Ciencia de Datos Aplicada a Ciencias de la Tierra

Tema: Análisis exploratorio de datos (AED)

Martín A. Díaz-Viera[1], Farhid M. Elisea Guerrero[2]

[1] *mdiazv@imp.mx*, [2] *felisea@imp.mx*

23 de febrero de 2024

Contenido I

1. Análisis Exploratorio de Datos
   - ¿Qué es el AED?
   - Importancia del AED
   - Etapas de un AED
   - Herramientas del AED
   - Datos del caso de estudio
2. Análisis estadístico univariado
   - Análisis estadístico básico
   - Histogramas
   - BoxPlots
   - Q-Q plots
   - Detección de valores atípicos (outliers)
3. Análisis estadístico bivariado

Contenido II

- Análisis de dependencia global
- Matriz de gráficos de dispersión y histogramas
- Mapas de calor de los coeficientes de correlación
- Análisis más detallado donde existan las mayores dependencias
- Gráficos de dispersión, histogramas, boxplots, medidas de dependencia, valores atípicos

4. Análisis estadístico multivariado
   - Detección de valores atípicos
   - Análisis de componentes principales
   - Análisis factorial
   - Análisis de varianzas
     - ANOVA
     - MANOVA

Contenido III

- Análisis discriminante

Análisis Exploratorio de Datos

Análisis

Exploratorio

de Datos

¿Qué es el AED?

- *Es un conjunto de técnicas estadísticas y gráficas que permiten establecer un buen entendimiento básico del comportamiento de los datos y de las relaciones existentes entre las variables que se estudian.*

¿Qué es el AED?

- El análisis exploratorio de datos (AED) es un paso previo e indispensable para la aplicación exitosa de cualquier método estadístico.
- En particular permite la detección de fallos en el diseño y toma de datos, el tratamiento y/o la evaluación de datos ausentes, la identificación de valores atípicos y la comprobación de los supuestos requeridos por parte de las técnicas geoestadísticas.

Etapas de un AED

- Realizar un examen gráfico de la naturaleza de las variables individuales y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
- Realizar un examen gráfico de las relaciones entre las variables y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
- Evaluar algunos supuestos básicos subyacentes a muchas técnicas estadísticas, por ejemplo, normalidad, linealidad y homocedasticidad.
- Identificar los posibles valores atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
- Evaluar, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

Herramientas del AED

- Análisis estadístico univariado
- Análisis estadístico bivariado
- Análisis estadístico multivariado

Datos del caso de estudio

- Los datos son de un yacimiento de aguas profundas en el Golfo de México.
- El yacimiento es siliciclástico y está formado principalmente por secuencias alternadas de arenas y arcillas.
- Propiedades petrofísicas (Phit, logKtimur, Phie, Sw, Vcl) y atributos sísmicos (Vp, Vs, Rhob, Ip, Is, Poisson, K, G, Lambda) a escala de registro de pozo con un intervalo de muestreo de 0.1 m.
- Una sección vertical (inline) con los mismos atributos sísmicos.
- La malla sísmica es de 33 intervalos de 12.5 m en X y 60 intervalos de 5.5 m aproximadamente en profundidad.

Datos de registros de pozo

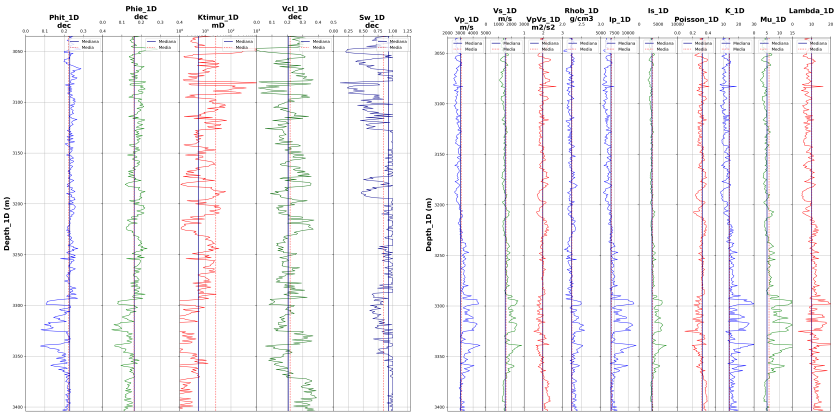

Figura 1: Petrofísica y atributos elásticos.

Sección sísmica vertical (inline) con atributos sísmicos

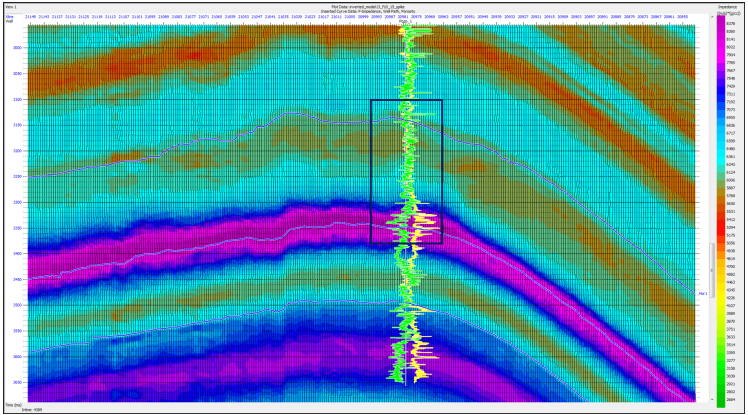

Figura 2: Sección sísmica vertical (inline) con Ip.

Análisis estadístico univariado

Análisis

Estadístico

Univariado

Por la estructura de datos:

De lo general a lo particular

- Global,
- por intervalos,
- por secciones,
- por clasificaciones, etc

Análisis estadístico univariado

- Análisis estadístico básico (mínimo, máximo, mediana, media, etc.)
- Histogramas
- Boxplots
- Q-Q y P-P plots
- Detección de valores atípicos (outliers)

Análisis estadístico básico

Estadística básica de los datos de registros de pozo de las propiedades petrofísicas

Statistics	PHIT_1D	PHIE_1D	KTIMUR_1D	KTIMUR_1D_LOG10	VCL_1D	SW_1D
Size	370	370	370	370.0000	370	370
Minimum	0.0769	0.0473	0.0244	-1.6126	0.0000	0.1115
1st quartile	0.2155	0.1412	2.2379	0.3498	0.1645	0.7331
Mean	0.2223	0.1632	25.6465	0.7590	0.2181	0.8436
Median	0.2294	0.1643	5.3507	0.7284	0.2051	0.9264
3rd quartile	0.2406	0.1879	14.7738	1.1695	0.2841	1.0000
Maximum	0.2878	0.2853	1602.6267	3.2048	0.3896	1.0000
Range	0.2109	0.2380	1602.6023	4.8174	0.3896	0.8885
Std Deviation	0.0326	0.0343	121.8057	0.6492	0.0814	0.1949
Variance	0.0011	0.0012	14836.6360	0.4214	0.0066	0.0380
Skewness	-1.9753	-0.3520	10.4805	0.1917	0.0194	-1.3335
Kurtosis	4.8156	0.6035	121.5155	1.5613	-0.4117	1.1896

Tabla 1: Resumen estadístico de los datos de registros de pozo de las propiedades petrofísicas.

