Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Grupo 101

3.1 Regresión No Lineal

DataForge

Jesús Eduardo Valle Villegas | A01770616

Manuel Eduardo Covarrubias Rodríguez | A01737781

Diego Antonio Oropeza Linarte | A01733018

Ithandehui Joselyn Espinoza Mazón | A01734547

Mauricio Grau Gutierrez Rubio | A01734914

Última edición el 13 de Octubre de 2025

1. Objetivo

Analizar la relación entre las variables *TaxonName, TaxonCode, SamplingOperations_code,CodeSite_SamplingOperations,Date_SamplingOperation, Abundance_nbcell, TotalAbundance_SamplingOperation* y *Abundance_pm* del conjunto de datos *01_DiatomInventories_GTstudentproject_B.csv*, aplicando y comparando dos modelos de regresión no lineal para determinar el grado de correlación y la capacidad explicativa de cada modelo a través de los coeficientes de determinación (R^2) y correlación.

2. Metodología

Para el desarrollo de esta actividad se empleó el conjunto de datos *01_DiatomInventories_GTstudentproject_B.csv*, el cual contiene información relacionada con códigos de muestreo, sitios de muestreo, fechas, taxones y valores de abundancia.

1. **Preparación y limpieza de datos:**
   Se realizó la carga del archivo en el entorno de trabajo, verificando su estructura, tipos de datos y la existencia de valores nulos. Posteriormente, las variables categóricas (*TaxonName*, *TaxonCode*, *SamplingOperations_code*, *CodeSite_SamplingOperations*) fueron codificadas numéricamente para permitir su tratamiento estadístico. La variable *Date_SamplingOperation* se transformó a un formato numérico (ordinal o continuo) para facilitar su uso en los modelos.

2. **Selección de variables:**
   Se consideraron como variables de estudio: *TaxonName*, *TaxonCode*, *SamplingOperations_code*, *CodeSite_SamplingOperations*, *Date_SamplingOperation*, *Abundance_nbcell*, *TotalAbundance_SamplingOperation* y *Abundance_pm*, definiendo distintas combinaciones para evaluar las correlaciones existentes.

3. **Aplicación de modelos no lineales:**
   Se seleccionaron y aplicaron dos modelos de regresión no lineal:

   - **Modelo polinómico**, utilizando funciones de segundo y tercer grado para capturar posibles curvaturas en los datos.

   - **Modelo exponencial o potencial**, empleado para explorar relaciones de crecimiento o decrecimiento no lineales entre las variables.

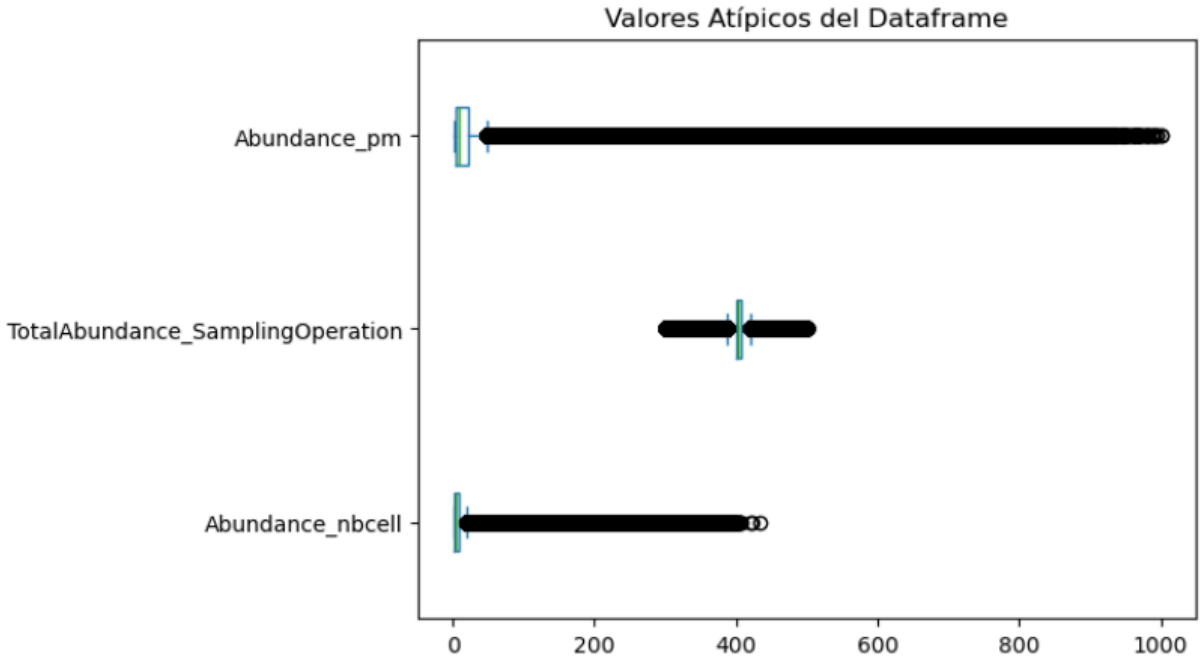
4. **Evaluación del ajuste:**
   Para cada modelo se calcularon los **coeficientes de determinación (R^2)** y los **coeficientes de correlación** (Pearson o Spearman, según el tipo de variable) con el fin de medir la fuerza y dirección de la relación. Además, se analizaron los **residuales** para comprobar la adecuación de los modelos y detectar posibles

desviaciones.

