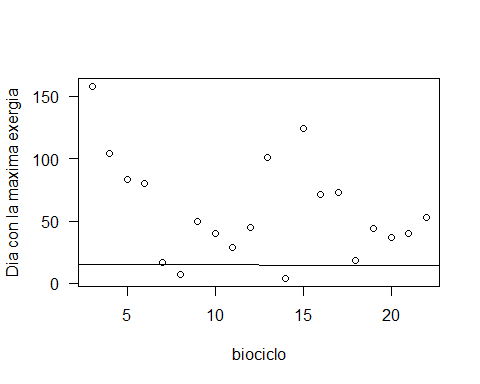
*PRIMER DIA*

*ULTIMO DIA*

**TENDENCIA EN EXERGIA**

*DIA DLE PICO MAXIMO*

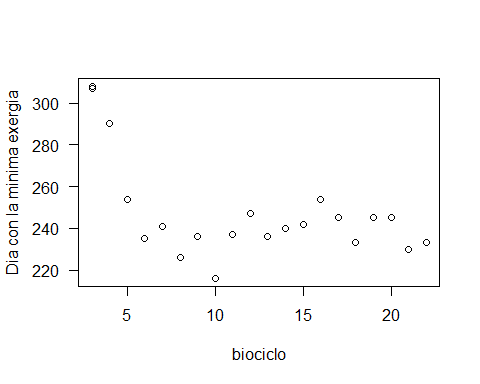
##   
## Call:  
## lm(formula = max\_exergia)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.588 -4.784 -1.214 5.338 9.206   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.43457 2.30020 6.710 2.72e-06 \*\*\*  
## yday -0.04982 0.03247 -1.534 0.142   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.716 on 18 degrees of freedom  
## Multiple R-squared: 0.1157, Adjusted R-squared: 0.06655   
## F-statistic: 2.355 on 1 and 18 DF, p-value: 0.1423



Tendencia negativa muy cercana a cero, pero sin significancia. Esto tampoco quiere decir que no hay tendencia (slope=0)

*DIA DEL PICO MINIMO*

##   
## Call:  
## lm(formula = min\_exergia)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.7588 -4.6232 -0.6232 4.5972 7.9700   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.62621 12.03194 3.792 0.00123 \*\*  
## yday -0.13561 0.04837 -2.804 0.01133 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.287 on 19 degrees of freedom  
## Multiple R-squared: 0.2926, Adjusted R-squared: 0.2554   
## F-statistic: 7.861 on 1 and 19 DF, p-value: 0.01133



Tendencia leve negativa y muy significativa.