# TEMA 5

MÓDULO:
TÉCNICAS AVANZADAS DE PREDICCIÓN

# INTRODUCCIÓN A MÉTODOS TEMPORALES Y ESPACIALES


Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

# ÍNDICE

# OBJETIVOS ESPECÍFICOS

- Entender cómo se rompen las hipótesis/supuestos iniciales del modelo lineal, con datos con retardos espaciales o temporales.
- Plantear un problema temporal o espacial y el tipo de predicción que podemos hacer con estos modelos.

# 1. INTRODUCCIÓN A MÉTODOS TEMPORALES Y ESPACIALES

## 1.1 INTRODUCCIÓN A MÉTODOS TEMPORALES

En esta subsección, vamos a adentrarnos, brevemente, en el mundo de las series temporales. Dispondremos de una base de datos, como antes, en la que vamos a encontrar nuestra variable respuesta. La única diferencia con respecto a lo que hemos visto anteriormente es que las observaciones sucesivas no son independientes entre ellas, es decir, el valor que toma nuestra variable, ahora, va a depender, entre otras cosas, de valores pasados que ha tomado.

Podemos dividir las series temporales en:

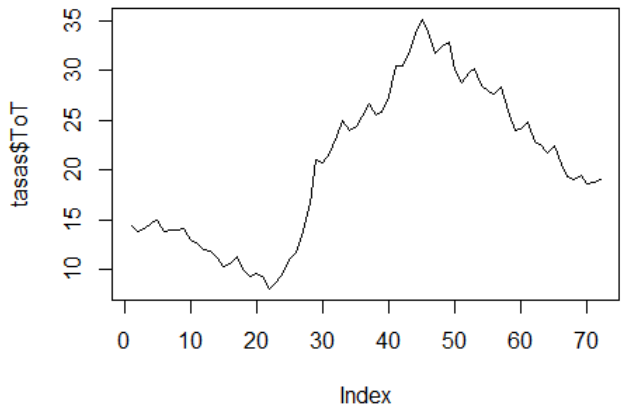
- **Series estacionarias:** la media y la varianza de la variable respuesta son constantes a lo largo del tiempo.
- **Series no estacionarias:** la media y la varianza de la variable respuesta no son constantes a lo largo del tiempo.

### ACTIVIDAD

#### TASAS DE PARO

Contamos con los datos procedentes del INE con las tasas de paro por trimestre:

```
```{r results='asis', size="small"}
tasas<-read.csv("./Data/table_5.06.csv",sep=",")[,-1]
inicio<-2000
fin<-2019
tasas<-dplyr::filter(tasas,Edad=="De 25 a 29 años",year<=fin,year>=inicio)
plot(tasas$ToT,type="l")
```
```



**Objetivos:**

- Hacer un análisis del paro en jóvenes.
- Hacer una descomposición de la serie temporal.
- Lograr hacer predicciones con un modelo de serie temporal.

Antes que nada, preguntémosnos **¿para qué queremos usar series temporales dentro de la modelización que hemos estado viendo?**

Pueden ser dos los motivos:

- El primero de ellos, nuestro interés en conocer o estimar la tasa de paro desde el punto de vista económico, para explicar ciertos patrones o comportamientos.
- El segundo, porque puede que la variable tasa de paro sea una variable explicativa de nuestro **modelo de compras online** o del **precio de la vivienda.** Para predecir lo que ocurrirá el próximo año, tendremos que estudiar nuestra variable explicativa y saber cómo se va a comportar dentro de una lógica estadística.

Partimos de la idea de que la serie temporal la podemos descomponer en **tendencia** y en **fluctuación cíclica.** Dentro de la ciclicidad, tendremos variaciones estacionales de la serie a lo largo de un periodo de tiempo y movimientos irregulares aleatorios o producidos por fenómenos concretos (terremotos, sequías, atentados, etc).

Entonces, nuestra variable objetivo:

$$Y_t = Tendencia_t + Estacionalidad_t + Error_t$$

Realizamos nuestra primera descomposición de la serie.

**¿Cómo la descomponemos?**

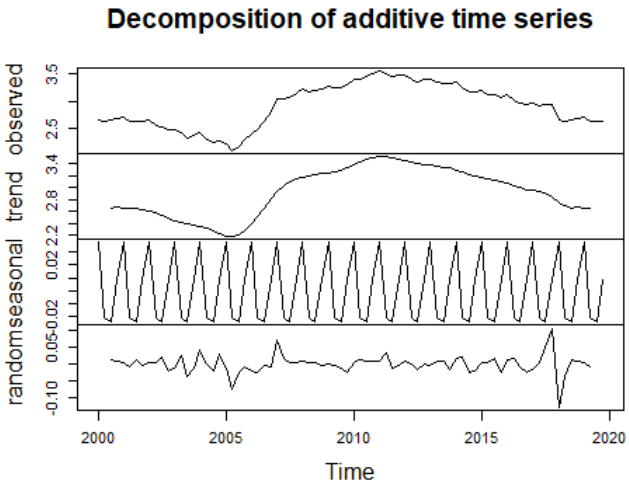
- Utilizando una media móvil para el cálculo de la tendencia.
- Una vez extraída la tendencia de la serie principal, se calcula la media de cada periodo para obtener la parte estacional.
- Por último, la parte aleatoria será la diferencia entre las dos anteriores y la serie real.