5. **Análisis comparativo de resultados:**
   Se elaboró una **tabla resumen** con los valores obtenidos de R^2 y de correlación para cada relación evaluada, identificando los modelos que presentaron el mejor desempeño. Finalmente, se interpretaron los resultados y se formularon conclusiones respecto al comportamiento de las variables y la eficacia de los modelos no lineales aplicados.

3. Resultados

Preprocesamiento:



Interpretación:
El análisis evidenció la presencia de valores atípicos en Abundance_pm y Abundance_nbcell, con una alta concentración en valores bajos pero con casos aislados muy elevados que generan colas largas en la distribución. Esto sugiere que, aunque la mayoría de los registros presentan abundancias reducidas, existen observaciones extremas que podrían influir en los resultados de los modelos.

En contraste, TotalAbundance_SamplingOperation mostró una distribución más compacta, aunque también con algunos valores fuera del rango esperado. La detección de estos casos es relevante, ya que pueden afectar la estabilidad y precisión de las regresiones lineales.

Transformación de variables:

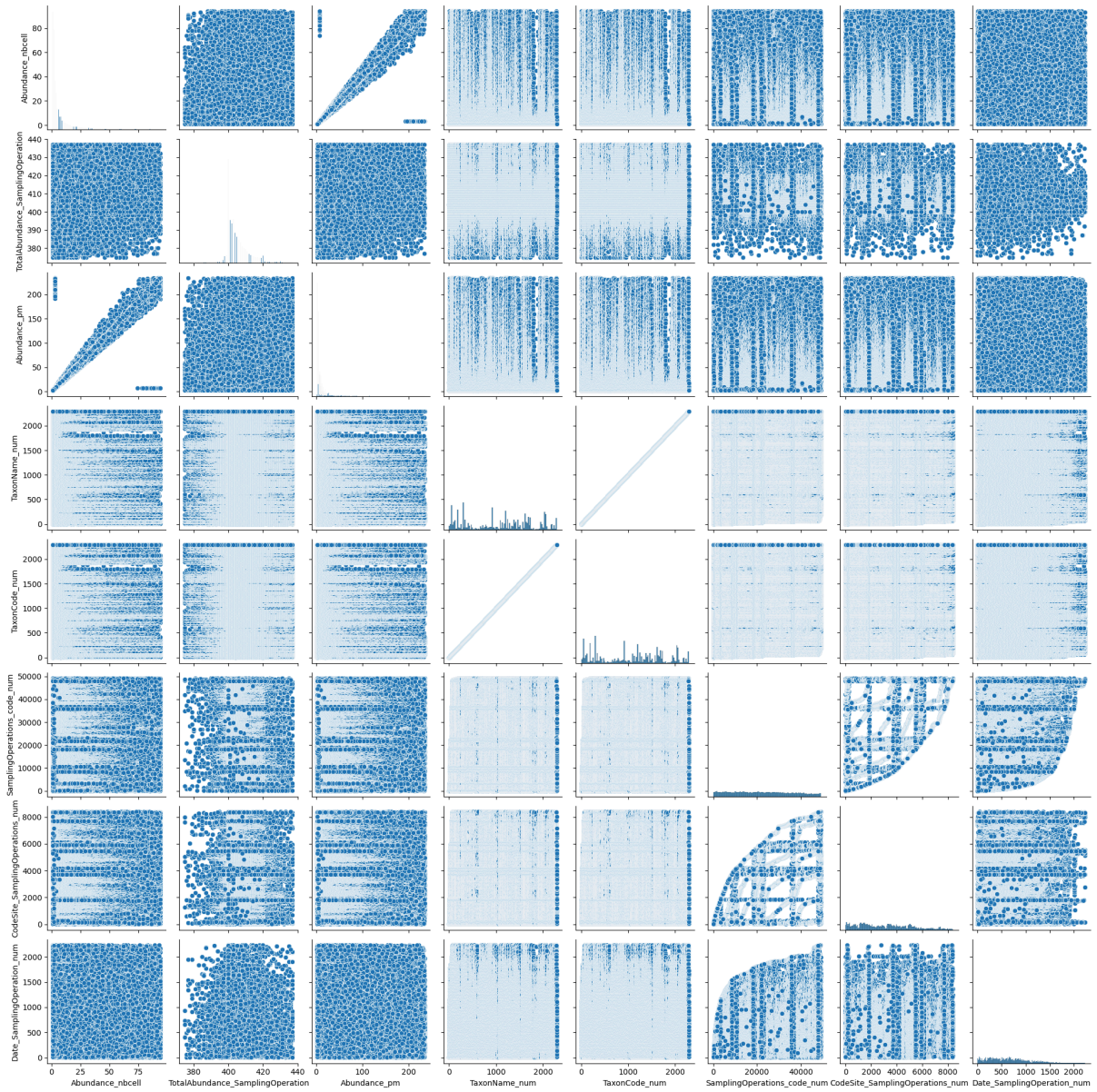
Index	TaxonName_num	TaxonCode_num	SamplingOperations_code_num	CodeSite_SamplingOperations_num	Date_SamplingOperation_num
0	1	1	1	1	1
1	1	1	2	2	2
2	2	2	2	3	3
3	2	2	3	4	4
4	2	2	4	5	5
5	2	2	5	6	6
6	2	2	6	7	7
7	2	2	7	8	8
8	2	2	8	9	9
9	2	2	9	10	10

