ISSN 1870-4069

Amputación de datos de registros geofísicos con mecanismo de pérdida MCAR

César E. Santos-Vázquez, Leonardo Morales-Collado,
Juana Canul-Reich, José Hernández-Torruco

Universidad Juárez Autónoma de Tabasco,
División Académica de Ciencia y Tecnologías de la Información,
México

{csv.santos92, lmc580, jcanulreich}@gmail.com,
jose.hernandezt@ujat.mx

Resumen. El tratamiento de datos faltantes en el área de registros geofísicos es una tarea crucial, por lo cual el desarrollo de una metodología que simule la presencia de los mismos, como lo es la amputación de datos, permite contar con un punto de referencia a la hora de implementar nuevas técnicas enfocadas en la solución de este tipo de situaciones. En esta investigación se propone un flujo de trabajo para el análisis, procesado y amputación de datos de registros geofísicos. Mediante este análisis se obtuvo una muestra de los datos amputados con el mecanismo de pérdida MCAR y se comparó con las distintas distribuciones logísticas para el mismo mecanismo. En esta propuesta se utiliza tanto un lenguaje de código abierto, así como librerías accesibles para todo público, por lo cual replicar esta experimentación no presenta ningún inconveniente.

Palabras clave: Amputación, datos perdidos, datos nulos, registros geofísicos.

Well Logs Data Amputation with MCAR Loss Mechanism

Abstract. The treatment of missing data in the area of geophysical well logs is a crucial task, which is why the development of a methodology that simulates the presence of the same, such as the data amputation, allows to have a point of reference at the time to implement new techniques focused on solving this type of situation. In this research, a workflow is prposed for the analysis, processing and data amputation from geophysical well logs. Through this analysis, a sample of the amputee data with the MCAR loss mechanism was obtained and compared with the different logistic distributions for the same mechanism. In this proposal, an open source language is used, as well as libraries accesible to all audiences, so replicating this experiment does not present any inconvenience.

Keywords: Amputation, missing data, null data, well logs.

1. Introducción

El manejo de valores faltantes en un set de datos es una problemática típica de la etapa de preprocesado. El autor [5] deja en claro las relaciones existentes entre las variables y su variabilidad exponiendo ejemplos sobre como se relaciona una variable con la taza de cambio de la otra. En el área de registros geofísicos también existen procesos que necesitan la preparación de los datos para poder realizar estudios o interpretaciones petrofísicas.

La pérdida de datos en curvas de registros geofísicos es un problema común que se debe corregir, cuando es posible, durante la etapa de control de calidad. Algunas de las causas de estas pérdidas de información se deben a las condiciones físicas que pueden sufrir los equipos de medición de registros, llevándolos a un mal funcionamiento, a las condiciones y características del agujero, a la velocidad excesiva en el momento en que se toma el registro, al manejo de la información previo a ser entregada para realizar los estudios, entre otras, como menciona [3].

La simulación de datos faltantes, permite desarrollar y evaluar métodos que abordan esta problemática. Contar con un procedimiento de amputación de datos que pueda aplicarse sobre datos completos es importante, ya que en primera instancia (antes de la etapa de procesado) estos suelen contar con valores faltantes. Particularmente la información de registros geofísicos es privada para algunas empresas, por lo que conseguir datasets públicos puede ser una tarea difícil.

En los casos de las empresas que sí hacen públicos sus datasets, estos ya están preprocesados y típicamente sus posibles valores faltantes ya han sido resueltos. Este estudio es parte de una investigación más amplia, para la cual no se disponía de datasets con valores faltantes, por lo que se decidió identificar una técnica para generarlos, es decir realizar amputación de datos. Al realizar la revisión de la literatura relacionada se observó que no se cuenta con una metodología establecida para la amputación de datos de registros geofísicos.

Por lo tanto, establecer tal metodología, que sirva de referencia para futuras investigaciones en la industria energética y en el área de investigación científica, es la principal contribución de este manuscrito. En el articulo [9] se menciona que algunas de las prácticas actuales para la amputación de datos no son del todo correctas, ya que la validez de la metodología y la estadística aplicada en dichos métodos de amputación pueden no ser apropiadas para dicho propósito.

Finalmente concluye que la amputación multivariable, suele generar amputaciones mucho más fiables, al mismo tiempo que permite regular la aplicación del método de amputación utilizado. En la literatura actual [7], señala que el proceso de amputación de datos es la base para los experimentos de imputación, pero si no se realiza adecuadamente puede resultar en conclusiones inválidas. Propone utilizar el desafiante mecanismo de pérdida, MCAR, para trabajos futuros.

Por otra parte, en [1], se menciona que los datos faltantes mas allá de ser casos aislados, son prácticamente un regla en la vida real. Se hace mención a estudios publicados durante los años 1998 a 2004, en los cuales se encontró que el 48 % de los estudios presentó datos faltantes, y el 36 % no, quedando un restante indeterminado. De este modo se justifican los esfuerzos realizados con el fin de mejorar los métodos para tratar este tipo de problemáticas.

Tabla 1. Definición de probabilidades de pérdida según el mecanismo de pérdida. Donde y: matriz de datos entera, y_{obs}: observaciones de y, y_{mis}: observaciones perdidas, R: matriz de pérdida y q: vector de parámetros que describen la relación entre la pérdida, R y el conjunto de datos y.

Mecanismo de pérdida	Definición de probabilidad
MNAR	$p(R\|y_{obs}, y_{mis}, q)$
MAR	$p(R\|y_{obs}, q)$
MCAR	$p(R\|q)$

Este artículo está estructurado de la siguiente manera: En la Sección 2 se realiza una definición formal de los mecanismos de pérdida involucrados en el manejo de datos faltantes. La Sección 3 presenta el conjunto de datos utilizado, algunas características de su estructura y formato; muestra el hardware y software utilizado para realizar los experimentos y establece la elección del mecanismo de pérdida para este experimento.

