Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Aplicación de técnicas de data mining para la evaluación de pérdidas no localizadas en pozos petroleros surgentes

Plan de tesis presentado para obtener el título de *Magíster en Explotación de Datos y Descubrimiento de Conocimiento*

Adriana Lucia Romero Quishpe

Director de Tesis: Dr. Julio Rodríguez Martino
Co - director de Tesis: Dr. Gabriel Horowitz

Buenos Aires, 2019

Aplicación de técnicas de data mining para la evaluación de pérdidas no localizadas en pozos petroleros surgentes

Cada sector de la industria petrolera ve justificado su esfuerzo e inversión cuando el petróleo esta en superficie, en condiciones de ser comercializado. De ahí que es imperativo el no perder producción, trabajando en la prevención de las pérdidas. El poder explicar e identificar el porqué y dónde ocurren las pérdidas es de suma importancia para los operadores de campo para poder actuar de manera temprana dando así soluciones técnicas/económicas. El propósito general de este proyecto consistió en acelerar el proceso de localización de pérdidas en la producción identificando cuando existe una desviación significativa cualquiera de las mediciones monitoreadas en los pozos surgentes. Para esto se utilizó la información disponible de la vida de un pozo: controles de pozos, datos estadísticos, las mediciones físicas, etc. Los resultados muestran que usando la metodología propuesta se podrían detectar los origenes de las pérdidas en una prueba de control de pozo de 24 horas.

Palabras clave: Pozo surgente, pérdidas, petróleo, Aprendizaje Automático, Reconciliación de datos.

Índice general

Índice de figuras

Índice de tablas

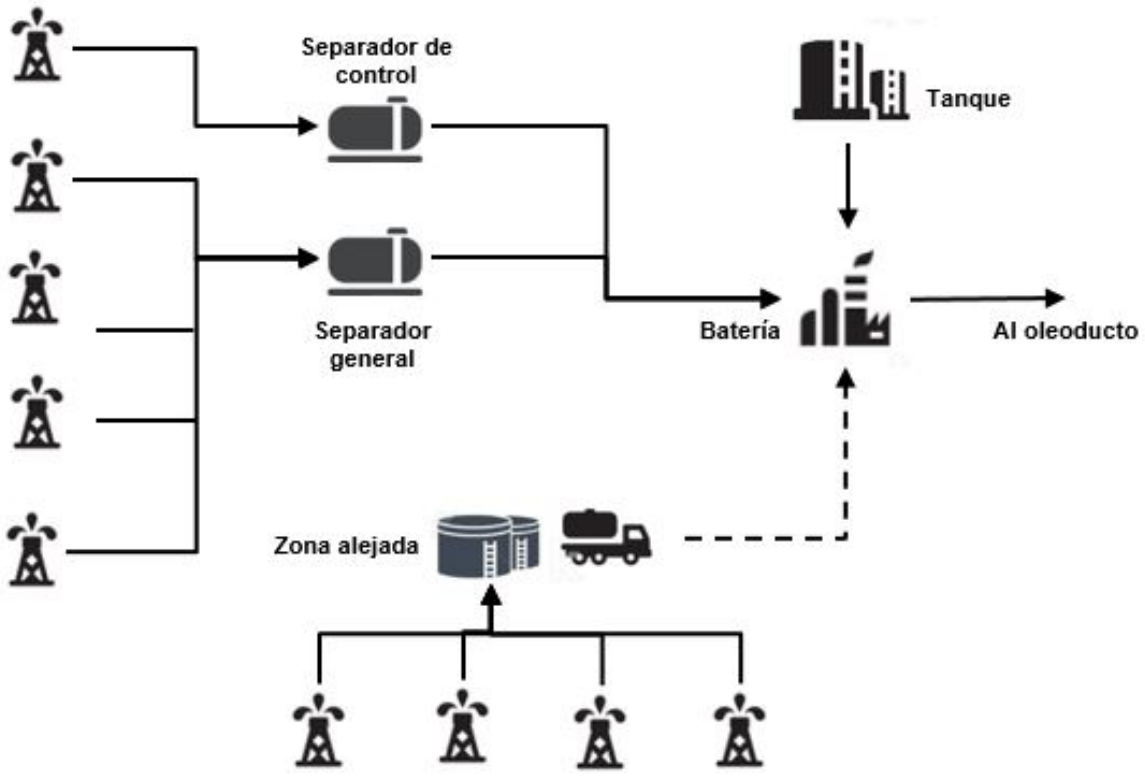
1. DESCRIPCIÓN DEL AMBIENTE DE ANÁLISIS

1.1 Acerca de las instalaciones de la batería

A la batería en análisis aportan 15 pozos surgentes, la producción llega a la misma de dos maneras:por un lado, la producción de 11 pozos es transportada mediante líneas o tuberías. La producción de los 4 pozos restantes es transportada mediante camiones, esto debido a la localización distante de los mismos a la batería.

La figura 1.1.1 muestra el diagrama representativo de la batería.

Figura 1.1.1: Diagrama representativo de la batería en estudio.



Fuente: elaboración propia en base a datos descriptivos de la batería en análisis.

Se puede observar los pozos que transportan la producción por línea y aquellos que la transportan por camiones. Dentro de las instalaciones de la batería se tienen dos separadores donde se realizan los controles a cada pozo. También se tiene un tanque donde se almacena la producción. Para los pozos que transportan crudo mediante camiones, los controles se realizan en el ubicación de los pozos, usando separadores móviles.

1.2 Descripción de las variables a utilizar

Debido a la confidencialidad de los datos, no se especifican los nombres reales de los pozos ni de las variables, salvo la variable de interés inherente que es la producción de petróleo (m^3).

Los datos a analizar provienen de dos fuentes principalmente: por un lado están aquellas medidas físicas tomadas mediante telemetría en la boca de los pozos surgentes, estos datos se almacenan con una granularidad de 1 a 10 minutos (esta configuración depende del ingeniero de producción, no existe algo preestablecido). Por otra parte se tienen datos de control de pozo, estos datos son guardados en planillas de Excel por los recorredores de campo. Las variables tomadas en estas planillas no son las mismas que las tomadas mediante telemetría. Estas mediciones son tomadas cada hora y la frecuencia con la que se realiza en el mes responde a la clasificación de los pozos.

A efectos de garantizar la consistencia del conjunto de datos en cuanto a las granularidad en el tiempo, se llevaron a cabo las siguientes tareas sobre las variables de análisis:

- Se realiza la consulta de los datos tomados con telemetría. Esta consulta se realiza con una granularidad de una hora, para que se corresponda con los datos tomados en los controles de pozo. Para esto se realiza una interpolación de los datos.
- Se realiza la unión de estos datos con los correspondientes al control de pozo, la correspondencia de las mediciones se realizan mediante la fecha de realización. De tal manera que la granularidad en tiempo es de una hora entre cada medición para el estudio.

Se establece una nomenclatura para diferenciar los pozos según como transportan el petróleo a la batería. Para el transporte por línea, se usa *tl* al final de la palabra *pozo*, seguida por el número de pozo. Para los pozos que transportan la producción se usa *tc* y se sigue la misma lógica que con los pozos anteriores.

Se realiza esta diferenciación debido a que los pozos que transportan la producción por camión cuentan únicamente con un sensor de telemetría y, dentro de los pozos de transporte por línea, existe uno que tiene dos sensores. No se descartan las variables

para que tengan el mismo número todos los pozos debido a que las mismas pueden ser importantes dentro del análisis.

Se establece la siguiente nomenclatura para las variables: para aquellas que se obtienen a través de telemetría se usa *telem* seguida por el número de variable y para las variables tomadas de las planillas de control se usa *control* seguida por el número de variable. La variable producción de petróleo se nombra *prod_ oil*.

La tabla 1.2.1 muestra los nombres de los pozos y de las variables a utilizar.

Tabla 1.2.1: Nomenclatura de las variables

		Origen de la variable	
Pozos	**Total de variables**	**Telemetría**	**Control de pozo**
pozo_tl1 al pozo_tl10	21	var_telem1 var_telem2 var_telem3	var_control1 var_control2 ...
pozo_tl11	20	var_telem1 var_telem2	var_control16 var_control17
pozo_tc1 al pozo_tc4	19	var_telem1	prod_oil

2. PROCESAMIENTO DE DATOS

2.1 Tratamiento de datos faltantes

Teniendo en cuenta que la mayor parte de las variables presentan datos faltantes previo al análisis se llevó a cabo imputación de datos. Esto con la finalidad de no perder variables que podrían ser representativas en el análisis. Se utilizó para esto algoritmos automáticos basados en los datos completos disponibles.

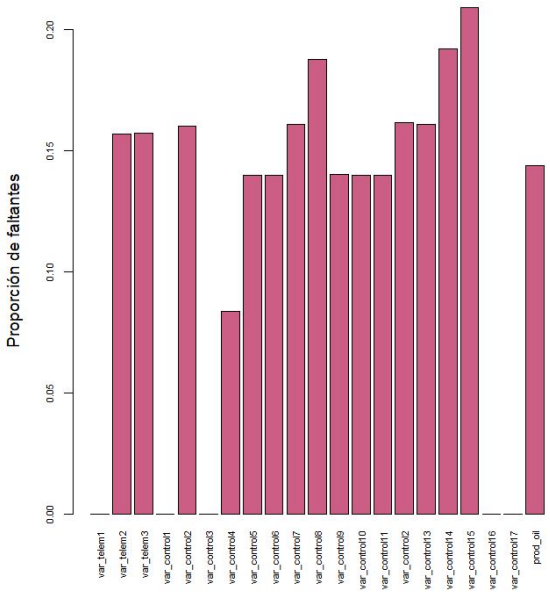
Las figura 2.1.1 presenta la información de los datos faltantes dentro del conjunto de datos correspondiente al *pozo_ tl1* y al pozo *pozo_ tc1*.

En primer lugar se muestra la proporción de datos faltantes de cada una de las variables (figura 2.1.1a y figura 2.1.1c) y un mapa con las combinaciones observadas (figura 2.1.1b y figura 2.1.1d) de valores faltantes en orden de frecuencia de aparición (distribución de faltantes). Los valores conocidos están representados con color azul y los datos faltantes con color violeta.

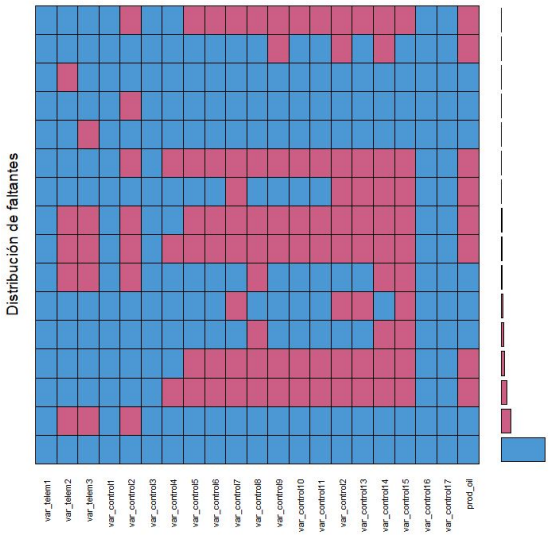
Las tres primeras variables del *pozo_ tl1*, corresponden a las tomadas por telemetría, como se puede observar, la primera no tiene valores faltantes, mientras que las dos restantes tienen igual número de valores faltantes. En porcentaje representa el 16 por ciento (301 casos faltantes de 1876 en total). Para el *pozo_ tc1*, la variable tomada por telemetría no presenta valores faltantes. Estos datos faltantes corresponden a fallas del equipo de telemetría por ejemplo: re-inicio de los equipos, cortes de electricidad, ruptura de antenas de transmisión, etc. Es decir, son eventos aleatorios debido a los cuales no se realiza el censo de la medición. Las variables siguientes (*var_ telem1* - *var_ telem17*) corresponden a los datos de control de pozo. La proporción de datos faltantes varía desde el 8 por ciento (*var_ control4*) hasta el 21 por ciento (*var_ control15*), para el caso del *pozo_ tl1*. Los rangos de valores faltantes para el pozo *pozo_ tc1* varían desde el 4 por ciento (*var_ control4*) hasta el 12 por ciento (*var_ control2*).

Como se puede observar en las gráficas correspondientes a las distribuciones, los datos faltantes se muestran en bloques por filas, esto es debido a que, en determinada fecha, no se completaron las 24 horas correspondientes al control del pozo. Este tipo de eventos tampoco sigue un patrón establecido, ya que depende del criterio del ingeniero de producción el número de horas en el cual se realiza el control.

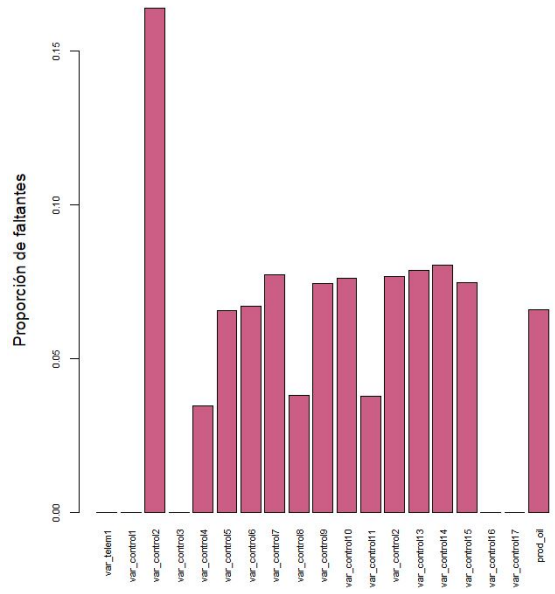
Figura 2.1.1: Información de datos faltantes de los pozos *pozo_ tl1* y *pozo_ tc1*



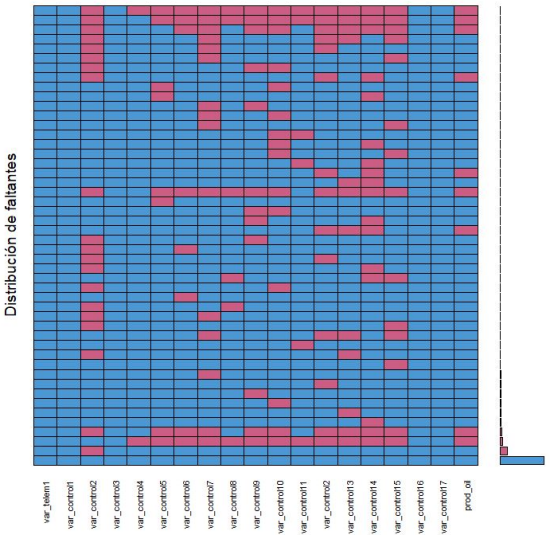
(a) Proporción de datos faltantes *pozo_ tl1*



(b) Distribución de datos faltantes *pozo_ tl1*



(c) Proporción de datos faltantes *pozo_ tc1*



(d) Distribución de datos faltantes *pozo_ tc1*

Fuente: elaboración propia en base a los datos de la batería en análisis.

En base a lo antedicho, los datos están perdidos al azar (MAR por sus siglas en inglés). Este comportamiento se repite para todos los pozos analizados.

A continuación, a efectos de completar los valores faltantes para un correcto análisis de los datos, se utilizan dos métodos popularmente utilizados en problemas de datos incompletos: imputación múltiple basada en ecuaciones encadenadas (MICE por sus siglas en inglés) y Miss Forest , basado en el algoritmo de clasificación Random Forest, compararlos y determinar la opción más conveniente.

2.2 Imputación de datos

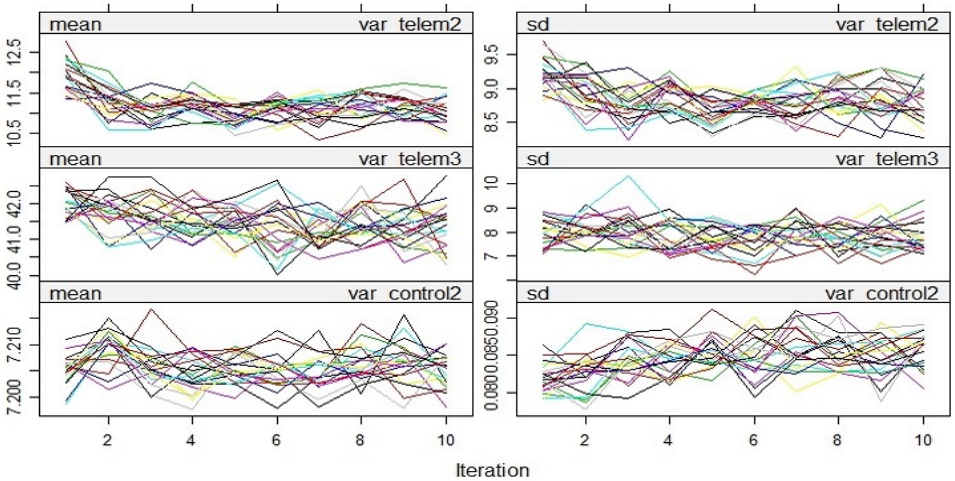
2.2.1. Algoritmo MICE

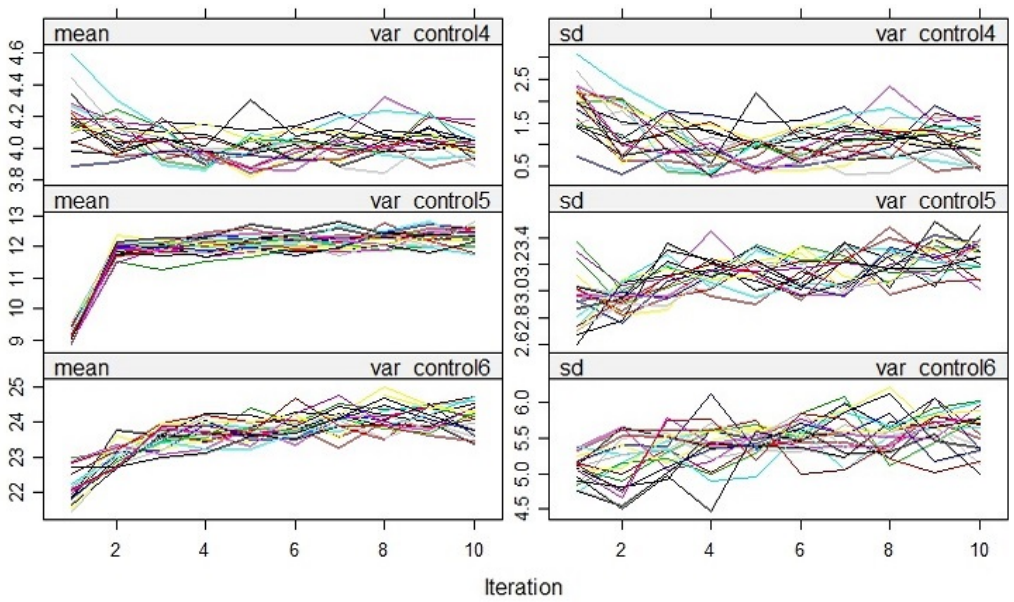
Se aplica el algoritmo MICE a los datos correspondientes al *pozo_ tl1*, para esto el modelo genera una regresión lineal para cada variable con datos faltantes, haciendo que las variables con datos completos actúen como regresoras en el modelo. El modelo generado utiliza todas las variables. Para este algoritmo se considera un número de itraciones igual a 10.

