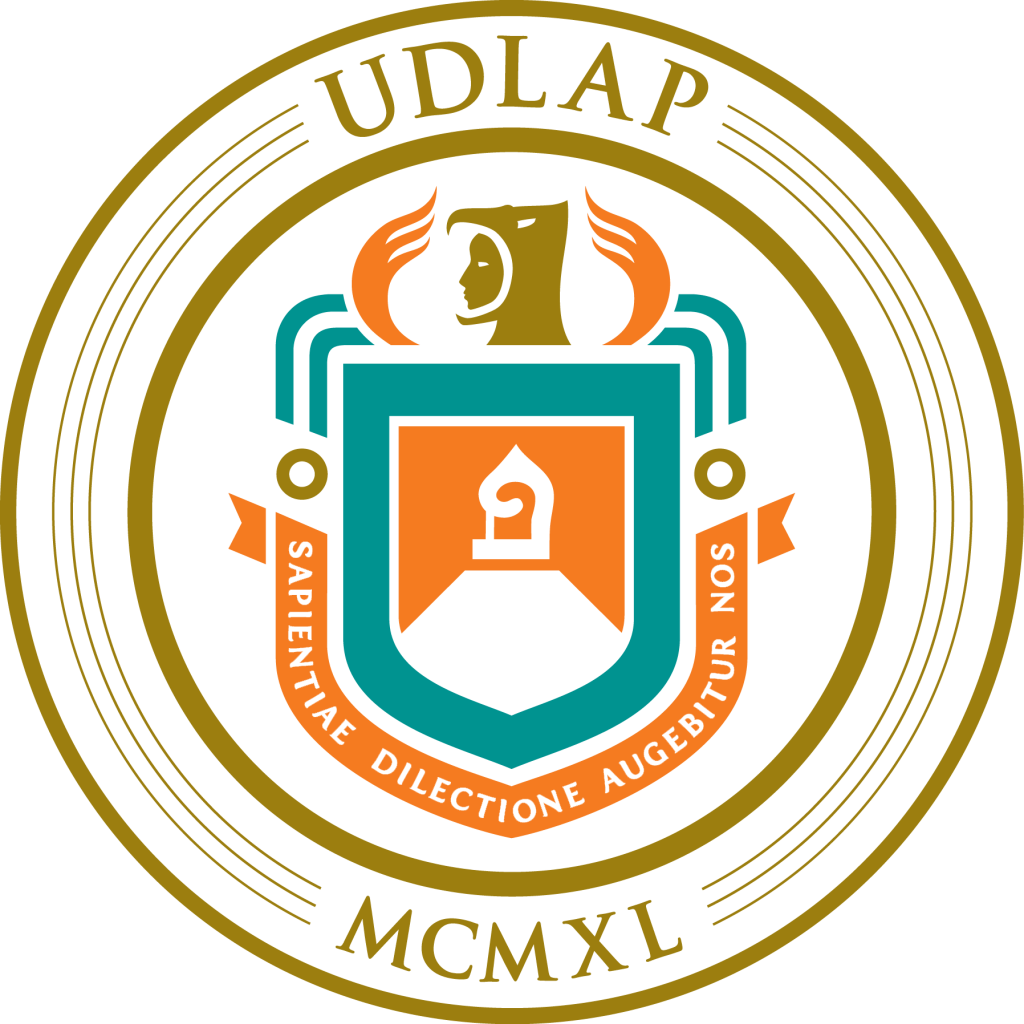
**UNIVERSIDAD DE LAS AMÉRICAS PUEBLA**

****

**TEMAS SELECTOS III: *BIG DATA***

**DR. DANIEL VALLEJO RODRÍGUEZ**

***“*PROYECTO FINAL: PREDICCIONES DE LAS ELECCIONES”**

ALAN PÉREZ CASTILLO 150294

ARTURO BRAVO ROVIROSA 150822

RAÚL GONZÁLEZ CRUZ 151211

ITZEL TLELO COYOTECATL 150154

**LICENCIATURA EN INGENIERÍA EN SISTEMAS COMPUTACIONALES**

SAN ANDRÉS CHOLULA, A 07 DE MAYO DE 2018

**ÍNDICE**

**[1. RESUMEN](#_Toc513567289)** [2](#_Toc513567289)

[**2. INTRODUCCIÓN** 2](#_Toc513567290)

[**3. DESCRIPCIÓN** 3](#_Toc513567291)

[**4. SIMULACIONES Y RESULTADOS** 4](#_Toc513567292)

[**5. CONCLUSIONES** 9](#_Toc513567293)

[**6. APÉNDICE A: CÓDIGO IMPLEMENTADO EN R** 11](#_Toc513567294)

[**7. REFERENCIAS** 13](#_Toc513567295)

# 

# 

# 

# 

# 

# 

# **1. RESUMEN**

Desarrollar un programa que sea capaz de predecir el resultado de las elecciones presidenciales de 2018 en México a partir de ciertas fuentes de información, resaltando la popularidad registrada por cada candidato en diversas encuestas.

Evaluar y graficar el comportamiento de este fenómeno mediante herramientas aprendidas durante el curso *Big Data* enfocado a la estadística.

