

¿Es posible inferir el comportamiento de un retweet?



MODELADO ESTADÍSTICO A LA COLA DE LA
DISTRIBUCIÓN DE FRECUENCIA



ANALÍTICA DIGITAL
DEP. NEGOCIO

 rlserrano@bcc.es
 [@Rafaelopez8](https://twitter.com/Rafaelopez8)



UNIVERSIDAD DE
MURCIA



¿QUIENES SOMOS?



❑ **Un equipo centrado en el valor de los datos:** Cajamar Data Lab es un **equipo de especialistas dedicado a maximizar el potencial de los datos** para ofrecer servicios y productos a nuestros clientes así como al resto de la ciudadanía.

❑ **Analítica Big Data:** Los **proyectos de Big Data** son uno de los núcleos de la estrategia de innovación del Grupo Cooperativo Cajamar.



PLANTEAMIENTO DEL PROBLEMA



¿Cómo saber si un tweet se convertirá en viral?

¿Es posible saber cuantos días se retweeteará?



¿A cuantos usuarios vas a llegar con su publicación?

PARÁMETROS ESTUDIADOS ANTERIORMENTE



Los parámetros incluidos en estudios anteriores se basan en medidas de la serie que tratan de explicar el comportamiento del tweet dependiendo de esos índices.

seguidores

días

Tamaño del
texto

Número
de tweets

Análisis
del
sentimiento

Hashtags

Menciones

ESTRATEGIA – Divide y vencerás



Temáticas

Desarrollo Rural

Sostenibilidad

Medioambiente

Refugiados

Derechos Humanos

Elecciones

Cine

Tweets

Los más retweeteados

Top 20

Es necesario considerar los tweets más retweeteados para poder analizar distintas características de la serie que no se podría estudiar sin suficientes casos.

Series Temporales

Descomposición

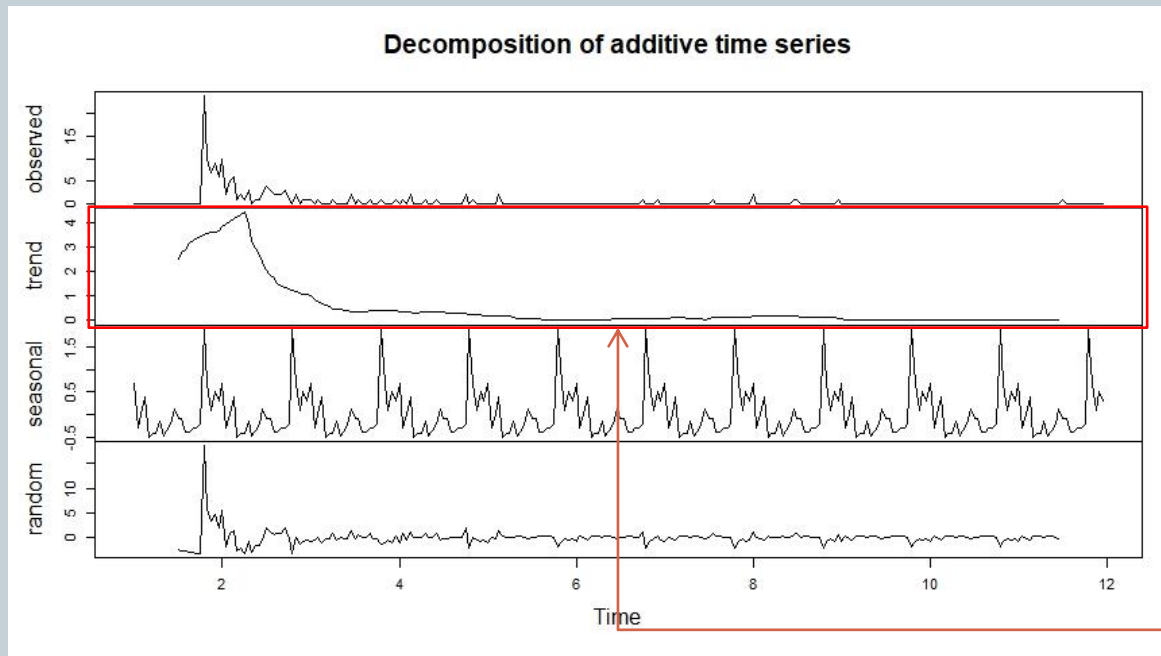
Serie de tendencia

Descomponer la serie nos aporta una visión general del comportamiento, sin interferencias entre la periodicidad y estacionalidad del evento.

