

INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE POSTGRADO

TÉCNICAS DE APRENDIZAJE PROFUNDO PARA EL PRONÓSTICO DE LA PRODUCCIÓN DE PETROLEO

Aplicación a pozos perforados en yacimientos no-convencionales

AUTOR: Ortega Arrieta, Alexis Enrique

TUTOR: Soliani, Valeria Inés

**TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN
CIENCIA DE DATOS**

BUENOS AIRES

Diciembre, 2021

Sinopsis

Dada la creciente importancia de los yacimientos no-convencionales en el suministro de energía mundial, y particularmente en Argentina, diversas técnicas analíticas y numéricas para el pronóstico de la producción de petróleo y gas de dichos yacimientos se han desarrollado durante las últimas décadas. Sin embargo, debido a las incertidumbres y la complejidad asociadas con los modelos de reservorios no-convencionales, se han propuesto modelos alternativos de pronósticos de producción, incluyendo aquellos basados en el análisis de datos (i.e., data-driven, deep-learning models), tales como redes neuronales convolucionales, redes neuronales recurrentes y, más recientemente, arquitecturas tipo “Codificador/Decodificador” (“Encoder/Decoder”) con “mecanismos de atención”.

El presente trabajo busca desarrollar una metodología basada en datos para pronosticar la producción mensual promedio de petróleo de pozos horizontales perforados y completados en el yacimiento no-convencional Vaca Muerta en Argentina. Se busca también modelar el efecto de varios parámetros de yacimiento y completamiento en la producción de pozos de petróleo y gas mediante la aplicación de una metodología probabilística. Para lo anterior, se propone el uso de una arquitectura novedosa denominada “Temporal Fusion Transformers” que, mediante el análisis de la compleja interacción entre covariables estáticas y covariables temporales, permite pronosticar múltiples pasos de tiempo futuros.

Usando un set de datos público provisto por la Secretaría de Gobierno de Energía, conteniendo alrededor de 65,000 registros de producción de más de 1,400 pozos en la cuenca, se ha implementado un modelo Temporal Fusion Transformers para la predicción de 6 meses de producción en el futuro de 275 pozos horizontales de petróleo en la formación no-convencional Vaca Muerta. El modelo muestra un buen desempeño cuando se compara contra un modelo base. Adicionalmente, haciendo uso de las características de interpretabilidad del modelo, se ha identificado que variables de completamiento como el volumen de fluido y apuntalante inyectado, la longitud horizontal y el número de etapas de fractura tienen la mayor importancia en la predicción de la producción de petróleo.

Tabla de Contenido

1.	Introducción	5
2.	Marco Conceptual y Antecedentes.....	5
3.	Definición del Problema	9
4.	Justificación del Estudio	9
5.	Alcances del trabajo y limitaciones.....	10
6.	Hipótesis.....	11
7.	Objetivos	12
7.1.	Objetivo General	12
7.2.	Objetivos Específicos.....	12
8.	Metodología	12
8.1.	Técnicas.....	12
8.1.1.	Temporal Fusion Transformers (TFT).....	13
8.2.	Herramientas.....	15
8.2.1.	PyTorch Forecasting	15
9.	Resultados	17
9.1.	Análisis Exploratorio de Datos.....	17
9.1.1.	Filtrado de registros	21
9.1.2.	Selección de variables	23
9.1.3.	Manejo de datos faltantes (NA's)	24
9.1.4.	Adición de “Índice de Tiempo” (time_idx)	24
9.1.5.	Ingeniería de covariables (Feature Engineering).....	24
9.2.	Análisis de series de tiempo con PyTorch Forecasting	25
9.2.1.	Creación de conjuntos de datos de entrenamiento y validación.....	25
9.2.2.	Configuración del método TimeSeriesDataSet	27
9.2.3.	Creación de Dataloaders	28
9.2.4.	Identificación de la tasa de aprendizaje óptima	29
9.2.5.	Ajuste de hiperparámetros	29
9.2.6.	Creación de instancia del modelo TFT	30
9.2.7.	Entrenamiento del modelo TFT.....	31
9.2.8.	Evaluación del modelo TFT.....	32
9.2.9.	Revisión de predicciones pobres (alto error)	34
9.2.10.	Predicción versus valores futuros por variable	35

9.2.11.	Interpretación del modelo entrenado	37
9.2.12.	Predicción de valores futuros.....	39
10.	Discusión	42
11.	Conclusiones.....	44
12.	Referencias Bibliográficas	45

1. Introducción

Modelar la producción de petróleo y gas en pozos no-convencionales requiere de poder capturar no sólo su heterogeneidad geológica, sino también las características de los diferentes regímenes de flujo que se derivan de la implementación de diversas técnicas de perforación y completamiento, comunes a estos yacimientos. Como resultado de lo anterior, existe alta incertidumbre en el pronóstico de la productividad de pozos no-convencionales, en especial si la evaluación hace uso de modelos “convencionales” (i.e., modelos desarrollados para uso en yacimientos convencionales).

El presente trabajo, partiendo del estudio de las técnicas actualmente utilizadas para la evaluación del desempeño de producción, propone una técnica alternativa aplicable, en particular, a formaciones no-convencionales. El estudio se enfoca en el uso de modelos basados en el análisis de datos, tales como aprendizaje automático (Machine Learning) y/o aprendizaje profundo (Deep Learning), incluyendo herramientas como el análisis de series de tiempo y las redes neuronales recurrentes. El objetivo es modelar el efecto de varios parámetros de yacimiento y completamiento en la producción de pozos de petróleo y gas mediante el desarrollo de una metodología probabilística.

A manera de validación, la técnica propuesta se aplicará en este estudio a pozos perforados en la formación Vaca Muerta, ubicada en la cuenca Neuquina en Argentina. Para la construcción del modelo se utilizarán datos públicos de producción mensual de petróleo y gas. El modelo obtenido permitirá verificar la relación entre diferentes variables de yacimiento y completamiento de pozos y la productividad de éstos, haciendo posible la optimización técnica y económica de los proyectos a desarrollar en Vaca Muerta.

2. Marco Conceptual y Antecedentes

Los recursos no-convencionales juegan un papel cada vez más importante en el suministro de energía mundial debido a que las reservas convencionales han comenzado a declinar (Zhou, W. et al., 2013). Sin embargo, dada su mayor complejidad geológica y estructural, asociada a una mayor heterogeneidad de propiedades, el desarrollo de este tipo de recursos requiere de un enfoque diferente al típicamente utilizado.

Resultado de la heterogeneidad intrínseca de los yacimientos no-convencionales, pozos perforados en estas formaciones muestran una gran variabilidad en su desempeño (definido como productividad de pozo), incluso aunque se perforen en una misma capa productiva y/o se completen con prácticas similares de terminación. Es importante entonces desarrollar técnicas específicas para la evaluación de pozos no-convencionales, que consideren aspectos propios de estas formaciones.

Además de lo anterior, un modelo predictivo de producción de yacimientos no-convencionales debe incorporar características probabilísticas (es decir, no determinísticas) a fin de indicar incertidumbre en la respuesta del sistema (en este caso, incertidumbre en el perfil de producción de petróleo y gas), asociada con el uso de modelos simplificados y/o la carencia de precisión en la estimación de las variables requeridas para describir el sistema.

La alta variabilidad en la distribución espacial de las propiedades de los yacimientos no-convencionales (es decir, heterogeneidad de propiedades), se debe primordialmente a los muchos procesos geológicos complejos (por ejemplo, sedimentación, erosión, migración, compactación y diagénesis) a los cuales están sometidos estas formaciones (Ballin, Journel & Aziz, 1992). Algunas de las propiedades afectadas son: tipo de roca, porosidad, permeabilidad, ocurrencia de fallas, grado de cementación y saturación de hidrocarburos. No es posible predecir con exactitud las condiciones exactas que resultan de estos complejos procesos geológicos.

La incertidumbre en la descripción geológica de los yacimientos no-convencionales afecta significativamente la confiabilidad de la predicción de los parámetros de desempeño de dichos yacimientos (por ejemplo, tasa de producción y volumen recuperado de petróleo, corte de agua, relación de gas y petróleo, etc.). Dicha incertidumbre geológica comúnmente se especifica mediante la generación de múltiples descripciones de yacimiento igualmente probables, haciendo uso de conocidas técnicas geoestadísticas como es, por ejemplo, la simulación estocástica. Otros tipos de incertidumbre a incluir en el pronóstico final son: la incertidumbre de modelo y la incertidumbre intrínseca (es decir, errores en la medición) en los datos obtenidos (Hong, Bratvold, Lake & Ruiz Maraggi, 2019).

El pronóstico confiable del volumen de recobro estimado (Estimated Ultimate Recovery, EUR), determinado a partir de la simulación del volumen de yacimiento y del pronóstico de producción, es de gran importancia e interés en la gestión y operación de los activos de gas y petróleo no-convencionales. Particularmente, el diseño y la optimización de varios aspectos de la construcción de un pozo, tales como la perforación, el completamiento y la estimulación, se

basan en los resultados de las técnicas de predicción de la producción. Por otra parte, durante los últimos años se ha observado un incremento en la productividad de pozos no-convencionales asociado a mejoras en las técnicas de perforación y completamiento. Un modelo de producción, ajustado a este tipo de formaciones, debe ayudar a determinar el efecto de dichos parámetros en el desempeño productivo de los pozos.

Un enfoque tradicional es pronosticar la producción a través de estudios de simulación numérica. Sin embargo, debido a las incertidumbres y la complejidad asociadas con el modelo de reservorio, flujos de trabajo analíticos como son el análisis de balance de materiales (Material Balance Analysis, MBA) y el análisis de tasa transitoria (Rate Transient Analysis, RTA), además de flujos de trabajo centrados en datos de producción, como el análisis de curvas de declinación (Decline Curve Analysis, DCA), se utilizan ampliamente. Cada uno de esos métodos tienen sus propias fortalezas y limitaciones inherentes, y requieren de desarrollos adicionales para su aplicación a formaciones no-convencionales.

Aunque las técnicas numéricas para predecir producción de hidrocarburos se han desarrollado muy rápidamente durante las últimas décadas (Hong et al., 2019), el análisis de las curvas de declinación (DCA) permanece como el método más aceptado y ampliamente utilizado en la industria del petróleo y gas, tanto para los campos convencionales como los no-convencionales. Su amplio uso se debe principalmente a su eficiencia computacional, ya que sólo requieren del uso de datos de producción para determinar algunos parámetros de ajuste mediante el mapeo del historial de producción. El modelo ajustado se utiliza para pronosticar la producción futura y las reservas de hidrocarburos.

A pesar de su extenso uso, los modelos tradicionales, basados en el macheo del historial de producción, pueden no ser capaces de capturar las características de todos los regímenes de flujo presentes durante la producción un pozo no-convencional. Por ejemplo, para yacimientos no-convencionales, en el cual la producción está siempre en estado “transitorio”, los métodos tradicionales de balance de materiales no tienen aplicación. Por otra parte, dado que los pozos no-convencionales usualmente se completan con fracturamiento hidráulico, de modo que pueden aparecer varios regímenes de flujo durante la vida útil de un pozo, modelos de análisis de declinación (por ejemplo, Arps) podrían no ser ideales (Hong et al., 2019). Como resultado de los factores mencionados, el modelo con el mejor macheo puede no ser un buen predictor del desempeño futuro.

Además de lo anterior, estos modelos tradicionales de pronóstico son determinísticos, resultando por lo tanto inadecuados para la cuantificación de la incertidumbre, lo cual puede conllevar a análisis económicos errados o a una gestión incorrecta de los activos productores. Una forma de corregir esta presunta deficiencia es mediante el uso de probabilidad para la integración de incertidumbre en los modelos de declinación de producción (por ejemplo, DCA probabilístico).

Se han propuesto modelos alternativos de pronósticos de producción para superar las carencias de los modelos tradicionales. Por ejemplo, existen algunas referencias de aplicación de modelos probabilísticos de curvas de declinación (DCA) para el pronóstico de producción de pozos no convencionales (Hong et al., 2019), aunque su uso en la industria no se ha extendido ampliamente. Este tipo de enfoque se basa en el concepto de “agregación de pronósticos” para incorporar incertidumbre a la predicción. En un enfoque similar, Rotondi et al. (2006), han propuesto el uso de algoritmos de “vecindad” (“neighborhood algorithm”) que realiza un muestreo estocástico para explorar el espacio de parámetros, a fin de encontrar un conjunto aceptable de modelos y extraer información de este conjunto dentro de un marco bayesiano.