En la Sección 4 se define la implementación del método de amputación de datos aplicado al conjunto de datos utilizado, empezando por el preprocesado de la información para adecuarla a las necesidades del experimento, seguido del desarrollo del método de amputación y finalizando con una verificación gráfica de los datos ya amputados.

La Sección 5 presenta los datos amputados, comparando los resultados del método utilizando distintos tipos de distribución de probabilidad. Finalmente, la Sección 6 concluye el artículo resaltando la importancia del uso del mecanismo de pérdida MCAR para la amputación de datos de registros geofísicos y propone algunas direcciones para futuras investigaciones.

2. Mecanismo de pérdida para amputación de datos

La investigación realizada se centra en la generación y posterior análisis de datos faltantes en datos de registros geofísicos. Regularmente las herramientas presentes en software petrofísico ofrecen una solución no del todo útil como lo es la imputación de datos faltantes utilizando la media de dicho atributo en el conjunto de datos, del mismo modo los enfoques clásicos en otras disciplinas ofrecen desde la eliminación de las instancias con presencia de valores nulos hasta la imputación por medio de una regresión lineal.

Por lo tanto, contar con una metodología para la experimentación en esta disciplina podría ser de gran ayuda para referenciar futuras experimentaciones o propuestas de solución a este tipo de problemática, al tener un punto de comparación.

2.1. Mecanismos de pérdida en datos faltantes

En primera instancia se debe definir el mecanismo de pérdida bajo el cual sería lo más lógico que se presentaran datos faltantes en un conjunto con respectiva naturaleza del mismo. Fox et al [2] definen tres categorías para clasificar los datos faltantes:

1. **MCAR** (Missing completely at random): cuando la probabilidad de que un dato se pierda no está relacionada a ninguna variable en el conjunto de datos.

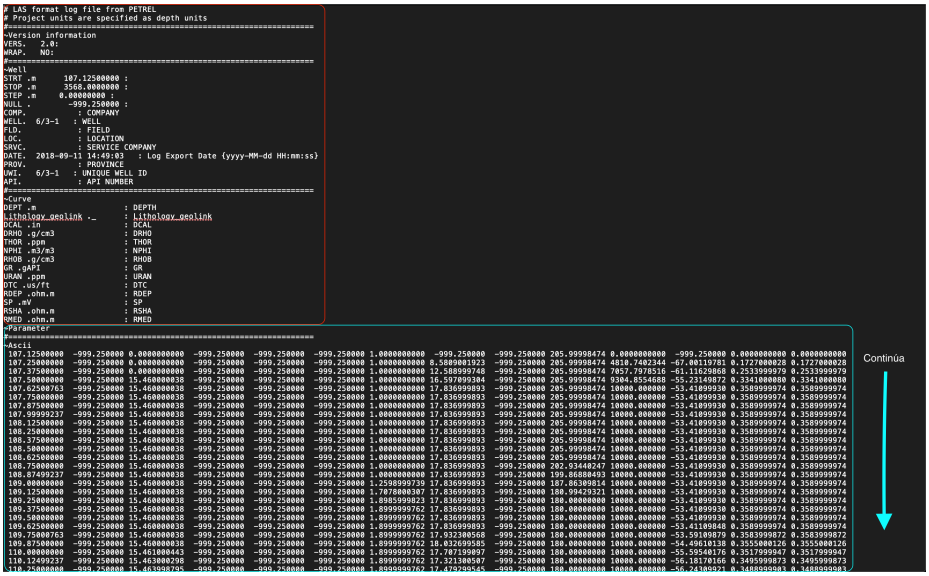


Fig. 1. Vista de un archivo con extensión `.las`.

2. **MAR** (Missing at random): cuando la probabilidad de que hagan falta datos en una variable está relacionada con otra(s) variable(s) del conjunto de datos.
3. **MNAR** (Missing not at random): sucede cuando la probabilidad de la pérdida de datos está relacionada con la variable en sí misma.

A pesar de que estos mecanismos tienen nombres que podrían sugerir una causa probable de datos perdidos, el concepto realmente engloba la relación estadística entre las observaciones (variables) y la probabilidad de pérdida de datos. En la tabla 1 se muestran las definiciones formales (matemáticas) de los mecanismos de pérdida.

Los autores en [4] ejemplifican tres tipos de mecanismos de pérdida. Basado en esa diferenciación se puede clasificar la problemática de este estudio (gaps en registros geofísicos) dentro del mecanismo de pérdida MCAR (missing completely at random), debido a que la probabilidad de que se presenten datos nulos es la misma para todo el conjunto de datos.

3. Métodos y materiales

3.1. Conjunto de datos

El conjunto de datos a utilizar pertenece al proyecto de plataforma continental de Noruega realizado por la empresa Geolink, el cual contiene datos de más de quinientos pozos. Estos datos contienen información geológica y petrofísica sobre hidrocarburos del sector noruego. Los datos se encuentran disponibles a través de GitHub en el repositorio de Lukas Mosser[1].