DESCOMPOSICIÓN DE LA SERIE

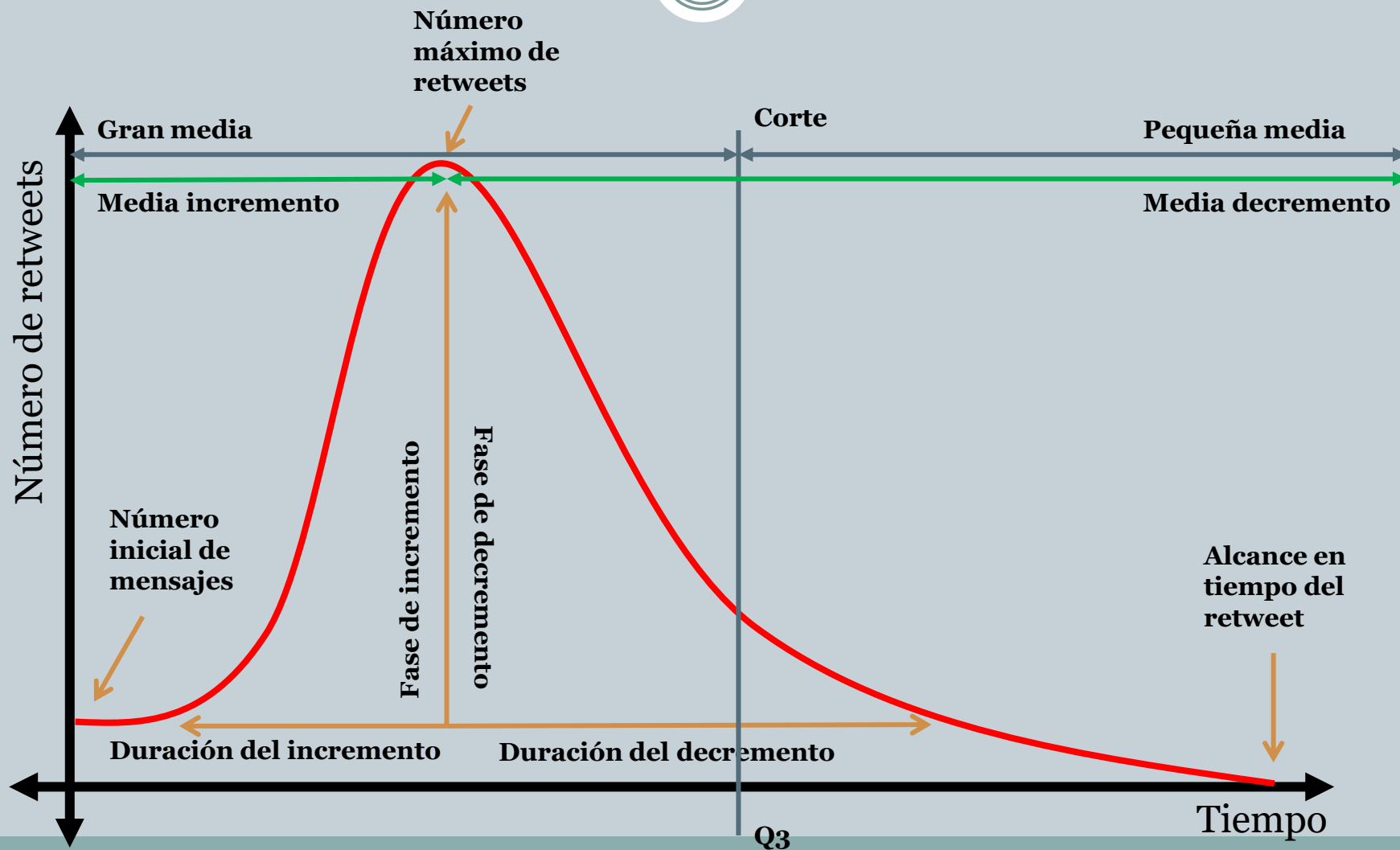


Aplicado a la distribución de frecuencia de retweet de un tweet concreto. Se descompone usando un modelo aditivo.