Más recientemente, diversas técnicas basadas en análisis de series de tiempo se han propuesto para predecir el desempeño de los pozos de hidrocarburos y cuantificar la incertidumbre de dicho pronóstico (Olominu & Sulaimon, 2014). Joshi, Awoleke y Mohabbat (2018), por ejemplo, presentaron una metodología aplicando Logistic Growth Analysis (LGA) y modelos de series de tiempo. Sagheer y Kotb (2019), por otro lado, han propuesto el uso de una arquitectura de redes neuronales recurrentes (LSTM’s) para el pronóstico de la producción de petróleo. Otros autores han desarrollado modelos híbridos que combinan varios tipos de redes neuronales (por ejemplo, redes neuronales convolucionales y redes neuronales recurrentes) (Abdullayeva & Imamverdiyev, 2019), para superar las limitaciones de cada uno de estos modelos individuales.

Para el caso particular de predicciones futuras “multi-horizonte”, es decir, la predicción de variables de interés en múltiples pasos de tiempo futuros, se han desarrollado arquitecturas novedosas de redes neuronales que incorporan los llamado “mecanismos de atención”, entre las que se destacan: DA-RNN: Dual-Stage Attention-Based Recurrent Neural Network (Qin, Y., et al., 2017), y DSTP-RNN: Dual-Stage Two-Phase Attention-Based Recurrent Neural Networks (Liu, Y., Gong, C., Yang, L., & Chen, Y., 2020). Estas arquitecturas han sido específicamente diseñadas para el análisis de series de tiempo multivariable y a largo plazo. Más recientemente, arquitecturas tipo “Transformers”, tales como Temporal Fusion Transformer (TFT), (Lim, B., Arık,

S. Ö., Loeff, N., & Pfister, T., 2021), se han introducido para “aprender” de manera eficiente dependencias a largo plazo en un conjunto de datos con series de tiempo multivariantes. Al momento de escribir este reporte, no se han encontrado publicaciones que muestren la aplicación de este tipo de metodologías para el pronóstico de producción de pozos de petróleo, y particularmente para la formación Vaca Muerta.

3. Definición del Problema

Los modelos analíticos y numéricos que en la industria petrolera se utilizan para evaluar el desempeño productivo de pozos perforados en reservorios no convencionales (i.e. formaciones productoras de hidrocarburos de lutita) no permiten capturar la variabilidad geológica ni la heterogeneidad de propiedades características de este tipo de formaciones. Sumado a esto, la aplicación de dichos modelos convencionales se dificulta por el fracturamiento hidráulico masivo al que, para viabilizar su producción comercial, se ven sometidos estos yacimientos de muy baja permeabilidad.

Como resultado de lo anterior, los pronósticos de producción que se obtienen al aplicar dichos modelos convencionales no cuentan con la precisión y confiabilidad requeridas para su uso en análisis técnicos y/o proyecciones económicas futuras, por lo que es necesario buscar enfoques alternativos de predicción de desempeño.

4. Justificación del Estudio

Para la optimización técnica y económica de sus proyectos de desarrollo, es fundamental que las empresas operadoras de petróleo y gas dispongan de métodos y técnicas que les permitan evaluar la productividad de pozos no convencionales, como es el caso de los pozos perforados en la formación Vaca Muerta en Argentina.

Por ejemplo, la optimización de los parámetros de diseño relacionados con la perforación y terminación de pozos no convencionales, incluyendo la longitud del ramal horizontal del pozo, el número de etapas de fracturamiento hidráulico y la cantidad de fluido y apuntalante por etapa de fractura, entre otros, requiere de un análisis adecuado del comportamiento productivo a largo plazo de los pozos.

Considerando lo anterior, la empresa operadora Wintershall Dea, con su unidad de negocios en Argentina, busca contar con un modelo para predecir la producción de sus pozos no convencionales en la formación Vaca Muerta. El desarrollo de dicho modelo permitirá a Wintershall Dea ajustar los planes y programas de perforación y terminación durante las etapas de predesarrollo y desarrollo de sus áreas operadas.

5. Alcances del trabajo y limitaciones

El presente estudio evaluará la producción de pozos no-convencionales perforados desde el año 2015 en la formación Vaca Muerta, ubicada en la Cuenca Neuquina (Argentina). Dado que la tecnología de terminación recomendada para este tipo de formaciones es de “pozos horizontales estimulados con fracturamiento hidráulico”, se limitará el análisis a pozos horizontales con una rama lateral mayor de 800 m. Además, debido a que los activos de Wintershall Dea en Argentina están ubicados en la ventana de petróleo, se considerará para el análisis sólo este tipo de fluidos (Black Oil y Volatile Oil).

Para el análisis de los métodos propuestos se utilizan datos públicos de producción mensual de petróleo, agua y gas, para cada pozo perforado en Vaca Muerta. Se incluyen también datos del completamiento de los pozos, tales como longitud de la rama horizontal, cantidad de etapas de fractura realizadas, y volúmenes de fluido y de arena de fractura. El estudio se enfoca sólo en la predicción de la producción futura (6 meses) de petróleo. Es decir, no se pronosticará la producción de otros fluidos como gas y agua.

A continuación, se resumen las principales limitaciones del proyecto:

- Poca historia de producción de los pozos en Vaca Muerta. La mayoría de los pozos horizontales han sido perforados durante los últimos 5 años, por lo que la cantidad de datos disponibles por pozo es reducida (e.g., número de meses en producción). Lo anterior puede limitar la aplicación de ciertas técnicas de análisis que se basan en datos (e.g., uso de técnicas de aprendizaje profundo para análisis de series de tiempo).
- No disponibilidad de datos clave de yacimiento y completamiento en base de datos de producción. La base de datos pública no contiene datos que pueden ser importantes al momento de predecir la producción futura de los pozos; por ejemplo, no se dispone de datos de presión de fondo y/o superficie, tampoco de los tamaños de las válvulas de orificio por la que los pozos han producido. Así mismo, no hay información de la cantidad

de grupos de perforaciones por etapa de fractura hidráulica en cada pozo. Este parámetro puede tener un impacto importante en la productividad final de los pozos fracturados. Por otra parte, la base de datos no especifica la zona específica, dentro de la formación Vaca Muerta, donde se han aterrizado los pozos horizontales.

6. Hipótesis

Para pozos horizontales de petróleo, perforados en la formación no-convencional Vaca Muerta, es posible pronosticar la producción futura de petróleo a partir del análisis de datos generales de pozo (e.g., área y operador), datos históricos de producción (e.g., producción de agua, gas y petróleo) y datos de perforación y completamiento de pozo (e.g., fecha/campaña de perforación/completamiento, longitud horizontal, número de etapas de fractura, volumen de fluido de estimulación, volumen de agente de sostén).

Se definen a continuación las principales variables incluidas en la hipótesis:

- Producción futura de petróleo: Variable dependiente y cuantitativa definida como el volumen diario de petróleo producido en un período de hasta 6 meses posteriores a la fecha del último registro de producción.
- Área: Variable dependiente cualitativa que indica la extensión de tierra utilizada con el propósito de extraer petróleo, como gas natural o petróleo crudo, del suelo.
- Operador: Variable dependiente cualitativa que indica la compañía a cargo de planear y ejecutar las actividades requeridas para desarrollar un campo de gas o petróleo.
- Producción histórica de agua: Variable dependiente cuantitativa definida como el volumen diario de agua producida desde el completamiento del pozo.
- Producción histórica de gas: Variable dependiente cuantitativa definida como el volumen diario de gas producido desde el completamiento del pozo.
- Producción histórica de petróleo: Variable dependiente cuantitativa definida como el volumen diario de petróleo producido desde el completamiento del pozo.
- Longitud horizontal: Variable dependiente cuantitativa que indica la extensión de la rama horizontal del pozo.
- Número de etapas de fractura: Variable dependiente cuantitativa que indica la cantidad total de etapas de fracturamiento hidráulico realizadas en el pozo.

- Volumen de fluido de estimulación: Variable dependiente cuantitativa que indica la cantidad de fluido inyectado por pozo durante la operación de fracturamiento hidráulico.
- Volumen de agente de sostén: Variable dependiente cuantitativa que indica la cantidad de arena inyectada a la formación durante la operación de fracturamiento hidráulico.

7. Objetivos

7.1. Objetivo General

Desarrollar una metodología basada en datos para pronosticar la producción mensual promedio de petróleo de pozos horizontales perforados y completados en el yacimiento no-convencional Vaca Muerta en Argentina.

7.2. Objetivos Específicos

- Evaluar la aplicabilidad de técnicas de aprendizaje profundo (Deep Learning) para el análisis y predicción de series temporales de producción de hidrocarburos en yacimientos no-convencionales.
- Incorporar métodos para estimación de incertidumbre en la predicción de producción futura de petróleo.
- Examinar la importancia de diversos parámetros de yacimiento, producción y completamiento en el desempeño de los pozos.

8. Metodología

8.1. Técnicas

El presente estudio hace uso de técnicas para *análisis de series de tiempo*, particularmente algoritmos de aprendizaje profundo (Deep Learning), incluyendo *Redes Neuronales Recurrentes (Recurrent Neural Networks-RNN)*, con capas de tipo *Long-Short Term Memory (LSTM)*. Estas redes son adecuadas para analizar datos secuenciales, como son los datos de producción de petróleo, porque contienen un componente de memoria que permite conservar información histórica al momento de hacer predicciones.

Sumado a esto, se incorpora el uso de “*mecanismos de atención*”, particularmente en modelos secuencia-a-secuencia (“*seq2seq*”), para manejar y cuantificar interdependencia entre variables. Dicha interdependencia puede ser entre los elementos de entrada (i.e., auto-atención) o entre los elementos de entrada y salida (i.e., atención general).

Adicionalmente, a fin de combinar la predicción a largo plazo de múltiples series de tiempo con conocimientos interpretables sobre la dinámica temporal, se introduce el uso de una arquitectura novedosa denominada “*Temporal Fusion Transformers*”.

8.1.1. Temporal Fusion Transformers (TFT)

El artículo publicado por Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021) describe la aplicación de Temporal Fusion Transformers para pronósticos interpretables de series de tiempo de múltiples horizontes. Dicha arquitectura, desarrollada por la Universidad de Oxford y Google, ha superado a DeepAR de Amazon en un 36% a un 69% en pruebas de referencia (Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T., 2021).

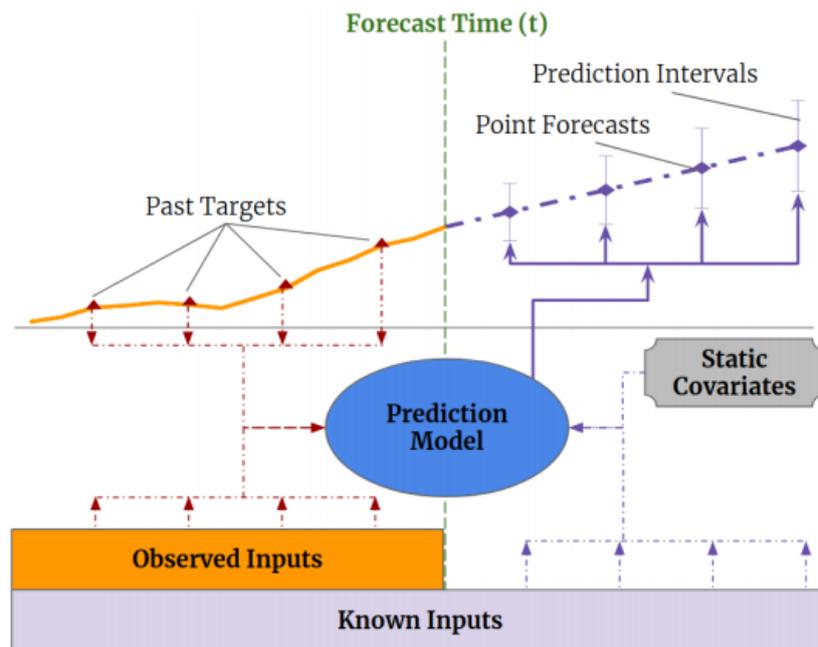


Figura 1. Ilustración del pronóstico multi-horizonte con uso de variables estáticas, variables “observadas” (conocidas) en el pasado y variables temporales futuras previamente conocidas. Tomado de Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021).

Con el fin de aprender las relaciones temporales a diferentes escalas, TFT utiliza capas (“layers”) recurrentes para el procesamiento local y capas de auto-atención interpretables para las dependencias a largo plazo. TFT utiliza componentes especializados para seleccionar características relevantes y una serie de capas de compuerta (“gates”) para suprimir componentes innecesarios, lo que mejora el rendimiento general del algoritmo.

Para mejorar el desempeño e interpretabilidad, TFT incorpora específicamente los siguientes componentes (Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T., 2021): (1) codificadores de covariables estáticos que codifican vectores de contexto para su uso en otras partes de la red, (2) mecanismos de activación y selección de variables para minimizar las contribuciones de entradas irrelevantes, (3) una capa de secuencia-a-secuencia para procesar localmente entradas conocidas y observadas, y (4) un decodificador de auto-atención temporal para aprender cualquier dependencia a largo plazo presente dentro del conjunto de datos. Un diagrama general de la arquitectura del modelo TFT se presenta en la **Figura 2**. Información más detallada sobre las características y consideraciones de aplicación del modelo TFT se pueden consultar en la referencia Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021).