# **2. INTRODUCCIÓN**

Durante las elecciones presidenciales que se realizan cada 6 años en México, se puede notar que la competencia entre candidatos comienza ser uno de los eventos más criticados cuando llega finalmente el día en el que el ciudadano debe votar; sin embargo, ¿Es posible predecir los resultados de una elección? ¿Con qué nivel de precisión? ¿Cuán útiles son para ello las encuestas de opinión pública? En este documento se busca responder a estas interrogantes a través de una estimación de los resultados electorales para las elecciones presidenciales de México 2018 a partir de la información provista por Oraculus, un modelo que da seguimiento a la intención de voto que es registrada por distintas fuentes de información como El Financiero, Reforma, Consulta, Suasor, entre otros.

En este caso, se pretende obtener una estimación de la preferencia electoral a partir de la combinación de los resultados de varias encuestas. La lógica detrás es que, aunque de manera individual las encuestas pueden producir mediciones con “ruido”, en conjunto pueden ayudarnos a recuperar de manera aproximada lo que intentamos medir.

Con una anticipación aproximada de dos meses previos a celebrarse las próximas elecciones federales en México de 2018, las cuales se esperan para el 1° de julio, mediante el uso de herramientas estadísticas relativamente sencillas, se pretende predecir el porcentaje de voto que recibirá cada uno de los candidatos para Presidente de la República. Los resultados del análisis cuantitativo son realizados usando R como principal herramienta de cálculo analítico y estadístico llevado a cabo en programación.

# **3. DESCRIPCIÓN**

La principal herramienta que usaremos para realizar el análisis estadístico es R, un entorno y lenguaje de programación con un enfoque al análisis estadístico. De manera más detallada, R es una implementación de software libre del lenguaje S pero con soporte de alcance estático. Se trata de uno de los lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy popular en el campo de la minería de datos, la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y gráficas. Se destaca que R está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

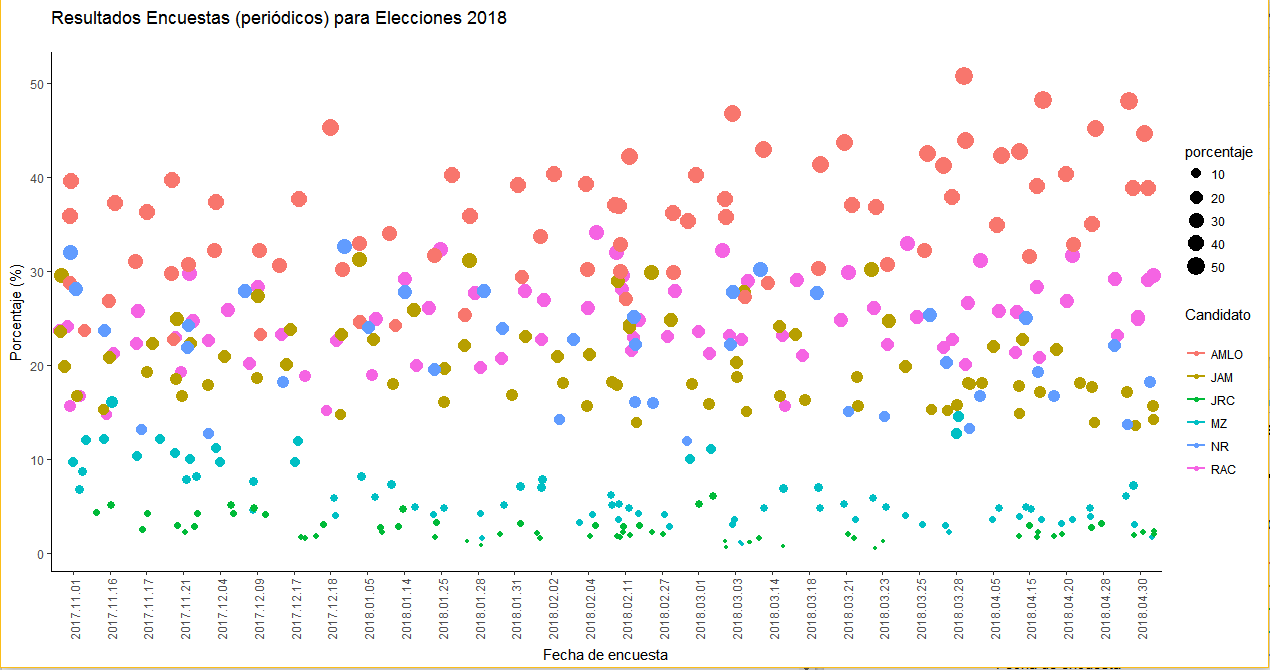
Dado el trabajo a realizar, R proporciona un amplio abanico de herramientas estadísticas (modelos lineales y no lineales, pruebas estadísticas, análisis de series temporales, entre otros) que nos serán de utilidad; así como el uso de paquetes que permiten representar la información en gráficas para facilitar el entendimiento del análisis.