Descomponemos la distribución de frecuencia y nos centramos en el estudio de la serie de tendencia.

PARÁMETROS DE INTERÉS



RECOPIACIÓN DE VARIABLES



variables	descripción
Pos	orden de retweets, se ordenan del más retweeteado al menor
Maxpos	posición en que se encuentra el máximo de la serie
Maxvalor	valor del máximo de la serie
Media_incremento	promedio de los valores de la serie desde el inicio hasta el máximo
Temática	codificación de la temática que trata el tweet
Corte	posición que ocupa el Q3
Granmedia	promedio de la serie desde el inicio hasta el Q3
Ultimapos	posición del último retweet
Ndias	número de días que abarca la serie
Total_rtw	suma total de la serie de tendencia

Antes del máximo

Antes del Q3

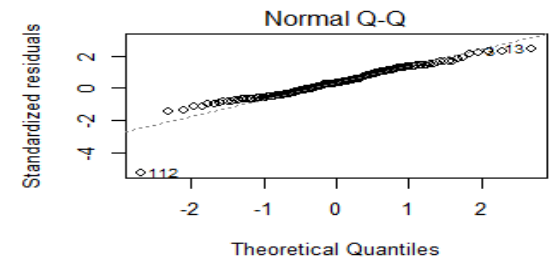
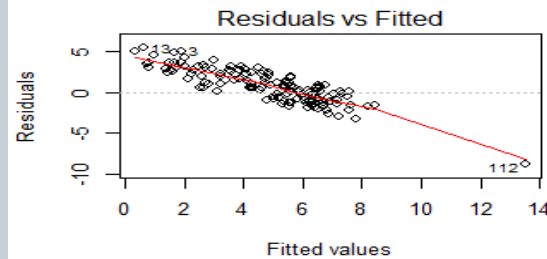
A estimar

MODELO ANTES DEL MÁXIMO



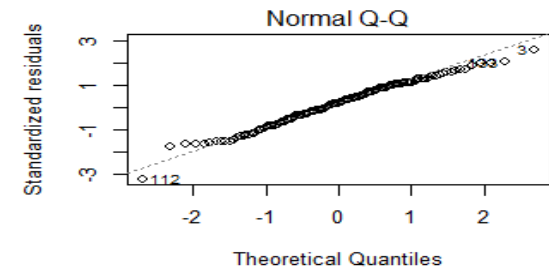
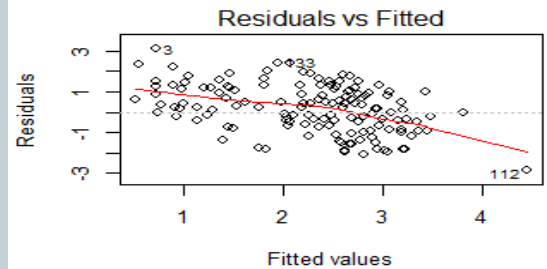
$$\ln(\text{Ultimapos}) \sim 0.08 * \text{Maxpos} + 0.86 * \ln(\text{Media_incremento}) + 0.43 * \text{Temática}$$

R2 ajustado = 0.85



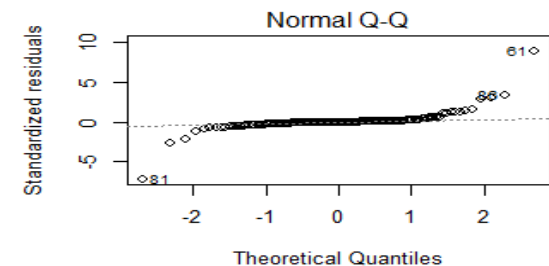
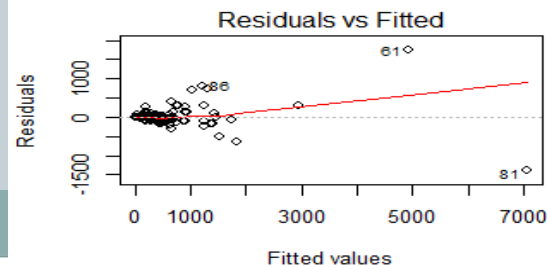
$$\ln(\text{Ndias}) \sim 0.055 * \text{Pos} + 0.027 * \text{Maxpos} + 0.618 * \ln(\text{Media_incremento})$$