8.2. Herramientas

El presente trabajo se ha realizado en un computador personal con sistema operativo Windows:

- 16 GB de RAM
- Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz
- NVIDIA GeForce GTX 1660 Ti GPU

Además, se ha configurado un ambiente Python (versión 3.7.9) con las siguientes librerías:

- PyTorch Forecasting, implementación general del modelo TFT.
- PyTorch Lightning, automatización del entrenamiento en CPUs y GPUs.
- Torch, Pandas, Numpy, Matplotlib, librerías base de Python.

8.2.1. PyTorch Forecasting

PyTorch Forecasting es una librería de última generación que utiliza redes neuronales para el pronóstico de series de tiempo. Se listan a continuación algunas características básicas de esta librería (adaptado de la documentación oficial de la librería PyTorch Forecasting):

- Abstracción del manejo de transformaciones de variables, valores faltantes, submuestreo aleatorio, múltiples longitudes (ventanas) de historia, etc., mediante la implementación de una clase “timeseries dataset”.
- Entrenamiento básico de modelos de series de tiempo y visualizaciones genéricas (e.g., visualización de valores reales versus predicciones, diagramas de dependencia) mediante la implementación de una clase “base model”.
- Disponibilidad de varias arquitecturas de redes neuronales optimizadas para el pronóstico de series de tiempo (e.g., DeepAR, NBeats, TemporalFusionTransformer) y con capacidades de interpretación integradas.
- Implementación de métricas para evaluación de series temporales de múltiples horizontes.
- Ajuste de hiperparámetros mediante integración con la librería Optuna.
- Entrenamiento en CPU y GPU mediante integración con librería PyTorch Lightning.
- Registro del progreso del entrenamiento mediante integración con librería Tensorboard.

El proceso general para entrenar y probar un modelo en PyTorch Forecasting es:

1. Crear un conjunto de datos de entrenamiento con el método *TimeSeriesDataSet*.
2. Crear un conjunto de datos de validación (a partir del conjunto de datos de entrenamiento) usando el método *from_dataset*.
3. Crear una instancia del modelo usando el método *from_dataset*.
4. Crear un objeto *pytorch_lightning.Trainer*.
5. Encontrar la tasa de aprendizaje (“*learning rate*”) óptima con el método *tuner.lr_find*.
6. Entrenar el modelo con parada temprana (“*early stopping*”), usando el conjunto de datos de entrenamiento, y utilizar los registros de Tensorboard para verificar si ha convergido con una precisión aceptable.
7. Ajustar los hiperparámetros del modelo usando *Optuna*.
8. Entrenar el modelo con el mismo programa de tasa de aprendizaje en todo el set de datos.
9. Cargar el mejor modelo guardado (“*checkpoint*”) y aplicarlo a nuevos datos.

El flujo de trabajo mostrado arriba se ha aplicado en el presente trabajo para la predicción de la producción de petróleo en pozos perforados en la formación Vaca Muerta.

9. Resultados

9.1. Análisis Exploratorio de Datos

En este trabajo se utiliza un set público de datos proporcionado por la “*Secretaría de Gobierno de Energía*” en Argentina (https://datos.gob.ar/dataset/energia-produccion-petroleo-gas-por-pozo-capitulo-iv/archivo/energia_cb5c0f04-7835-45cd-b982-3e25ca7d7751). Los datos contienen volúmenes mensuales de producción de petróleo, gas y agua para los pozos perforados en la formación no convencional Vaca Muerta (alrededor de 65.000 registros históricos de producción). Además, se incluyen datos sobre la técnica de terminación (fracturamiento hidráulico) utilizada por cada empresa (operador) en cada área operada (campo), así como información general de perforación de cada pozo (por ejemplo, localización geográfica, fecha de perforación, profundidad total, etc.).

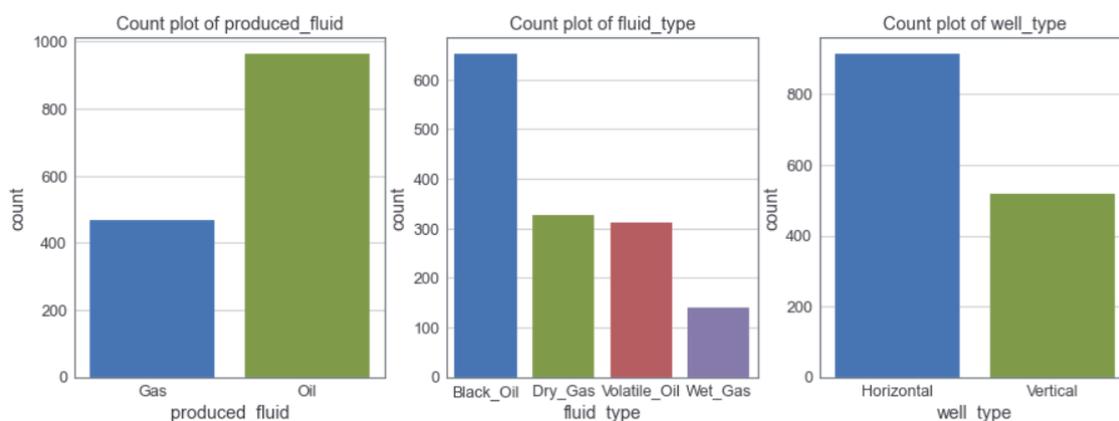


Figura 3. Distribución de pozos en VM por tipo de fluido producido y configuración geométrica.

Desde el año 2010, hasta la fecha del presente reporte, se han perforado un total de 1,432 pozos en Vaca Muerta, incluyendo pozos horizontales y verticales. Casi dos tercios de los pozos perforados se han ubicado en la ventana de petróleo (965 pozos), particularmente en la zona de petróleo negro (“Black Oil”, 653 pozos). La **Figura 3** ilustra la distribución total de pozos en Vaca Muerta, con respecto a la geometría de pozo y tipo de fluido. A la fecha de este reporte, un total de 1,146 pozos se encuentran en producción. La **Figura 4** presenta la evolución del número de pozos en producción en Vaca Muerta.

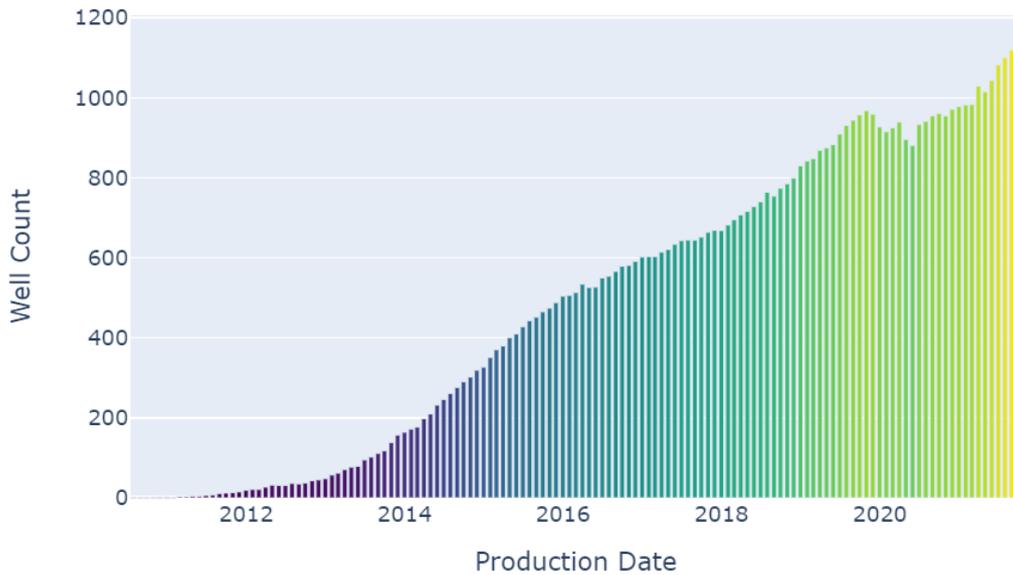


Figura 4. Evolución anual de pozos en producción en Vaca Muerta.

La proporción de pozos horizontales a verticales en la cuenca es de 7:4 (915 pozos horizontales vs. 517 pozos verticales). Sin embargo, según se muestra en la **Figura 5**, a partir del año 2016, el foco ha estado en la perforación de pozos horizontales, lo cuales son completados con fracturamiento hidráulico multietapa.

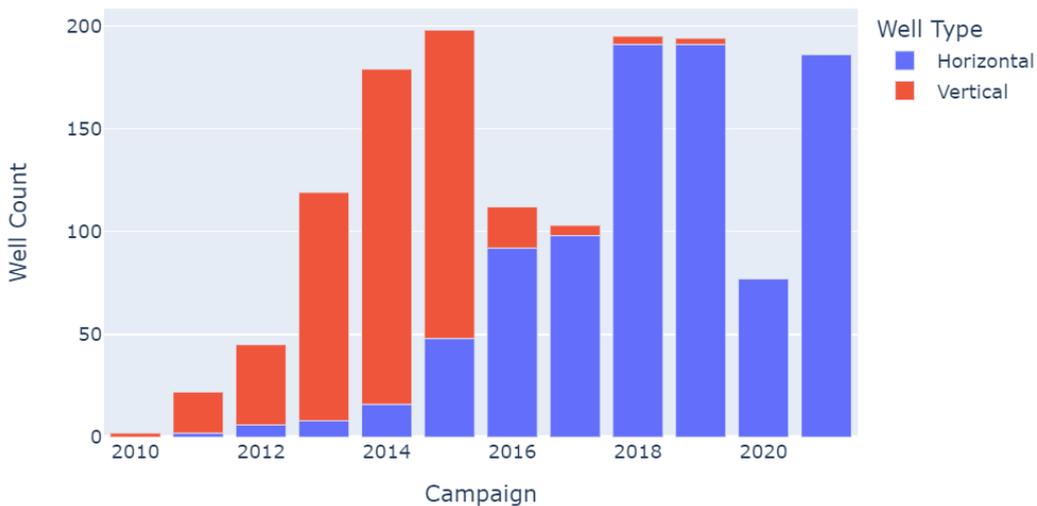


Figura 5. Evolución del número de pozos perforados por año, discriminados por geometría de pozo.

La compañía estatal YPF ha perforado y completado dos tercios de los pozos en Vaca Muerta (954 pozos). El Top-10 de los operadores por actividad en el área (**Tabla 1**) está integrado tanto por compañías locales (Vista, Tecpetrol, PAE y Pluspetrol), como por importantes jugadores internacionales (Total, Shell, Exxon Mobil, Chevron, Phoenix y Wintershall Dea). A manera de

ejemplo, se puede observar que la compañía alemana Wintershall Dea ha perforado y completado un total de 14 pozos (10 pozos horizontales) en sus activos operados en los campos Aguada Federal (petróleo negro) y Bandurria Norte (petróleo volátil).

TABLA 1. Top-10 de operadores por pozos perforados.

	operator	count
0	YPF	954
1	TECPETROL	103
2	TOTAL	76
3	PAE	69
4	SHELL	63
5	VISTA	42
6	PLUSPETROL	36
7	EXXON_MOBIL	28
8	WINTERSHALL	14
9	CHEVRON	12
10	PHOENIX	11

Para facilitar el acceso y consulta de los datos públicos de producción, se ha construido una base de datos Postgres, cuyo diagrama de tablas relacionales se presenta en la **Figura 6**. Utilizando dicha base de datos es posible realizar consultas exploratorias como la que, a manera de ejemplo, se presenta a continuación.