Con el fin de aplicar técnicas de regresión lineal, fue necesario transformar las variables categóricas en variables numéricas. Para ello, se utilizó la jerarquía de frecuencias, asignando valores más bajos a las categorías con mayor frecuencia de aparición.

La tabla muestra el resultado de esta transformación para las variables *TaxonName*, *TaxonCode*, *SamplingOperations_code*, *CodeSite_SamplingOperations* y *Date_SamplingOperation*. De esta forma, se generaron nuevas columnas numéricas (*_num*) que permiten el tratamiento estadístico y la construcción de modelos de regresión.

Heatmap de correlaciones

Dispersión de variables

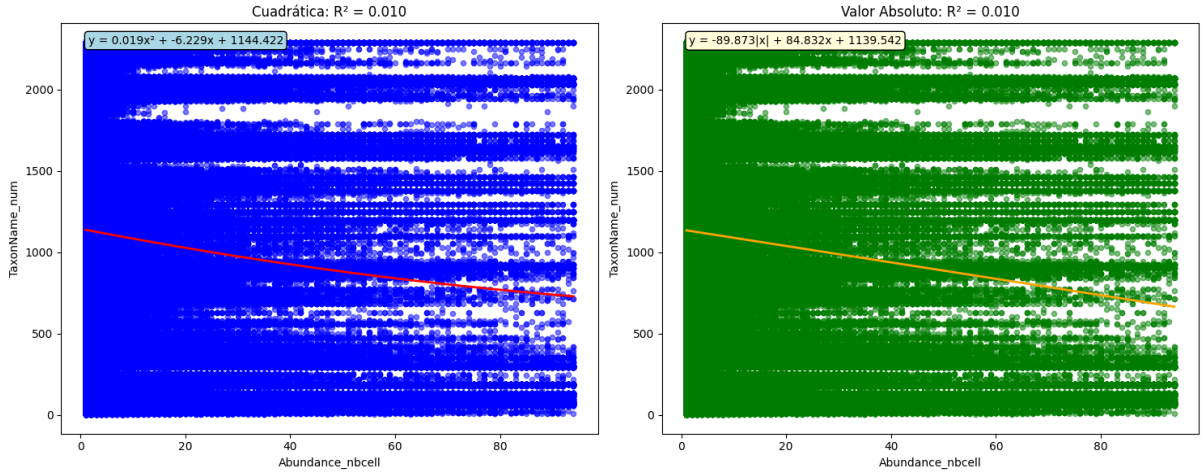


4. Aplicación de modelos no lineales

En esta sección se presentan los resultados obtenidos tras aplicar los modelos de regresión no lineal polinómica y exponencial al conjunto de datos *01_DiatomInventories_GTstudentproject_B.csv*. El análisis tuvo como propósito identificar el tipo y la fuerza de las relaciones existentes entre las variables *TaxonName, TaxonCode, SamplingOperations_code, CodeSite_SamplingOperations, Date_SamplingOperation, Abundance_nbcell, TotalAbundance_SamplingOperation* y *Abundance_pm*, con el fin de determinar cuál modelo ofreció un mejor ajuste a los datos observados.

4.1 TaxonName Vs Abundance_nbcell

Para analizar la relación entre el identificador del taxón (*TaxonName_num*) y la abundancia celular (*Abundance_nbcell*), se aplicaron dos modelos de regresión no lineal: una función cuadrática ($y = ax^2 + bx + c$) y una función de valor absoluto ($a|x| + bx + c$)