## IMPORTANTE

**¡Cuidado!, hay que verificar que la serie tenga varianza constante y sea estacionaria.**

- En el caso de que la varianza de la serie no fuese constante en el tiempo, se recomienda una transformación de la serie al igual que explicábamos en las bases de datos anteriores. Por ejemplo, tomando logaritmos sobre la serie.

```{r results='asis', size="small"}
x <- ts(tasas$ToT, start = c(inicio, 1), end = c(fin, 4), frequency = 4)
x<-log(x)
plot(decompose(x))
```




Con esto, tenemos una idea de cómo se comporta nuestra serie. En este caso, tiene una tendencia que va unida al ciclo económico y, luego, tiene una estacionalidad en la que la tasa de paro cae en los meses de primavera-verano y, luego, va cayendo en meses posteriores debido, principalmente, al sistema económico español.

**¿Cómo podemos hacer predicciones con esta información?**

Hay que recurrir a los **modelos ARIMA** los cuales se basan en el principio de que vamos a conseguir hacer una predicción de una variable en el tiempo, únicamente, con la información de su pasado. No vamos a recurrir a variables exógenas, simplemente vamos a realizar el análisis de la serie frente a sí misma en el pasado.

## MODELOS AUTOREGRESIVO (AR)

El valor de la variable depende de los "p" valores anteriores:

$$Y_t = \beta_0 + \beta_1 * Y_{t-1} + \ldots + \beta_n * Y_{t-p} + u$$

## MODELOS MEDIA MÓVIL (MA)

El valor de la variable depende de la media ponderada de las "q" perturbaciones aleatorias precedentes:

$$Y_t = \rho_0 + e_t - \rho_1 * e_{t-1} - \ldots - \rho_n * e_{t-n}$$

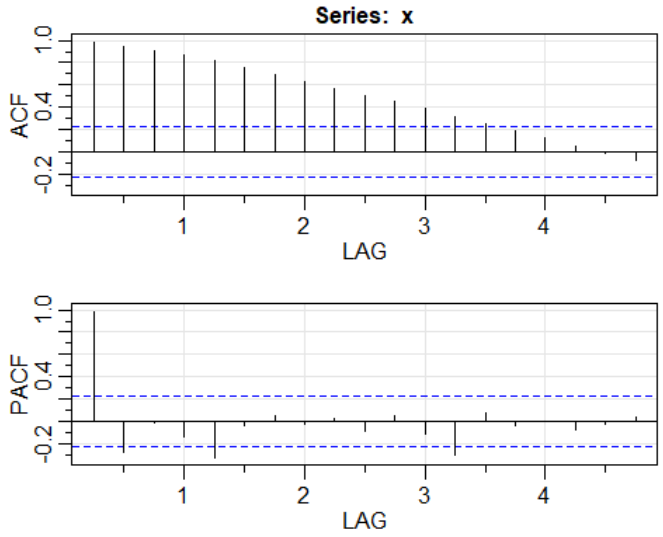
## AUTOCORRELACIÓN Y AUTOCORRELACIÓN PARCIAL

Una medida para ver, a priori, el tipo de modelo que tenemos entre manos consiste en ver la autocorrelación de las variables:

$$correlacion(time1, time2) = \frac{cov(time1, time2)}{(var(time1)var(time2))^{0.5}}$$

La correlación parcial, en lugar de tener en cuenta dos "lags" temporales, se corrige con la correlación entre los tiempos entre ellos. La única que permanecerá inalterada será la correlación de orden 1, que como no tiene tiempos entre medias, permanece equivalente:

```{r results='asis', size="small"}
auto.arima(x)
```




En la primera parte, tenemos la autocorrelación retardo a retardo y en la segunda parte, tenemos la autocorrelación parcial (corregida):

- Si la función de autocorrelación decrece rápidamente cuando incrementa el retardo, es una señal de modelo autorregresivo. Si es un modelo autorregresivo de orden p, entonces, encontraremos en la autocorrelación parcial los p primeros coeficientes distintos a cero, y el resto cero.
- En los modelos de medias móviles, la función de autocorrelación es cero para retardos superiores a q, y la parcial decrece rápidamente.

**¿Qué tenemos aquí, un AR, MA o Mix ARIMA? ¿qué tipo de retardo podemos establecer?**

Tenemos que encontrar el mejor modelo, de tal forma que se encuentre qué grado de AR y MA son los mejores (para una serie estacionaria).

**Si la serie no es estacionaria** (Media Constante), tendremos que buscar un parámetro adicional, un parámetro "d" que nos diga qué grado diferencial tendríamos que meter en la serie para corregir su media. De tal forma que, si es estacionaria la serie, solamente tendremos que ajustar AR+MA, si no es estacionaria la serie, tendremos que ajustar AR+MA+d.

ARIMA(AR=p,d=d,MA=q) -> Si hacemos algún valor cero, estaríamos ante un AR,ARMA,ARd,MA, ...

## RECUERDA

Como en cualquier modelo que hemos visto, tenemos que definir p,d,q hasta que los residuos sean normales, ruido blanco, incorrelacionados y mínimos.

**¿Da igual que tengamos datos anuales o trimestrales?**

No, cuando la frecuencia es menor a un año, tenemos más complejidad en el modelo. La detección de un comportamiento estacional es clave, ya que es posible incorporar a un modelo ARIMA (p,d,q) las correlaciones existentes entre observaciones separadas por periodos estacionales.