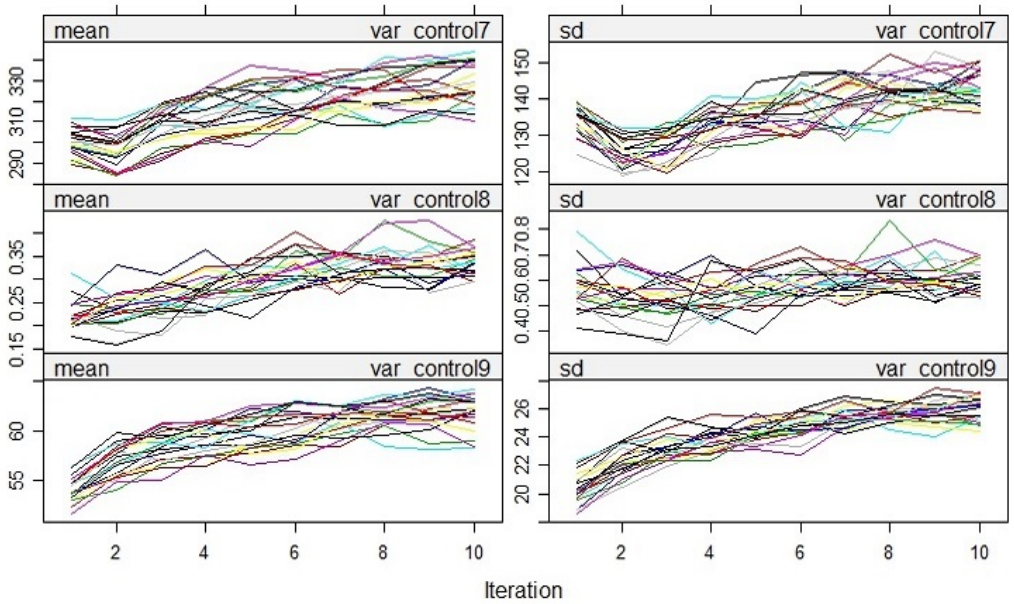
Como se vió anteriormente, la distribución de datos faltantes es similar en todos los pozos, por lo cual, se aplica el mismo algortimo de imputación al conjunto de datos de estudio.

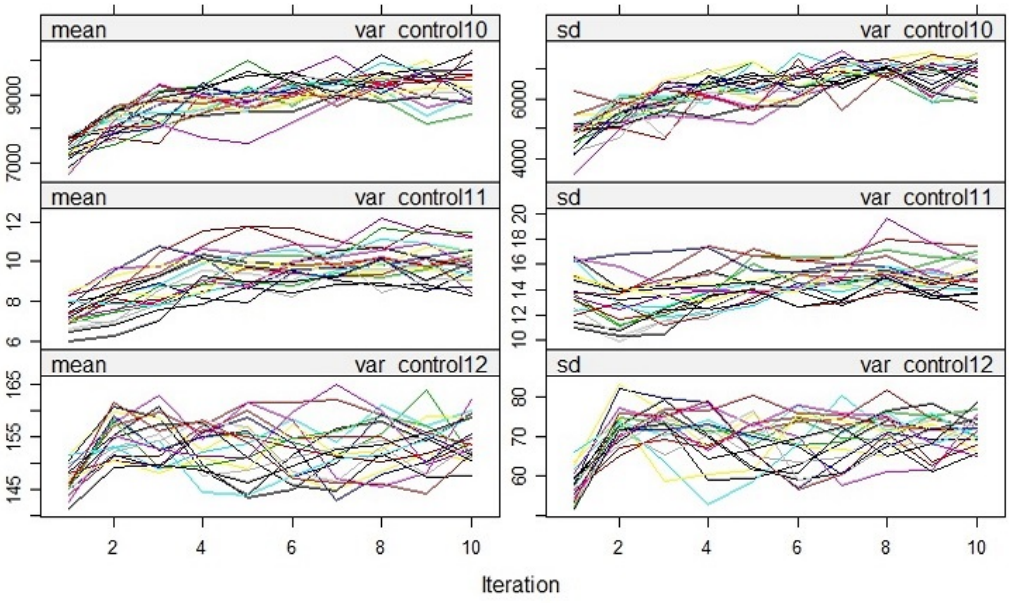
La figura 2.2.1 muestra la media y la desviación estándar de cada una de las variables conforme se incrementa el número de iteraciones. Cada una de las lineas representa una de las versiones de los datos imputados y el eje x representa las iteraciones realizadas por el algoritmo en busca de la convergencia.

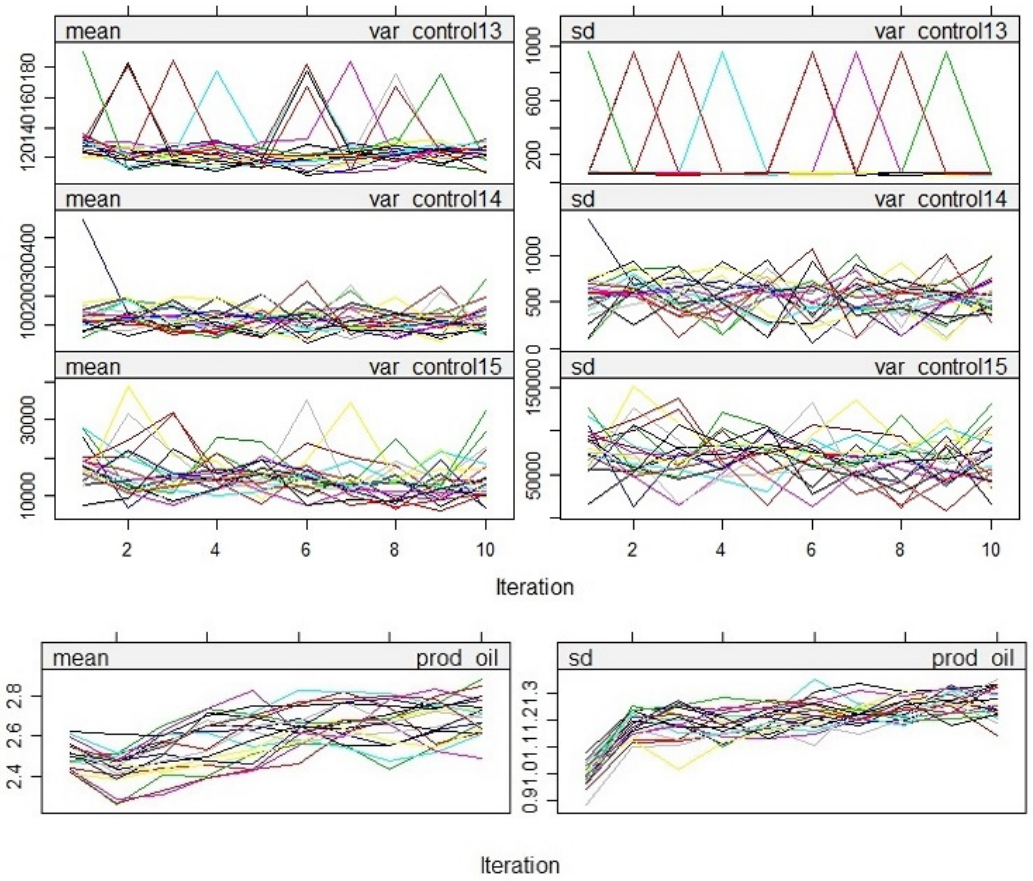
Figura 2.2.1: Media (mean) y desviación estándar (sd) de cada conjunto de variables generada por el algoritmo MICE durante 10 iteraciones (iterations)




mean
var_control4
sd
var_control4
mean
var_control5
sd
var_control5
mean
var_control6
sd
var_control6
Iteration


mean
var_control7
sd
var_control7
mean
var_control8
sd
var_control8
mean
var_control9
sd
var_control9
Iteration


mean
var_control10
sd
var_control10
mean
var_control11
sd
var_control11
mean
var_control12
sd
var_control12
Iteration



Fuente: elaboración propia en base a los datos de la batería en análisis.

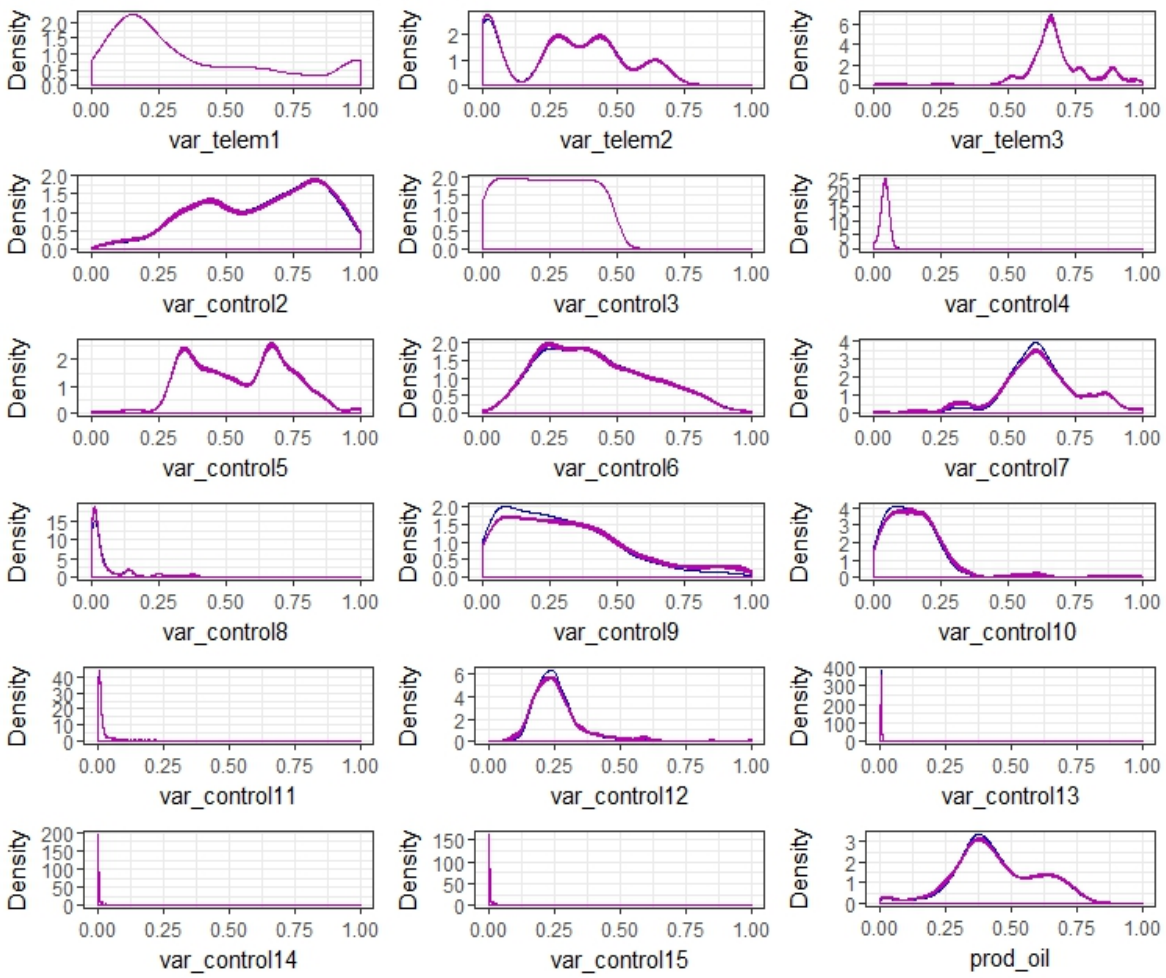
Como se puede observar la líneas se entremezclan entre sí (indicando convergencia). En el caso de la variable *prod_ oil* se observa una tendencia inicial hacia valores más altos que los iniciales, en busca de la convergencia.Esto se da debido a que el algortimo elige valores muy bajos para iniciar el proceso, este comportamiento se hace notorio en la variable *var_ control5* en donde el salto de valores entre la primera y segunda iteración es grande.Una gráfica que llama la atención es la correspondiente a la variable *var_ control13* donde los valores tomados en cada iteración tienen saltos muy grandes, pero a medida que se van generando nuevos conjuntos de datos esto mejora.

La figura 2.2.3 presenta las curvas de densidad de cada variable a lo largo de las iteraciones realizadas por el algoritmo, el conjunto de datos original se muestra en color azul, mientras que los datos imputados se muestran en color violeta.

Como se observa los conjuntos de datos generados por el algortimo siguen en general la naturaleza de los datos originales. No se observan diferencias significativas en las gráficas.

La variable *var_ control9* y la variable *var_ control12* muestran valores más bajos que la distribución original. La variable de *prod_ oil* sigue la distribución de la variable original.

Figura 2.2.3: Densidad (density) de las variables para los conjuntos de datos imputados con MICE.



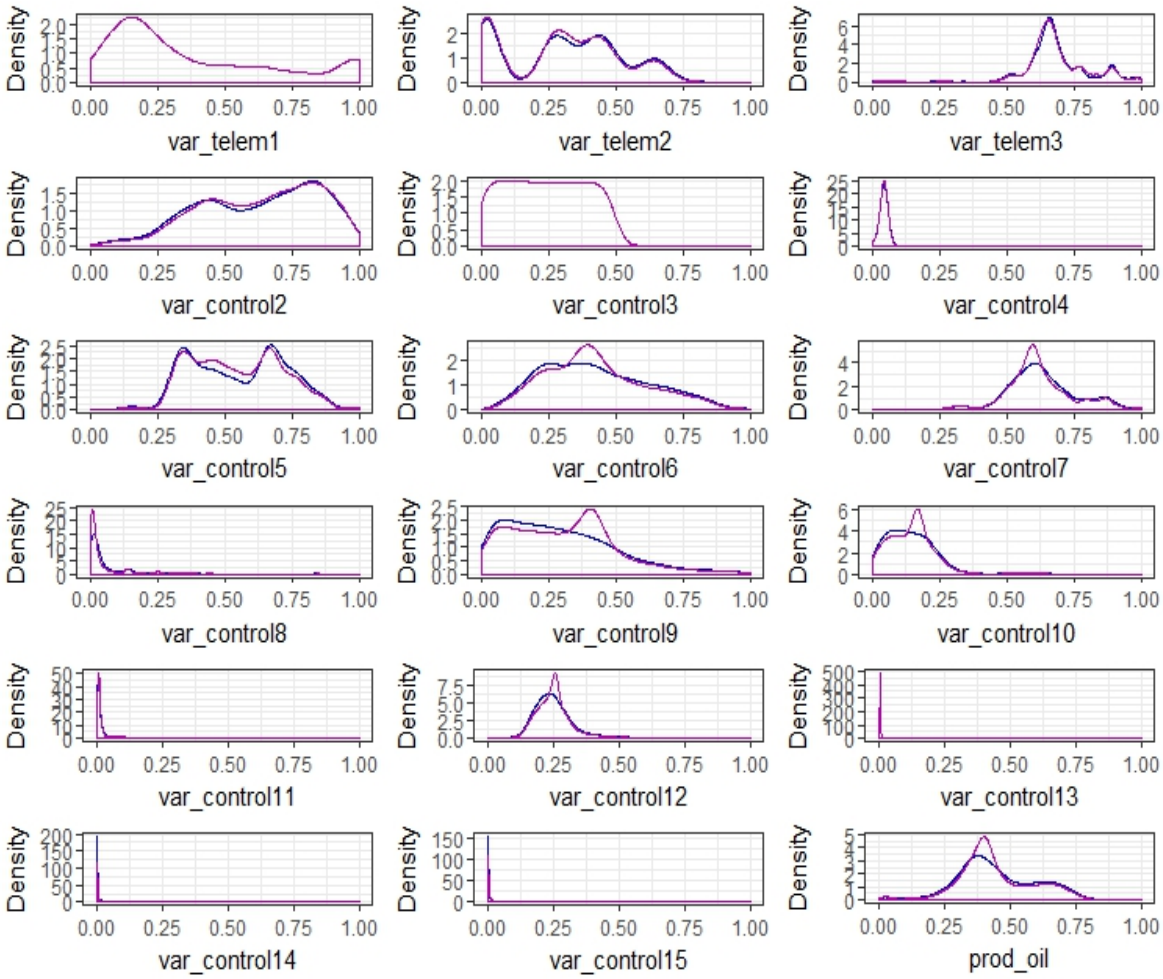
Fuente: elaboración propia en base a los datos de la batería en análisis.

2.2.2. Algoritmo missForest

A continuación se evalúa el algoritmo missForest, al igual que con el algortimo anterior, se establece un número de iteraciones igual a 10. El conjunto de datos imputado a partir del modelo, presenta una estimación del error para cada conjunto de datos imputado en las iteraciones, este error esta basado en el método out-of- bag (OOB por sus siglas, fuera de la bolsa). En este caso, inicia con un valor de 0.56 (56 por ciento) en la primera iteración, y fue disminuyendo, hasta que su valor final es de 0.27 (27 por ciento).

La figura 2.2.4 muestra la curva de densidad del conjunto de datos generado para cada variable. La distinción de colores es igual a la anterior.

Figura 2.2.4: Densidad (density) de las variables para los conjuntos de datos imputados con Miss Forest.



Fuente: elaboración propia en base a los datos de la batería en análisis.

Como se puede observar, si bien la curva de densidad del conjunto de datos generado por el algortimo se ajusta bien a la distribución original, parecería que los datos imputados por el algoritmo *MICE* tiene un comportamiento más cercano.