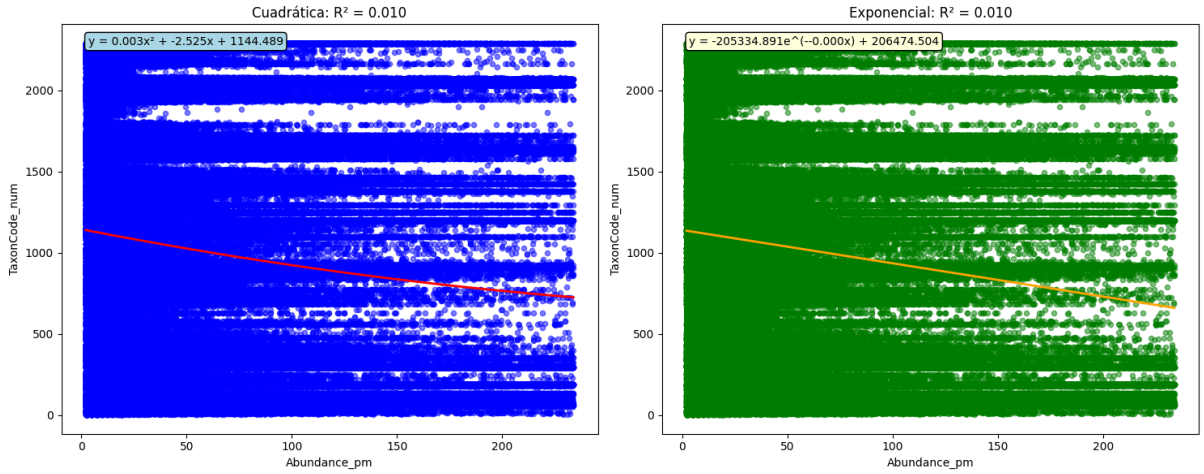


El modelo cuadrático obtuvo un coeficiente de determinación de $R^2 = 0.0101$ y un coeficiente de correlación de $r = 0.1007$, mientras que el modelo de valor absoluto presentó valores prácticamente equivalentes ($R^2 = 0.0101$ y $r = 0.1003$).
Estos resultados indican que la relación entre ambas variables es muy débil, con una capacidad explicativa cercana al 1 % de la variabilidad total en los datos.

Visualmente, las gráficas confirman una dispersión amplia y sin tendencia marcada, donde la mayoría de los puntos se concentran en un rango intermedio de abundancia, sin un patrón definido de incremento o decrecimiento asociado a los distintos taxones.
Aunque ambos modelos ajustan una ligera pendiente negativa, la función cuadrática presenta un trazo más estable y un mejor ajuste local, por lo que puede considerarse marginalmente superior para describir esta relación.

4.2 TaxonCode vs Abundance_pm

En esta relación se buscó evaluar la dependencia entre el código taxonómico (*TaxonCode*), que identifica de forma numérica las especies o morfotipos registrados, y la abundancia promedio por muestra (*Abundance_pm*). Para ello se aplicaron dos modelos de regresión no lineal: una función cuadrática $ax^2 + bx + c$ y una función exponencial decreciente $y = ae^{-bx} + c$, con el propósito de determinar el comportamiento general de los datos y su nivel de ajuste.



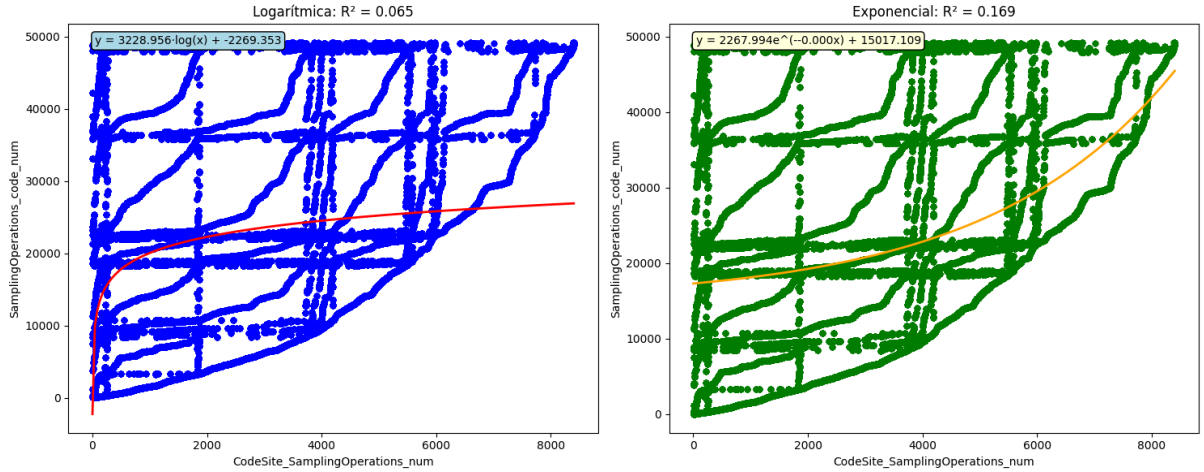
El modelo cuadrático presentó un coeficiente de determinación R² = 0.0102 y una correlación r = 0.1011, mientras que el modelo exponencial obtuvo valores similares (R² = 0.0101, r = 0.1006).Estos resultados reflejan una relación extremadamente débil entre las variables, explicando apenas el 1% de la variabilidad en los datos. La ligera pendiente negativa observada en ambos modelos indica una tendencia decreciente muy tenue: conforme aumenta el código taxonómico, la abundancia promedio tiende a disminuir, aunque el efecto es prácticamente nulo.

Gráficamente, se observa una alta dispersión de los puntos, sin una tendencia clara ni agrupamientos definidos, lo que confirma la ausencia de una relación significativa.

No obstante, la función cuadrática se ajusta ligeramente mejor a los datos, describiendo de manera más estable las variaciones locales y mostrando un trazo más coherente con la distribución observada.