Por ello es que se propone el uso de la librería ggplot2, un sistema de trazado para R basado en la gramática de los gráficos. Con esta dependencia buscamos proporcionar de una manera fácil un modelo de gráficos que produzca gráficos complejos de varias capas. A continuación se mencionara brevemente las siguientes funciones utilizadas en R:

* **head():** Regresa la primera o última parte de una lista, tabla o matriz.
* **ggplot():** Crea una gráfica de manera declarativa bajo subsecuencias de capas y sobreescritura especificada. Como se verá más adelante, con base en los datos es posible generar gráficas por capas.
* **labs():** Permite el acceso a las capas de las variables definidas dentro de una gráfica para que estas puedan ser desplegadas y ser vistas a detalle. Principalmente la usamos para establecer etiquetas de texto dentro del gráfico.
* **theme\_set():** Establece un color a un elemento asignado en una gráfica.
* **geom\_set():** Establece una forma geométrica a un elemento asignado en una gráfica.
* **geom\_smooth() y stat\_smooth():** Nos ayuda en la visualización de patrones al momento de generar predicciones sobre valores. Utilizamos geom\_smooth para mostrar los resultados con una geometría no estándar. El cálculo es realizado por el predictdf generic (actualmente no documentado) y sus métodos. Para la mayoría de los métodos, los límites de error estándar se calculan utilizando el método de predicción por *loess* para tener una aproximación basada en t. Finalmente usamos en algunos casos el método lm para generar líneas suaves lineales.
* **geom\_point():** Se usa para crear diagramas de dispersión. El diagrama de dispersión es más útil para mostrar la relación entre dos variables continuas; aunque también se puede usar para comparar una variable continua y una variable categórica, o dos variables categóricas.
* **geom\_jitter():** Es un atajo conveniente para geom\_point (position = "jitter"). Agrega una pequeña cantidad de variación aleatoria a la ubicación de cada punto y es una forma útil de manejar la sobreimpresión causada por la discreción en conjuntos de datos.
* **facet\_grid():** Forma una matriz de paneles definidos por variables de facetas de fila y columna. Es más útil cuando se tienen dos variables discretas, y todas las combinaciones de las variables existen en los datos.
* **theme(axis.text.x = element\_text(angle = 90, hjust = 1)):** Nos permite cambiar la orientación de las etiquetas de texto del eje X dentro de la gráfica. En este caso, el texto se mostrará de marea vertical.
* **xlim():** Nos permite hacer una determinada proyección de datos con respecto al eje X. Puede ser numérica (cantidad de encuestas), o bien, de acuerdo al objecto Date como se verá más adelante.

# **4. SIMULACIONES Y RESULTADOS**

A partir de información extraída de los sitios web indicados, se tuvo que recolectar la mayor cantidad de datos posible en un archivo CSV (comma-separated values) para posteriormente trabajar con el software *R,* graficar dicha información y representarla esquemáticamente con base en las encuestas llevadas a cabo (popularidad entre los candidatos aspirantes a las elecciones México 2018). A continuación se muestra una gráfica, Figura 4.1, que resume los resultados obtenidos después de que algunos periódicos publicarán el resultado de varias encuestas realizadas desde noviembre del 2017 a abril del 2018. Los datos son representados de forma que se aprecia el porcentaje de personas que dijeron que votarían por cierto candidato de los cinco disponibles. Cada punto se encuentra representado en proporción al porcentaje de acuerdo a las encuestas realizadas, por lo que las escalas van desde el 10% hasta el 50% aproximadamente:



*Figura 4.1.* Gráfica obtenida en R. Describe el resultado de las encuestas respecto a los candidatos realizados por un periodo de tiempo

Observando la gráfica en la Figura 4.1 puede notarse que el mayor porcentaje en cuanto a preferencia se encuentra relacionado con (AMLO) Andrés Manuel López Obrador, seguido por (RAC) Ricardo Anaya Cortés, (JAM) José Antonio Meade y (MZ) Margarita Zavala, de modo que en las posiciones con menor porcentaje se encuentran (JRC) Jaime Rodríguez Calderón y la opción (NR) que hace referencia a aquellas personas que no tuvieron alguna preferencia en cuanto a su voto.

Ahora, nos proponemos a analizar la preferencia que obtuvo el candidato Andrés Manuel López Obrador por separado dado su porcentaje de popularidad. Con esto pretendemos predecir los resultados que obtendrá en las siguientes encuestas; por lo cual generamos un dataframe (una estructura de datos bidimensional en R) que contenga sólo los datos de interés relacionados con el candidato AMLO. Después aplicamos un modelo de regresión lineal simple con los datos obtenidos de las fuentes, por lo que una vez obtenida la recta de regresión de la variable Y respecto a X, se puede estimar el parámetro xt = E(Y /X = xt) con base en la ecuación de la recta:

**

*Figura 4.2.* Fórmula de la ecuación de una recta

Recordando la hipótesis planteada de que se puede producir “ruido”, se puede establecer que la varianza se mide como:

*             (             2   ) V
ar(^m ) = s2 1-+ ---(xt---x)---  = s2-= s2h ,
      t       n    sum ni=1(xi- x)2   nt      tt
     *

