Clasificación y análisis de emisiones de gases de efecto invernadero

Helena Patricia Carrillo Soto*

Resumen

Se utilizaron métodos de aprendizaje de máquina no supervisado y supervisado para analizar datos de emisiones de gases de efecto invernadero. Se utiliza el algoritmo de k-medias para separar las emisiones en cuatro grupos y se revisan un conjunto de modelos para encontrar el que mejor se ajusta para predicciones futuras resultando en una cresta bayesiana como mejor modelo elegido.

Palabras clave: gases de efecto invernadero, aprendizaje no supervisado, aprendizaje de máquina

1. Introducción

Los gases de efecto invernadero (GEI) son los componentes de la atmósfera que absorben la radiación emitida por la superficie de la Tierra y la emiten de regreso.

Los principales GEI son el vapor de agua (H_2O), el dióxido de carbono (CO_2), el óxido nitroso (N_2O), el metano (CH_4) y el ozono (O_3).

También se encuentran GEI creados en su totalidad por el ser humano, como los halocarbonos (CFCs, HCFCs, HFCs y PFCs) los cuales contienen cloro, bromo o flúor y carbono. Estos compuestos son una de las causas del agotamiento de la capa de ozono en la atmósfera (Ballesteros and Aristizabal, 2007) .

Los GEI están directamente relacionados con el calentamiento global y este a su vez con el cambio climático.

Este tema ha cobrado relevancia en los último años con el rápido aumento de la temperatura media global, los intentos fallidos de distintos gobiernos de reducir sus emisiones y la llamada de científicos alrededor del mundo de tomar medidas urgentes para controlar el cambio climático.

2. Metodología y Datos

En el presente trabajo se utilizarán métodos de aprendizaje de máquina para analizar los datos correspondientes al Ïnventario Nacional de Emisiones de Gases y Compuestos de Efecto Invernadero"del INECC que se puede encontrar en la siguiente liga: https://datos.gob.mx/busca/dataset/inventario-nacional-de-emisiones-de-gases-y-compuestos-de-efecto-invernadero-inegycei.

Los datos consisten en las emisiones anuales totales, y desglosadas por industria, de los diferentes GEI para los años de 1990 a 2021.

*Facultad de Ciencias Fisico Matemáticas, Universidad Autónoma de Nuevo León (UANL). Email: helena.carrilloso@uanl.edu.mx

3. Aprendizaje no supervisado

El aprendizaje no supervisado, usa algoritmos de aprendizaje de máquina para analizar y clasificar datos no etiquetados.

Se utilizan principalmente para encontrar patrones ocultos en los datos.

3.1. K-medias

El algoritmo K-medias es un método popular de agrupamiento por aprendizaje no supervisado.

El algoritmo se basa en agrupar los datos en k grupos en base a su media.

En él asumimos que se tiene un conjunto de objetos $D = \{ p_1, p_2, \ldots, p_n\}$ donde cada objeto $p_i = \{ p_{i1}, p_{i2}, \ldots, p_{in}\}$ representa un punto en el espacio R^m donde m es el número de atributos o dimensiones de los objetos del conjunto D.

Se define k como el número de grupos y m_k la media del grupo C_k. El objetivo general del método es minimizar la función criterio $e(k)$.

Normalmente la función criterio que se minimiza es la suma de cuadrados del error:

$$e(k) = \sum_{i=1}^{k} \sum_{p_j \in C_i} dist(p_j, m_i)^2$$

donde

p_j - punto en el espacio R^m que representa el punto p_j ,

m_i - media del grupo C_i ,

$dist(p_j, m_i)$ - distancia Euclidiana del punto p_j a la media (centro) del grupo más cercano C_i

(Kijewska and Bluszcz, 2016).

3.2. K-medias en el análisis de GEI

El algoritmo de K-medias se ha usado para el análisis de emisiones de GEI en múltiples ocasiones.

Anna Kijewska y Anna Bluszcz hicieron uso de él en 2016 para analizar los niveles de emisiones de GEI variantes en países de Europa (Kijewska and Bluszcz, 2016)

En 2018, Lozza y Bellini lo utilizaron para realizar la clasificación de sistemas ganaderos para estimaciones de GEI (Lozza and Bellini Saibene, 2018)

En esta ocasión se usará el algoritmo para categorizar las actividades que aportan al total de emisiones de GEI en México de los años de 1990 a 2021.