Esto se puede apreciaren las variables *var_ control6*, *var_ control9* y *var_ control12*, donde los datos imputados cambian de manera significativa. Es de especial consideración notar que la *prod_ oil* no se ajusta a la original. Esto se debe considerar especialmente ya que es la variable de interés.

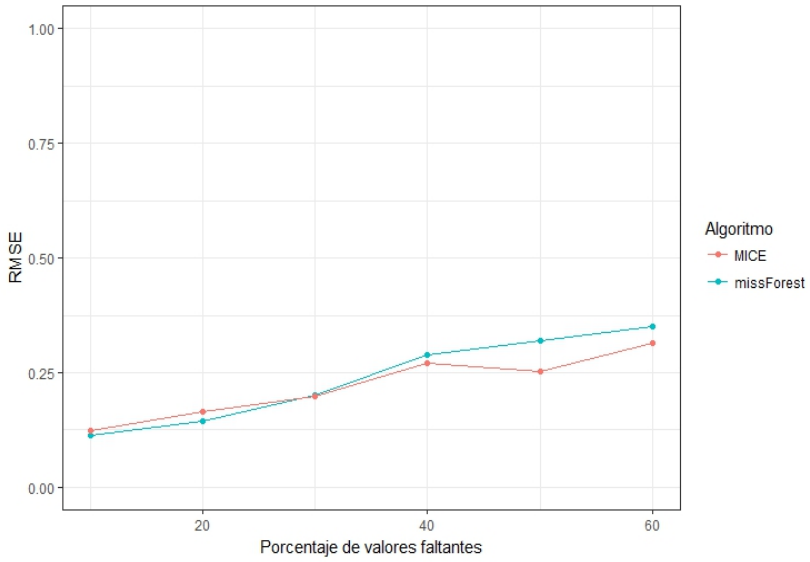
2.2.3. Evaluación y selección del algoritmo de imputación

Con la finalidad de seleccionar el algoritmo a utilizar para el conjunto de datos, se realiza el siguiente procedimiento:

- Se toman los casos completos del conjunto de datos, de tal manera que el número de casos sea igual a la cantidad de faltantes.
- A este nuevo conjunto de datos se le generan datos faltantes en porcentajes del 10 al 60 por ciento, intentando que sigan la misma distribución de faltantes para que éste sea lo más representativo posible.
- Se realiza el proceso de imputación utilizando los dos algoritmos descritos.
- Se calcula la raíz cuadrada del error cuadrático medio ($RMSE$) para cada conjunto de datos imputados y se realiza la comparación según esta métrica. Será mejor aquel que posea un $RMSE$ menor.

La figura 2.2.5 muestra los resultados de la evaluación propuesta. Para porcentajes de valores faltantes menores del 30 por ciento, missForest presenta un valor menor de $RMSE$, sin embargo la diferencia con MICE no es significativa, para un porcentaje de faltantes del 30 por ciento son práticamente iguales (missForest con 20.01 por ciento y MICE con 19.70 por ciento). Cuando el porcentaje de valores faltantes aumenta, el algoritmo MICE tienen un valor de error cuadrático medio menor que missForest.

Figura 2.2.5: $RMSE$ generado por los dos algortimos de imputación de datos para distintos porcentajes de valores faltantes



Fuente: elaboración propia en base a los resultados obtenidos.

En base a la prueba realizada y con las características observadas en los gráficos de densidad, es el algoritmo *MICE* el que se utilizar para efectuar la imputación de datos en el conjunto completo de datos.

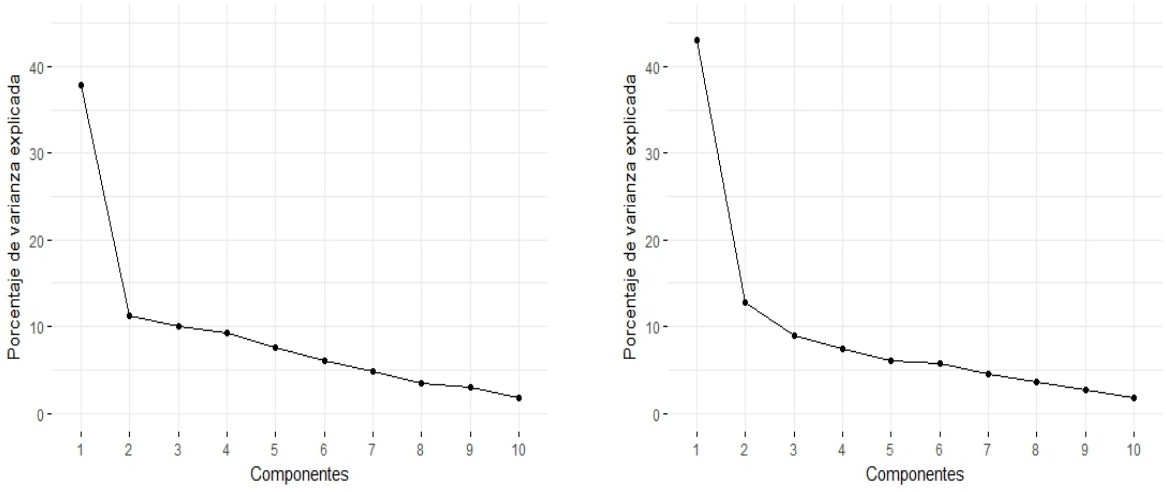
3. ANÁLISIS MULTIVARIADO

3.1 Análisis de componentes principales

A continuación se realiza un análisis de componentes principales con la finalidad de comprender el comportamiento de las variables, y no para realizar una reducción dimensional. Esta técnica permite además la visualización de las correlaciones que existen entre variables. Para esta parte del análisis se diferenciarán los pozos que transportan la producción por línea y por camión.

La figura 3.1.1 muestra la varianza asociada a cada componente principal, esto ayuda a determinar cuántas componentes principales pueden utilizarse según la variabilidad que describen.

Figura 3.1.1: Gráfica de sedimentación



(a) Transporte de producción por línea *(b)* Transporte de producción por camión

Fuente: elaboracion propia basado en los datos de la batería en análisis.

Se puede observar una ruptura en la pendiente de ambas curvas entre la segunda y cuarta componente, después se observa el descenso gradual de las restantes (los sedimentos) a medida que aumentan las mismas. Esta ruptura no se visualiza de manera muy clara, de tal manera que se podrían tomar tres o cuatro componentes principales.

Este criterio de selección del número de componentes principales se conoce como çriterio del bastón roto".

La tabla 3.1.1 muestra los autovalores, porcentaje y acumulada de la varianza explicada. Según el criterio de Kaiser, se deben tomar 6 componentes para el caso de los pozos que transportan la producción por línea y para el caso de transporte por camión sugiere 5 componentes (autovalores >1) . En cada uno de los casos la varianza explicada sería del 81.90 por ciento y del 78.20 por ciento respectivamente.

Tabla 3.1.1: Porcentaje de la varianza total explicada por cada componente

	Transporte por línea			**Transporte por camión**		
Componente	Autovalor	%	%acum	Autovalor	%	%acum
Componente 1	7.20	37.90	37.90	7.32	43.10	43.10
Componente 2	2.13	11.20	49.10	2.17	12.80	55.90
Componente 3	1.90	10.00	59.10	1.52	8.90	64.80
Componente 4	1.75	9.20	68.30	1.25	7.40	72.20
Componente 5	1.44	7.60	75.90	1.03	6.00	78.20
Componente 6	1.15	6.00	81.90	0.98	5.70	83.90
Componente 7	0.90	4.80	86.70	0.76	4.50	88.40
Componente 8	0.66	3.50	90.20	0.62	3.70	92.10
Componente 9	0.57	3.00	93.20	0.46	2.70	92.10
Componente 10	0.33	1.70	94.90	0.31	1.80	93.90

Tomando en consideración que se cuenta con 21 variables, el hecho de que se pueda explicar más del 70 por ciento de la varianza con apenas 6 nuevas variables, indica que efectivamente existen correlaciones altas entre las variables.

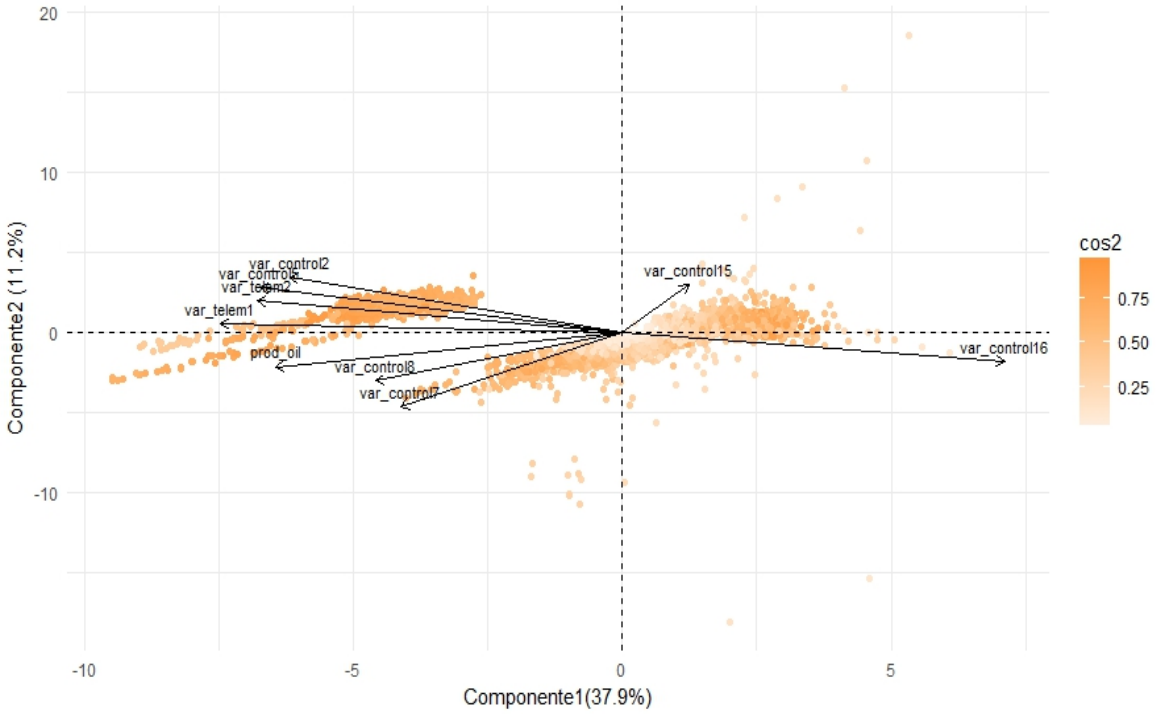
Para realizar la visualización de las variables y sus contribuciones se toman las dos primeras componentes principales.

La figura 3.1.2 muestra las variables sobre las cuales se miden los datos en un espacio de dos dimensiones. Las longitudes de los vectores que representan las variables indican el grado de contribución de la variable y los ángulos de los vectores que representan a las variables indican la correlación que existe entre las mismas. Esta gráfica se denomina biplot.

Se toma especial atención en la variable producción de petróleo *prod_ oil* ya que es la variable de interés.

Analizando el eje de la Componente 1, se observa que las variables se posicionan en

Figura 3.1.2: Biplot para las dos primeras componentes principales. Transporte de producción por línea.



Fuente: elaboracion propia basado en los datos de la batería en análisis.

la parte positiva y negativa del eje, lo que indica que ésta componente es de forma. Las variables ubicadas en el eje negativo tienen valores absolutos altos y son la mayoría. Es decir que, a medida que los valores de las variables crecen el valor de esta componente disminuye. Es interesante resaltar que la *var_ telem1* está ubicada prácticamente sobre el eje negativo de ésta componente y su contribución es la más grande (longitud del vector).

Con lo que respecta al eje de la Componente 2, también de forma, no tiene un valor preponderante entre las variables.

Las variables *var_ control15* y *var_ control16*, ubicadas en el primer y cuarto cuadrante forman un ángulo de casi 180 grados con la variable producción de petróleo *prod_ oil*, lo que indica que su correlación negativa es alta, en menor proporción la variable*var_ control15*. Una explicación a este comportamiento de las variables puede darse tomando en consideración el comportamiento de un pozo surgente[1]. En las primeras etapas de producción se desaloja una gran cantidad de agua, a medida que pasa el tiempo, la cantidad de agua disminuye y la producción de petróleo y/o gas aumenta. Si los valores de variable *prod_ oil* disminuyen, se esperaría que los valores

[1]Referirse al capitulo 1 de este documento: Materiales - Pozos surgentes

de las variables *var_ control15* y *var_ control16* aumenten, de no ser así, se tendría que tomar especial atención a este comportamiento.

Las variables ubicadas en el segundo y tercer cuadrante forman ángulos agudos con la varaible producción de petróleo *prod_ oil*, lo que indica que existe una correlación con la misma. Esto ayuda al momento de seleccionar las variables que se usarán en el modelado.

La calidad de la representación de las observaciones se mide a través del coseno cuadrado[2]. Los valores próximos a 1 indican que las componentes representan bien la observación, y los valores cercanos a 0 indican lo contrario. Esto se hace evidente en el cuarto cuadrante, donde existen puntos con poca representatividad.

Los valores con un color naranja más intenso (representativos) se encuentran distribuidos en el segundo cuadrante. La distribución de representatividad de las observaciones se ubican en los bordes de la nube de puntos, siendo los casos más dispersos en el primer cuadrante.

La figura3.1.3 muestra el biplot correspondiente a las observaciones de los pozos que transportan su producción por camión. Si bien las variables medidas son las mismas para todos los pozos (con excepción de las variables medidas con telemetría, debido a la ausencia de sensores), el comportamiento del conjunto de datos difiere. Esto puede justificarse con la ubicación de los pozos, caracteristicas propias del yacimiento (tipo de suelo, densidad del petróleo, del agua, presencia de minerales, etc), el tipo de separador de control utilizado, el tiempo de ensayo realizado por pozo, el tiempo desde que inicio su producción, etc.

De manera similar al caso anterior, la componente 1 es de forma,teniendo la mayoría de las variables ubicadas en el eje positivo.

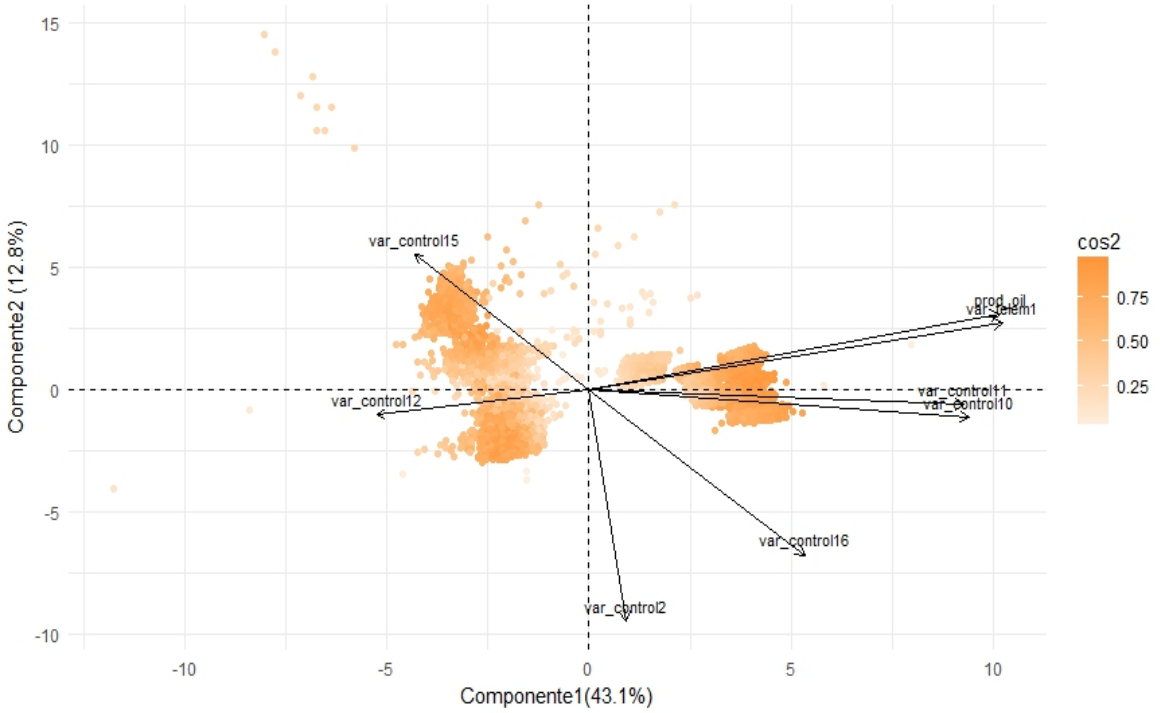
La variable *prod_ oil* está altamente correlacionada con las demás variables que ocupan su mismo cuadrante y el cuarto cuadrante, con excepción de la variable *var_ control2* con la cual muestra independencia (son cuasi perpendiculares entre si), este comportamiento es similar al grupo anterior. Lo que indica que esta variable no sería un indicador para un comportamiento anómalo.

Para el caso de estos pozos un valor alto de la componente 1 representa una producción mayor, al igual que con el grupo anterior está altamente correlacionado con la variable *var_ telem1*.

La variable *var_ control12* forma un ángulo de 180 grados con la variable producción de petróleo *prod_ oil*, lo que implica alta correlación negativa. En el caso anterior se veía este comportamiento con la variable *var_ control16*. Esto indica que, cuando el valor de ésta variable crezca, la producción de petróleo se verá afectada de manera

[2]Referirse al capitulo 1 de este documento: Métodos y Materiales - Métricas de evaluación utilizadas - Coseno cuadrado

Figura 3.1.3: Biplot para las dos primeras componentes principales. Transporte de producción por camión.



Fuente: elaboracion propia basado en los datos de la batería en análisis.

inversa. En caso de que la variable *var_ control12* tenga un comportamiento habitual, en cuanto a sus mediciones, y la producción decaiga, se debe prestar especial atención.