*Figura 4.3.* Fórmula de la varianza

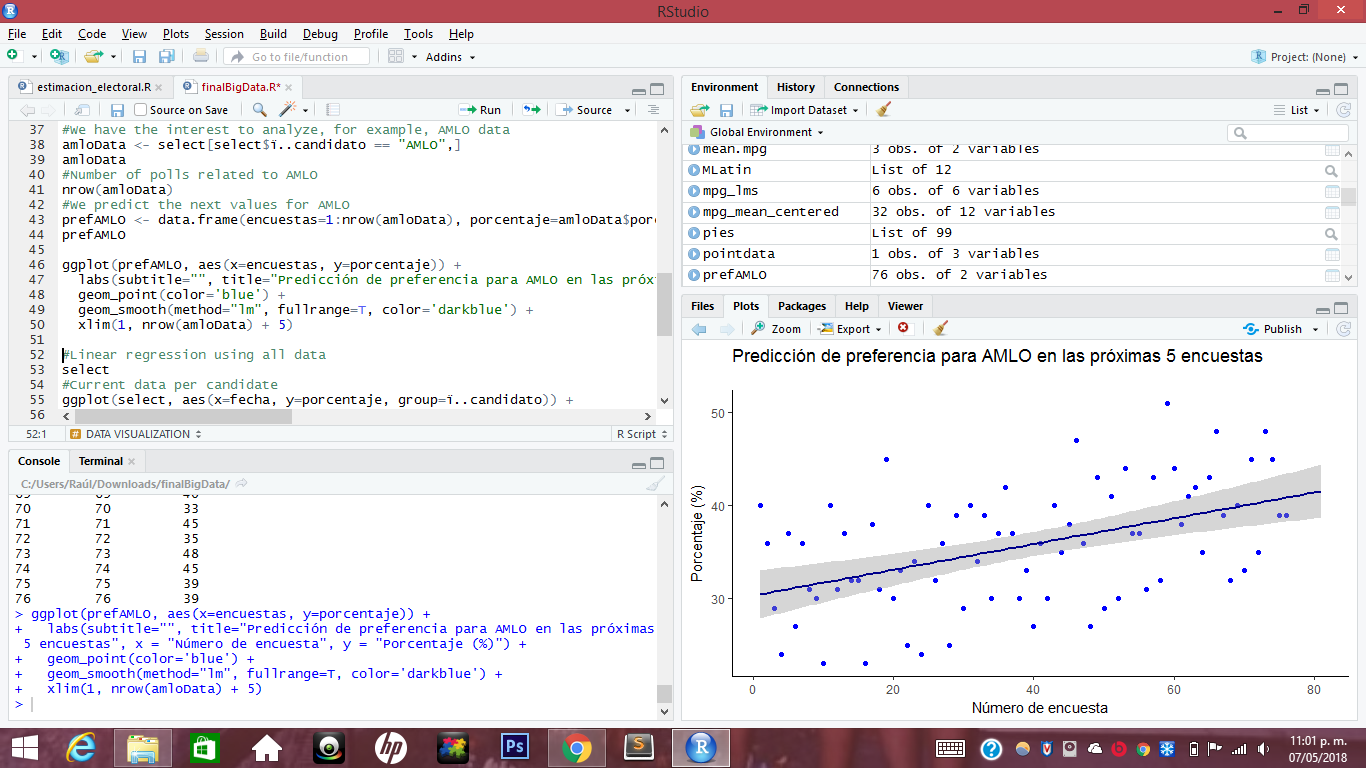
donde

*nt = ---(-n----)2-
     1+  xt---x
           sX
     *

*Figura 4.4.* Fórmula del valor de nt (unidad de dispersión)