[1] https://github.com/LukasMosser/geolink_conjuntodedatos

```
$ VERSION   : num 2
$ WELL      :'data.frame':    13 obs. of  4 variables:
 ..$ MNEM       : chr [1:13] "STRT" "STOP" "STEP" "NULL" ...
 ..$ UNIT       : chr [1:13] "m" "m" "m" "" ...
 ..$ VALUE      : chr [1:13] "107.12500000" "3568.0000000" "0.00000000" "-999.250000" ...
 ..$ DESCRIPTION: chr [1:13] "" "" "" "" ...
$ CURVE     :'data.frame':    14 obs. of  4 variables:
 ..$ MNEM       : chr [1:14] "DEPT" "Lithology_geolink" "DCAL" "DRHO" ...
 ..$ UNIT       : chr [1:14] "m" "_" "in" "g/cm3" ...
 ..$ API.CODE   : chr [1:14] "" "" "" "" ...
 ..$ DESCRIPTION: chr [1:14] " DEPTH" " Lithology_geolink" " DCAL" " DRHO" ...
$ PARAM     : NULL
$ OTHER     : chr ""
$ LOG       :'data.frame':    27688 obs. of  14 variables:
 ..$ DEPT            : num [1:27688] 107 107 107 108 108 ...
 ..$ Lithology_geolink: num [1:27688] NA NA NA NA NA NA NA NA NA ...
 ..$ DCAL            : num [1:27688] 0 0 0 15.5 15.5 ...
 ..$ DRHO            : num [1:27688] NA NA NA NA NA NA NA NA NA ...
 ..$ THOR            : num [1:27688] NA NA NA NA NA NA NA NA NA ...
 ..$ NPHI            : num [1:27688] NA NA NA NA NA NA NA NA NA ...
 ..$ RHOB            : num [1:27688] 1 1 1 1 1 1 1 1 1 1 ...
 ..$ GR              : num [1:27688] NA 8.58 12.59 16.6 17.84 ...
 ..$ URAN            : num [1:27688] NA NA NA NA NA NA NA NA NA ...
 ..$ DTC             : num [1:27688] 206 206 206 206 206 ...
 ..$ RDEP            : num [1:27688] 0 4811 7058 9305 10000 ...
 ..$ SP              : num [1:27688] NA -67 -61.1 -55.2 -53.4 ...
 ..$ RSHA            : num [1:27688] 0 0.173 0.253 0.334 0.359 ...
 ..$ RMED            : num [1:27688] 0 0.173 0.253 0.334 0.359 ...
$ PATH      : chr  "**************************************** /6_3-1.las"
$ ATTRIBUTES:'data.frame':    1 obs. of  8 variables:
 ..$ well       : chr "6/3-1"
 ..$ null       : num -999
 ..$ start      : num 107
 ..$ start_units: chr "m"
 ..$ stop       : num 3568
 ..$ step       : num 0
 ..$ step_units : chr "m"
```

Fig. 2. Estructura de un archivo `.las` en R Studio.

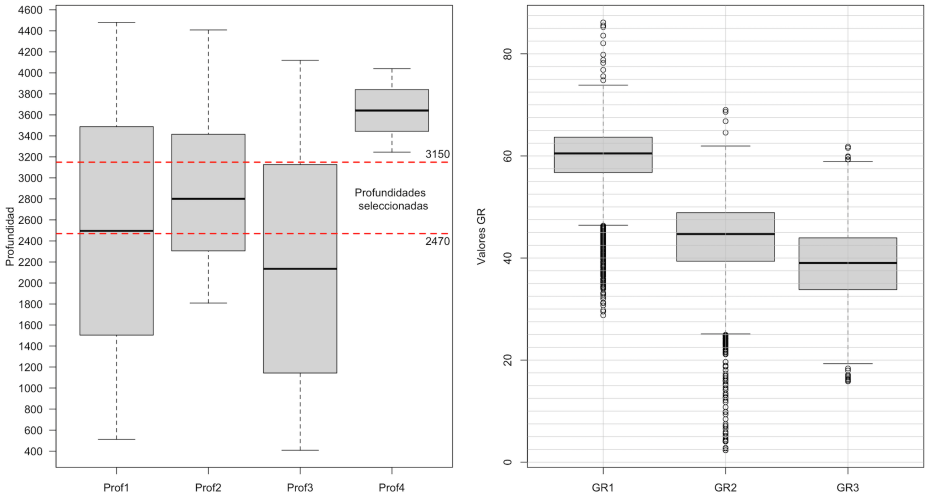
La información de cada pozo se almacena en un archivo con formato **.LAS**, dentro de tal archivo se encuentran los parámetros para diversas respuestas de señal (curvas). La figura 1 muestra la estructura interna de un archivo **.LAS**, el recuadro rojo delimita la zona de encabezado del pozo, dentro de esa zona se puede encontrar información como, por ejemplo, la versión del formato, el inicio y fin del pozo, la localización, fecha, nombre y se hace especial referencia a la nomenclatura utilizada para los datos faltantes o nulos, señalada con el número negativo “-999.250000”.

Del mismo modo se observa que dentro del encabezado de pozo (recuadro rojo) se encuentra información sobre el arreglo de curvas (señales) que incluye este archivo, estas situadas en la parte inferior del encabezado (apartado señalado como curve). Por otra parte en el recuadro azul se muestran los datos apilados por columna, estas columnas llevan un orden descendente (como se encuentra indicado en el apartado curve).

3.2. Hardware

Para la realización de esta investigación se utilizó un equipo de computo Apple MacBook Air 13” 2018 el cual cuenta con las siguientes especificaciones:

– Procesador: 1.6 GHz Intel Core i5 de dos núcleos.
– Memoria: 8GB 2133 MHz LPDDR3.
– Almacenamiento: 256 GB SSD.
– Gráficos: Intel UHD Graphics 617, 1536 MB.
– Sistema Operativo: macOs Catalina 10.15.4.



(a) (b)

Fig. 3. Distribución de datos en el intervalo seleccionado.

Tabla 2. Proporción de datos nulos generados.