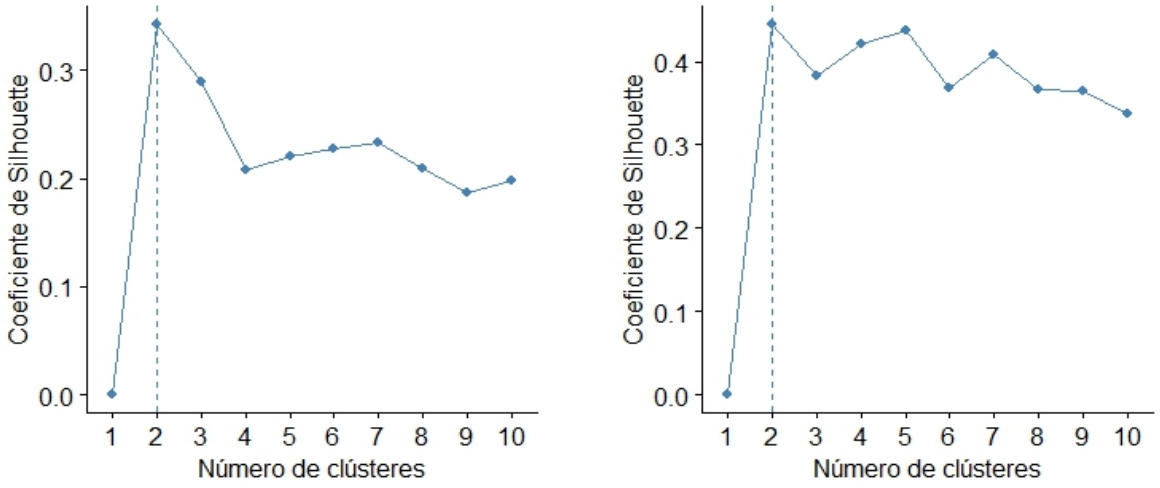
En cuanto a las observaciones, la gráfica muestra puntos con mayor calidad hacia la periferia de la nube de puntos, con mayor preponderancia en el primer cuadrante. Se observa también que los puntos siguen la dirección y el sentido de las variables en el primer y cuarto cuadrante. Los casos poco representativos se observan en el segundo cuadrante y parecerían seguir la dirección y el sentido de la varaible *var_ control15* .

3.2 Análisis de conglomerados o clústeres

Continuando con el análisis del comportamiento de las variables se realizan clústeres sobre el conjunto de datos, dado que podrían existir características de cada agrupación las cuales, al conocerlas, ayuden al diseño del modelo.

Se plantea dibujar clústeres sobre dos primeras componentes principales, para los dos conjuntos de datos. El número de clústeres se selecciona en base al criterio de Silhouette como lo muestra la figura 3.2.1:

Figura 3.2.1: Coeficiente de Silhouette para la selección del número de clústeres



(a) Transporte de producción por línea *(b)* Transporte de producción por camión

Fuente: elaboracion propia basado en los datos de la batería en análisis.

Con dos clústeres se tiene un índice de silueta de 0,36 (figura 3.2.1a) y 0.44 (figura 3.2.1b) respectivamente. Este coeficiente indica que mientras más cercano a 1, las agrupaciones son más sólidas. Si se aumenta el número de clústeres, este coeficiente baja, es decir que los nuevos grupos que se formen no aportarían mayor información al análisis.

Dado que el coeficiente es menor a 0.50, las estructuras encontradas son débiles y podrían ser artificiales. Con el fin de establecer si los datos tienen realmente tendencia a la agrupación, se usa el estadístico de Hopkins.

La hipotesis nula propone que propone que los datos se distribuyen de manera uniforme, es decir, sin evidencia de agrupaciones o clústeres. La hipótesis alternativa establece que los datos no se distribuyen de manera uniforme, es decir, que contiene grupos estadísticamente significativos.

El resultado de la prueba entrega un valor de 0.0202, para el primer conjunto de datos y de 0.0018 para el segundo grupo de datos. Con un nivel de confianza del 90 por ciento.

Por lo tanto se corroboran los resultados obtenidos con el criterio de Silhouette y los datos no tienen una tendencia al agrupamiento.

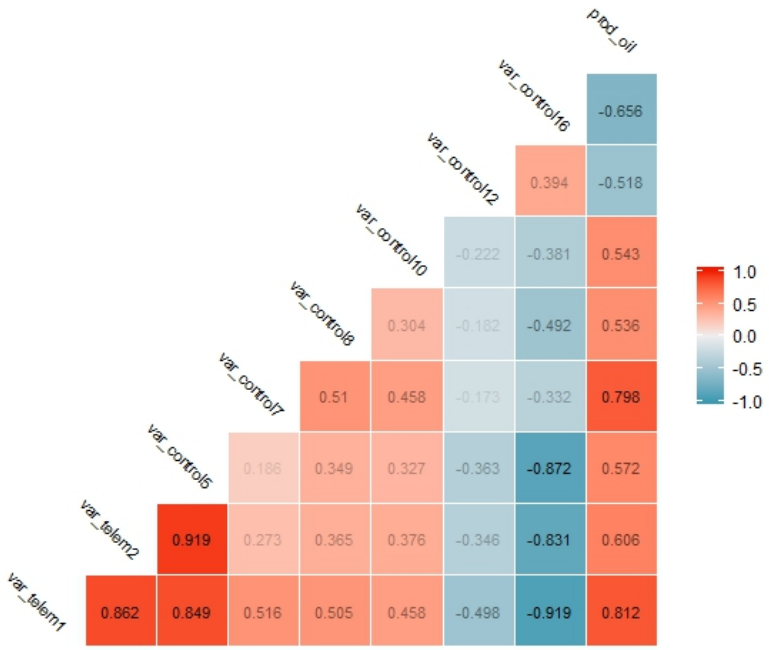
3.3 Análisis de correlaciones

Del Análisis de componentes principales se puede destacar que, para los dos grupos de pozos, las variables medidas con telemetría tienen buena correlación con la producción de petróleo *prod_ oil*. A continuación se seleccionan las variables a utilizar en el modelado del predictor de caudal.

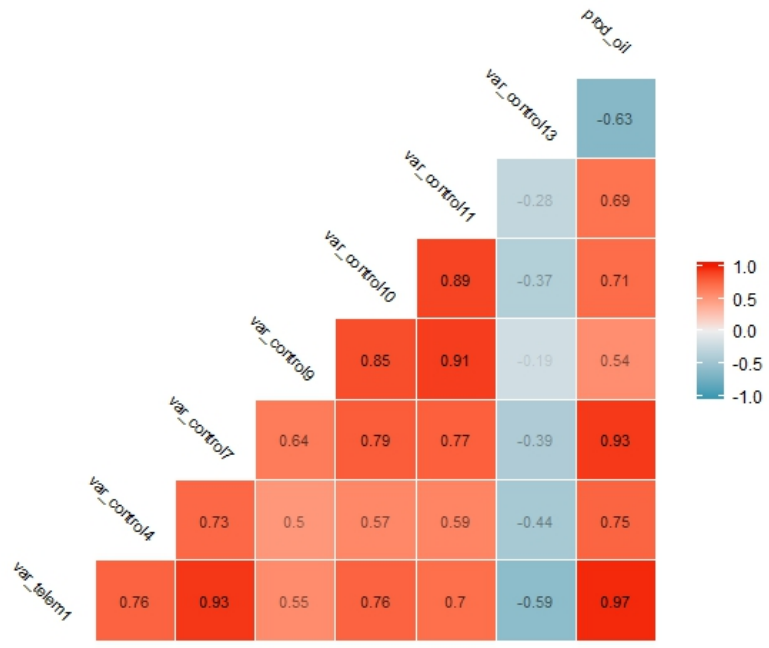
Se presenta en la figura 3.3.1 las matrices de correlación asociadas a las variables. Allí las variables que tienen una correlación positiva alta están asociadas a las celdas rojas, mientras que para el caso negativo corresponden las celdas más azules.

A partir de analizar la misma puede verse, para el caso de los pozos que transportan producción por línea (figura 3.3.1a), las varaibles como las variables *var_ telem1* y *var_ telem2* se encuentran correlacionadas con la variable producción de petróleo en más del 60 por ciento.

Figura 3.3.1: Matriz de correlaciones correspondiente a las variables consideradas.



(a) Transporte de producción por línea



(b) Transporte de producción por camión

Fuente: elaboracion propia basado en los datos de la batería en análisis.

La variable *var_ telem2* y la variable *var_ control5* tienen una correlación muy alta, al momento de seleccionar cual entraría al modelo conviene seleccionar aquella que se mide mediante telemetría, por la facilidad de obtener el dato y no se pierde calidad de la información debido a su alta correlación.

Para el caso del grupo de pozos que transporta su producción mediante camiones (figura 3.3.1b), también se puede ver una correlación alta con la variable medida por telemetría. Se destaca también la correlación existente entre las variables *var_ telem1* y *var_ control7* (0.93), por lo cual, al momento de seleccionar una variable se prefiere aquella que es medida mediante telemetría.

La tabla 3.3.1 a continuación presenta las variables seleccionadas según los análisis previamente realizados. Se toma en consideración además la facilidad de adquisición de la variable y calidad del dato. Se presentan las variables más correlacionadas con la variable de interés y el tipo de asociación existente (negativa o positiva).

Tabla 3.3.1: Variables consideradas con mayor correlación positiva y negativa

Transporte por línea			**Transporte por camión**		
Variable 1	Variable 2	Correlación	Variable 1	Variable 2	Correlación
prod_oil	var_telem1	0.81	prod_oil	var_telem1	0.97
prod_oil	var_telem2	0.61	prod_oil	var_control4	0.75
prod_oil	var_control7	0.78	prod_oil	var_control10	0.71
prod_oil	var_control16	-0.66	prod_oil	var_control13	-0.63

4. MODELADO

A continuación se aplican técnicas de aprendizaje automático para obtener un modelo predictor de caudal (caudalímetro virtual), utilizando el conocimiento adquirido del conjunto de datos realizado anteriormente. Siguiendo la misma línea de trabajo, se realizará un modelo de caudalímetro virtual para los pozos que transportan su producción por línea y otro para aquellos que transportan su producción por camión.

4.1 Conjuntos de datos de entrenamiento y prueba

Del conjunto de datos se separó un subconjunto de los mismos equivalente al tiempo de un control de de pozo (24 horas), sobre la cual se realizará la prueba de reconciliación de datos, con los caudales inferidos por los caudalímetros virtuales.

Los datos restantes se dividieron en dos partes, 70 por ciento para entrenamiento y 30 por ciento para prueba. Las variables utilizadas se describen en la tabla 4.1.1. Estas varibles fueron seleccionadas según el análisis multivariado realizado anteriormente.

Tabla 4.1.1: Variables consideradas para los modelos

Transporte por línea	**Transporte por camión**
var_telem1	var_telem1
var_telem2	var_control4
var_telem3	var_control7
var_control4	var_contro10
var_control7	prod_oil
var_control16	
prod_oil	

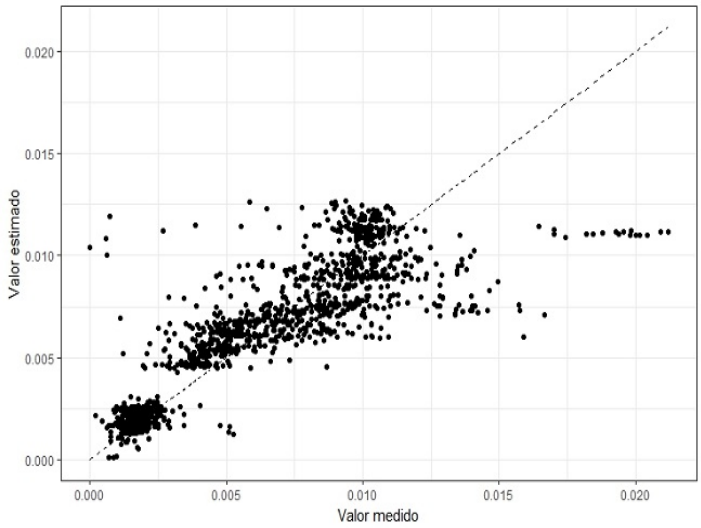
Se calculan tres modelos usando regresión lineal múltiple, support vector regression y redes neronales con las variables especificadas. Una vez realizados los modelos, se selecciona aquel que tenga una menor valor en la raiz del error cuadrático medio, para

continuar con la fase de reconciliación de datos.

4.2 Modelo predictor de caudal. Transporte de producción por línea

La figura 4.2.1 a continuación muestra los valores estimados por el modelo de regresión lineal sobre el conjunto de datos de prueba.

Figura 4.2.1: Modelo usando regresión lineal múltiple



Fuente: elaboración propia en base a datos de la batería en análisis.

Se puede ver que para los estimados con la regresión lineal multiple están dispersos y no tienen tendencia visible sobre la línea ideal de medición.

Sobre este modelo se realizó el procedimiento "paso a paso"(*stepwise*) que permite elegir un subconjunto de variables regresoras, esto con la finalidad de conocer si en realidad las variables que están presentes en el modelo estimarán los valores de manera eficiente; dado que el número de variables influye sobre la varianza del conjunto de datos.

Se aplica el procedimento paso a paso hacia atrás. Éste descarta la variable var_telem3 del modelo (el AIC disminuye de -37386.95 a -37388.22), con esto se consigue que el valor del $R^2 ajustado$ sea de 0.74.

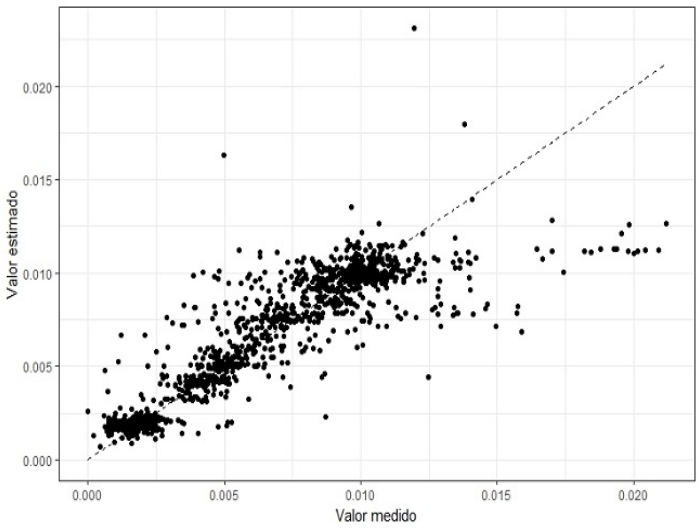
La utilización de un modelo de regresión lineal necesita la verificación de las hipótesis del modelo. Una de estas hipótesis es que los residuos deben ser independientes (no estar correlacionados). Para corroborar esto se realiza la prueba estadística Durbin-Watson, la cual indica si los residuos son o no indepenedientes. La prueba tiene la hipótesis nula de que la autocorrelación de los residuos es 0.

En este caso el valor de la prueba estadística muestra que existe una auto-correlación negativa (-0.018 con p-valor igual 0.304), por lo cual se descarta la hipótesis nula. Por este motivo se descarta el uso del modelo de regresión lineal múltiple para este grupo de pozos, ya que sus estimadores son ineficientes.

La figura 4.2.2 muestra los valores estimados usando Support Vector Regression. Para el caso de SVR los valores de caudal más bajos se encuentran más aglutinados y no dispersos sobre la recta ideal. Destaca también en la gráfica unas observaciones de caudal cuya estimación es alta en comparación con su valor real.

El ajuste del modelo SVR se puede realizar ya que la técnica proporciona flexibilidad con respecto al error máximo y el costo de penalización. Se realizaron estos ajustes al modelo con el fin de optimizar los parámetros para mejorar las predicciones.

Figura 4.2.2: Modelo usando*Support Vector Regression*



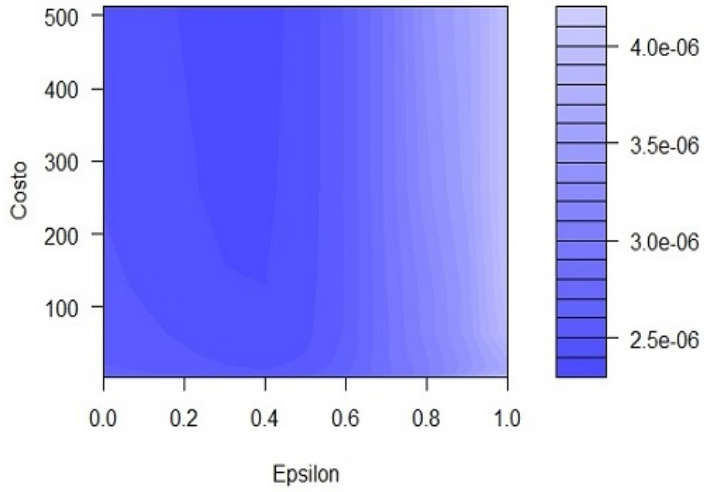
Fuente: elaboración propia en base a datos de la batería en análisis.

Este ajuste se realizó usando el enfoque de búsqueda por cuadrícula. Se realizaron 168 modelos variando el parámetro *epsilon* entre 0 hasta 1 en pasos de 0.1 y el parámetro *costo* entre 2^2 hasta 2^9. La figura 4.2.3 muestra el rendimiento de cada uno de los modelos.

La métrica utilizada en esta figura es la raíz del error cuadrático medio ($RMSE$ por sus siglas en inglés). El mejor modelo es el que tiene el $RMSE$ más bajo.

La región más oscura de la gráfica representa un $RMSE$ más bajo, en este caso se tiene esa característica en la sección definida entre 0.2 hasta 0.4 en el eje x y entre 400 hasta 520 para el eje y. Se vuelve a repetir el ajuste del modelo estableciendo estos nuevos valores. Finalmente los parámetros fijados entonces para el modelo son: *epsilon* $= 0.3$ y *costo* $= 500$.

Figura 4.2.3: Desempeño de Support Vector Regression, variando los parámetros de costo y epsilon



Fuente: elaboración propia en base a datos de la batería en análisis.

Por último, la figura 4.2.4 muestra los valores estimados usando una red neuronal.

Para establecer el número de neuronas en la capa oculta se realizaron modelos variando el número de las mismas. La recomendación habitual es usar $n - 1$ neuronas en la capa oculta, donde n es el número de variables.

La tabla 4.2.1 muestra los valores de $RMSE$ para 5 modelos entrenados. Se seleccionó aquel con menor valor de $RMSE$.