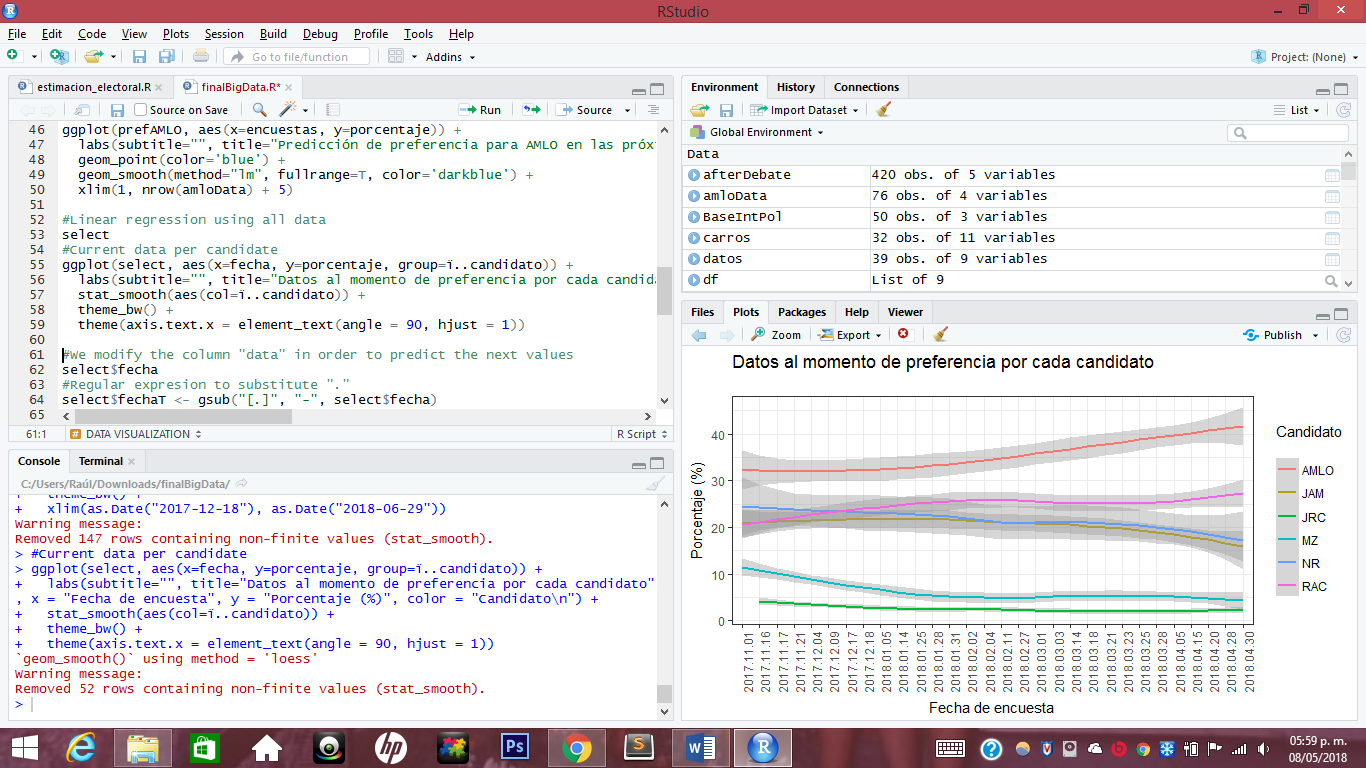
nt se denomina número equivalente de observaciones para estimar xt.

En este caso, el número de observaciones consiste en determinar la cantidad de encuestas registradas para el candidato (siendo en total 76). A partir de esta información procedemos a predecir el resultado que se tendrán en las próximas 5 encuestas para AMLO, obteniendo como resultado lo siguiente:



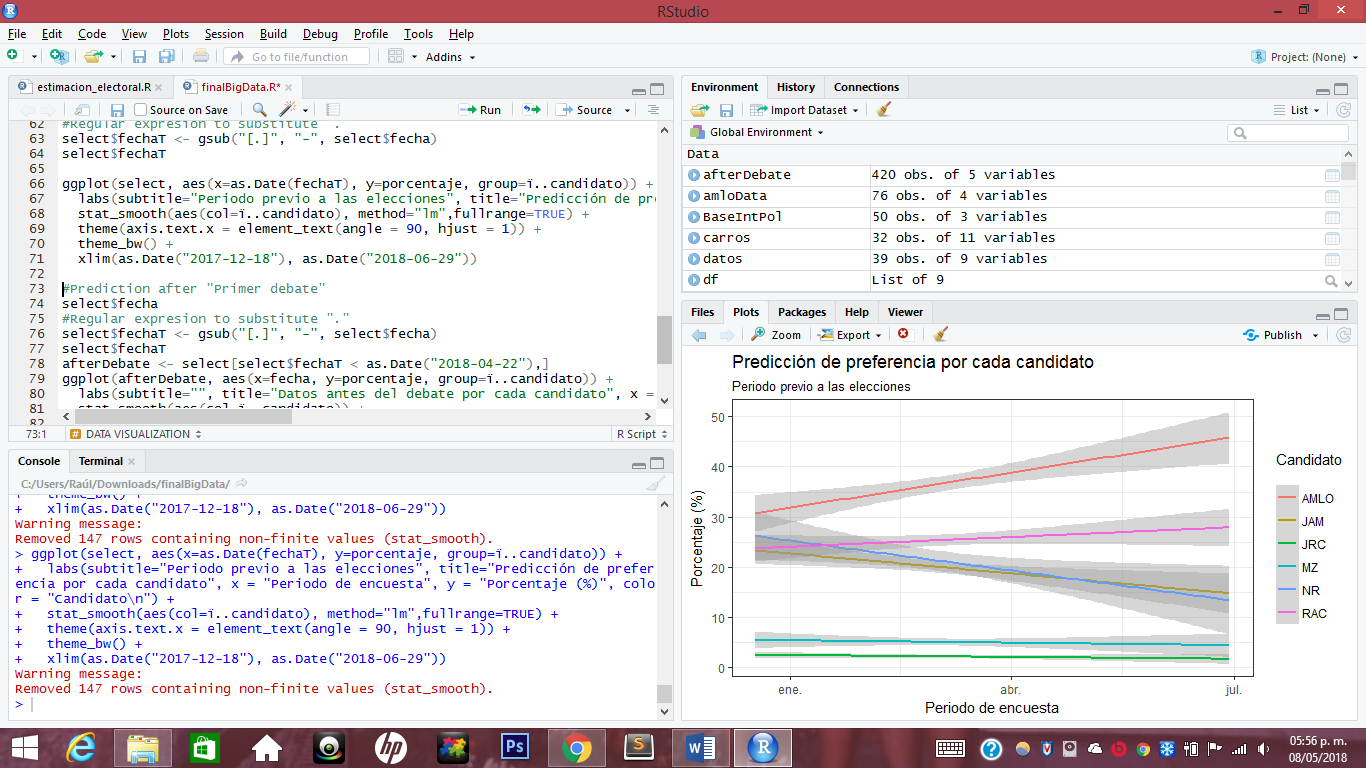
*Figura 4.5.* Gráfica obtenida en R. Predicción con respecto al candidato Andrés Manuel López Obrador.