Proporción	Nulos
3 %	139
10 %	443
15 %	683
25 %	1084
30 %	1301

3.3. R y RStudio

El lenguaje utilizado fue R en su versión 4.0.1. Por otro lado la utilización del lenguaje se realizó a través del IDE (entorno de desarrollo integrado) RStudio [6], con la versión Orange Blossom 1.2.5033.

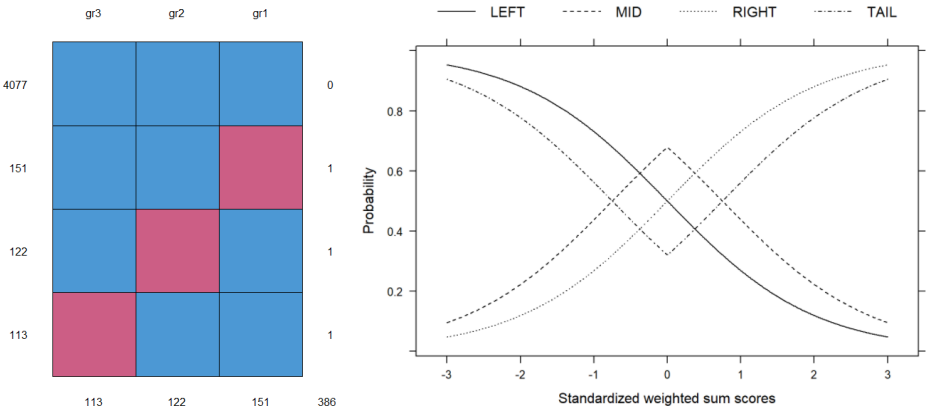
3.4. Missing Completely at Random (MCAR)

Dentro de los mecanismos de pérdida, se denomina como Missing Completely at Random (MCAR) a aquellas situaciones en las que una matriz de pérdida **M** no presenta ninguna relación con la matriz de datos **X**, como se aprecia en [7] y los parámetros que describen la relación entre ambas matrices **q**.

Atendiendo a la naturaleza de los datos, de registros geofísicos, este mecanismo es el que mejor se ajusta a un panorama real, ya que la aleatoriedad de las situaciones presentes durante la recolección de los registros son situaciones muy comunes y se presentan con facilidad:

$$p(\mathbf{M}|\mathbf{X},\mathbf{q}) = p(\mathbf{M}|\mathbf{q}). \tag{1}$$

Del mismo modo la configuración y la tasa de pérdida, son dos factores muy importantes cuando se amputan valores a un conjunto de datos.



(a) Patrón de pérdida del 3 %. (b) Distribución logística.

Fig. 4. Patrón de Pérdida y Tipos de Distribución Logística.

La configuración se refiere a si los datos faltantes se presentan solamente en una variable o se presentan en dos o más variables del conjunto de datos, pudiendo ser univariable o multivariable. Por otra parte la tasa de pérdida representa el porcentaje de datos faltantes, es posible referirse a estos por variable o por el conjunto de datos completo.

Por ejemplo si se tuviera un conjunto de datos constituido por una matriz de 5 variables y 100 observaciones ($X[i_{1..,100},\ j_{1..,5}]$) y cada variable tuviera 20 valores faltantes, la tasa de pérdida sería 20 %, ya que si se cuenta con 100×5 se tendrían 500 valores en el conjunto de datos, de los cuales $20 \times 5 = 100$ son faltantes, esto es $100/500 = 0{,}2$ es decir, el 20 % de los valores presentes en la matriz de datos **X** son valores faltantes.

4. Implementación del método de amputación

4.1. Preprocesado de datos

La librería **lastools** permite utilizar archivos con extensión **.LAS** en R. La figura 2 muestra la estructura interna de un archivo con esta extensión dentro del ambiente. Los datos se extrajeron en un dataframe, que contiene los valores del registro y su correspondiente columna de profundidad.

En la figura 3a se muestra un boxplot en el cual se comparan las profundidades, donde se observa que los pozos 1, 2 y 3 coinciden en el intervalo de los 2470 a los 3150 metros, por lo cual se utilizó esta sección para delimitar un conjunto de datos de trabajo. Por otra parte en la figura 3b se muestra el boxplot de los valores de la curva GR de cada pozo seleccionado, en el intervalo definido anteriormente.

4.2. Amputación de datos

La función **ampute()** de la librería **mice** permite realizar la amputación multivariable, es decir amputar datos en las diferentes variables del dataset, de una sola vez y seleccionar el mecanismo, patrón y proporción de pérdida de datos.