Análisis estadístico básico

Estadística básica de los atributos elásticos a escala de registro de pozo

Statistics	VP_1D	VS_1D	VPVS_1D	RHOB_1D	IP_1D	IS_1D	POISSON_1D	K_1D	MU_1D	LAMBDA_1D
Size	370.00	370.00	370.00	370.00	370.00	370.00	370.00	370.00	370.00	370.00
Minimum	2473.33	1085.63	1.50	2.03	5301.02	2342.89	0.10	7.69	2.55	3.06
1st quartile	2788.53	1389.18	1.83	2.19	6151.65	3120.76	0.29	11.05	4.31	7.90
Mean	3085.22	1583.76	1.97	2.25	6960.85	3573.88	0.32	14.08	5.87	10.17
Median	3016.67	1517.67	1.96	2.24	6856.05	3365.04	0.32	13.78	5.10	10.01
3rd quartile	3280.73	1695.48	2.08	2.31	7293.79	3744.49	0.35	15.83	6.35	11.96
Maximum	4548.46	2815.88	2.67	2.58	11469.07	7135.46	0.41	31.63	20.09	26.32
Range	2075.13	1730.26	1.17	0.55	6168.05	4792.57	0.32	23.94	17.54	23.26
Std Deviation	400.23	285.12	0.17	0.09	1127.62	753.66	0.04	3.99	2.56	2.98
Variance	160180.22	81292.96	0.03	0.01	1271527.84	568002.46	0.00	15.96	6.55	8.88
Skewness	1.24	1.32	0.38	0.75	1.59	1.82	-0.74	1.36	2.19	0.81
Kurtosis	1.64	2.00	0.04	0.93	2.96	3.66	1.22	2.71	5.52	2.22

Tabla 2: Resumen estadístico de los atributos elásticos a escala de registro de pozo.

Histograma

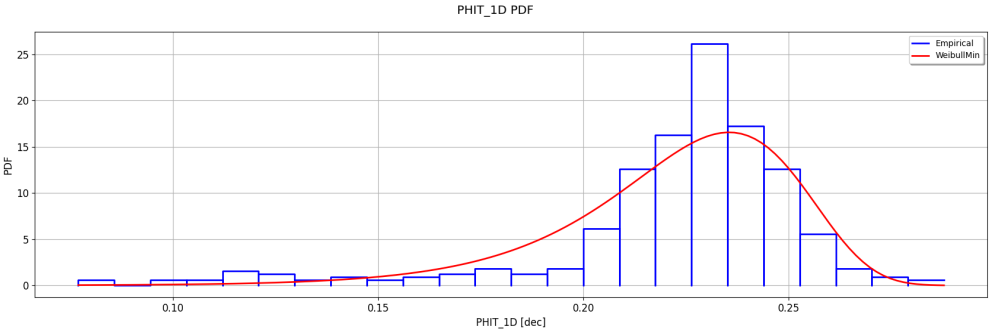

Figura 3: Histograma y función de densidad de probabilidad paramétrica ajustada para PHIT.

Histograma

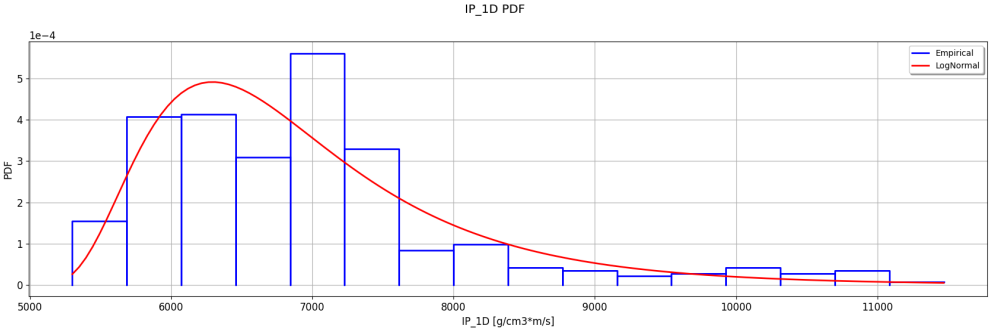

Figura 4: Histograma y función de densidad de probabilidad paramétrica ajustada para IP.

Histograma

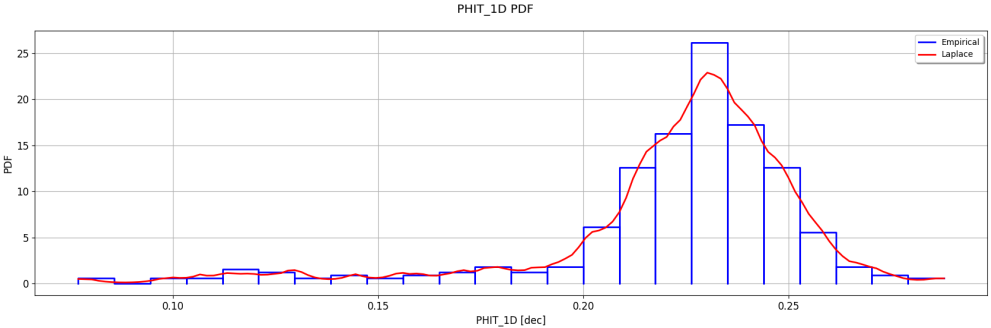

Figura 5: Histograma y función de densidad de probabilidad no paramétrica ajustada para PHIT.

Histograma

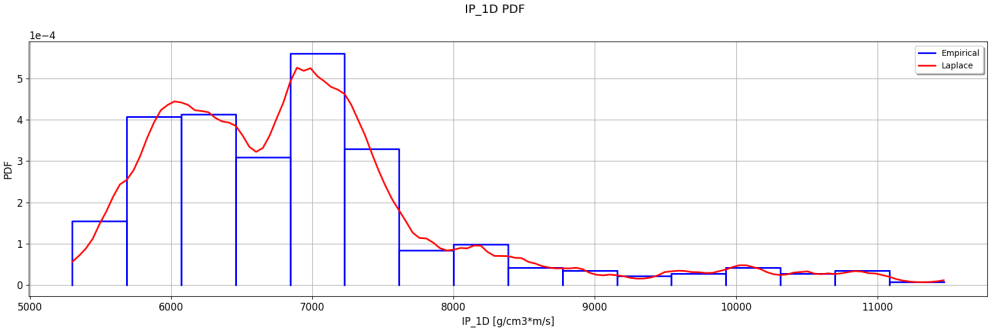

Figura 6: Histograma y función de densidad de probabilidad no paramétrica ajustada para IP.

Histograma acumulativo

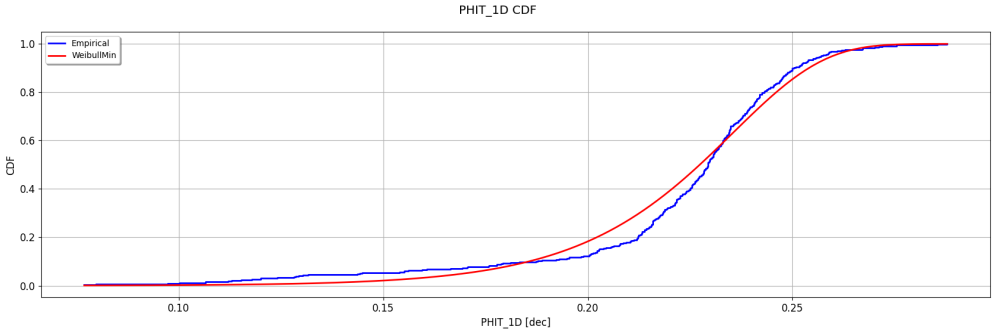

Figura 7: Histograma y función de probabilidad acumulada paramétrica ajustada para PHIT.