Daria lugar al proceso:

$$ARIMA(p,d,q) * XARIMA(P,D,Q)_{frecuecia\ anual}$$

Tendríamos que calibrar:

- p retardos del AR a la parte regular.
- q retardos del MA a la parte regular.
- d diferencias a la parte regular.
- P retardos AR a la parte estacional.
- Q retardos MA a la parte estacional.
- D diferencias a la parte estacional.

**¿Tenemos que ir a manubrio probando?**

Es una opción. Esto nos permite ir aprendiendo del proceso y de la serie, pero hay otra forma, se trata del **Paquete Auto-Arima.** Igual que el Stepwise que vimos, pero en series temporales:

```
```{r results='asis', size="small"}
auto.arima(x)
```
```

```
Series: x
ARIMA(1,1,0)(1,0,1)[4]

Coefficients:
         ar1    sar1     sma1
      0.4049  0.9380  -0.7461
s.e.  0.1107  0.0817   0.1788

sigma^2 estimated as 0.003701:  log likelihood=109.45
AIC=-210.9   AICc=-210.36   BIC=-201.43
```

El modelo que minimizaría el AIC sería un modelo con un autoregresivo de orden 1 y, en la parte estacional, una media móvil sobre los residuos con 1 retardo. Con este modelo, ya podemos hacer una evaluación del paro previsible para el próximo año dada nuestra serie histórica.

Hay diferentes metodologías para estimar series temporales. Esta es la original y la que lleva la esencia de las series temporales. Metodologías ligeramente más avanzadas proponen otro tipo de soluciones, como incorporar variables explicativas adicionales a valores de la propia serie a regresionar. También es bastante frecuente, encontrar, últimamente, series temporales modelizadas con redes neuronales.

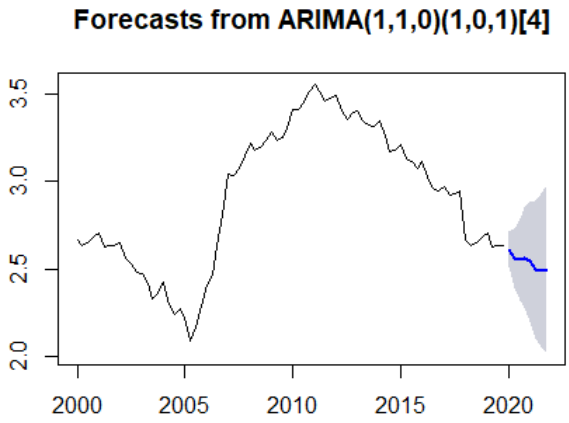
## SERIES VS REDES

**¿Cuáles serán las tasas de desempleo más probables para el próximo año?**

Tendríamos que utilizar el modelo, anteriormente, calibrado o el ajustado manualmente para hacer la

predicción correcta. El proceso es similar al que hacíamos con los glm. Multiplicando los coeficientes por el regresor elegido obtendremos una predicción:

```
```{r results='asis', size="small"}
futurval <- forecast(auto.arima(x),  level=c(90),8)
plot(futurval)
```
```




Aun teniendo un intervalo de confianza bastante amplio, analizando la serie temporal y lo vivido hasta el cierre de los datos, la serie estima que durante 2020 el paro en edades jóvenes debiera seguir disminuyendo.

Podéis imaginaros que ni este ni ningún modelo estadístico podría haber hecho una valoración teniendo en cuenta los efectos de una crisis sanitaria.

## RECUERDA

Los modelos estadísticos calcularán predicciones en base a lo vivido. Si un evento no ha ocurrido anteriormente, necesitaríamos muchas hipótesis y supuestos para poder hacer una valoración de su impacto. No sería posible hacer una valoración con un modelo estadístico tal y como los estamos viendo, puesto que nuestros datos observados no contemplan ninguna crisis de este tipo.

Un supuesto podría ser que esta crisis actual va a tener unos efectos similares a la de 2008, entonces nuestras predicciones sí que podrían hacerse.

## PIENSA UN MINUTO

¿Deberíamos incluir lo ocurrido este año 2020 para extrapolar e inferir lo que ocurrirá en 2021 o 2022?

Dependerá de la persona a la que se le pregunte. En mi opinión, no debiera introducirse, puesto que los datos estarían totalmente impactados por la crisis sanitaria y se deberían hacer ajustes sobre los datos para poder utilizarlos.

## 1.2 INTRODUCCIÓN A MÉTODOS ESPACIALES

La última pregunta que nos hacíamos, anteriormente, era:

**Cuando hemos testado la autodependencia de los residuos, lo hemos hecho con un entorno estático o bien visto a lo largo del tiempo ¿y si los residuos están relacionados espacialmente?**

Volvemos a nuestra primera base de datos de compras online. Con esta base de datos, comenzábamos a aprender el mundo de las regresiones lineales y de la inferencia. Aprendimos entre otras cosas que no podemos dar por bueno un modelo hasta estar convencidos de que las hipótesis, previa resolución (estimación de las betas), son correctas.

**¿Para qué hacíamos esto?**

**Autocorrelación Temporal ->** No queríamos que nuestro modelo tuviese autocorrelación temporal puesto que estaríamos diciendo que nuestro modelo vale para este año, sin embargo, para años venideros no va a funcionar puesto que el error depende del tiempo.