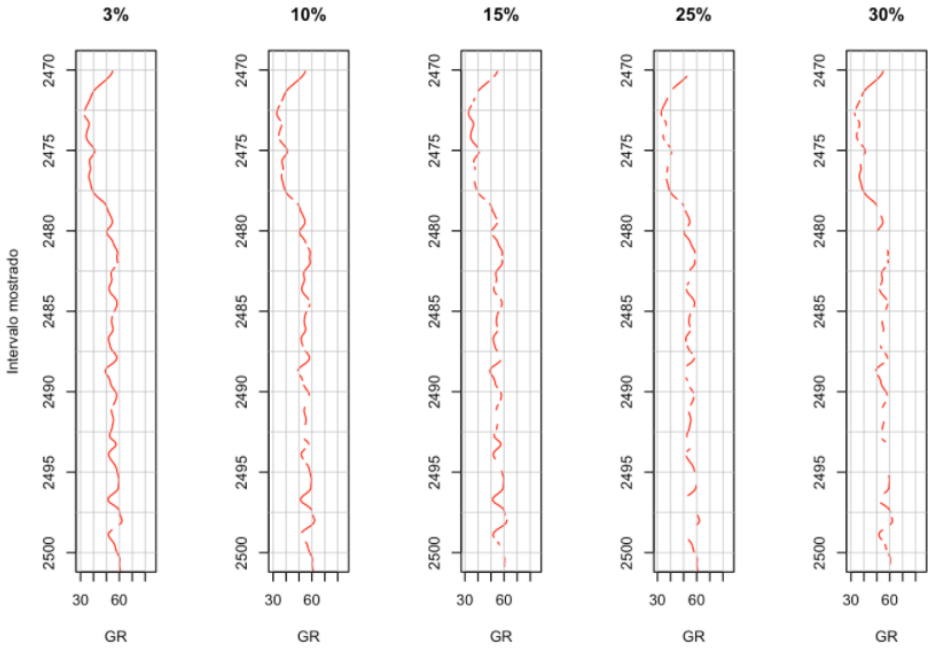


Fig. 5. Curvas con diferentes porcentajes amputados.

Genera la pérdida de datos deseada, de acuerdo a los parámetros indicados. Las proporciones de pérdida para este experimento fueron del 3, 10, 15, 25 y 30 por ciento, con un mismo tipo de distribución (**MID**). Es importante mencionar que la función **ampute()** no admite conjuntos con NA's. El resultado es un objeto de clase **mads** (multivariate amputed data set), el cual contiene los datos amputados, los originales, el patrón de pérdida y el tipo de distribución.

En la figura 4a se observa el patrón de pérdida generado. La primer fila representa todos los casos completos, mientras que la segunda fila representa 151 casos en los que la variable "gr1" contiene datos faltantes, la tercer fila representa 122 casos incompletos para la variable "gr2" y finalmente la cuarta fila representa 113 casos en los que la variable "gr3" cuenta con faltantes. Los datos faltantes se muestran en rojo y los datos observados en azul.

Otro parámetro importante para la función **ampute()** es **type**, el cual como explican los autores en [9], se refiere al tipo de distribución de probabilidad logística, la cual puede ser de 4 tipos: LEFT, MID, RIGHT, TAIL. Según el valor que se escoja para este parámetro se determinará el comportamiento de la función **ampute()**, basado en la distribución logística. La figura 4b nos muestra la relación entre la distribución logística y el tipo de pérdida escogido.

Verificación de datos amputados. Una vez que el conjunto de datos fue amputado, se verifica si realmente se generaron Na´s en su contenido. El número de observaciones totales es de 4463 y se cuenta con 3 variables; la tabla 2 muestra la cantidad de datos amputados por porcentaje de pérdida indicado en la función **ampute()**.

Tabla 3. Localización de Valores Amputados.

Loc. Valores Amputados por porcentaje							Valores Amputados por porcentaje				
Prof.	**GR 3 %**	**GR 10 %**	**GR 15 %**	**GR 25 %**	**GR 30 %**	**GR_Orig**	**gap 3 %**	**gap 10 %**	**gap 15 %**	**gap 25 %**	**gap 30 %**
12850	54.86	54.86	54.86	54.86	54.86	54.86	×	×	×	×	×
12851	53.98	53.98	53.98	×	53.98	53.98	×	×	×	53.98	×
12852	52.48	52.48	52.48	52.48	52.48	52.48	×	×	×	×	×
12853	50.55	50.55	×	50.55	50.55	50.55	×	×	50.55	×	×
12854	48.42	48.42	48.42	48.42	48.42	48.42	×	×	×	×	×
12855	46.20	46.20	46.20	46.20	46.20	46.20	×	×	×	×	×
12856	43.95	43.95	43.95	43.95	43.95	43.95	×	×	×	×	×
12857	41.87	41.87	41.87	41.87	41.87	41.87	×	×	×	×	×
12858	40.24	40.24	40.24	×	40.24	40.24	×	×	×	40.24	×
12859	39.13	39.13	×	39.13	39.13	39.13	×	×	39.13	×	×
12860	38.32	38.32	×	×	38.32	38.32	×	×	38.32	38.32	×
12861	37.56	37.56	37.56	37.56	×	37.56	×	×	×	×	37.56
12862	36.72	36.72	36.72	36.72	36.72	36.72	×	×	×	×	×
12863	35.80	×	×	35.80	35.80	35.80	×	35.80	35.80	×	×
12864	34.83	34.83	34.83	34.83	34.83	34.83	×	×	×	×	×
12865	33.92	33.92	33.92	33.92	33.92	33.92	×	×	×	×	×
12866	33.16	33.16	33.16	33.16	×	33.16	×	×	×	×	33.16
12867	×	32.77	32.77	32.77	32.77	32.77	32.77	×	×	×	×
12868	33.05	33.05	33.05	33.05	33.05	33.05	×	×	×	×	×
12869	34.11	34.11	34.11	×	×	34.11	×	×	×	34.11	34.11