De la anterior imagen, se observa que AMLO tenderá a subir el porcentaje de preferencia con base en los datos que se tienen hasta el momento. Por ello sería interesante generar la predicción para el resto de los candidatos en conjunto. Aplicando de nuevo un modelo de regresión lineal simple, primero procedemos a analizar la recta de regresión hasta la última fecha de la encuesta realizada; obteniendo lo siguiente:



*Figura 4.6.* Gráfica obtenida en R. Muestra un grado de preferencia hacia cada candidato a partir de las encuestas obtenidas por diferentes fechas.

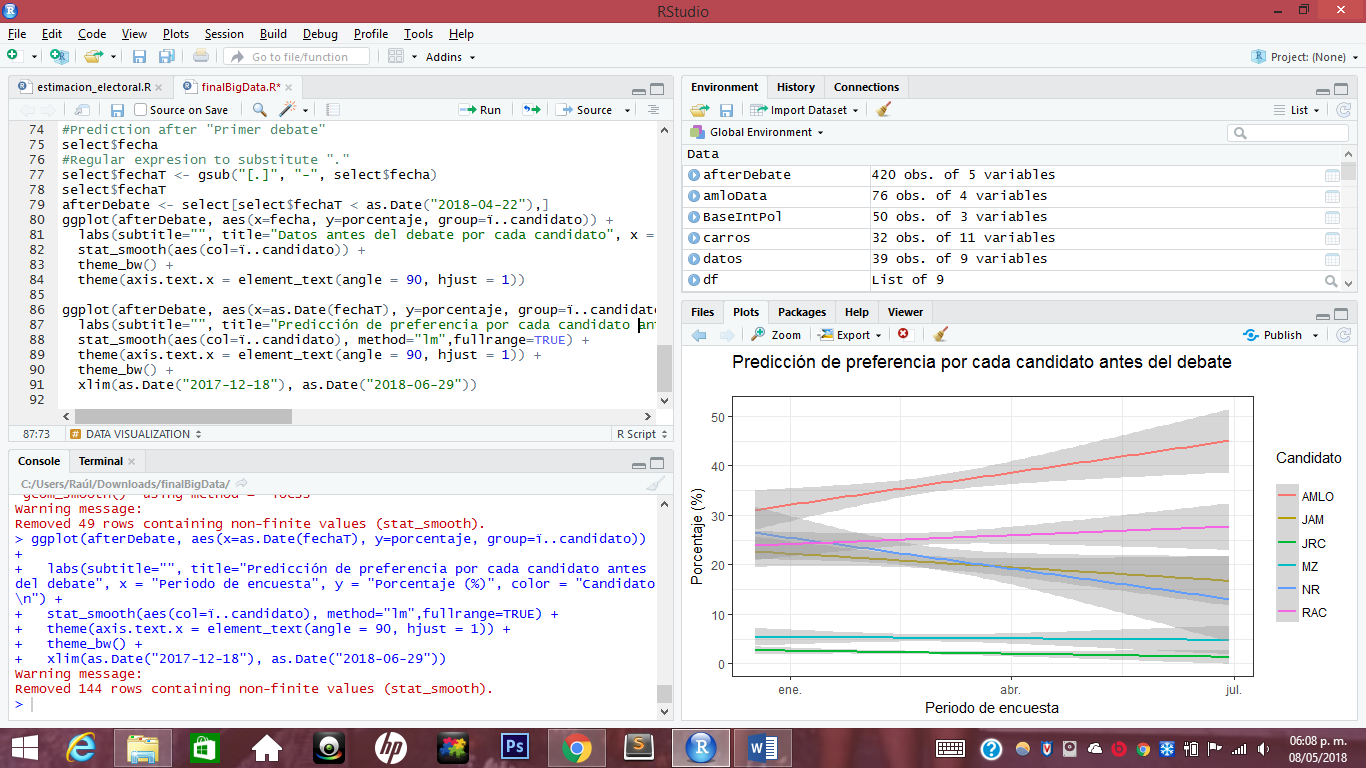
De la anterior gráfica podemos observar que AMLO, RAC y JRC tienden a subir; encabezando AMLO el porcentaje de preferencia. Con respecto al resto de los candidatos, se muestra una pendiente negativa, lo cual representa malas noticias para JAM y MZ. A partir de esta información, si tratamos de predecir el porcentaje que se tendrá hasta el 29 de junio del presente año, la recta de regresión nos mostrará lo siguiente:



*Figura 4.7.* Gráfica obtenida en R. Predicción sobre la preferencia de acuerdo a cada candidato antes de realizarse el día de las elecciones (1° de julio).

Con esto, se predice que AMLO obtendrá el mayor porcentaje de preferencia, lo cual lo convertirá en el candidato para ganar las elecciones presidenciales de México en 2018.