3.3. Resultados análisis K-medias

Para establecer el número de grupos a utilizar se utilizó el método del codo en el cuál se calcula la suma de las distancias al cuadrado para diferentes valores de k y se busca el valor de k donde la disminución en la suma se ralentiza o se acerca a su límite.

El resultado obtenido se puede ver en la figura 1 (p. 3):

Para tener un criterio más claro se puede calcular la distancia entre la recta creada por el mínimo y máximo número de grupos y sabemos que se optimiza k en el número que tenga una mayor distancia a la recta. Podemos observarlo en la figura 2 (p.4).

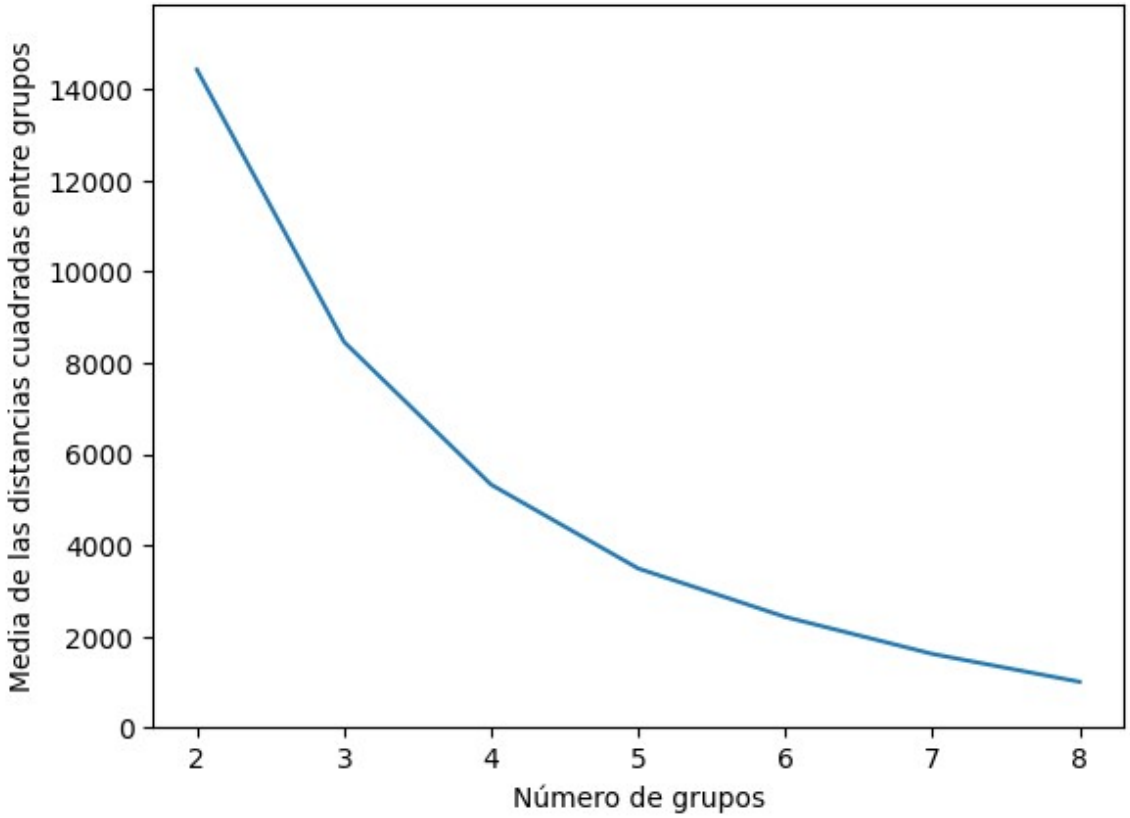

FIGURA 1: Método del codo

En base a esto se usan 4 grupos.

Los resultados arrojan los centros encontrados en la tabla 1 (p.3):

CUADRO 1: Centros: Puntos donde se encuentran los centros de cada uno de los grupos C_k.