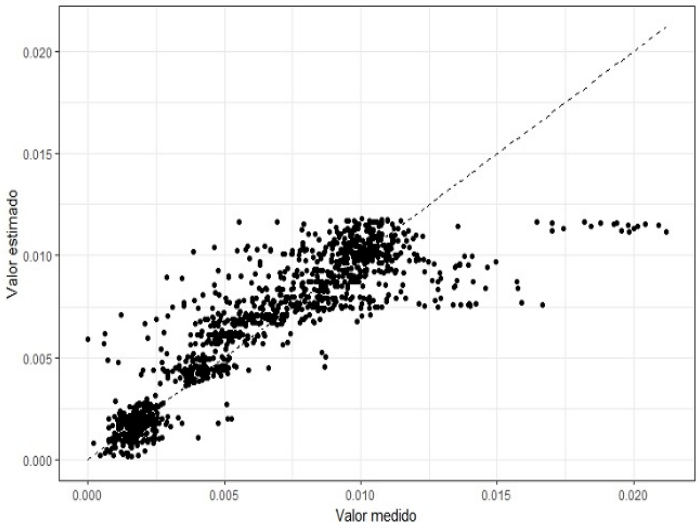
Tabla 4.2.1: Desempeño de la red neuronal, cambiando el número de neuronas en la capa oculta

# de neuronas en la capa oculta	**Valor de RMSE**
6	0.001830
7	0.001841
8	0.001837
9	0.001851
10	0.001883

A pesar de que el valor de $RMSE$ no tiene cambios significativos el más bajo se encuentra usando 6 neuronas en la capa oculta, de tal manera que se fijará este valor para realizar las estimaciones del modelo.

Los valores estimados de caudal se distribuyen de mejor manera a los largo de la

Figura 4.2.4: Modelo usando redes neuronales



Fuente: elaboración propia en base a datos de la batería en análisis.

recta ideal, en comparación con los modelos anteriores, sin embargo sigue presentando valores que no se encuentran estimados de manera correcta. Sin embargo no se observan valores estimados mayores a los valores medidos como en el caso del modelo con Support Vector Regression.

A continuación se muestran los valores de $RMSE$ para cada uno de los modelos (tabla 4.2.2):

Tabla 4.2.2: Raíz del error cuadrático medio para los modelos

Modelo	**Valor de RMSE**
Regresión lineal múltiple	0.002101
Support Vector Regression	0.001902
Red neuronal	0.001837

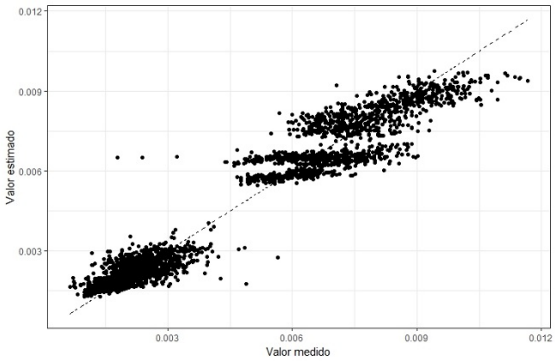
De tal manera que el modelo seleccionado para los pozos que transportan su producción mediante línea es la red neuronal.

4.3 Modelo predictor de caudal. Transporte de producción por camión

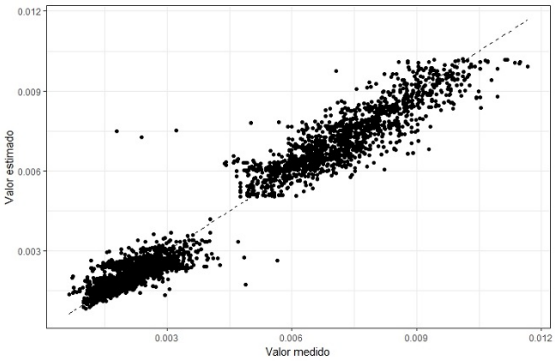
Se realiza un análisis similar para los pozos ubicados en la zona alejada de la batería en análisis.

La figura 4.3.1 muestra los valores estimados vs. los valores medidos para los pozos que transportan la producción mediante camiones.

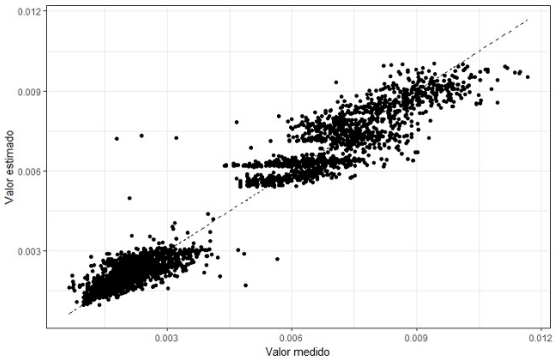
Figura 4.3.1: Modelos predictores de caudal. Conjunto de datos de prueba.



(a) Regresión lineal múltiple



(b) Support Vector Regression SVR



(c) Red neuronal

Fuente: elaboración propia en base a datos de la batería en análisis.

La regresión lineal múltiple 4.3.1a presenta valores muy dispersos, se puede ver que para caudales bajos, los estimados son mayores. Se realizó el procedimiento paso a paso para descartar variables regresoras innecesarias, sin embargo todas fueron tomadas en cuenta, el valor de $R^2 ajustado$ es de 0.94.

Nuevamente se realizó una prueba de independencia de los residuos de la regresión lineal múltiple. La prueba estadística Durbin-Watson indica que existe una autocorrelación positiva en los residuos de la regresión, por lo cual se descarta este modelo.

Para Support Vector Regression los valores de los parámetros *epsilon* y *costo* son de 0.2 y 512 respectivamente. La figura 4.3.1b muestra que los valores estimados se distribuyen de mejor manera sobre la recta ideal, existen puntos muy alejados, pero en menor proporción que con la regresión lineal múltiple.

Finalmente la figura 4.3.1c muestra los datos estimados usando una red neuronal, el número de capas ocultas es de 7 ($NRSME$ de 0.0006).

A continuación se muestran los valores de $RMSE$ para cada uno de los modelos (tabla 4.3.1):

Tabla 4.3.1: Raíz del error cuadrático medio para los modelos

Modelo	**Valor de RMSE**
Regresión lineal múltiple	0.006262
Support Vector Regression	0.005937
Red neuronal	0.005397

De tal manera que el modelo seleccionado para los pozos que transportan su producción mediante camión es la red neuronal.

4.4 Reconciliación de datos

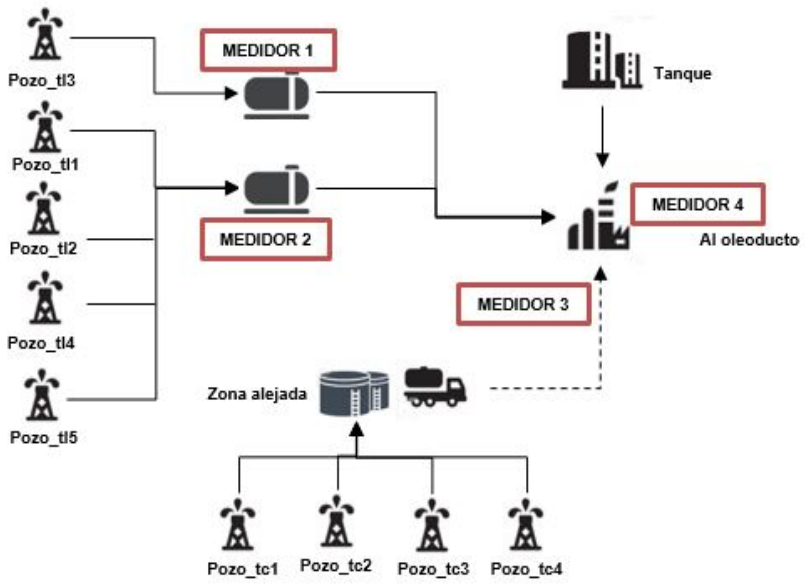
Con el fin de contrastar los datos medidos con los entregados por los caudalímetros virtuales se realiza una simulación del proceso de cierre de producción en la batería. Para esto se toma como ejemplo los datos recolectados en los separadores móviles, tanques, sensores de telemetría y controles de pozo de un día específico. Se deben tomar en cuenta las siguientes consideraciones:

- Para realizar esta simulación se toman los datos correspondientes a una prueba de control de pozo realizada durante 24 horas tener una mejor aproximación a la realidad en cuanto al proceso de cierre de producción. Este lapso de tiempo se seleccionó tomando en cuenta que se contaba con los datos para el separador general, separador de control, controles de pozo y telemtría, así como del transporte de crudo por camión.
- Se asigna un porcentaje de error a los caudalímetros, el correspondiente a $RSME$. En el caso de que los valores estimados por el modelo presenten un error mayor a dos veces el $RSME$ se utiliza los datos correspondientes a los controles de pozo en dicha ubicación.
- La batería cuenta con un tanque en sus instalaciones, no se cuenta con el volumen histórico del tanque en el sistema por lo cual se infiere el volumen de crudo que almacena el tanque en el lapso de tiempo de análisis.

La figura 4.4.1 presenta un diagrama de cómo se transporta el petróleo desde los pozos hasta llegar a la batería. El diagrama muestra además las etiquetas de los medidores de caudal que censan actualmente los datos para el sistema de telemetría.

Para la fecha seleccionada se encontraban en producción los siguientes pozos: *pozo_ tl1*,*pozo_ tl2*,*pozo_ tl3*,*pozo_ tl4*,*pozo_ tl5*,*pozo_ tc1*,*pozo_ tc2*,*pozo_ tc3* y*pozo_ tc4*.

Figura 4.4.1: Diagrama de funcionamiento de la batería en estudio



Fuente: elaboración propia en base a datos de la batería en análisis.

4.4.1. Descripción del ambiente de prueba

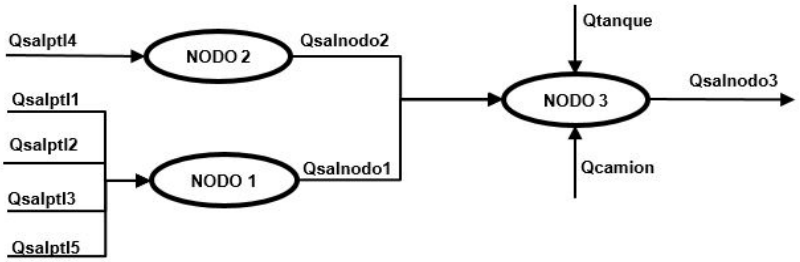
La figura 4.4.2 muestra el sistema de reconciliación a analizar. El sistema cuenta con 3 nodos, correspondientes a 3 lugares donde se miden la producción al cierre. Las variables del sistema son los pozos, el tanque y el colector de la batería.

- El **NODO 1** recibe la medida de producción de petróleo que corresponde a los pozos cuyo caudal se deriva al separador general.
- El **NODO 2** recibe la medida de producción de petróleo del pozo que se encuentre en control en esa fecha.
- El **NODO 3** recibe las medidas de producción de petróleo correspondiente a los pozos que lo transportan por camiones, la producción a la salida a la salida de los separadores de control y general y lo que aporta el tanque de la batería.

Según el registro del ensayo de control de pozo (línea) se tienen los siguientes datos:

- El ensayo inicia a las 10:00 am del día uno y su duración fue de 24 horas.
- Durante ese tiempo la producción del *pozo_ tl4* estuvo derivandose al separador de control, el resto de pozos derivan su producción al separador de general.

Figura 4.4.2: Esquema del sistema de reconciliación

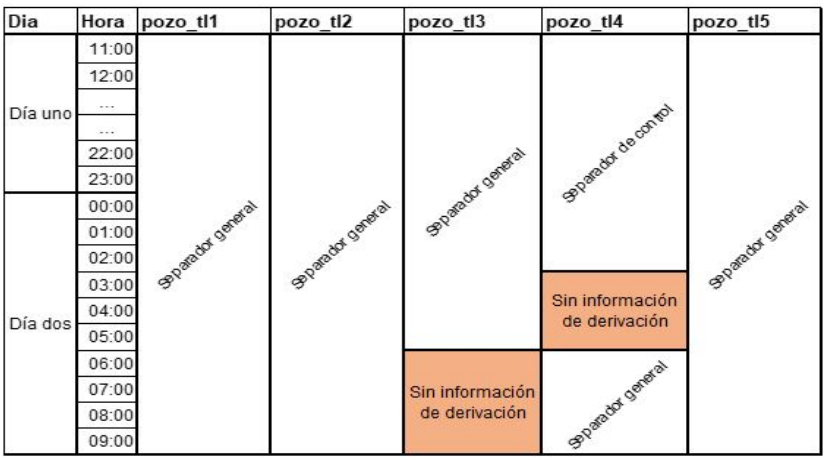


Fuente: elaboración propia en base a datos de la batería en análisis.

- En la planilla de control de pozo del separador de control, no se tienen registros del caudal medido desde las 03:00 am hasta las 06:00 am del día dos.
- A las 06:00 am del día dos se tienen registros del pozo *pozo_ tl3* en el separador general. Por lo cual se infiere que éste pozo paso a la prueba de ensayo de control. El pozo *pozo_ tl4* no completó las 24 horas de ensayo.
- En la planilla de control de pozo del separador general se empieza a censar el caudal del pozo *pozo_ tl4* (antes en control) desde las 06:00 am hasta las 09:00 del día dos (finalización de la prueba).
- A las 06:00 am sale del registro del separador general los datos censados del pozo *pozo_ tl3*, por lo cual se infiere que paso al separador de control.

La figura 4.4.3 muestra los pozos que derivan su producción al separador general y al separador de control.

Figura 4.4.3: Registro de control de pozo realizado

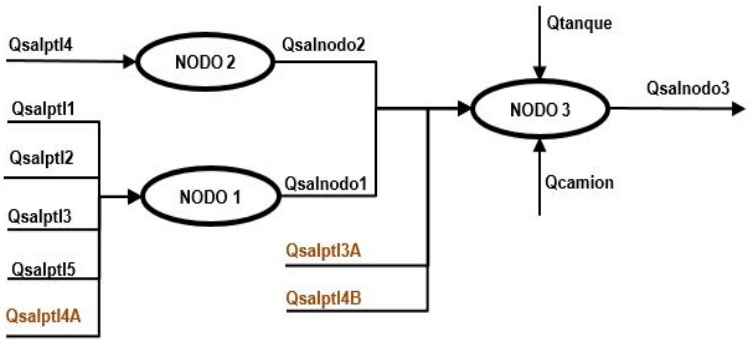


Fuente: elaboración propia en base a datos de la batería en análisis.

Las celdas resaltadas corresponden a aquella ventanas de tiempo en las cuales no se tienen registros en las planillas de control de pozo. A fin de que la prueba sea lo más cercana a la realidad, se realiza una modificación al esquema de prueba. La producción correspondiente a los pozos *pozo_tl3* y *pozo_tl4* donde no se tiene se deriva al nodo 3 directamente.

El esquema final que se utiliza para la reconciliación se muestra en la figura 4.4.4:

Figura 4.4.4: Esquema del sistema de reonciliación modificado



Fuente: elaboración propia en base a datos de la batería en análisis.

4.4.2. Proceso de reconciliación de datos

Para el proceso de reconciliación se toman los valores entregados por telemetría en las salidas de los separadores de control y general, en el tanque y por camiones, en la ventana de tiempo seleccionada para la simulación de cierre de control de producción.

4.4.2.1. Valores de caudal estimados por el modelo

La tabla 4.4.1 muestra los valores de caudal entregados por los caudalímetros virtuales, la desviación estánda correspondiente a cada pozo (transporte por línea y por camión) y los caudales de salida de los nodos 1 y 2 hacia el nodo 3 (batería).

Los caudales inferidos por los caudalímetros virtuales en la zona alejada de la batería tienen un error considerable, por esta razón se toma el valor ingresado por en la planilla de control de pozo para esa fecha.

Finalmente, se necesita conocer el valor que ingresa al nodo 3 desde el tanque en la batería. No se cuenta con datos históricos en el sistema de telemetría, por lo cual, se establece un valor correspondiente a la mitad de su capacidad total.Se selecciona este valor con el fin de minimizar el error introducido, es decir, que se tendría un error de máximo la mitad de la capacidad del tanque.

Tabla 4.4.1: Valores de caudal desde y hacia los nodos del sistema de reconciliación y el error correspondiente. Los valores se encuentran escalados.

Etiqueta	**Desde**	**Hacia**	**Caudal(m^3)**	**Desviación estándar**
$Qsal_{ptl1}$	pozo_tl1	Nodo_1	1.442	0.403
$Qsal_{ptl2}$	pozo_tl2	Nodo_1	1.496	0.224
$Qsal_{ptl3}$	pozo_tl3	Nodo_1	1.312	0.240
$Qsal_{ptl4}$	pozo_tl4	Nodo_2	1.238	0.190
$Qsal_{ptl5}$	pozo_tl5	Nodo_1	1.204	0.022
$Qsal_{ptl3A}$	pozo_tl3	Nodo_3	1.156	0.204
$Qsal_{ptl4A}$	pozo_tl4	Nodo_1	1.149	0.172
$Qsal_{ptl4B}$	pozo_tl4	Nodo_3	1.136	0.170
Q_{camion}	Zona alejada	Nodo_3	2.00	0.200
Q_{tanque}	Tanque	Nodo_3	1.116	0.56
$Qsal_{nodo1}$	Nodo_1	Nodo_3	6.603	
$Qsal_{nodo2}$	Nodo_2	Nodo_3	1.238	
$Qsal_{nodo3}$	Nodo_3	Oleoducto	10.957	

4.4.2.2. Matriz de incidencia y matriz de varianza

Para calcular los valores reconciliados son necesarias las matrices de incidencia y varianza

Primero se crea la matriz A de incidencia. La matriz será una de 3x13 con valores de 0 y 1, donde las filas representan los nodos ordenados y las columnas los distintos flujos (según el orden indicado en la tabla 4.3.1 en la columna etiquetas), con signo positivo si son entrantes y negativo si son salientes, como lo muestra la figura 4.4.5:

Figura 4.4.5: Matriz de incidencia

	$Qsal_{ptl1}$	$Qsal_{ptl2}$	$Qsal_{ptl3}$	$Qsal_{ptl4}$	$Qsal_{ptl5}$	$Qsal_{ptl3A}$	$Qsal_{ptl4A}$	$Qsal_{ptl4B}$	Q_{camion}	Q_{tanque}	$Q_{salnodo1}$	$Q_{salnodo2}$	$Q_{salnodo3}$
Nodo 1	1	1	1	0	1	0	1	0	0	0	-1	0	0
Nodo 2	0	0	0	1	0	0	0	0	0	0	0	-1	0
Nodo 3	0	0	0	0	0	1	0	1	1	1	1	1	-1

Fuente: elaboración propia en base a datos de la batería en análisis.