R2 ajustado = 0.801



$$\text{Total_rtw} \sim 75.60 * \text{Maxvalor} - 58.28 * \text{media_incremento}$$

R2 ajustado = 0.9357

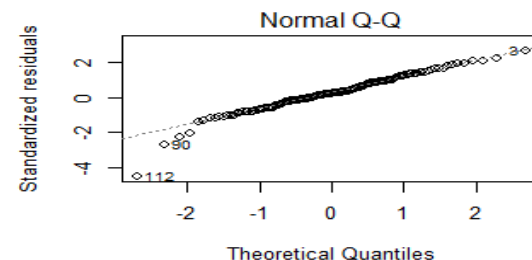
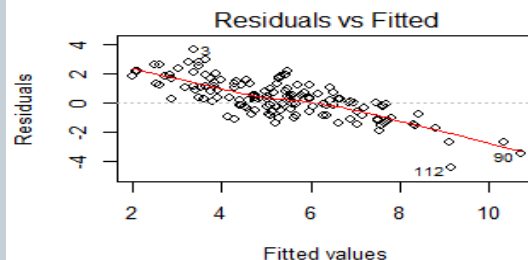


MODELO ANTES DEL Q3



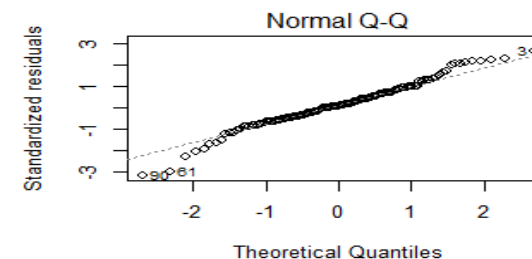
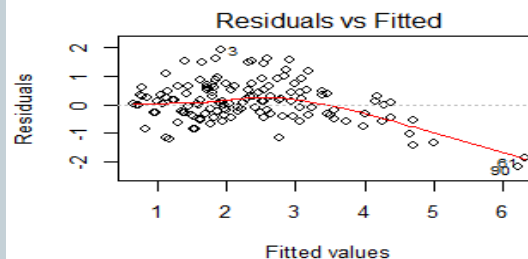
$$\ln(\text{Ultimapos}) \sim 0.154 * \text{Pos} + 0.058 * \text{Maxpos} - 0.014 * \text{Maxvalor} + 0.676 * \ln(\text{Media_incremento}) + 0.026 * \text{Corte}$$

R2 ajustado = 0.9443



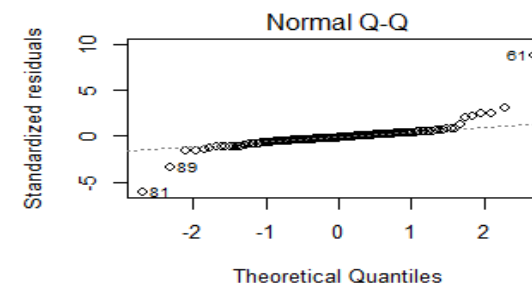
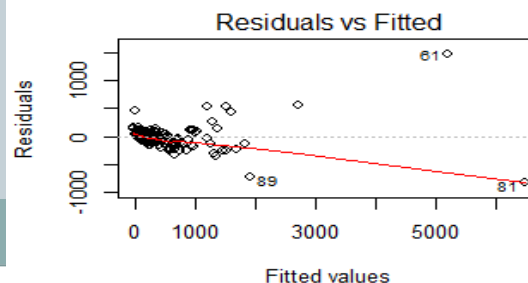
$$\ln(\text{Ndias}) \sim 0.037 * \text{Pos} + 0.018 * \text{Maxpos} - 0.010 * \text{Maxvalor} + 0.282 * \ln(\text{Media_incremento}) + 0.020 * \text{Corte}$$

R2 ajustado = 0.9307



$$\text{Total_rtw} \sim 71.92 * \text{Maxvalor} - 62.82 * \text{Media_incremento} - 38.15 * \text{temática} + 2.292 * \text{Corte} + 28.84 * \text{Granmedia}$$