### RECUERDA

Si aparece autocorrelación en los residuos, o residuos no independientes, además de anular matemáticamente la optimización que hemos hecho, puede traer graves consecuencias para nuestro negocio. Podemos hacer campañas comerciales o tarifas específicas para un segmento de personas que hubiese funcionado el año pasado, pero que el año que viene no va a funcionar.

Gracias a nuestra destreza modelizando, esto no ocurría y, de acuerdo a los test, todos los test de residuos estaban bien posicionados.

**¡Cuidado! si cambiamos ligeramente el párrafo anterior, vamos a ver lo que pasa.**

**Autocorrelación Espacial ->** No queríamos que nuestro modelo tuviese autocorrelación espacial puesto que estaríamos diciendo que nuestro modelo vale para una zona, sin embargo, para otras zonas no va a funcionar puesto que el error depende del espacio.

### SABÍAS QUE...

Si aparece autocorrelación espacial en los residuos, o residuos no independientes, además de anular matemáticamente la optimización que hemos hecho, puede traer graves consecuencias para nuestro negocio. Podemos hacer campañas comerciales o tarifas específicas para un segmento de personas que hubiesen funcionado en una zona, pero en otra zona no va a funcionar.

### PIENSA UN MINUTO

**¿Tendría sentido combinar ambos párrafos en uno?**

Por supuesto, es un tema algo más avanzado que no incorporaré al temario, pero podéis encontrar modelos de corte trasversal espacial o lo que es lo mismo modelos espacio-temporales en la literatura estadística.

## PARA SABER MÁS

Te recomendamos la siguiente lectura:
Elhorst, J. P. (2014). Spatial econometrics: from cross-sectional data to spatial panels (Vol. 479, p. 480). Heidelberg: Springer.

**¿Qué es la dependencia espacial?**

- La dependencia espacial es el grado de asociación/correlación entre observaciones próximas entre sí.
- Existencia de una relación en un punto del espacio y lo que ocurre en otro lugar.
- El valor de una variable depende del valor de sus vecinos.
- Coincidencia de valores altos/bajos en un lugar.

**¿Quién es vecino de quién?**

Matriz de vecindad = W


*Fuente www.scielo.org.co*

Simplemente, vamos a indicar dentro de una matriz:

**1:** El individuo i y el individuo j son vecinos.
**0:** El individuo i y el individuo j no son vecinos.

Dependiendo del criterio de adyacencia, la matriz será simétrica o no.

## SABÍAS QUE...

Existen todas las combinaciones de vecindad que queráis meter. Únicamente, tened en mente cosas que tengan sentido estadístico o que puedan meterse en una matriz nxn.

**¿Hay alguna forma de medir la dependencia espacial?**

Con el **test de I-Moran:**

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \overline{x})(x_j - \overline{x})}{\sum_i (x_i - \overline{x})^2}$$

donde W es la adyacencia entre Xi y Xj, la matriz W forma la matriz de pesos espaciales.

El test I-Moran tomará valores entre -1 y 1:

- 1 querrá decir que tenemos dependencia espacial positiva.
- -1 querrá decir que tenemos dependencia espacial negativa.
- 0 indicará que no hay dependencia espacial.

**¿Tenemos dependencia espacial en los residuos de nuestro modelo de compras online?**

Vamos a comprobarlo con R. Vamos a probar seleccionando 5 vecinos y 10 vecinos de cada persona, por cercanía:

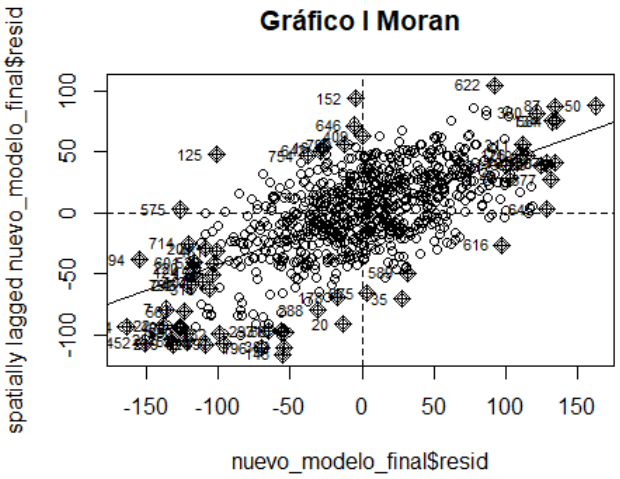
```
```{r results='asis', size="small",warning=FALSE,message=FALSE}
nb <- knn2nb(knearneigh(cbind(tabla$LONG, tabla$LAT), k=10))

moran.test(x = nuevo_modelo_final$resid, listw = nb2listw(nb, style="W"))
moran.plot(x = nuevo_modelo_final$resid, listw = nb2listw(nb, style="W"),main="Gráfico I Moran")
```
```

```
        Moran I test under randomisation

data:  nuevo_modelo_final$resid
weights: nb2listw(nb, style = "W")

Moran I statistic standard deviate = 28.451, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.4256907434     -0.0012406948      0.0002251804
```




Solamente, muestro 10 vecinos. Para 5 lo podéis hacer vosotros.