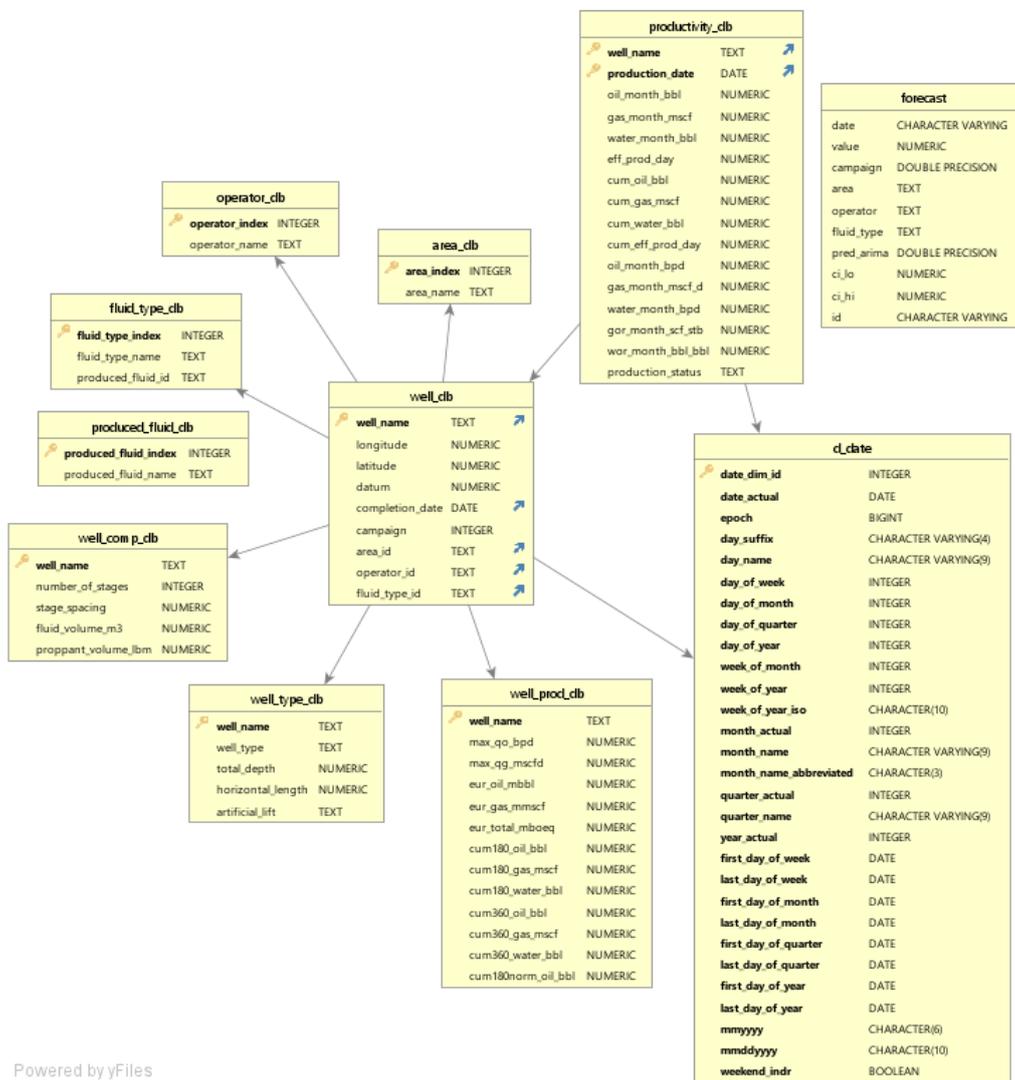
Evolución del número de pozos perforados por trimestre, agrupados por tipo de fluido producido (Petróleo y Gas).

```

SELECT Y.year_actual AS YEAR, Y.quarter_name AS QUARTER, PF.produced_fluid_name AS
PRODUCED_FLUID, COUNT(DISTINCT P.well_name) AS WELL_TOTAL
FROM well_db P, d_date Y, fluid_type_db TF, produced_fluid_db PF
WHERE PF.produced_fluid_name = TF.produced_fluid_id AND
TF.fluid_type_name = P.fluid_type_id AND
P.completion_date = Y.date_actual
GROUP BY Y.year_actual, Y.quarter_name, PF.produced_fluid_name

```

La **Tabla 2** presenta los registros obtenidos como resultado de la consulta anterior, para los años 2020 y 2021.



Powered by yFiles

Figura 6. Esquema de tablas relacionales para consulta de datos de producción de Vaca Muerta.

TABLA 2. Ejemplo de resultado a consulta de base de datos Postgres.

	year integer	quarter character varying (9)	produced_fluid text	well_total bigint
74	2020	First	Gas	5
75	2020	First	Oil	40
76	2020	Fourth	Oil	26
77	2020	Second	Gas	5
78	2020	Second	Oil	1
79	2020	Third	Oil	12
80	2021	First	Gas	22
81	2021	First	Oil	38
82	2021	Second	Gas	19
83	2021	Second	Oil	27
84	2021	Third	Gas	23
85	2021	Third	Oil	45

9.1.1. Filtrado de registros

Anteriormente se mencionó que el presente trabajo busca pronosticar la tasa de producción mensual promedio de petróleo de pozos horizontales perforados y completados en el yacimiento no-conventional Vaca Muerta. Para el análisis correspondiente se dispone de la producción mensual de petróleo, según se muestra en la **Figura 7**. Del total de pozos perforados hasta la fecha en Vaca Muerta, se han seleccionado pozos para el estudio de acuerdo con los siguientes filtros:

- Tipo de geometría de pozo: **Horizontal**
- Longitud mínima de la rama horizontal: **800 metros**
- Tipo de fluido producido: **Petróleo (Black Oil y Volatile Oil)**
- Año de perforación/completamiento (campana): **2015 en adelante**

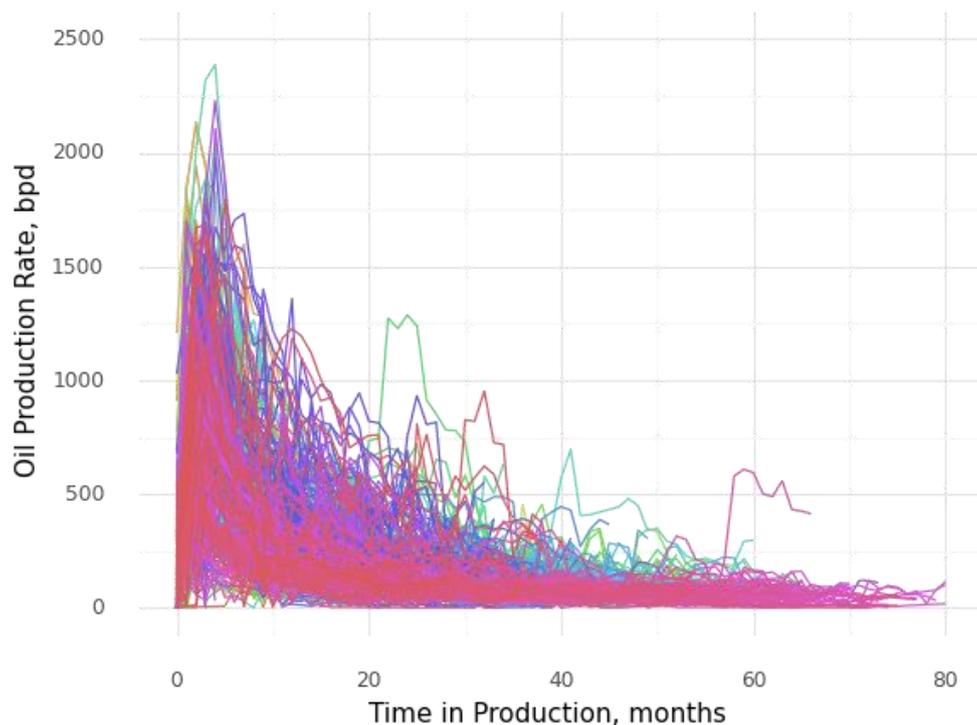


Figura 7. Tasa de producción de petróleo de los pozos considerados en el estudio.

Adicionalmente, se han excluido del estudio pozos en el área de *Bandurria Centro (PAE)* que han sido terminados con técnicas de completamiento distinta a la comúnmente utilizada “*Plug & Perf*”.

Por otra parte, se han seleccionado para el presente estudio pozos con **al menos 24 meses de producción** (longitud mínima del historial de la serie temporal). Lo anterior considera los requerimientos del modelo TFT en cuanto a la longitud de *encoder* (codificador) necesaria para un período de **pronóstico objetivo de 6 meses**.

Como resultado de esta selección de filtros, para el análisis con TFT se dispone de un total de 275 pozos horizontales de petróleo, ubicados en 21 áreas y operados por 9 compañías de petróleo y gas. Dichos pozos fueron perforados y completados durante las campañas de 2015 a 2019.

Para ilustrar el comportamiento típico de producción de los pozos en Vaca Muerta, la **Figura 8** muestra la historia de producción de dos pozos (*AF-5h* y *AF-6h*) perforados y completados por la compañía Wintershall Dea durante su última campaña piloto en el año 2019, en el área Aguada Federal. Ambos pozos tienen una historia de producción de 25 meses a la fecha. La **Tabla 3** abajo contiene la información general y de completamiento de estos pozos que será utilizada como datos de entrada en el modelo de predicción de producción.

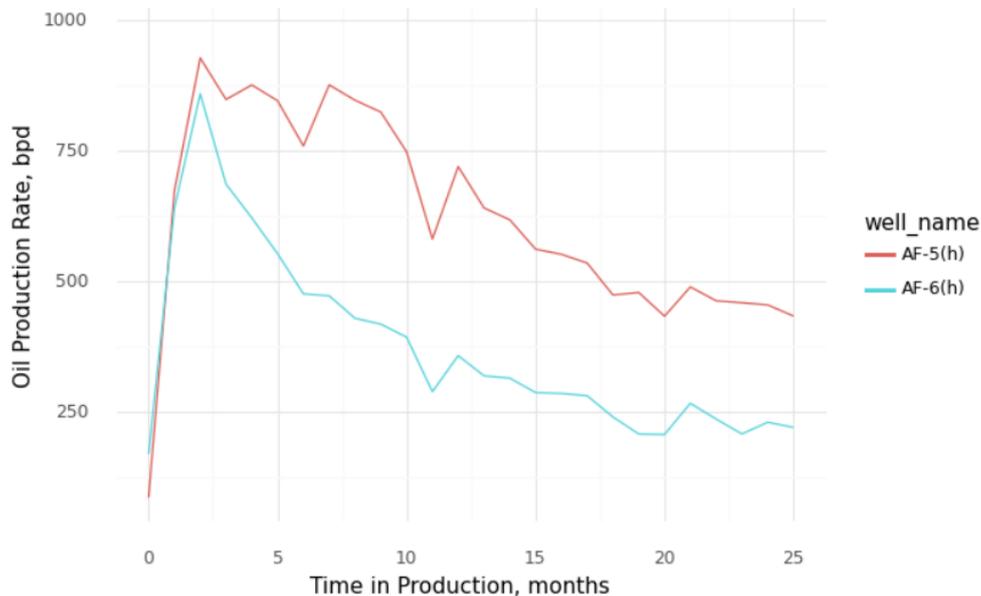


Figura 8. Tasa de producción de petróleo de los pozos considerados en el estudio.

Se puede observar cómo durante una primera fase, que puede durar entre 3 a 6 meses, la tasa promedio de producción diaria de petróleo incrementa a medida que el pozo se va “limpiando”, es decir, va liberando el agua inyectada durante la fractura. Después de alcanzar un valor de producción máxima, la tasa diaria de petróleo comienza a decrecer consistentemente como resultado del depletamiento de la presión de yacimiento en la región cercana al pozo. El objetivo

del presente trabajo será desarrollar un modelo que pueda predecir la producción de los próximos 6 meses en estos pozos (por ejemplo, producción de los meses 26 a 31).

TABLA 3. Información general y de completamiento de los pozos AF-5h y AF-6h.

Parámetro	AF-5(h)	AF-6(h)
Area	Aguada Federal	Aguada Federal
Operador	Wintershall Dea	Wintershall Dea
Campaña	2019	2019
Tipo de Fluido	Black Oil	Black Oil
Longitud Horizontal	2500 m	2500 m
Número de etapas de fractura	35	35
Volumen de fluido de fractura inyectado	41,786 m ³	42,432 m ³
Volumen de apuntalante de fractura inyectado	8,925 Tn	8,990 Tn

9.1.2. Selección de variables

La **Tabla 4** presenta una descripción de las variables que sean considerado relevantes para el presente estudio.

TABLA 4. Lista de variables seleccionadas para el estudio

Nombre de Variable	Descripción	Tipo
well_name	Nombre del pozo	Catórica (UID)
area	Campo donde el pozo está ubicado	Catórica
operator	Compañía operadora	Catórica
campaign	Año de perforación y completamiento	Catórica
fluid_type	Tipo de fluido producido	Catórica
horizontal_length	Longitud de la rama horizontal, m	Numérica
number_of_stages	Número de etapas de fractura	Numérica
fluid_volume_m3	Cantidad de fluido inyectado, m ³	Numérica
proppant_volume_Tn	Cantidad de arena inyectada, Tn	Numérica
oil_month_bpd	Tasa de producción de petróleo, bbl/día	Numérica (Objetivo)
gas_month_mscf_d	Tasa de producción de gas, Mscf/día	Numérica
water_month_bpd	Tasa de producción de agua, bbl/día	Numérica

9.1.3. Manejo de datos faltantes (NA's)

Dado que no todos los pozos en la base de datos cuentan con datos de completamiento, existen registros faltantes (NA's) en los campos *number_of_stages*, *fluid_volume_m3* y *proppant_volume_Tn*. Para efectos del presente estudio, y debido a requerimientos del algoritmo TFT, se ha procedido a remover los registros faltantes.

9.1.4. Adición de “Índice de Tiempo” (time_idx)

Por requerimiento del modelo TFT, se incluye para cada pozo un “índice de tiempo” (*time_idx*) que se incrementa una unidad (i.e., un mes) en cada paso de tiempo.