R2 ajustado = 0.9507

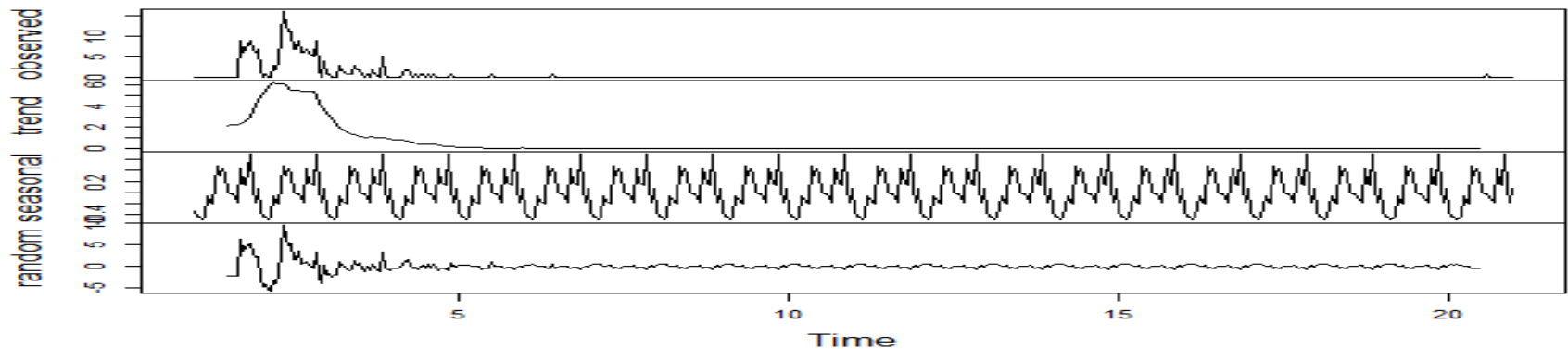


CASO PRÁCTICO



ultimapos	est_ultimapos_media	est_ultimapos_Q3
1008	4493,92	773,28
total_rtw	est_total_rtw_media	est_total_rtw_Q3
296	303,76	282,18
ndias	est_ndias_media	est_ndias_Q3
4	7,669	4,264

Decomposition of additive time series



OBSERVACIONES



- Parece que la temática tratada en el tweet influye en el modelo de predicción para algunas variables dependientes. Que no salga en todas podría ser porque tenemos pocas temáticas.
- La posición que ocupa el tweet ordenando por frecuencia de retweet se hace significativa para algunos modelos, pero no para todos.
- Se debería hacer un nuevo estudio ampliando el número de tweets estudiados.
- Los valores promedio predominan entre las variables seleccionadas como significativas para explicar la variable dependiente.

CONCLUSIONES



- Es posible predecir con cierta fiabilidad el comportamiento del retweet contemplando información en su fase inicial (antes del máximo)
- Es posible aumentar la fiabilidad de las estimaciones a costa de hacer las predicciones en una etapa más tardía (antes del Q3). Para que este caso sea útil es necesario que el periodo de estudio de retweet sea amplio.

REFERENCIAS



- Birkin, Mark and Malleson, Nicholas (2012) *Investigating the Behaviour of Twitter Users to Construct an Individual-Level Model of Metropolitan Dynamics*.
- Tan, Chenhao & Lee, Lillian & Pang, Bo. (2014). *The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter*. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 1. 10.3115/v1/P14-1017.
- Quanzhi Li, Xiaomo Liu, Rui Fang, Armineh Nourbakhsh, Sameena Shah (2016). *User Behaviors in Newsworthy Rumors: A Case Study of Twitter*. AAAI Publications, Tenth International AAAI Conference on Web and Social Media.
- Maximilian Jenders, Gjergji Kasneci, and Felix Naumann. (2013). *Analyzing and predicting viral tweets*. In Proceedings of the 22nd International Conference on World Wide Web (WWW '13 Companion). ACM, New York, NY, USA, 657-664. DOI: <https://doi.org/10.1145/2487788.2488017>.

M#CH@\$ GR@Ci@\$



Si quieres formar parte del equipo,
envía tu CV a [Trabaja con nosotros](#) de Cajamar.

 @Rafaelopez8