C_k	x	y	z	w	v	u	r	s
1	-0.0907	-0.0822	-0.0089	0.0020	-0.0469	0.0019	0.0019	-0.1110
2	-0.1175	10.6739	-0.1918	-0.0760	-0.0551	-0.0724	-0.0724	3.0160
3	5.7762	-0.1407	0.6824	-0.0760	-0.0551	-0.0724	-0.0724	5.4903
4	-0.1129	-0.1810	-0.1918	-0.0760	19.3721	-0.0724	-0.0724	-0.1575

También se obtienen los tamaños de grupo mostrados en la tabla 2 (p.3):

CUADRO 2: Tamaño: Tamaño de cada uno de los grupos C_k.

Grupo	nk
1	4022
2	32
3	64
4	10

Debido a que el espacio es de 8 dimensiones solo se muestran una de las caras en la figura 3 (p.5).

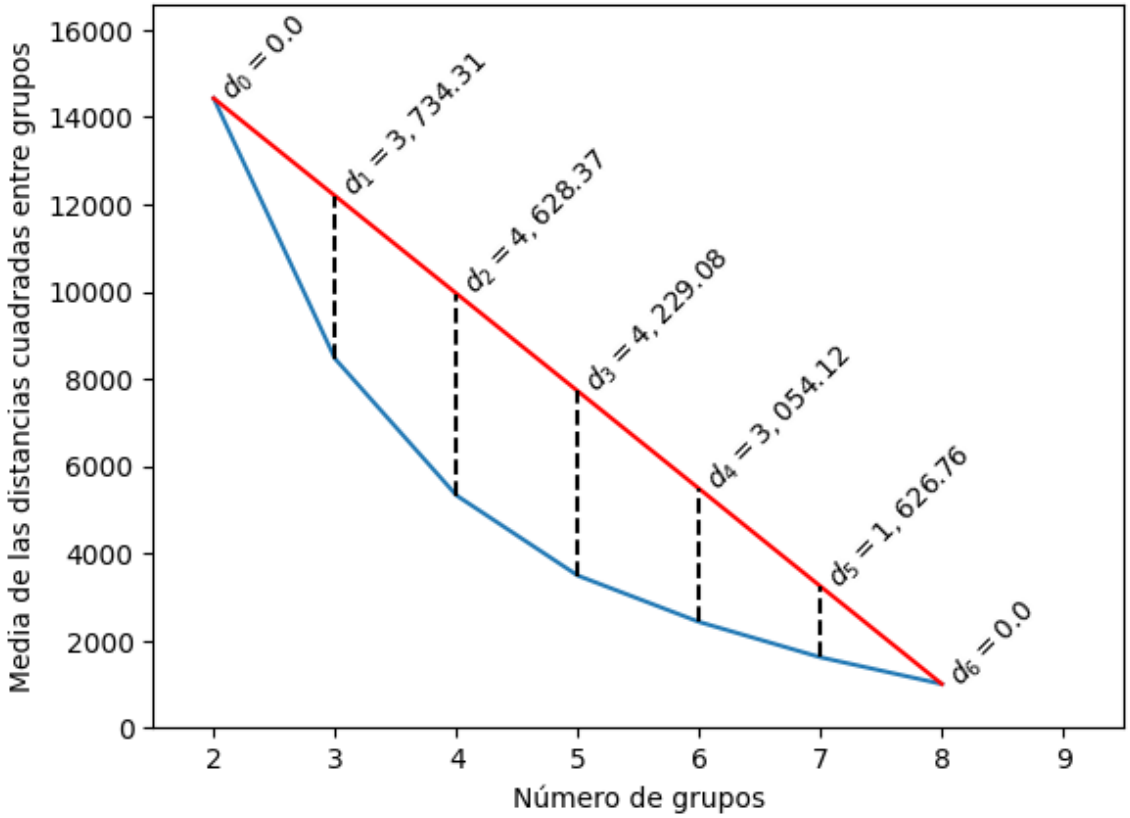

FIGURA 2: Método del codo con distancias

4. Aprendizaje supervisado

A diferencia del aprendizaje no supervisado, el aprendizaje supervisado utiliza datos etiquetados. Este utiliza datos de entrenamiento para enseñar a los modelos a generar la salida deseada.

4.1. Modelos de aprendizaje supervisado en el análisis de GEI

Distintos modelos de aprendizaje supervisado se han usado a lo largo del tiempo para analizar emisiones de Gases de Efecto Invernadero.