4.3 SamplingOperations_code vs CodeSite_SamplingOperetions_num

En esta sección se analizó la relación entre el código de operación de muestreo (*SamplingOperations_code*) y el código del sitio de muestreo asociado (*CodeSite_SamplingOperations_num*), con el objetivo de identificar si existe una correspondencia sistemática entre ambos.Dado que ambas variables están estructuradas numéricamente y podrían reflejar patrones secuenciales o espaciales, se aplicaron dos modelos de regresión no lineal: una **función logarítmica** y una **función exponencial**

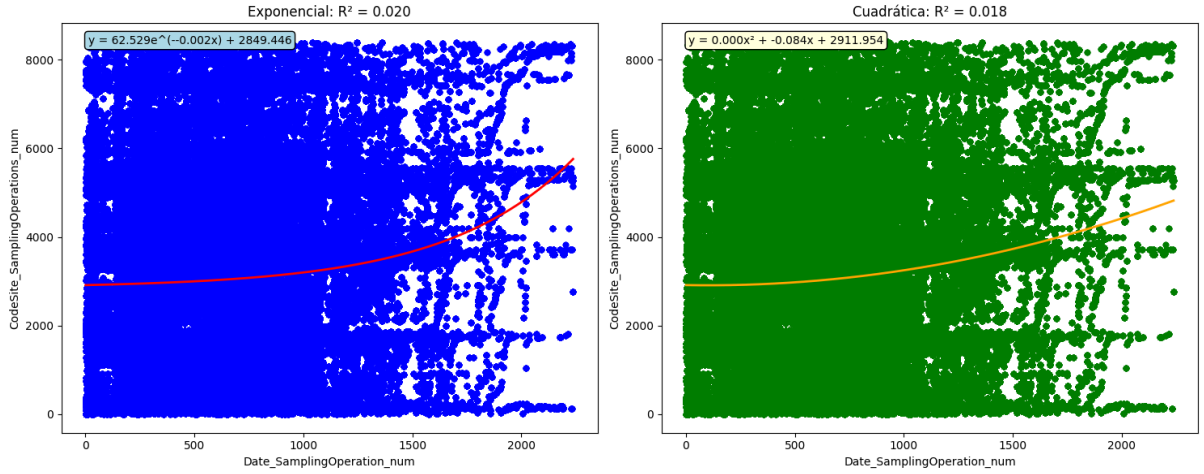


El modelo logarítmico presentó un coeficiente de determinación $R^2 = 0.0654$ y una correlación $r = 0.2558$, lo cual indica una relación débil pero positiva entre ambas variables. La curva ajustada refleja un crecimiento rápido en los valores iniciales que se estabiliza conforme aumenta el número de sitios de muestreo, lo que sugiere un comportamiento asintótico moderado.Por otro lado, el modelo exponencial logró un ajuste significativamente mejor, con $R^2 = 0.1685$ y $r = 0.4105$, evidenciando una correlación positiva moderada. Este resultado sugiere que, a medida que se incrementan los códigos de sitio de muestreo, los códigos de operación asociados también tienden a aumentar de manera progresiva y no lineal.

Desde un punto de vista interpretativo, la relación detectada indica que existe cierta correspondencia estructural entre los códigos de sitio y las operaciones de muestreo, posiblemente derivada del orden sistemático con que se realizaron las campañas o del modo en que se codificaron los sitios en la base de datos. El modelo exponencial, al reflejar un patrón de crecimiento progresivo, puede representar mejor el vínculo entre la frecuencia de muestreo y la distribución de los sitios a lo largo del proyecto.

4.4 CodeSite_SamplingOperations vs Date_SamplingOperation_num

En esta relación se buscó analizar la posible dependencia entre el código del sitio de muestreo (*CodeSite_SamplingOperations_num*) y la fecha de muestreo (*Date_SamplingOperation_num*), con el propósito de identificar si existe un patrón temporal en la asignación o comportamiento de los sitios a lo largo del proyecto.
Para ello se aplicaron dos modelos de regresión no lineal: una **función exponencial** y una **función cuadrática**



El modelo exponencial obtuvo un coeficiente de determinación $R^2 = 0.0199$ y una correlación $r = 0.1409$, mientras que el modelo cuadrático presentó valores ligeramente inferiores ($R^2 = 0.0185$, $r = 0.1360$).En ambos casos, los coeficientes indican una relación muy débil y de baja capacidad explicativa, aunque el modelo exponencial muestra un ajuste marginalmente mejor.

Visualmente, el modelo exponencial refleja una ligera tendencia creciente en los valores del código del sitio conforme avanzan las fechas de muestreo, lo que podría interpretarse como un incremento gradual en la cobertura o en la secuencia de sitios muestreados a lo largo del tiempo. No obstante, la alta dispersión observada en los datos evidencia una falta de estructura clara o regularidad temporal marcada.