Histograma acumulativo

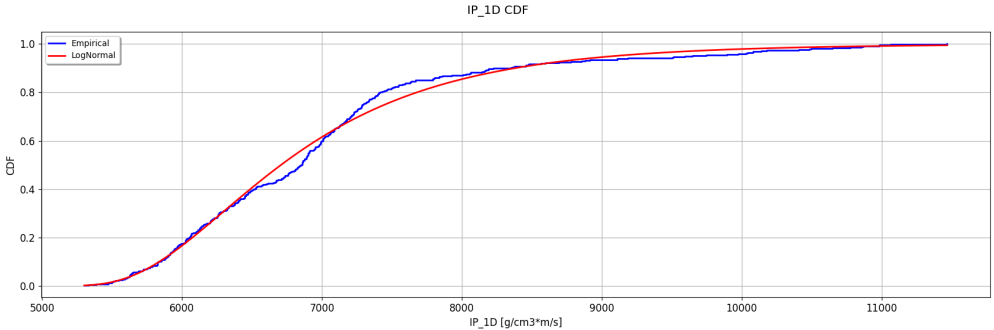

Figura 8: Histograma y función de probabilidad acumulada paramétrica ajustada para IP.

Histograma acumulativo

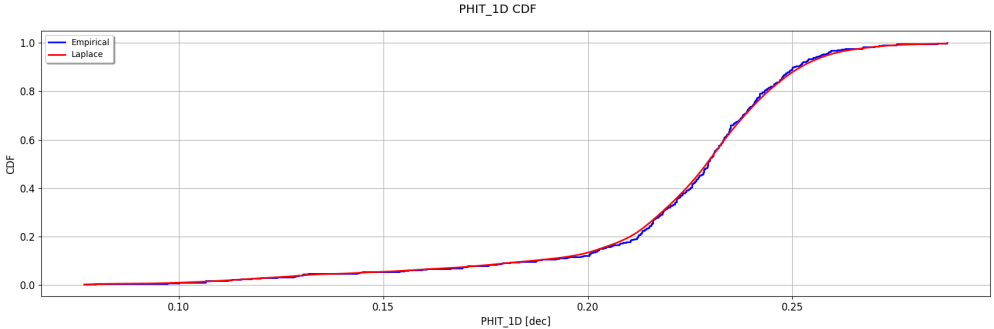

Figura 9: Histograma y función de probabilidad acumulada no paramétrica ajustada para PHIT.

Histograma acumulativo

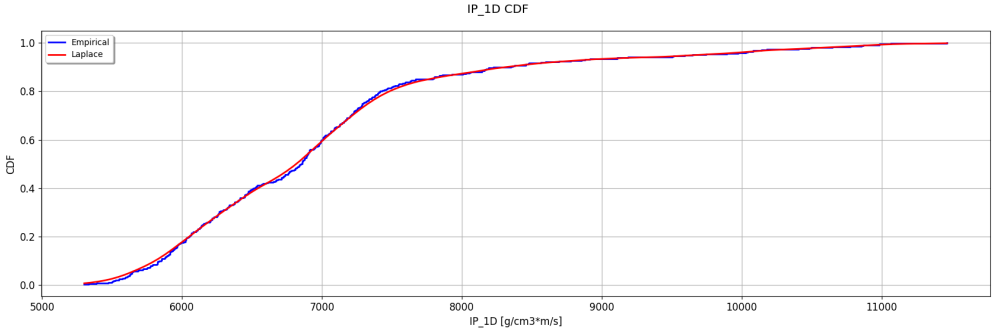

Figura 10: Histograma y función de probabilidad acumulada no paramétrica ajustada para IP.

BoxPlot

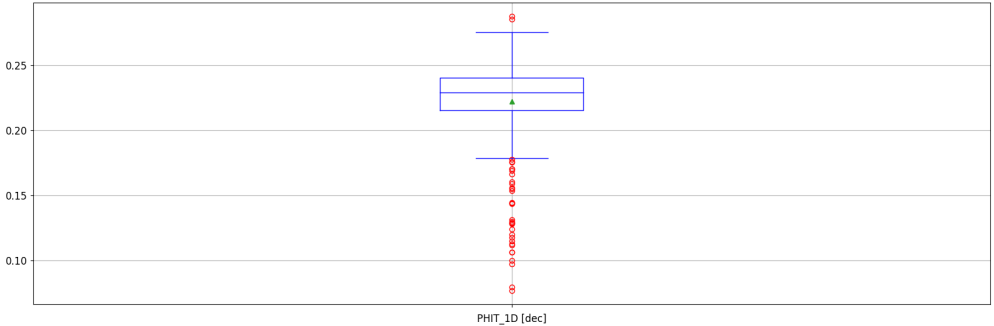

Figura 11: Boxplot para PHIT.

BoxPlot

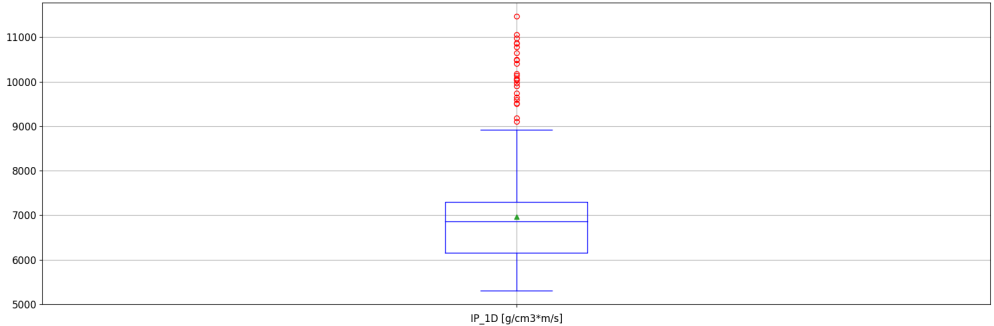

Figura 12: Boxplot para IP.

Q-Q plot

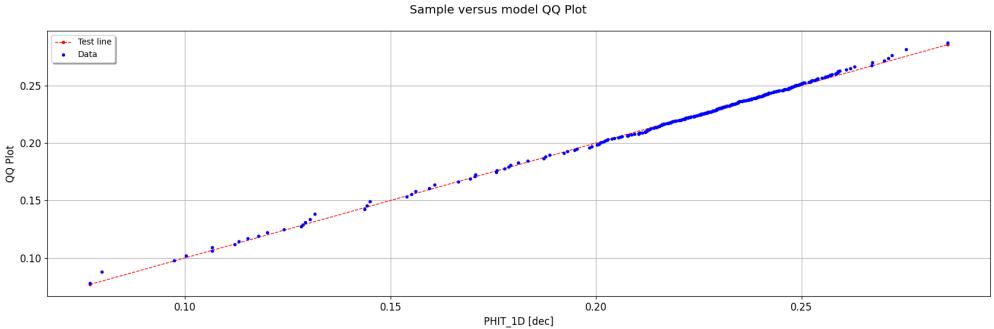

Figura 13: Q-Q plot para PHIT.

Q-Q plot

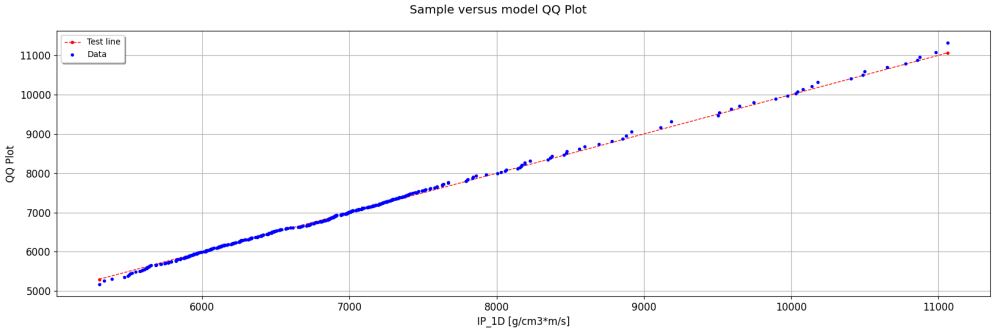

Figura 14: Q-Q plot para IP.