**¿Qué vemos?**

Vemos que el I-Moran es altamente significativo (p-valor), lo cual es mal síntoma. Hay dependencia espacial en el modelo que habíamos valorado como bueno, y nadie nos había dicho nada.

Esto significa que, por muy bueno que fuese nuestro modelo, no está teniendo en cuenta el tema de la cercanía entre personas. Quiere decir que, si alguien de mi entorno compra online, yo voy a estar altamente influido para comprar online y si alguien de mi entorno deja de comprar en esta web, yo voy a estar influido para dejar de comprar. Y todo esto, tan simple como es contarlo, lo estamos olvidando en la modelización.

Lo que nos está sacando el I-Moran es que los residuos están interconectados: los residuos altos o positivos están cerca, geográficamente, de los altos, y los bajos están cerca, geográficamente, de los bajos:

- Valores Altos del gasto se encuentran cerca de otros que tienen valores altos.
- Valores Bajos del gasto se encuentran cerca de otros que tienen valores bajos.

**Muy Importante: interpretación ¿Por qué puede estar sucediendo?**

- **La geolocalización puede estar enmascarando otras variables relevantes.** Puede ser que la zona superior sea una zona rica, y la parte de la derecha una parte deprimida, económicamente.
- **Puede haber efecto llamada.** La gente habla entre sí, y habla sobre precios y productos. Puede que ciertas personas se influyan entre ellos en el precio que pagarían por un determinado producto.
- Puede ser que haya factores geográficos muy distintos por ubicación. Puede que el número de centros comerciales en un sitio y en otro sea muy distinto y esto influya en el precio de búsqueda y efectivo que pagaría una persona.

**¿Sabemos si están concentrados estos casos, en una zona en concreto del mapa?**

No. Lo que nos está diciendo es que, en todo el mapa, está pasando esto. Quiere decir que todas las personas de nuestra base de datos están influenciadas por gente de su alrededor tanto positiva como negativamente.

Para ver, si además de esto, hay alguna zona en el mapa que presenta un alto grado de dependencia espacial, es decir, una dependencia espacial local, podemos llamar al **test LISA** que es equivalente al I-Moran pero lo vamos a hacer a nivel regiones:

```
```{r results='asis', size="small",warning=FALSE,message=FALSE}
nb <- knn2nb(knearneigh(cbind(tabla$LONG, tabla$LAT), k=10))
imoranlocal<-as.data.frame(localmoran(x = nuevo_modelo_final$resid, listw = nb2listw(nb, style="W")))
tabla$registo<-1
#pl_pt(tabla,color2 = imoranlocal$Z.Ii,size2 =tabla$registo ,dd = 6)
```
```

Lo que saca el test I-Moran es el valor del estadístico para cada una de las zonas.

Nos interesan aquellos valores del estadístico que sean suficientemente grandes como para decir que es significativo. Por los datos que extrae, centrándonos en aquellos valores en color rojo, podemos ver cierta dependencia espacial en la zona sur-oeste.

Aunque mi única intención es alimentar vuestra curiosidad por este tipo de modelos y ver las posibilidades que plantean sin llegar a un análisis muy detallado de la materia. Hay que decir que, a veces, la dependencia espacial confunde la heterocedasticidad espacial.

**¿Qué quiere decir esto?**

Antes estábamos comentando que el problema planteaba una dependencia espacial severa. En algunos casos, puede que la dependencia espacial se confunda con heterocedasticidad espacial, o puede que convivan autocorrelación espacial y heterocedasticidad espacial. La heterocedasticidad espacial es la diferencia en la varianza del error a lo largo del mapa.

En el caso que estamos analizando, parece que, además de existir una dependencia espacial en el mapa, algo de heterocedasticidad espacial hay también. (Foco Sur-Oeste de Madrid).

**¿Qué hacemos ante tal problema?**

Defino una serie de opciones que tenemos para ir venciendo a la dependencia espacial:

1. Incorporo más variables relacionadas con el espacio. Ejemplo: renta por municipios, distancia a principales redes de carreteras, densidad de población, etc.

   Check I-Moran.

2. Defino clusters espaciales y los incorporo en el modelo estadístico. Ejemplo: ciertos barrios tienen propensión a comprar mucho, entonces, creo una variable que, para estos barrios, tome valor 1 y para el resto 0. Algoritmos interesantes para proponer un cluster espacial son Satscan o modelos GWR.

   Check I-Moran.

3. Cambiar nuestro GLM, por un GLMSpacial.

**Spatial Autorregresive Model**

Lo definimos como:

$$Y = X\beta + \rho WY + u$$

donde W es la matriz de pesos espaciales.

**¿Qué quiere decir este modelo?**

Quiere decir que la Y se explica con las variables exógenas, como siempre, pero hay un factor más que es “rho” que es el impacto “boca a boca”. Esto quiere decir que las personas están impactadas por lo que sucede a su alrededor.