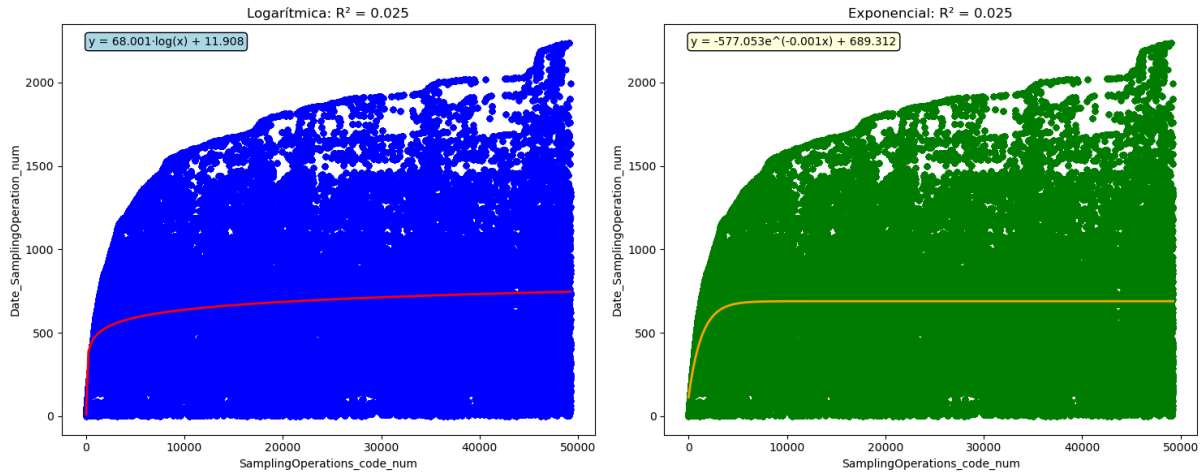
El modelo cuadrático, por su parte, también sugiere un crecimiento suave en los valores del sitio de muestreo conforme a la fecha, pero con un ajuste inferior, lo que confirma que la relación entre ambas variables no sigue un patrón parabólico ni presenta puntos de inflexión relevantes.

4.5 Abundance_nbcell vs SamplingOperations_code_num

En esta relación se buscó identificar si la abundancia celular total (*Abundance_nbcell*) muestra alguna dependencia respecto al código de operación de muestreo (*SamplingOperations_code_num*), es decir, si el número de células registradas por muestra presenta algún patrón asociado a la secuencia o tipo de

muestreo. Para ello se aplicaron dos modelos de regresión no lineal: una **función de valor absoluto** y una **función cociente de polinomios**

4.5 Date_SamplingOperation_num vs SamplingOperations_code_num



Se analizaron los datos de las variables Date_SamplingOperation_num y SamplingOperations_code_num aplicando modelos de regresión no lineal logarítmica y exponencial.

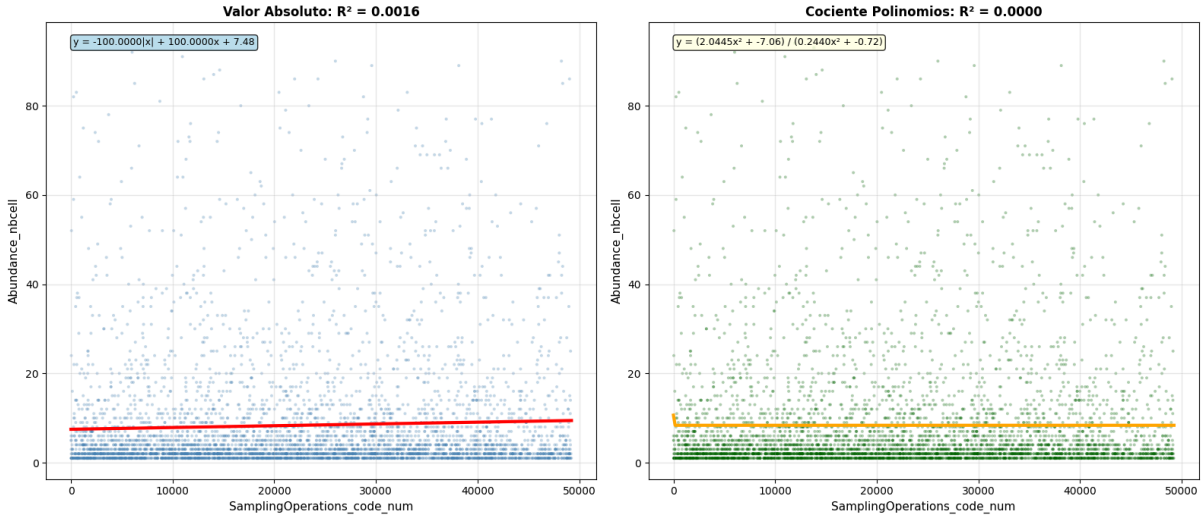
El modelo logarítmico, descrito por la ecuación:
$Y = a \cdot \log(X) + b$,
presentó un coeficiente de determinación $R^2 = 0.025$ y una correlación no lineal $r = 0.1580$, mientras que el modelo exponencial, definido como:
$Y = a \cdot \exp(-bX) + c$,
mostró resultados prácticamente idénticos ($R^2 = 0.025$, $r = 0.1580$).

Estas correlaciones, aunque bajas, superan ligeramente la correlación lineal inicial ($r = 0.1207$), lo que indica una leve mejora al aplicar modelos no lineales. Sin embargo, los valores de R^2 sugieren que la relación entre ambas variables es muy débil y no significativa.