Zewei Jiang y Shihong Yang junto con otros investigadores usaron los modelos de bosques aleatorios (por sus siglas en inglés, RF), regresiones K-vecino mas cercano (KNN), regresiones aumento de gradiente(GBR), y regresiones lineales (LR), así como un regresor de ampliamiento conjunto de un GBR y RF, para modelar emisiones de GEI en diferentes escalas de tiempo de campos de arroz (Jiang et al., 2023) .

Por otro lado, Sabrina Hempel y Julian Adolph junto con su equipo usaron regresiones GBR, RF y regresiones lineales múltiples, regresiones de cresta bayesiana (BR, por sus siglas en inglés), una red neuronal con una capa oculta así como máquinas de soporte vectorial y procesos gaussianos para evaluar las emisiones de metano de un edificio lechero con ventilación natural (Hempel et al., 2020).

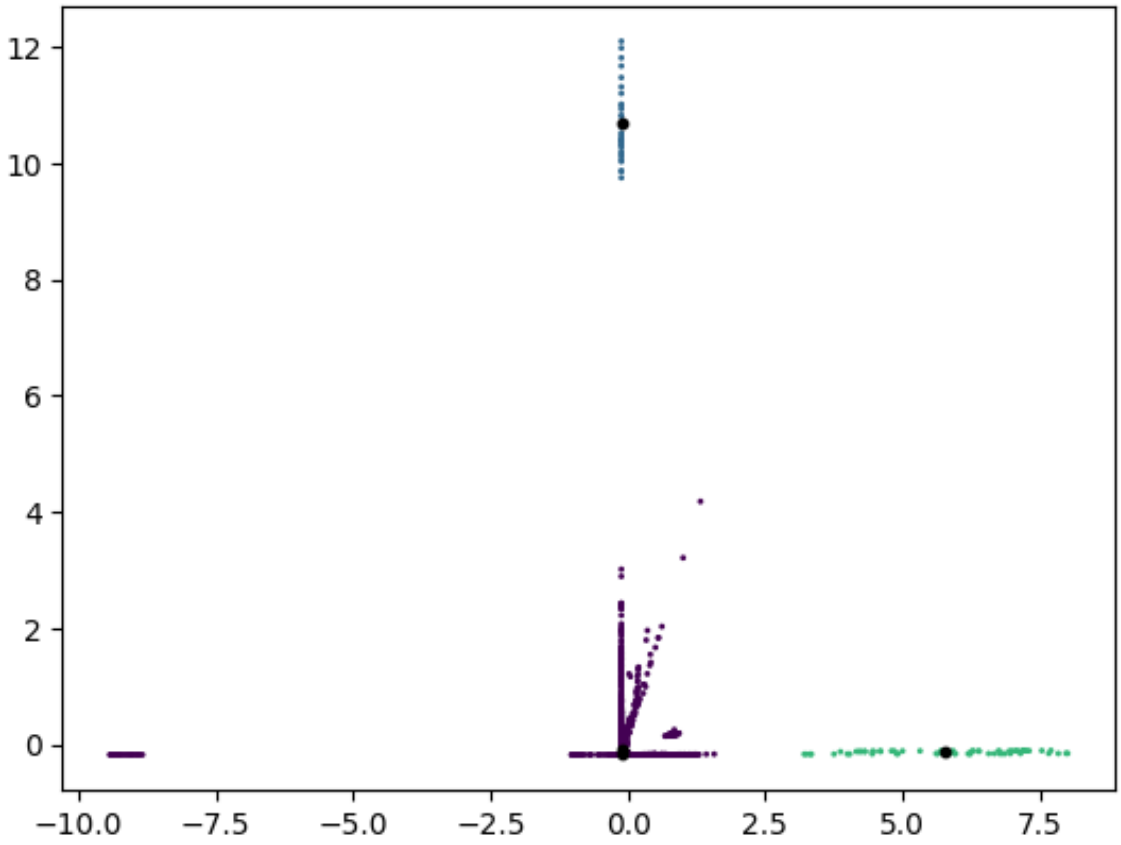

FIGURA 3: Grupos de CO_2 vs CH_4

4.2. Análisis y Resultados

Para este estudio primero se analizaron diferentes modelos de los encontrados en la literatura para evaluar cuál de ellos era el de mejor desempeño. Se usó el error de raíz cuadrada media (RMSE por sus siglas en inglés) y el error absoluto medio (por sus siglas e inglés, MAE) para compararlos.