Detección de valores atípicos (outliers)

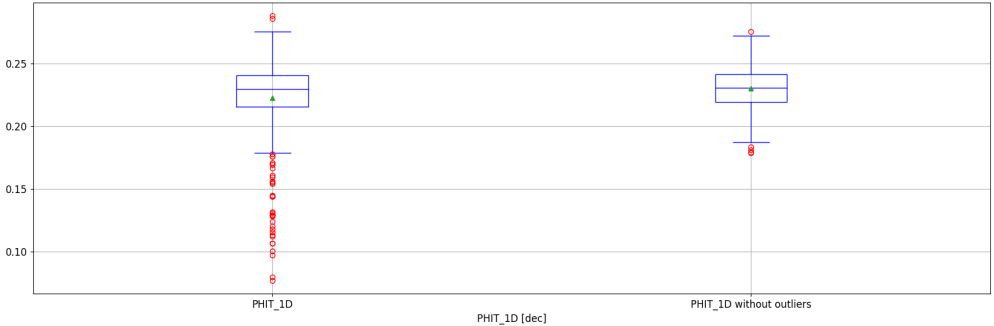

Figura 15: Boxplot con y sin valores atípicos para PHIT.

Detección de valores atípicos (outliers)

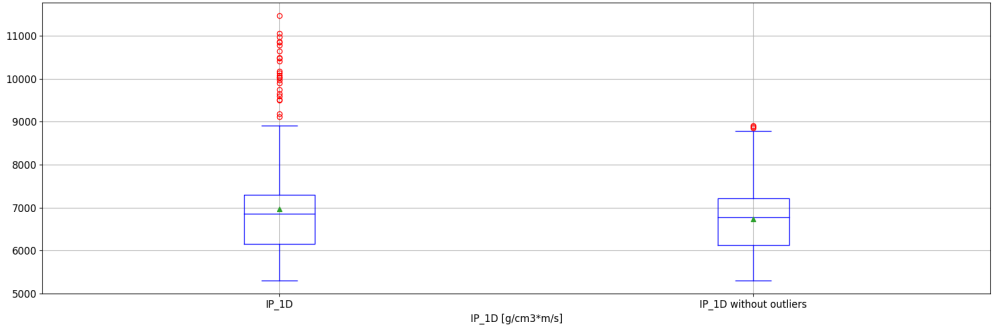

Figura 16: Boxplot con y sin valores atípicos para IP.

Análisis estadístico bivariado

Análisis

Estadístico

Bivariado

Por la estructura de datos:

De lo general a lo particular

- Global,
- por intervalos,
- por secciones,
- por clasificaciones, etc

Análisis estadístico bivariado

- Análisis de dependencia global
  - Matriz de gráficos de dispersión, histogramas
  - Mapas de calor (con medidas de dependencia)
  - Medidas de dependencia (Pearson, Spearman y Kendall)

Matriz de gráficos de dispersión y histogramas

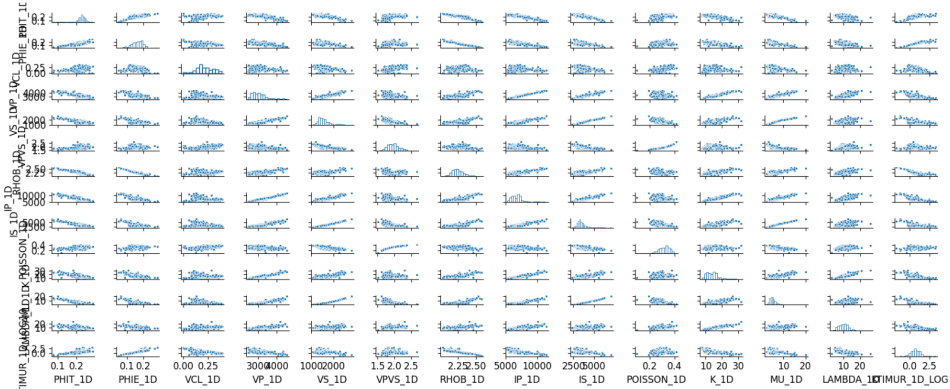

Figura 17: Matriz de gráficos de dispersión y histogramas.

Mapa de calor del coeficiente de correlación de Pearson

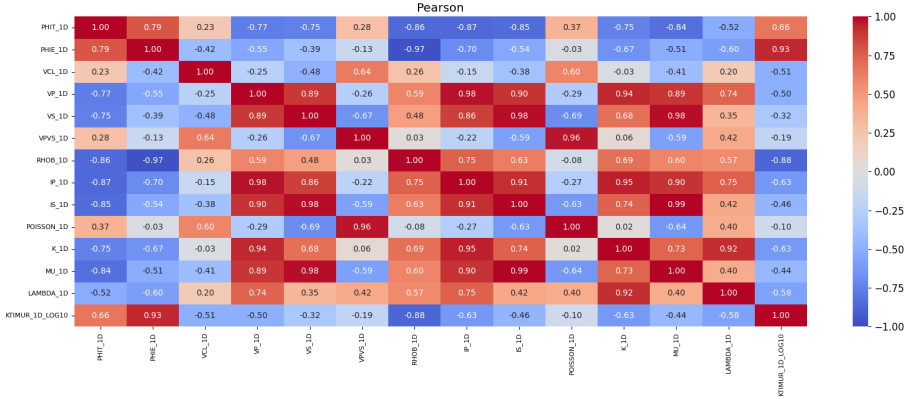

Figura 18: Mapa de calor del coeficiente de correlación de Pearson.

Mapa de calor del coeficiente de correlación de Spearman

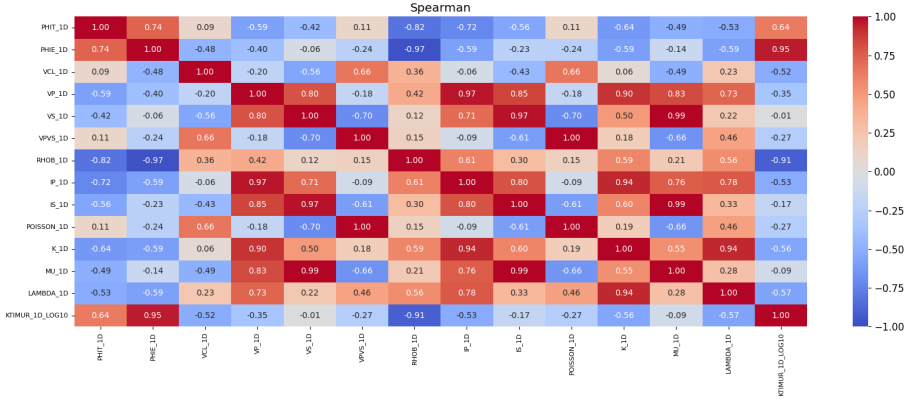

Figura 19: Mapa de calor del coeficiente de correlación de Spearman.