Visualmente, las gráficas exhiben bandas horizontales discretas y una alta concentración de puntos en ciertos rangos, reflejando una expansión gradual pero sin una tendencia definida. Esto sugiere que las operaciones de muestreo (SamplingOperations_code_num) no tienen una dependencia directa o sistemática respecto al tiempo o número de muestreo (Date_SamplingOperation_num).

En conclusión, aunque ambos modelos muestran un leve incremento en la correlación respecto a la lineal, ninguno evidencia una relación estadísticamente significativa, lo que implica que la variabilidad en las operaciones de muestreo ocurre de manera independiente al número o fecha de muestreo.

4.6 Abundance_nbcell vs SamplingOperations_code_num



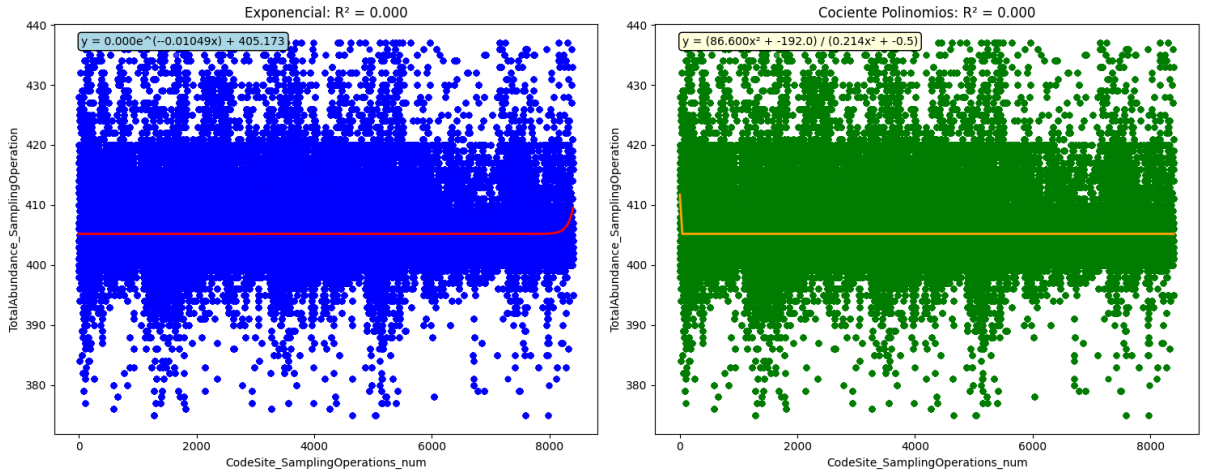
El modelo de valor absoluto obtuvo un coeficiente de determinación $R^2 = 0.0016$ y una correlación r = 0.0395, mientras que el modelo cociente de polinomios presentó valores aún menores ($R^2 = 0.0000$, r = 0.0015). Estos resultados evidencian que la relación entre ambas variables es prácticamente nula, con una capacidad explicativa inferior al 0.2% de la variabilidad total.

En las gráficas, se observa una alta dispersión de los datos, concentrados mayormente en valores bajos de abundancia, sin una tendencia clara de incremento o decrecimiento conforme aumentan los códigos de muestreo. La línea de ajuste del modelo de valor absoluto presenta una pendiente mínima positiva, indicando un incremento casi imperceptible de la abundancia con respecto al número de muestreo, mientras que el modelo de cociente de polinomios genera una curva prácticamente constante.

4.6 TotalAbundance_SamplingOperation vs CodeSite_SamplingOperations_num

En esta sección se analizó la relación entre la abundancia total por operación de muestreo (*TotalAbundance_SamplingOperation*) y el código del sitio de muestreo (*CodeSite_SamplingOperations_num*), con el objetivo de determinar si existe una asociación entre el número de sitio y la magnitud total de organismos observados en cada muestreo.Se evaluaron dos modelos de regresión no lineal: una **función exponencial** y una **función cociente de polinomios**