A continuación se determina la matriz de varianza de los caudales inferidos, ésta

matriz se obtiene multiplicando el vector de varianzas (desviación estándar al cuadrado) por la matriz identidad, de la misma dimensión que la matriz de incidencia, de tal manera que las varianzas corresponden a la diagonal de la matriz.

Los resultados del cálculo de los valores reconciliados[1] para cada medición se enumeran en la tabla 4.4.2

Tabla 4.4.2: Resultados de la reconciliación

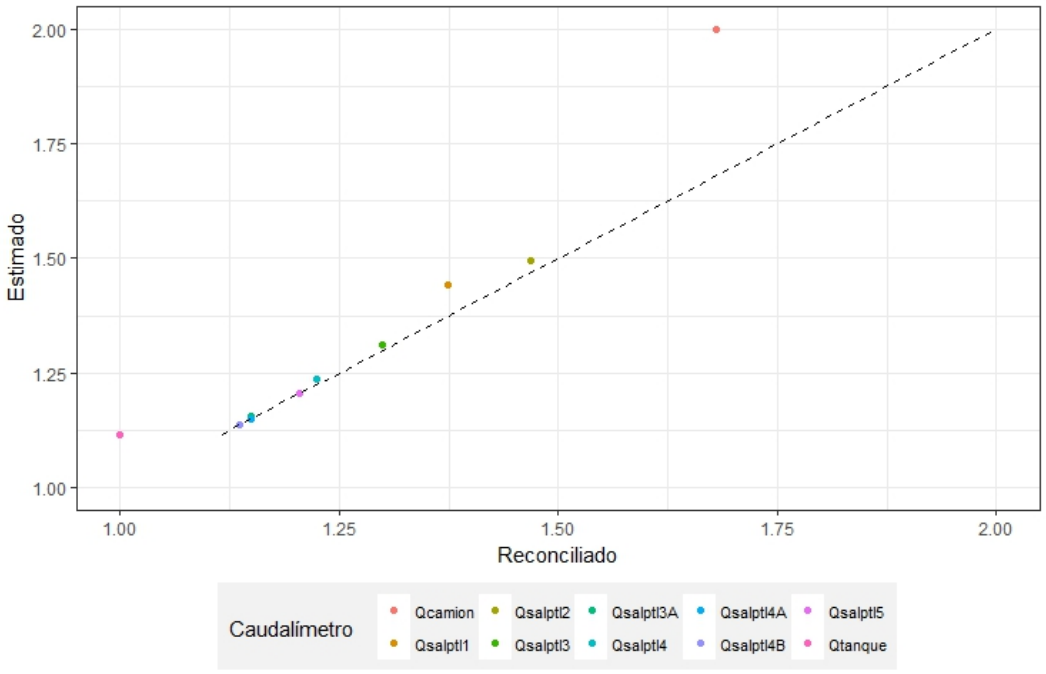
Etiqueta	**Valor estimado** (m^3)	**Valor reconciliado**(m^3)	**Ajuste**
$Qsal_{ptl1}$	1.442	1.374	-0.068
$Qsal_{ptl2}$	1.496	1.469	-0.027
$Qsal_{ptl3}$	1.312	1.299	-0.013
$Qsal_{ptl4}$	1.238	1.224	-0.014
$Qsal_{ptl5}$	1.204	1.204	0
$Qsal_{ptl3A}$	1.156	1.149	-0.007
$Qsal_{ptl4A}$	1.149	1.149	0
$Qsal_{ptl4B}$	1.136	1.136	0
Q_{camion}	2.00	1.680	-0.320
Q_{tanque}	1.116	1.000	-0.116
$Qsal_{nodo1}$	6.603	6.494	-0.109
$Qsal_{nodo2}$	1.238	1.224	-0.014
$Qsal_{nodo3}$	10.957	10.473	-0.484

Observando la figura 4.4.6a, de los valores estimados vs. los valores reconciliados, se identifica a priori que, de existir un error, éste estará en el caudal que ingresa al Nodo3 mediante camiones desde la zona alejada de la batería. También resalta el valor entregado por el tanque. Cabe recalcar que, como no se contaba con los datos históricos de los aportes del tanque, se esperaría este resultado.

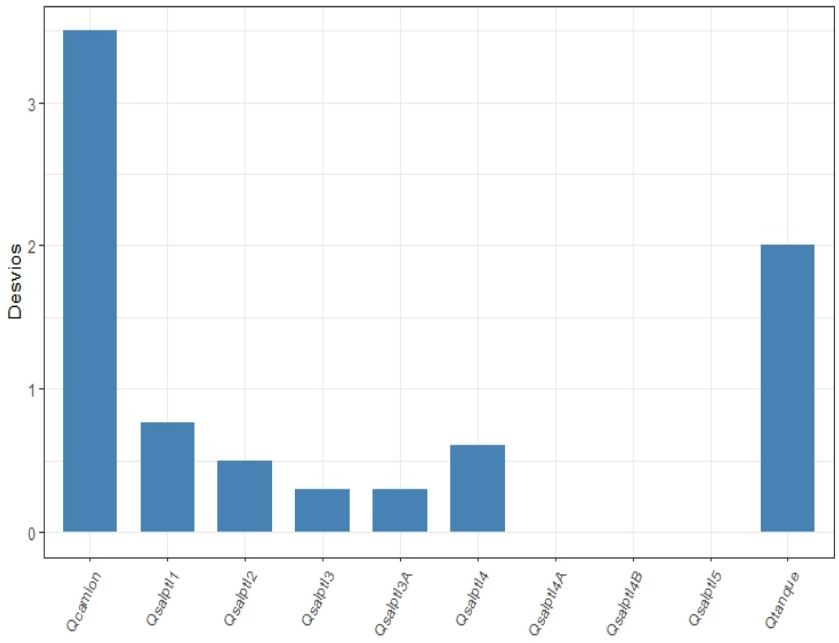
La figura de barras 4.4.6b muestra también de manera clara aquellas mediciones que indican un error.

[1]Ecuación para cálculo de valores reconcilados $\hat{y} = y - V * A^T * (A * V * A^T)^-1 * A * y$ donde A es la matriz de incidencia, V la matriz de varianza y y los valores medidos

Figura 4.4.6: Resultados del proceso de reconciliación.



(a) Valores estimados vs. valores reconciliados



(b) Desviaciones estándar entre el valor medido y el valor reconciliado

Fuente: elaboración propia en base a datos de la batería en análisis.

4.5 Discusión y conclusiones

El propósito general de este proyecto consistió en acelerar el proceso de localización de pérdidas en la producción de petróleo, identificando cuando existe una desviación significativa cualquiera de las mediciones monitoreadas en los pozos. Para esto se elaboró un modelo estimador de caudal o çaudalímetro virtual"basado en redes neuronales que entrega un valor de caudal de petróleo basado las mediciones tomadas en el pozo. Estas mediciones se toman mediante telemetría y controles realizados en el pozo.

Se realizó una simulación del proceso de cierre de producción utilizando los valores entregados por los caudalímetros virtuales. En cada uno de los puntos de medición se tomó en cuenta el error generado por el modelo propuesto. El error considerado es el que existe al comparar la medición en los separadores de control (fijo y/o móvil) registrado en las planillas de control de pozo con la medición entregada con los caudalímetros virtuales. Al momento de realizar la simulación se añade el error generador propiamente por el proceso de cierre de producción, visto así, el error que genera el modelo sería despreciable.

El proceso de reconciliación de datos necesita conocer exactamente el funcionamiento de la batería en análisis, ya que se necesitan fijar puntos (nodos) para aplicar la metodología de este procedimiento. En este caso se hace necesario conocer que pozos enviaban su producción al separador general y cuales al separador de control, la capacidad del tanque de almacenamiento del tanque de la batería, así como los valores históricos.

El proceso de reconciliación de datos entrega valores estadísticamente más confiables de los valores medidos. El tomar estos valores para recalibrar los caudalímetros disminuirían el error generado con los mismos. De la misma manera, el proceso de reconciliación puede realizarse en cualquier momento, no es necesaria una ventana de tiempo para su realización.

El porcentaje de error más alto se encontró en las medidas de caudal del petróleo transportado por camión y en el nivel del tanque de la batería. Para el proceso de reconciliación de datos se tomaron las mediciones correspondientes a una prueba de pozo de 24 horas. Se separaron del conjunto de datos éstas mediciones, de tal manera que no haya un sesgo en la elaboración del modelo.

Los pozos ubicados en la zona alejada de la batería no cuentan con todos los sensores instalados, es por esa razón que utiliza un modelo diferente a los pozos que transportan su producción por líneas. De contar con esta información se podría mejorar el ajuste de los caudalímetros virtuales para esta locación utilizando estas nuevas varaibles.

El modelo propuesto en este estudio aplica únicamente a pozos surgentes, ya que las características de las variables que se miden para este tipo de pozos contribuyen al modelo.

Para mejorar el ajuste de estos modelos propuestos se recomienda que, al realizar los controles de pozo en el separador, tanto móvil como en la batería, se tomen las 24 horas de control. Al mejorar la calidad de los datos se podrían recalibrar los caudalímetros virtuales.

Este modelo aplica únicamente para pozos surgentes pero, es posible el desarrollo de un modelo para pozos que funcionen con otros métodos de extracción. Para esto es necesario un conjunto de datos apropiado para realizar este el análisis.

5. ANEXO

5.1 Métodos y Materiales

5.1.1. Yacimiento de petróleo

En Geología se denomina yacimiento al lugar donde se encuentra de forma natural una roca, un mineral o un fósil. De ahí que un yacimiento de petróleo es un lugar donde se ha acumulado de forma natural petróleo crudo o ligero. Este mineral se encuentra retenido por formaciones de rocas suprayacentes con baja permeabilidad [1]. Suele ser denominado también reservorio o depósito.

5.1.2. Pozo petrolífero

Un pozo petrolífero es una obra de ingeniería encaminada a poner en contacto un yacimiento de hidrocarburos con la superficie. Es una perforación efectuada en el subsuelo con barrenas de diferentes diámetros y con revestimiento de tuberías, a diversas profundidades, para la prospección o explotación de yacimientos[2].

5.1.2.1. Pozo surgente

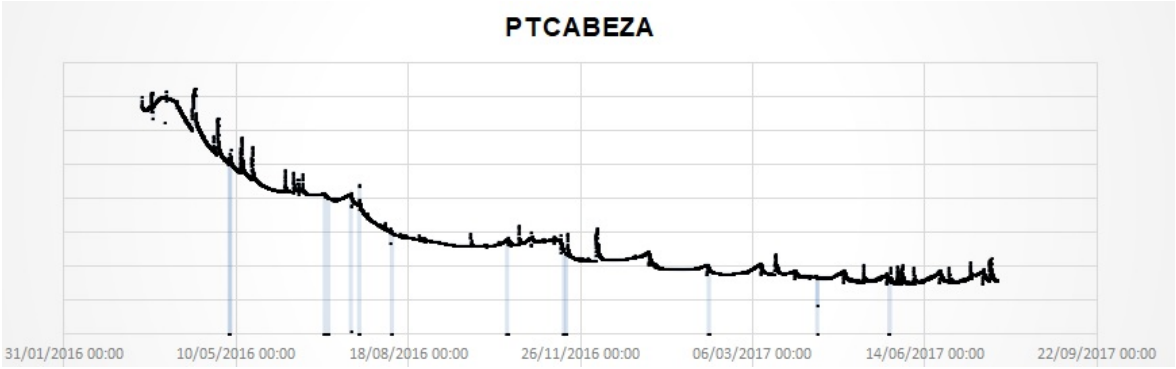
Figura 5.1.1: Pozo surgente. Árbol de navidad

Fuente: ABC de la Industria del Petroleo y Gas del IAPG (Instituto Argentino de Petroleo y Gas).

Se dice que un pozo es surgente o está en surgencia natural cuando la energía del yacimiento (presión de fondo) es suficiente para impulsar los fluidos hasta la superficie. Esta presión se produce en las primeras etapas de vida productiva del pozo y va decayendo hasta que se hace necesario instalar un sistema de extracción artificial. La figura 5.1.1 muestra la instalación de un conjunto de válvulas conocido como "Árbol de Navidad", característico de pozos en surgencia.

En la primera etapa de producción del pozo, los tres primeros meses, la presión de formación es inestable, debido al desalojo de agua y gas en exceso resultado de la perforación del pozo. Una vez que se estabiliza el corte de agua, la presión empieza a decaer de manera exponencial decreciente, como muestra la figura 5.1.2 a continuación:

Figura 5.1.2: Curva de presión de pozo surgente



Fuente: elaboración propia en base a datos de la batería en análisis.

5.1.2.2. Utilización de orificios

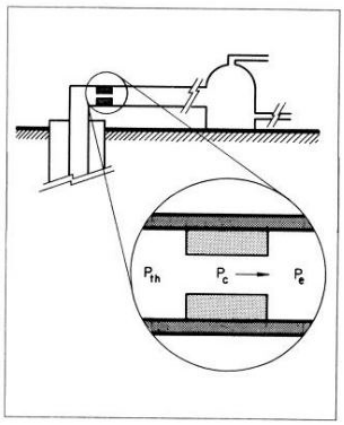
La surgencia se regula en la cabeza (boca) de pozo[1] mediante un pequeño orificio (estrangulador, choke o bean) como muestra la figura 5.1.3, cuyo diámetro dependerá del régimen de producción elegido. Los orificios se utilizan, en general, durante la primera etapa de producción. Con el correr del tiempo los diámetros se van incrementando y eventualmente el orificio es retirado. Si bien la mayoría de los pozos surgentes utilizan orificios en boca de pozo, en casos especiales los mismos se instalan en el fondo.

El orificio o estrangulador se instala en la boca del pozo y su funcionamiento se basa en el principio de flujo crítico[2], por lo cual la elección de su diámetro se realiza pensando en que no afecte a la presión en la cabeza del pozo y como consecuencia no repercuta en la producción[5].

[1] La boca de pozo involucra la conexión de las cañerías de subsuelo con las existentes en superficie que se dirigen a las instalaciones de producción.

[2] El flujo crítico es aquel que viaja a una velocidad equivalente a la velocidad de propagación, de manera que el cambio en la presión después del estrangulador no afecte la presión antes del mismo.

Figura 5.1.3: Estrangulador u orificio instalado en la cabeza de pozo



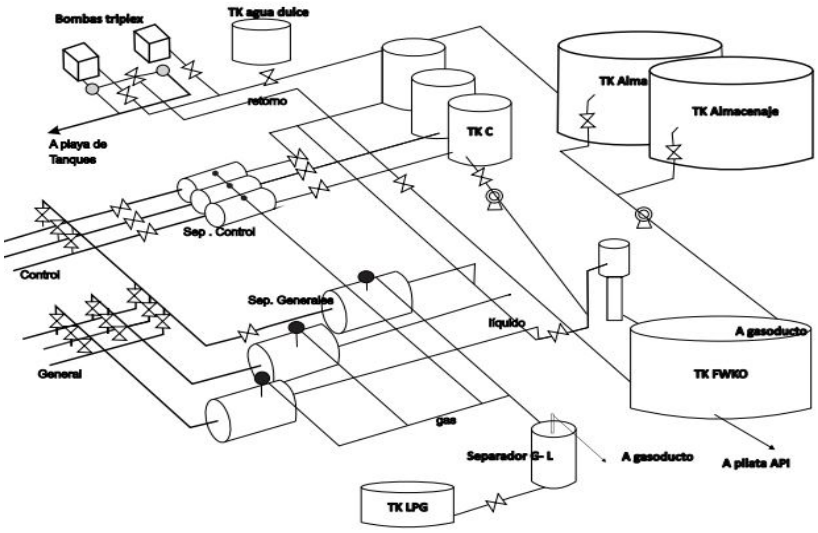
Fuente: Wellhead Surface Equipment.Oil & Gas Process Engineering.

5.1.3. Batería de producción

Las *baterías de producción* también conocidas como facilidades de producción y/o estaciones de producción, son instalaciones que reciben o recolectan la producción de varios pozos de un yacimiento.

La figura 5.1.4 muestra el esquema de una batería estándar, este esquema puede variar dependiendo de las necesidades del lugar. Es aquí donde se realiza la separación primaria de los fluidos. En el caso de petróleos viscosos, aquí es donde se efectúa su calentamiento para facilitar el bombeo a la planta de tratamiento[3].

Figura 5.1.4: Diagrama de una batería estándar



Fuente: Producción y transporte en la industria petrolera. Expo-Petrol 2012 – SPE-UNP.

También son en estas instalaciones donde se realiza la medición diaria del volumen total producido y el control de pozos individual.

5.1.3.1. Colectores

El fluido de cada pozo es transportado por medio de líneas de conducción hasta el colector de la batería. El colector se compone de varias bocas de entrada y está diseñado para derivar el fluido del pozo a la línea general o a cualquiera de la/las líneas de control (figura 5.1.5):

Figura 5.1.5: Colector de la batería.

Fuente: Producción y transporte en la industria petrolera. Expo-Petrol 2012 – SPE-UNP.

5.1.3.2. Separadores

Los separadores de petróleo y gas (figura 5.1.6) tienen la función de separar los componentes líquidos y de gas existentes bajo una temperatura y presión específica mecánicamente, para eventualmente procesarlos en productos vendibles. Estos separadores son cilindros metálicos, cerrados, con casquetes en los extremos, pueden ser verticales y horizontales según su disposición geométrica.

Una batería, por lo general, tiene un separador general y uno de control. Estos se diferencian únicamente por los métodos de medición en cada uno. Un separador general censa el volumen de agua, gas y petróleo de todos los pozos que convergen a la batería con excepción de uno, cuya producción se deriva al separador de control. Dependiendo del esquema de planificación de controles de pozo, se medirá el volumen de petróleo, agua y gas que produce en una ventana de control establecida por el supervisor de producción. Más adelante se describe un control de pozo con más detalle.

Figura 5.1.6: Separadores en la batería.

Fuente: Separadores de Líquidos, Gases y Sólidos.Bolland y Cía S.A. servicios petroleros.

Figura 5.1.7: Tanques en la batería.