Los modelos usados fueron:

- Cresta bayesiana (BR)
- Aumento de gradiente (GBR)
- Regresión Lineal (LR)
- Bosques aleatorios (RF)
- *K*-vecino más cercano (KNN)
- Regresor de ampliamiento conjunto usando GBR y RF

Todos se revisaron usando una división de entrenamiento y prueba especializada para series de tiempo encontrada en *scikit-learn* para tres distintos tamaños de prueba. Los mejores cinco resultados fueron los presentados en la tabla 3 (p.7).

Se realizó también un análisis gráfico que puede observarse en la figura 4 (p.6) y la figura 5 (p.6).

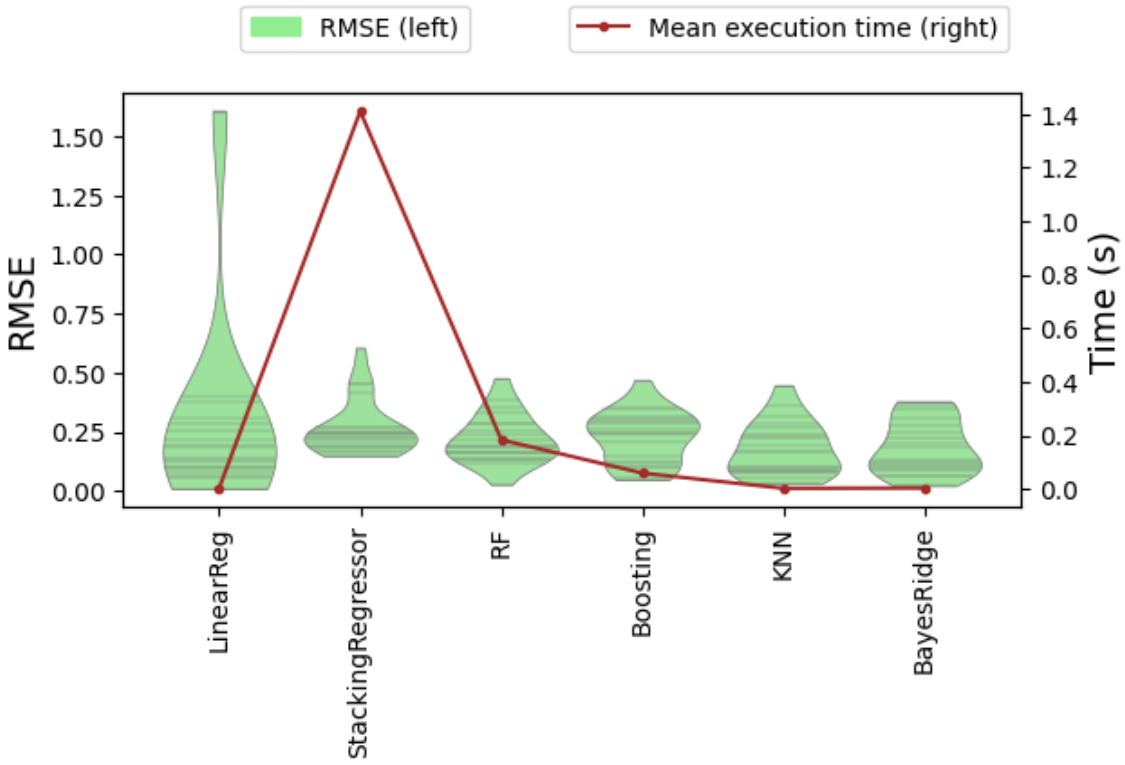

FIGURA 4: Modelos por RMSE y tiempo de ejecución

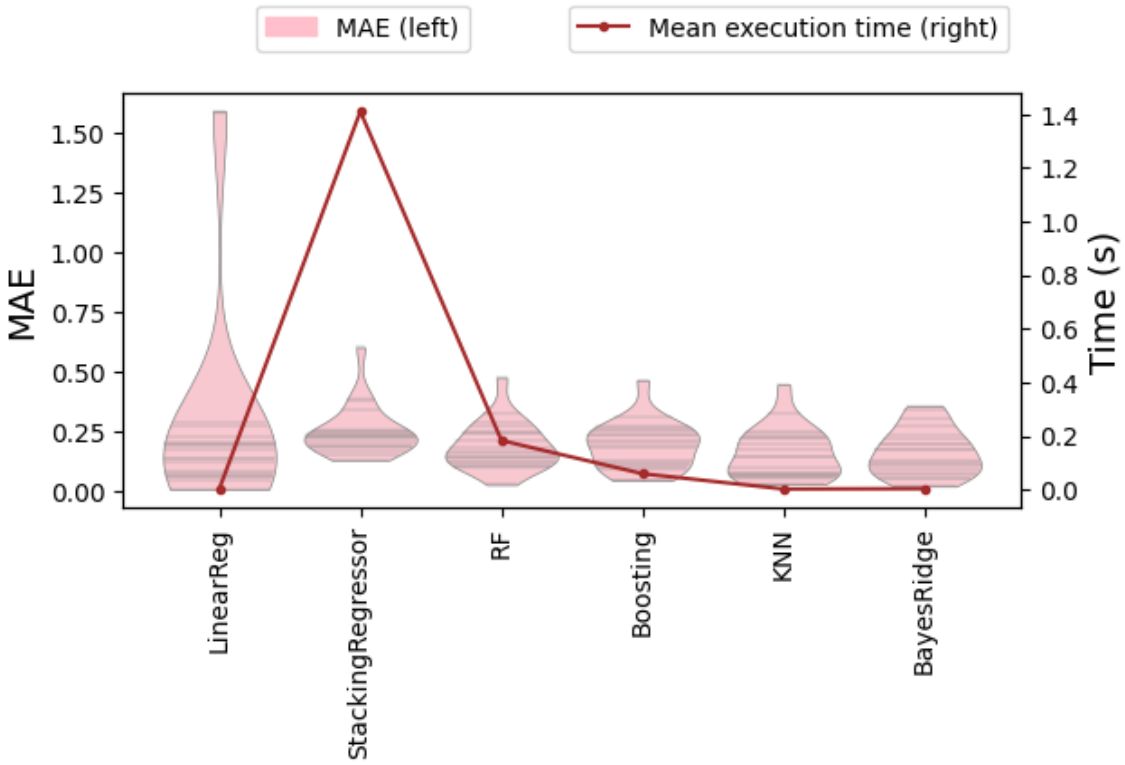

FIGURA 5: Modelos por MAE y tiempo de ejecución

CUADRO 3: Mejores cinco resultados de pruebas de regresiones por RMSE y MAE

	Modelo	RMSE	MAE	Tiempo
1	LR	0.0105	0.0094	0.0028
2	BR	0.0216	0.0214	0.0026
3	RF	0.0275	0.0274	0.2002
4	KNN	0.0301	0.0301	0.0022
5	Boosting	0.0477	0.0476	0.0598

Basandonos en los gráficos podemos ver que a pesar de que el mejor modelo es una regresión lineal esta en realidad presenta mucha variabilidad respecto a las métricas obtenidas por lo cual no es ideal como modelo.