9.1.5. Ingeniería de covariables (Feature Engineering)

TABLA 5. Lista de variables adicionales incluidas en el estudio.

Nombre de Variable	Descripción	Tipo
log_oil_month_bpd	Log tasa de producción de petróleo	Numérica
oil_month_bpd_roll	Promedio móvil tasa de producción de petróleo	Numérica
avg_oil_by_fluid	Tasa promedio de producción de petróleo, por fluido	Numérica
avg_oil_by_area	Tasa promedio de producción de petróleo, por área	Numérica
avg_oil_by_operator	Tasa promedio de producción de petróleo, por operador	Numérica
avg_oil_by_idx	Tasa promedio de producción de petróleo, por mes	Numérica
avg_hor_length_by_area	Longitud horizontal promedio, por área	Numérica
avg_hor_length_by_operator	Longitud horizontal promedio, por operador	Numérica
avg_stages_by_area	Etapas de fractura promedio, por área	Numérica
avg_stages_by_operator	Etapas de fractura promedio, por operador	Numérica
avg_prop_by_area	Volumen de arena promedio, por área	Numérica
avg_prop_by_operator	Volumen de arena promedio, por operador	Numérica
avg_fluid_by_area	Volumen de fluido promedio, por área	Numérica
avg_fluid_by_operator	Volumen de fluido promedio, por operador	Numérica

Se agregan algunas características (variables) adicionales al conjunto de datos a fin de extraer relaciones de productividad entre los pozos incluidos (**Tabla 5**). Por ejemplo, se ha agregado

como covariable una *transformación logarítmica* de la variable objetivo, las cuales son usadas frecuentemente en distribuciones sesgadas a fin de disminuir la escala de la distribución, haciéndolas más parecidas a una distribución normal. Así mismo, se incluye como covariable el *promedio móvil* de la producción calculado usando una *ventana de 3 meses*. El propósito de incluir esta variable es identificar/enfatizar la dirección de una *tendencia* en la producción.

Adicionalmente, se han incluido también para cada pozo variables que permiten *agregar* a nivel de área, operador o tipo de fluido, el *promedio* de producción, longitud horizontal, número de etapas y cantidad de fluido/apuntalante. Lo anterior busca descubrir posibles tendencias o patrones en el comportamiento de las variables numéricas cuando se agrupan (i.e., “agregar”) por categoría.

9.2. Análisis de series de tiempo con PyTorch Forecasting

PyTorch Forecasting utiliza el método *TimeSeriesDataSet* para crear conjuntos de datos de series temporales (“*datasets*”) que contienen los datos a ser usados en los modelos de series de tiempo. Estos *PyTorch datasets* permiten automatizar tareas comunes, tales como:

- Escalado y codificación de variables.
- Normalización de la variable objetivo.
- Eficiente conversión de series de tiempo en *Pandas* dataframe a tensores de *PyTorch*.
- Almacenamiento de información sobre variables estáticas y variables en el tiempo, conocidas y desconocidas en el futuro
- Almacenamiento de información sobre categorías relacionadas (como días festivos).
- Reducción de resolución (“*downsampling*”) para aumento de datos.
- Generación de conjuntos de datos de inferencia, validación y prueba.

9.2.1. Creación de conjuntos de datos de entrenamiento y validación

En este trabajo se han dividido las series de tiempo en 2 partes: un conjunto de *entrenamiento*, utilizado para el aprendizaje, y un conjunto de *validación*, para el ajuste de *hiperparámetros* y y evaluación del desempeño. Para la creación de los conjuntos de entrenamiento y validación, se han definido las siguientes ventanas de tiempo (**Tabla 6**).

TABLA 6. Ventanas de tiempo para análisis.

Variable	Valor	Definición
<i>max_encoder_length</i>	20	Longitud máxima para codificar. Duración máxima del historial utilizada por el conjunto de datos de la serie temporal.
<i>max_prediction_length</i>	6	Longitud máxima de predicción / decodificador

Basado en las definiciones anteriores, el *conjunto de entrenamiento* se conforma por todos los registros de producción *hasta 6 meses antes del último registro*. Es decir, se guardan los 6 últimos meses de producción de cada pozo para conformar el *conjunto de validación*. Las **Figuras 9 y 10** muestran los registros correspondientes a los *conjuntos de entrenamiento* y *validación*, respectivamente. Se observa que sólo los últimos 6 meses de producción de cada pozo se usan para validación.

Una observación adicional de las Figuras 9 y 10 es que las series de tiempo correspondientes a la producción mensual de petróleo de cada pozo no tienen la misma duración en meses. Dependiendo de la campaña (año) en el que fueron perforados y completados, el historial de producción de los pozos incluidos varía desde 24 meses (producción mínima definida) hasta más de 70 meses para los pozos más antiguos.

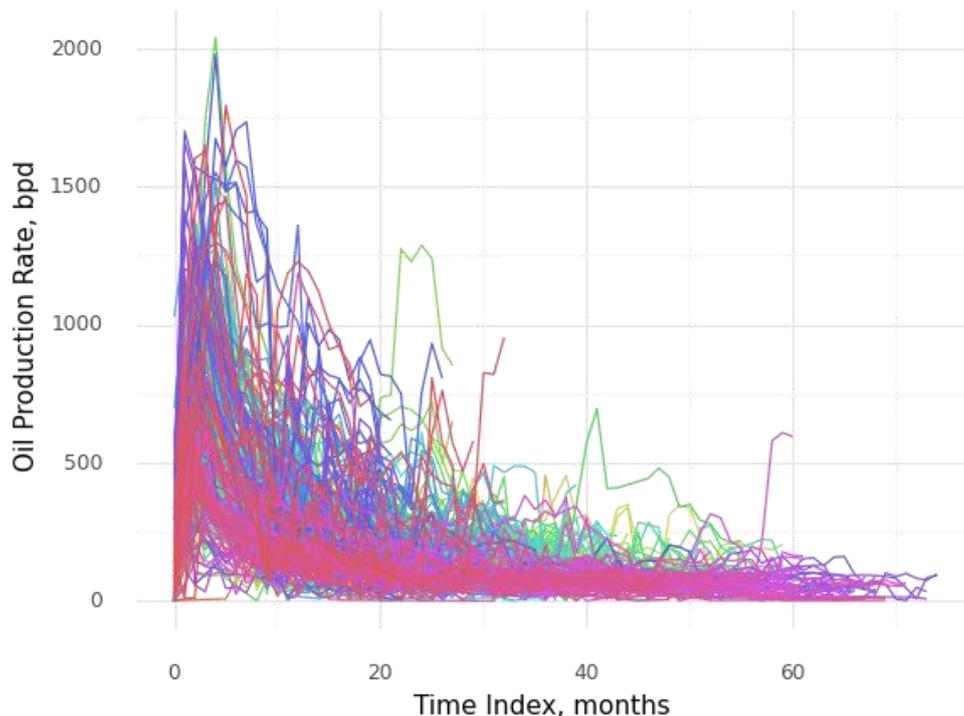


Figura 9. Tasa de producción de petróleo de los pozos incluidos en el set de entrenamiento.

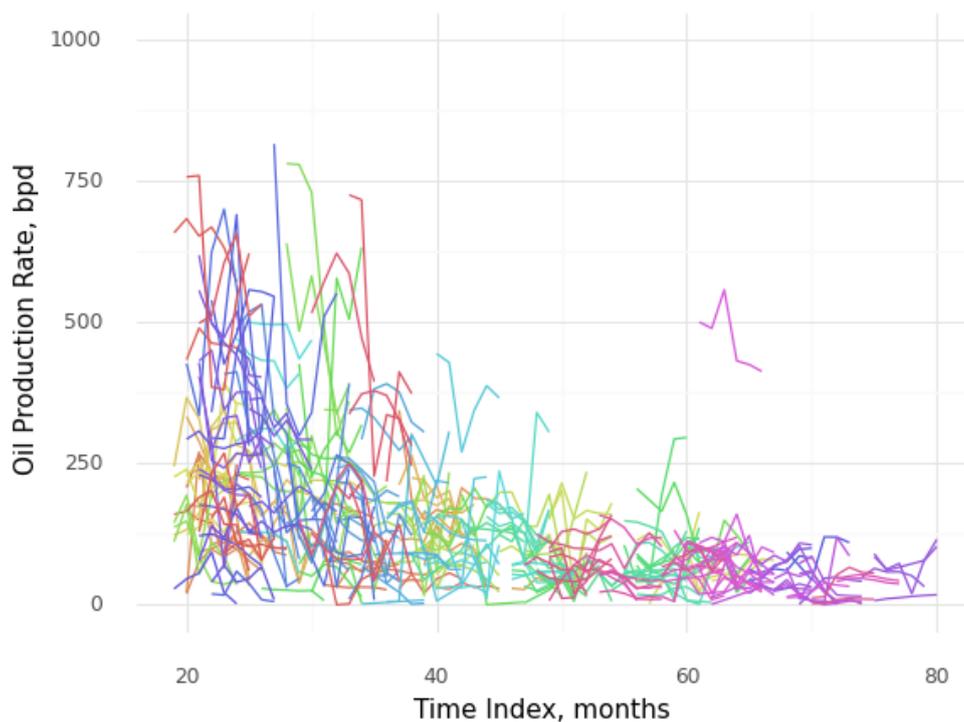


Figura 10. Tasa de producción de petróleo de los pozos incluidos en el set de validación.

9.2.2. Configuración del método TimeSeriesDataSet

El paso siguiente en el análisis consiste en convertir los dataframes de entrenamiento y validación en *PyTorch Forecasting datasets*. Esto incluye especificar cuáles variables son *categoricas* versus *continuas* y cuáles son *estáticas* versus *dependientes del tiempo*. También se debe definir un método para *normalizar* los datos. En este caso, se ha utilizado un *escalador estándar*. La **Tabla 7** muestra los parámetros principales utilizados en la configuración del método *TimeSeriesDataSet*.

TABLA 7. Configuración del método *TimeSeriesDataSet*.

Variable	Valor	Definición
time_idx	"time_idx"	Columna que indica el índice de tiempo
target	"oil_month_bpd"	Columna que indica la variable objetivo
group_ids	"well_name"	Columnas que identifican una serie de tiempo
static_categoricals	"well_name" "area" "operator" "fluid_type" "campaign"	Lista de variables categóricas que no cambian con el tiempo

static_reals	"horizontal_length" "number_of_stages" "fluid_volume_m3" "proppant_volume_Tn" "avg_hor_length_by_area" "avg_hor_length_by_operator" "avg_stages_by_area" "avg_stages_by_operator" "avg_prop_by_area" "avg_prop_by_operator" "avg_fluid_by_area" "avg_fluid_by_operator"	Lista de variables continuas que no cambian con el tiempo
time_varying_known_reals	"time_idx"	Lista de variables continuas que cambian con el tiempo y se conocen en el futuro
time_varying_unknown_reals	"oil_month_bpd" "log_oil_month_bpd" "oil_month_bpd_roll" "water_month_bpd" "gas_month_mscf_d" "avg_oil_by_area" "avg_oil_by_operator" "avg_oil_by_fluid" "avg_oil_by_idx"	Lista de variables continuas que cambian con el tiempo y no se conocen en el futuro

9.2.3. Creación de Dataloaders

El método `to_dataloader` especifica cómo se seleccionan las muestras en lotes (“*batches*”) para el entrenamiento del modelo. Como resultado de la aplicación de este método, objetos de tipo “*dataloader*” se obtienen a partir de los *datasets* de entrenamiento y validación. La **Tabla 8** presenta el valor utilizado para el hiperparámetro “*batch_size*” que debe ser especificado durante la creación de los *dataloaders*.

TABLA 8. Configuración del método `to_dataloader`.

Variable	Valor	Definición
<code>batch_size</code>	64	Tamaño de lote (batch) para el modelo de entrenamiento.

Es importante mencionar que, como parte del proceso de afinación de los hiperparámetros, se realizaron corridas considerando valores adicionales de 32, 48 y 128 para el parámetro de `batch_size`, resultando óptimo el valor seleccionado de 64.

9.2.4. Identificación de la tasa de aprendizaje óptima

Antes del entrenamiento, se puede identificar una *tasa de aprendizaje óptima* con *PyTorch Lightning*. Según se muestra en la **Figura 11** se sugiere una tasa de *0.004365*.