Sin embargo, una de las preguntas que nos hemos planteado con respecto a esta predicción es: Dado el primer debate presidencial que se llevó a cabo el 22 de abril del presente año en la ciudad de México, ¿cuál fue el impacto que tuvo en las preferencias de cada candidato? Con esto buscamos conocer la influencia que tuvo este evento para aumentar o disminuir el porcentaje de preferencia. Entonces, se aplicó nuevamente un modelo de regresión lineal estableciendo el análisis de encuestas cuya fecha no fuera mayor al 20 de abril de 2018. A continuación se mostrará la gráfica que revela cuál hubiera sido el porcentaje de preferencia por candidato en caso de no considerar las encuestas posteriores el debate mencionado previamente:



*Figura 4.8.* Gráfica obtenida en R. Predicción sobre la preferencia de acuerdo a cada candidato antes de realizarse el primer debate presidencial.

En comparación con los resultados previos, es posible ver a primera instancia que el debate realmente favoreció a MZ para disminuir la desventaja que tenía con su competidor más cercano JAM. Por otra parte, en general todos los candidatos resultaron favorecidos con el debate a excepción de JAM, quien lamentablemente perdió preferencia vista en la Figura 4.6. En comparación con los otros candidatos, AMLO fue el mayor favorecido ya que la expectativa antes de ser el debate provocó un aumento de 3% aproximadamente, mientras que para RAC, MZ y JRC fue de apenas un 1% aproximadamente.

# **5. CONCLUSIONES**

El objetivo principal del trabajo fue brindar una predicción con respecto a las elecciones que se llevarán a cabo en México este 2018. Para esto fue necesario basarse en los conocimientos obtenidos a través del curso *Big Data*, de modo que haciendo uso de conceptos de estadística fue posible el análisis de los datos recolectados.

A grandes rasgos las métricas utilizadas corresponden a modelos probabilísticos, especialmente empleando el modelo de regresión lineal que permitió la observación de la tendencia con respecto a los candidatos y los porcentajes obtenidos en cuanto a los votos que probablemente determinarían su elección. De acuerdo a estos métodos, la fuente sobre la que se extrajeron los datos fue Oraculus, un sitio dedicado a la recolección de datos y publicación de resultados de encuestas relacionadas la preferencia entre candidatos. Dichas encuestas fueron realizadas por diferentes periódicos, así que la colección fue a nuestro parecer bastante completa, permitiéndonos generar un archivo CSV con la información que necesitamos.

Después de haber procesado los datos, estos dieron una tendencia a que entre los candidatos y de acuerdo a los resultados publicados, Andrés Manuel López Obrador será quien ganará las elecciones presidenciales de este año. Las gráficas muestran que este candidato, en caso de realizarse más encuestas, tendrá una mayor cantidad de personas que voten por él con base en los resultados procesados, siendo el vencedor dado su porcentaje de preferencia con respecto a los otros candidatos. Probablemente hayan tenido que ver otros factores como popularidad, propuestas y promesas realizadas con respecto a su campaña, entre otros. Sin embargo, en esta ocasión al tomarse únicamente en cuenta datos que ya fueron publicados de encuestas realizadas, no sabemos con exactitud el cómo podría afectar la predicción el agregar otros factores; por lo cual sería interesante el incluir otros elementos para obtener distintos escenarios y así generar nuevas perspectivas en cuanto al resultado.

Finalmente pudimos observar que el debate si provocó cambios en el porcentaje de preferencia para cada candidato. Muestra de ello es que, sin tomar en cuenta las encuestas previas al debate, se tenía la expectativa de que JAM tendría un cierta ventaja contra MZ, la cual desapareció al consultar todas las encuestas registradas en el archivo CSV, permitiéndonos inferir que el debate tuvo influencia en este evento. También es que el debate, al parecer, permitió aumentar un poco más la expectativa sobre la preferencia que se tenía, resultando AMLO el más favorecido de esto. Sin embargo, sería interesante volver a realizar un análisis con nuevos datos de encuestas; así como esperar los dos debates restantes que quedan previo al día de las elecciones con el fin de predecir cuál será el impacto que tendrán estos eventos en cada candidato.

# **6. APÉNDICE A: CÓDIGO IMPLEMENTADO EN R**

A continuación se muestra el código (elaborado en lenguaje R), el cual describe ciertos componentes clave (documentación) acerca de cómo se emplean las funciones que se explicaron en el apartado Descripción, así como la parte relacionada con la graficación de datos y cálculo estadístico con base en el modelo de datos recopilado de los sitios mencionados en el apartado Introducción.