4.7TotalAbundance_SamplingOperation vs CodeSite_SamplingOperations_num



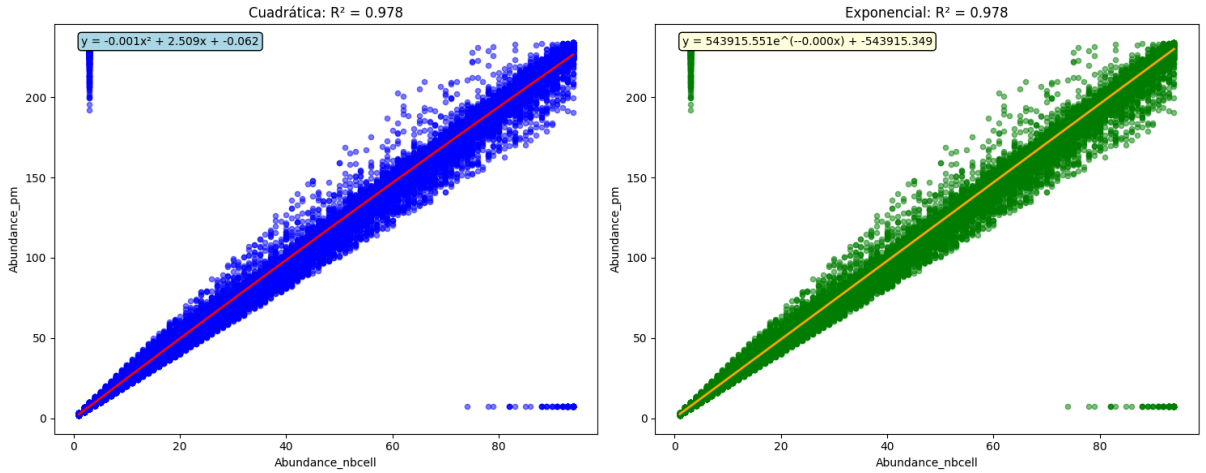
El modelo exponencial presentó un coeficiente de determinación $R^2 = 0.0004$ y una correlación $r = 0.0211$, mientras que el modelo cociente de polinomios obtuvo un $R^2 = 0.0003$ y $r = 0.0174$.Ambos resultados indican una relación prácticamente nula entre las variables, con una capacidad explicativa inferior al 0.1 % de la variabilidad total.

Visualmente, las gráficas muestran una nube de puntos con una distribución homogénea y sin tendencia discernible. Las curvas de ambos modelos se mantienen prácticamente constantes en torno al promedio general de abundancia, sin mostrar un patrón ascendente ni descendente claro. El modelo exponencial, aunque con diferencias mínimas, logra un ajuste ligeramente más estable a los valores medios, lo que justifica su consideración como el modelo con mejor desempeño relativo.

4.7 Abundance_pm vs Abundance_nbcell

Esta relación evaluó la correspondencia entre la abundancia promedio por muestra (*Abundance_pm*) y la abundancia total de células por muestra (*Abundance_nbcell*), con el propósito de determinar el grado de dependencia funcional entre ambas variables biológicas. Dado que ambas representan mediciones directas de la abundancia, se esperaba observar una relación estrecha y positiva. Para comprobarlo, se aplicaron dos modelos de regresión no lineal: una **función cuadrática** y una **función exponencial**

4.8 Abundance_pm vs Abundance_nbcell



El modelo cuadrático obtuvo un coeficiente de determinación $R^2 = 0.9783$ y una correlación $r = 0.9891$, mientras que el modelo exponencial alcanzó valores muy similares ($R^2 = 0.9782$, $r = 0.9890$). Estos resultados evidencian una relación casi perfecta entre ambas variables, con un nivel de ajuste superior al 97%, lo que indica que el comportamiento de *Abundance_pm* puede explicarse casi completamente en función de *Abundance_nbcell*.

En la representación gráfica, ambos modelos generan curvas muy similares que siguen de manera precisa la tendencia lineal ascendente de los datos. No obstante, la función cuadrática muestra un ajuste ligeramente superior, especialmente en los valores intermedios y altos de abundancia, donde logra capturar pequeñas desviaciones del comportamiento estrictamente lineal.

Conclusiones

Variables	Función	R^2	Correlación (r)
TaxonName vs Abundance_nbcell	Cuadrática	0.010	0.101
TaxonName vs Abundance_nbcell	Valor Absoluto	0.010	0.100
TaxonCode vs Abundance_pm	Cuadrática	0.010	0.101
TaxonCode vs Abundance_pm	Exponencial	0.010	0.101

SamplingOperations vs CodeSite	Logarítmica	0.065	0.256
SamplingOperations vs CodeSite	Exponencial	0.169	0.411
CodeSite vs Date_SamplingOperation	Exponencial	0.020	0.141
CodeSite vs Date_SamplingOperation	Cuadrática	0.018	0.136
Date_SamplingOperation vs SamplingOperations	Logarítmica	0.025	0.158
Date_SamplingOperation vs SamplingOperations	Exponencial	0.025	0.158
Abundance_nbcell vs SamplingOperations	Valor Absoluto	0.002	0.040
Abundance_nbcell vs SamplingOperations	Cociente Polinomios	0.000	0.001
TotalAbundance vs CodeSite	Exponencial	0.000	0.021
TotalAbundance vs CodeSite	Cociente Polinomios	0.000	0.017
Abundance_pm vs Abundance_nbcell	Cuadrática	0.978	0.989
Abundance_pm vs Abundance_nbcell	Exponencial	0.978	0.989

El análisis de regresión no lineal evidenció que la mayoría de las relaciones entre variables presentan baja correlación ($R^2 < 0.2$), indicando una escasa dependencia entre los factores de muestreo y las medidas de abundancia. Sin embargo, la relación entre Abundance_pm y Abundance_nbcell mostró un ajuste casi perfecto ($R^2 \approx 0.978$, $r \approx 0.989$), confirmando la coherencia interna y consistencia biológica de los datos. El modelo cuadrático se destacó como el más preciso, mientras que el exponencial tuvo un desempeño moderado en algunas relaciones estructurales. En conjunto, los resultados sugieren que las variaciones de abundancia dependen principalmente de factores externos no representados en las variables analizadas.