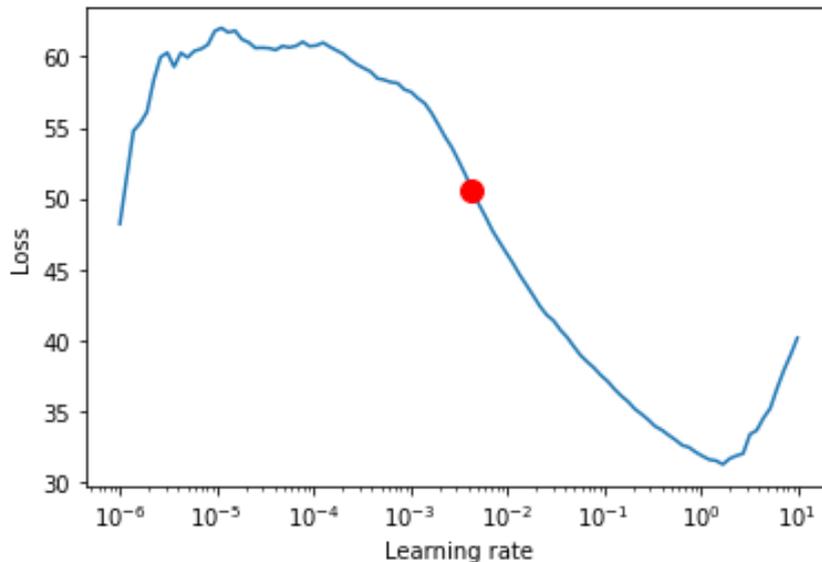


Figura 11. Identificación de la tasa de aprendizaje óptima.

9.2.5. Ajuste de hiperparámetros

La *optimización de hiperparámetros* se realiza a través de un procedimiento de búsqueda aleatoria ("*grid search*"). Se utiliza el método *optimize_hyperparameters* del modelo TFT, integrado con el paquete *Optuna* (<https://optuna.org/>). La lista abajo muestra los rangos de búsqueda definidos por defecto por la librería *PyTorch Forecasting* para todos los *hiperparámetros* a optimizar. Estos rangos se definen de manera que sean apropiados para diferentes configuraciones/aplicaciones de pronóstico de series de tiempo. Los valores determinados a partir de la optimización se muestran en la **Tabla 9**.

- `gradient_clip_val_range=(0.01, 1.0)`,
- `hidden_size_range=(8, 128)`,
- `hidden_continuous_size_range=(8, 128)`,
- `attention_head_size_range=(1, 4)`,
- `learning_rate_range=(0.001, 0.1)`,
- `dropout_range=(0.1, 0.3)`

TABLA 9. Resultado de optimización de hiperparámetros (Loss: 13.891273498535156)

Variable	Valor	Definición
gradient_clip_val	0.91899	Importante para evitar la divergencia del gradiente para redes neuronales recurrentes.
hidden_size	20	Hiperparámetro más importante además de la tasa de aprendizaje.
dropout	0.16652	Importante para prevenir overfitting.
hidden_continuous_size	11	
attention_head_size	1	Número de cabezas de atención.
learning_rate	0.00501	Tasa de aprendizaje.

9.2.6. Creación de instancia del modelo TFT

Name	Type	Params	
0	loss	QuantileLoss	0
1	logging_metrics	ModuleList	0
2	input_embeddings	MultiEmbedding	5.8 K
3	prescalers	ModuleDict	528
4	static_variable_selection	VariableSelectionNetwork	14.6 K
5	encoder_variable_selection	VariableSelectionNetwork	12.8 K
6	decoder_variable_selection	VariableSelectionNetwork	1.8 K
7	static_context_variable_selection	GatedResidualNetwork	1.7 K
8	static_context_initial_hidden_lstm	GatedResidualNetwork	1.7 K
9	static_context_initial_cell_lstm	GatedResidualNetwork	1.7 K
10	static_context_enrichment	GatedResidualNetwork	1.7 K
11	lstm_encoder	LSTM	3.4 K
12	lstm_decoder	LSTM	3.4 K
13	post_lstm_gate_encoder	GatedLinearUnit	840
14	post_lstm_add_norm_encoder	AddNorm	40
15	static_enrichment	GatedResidualNetwork	2.1 K
16	multihead_attn	InterpretableMultiHeadAttention	1.7 K
17	post_attn_gate_norm	GateAddNorm	880
18	pos_wise_ff	GatedResidualNetwork	1.7 K
19	pre_output_gate_norm	GateAddNorm	880
20	output_layer	Linear	147

56.8 K	Trainable params		
0	Non-trainable params		
56.8 K	Total params		
0.227	Total estimated model params size (MB)		

Figura 12. Arquitectura del modelo TFT.

Usando el método *TemporalFusionTransformer.from_dataset* de la librería *PyTorch Forecasting* se puede crear una instancia del modelo *Temporal Fusion Transformer*. La **Figura 12** muestra la configuración del modelo TFT generado, especificando el número de parámetros a entrenar.

Se observa la definición de las diferentes capas (“layers”) incluidas en la arquitectura encoder/decoder, según se explicó en una sección anterior (Figura 2).

9.2.7. Entrenamiento del modelo TFT

Para el entrenamiento del modelo, usando sólo el conjunto de entrenamiento, se define el objeto *Trainer* de la librería *PyTorch Lightning*, con las siguientes configuraciones:

- “*early_stop_callback*”, permite detener el entrenamiento, cuando la métrica de pérdida (“*loss*”) no mejora en el conjunto de validación.
- “*lr_logger*”, permite registrar la tasa de aprendizaje.
- “*logger*”, permite registrar los resultados a *Tensorboard*.

El modelo se inicializa con el método *from_dataset*, empleando los hiperparámetros óptimos determinados anteriormente. La librería *Tensorboard* se utiliza durante el entrenamiento para rastrear métricas de experimentos como pérdida y precisión, visualizar el gráfico del modelo, o para realizar un seguimiento de la velocidad de entrenamiento, la tasa de aprendizaje y otros valores escalares. Dado que *PyTorch Lightning* controla automáticamente el entrenamiento, es posible recuperar y cargar el mejor modelo, basado en la métrica seleccionada.

La **Figura 13**, por ejemplo, muestra la variación en la métrica “pérdida” medida en los conjuntos de entrenamiento y validación (“*train_loss*”, “*val_loss*”), correspondiente al *experimento 224* que tuvo el mejor desempeño durante el estudio. En esta figura se observa cómo, aunque la pérdida sigue disminuyendo en el conjunto de entrenamiento (izquierda), ésta alcanza un valor mínimo y comienza a incrementar en el conjunto de validación (derecha). El entrenamiento se detiene de manera automática después de 20 épocas por activación del “*early_stop_callback*”. Como es de esperarse, se observa también un valor de pérdida menor en el conjunto de entrenamiento que en el conjunto de validación.

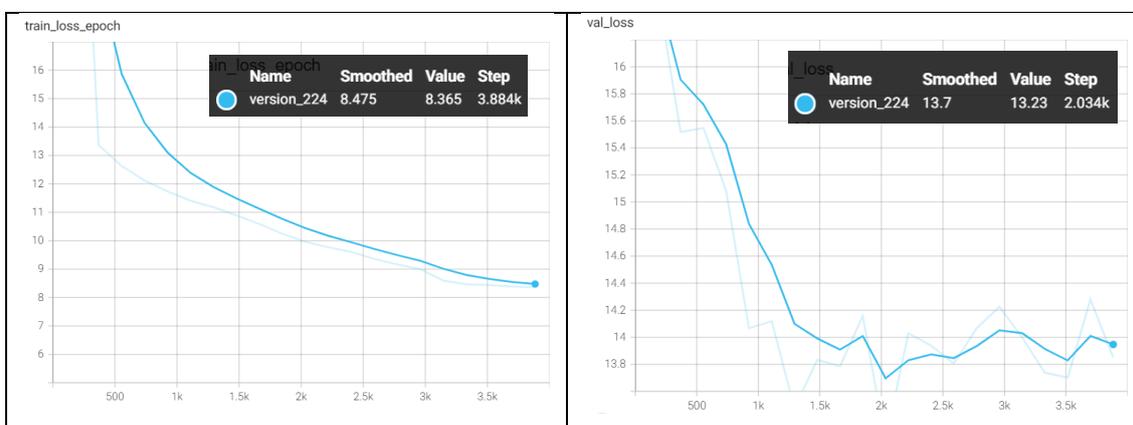
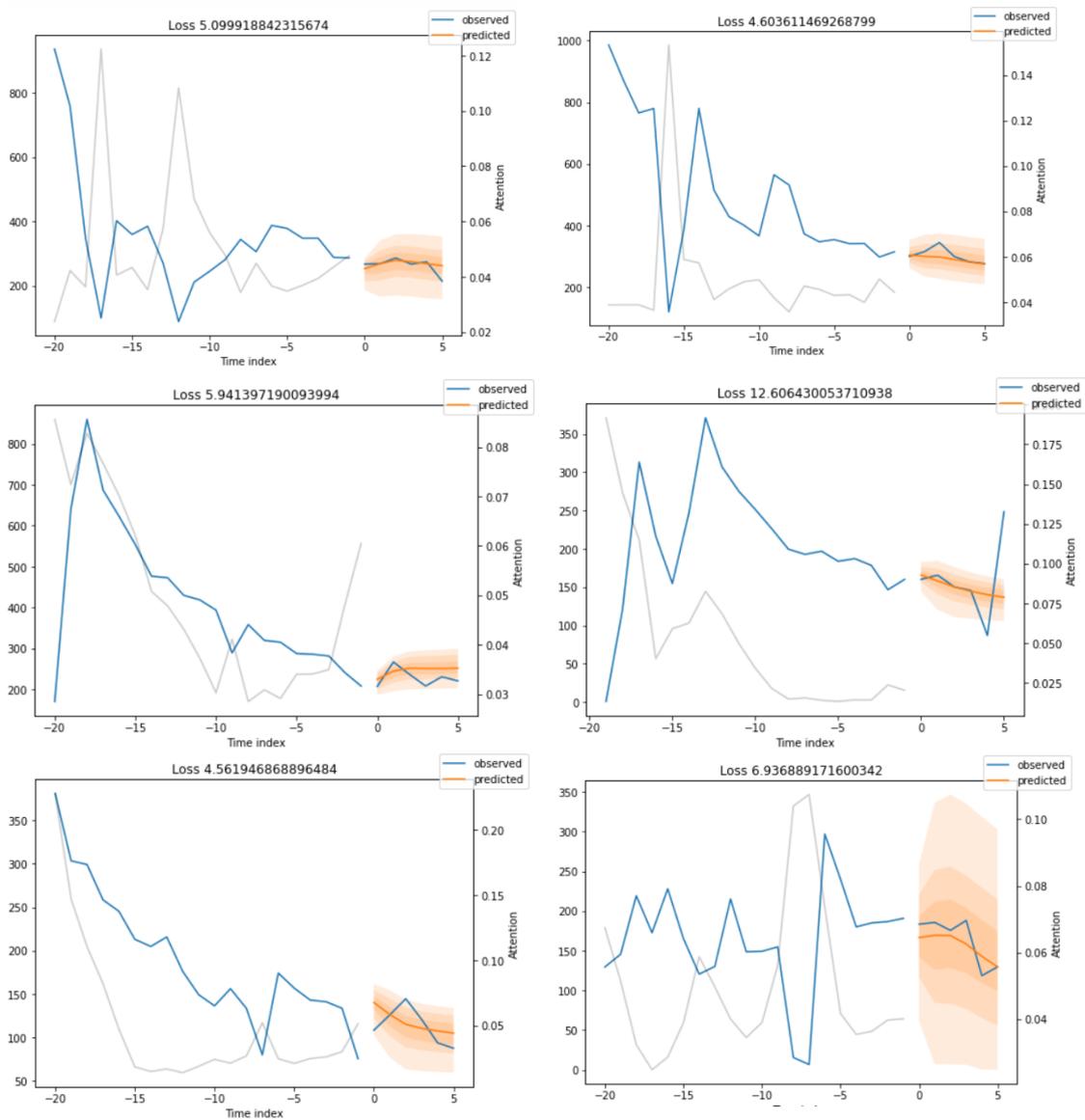


Figura 13. Variación de la métrica pérdida (*loss*) medida en los conjuntos de entrenamiento y validación para el experimento número 224.

9.2.8. Evaluación del modelo TFT

Una vez entrenado el modelo, es posible realizar predicciones usando el método *predict* de *PyTorch Forecasting*. La **Figura 14** muestra ejemplos de predicciones en el conjunto de entrenamiento. Se evalúan las métricas en el conjunto de validación. La línea gris cada ejemplo indica la cantidad de “atención” que el modelo presta a diferentes puntos en el tiempo al momento de hacer la predicción.