Fuente: Tanques en la Refinería Ricardo Elicabe (Bahía Blanca) de Petrobras Energía.

5.1.3.3. Tanques

La figura 5.1.7 muestra los tanques en una batería. Están como etapa intermedia entre la separación y el bombeo. La capacidad del tanque debe tener relación directa con el volumen total de fluido con que opera la batería. Estos tanques deben tener la capacidad de almacenar al menos de 10 a 12 horas de producción. De acuerdo al tamaño de la batería y al parámetro de capacidad de almacenaje se eligen las dimensiones de los tanques. Las baterías, además, están equipadas con tanques de menor tamaño para controlar pozos y/o recibir la descarga de los separadores de control.

5.1.4. Control de Producción

Producción es el sector de la Industria Petrolera que hace realidad todo el esfuerzo y la inversión llevada a cabo desde que se empieza a explorar una zona. Cada sec-

tor (exploración, geología, perforación, reservorios, etc.) ven justificados sus esfuerzos cuando el petróleo esta en superficie, en condiciones de ser comercializado.

El supervisor de Producción debe realizar el seguimiento continuo de la producción de fluidos de la batería a su cargo. El objetivo de esta tarea es detectar pérdidas y generar, rápidamente, acciones correctivas. Para ello es necesario hacer una rutina diaria que permita dar con estas anormalidades y solucionarlas en el momento. Y si no se puede remediar inmediatamente, es importante conocerlas para programar su solución. No es técnicamente correcto perder producción sin saber (tener detectado) el origen de la pérdida.

5.1.4.1. Control operativo (control de pozo)

Un control de pozo se realiza con la finalidad de conocer el estado de producción de los pozos de la batería. Se realiza un censo de las diferentes variables físicas de un pozo y su evolución, con una frecuencia de una hora. Estos controles de pozo tienen, por lo general, una duración de 24 horas. Este último parámetro podría variar dependiendo del criterio del supervisor de produción.

La frecuencia con la que se realizan estos controles de pozo puede quedar determinada por la diferente jerarquía fijada de acuerdo al criterio ABC. Este criterio consiste en clasificar a los pozos de acuerdo a su producción neta de la siguiente manera:

- **Tipo A**: Aquellos pozos cuya producción suma del 60 al 70 por ciento de la producción de petróleo de la batería a la que corresponden.
- **Tipo B**: Aquellos pozos cuya producción suma del 20 al 30 por ciento de la producción de petróleo de la batería a la que corresponden.
- **Tipo C**: Aquellos pozos cuya producción suma el 10 por ciento restante de la producción de petróleo de la batería a la que corresponde .

Una vez establecida la clasificación de los pozos, se opta por asignar a los pozos más importantes (Tipo A) una frecuencia mayor de controles, para este caso de estudio es de 3 veces al mes, los pozos clasificación Tipo B se controlarán 2 veces al mes y, por último, los pozos clasificiación Tipo C una vez al mes.

5.1.4.2. Cierre de producción

Todos los días, a una hora prefijada se realiza el cierre de existencias de producción, con la finalidad de conocer lo producido al día. Este valor debe coincidir con el valor previsto o comprometido. De no ser así, o ser menor, la diferencia se considera *merma o*

pérdida de producción. Esta puede ser *pérdida localizada* (paros de pozos, disminución de producción, intervenciones, etc); o *pérdida no localizada* (desconocida), que debe justificarse lo mas pronto posible.

La *pérdida localizada* tiene el compromiso de que una vez solucionado el problema, la producción se recupera. La *pérdida no localizada* debe justificarse, pero como no se conoce, debe salir a buscarse (pozos sin producir o con menor producción no detectada, roturas de líneas no detectadas, etc.)[10]

5.1.5. Tratamiento de datos faltantes

5.1.5.1. Mecanismo de pérdida de datos

Para realizar el tratamiento de datos faltantes es importante conocer el mecanismo de pérdida de datos. Se distinguen tres casos: [11]

- **Datos perdidos completamente al azar (MCAR)**: La variable presenta valores perdidos completamente al azar, es decir, que su probabilidad de ausencia es la misma para todos los sujetos de la muestra.
- **Datos perdidos al azar (MAR)**: La variable presenta valores perdidos aleatoriamente, pero su probabilidad de ausencia no es la misma para todos los sujetos de la muesta, depende de los datos disponibles.
- **Datos no perdidos al azar (NMAR)**: La variable presenta valores perdidos no aleatoria, su probabilidad de ausencia no depende de los datos disponibles, depende, por ejemplo, del mismo valor ausente.

La mayoría de los casos de datos faltantes pertenecen al grupo MAR, de ahí que las técnicas de imputación toman esta hipótesis para realizar sus procedimientos.

A continuación se describen dos algoritmos de imputación de datos faltantes.

5.1.5.2. Algoritmo MICE

Este método es utilizado comúnmente debido a su robustez en el tratamiento de datos incompletos complejos.

Este algoritmo es un método basado en ecuaciones encadenadas, se implementa generando modelos de regresiones lineales para cada una de las variables, donde la variable dependiente es aquella que tiene datos faltantes a imputar y las variables que tengan sus datos completos serán las regresoras en cada caso. El algoritmo MICE utiliza el método de ajuste de media predictiva, en donde el propósito de la regresión lineal no es generar realmente valores imputados. Más bien, sirve para construir una métrica para unir casos con datos faltantes a casos similares con datos presentes.

Así, se reemplazan los valores faltantes por los valores obtenidos por el método del ajuste de la media predictiva. Este procedimiento se repite n veces, fijando un máximo de iteraciones, logrando varias versiones de datos imputados.

La evaluación de la convergencia de los valores imputados se pueden realizar de dos maneras:

1. Revisando las medias y las desviaciones de las distribuciones de los datos imputados durante cada ciclo. Si los trazos se mezclan bien, se puede decir que los valores imputados alcanzaron un estado estable.

2. Otra manera de evaluar la convergencia es generar gráficos de densidad de cadenas múltiples y comprobar si todos producen la misma distribución (aproximadamente).

5.1.5.3. Algoritmo Miss Forest

El algoritmo missForest[12] es un método de imputación múltiple de datos, su funcionamiento está basado en Random Forest, considerado un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.

Este algoritmo en primera instancia realiza una estimación de valores faltantes utilizando el método de imputación mediante media. Luego, entrena el conjunto de datos usando Random Forest con la parte observada y realiza la imputación de datos con los valores predichos por el algoritmo Random Forest. Este proceso se realiza de forma iterativa hasta que se hayan imputado todos los valores faltantes en el conjunto de datos o hasta un punto de parada establecido. Este criterio de paro, consiste en evaluar las diferencias entre el resultado de las imputaciones de la iteración previa, y el correspondiente a la iteración actual, deteniendo procesamiento tan pronto como estas diferencias se incrementen.

El algoritmo entrega como medida de la calidad de la imputación realizada el error estimado OOB (out-of-bag) obtenido, éste se presenta como dos estadísticos, el primero es el error cuadrático medio normalizado (NRMSE) para las variables continuas, el segundo es la proporción de valores clasificados erróneamente (PFC) para las variables categóricas.

5.1.6. Análisis Multivariado

El Análisis Multivariado es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjunto de datos multivariantes en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado [13]. El objetivo principal, al igual que el análisis univariado [3] es tener un mejor entendimiento del conjunto de datos en estudio.

[3] Consiste en el análisis de cada una de las variables estudiadas por separado.

5.1.6.1. Análisis de Correlación Lineal

El concepto de correlación se refiere al grado de variación conjunta existente entre dos o más variables. Para cuantificar el grado de correlación existente, el mejor y más utilizado coeficiente es el de Pearson (1896), se suele representar con r y se obtiene tipificando el promedio de los productos de las puntuaciones diferenciales de cada caso (desviaciones de la media) entre dos variables correlacionadas[14](ecuación 5.1)

$$r_{xy} = \frac{\sum x_i y_i}{n S_x S_y} \tag{5.1}$$

Donde:

- x_i, y_i =Puntuaciones diferenciales de cada par.
- n = Número de datos.
- S_x, S_y =Desviaciones típicas de cada variable.

El valor de este estadístico varía en el intervalo [-1,1], y se interpreta de la siguiente manera:

- Si $r = 1$,existe una correlación positiva perfecta. Si una de las variables aumenta, la otra lo hará en una proporción contante.
- Si $0 < r < 1$, existe una correlación positiva.
- Si $-1 < r < 0$, existe una correlación positiva.
- Si $r = -1$, existe una correlación negativa perfecta. Cuando una de ellas aumenta, la otra disminuye en proporción constante.

5.1.6.2. Análisis de Componentes Principales

El Análisis de Componentes principales es una técnica que consiste en encontrar combinaciones lineales de las variables originales para conseguir un nuevo conjunto de variables ortogonales, no correlacionadas entre sí, ordenadas decrecientemente de acuerdo a su varianza.[14]

El propósito fundamental de la técnica consiste en la reducción de la dimensión de los datos (menor número de variables) con el fin de simplificar el problema de estudio.

Formalmente[15], sea $\mathbf{X} = [X1, ..., Xp]$ una matriz con p variables, las componentes principales son nuevas variables compuestas no correlacionadas tales que un menor grupo explica la variabilidad de $\mathbf{X}$

Las componentes principales son las variables compuestas $Y1 = X_{t1}, Y2 = X_{t2}, ..., Y_p = X_{tp}$, tal que:

- $Var(Y_1)$ es máxima condicionado a $t_1' t_1 = 1$
- Entre todas las variables compuestas Y tales que $Cov(Y_1, Y_2) = 0$, la variable Y_2 es tal que $Var(Y_2)$ es máxima condicionado a $t_2' t_2 = 1$
- Si $p \geq 3$, la componente Y_3 es una variable no correlacionada con Y_1, Y_2 con varianza máxima.
- Análogamente se definen las demás componentes principales si $p \geq 3$.

Los coeficientes que acompañan a la componente principal calculada se denominan cargas o *loadings*. Representan la correlación entre las variables (filas) y variables (columnas).

El análisis de componentes principales construye una transformación lineal que elige un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño en el conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Para reducir la dimensionalidad de un grupo de datos, retiene aquellas características del conjunto de datos que contribuyen más a su varianza.

Existen varios criterios utilizados para seleccionar el número de componentes que pasarán a ser el nuevo conjunto de variables; a continuación se describen dos de ellos:

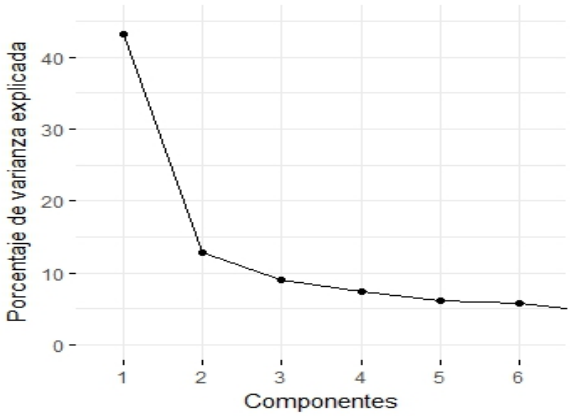
5.1.6.2.1. Criterio del bastón roto

La representación de la secuencia de valores propios de la matriz de covarianzas, ordenados de mayor a menor, recibe el nombre de gráfico de sedimentación o Scree Plot (figura 5.1.8). Una sugerencia gráfica nos provee este criterio indicando seleccionar los valores propios hasta que el descenso se estabiliza.[14]

5.1.6.2.2. Criterio de Kaiser

El criterio de Kaiser sugiere que obtener las componentes principales a partir de la matriz de correlaciones R equivale a suponer que las variables observables tengan varianza 1. Por lo tanto una componente principal con varianza inferior a 1 explica menos variabilidad que una variable observable. Por lo tanto, se seleccionarán las componentes cuyos autovalores sean mayores a 1.

Figura 5.1.8: Gráfico de sedimentación.

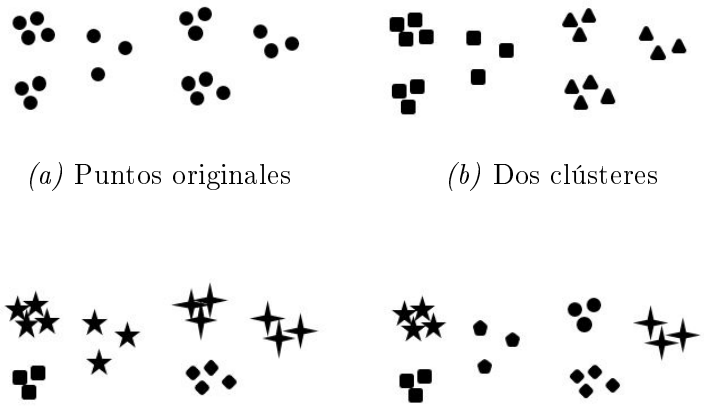


Fuente: Elaboración propia en base a datos de la batería en análisis.

5.1.6.3. Análisis de Conglomerados o Clústeres

Es una técnica estadística multivariada que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad entre los miembros de cada grupo y la mayor diferencia entre los grupos. Estos grupos deben capturar el estructura natural de los datos.

Figura 5.1.9: Diferentes formas de agrupar el mismo conjunto de puntos.



(a) Puntos originales *(b)* Dos clústeres

(c) Cuatro clústeres *(d)* Seis clústeres

Fuente: Cluster Analysis. Basic concepts and algorithms. (Introduction to Data Mining; Pang Tan).

Este análisis es un método basado en criterios geométricos y se utiliza como una técnica exploratoria, descriptiva, más no explicativa.

5.1.6.4. Conglomerados de k-medias

K-medias (MacQueen, 1967) es un algoritmo de aprendizaje no supervisado[4] muy utilizado al dividir un conjunto de datos dado en un conjunto de k grupos (es decir k clústeres), donde k representa la cantidad de grupos previamente especificado.[5]

El algoritmo funciona de la siguiente manera[19]:

- El algoritmo comienza seleccionando aleatoriamente k puntos del conjunto de datos que servirán como centros inicales para los clústeres. A estos puntos inciales se los conoce centroides.
- A continuación, cada uno de los puntos restantes se asigna a su centroide más cercano, donde más cercano está definido usando la distancia euclidiana [6] entre el punto y el centroide del clúster.
- Después del paso de asignación, el algoritmo calcula el nuevo valor medio de cada clúster. Una vez que los centroides han sido recalculados, se verifican nuevamente los puntos para saber si efectivamente pertenecen a ese grupo o se asigna a otro clúster.
- La asignación de clúster y los pasos de actualización del centroide se repiten iterativamente hasta que las asignaciones de clústeres haya llegado a una convergencia.

Una vez que se obtinene los clústeres, se puede combinar con el análisis de componentes principales (ACP), ya que mediante esta técnica se pueden homogeneizar los datos, lo cual permite posteriormente un análisis de clústeres sobre los componentes obtenidos[17][18].

5.2 Análisis predictivo

El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos[16].

A continuación se describen los algoritmos utilizados en este trabajo:

[4] Aprendizaje no supervisado es un método de aprendizaje automático donde un modelo es ajustado a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori.

[5] Una de las técnicas utilizadas para la determinación del k óptimo es Índice Silhouette[20].

[6] Distancia euclidiana o euclídea es la distancia .ºrdinaria"(que se mediría con una regla) entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.

5.2.1. Regresión lineal múltiple

En el modelo de regresión lineal múltiple[21], las variables independiente es una función lineal de k variables regresoras correspondientes a las variables explicativas (o a transformaciones de las mismas) y una perturbación aleatoria u error. El modelo también incluye un término independiente. La regresión lineal múltiple vendrá dado por la siguiente expresión (ecuación 5.2):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u \tag{5.2}$$

Donde:

- $\beta_1, \beta_2, ..., \beta_k$ son coeficientes de la regresión.
- u representa el error generado.

Para realizar un análisis de regresión lineal múltiple se hacen las siguientes consideraciones sobre los datos:

- **Linealidad.** La relación entre cada variable regresora con la variable dependiente debe ser lineal. Debe haber una correlación entre las variables regresoras y la dependiente.
- **Homocedasticidad.** La homocedasticidad es cuando la varianza de los errores de medición de nuestro análisis es igual para todas las variables independientes.
- **Independencia.** Los residuos (error) deben tener una distribución normal. La regresión es un análisis lineal y por ello, trabaja con relaciones lineales. Cuando los errores de las variables tienen distribución no normal, pueden afectar a los valores estimados.

  La independencia entre los residuos se mide utilizando el estadístico de *Durbin-Watson*. La hipótesis nula afirma que la autocorrelación de los residuos es cero. Si el valor del estadístico varia entre 1.5 y 2.5 se considera que existe independencia entre los residuos. Si el valor es menor a 2 indica autocorrelación positiva y si DW es mayor a 2 la autocorrelación es negativa.

5.2.2. Técnicas automáticas para selección de variables

A pesar de que en un modelo de regresión se pueden utilizar todas las variables se debe elegir un subconjunto de ellas que den un modelo adecuado. Existen métodos

automáticos para elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando una regla de parada se satisface.

Tres algoritmos muy usados para seleccionar variables son[22]:

- Backward Elimination. O eliminación para atrás. Se comienza con el modelo completo y en cada paso se va eliminando una variable. Si resultara que todas las variables predictoras son importantes, es decir, tienen "p-value"pequeños para la prueba T, entonces no se hace nada y se concluye que el mejor modelo es el que tiene todas las variables predictoras disponibles.

- Forward Selection. Se empieza con la regresión lineal simple que considera como variable predictora a aquella que está más altamente correlacionada con la variable de respuesta. Si esta primera variable no es significativa entonces se reconsidera este modelo y se para el proceso. Si hay variables que son significativas se añade al modelo.

- Stepwise Selection. Se puede considerar como una modificación del método "Forward". Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada.

5.2.3. Support Vector Regression

Support Vector Machine[7] también se puede utilizar como un método de regresión, manteniendo todas las características principales que caracterizan el algoritmo. Support Vector Regression (SVR) usa los mismos principios que SVM para la clasificación.