Lo resolvemos:

$$Y = (I - \rho W)^{-1}(X\beta + u)$$

La estimación de las betas la realizamos maximizando la verosimilitud:

```r
```{r results='asis', size="small",warning=FALSE,message=FALSE}
nb <- knn2nb(knearneigh(cbind(tabla$LONG, tabla$LAT), k=10))

formula<-as.formula('COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD_hasta_57 + EDAD_despues_57 + GENERO + Log_Ing +
Dens_h')

nuevo_modelo_final<-glm(formula = formula,data =tabla,family=gaussian)
modelo_espacial_sar <- lagsarlm(formula = formula,data=tabla, listw = nb2listw(nb, style="W"))
summary(modelo_espacial_sar)

paste("residuos modelo GLM",sum((nuevo_modelo_final$resid)**2))
paste("residuos modelo GLMEspacial",sum((modelo_espacial_sar$residuals)**2))
```
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-134.8161  -29.3394    0.3429   28.2863  134.1911

Type: lag
Coefficients: (asymptotic standard errors)
                Estimate Std. Error  z value  Pr(>|z|)
(Intercept)   131.11441   23.39513   5.6043 2.090e-08
Dist_Min        9.87270    1.08868   9.0685 < 2.2e-16
ANTIGUEDAD    -14.66951    0.88420 -16.5907 < 2.2e-16
EDAD_hasta_57  -2.93478    0.13796 -21.2720 < 2.2e-16
EDAD_despues_57 -6.73829   0.37309 -18.0606 < 2.2e-16
GENERO        -24.43679    3.27979  -7.4507 9.281e-14
Log_Ing        16.76927    2.28198   7.3486 2.003e-13
Dens_h         -0.36093    0.35206  -1.0252    0.3053

Rho: 0.66523, LR test value: 334.28, p-value: < 2.22e-16
Asymptotic standard error: 0.032552
    z-value: 20.436, p-value: < 2.22e-16
Wald statistic: 417.63, p-value: < 2.22e-16

Log likelihood: -4198.28 for lag model
ML residual variance (sigma squared): 1828.9, (sigma: 42.766)
Number of observations: 807
Number of parameters estimated: 10
AIC: 8416.6, (AIC for lm: 8748.8)
LM test for residual autocorrelation
test value: 0.73108, p-value: 0.39253

[1] "residuos modelo GLM 2360375.66615741"
[1] "residuos modelo GLMEspacial 1475913.46349477"
```

Como podemos ver, el modelo rebaja mucho el error medio del modelo. Es algo realmente interesante porque nos acercamos al "verdadero" valor de las betas.

Desmigando los datos y conociendo este tipo de técnicas podemos acercarnos a la realidad.

En este **modelo SAR,** la variable dependiente está autocorrelacionada, espacialmente hablando. Esto quiere decir que el valor que toma una determinada variable está influenciado por el valor que toman sus vecinos. Por eso, tenemos que introducir, en la ecuación, la matriz de pesos espaciales y la respuesta autocorrelacionada.

**Probablemente, a estas alturas, ya casi finalizando el tema os preguntaréis: ¿Cómo puedo saber qué modelo tengo que utilizar en cada momento?**

Aunque se han desarrollado algoritmos que te dan una indicación del tipo de modelo como función link, distribución a elegir, etc., la realidad es que, el mejor aliado para conocer el verdadero modelo, es el conocimiento de la materia. La intuición basada en experiencia es lo que nos hace llegar a modelos cada vez mejores.

IMPORTANTE

**Tenéis que dedicar tiempo a los datos, a conocer el problema y, antes de poneros a programar, y a probar algoritmos, hay que pensar en qué tipo de solución podemos darle a dicho problema.** Probablemente, después de un tiempo de reflexión, no va a salir el mejor modelo a la primera, sino que habrá que probar con varios, pero llevaremos mucho terreno ganado.

**Spatial Error Model**

Lo definimos como:

$$Y = X\beta + e$$
$$e = \rho We + \epsilon$$

donde W es la matriz de pesos espaciales.

**¿Qué quiere decir este modelo?**

Quiere decir que el error lleva implícito una estructura espacial. La existencia de factores o variables no considerados en la especificación del modelo trasladan la dependencia espacial al término de error.

Lo resolvemos:

$$Y = X\beta + (I - \rho W)^{-1}\epsilon$$

La estimación de las betas la realizamos maximizando la verosimilitud:

```{r results='asis', size="small",warning=FALSE,message=FALSE}
nb <- knn2nb(knearneigh(cbind(tabla$LONG, tabla$LAT), k=10))

formula<-as.formula('COMPRAS ~ Dist_Min + ANTIGUEDAD + EDAD_hasta_57 + EDAD_despues_57 + GENERO + Log_Ing + Dens_h')

nuevo_modelo_final<-glm(formula = formula,data =tabla,family=gaussian)
modelo_espacial_sar <- lagsarlm(formula = formula,data=tabla, listw = nb2listw(nb, style="W"))
modelo_espacial_sem <- errorsarlm(formula = formula,data=tabla, listw = nb2listw(nb, style="W"))
summary(modelo_espacial_sem)

paste("residuos modelo GLM",sum((nuevo_modelo_final$resid)**2))
paste("residuos modelo GLMEspacial SAR",sum((modelo_espacial_sar$residuals)**2))
paste("residuos modelo GLMEspacial SEM",sum((modelo_espacial_sem$residuals)**2))
```

```
Type: error
Coefficients: (asymptotic standard errors)
               Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    454.72356   19.48512  23.3370 < 2.2e-16
Dist_Min        20.32326    2.29353   8.8611 < 2.2e-16
ANTIGUEDAD     -14.46463    0.86622 -16.6986 < 2.2e-16
EDAD_hasta_57   -2.85092    0.13180 -21.6300 < 2.2e-16
EDAD_despues_57 -6.51825    0.35689 -18.2639 < 2.2e-16
GENERO         -25.03627    3.17150  -7.8941 2.887e-15
Log_Ing         15.94709    2.16933   7.3512 1.965e-13
Dens_h          -1.40367    0.96464  -1.4551    0.1456

Lambda: 0.74984, LR test value: 334.19, p-value: < 2.22e-16
Asymptotic standard error: 0.032798
    z-value: 22.862, p-value: < 2.22e-16
Wald statistic: 522.68, p-value: < 2.22e-16

Log likelihood: -4198.326 for error model
ML residual variance (sigma squared): 1790.1, (sigma: 42.31)
Number of observations: 807
Number of parameters estimated: 10
AIC: 8416.7, (AIC for lm: 8748.8)

[1] "residuos modelo GLM 2360375.66615741"
[1] "residuos modelo GLMEspacial SAR 1475913.46349477"
[1] "residuos modelo GLMEspacial SEM 1444613.56497393"
```