Mapa de calor del coeficiente de correlación de Kendall

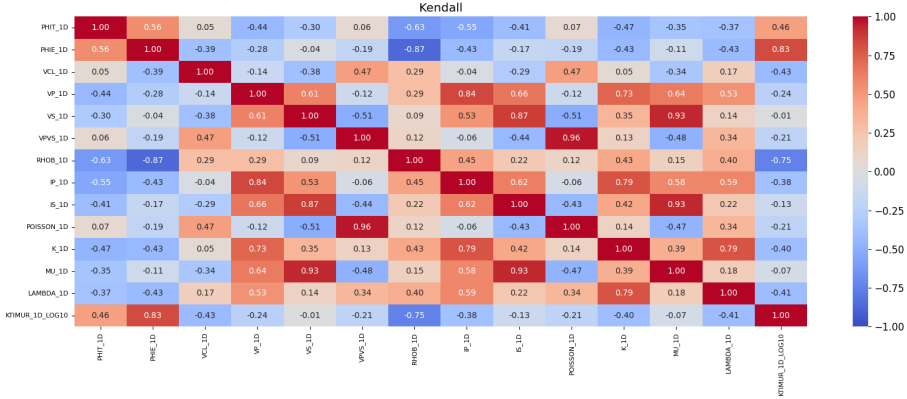

Figura 20: Mapa de calor del coeficiente de correlación de Kendall.

Análisis de dependencia

Mayores dependencias entre propiedades petrofísicas (variable primaria) y atributos elásticos (variable secundaria)

Primary variable	Secondary variable	Pearson	Kendall	Spearman
Phit	Ip	-0.8669	-0.5474	-0.7244
Phie	Rhob	-0.9742	-0.8712	-0.9747
logKtimur	Rhob	-0.8767	-0.7481	-0.9078
Vcl	Poisson	0.5986	0.4748	0.6640

Tabla 3: Coeficientes de correlación.

Análisis estadístico bivariado

- Análisis más detallado donde existan las mayores dependencias
  - Gráficos de dispersión,
  - histogramas,
  - boxplots,
  - medidas de dependencia,
  - valores atípicos

Gráfico de dispersión de Phit vs Ip

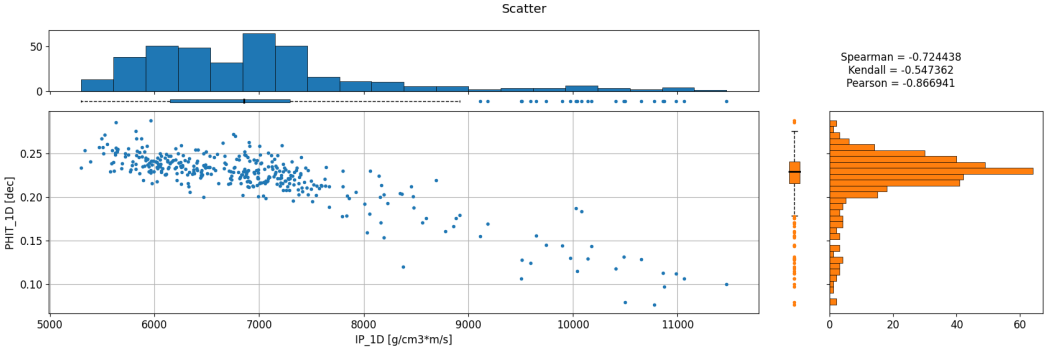

Figura 21: Scatterplot de Phit vs Ip.

Gráfico de dispersión de Phie vs Rhob

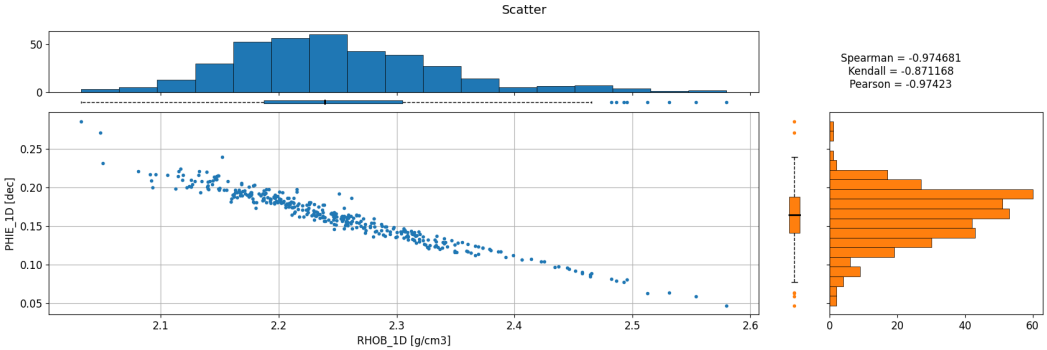

Figura 22: Scatterplot de Phie vs Rhob.

Gráfico de dispersión de logKtimur vs Rhob

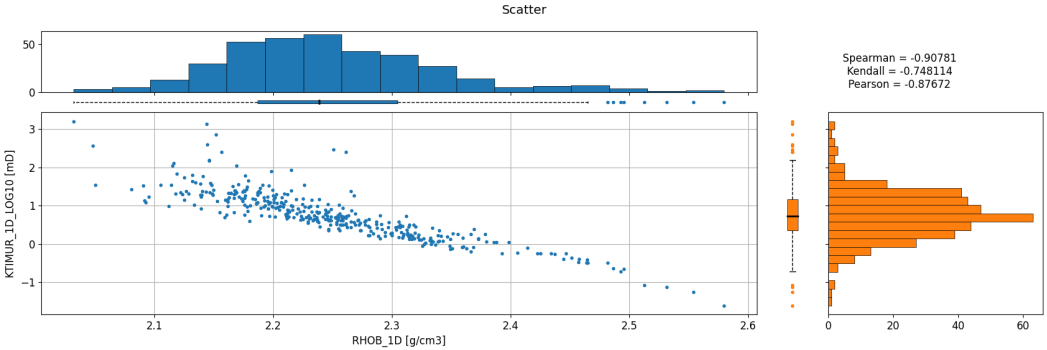

Figura 23: Scatterplot de logKtimur vs Rhob.

Gráfico de dispersión de Vcl vs Poisson

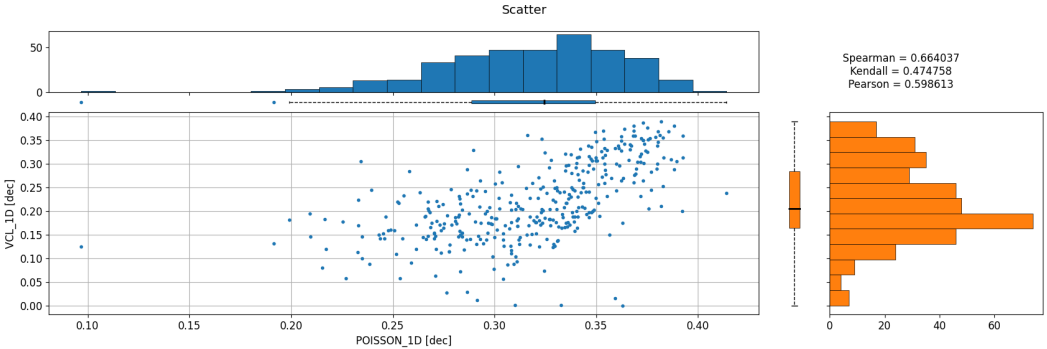

Figura 24: Scatterplot de Vcl vs Poisson.

Análisis estadístico multivariado

Análisis

Estadístico

Multivariado

Por la estructura de datos:

De lo general a lo particular

- Global
- Intervalos (o unidades)
- Secciones
- Clasificaciones
- Combinado. Considera dos o más criterios simultáneos.