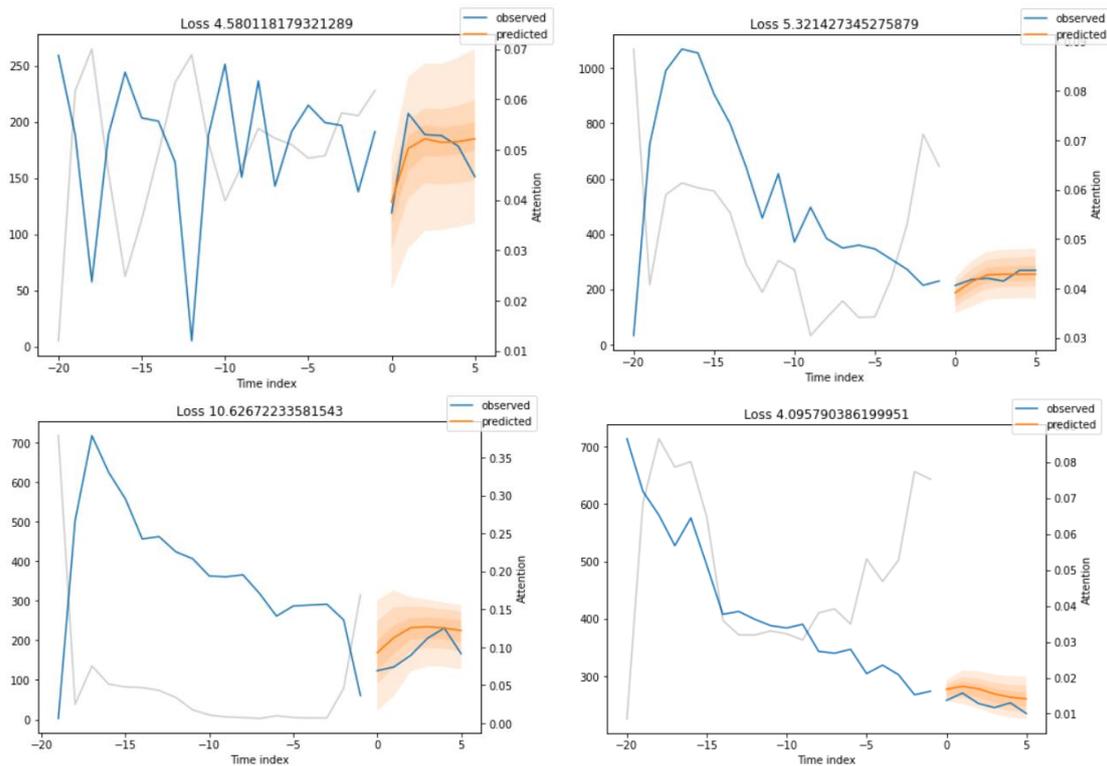


Figura 14. Ejemplos de predicciones usando el set de entrenamiento.

Es importante notar como TFT provee una estimado de incertidumbre dado que no genera un valor único como predicción, sino un pronóstico de *cuantiles (7 cuantiles por defecto)* que representa el rango de valores objetivo para cada horizonte de predicción. Las zonas sombreadas en la predicción de la Figura 14 indican el rango de valores probables entre los cuantiles pronosticados.

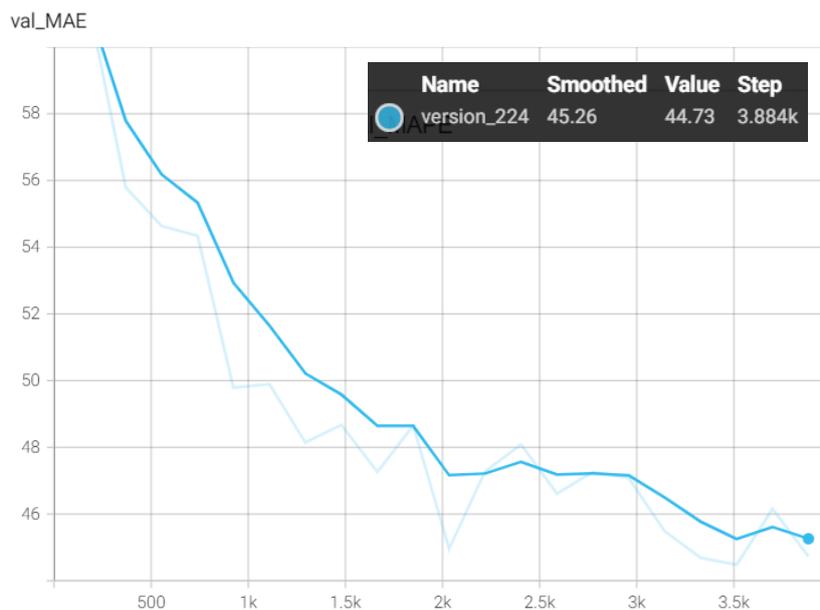


Figura 15. Error Absoluto Medio (MAE) estimado a partir del set de validación.

Usando el set de validación se determina un *Error Absoluto Medio (Mean Absolute Error, MAE)* de 44.73, ver **Figura 15**. Si se compara este error contra el error que resulta de aplicar un modelo de referencia base (i.e., modelo que predice los próximos 6 meses repitiendo las últimas tasas de producción observadas, *MAE: 51.4*) se observa que el modelo TFT tiene un desempeño muy superior.

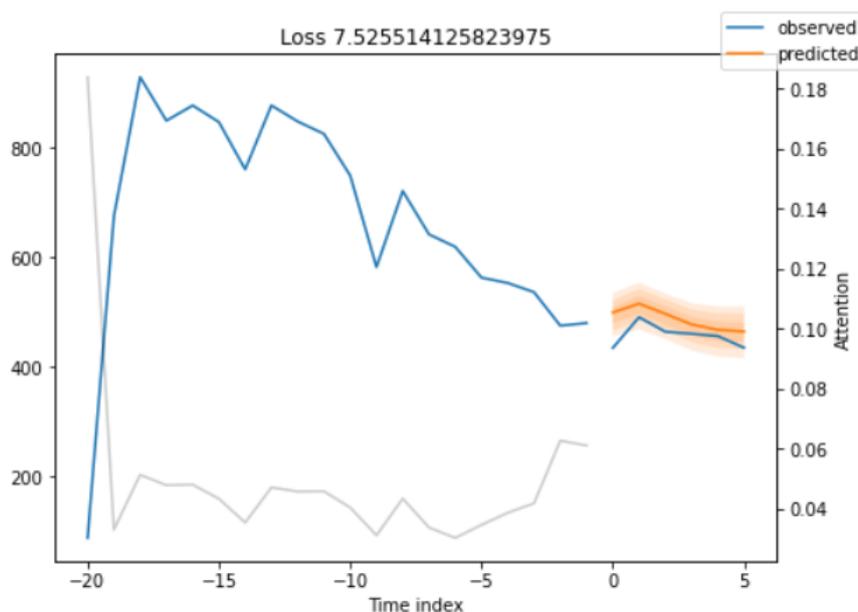


Figura 16. Resultados de entrenamiento y validación, pozo AF-5(h).

Para ilustrar con más detalle los resultados obtenidos del proceso de entrenamiento, la **Figura 16**, a continuación, muestra nuevamente el caso correspondiente al pozo AF-5(h), operado por la compañía Wintershall Dea en el campo Aguada Federal. En primer lugar, se nota claramente en la figura la división de los datos de producción en los conjuntos de entrenamiento y validación (6 últimos meses). Se puede observar una buena correspondencia entre los valores predichos (línea naranja) versus los datos reales (línea azul, conjunto de validación). Se observa también que el rango de incertidumbre en la predicción para el caso particular de este pozo es reducido (franja sombreada alrededor de la predicción), estando la producción real acotada entre los valores máximo y mínimo de los cuantiles de producción pronosticados. El valor de pérdida calculada durante el entrenamiento/validación es de 7.5.

9.2.9. Revisión de predicciones pobres (alto error)

Examinar los registros con el peor desempeño, en términos de error, permite identificar dónde el modelo tiene problemas de confiabilidad en la predicción, sirviendo esto para determinar

mejoras en el mismo. Para describir el error se puede emplear distintas métricas incluidas en la librería *PyTorch Lightning*, tales como *MAE*, *MAPE*, *SMAPE*, y *RMSE*. En la **Figura 17** se muestran algunos ejemplos donde el modelo se ha desempeñado pobremente en el conjunto de validación.

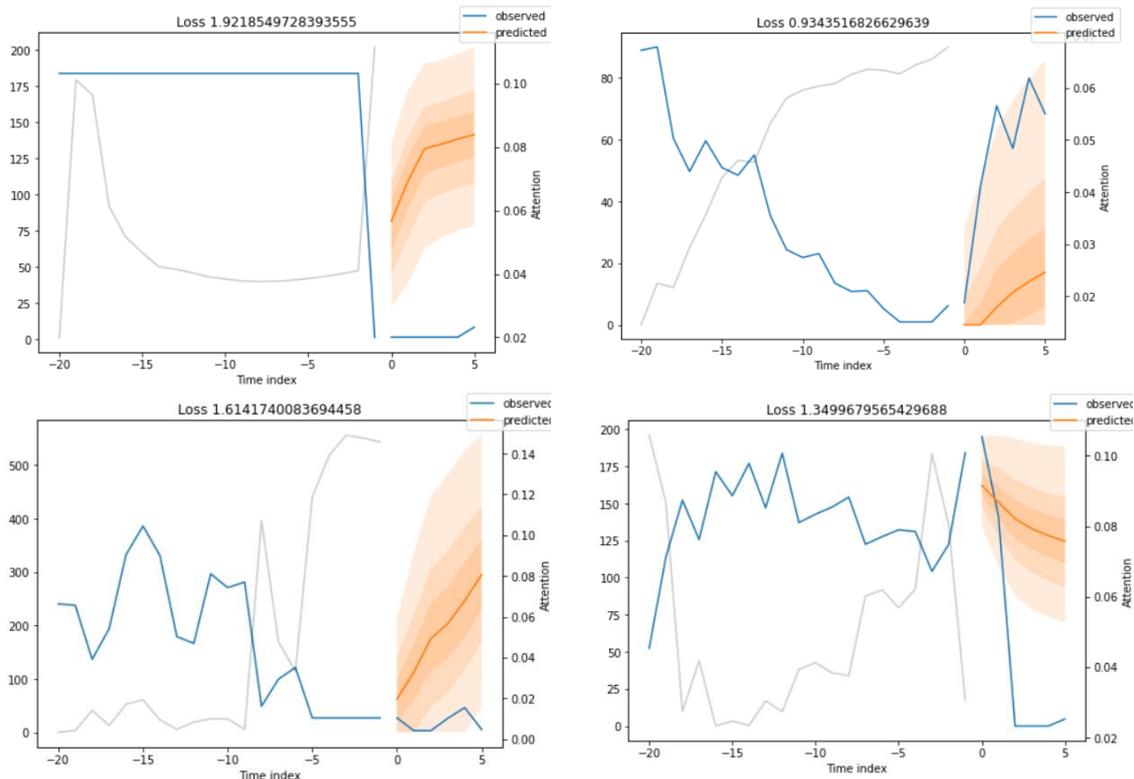
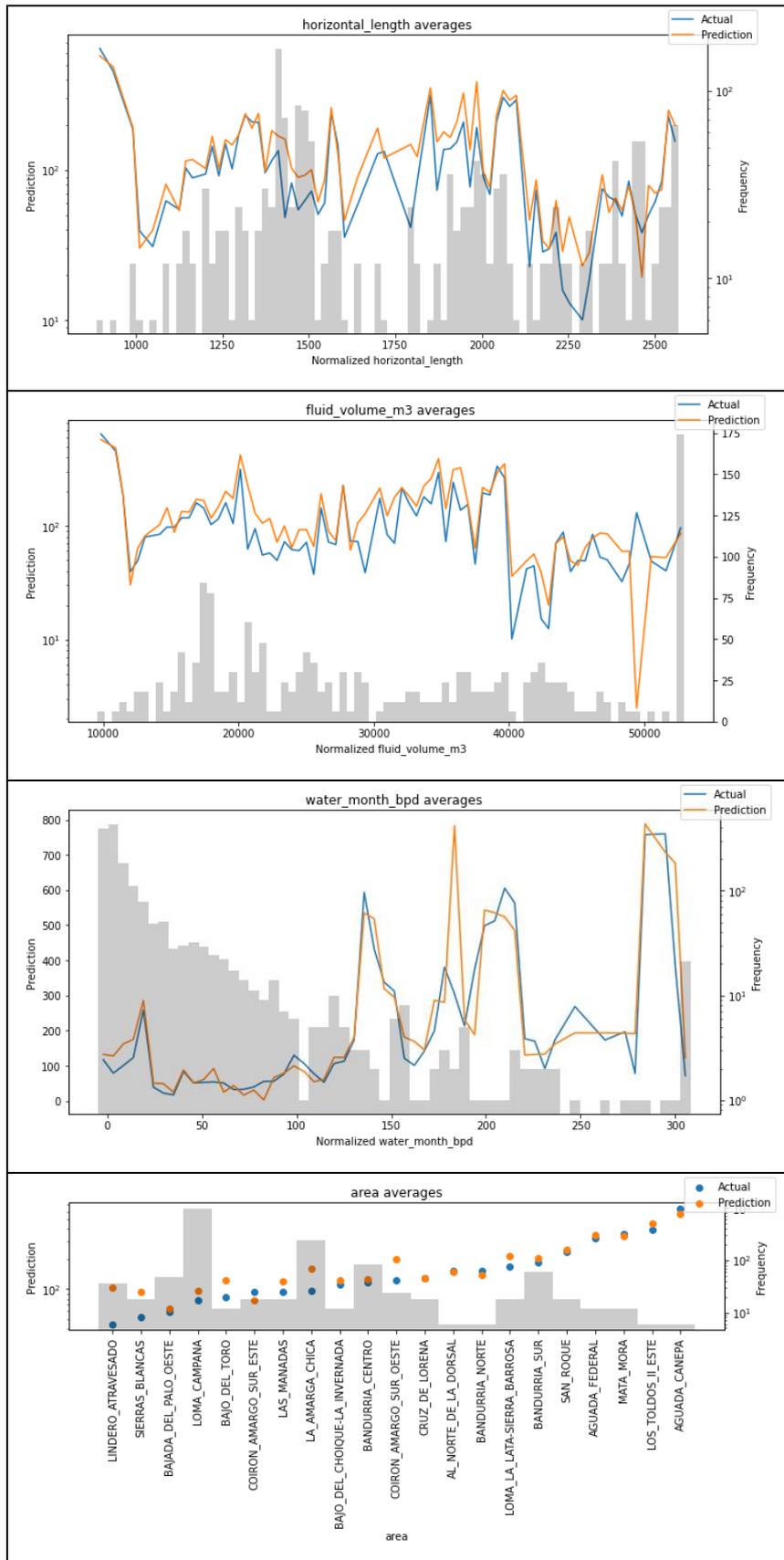


Figura 17. Ejemplos de pozos donde el modelo ha mostrado pobre desempeño según el error estimado con la métrica *SMAPE*, usando el conjunto de validación.