Podemos ver que SAR y SEM dan un ajuste bastante parecido dados los residuos.

**¿Destruyen ambos la dependencia espacial de los residuos?**

```
```{r results='asis', size="small",warning=FALSE,message=FALSE}
nb <- knn2nb(knearneigh(cbind(tabla$LONG, tabla$LAT), k=10))
#Dependencia espacial del SAR
moran.test(x = modelo_espacial_sar$residuals, listw = nb2listw(nb, style="W"))
#Dependencia espacial del SEM
moran.test(x = modelo_espacial_sem$residuals, listw = nb2listw(nb, style="W"))
```
```

```
        Moran I test under randomisation

data:  modelo_espacial_sar$residuals
weights: nb2listw(nb, style = "W")

Moran I statistic standard deviate = 0.64108, p-value = 0.2607
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.0083791448     -0.0012406948      0.0002251706


        Moran I test under randomisation

data:  modelo_espacial_sem$residuals
weights: nb2listw(nb, style = "W")

Moran I statistic standard deviate = -1.7746, p-value = 0.962
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
    -0.0278706290     -0.0012406948      0.0002251825
```

Ambos la destruyen. Parece que el SEM es el que mejores resultados otorga.

Antes de ver el último modelo espacial que me interesaría que conozcáis, deciros que, estos modelos que hemos visto en entorno gaussiano están igualmente probados y desarrollados en entorno generalizado con función exponencial.

Evidentemente el gasto computacional para calcular estos modelos es mucho mayor, pero los avances en computación, en GIS y en formulación espacial hacen que el futuro vaya encaminado a estos modelos más completos que pueden sacar patrones más interesantes y completos.

**Modelos de Regresión geográficamente ponderados**

Para finalizar os voy a enseñar otro método interesante para ver si, aún, nos queda heterocedasticidad en el modelo, aunque lo podéis aplicar a una infinidad de problemas.

Se trata de los modelos de regresión geográficamente ponderados.

Son modelos que responden a la siguiente pregunta: **¿El efecto de una variable sobre nuestra respuesta es independiente del espacio?**

Quiere decir que la edad es un factor clave (ya lo sabíamos), pero ¿afecta de la misma manera el cambio en comportamiento del cliente en el sur que en el norte? Este tipo de algoritmos lo pone a prueba.

La idea detrás de los **modelos GWR** es la medición de la relación entre la variable dependiente y las independientes a través del espacio. En lugar de calibrar un modelo único, miraremos un modelo global a través de la combinación de las diferentes áreas geográficas. La modelización GWR calibra tantos modelos como puntos hay en nuestra base de datos. Estima un modelo por punto cogiendo los puntos

que hay a su alrededor, dando mayor importancia a los que están en el centro. La técnica GWR no está desarrollada como algoritmo econométrico puro, sino que surge y tiene más empleabilidad para suavización o interpolación de datos (véase también métodos krigging).

Pasamos de:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + u$$

a:

$$Y_s = \beta_{s1} X_1 + \cdots + \beta_{s1} X_p + u$$

Donde s es cada zona geográfica que queremos representar.

Resolvemos:

$$\beta = (X^t W_s X)^{-1} X^t W_s Y$$

Con el modelo global general, tendremos valores únicos de cada estimador para todos los puntos de nuestra muestra, asumiremos independencia espacial de los residuos y no tendrá en cuenta la distancia entre puntos a la hora de hacer una valoración.

En el modelo ponderado geográficamente, tendremos diferentes estimadores para cada una de las variables. Dependiendo del área geográfica, reduciremos o eliminaremos la dependencia espacial de los residuos y se tendrá en cuenta la distancia entre puntos a la hora de predecir.