Análisis estadístico multivariado

- Detección de valores atípicos
- Análisis de Regresión
- Análisis de Agrupamiento
- Análisis de Componentes Principales (ACP)
- Análisis Factorial (AF)
- Análisis de Varianzas (ANOVA/MANOVA)
- Análisis Discriminante (LDA), etc

Análisis estadístico multivariado

VARIABLES DEPENDIENTES		VARIABLES INDEPENDIENTES	MÉTODO	OBJETIVOS
CONTINUAS	1	CATEGÓRICA	ANOVA	DEPENDENCIA
		CONT/CATEGÓR.	REGRESIÓN	DEPENDENCIA
	VARIAS	CATEGÓRICA	MANOVA	DEPENDENCIA
		CONT/ CATEGÓR.	REGRESIÓN MULTIPLE	DEPENDENCIA
		CONT/ CATEGÓR.	ANÁLISIS CANÓNICO	INDEPENDENCIA
CATEGÓRICAS	2 CAT.	CONT/CATEGOR.	A. DISCRIMINANTE.	CLASIFICACIÓN
		CONT/CATEGOR.	A. LOGIT BINOMIAL	CLASIF/DEPENDENCIA
	> 2 CAT.	CONT/CATEGOR.	A. DISCRIMINANTE.	CLASIFICACIÓN
		CONT/CATEGOR.	A. LOGIT POLINOMIAL	CLASIF/DEPENDENCIA
NINGUNA		CONT/CATEGOR	A. COMP.PRINCIPALES	REDUCC. DIMENSIÓN
		CONT/CATEGOR	A. FACTORIAL	REDUCC. DIMENSIÓN
		CONT/CATEGOR	A. CORRESPONDENCIAS	REDUCC. DIMENSIÓN
		CONT/CATEGOR	ESCALAS MULTIDIMENSIONALES	AGRUPACIÓN
		CONT/CATEGOR	ANÁLISIS CLUSTER	AGRUPACIÓN

Detección de valores atípicos

Detección de valores atípicos

Detección de valores atípicos

1. **Mediante la distancia de Mahalanobis (DM)**
   A la distancia de Mahalanobis previamente generada, se le calcula el primer cuartil (1Q), el tercer cuartil (3Q) y el rango intercuartil (RIQ), considerando los siguientes rangos:
   - Es atípico si DM $<$ 1Q – 1.5*RIQ.
   - Es atípico si DM $>$ 3Q + 1.5*RIQ.
2. **Mediante la Prueba Chi cuadrada (χ^2)**
   A la DM obtenida, se le calcula la Prueba Chi cuadrada (χ^2) con n-1 grados de libertad (n = no. de variables), y posteriormente se selecciona alguno de los criterios siguientes:
   - Etiquetar como atípicos las instancias con (1-χ^2) $<$ 0.001, ó
   - Calcular el cuantil 97.5 % (Q97_5) de χ^2 y etiquetar como atípicos las instancias donde DM $>$ DM en Q97_5.

Otros métodos avanzados

- **Factor local de valores atípicos (LOF)**
  Es un algoritmo propuesto por Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng y Jörg Sander en 2000 para encontrar puntos de valores atípicos midiendo la desviación local de un punto de datos dado con respecto a sus vecinos. LOF comparte algunos conceptos como "distancia central" y "distancia de alcance", que se utilizan para la estimación de la densidad local.
- **Bosque de aislamiento (IF)**
  Detecta anomalías mediante el aislamiento (qué tan lejos está un punto de datos del resto de los datos), en lugar de modelar los puntos normales. Desarrollado por Fei Tony Liu y Zhi-Hua Zhou en 2007. Aísla explícitamente valores atípicos utilizando árboles binarios. El algoritmo tiene una complejidad de tiempo lineal con una constante baja y un requisito de memoria bajo, lo que funciona bien con datos de gran volumen.

Análisis de componentes principales

Análisis de Componentes Principales

Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA)

Es una técnica de reducción de la dimensionalidad de los datos, que trata de tranformar un conjunto de variables en otro conjunto, de menor cantidad de variables, con la particularidad de que las nuevas variables no estén correlacionadas entre sí.

Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA)

- Permite analizar grandes conjuntos de datos
- Con gran cantidad de variables por observación
- Aumenta la interpretabilidad de los datos
- Conserva la máxima cantidad de información
- La variación de los datos se puede describir con menos variables.
- Facilita la visualización de datos multidimensionales.

Análisis de componentes principales (PCA)

- Reduce la dimensionalidad de un conjunto de datos.
- Es una transformación lineal de los datos en un nuevo sistema de coordenadas (componentes).
- Las componentes son combinaciones lineales de las variable y son independientes entre sí
- Los primeros 2 o 3 componentes contienen el mayor porciento de la varianza.
- Facilita la visualización de grupos de datos estrechamente relacionados.

Análisis de componentes principales (PCA)

- Está estrechamente relacionado con el análisis factorial.
- El análisis factorial incorpora suposiciones más específicas sobre la estructura subyacente.
- PCA define un nuevo sistema de coordenadas ortogonales que describe de manera óptima la varianza en un único conjunto de datos.

Matriz de gráficos de dispersión y histogramas

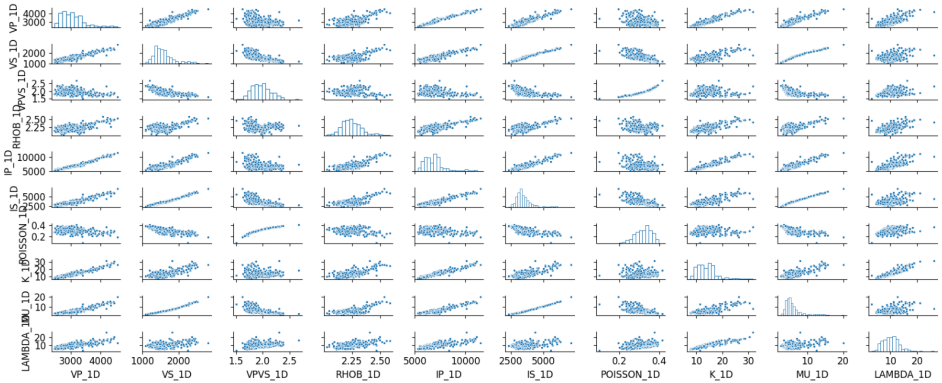

Figura 26: Matriz de gráficos de dispersión y histogramas.

Mapa de calor del coeficiente de correlación de Pearson

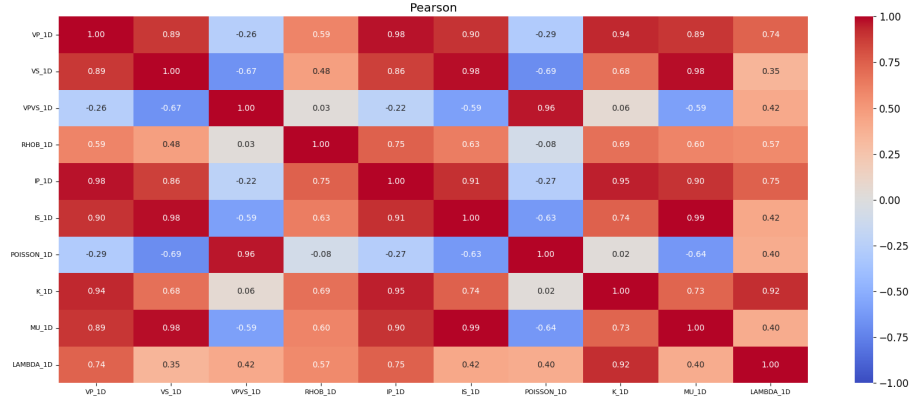

Figura 27: Mapa de calor del coeficiente de correlación de Pearson para los atributos elásticos.