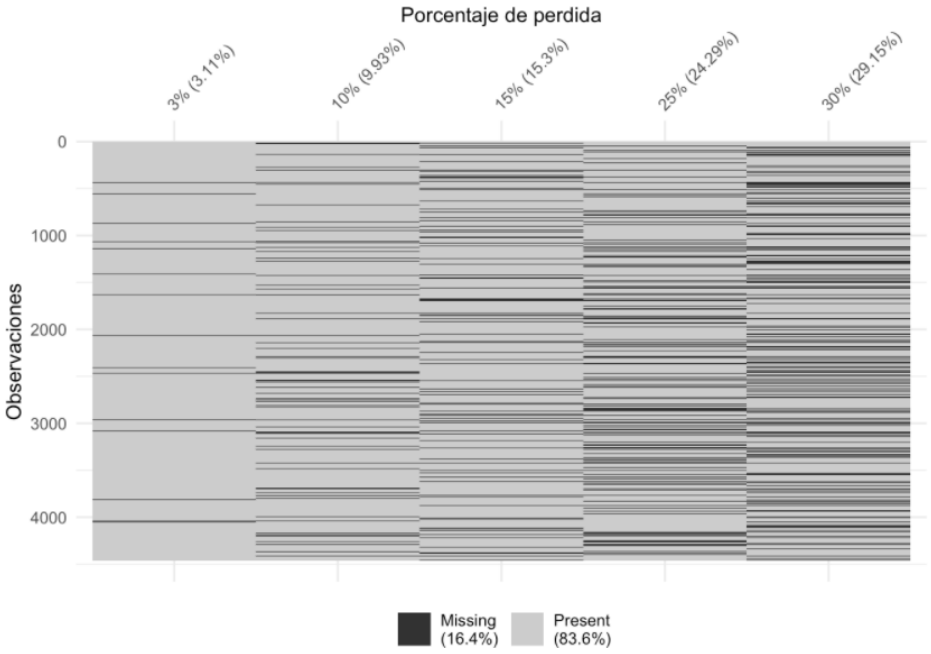


Fig. 6. Patrón de pérdida MCAR.

4.3. Representación gráfica de los datos amputados

La figura 5 muestra una gráfica de todas las curvas amputadas con los diferentes porcentajes de pérdida seleccionados, se observan las partes en las que se generaron datos nulos. Los datos corresponden a la misma curva, del mismo pozo.

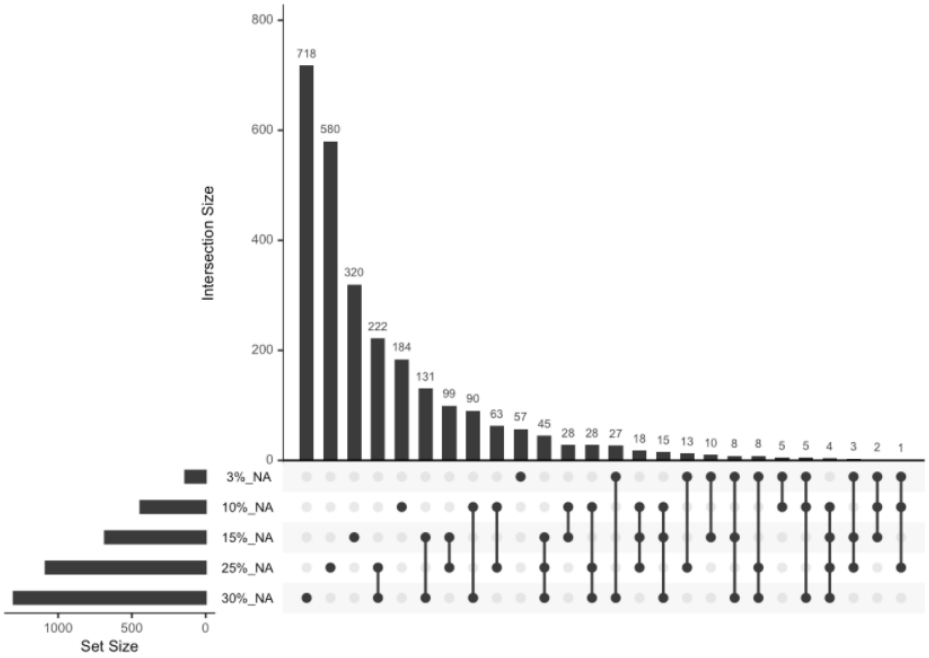


Fig. 7. Acumulación de datos amputados por columna (Porcentaje de pérdida).

La tabla 3 muestra la impresión de los datos con los valores amputados, en comparación con el valor original en cada parte. Nótese que las instancias de GR que contienen NA's coinciden con las instancias de gaps que contienen valores, es decir son inversas entre sí. De este modo se identifican los intervalos con datos faltantes y se señalan dentro del conjunto de datos.

La figura 6 muestra los patrones de pérdida generados por la función **ampute()**. Es posible apreciar los porcentajes de pérdida por columna (variables), esto también permite comprobar que el mecanismo de pérdida no presenta ningún patrón reconocible y que el mecanismo de pérdida aplicado por la función **ampute()** es **MCAR**.

Por otra parte, la figura 7 muestra las veces que el conjunto de datos presenta NA's, se observa que las columnas 15 % y 30 % coinciden 131 veces, mientras que 25 % y 30 % coinciden 222 veces. Conforme el tamaño de intersección crece, la incidencia por columna se va aislando.

5. Resultados

Una vez realizada la amputación, en la tabla 4 se puede observar una muestra de las instancias 30-60; a simple vista se puede percibir la ausencia de un patrón sistemático, lo cual indica que las instancias fueron amputadas (celdas sombreadas) aleatoriamente. En la figura 8 se presentan gráficamente los datos amputados, con la distribución logística tipo Right, mostrados en la tabla 4 (Right). Se aprecia que los gráficos de las curvas coinciden en las zonas con presencia de datos nulos.

Tabla 4. Muestra de los datos amputados con mecanismo MCAR por tipos de distribución de probabilidad logística Right, Left, Mid y Tail.