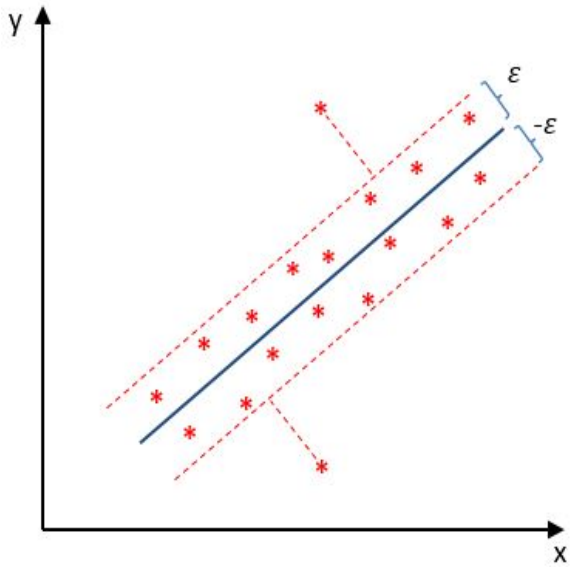
Cuando se tiene un conjunto linealmente separable de puntos de dos clases diferentes, el objetivo de SVM es encontrar un hiperplano que separe estas clases con un error mínimo, asegurándose también que el hiperplano tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo.

Con esta premisa, SVR busca este hiperplano, pero asegurándose de que la distancia entre estos puntos y el hiperplano no sea mayor a un valor establecido (epsilon), como muestra la figura 5.2.1. En otras palabras, es como dibujar a mano una línea en algún lugar de un conjunto de puntos, asegurándosee de estar lo más cerca posible de ellos.

Se sugiere ver documento de Alex J. Smola y Bernhard Scholkopf[24], para una aproximación más formal al tema.

[7]Support Vector Machine (SVM) es un clasificador discriminativo definido formalmente por un hiperplano de separación. En otras palabras, dados los datos de entrenamiento etiquetados (aprendizaje supervisado), el algoritmo genera un hiperplano óptimo que clasifica los nuevos ejemplos.[23]

Figura 5.2.1: Funcionamiento de Support Vector Regression.



Fuente: Tutorial on Support Vector Regression. NeuroCOLT Technical Report Series.

5.2.4. Redes Neuronales Artificiales

Una Red Neuronal Artificial (RNA) está constituida por una colección de elementos de procesamiento (nodos o neuronas) altamente interconectados que transforma un conjunto de datos de entrada en un conjunto de datos de salida deseado.

Es una técnica de estimación y clasificación perteneciente a la inteligencia artificial, que tiene como principio el proceso de aprendizaje que intenta simular la conducta cognitiva del cerebro humano[25].

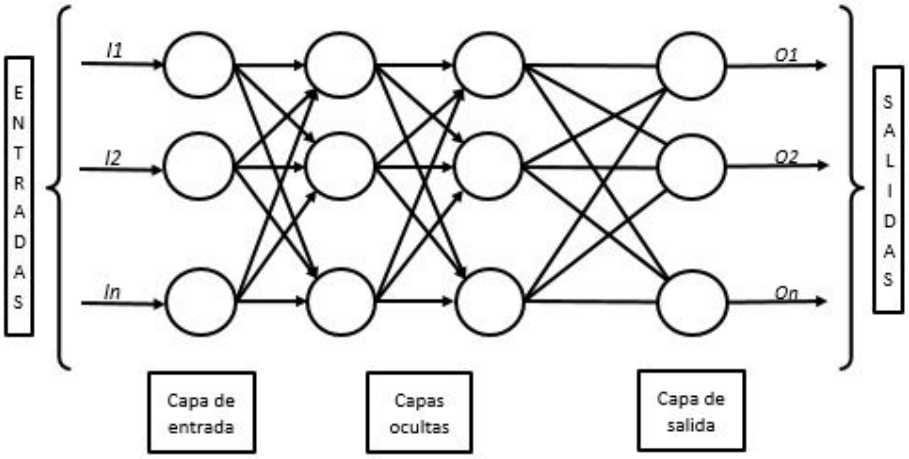
Una red neuronal artificial está constituida por capas de información, donde generalmente se puede distinguir una capa de entrada (variables independientes), una o varias capas intermedias u ocultas (que realizan la determinación de las relaciones entre las variables de entrada y salida) y una capa de salida que recibe el resultante de las variables independientes, como se muestra en la figura 5.2.2.

5.2.4.1. Red Backpropagation

El funcionamiento de la red backpropagation (BPN) consiste en el aprendizaje de un conjunto predefinido de pares de entradas-salidas dados como ejemplo:

- Primero se aplica un patrón de entrada como estímulo para la primera capa de las neuronas de la red.

Figura 5.2.2: Esquema general de una red neuronal artificial



Fuente: Algoritmo Backpropagation para Redes Neuronales: conceptos y aplicaciones. CIC-IPN.

- Se va propagando a través de todas las capas superiores hasta generar una salida.
- Se compara el resultado en las neuronas de salida con la salida que se desea obtener y se calcula un valor de error para cada neurona de salida.
- A continuación, estos errores se transmiten hacia atrás, partiendo de la capa de salida hacia todas las neuronas de la capa intermedia que contribuyan directamente a la salida.

Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido un error que describa su aportación relativa al error total. Basándose en el valor del error recibido, se reajustan los pesos de conexión de cada neurona, de manera que en la siguiente vez que se presente el mismo patrón, la salida esté más cercana a la deseada[26][27].

5.3 Métricas de evaluación utilizadas

5.3.1. Raíz del Error Cuadrático Medio

La Raíz del Error Cuadrático Medio o (RSME por sus siglas en ingles) es una medida de desempeño cuantitativa, la cual entrega la diferencia existente entre los valores pronosticados y los valores observados[28]. Será mejor cuanto menor sea su valor. Se expresa como(ecuación 5.3):

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(\hat{y} - y)^2} \tag{5.3}$$

Donde:

- $\hat{y}$ es el valor pronosticado.
- y es el valor observado
- n número de valores analizados.

5.3.2. Coseno cuadrado Cos^2

En el contexto de componentes principales Cos^2 indica la contribución de una componente a la distancia cuadrada de la observación al origen. Corresponde al cuadrado del coseno del ángulo del triángulo rectángulo conformado por el origen, la observación (en el espacio original), y su proyección sobre la componente, y se calcula como:

$$Cos^2_{i,l} = \frac{f^2_{i,l}}{\sum_l (f^2_{i,l})} = \frac{f^2_{i,l}}{\sum_l (d^2_{i,g})} \tag{5.4}$$

Donde:

- i es la observación.
- l es la componente.
- $d^2_{i,g}$ es la distancia al cuadrado de una observación dada al origen.

$d^2_{i,g}$ se calcula, en base al teorema de Pitágoras, como la suma de los valores al cuadrado de las puntuaciones de la observación en términos de cada una de las componentes $f^2_{i,g}$ (observación "valuada.^en las componentes). Por último, el hecho de utilizar coseno cuadrado, permite sumar las contribuciones con respecto a cada componente para obtener la ponderación de una observación en el (hiper)plano reducido por las componentes.[29]

5.3.3. Estadístico de Hopkins

El estadístico de Hopkins [18] se utiliza para evaluar la tendencia a la agrupación de un conjunto de datos al medir la probabilidad de que un conjunto de datos determinado

se genere mediante una distribución de datos uniforme. En otras palabras, prueba la aleatoriedad espacial de los datos. El valor del estadístico se calcula como (ecuación 5.5):

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{5.5}$$

Donde:

- $x_i = dist(p_i, p_j)$ es la distancia al vecino más cernano entre p_i, p_j, que son observaciones del conjunto de datos.
- $x_i = dist(q_i, q_j)$ es la distancia al vecino más cernano entre q_i, q_j. Aquí q_i, q_j son puntos de un nuevo conjunto de datos generado a partir del original. Este nuevo conjunto de datos se genera tomando datos aleatorios, tiene una distribución uniforme y la misma varainza.

De esta manera si el conjunto de datos original se distribuyera de manera uniforme, entonces $\sum_{i=1}^{n} y_i$ y $\sum_{i=1}^{n} x_i$ estarían cerca el uno del otro, y así H sería aproximadamente 0.5. Esto indicaría que la distribución de datos es uniforme.

Sin embargo, si existe una tendencia al agrupamiento en el conjunto de datos, entonces las distancias para los puntos artificiales $\sum_{i=1}^{n} y_i$ serían sustancialmente mayores que para los reales $\sum_{i)=1}^{n} x_i$. Este comportamiento se da debido a que los puntos artificiales están distribuidos homogéneamente mientras que los puntos reales están agrupados, y por ende el valor del estadístico será más grande.

5.3.4. Coeficiente de determinación corregido $R^2 ajustado$

El coeficiente de determinación corregido o $R^2 ajustado$ se define como la proporción de la varianza total de la variable explicada por la regresión. Reflej ala bondad del ajuste de un modelo a las variable que pretende explicar[30]. Su valor varia entre 0 y 1. Es mejor mientras mas cercano a 1 se encuentre. Se expresa como(ecuación 5.6):

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \tag{5.6}$$

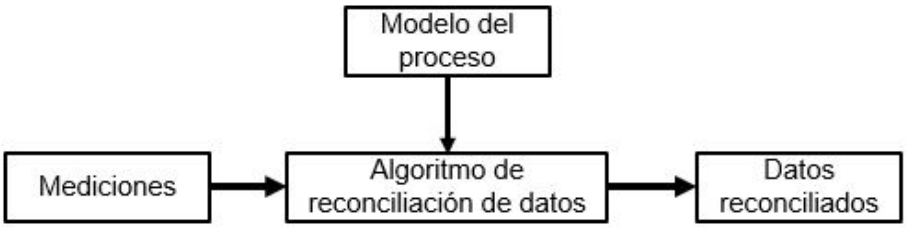
Donde:

- σ_{XY} es la covarianza de (X, Y).
- σ_X es la desviación típica de la variable X
- σ_Y es la desviación típica de la variable Y

5.4 Reconciliación de datos

La reconciliación consiste en ajustar las medidas redundantes de modo que obedezcan las leyes de conservación de la materia o cualquier otro tipo de restricción que incorpore el modelo matemático del sistema.[31]. Esto permite reducir el efecto de errores sistemáticos en las observaciones y la estimación de valores y variables no medidos. Los datos reconciliados son estadísticamente más confiables que los datos experimentales (no reconciliados).

Figura 5.4.1: Esquema del funcionamiento de reconciliación de datos



Fuente: Técnicas de reconciliación de datos: Aplicación a industria.

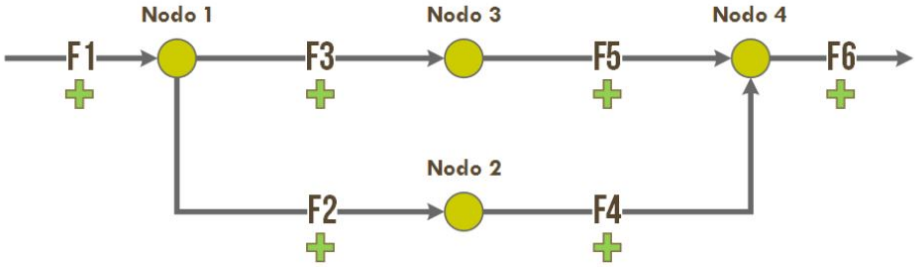
5.4.1. Reconciliación de datos en sistemas lineales y estacionarios

5.4.1.1. Reconciliación lineal con todos los flujos medidos.

La reconciliación de lineales y estacionarios corresponde al ajuste de datos experimentales sujeto a que se cumplan las restricciones impuestas por las ecuaciones de balance de masa. Los datos reconciliados deben parecerse tanto como sea posible a los datos medidos.

Un ejemplo de un esquema de reconciliación se muestra en la figura 5.4.2. Es un sistema con 4 nodos, correspondientes a 4 puntos donde se toman las mediciones, y 6 variables, todas medidas por un sensor. Las ubicaciones de los sensores estarán representadas por las aspas verdes.

Figura 5.4.2: Esquema del ejemplo



Fuente: Técnicas de reconciliación de datos: Aplicación a industria.

En el estado estable del proceso, los datos conciliados se obtienen de la siguiente manera (ecuación 5.7):

$$\hat{y} = y - V * A^T * (A * V * A^T)^{-}1 * A * y \tag{5.7}$$

:

Donde:

- $\hat{y}$ son las medidas reconciliadas.
- V es una matriz cuyas diagonales son las varianzas de las medidas.
- A la matriz de la incidencia.[8]

La matriz de incidencia es una matriz formada únicamente por ceros, unos y menos uno, cuyo número de columnas es el número de variables diferentes y el número de filas los nodos del sistema. Los nodos son los lugares donde se realizan las mediciones del sistema.

Tomando en consideración el esquema anterior la matriz de incidencia sería como se muestra a continuación:

Se sugiere ver documento de Shankar Narasimhan[33], para una aproximación más formal al tema.

[8]La matriz de incidencia que representa los balances másicos del sistema en estado estacionario y, en este caso, considera que todas las variables involucradas han sido medidas, o lo que es lo mismo, que todas sus filas son linealmente independientes.[32]

$$\begin{bmatrix} \uparrow & \leftarrow & N^{ro}variables & \rightarrow \\ N^{ro}nodos & \dots & \dots & \dots \\ \downarrow & \dots & \dots & \dots \end{bmatrix}$$

Figura 5.4.3: Matriz de incidencia

$$A = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \end{bmatrix}$$

Fuente: Técnicas de reconciliación de datos: Aplicación a industria.

Bibliografía

[1] Disponible en: https://www.significados.com/yacimiento/ Consultado: 14 de enero de 2018, 09:38 pm.

[2] Castro, Armando. *Aspectos de producción*. Instituto Mexicano del Petróleo. 2013

[3] Society of Petroleum Engineers UNP Student Chapter *Producción y transporte en la industria petrolera*. Universidad Nacional de Piura. 2012

[4] El abecé del Petróleo y del Gas *Capítulo 9: Producción*. Instituto Argentino del Petróleo y del Gas. 2013

[5] Bahena, Arcadio *Ajuste de correlaciones para estranguladores de pozos de gas y condensado del activo Muspac*. México. 2008

[6] Gilbert, W.E. *Flowing and Gas-lift Well Performance*. Drilling and production practice, 1954.

[7] Ros, N. C. J. *An Analysis of Critical Simultaneous Gas/liquid Flow through a Restriction and its Application to Flow Metering*. Applied Science Research, 1960.

[8] Baxendall, P. B. *Bean Performance–Lake Wells*. Internal Report (October), 1957.

[9] Achong, I. *Revised Bean Performance Formula for Lake Maracaibo Wells*. Shell Internal Report (October), 1961.

[10] Sánchez, M *Curso de producción II*. Disponible en: https://www.scribd.com/presentation/288345013/Tema-3-Control-de-Produccion-de-Petroleo/ Consultado: 15 de enero de 2018, 06:38 pm.

[11] Gómez, A. *Comparativa de análisis de imputación de datos faltantes con análisis de casos completos en pruebas diagnósticas*. Julio, 2017.

[12] Stekhoven, D *Using the missForest Package*. Mayo, 2011.

[13] Diaz, M. , Rojas, A. *Análisis multivariado*. Disponible en: https://es.slideshare.net/federicodonneysg/analisis-multivariado-8829215/ Consultado: 15 de enero de 2018, 11:42 pm.

[14] Chan, D. , Badano, C. , Gambini, J. *Análisis Inteligente de datos con lenguaje R*. Curso Introductorio Buenos Aires, Argentina, 2017

[15] Hastie,T. , Tibshirani, R., y Friedman, J. *Unsupervised learning. In The elements of statistical learning.* Springer, 2009.

[16] Nyce, C. *Predictive Analytics White Paper.* American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America. 2007

[17] Fernandez, S. *Análisis de conglomerados.* Facultad de ciencias económicas y empresariales Madid 2011

[18] Pang-Ning Tan, Michael Steinbach, Vipin Kumar *Introduction to Data Mining. Chapter 8. Cluster Analysis: Basic Concepts and Algorithms.* Edición 2, Marzo 2006

[19] Alboukadel Kassambara *Practical Guide To Cluster Analysis in R.* Disponible en: http://www.sthda.com/english/web/5-bookadvisor/17-practical-guide-to-cluster-analysis-in-r/ Consultado: 20 de enero de 2018, 11:39 pm.

[20] Peter J. Rousseeuw *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics.* Disponible en: http://dx.doi.org/10.1016/0377-0427(87)90125-7 Consultado: 20 de enero de 2018, 11:31 pm.

[21] Douglas C. Montgomery *Introduction to Linear Regression Analysis.* Quinta edición.

[22] Douglas C. Montgomery *Regresión Lineal Múltiple.* Apuntes de la materia Enfoque estadístico. 2016

[23] Patel, Savan *SVM (Support Vector Machine) Theory.* Disponible en: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 Consultado: 21 de enero de 2018, 01:31 pm.

[24] Alex J. Smola y Bernhard Scholkopf *A Tutorial on Support Vector Regression.* Disponible en: http://www.svms.org/regression/SmSc98.pdf Consultado: 21 de enero de 2018, 01:31 pm.

[25] Mena,Carlos , Guajardo, Rodrigo *Comparison of neural networks and linear regression to estimate site productivity in forest plantations, using geomatic.* Disponible en: https://scielo.conicyt.cl/scielo.php Consultado: 21 de enero de 2018,08:24 pm.

[26] José R. Hilera y Victor J. Martinez *Redes Neuronales Artificiales.* Alfaomega-Rama, 2000

[27] E. Castillo Ron, Á. Cobo Ortega, J. M. Gutiérrez Llorente, R. E. Pruneda González *Introducción a la Redes Funcionales con Aplicaciones.* Paraninfo, 1999

[28] *Metodología para la evaluación del modelo de pronóstico metereológico.* Disponible en: http://www.tdx.cat/bitstream/handle/10803/6836/11Ojc11de12.pdf Consultado: 23 de enero de 2018,12:24 pm.

[29] Abdi, H. *Principal Component Analysis. Vol 2. John Wiley & Sons, Inc.*. 2010

[30] *Coeficiente de determinación corregido.* Disponible en: http://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html Consultado: 23 de enero de 2018,12:24 pm.

[31] Miguel A. Lozano y Jesus A. Remerio. *Clasificación de variables y reconciliación de datos en ingeniería de procesos.* España,2002

[32] Marcos González Fernández *Técnicas de reconciliación de datos: Aplicación a industrias.* Sevilla, 2016

[33] Shankar Narasimhan *Data Reconciliation and Gross Error Detection An Intelligent Use of Process Data* . páginas 113-114 Houston, Texas. 2000