Mapa de calor del coeficiente de correlación de Spearman

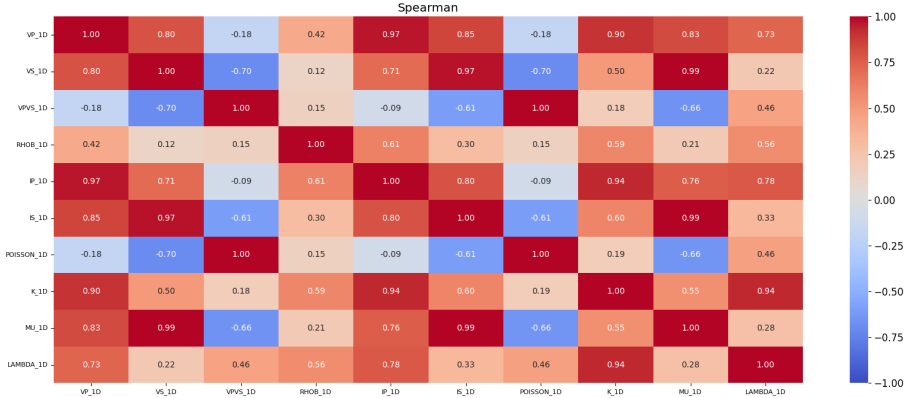

Figura 28: Mapa de calor del coeficiente de correlación de Spearman para los atributos elásticos.

Mapa de calor del coeficiente de correlación de Kendall

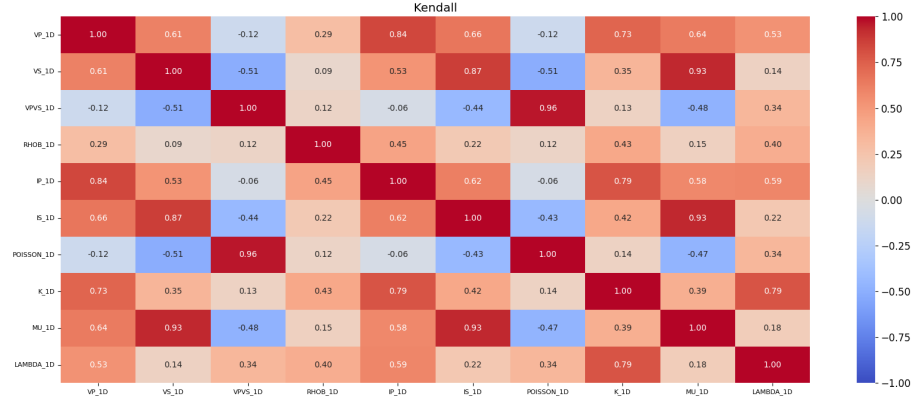

Figura 29: Mapa de calor del coeficiente de correlación de Kendall para los atributos elásticos.

Componentes principales de los atributos elásticos

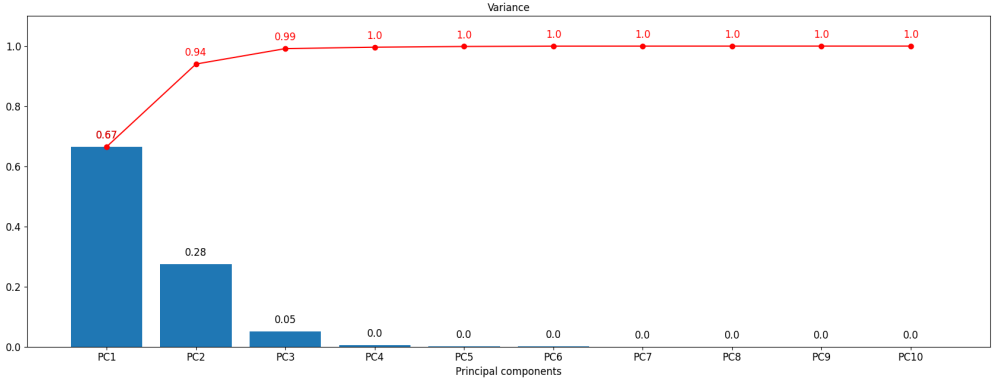

Figura 30: Componentes principales de los atributos elásticos.

Matriz de pesos de los componentes principales de los atributos elásticos

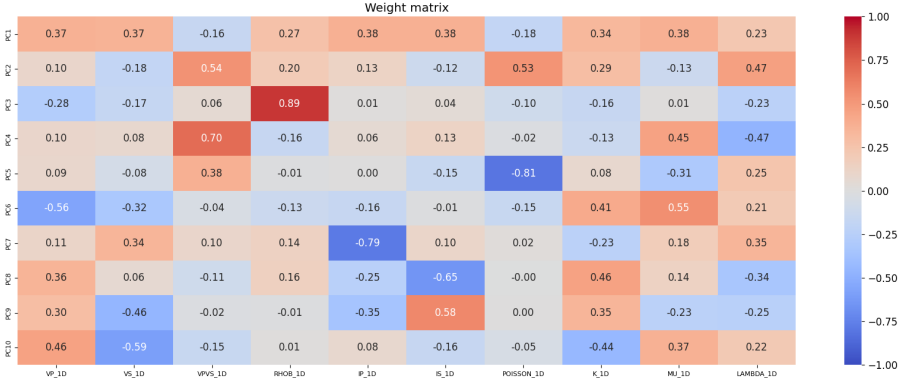

Figura 31: Matriz de los pesos de los componentes principales de los atributos elásticos.

Matriz de dispersión y histogramas de PC1 vs Propiedades Petrofísicas

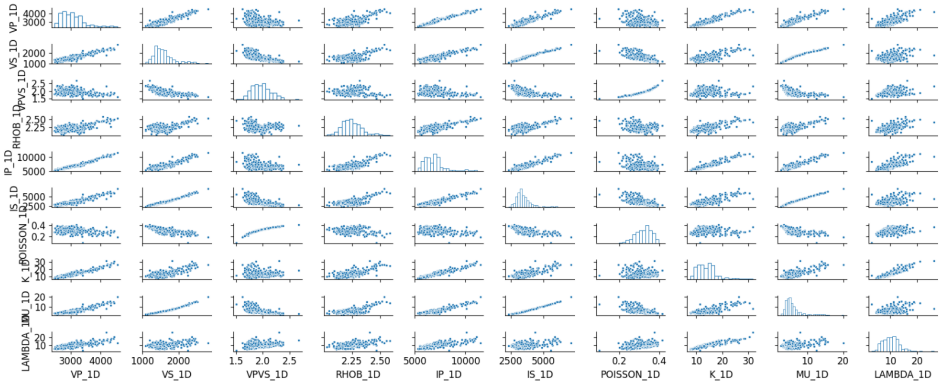

Figura 32: Matriz de dispersión y histogramas de PC1 vs Propiedades Petrofísicas.

Mapa calor coef. corr. de Pearson de PC1 vs Propiedades Petrofísicas

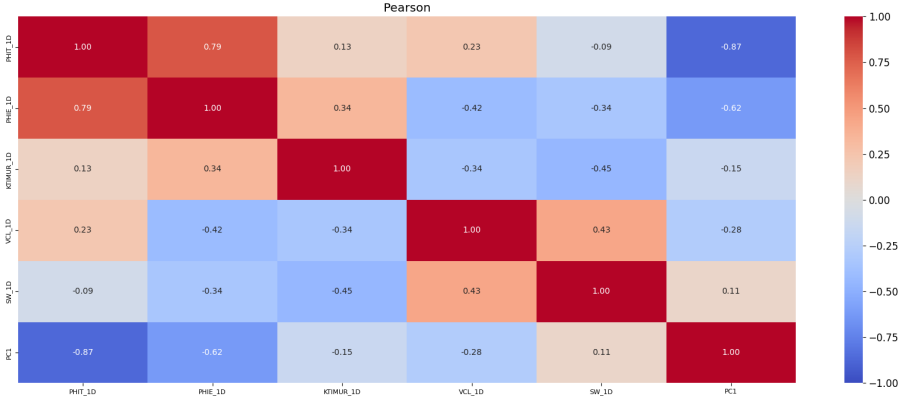

Figura 33: Mapa calor coef. corr. de Pearson de PC1 vs Propiedades Petrofísicas.