Right				Left				Mid				Tail			
Obs.	**GR1**	**GR2**	**GR3**	**Obs.**	**GR1**	**GR2**	**GR3**	**Obs.**	**GR1**	**GR2**	**GR3**	**Obs.**	**GR1**	**GR2**	**GR3**
30	36.47	46.57	51.15	30	×	46.57	51.15	30	36.47	46.57	51.15	30	36.47	46.57	51.15
31	38.15	49.38	52.41	31	38.15	49.38	52.41	31	38.15	49.38	52.41	31	×	49.38	52.41
32	39.78	49.40	53.92	32	39.78	49.40	53.92	32	39.78	49.40	53.92	32	39.78	49.40	53.92
33	40.86	×	54.33	33	×	46.88	54.33	33	×	46.88	54.33	33	40.86	46.88	54.33
34	40.95	46.88	53.19	34	40.95	46.88	53.19	34	40.95	46.88	53.19	34	×	46.88	53.19
35	39.92	49.66	51.46	35	39.92	49.66	51.46	35	39.92	49.66	51.46	35	39.92	×	51.46
36	38.29	50.16	50.47	36	38.29	50.16	50.47	36	38.29	50.16	50.47	36	38.29	50.16	50.47
37	37.06	×	50.73	37	37.06	50.25	50.73	37	37.06	50.25	50.73	37	37.06	50.25	50.73
38	36.81	×	51.55	38	36.81	50.72	51.55	38	36.81	50.72	51.55	38	36.81	50.72	51.55
39	37.23	48.81	51.96	39	37.23	48.81	51.96	39	×	48.81	51.96	39	37.23	48.81	51.96
40	37.69	×	51.61	40	37.69	46.88	51.61	40	37.69	46.88	51.61	40	37.69	46.88	51.61
41	37.69	43.01	51.03	41	37.69	43.01	51.03	41	37.69	43.01	51.03	41	37.69	43.01	51.03
42	×	46.28	51.00	42	×	46.28	51.00	42	37.15	46.28	51.00	42	×	46.28	51.00
43	36.56	×	51.61	43	36.56	45.19	×	43	36.56	45.19	51.61	43	36.56	45.19	51.61
44	36.42	×	52.15	44	36.42	46.59	52.15	44	36.42	46.59	52.15	44	36.42	46.59	52.15
45	36.80	47.91	51.65	45	36.80	×	51.65	45	36.80	47.91	51.65	45	36.80	47.91	×
46	37.38	51.19	49.51	46	×	51.19	49.51	46	37.38	51.19	49.51	46	37.38	51.19	49.51
47	37.85	54.43	45.72	47	37.85	54.43	45.72	47	37.85	54.43	×	47	37.85	×	45.72
48	38.23	49.75	40.70	48	38.23	×	40.70	48	38.23	49.75	40.70	48	38.23	49.75	40.70
49	38.81	48.41	35.34	49	38.81	48.41	35.34	49	38.81	48.41	35.34	49	38.81	48.41	35.34
50	39.70	47.54	30.96	50	39.70	47.54	30.96	50	×	47.54	30.96	50	39.70	×	30.96
51	40.90	50.75	28.75	51	40.90	50.75	28.75	51	40.90	50.75	28.75	51	×	50.75	28.75
52	42.61	50.69	×	52	42.61	50.69	×	52	42.61	×	29.23	52	42.61	50.69	×
53	44.85	53.19	32.02	53	44.85	53.19	32.02	53	44.85	×	32.02	53	44.85	53.19	32.02
54	×	51.00	36.15	54	47.19	51.00	×	54	×	51.00	36.15	54	47.19	51.00	36.15
55	49.04	53.59	40.61	55	49.04	53.59	40.61	55	49.04	×	40.61	55	49.04	53.59	40.61
56	50.09	52.09	44.94	56	50.09	52.09	44.94	56	50.09	52.09	44.94	56	×	52.09	44.94
57	50.69	52.66	48.96	57	50.69	52.66	×	57	50.69	52.66	48.96	57	×	52.66	48.96
58	51.39	52.09	52.27	58	51.39	52.09	52.27	58	51.39	52.09	52.27	58	51.39	52.09	×
59	52.32	51.07	54.19	59	52.32	51.07	54.19	59	52.32	×	54.19	59	52.32	51.07	54.19
60	53.27	52.35	54.59	60	53.27	52.35	54.59	60	53.27	52.35	54.59	60	×	52.35	54.59

Por ejemplo, la curva **gr1** (negro) presenta 2 gaps, que corresponden a las observaciones 42 y 54. La curva **gr2** (rojo) presenta los gaps correspondientes a las observaciones 33, 37, 38, 40, 43 y 44, en el caso de la observación 39, esta no es graficada puesto que no tiene datos antes y después.

Finalmente, la curva **gr3** (azul) presenta 1 gap correspondiente a la observación 52. La experimentación realizada para llegar a estos resultados es replicable en su totalidad, sin embargo es importante mencionar que los resultados pueden variar un poco, dado que el mecanismo de pérdida es completamente aleatorio. No obstante, la metodología utilizada funciona para cualquier conjunto de datos de registros de pozos.

6. Conclusiones

Se logró establecer un procedimiento completamente replicable, adecuado y de fácil acceso para la comunidad científica y petrofísica, con el cual es posible simular la presencia de datos faltantes. Sin importar cuál sea la finalidad que tenga cada usuario, puede aplicar este procedimiento.