Tomando en cuenta ambas métricas y el tiempo de ejecución se decide que el mejor modelo es el de cresta bayesiana.

Referencias

Ballesteros, H. B. and Aristizabal, G. L. (2007). Información técnica sobre gases de efecto invernadero y el cambio climático. *Instituto de Hidrología, Meteorología y Estudios Ambientales-IDEAM. Subdirección de Meteorología (Bogotá, Colombia).*

Hempel, S., Adolphs, J., Landwehr, N., Willink, D., Janke, D., and Amon, T. (2020). Supervised machine learning to assess methane emissions of a dairy building with natural ventilation. *Applied Sciences*, 10(19).

Jiang, Z., Yang, S., Smith, P., and Pang, Q. (2023). Ensemble machine learning for modeling greenhouse gas emissions at different time scales from irrigated paddy fields. *Field Crops Research*, 292:108821.

Kijewska, A. and Bluszcz, A. (2016). Research of varying levels of greenhouse gas emissions in European countries using the k-means method. *Atmospheric Pollution Research*, 7(5):935–944.

Lozza, A. and Bellini Saibene, Y. (2018). Clasificación de sistemas ganaderos para estimación de gases de efecto invernadero. In *Conferencia Latinoamericana sobre Uso de R en Investigación+ Desarrollo (LatinR 2018)-JAIIO 47 (CABA, 2018).*