Mapa calor coef. corr. de Spearman de PC1 vs Propiedades Petrofísicas

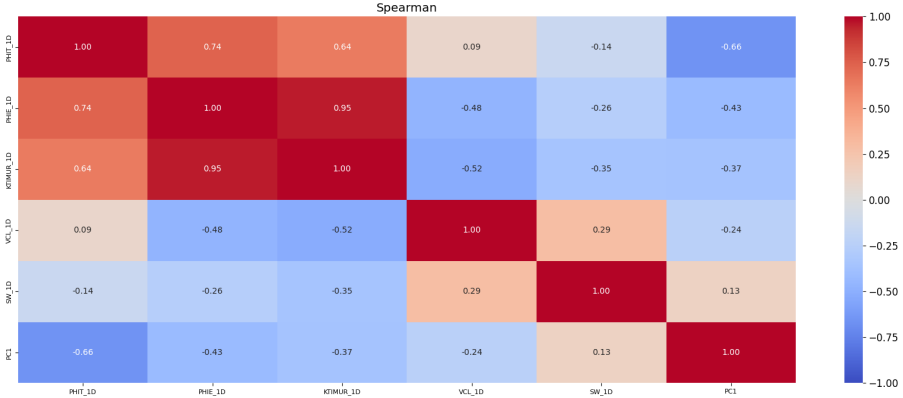

Figura 34: Mapa calor coef. corr. de Spearman de PC1 vs Propiedades Petrofísicas.

Mapa calor coef. corr. de Kendall de PC1 vs Propiedades Petrofísicas

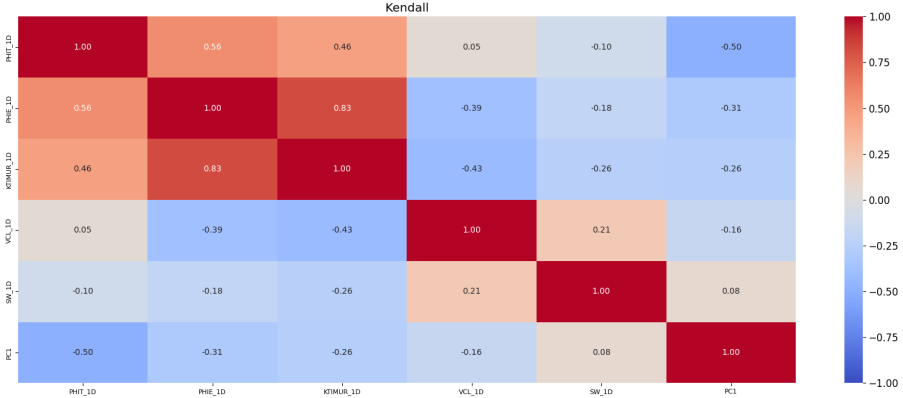

Figura 35: Mapa calor coef. corr. de Kendall de PC1 vs Propiedades Petrofísicas.

Análisis factorial

Análisis Factorial

Análisis factorial (FA)

El análisis factorial (FA)

Su pretensión es similar al A.C.P. pero aquí se formula un modelo teórico en el que se explica el comportamiento de las variables observables en función de unos factores (comunes) que se pretenden obtener y unos factores específicos.

Análisis factorial (FA)

El análisis factorial (FA)

- Describe la variabilidad de las variables observadas en términos de un número menor no observadas, llamadas factores.
- Se modelan las variables como combinaciones lineales de los factores.
- La carga factorial de una variable cuantifica el grado en que la variable está relacionada con un factor dado.
- Es una de las técnicas de interdependencia más utilizadas.
- El objetivo es descubrir los factores latentes.

Análisis factorial (FA)

- FA es similar al PCA, en que ambos implican combinaciones lineales.
- FA es un enfoque centrado en la correlación entre variables (términos fuera de la diagonal de la matriz de covarianza)
- PCA se centra en explicar los términos que se encuentran en la diagonal
- Se prefiere PCA para fines de reducción de datos
- FA se usa cuando el propósito es detectar la estructura de datos (factores latentes)

Análisis de varianzas

Análisis de Varianzas

Análisis de varianzas (ANOVA)

El análisis de varianza (ANOVA)

Tiene por objeto determinar en qué medida una variable dependiente de naturaleza continua está condicionada por los valores que toman variables independientes de naturaleza categórica, llamadas factores.

Análisis de varianzas (ANOVA)

El análisis de varianza (ANOVA)

- Es una colección de procedimientos estadísticos de estimación asociados a la variación entre grupos
- Se basa en la ley de la varianza total,
- La varianza observada en una variable particular se divide en componentes atribuibles a diferentes fuentes de variación
- Proporciona una prueba estadística de si dos o más medias de población son iguales

Análisis de varianzas (MANOVA)

El análisis de varianza multivariado (MANOVA)

Es la generalización del ANOVA, para el caso en que el número de variables dependientes de naturaleza continua sea superior a uno.

Análisis de varianzas (MANOVA)

El análisis de varianza multivariado (MANOVA)

- Es una forma generalizada de análisis de varianza univariado (ANOVA)
- Es un procedimiento para comparar medias muestrales multivariadas
- Se utiliza cuando hay dos o más variables dependientes
- Usa pruebas de significancia que involucran variables dependientes individuales por separado

Análisis discriminante

Análisis Discriminante

Análisis discriminante

El análisis discriminante lineal (LDA)

Se utiliza para caractizar mediante un conjunto de variables independientes, las diferencias existentes entre individuos de distintos grupos y también para clasificar nuevos casos en uno de esos grupos a partir de la información sobre las variables consideradas.

Análisis discriminante

El análisis discriminante lineal (LDA)

1. Es utilizado para encontrar una combinación lineal de variables que separa dos o más clases.
2. La combinación resultante puede usarse como un clasificador lineal.
3. Se usa comúnmente para la reducción de la dimensionalidad antes de una clasificación posterior.

Análisis discriminante

El análisis discriminante lineal (LDA)

1. Está estrechamente relacionado con el análisis de varianza (ANOVA) y el análisis de regresión
2. ANOVA usa variables independientes categóricas y una variable dependiente continua, mientras que el análisis discriminante tiene variables independientes continuas y una variable dependiente categórica
3. La regresión logística y la regresión probit son más similares a LDA que ANOVA, ya que también explican una variable categórica por los valores de las variables independientes continuas.
4. También está estrechamente relacionado con el análisis de componentes principales (PCA) y el análisis factorial

Referencias

[1] DellÁversana, P. , "Cross-disciplinary Machine Learning - Part 1: Applications to rock classification, well log analysis and integrated geophysics", 182 pags. (2020).

[2] Leijnen, S. and Veen, F. V., *The neural network zoo*. Proceedings, **47(1)** (2020).

[3] Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain", *Psychological Review* **65 (6)** 386-408 (1958).

[4] Samuel, A. L. . "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, **3:3** 210–229 (1959).

Siguiente tema:

Modelos de dependencia

basados en cópulas