Este tipo de modelos se genera en dos partes:

1. Primero hay que definir s. ¿Cuál es el ancho espacial óptimo para ponderar nuestro modelo? El algoritmo probará con diferentes distancias y decidirá cuál es el ancho ideal:

```r
{r results='asis', size="small",warning=FALSE,message=FALSE}
#Convierto mi base de datos en base de datos espacial
tabla$residuos<-modelo_espacial_sem$residuals
puntos_sp<-tabla
coordinates(puntos_sp)<- c("LONG","LAT")
proj4string(puntos_sp) <- CRS("+proj=longlat +datum=WGS84")
#Obtenemos el mejor BW
bw <- gwr.sel(residuos~1, data=puntos_sp)

paste("El mejor ancho de banda es:",bw)
```

```
Bandwidth: 18.21512 CV score: 1446566
Bandwidth: 29.44326 CV score: 1447597
Bandwidth: 11.27575 CV score: 1443821
Bandwidth: 6.986983 CV score: 1436759
Bandwidth: 4.336378 CV score: 1428663
Bandwidth: 2.698215 CV score: 1435177
Bandwidth: 4.730161 CV score: 1429453
Bandwidth: 4.192511 CV score: 1428504
Bandwidth: 3.621741 CV score: 1428886
Bandwidth: 4.042059 CV score: 1428433
Bandwidth: 4.008278 CV score: 1428432
Bandwidth: 4.019386 CV score: 1428432
Bandwidth: 4.019691 CV score: 1428432
Bandwidth: 4.01965 CV score: 1428432
Bandwidth: 4.01961 CV score: 1428432
Bandwidth: 4.01965 CV score: 1428432
[1] "El mejor ancho de banda es: 4.01965045240776"
```

2. Con este ancho de banda vamos a estimar el modelo.

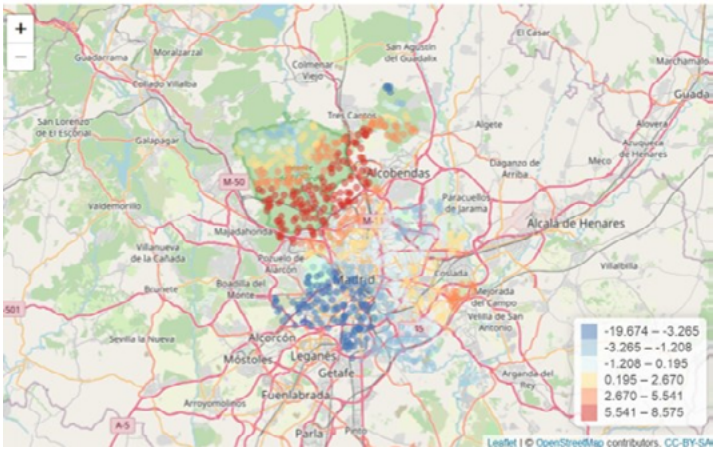
```r
```{r results='asis', size="small",warning=FALSE,message=FALSE}
#Modelizamos
g <- gwr(residuos~1, data=puntos_sp, bandwidth=bw)
```
```

**¿Qué hemos hecho?**

¿Dependen los residuos del espacio? Justamente estamos viendo si podríamos meter algún aspecto espacial adicional para completar nuestra modelización.

**Nota:** estamos, solamente, metiendo el intercept en el modelo, pero este tipo de modelos también permiten meter más variables para ver si tienen algún comportamiento espacial:

```r
```{r results='asis', size="small",warning=FALSE,message=FALSE}
tabla$intercept<-g$SDF$`(Intercept)`
#pl_pt(tabla,color2 = tabla$intercept,size2 =tabla$registo ,dd = 6)
```
```




Interesante el patrón que estamos viendo en los datos. Resulta que después de haber hecho una modelización casi perfecta, nos hemos dado cuenta de que los datos tenían dependencia espacial y, ahora, nos damos cuenta de que, incluso después de vencer la dependencia espacial, tenemos una heteroscedasticidad espacial importante.

## PIENSA UN MINUTO

**¿Y si sois directores de esta empresa? ¿Os hubieseis quedado con el SuperModelo que veíamos en el apartado 2? ¿Hasta qué punto es importante predecir bien las betas y contemplar todos estos puntos?**

Hasta el punto de que está en juego la supervivencia de la empresa.

Imaginemos que, antes de conocer las técnicas espaciales, ponemos en marcha una acción comercial por la que las personas más propensas a comprar (las que creemos que son las más propensas) van a pagar menor coste de servicio. Otra compañía (competencia) va a dar, exactamente, la misma noticia a los clientes, pero seleccionando, con este último modelo, sus clientes más propensos.

**¿Qué va a ocurrir?**

Los clientes propensos de verdad, van a irse directamente a la compañía que mejor los elige. Los menos propensos, si creen que van a ser elegidos como buenos consumidores, van a irse a la compañía que peor elige a sus clientes y así en bucle. Al final del bucle, la compañía que mejor elige a sus clientes, dado

que ha ido creciendo con clientes buenos, puede abaratar los costes puesto que trabaja con economías de escala y, dada toda la producción que tiene, puede establecer unos costes muy bajos.

# IDEAS CLAVE

- Las series temporales cambian la perspectiva del modelo corte trasversal que veníamos trabajando. En ellas, se incumple el supuesto de que las observaciones de nuestra variable respuesta son independientes. Para ello, la variable respuesta con el nuevo modelo propuesto se explica sobre retardos sobre sí misma.

- Los modelos ARIMA combinan retardos temporales sobre la variable respuesta, directamente, y sobre los residuos de la serie. El grado del retardo lo conseguiremos con conocimiento de la propia serie y con la minimización de algún indicador (AIC, R-cuadrado, BIC).

- Los modelos espaciales cambian de nuevo la perspectiva y proponen soluciones cuando la dependencia de la variable dependiente es espacial.