# -------------------------------------------------------------------------

# ---------- ELECTION PREDICTION ---------- BIG DATA ----------------------

# -------------------------------------------------------------------------

# Data source: http://oraculus.mx/poll-of-polls/

# ----------- DATA VISUALIZATION ------------------------------------------

library(ggplot2)

setwd("C:/Users/Raúl/Desktop/Big Data/finalBigData")

encuestas <- read.csv("encuestas.csv")

dim(encuestas)

head(encuestas)

candidatos <- unique(encuestas$ï..candidato)

candidatos

select <- encuestas[encuestas$ï..candidato %in% candidatos, ]

select

# Grafica porcentaje de votos (encuestas) por años dependiendo del candidato

theme\_set(theme\_classic())

g <- ggplot(select, aes(x=encuestas$fecha, y=encuestas$porcentaje)) +

labs(subtitle="", title="Resultados Encuestas (periódicos) para Elecciones 2018") +

labs(x = "Fecha de encuesta", y = "Porcentaje (%)") + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

g + geom\_jitter(aes(col=ï..candidato, size=porcentaje)) +

geom\_smooth(aes(col=encuestas$ï..candidato), method="lm", se=F) +

labs(x = "Fecha de encuesta",y = "Porcentaje (%)", color = "Candidato\n")

gridEncuestas <- ggplot(encuestas, aes(x = encuestas$fecha, y = encuestas$porcentaje, colour=encuestas$ï..candidato)) +

geom\_point() + facet\_grid(~ encuestas$ï..candidato) +

labs(x = "Fecha de encuesta", y = "Porcentaje (%)", color = "Candidato\n") +

theme(axis.text.x = element\_text(angle=90, vjust=0.6))

gridEncuestas

#We have the interest to analyze, for example, AMLO data

amloData <- select[select$ï..candidato == "AMLO",]

amloData

#Number of polls related to AMLO

nrow(amloData)

#We predict the next values for AMLO

prefAMLO <- data.frame(encuestas=1:nrow(amloData), porcentaje=amloData$porcentaje)

prefAMLO

ggplot(prefAMLO, aes(x=encuestas, y=porcentaje)) +

labs(subtitle="", title="Predicción de preferencia para AMLO en las próximas 5 encuestas", x = "Número de encuesta", y = "Porcentaje (%)") +

geom\_point(color='blue') +

geom\_smooth(method="lm", fullrange=T, color='darkblue') +

xlim(1, nrow(amloData) + 5)

#Linear regression using all data

select

#Current data per candidate

ggplot(select, aes(x=fecha, y=porcentaje, group=ï..candidato)) +

labs(subtitle="", title="Datos al momento de preferencia por cada candidato", x = "Fecha de encuesta", y = "Porcentaje (%)", color = "Candidato\n") +

stat\_smooth(aes(col=ï..candidato)) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

#We modify the column "data" in order to predict the next values

select$fecha

#Regular expresion to substitute "."

select$fechaT <- gsub("[.]", "-", select$fecha)

select$fechaT

ggplot(select, aes(x=as.Date(fechaT), y=porcentaje, group=ï..candidato)) +

labs(subtitle="Periodo previo a las elecciones", title="Predicción de preferencia por cada candidato", x = "Periodo de encuesta", y = "Porcentaje (%)", color = "Candidato\n") +

stat\_smooth(aes(col=ï..candidato), method="lm",fullrange=TRUE) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

xlim(as.Date("2017-12-18"), as.Date("2018-06-29"))

#Prediction after "Primer debate"

select$fecha

#Regular expresion to substitute "."

select$fechaT <- gsub("[.]", "-", select$fecha)

select$fechaT

afterDebate <- select[select$fechaT < as.Date("2018-04-22"),]

ggplot(afterDebate, aes(x=fecha, y=porcentaje, group=ï..candidato)) +

labs(subtitle="", title="Datos antes del debate por cada candidato", x = "Fecha de encuesta", y = "Porcentaje (%)", color = "Candidato\n") +

stat\_smooth(aes(col=ï..candidato)) +

theme\_bw() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

ggplot(afterDebate, aes(x=as.Date(fechaT), y=porcentaje, group=ï..candidato)) +

labs(subtitle="", title="Predicción de preferencia por cada candidato antes del debate", x = "Periodo de encuesta", y = "Porcentaje (%)", color = "Candidato\n") +

stat\_smooth(aes(col=ï..candidato), method="lm",fullrange=TRUE) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

theme\_bw() +

xlim(as.Date("2017-12-18"), as.Date("2018-06-29"))

# -------------------------------------------------------------------------

*Script* del código creado en lenguaje R. Lógica implementada para predecir los resultados de las elecciones presidenciales en México 2018.

# **7. REFERENCIAS**

Moreno, B. (2015). ¿Cómo predecir las elecciones?. The social s*cience post.* (abril 04, 2018), recuperado de: <http://thesocialsciencepost.com/es/2015/10/como-predecir-las-elecciones/>

Aguilar, J., et. al. (2015). Modelo de predicción electoral: el caso de la elección municipal 2015 de León de los Aldama, Guanajuato. *Science Direct.* (abril 04, 2018), recuperado de: <https://www.sciencedirect.com/science/article/pii/S0185161615000256>

Márquez, J., et. al. (2018). Poll of polls. *Oraculus.* (abril 04, 2018), recuperado de: <http://oraculus.mx/poll-of-polls/>

[s.a]. (2017). Predicción en regresión lineal simple. Recuperado de: <http://dm.udc.es/asignaturas/estadistica2/sec6_9.html>

RStudio. (2018). Documentation Guide. *RStudio.* (abril 06, 2018), recuperado de:<https://www.rstudio.com/wp-content/uploads/2016/11/ggplot2-cheatsheet-2.1.pdf>