Right			
Obs.	**GR1**	**GR2**	**GR3**
30	36.47	46.57	51.15
31	38.15	49.38	52.41
32	39.78	49.40	53.92
33	40.86	×	54.33
34	40.95	46.88	53.19
35	39.92	49.66	51.46
36	38.29	50.16	50.47
37	37.06	×	50.73
38	36.81	×	51.55
39	37.23	48.81	51.96
40	37.69	×	51.61
41	37.69	43.01	51.03
42	×	46.28	51.00
43	36.56	×	51.61
44	36.42	×	52.15
45	36.80	47.91	51.65
46	37.38	51.19	49.51
47	37.85	54.43	45.72
48	38.23	49.75	40.70
49	38.81	48.41	35.34
50	39.70	47.54	30.96
51	40.90	50.75	28.75
52	42.61	50.69	×
53	44.85	53.19	32.02
54	×	51.00	36.15
55	49.04	53.59	40.61
56	50.09	52.09	44.94
57	50.69	52.66	48.96
58	51.39	52.09	52.27
59	52.32	51.07	54.19
60	53.27	52.35	54.59

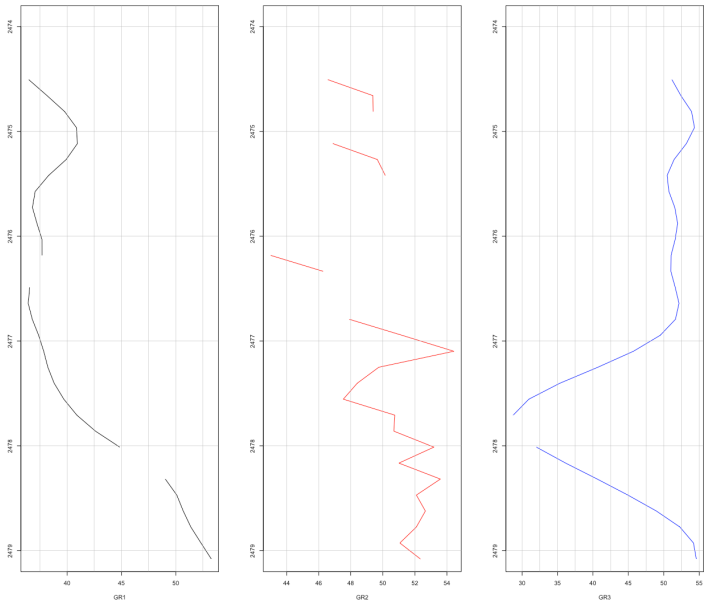


Fig. 8. Comparación de las curvas amputadas con probabilidad logística *Right* y la ocurrencia de datos nulos en el conjunto de datos.

Se pretende extender esta investigación hasta la aplicación de métodos de imputación, tomando los resultados obtenidos como base fundamental para el establecimiento de una metodología que sirva para replicar un procedimiento adecuado de amputación (simulación de datos faltantes) con fines de desarrollo de nuevas técnicas de imputación y el mejoramiento de las actuales.

Los datos utilizados aseguran la reproducibilidad de esta metodología, al ser de libre acceso. La función de amputación (**ampute()**) presenta un límite para el porcentaje de pérdida deseado (no más del 30 %), puesto que una presencia muy alta de faltantes en un dataset, lo convierte en un conjunto de datos de mala calidad.

Por otro lado los tipos de probabilidad logística juegan un papel sumamente importante a la hora de la amputación, ya que definen los rangos dentro de los cuales se encuentran los datos con mayor probabilidad de ser amputados según el tipo de esta distribución de probabilidad.

El estudio y la aplicación de nuevos paradigmas de la computación en conjunto con técnicas clásicas implica el desarrollo de experimentación controlada y referenciada, por lo que contar con un proceso de muestreo y análisis de los datos disponibles es una ventaja y un avance en la prueba de nuevas alternativas para el tratamiento de datos faltantes en registros geofísicos. De este modo, podrá ser más certera la comparación de resultados.

El mecanismo de pérdida MCAR es, en la opinión de los autores, el que más se ajusta a la naturaleza de los datos de registros geofísicos, debido a que los procesos naturales implican la incidencia de muchas variables, las cuales son difíciles (por no decir imposibles) de modelar de manera exacta.

Aunque este trabajo se enfocó al mecanismo de pérdida MCAR, es posible para otros conjuntos de datos, simular los mecanismos, patrones (MNAR, MAR, según la naturaleza de los datos), frecuencia, proporción y distribución logística de pérdida, relacionados a la propia naturaleza de cada conjunto de datos.

Referencias

1. Dong, Y., Joanne-Peng, C. Y.: Principled missing data methods for researchers. SpringerPlus, Springer Science and Business Media LLC, vol. 2, no. 1 (2013) doi: 10.1186/2193-1801-2-222
2. Fox, G.A., Negrete-Yankelevich, S., Sosa, V. J.: Ecological statistics: Contemporary theory and application. Oxford University Press (2015) doi: 10.1093/acprof:oso/9780199672547.001.0001
3. Lopes Rui L., Alípio, J.: Mind the Gap: A well log data analysis (2017)
4. Mallol Roselló, P. J.: Importancia del tratamiento de datos perdidos. Aplicación en estudios longitudinales pequeños, Universitat Oberta de Catalunya (2017)
5. Pyle, D.: Data preparation for data mining (1999)
6. RStudio Team. RStudio: Integrated Development Environment for R. RStudio (2020)
7. Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J .P., Santos, J., Abreu, P. H.: Generating synthetic missing data: A review by missing mechanism. IEEE Access, vol. 7, pp. 11651–11667 (2019) doi: 10.1109/access.2019.2891360
8. Schouten, R. M., Lugtig, P., Vink, G.: Generating missing values for simulation purposes: A multivariate amputation procedure. Journal of Statistical Computation and Simulation, vol. 88, no. 15, pp. 2909–2930 (2018) doi: 10.1080/00949655.2018.1491577
9. Schouten, R. M., Vink, G.: The dance of the mechanisms: How observed information influences the validity of missingness assumptions. Sociological Methods and Research, SAGE Publications, vol. 50, no. 3, pp. 1243–1258 (2018) doi: 10.1177/0049124118799376