De los ejemplos mostrados en la Figura 17 se observa que el modelo tiene dificultades en la predicción cuando el valor de la producción se hace igual a cero. En caso de ser necesario, esta condición deberá investigarse con mayor detalle para buscar estrategias que permitan minimizar este error.

9.2.10. Predicción versus valores futuros por variable

Otra forma de detectar debilidades en el modelo es mediante la verificación de los resultados en diferentes segmentos de los datos de entrada. La **Figura 18** muestra ejemplos de los promedios de las predicciones contra los datos reales, para el rango de valores de cada variable (100 divisiones/bins). Los histogramas indican la frecuencia de la variable por intervalo.



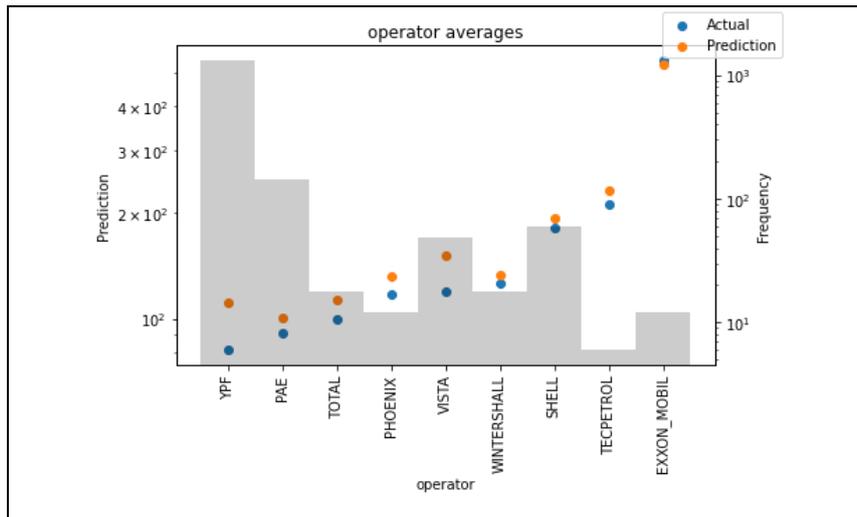


Figura 18. Ejemplos de promedios de las predicciones contra datos reales, dentro del rango de valores de cada variable. Los histogramas indican la frecuencia de la variable por intervalo.

De la Figura 18, se muestra una buena correspondencia entre el promedio de las predicciones y los valores reales a lo largo de todo el rango de valores para las variables “horizontal_length”, “fluid_volume_m3” y “water_month_bpd”. Dentro de las áreas estudiadas, se observa una mayor discrepancia entre la predicción y el valor real en los campos Lindero Atravesado (PAE), La Amarga Chica (YPF) y Coirón Amargo Sur Oeste (Shell). En el caso de los operadores, el error es mayor en los casos de YPF y Vista Oil&Gas. Las discrepancias indicadas deberán ser investigadas con el fin de mejorar el desempeño predictivo del modelo.

9.2.11. Interpretación del modelo entrenado

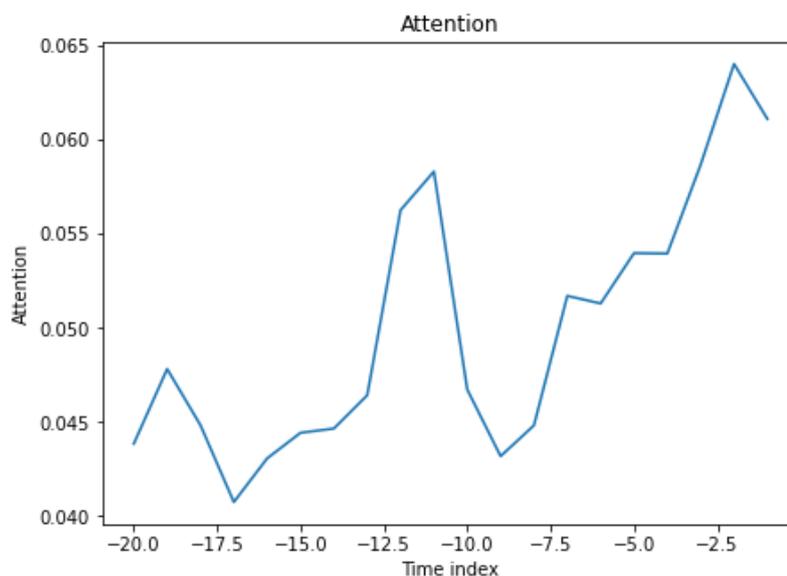


Figura 19. Interpretabilidad del modelo TFT, mecanismo de atención.

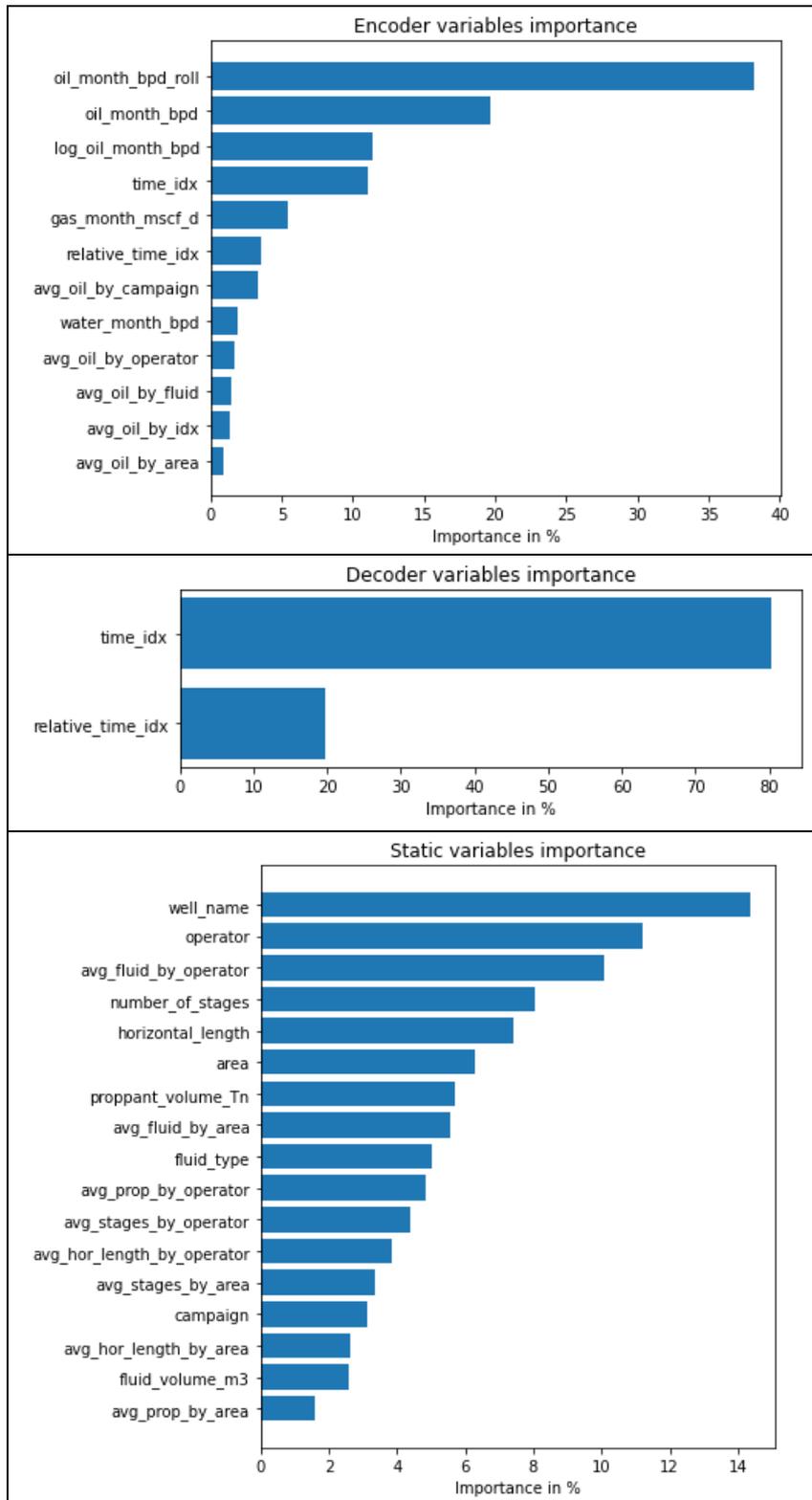


Figura 20. Interpretabilidad del modelo TFT, importancia de variables.

A diferencia de los modelos de aprendizaje profundo convencionales (i.e., Redes Neuronales), el modelo *Temporal Fusion Transformer (TFT)*, implementado mediante la librería *PyTorch*

Forecasting, cuenta con capacidades de interpretación incorporadas. Por ejemplo, la **Figura 19** presenta el comportamiento del nivel de atención durante la predicción. La **Figura 20**, por otra parte, muestra la importancia de las variables que se definieron como variables estáticas y también de las variables que se incluyeron en el *codificador (encoder)/decodificador (decoder)*.

Como era de esperarse, las variables relacionadas con las tasas de producción de petróleo observadas en el pasado (e.g., “*oil_month_bpd_roll*”, “*oil_month_bpd*”, “*log_oil_month_bpd*”) aparecen como las variables principales en el *codificador (encoder)*. Por otra parte, dado que el índice de tiempo (“*time_idx*”) y el índice de tiempo relativo (“*relative_time_idx*”) son las únicas variables “*conocidas*” en el futuro, también aparecen como los principales predictores del *decodificador (decoder)*. Entre las variables estáticas, los lugares más altos están ocupados por variables relacionadas con el diseño de completamiento de los pozos, tales como “*number_of_stages*”, “*proppant_volume_Tn*” y “*horizontal_length*”. Dado que la tercera variable en importancia está relacionada con la cantidad promedio de fluido de fractura inyectado por cada operador (“*avg_fluid_by_operator*”), podría esperarse que la variable “*fluid_volume_m3*” (i.e. volumen de fluido inyectado por etapa) ocupara un lugar más alto si la primera no estuviera incluida en el modelo.

9.2.12. Predicción de valores futuros

Siguiendo el proceso general descrito en la sección 8.2.1, una vez que se ha entrenado, evaluado y seleccionado el “mejor modelo” (mediante el uso de los conjuntos de entrenamiento y validación), con el fin de predecir valores futuros se requiere “re-entrenar” el (mejor) modelo, utilizando el mismo programa de tasa de aprendizaje, pero incorporando ahora toda la historia de producción (entrenamiento + validación). Además de la tasa de aprendizaje (learning rate), durante el re-entrenamiento del modelo también se usan los mismos hiperparámetros optimizados, lo mismo que el máximo número de épocas del entrenamiento inicial.

La **Figura 21** muestra una comparación de la evolución de la métrica “*train_loss*” durante el entrenamiento inicial (usando sólo el conjunto de entrenamiento) y el re-entrenamiento (usando todo el set de datos) del mejor modelo. En ambos casos, el entrenamiento se realizó por 20 épocas, con una *tasa de aprendizaje inicial de 0.005* (**Figura 22**). El mayor número de pasos que se observa durante el re-entrenamiento se debe a la mayor cantidad de registros utilizados. El número de épocas, sin embargo, es consistente en ambos experimentos.

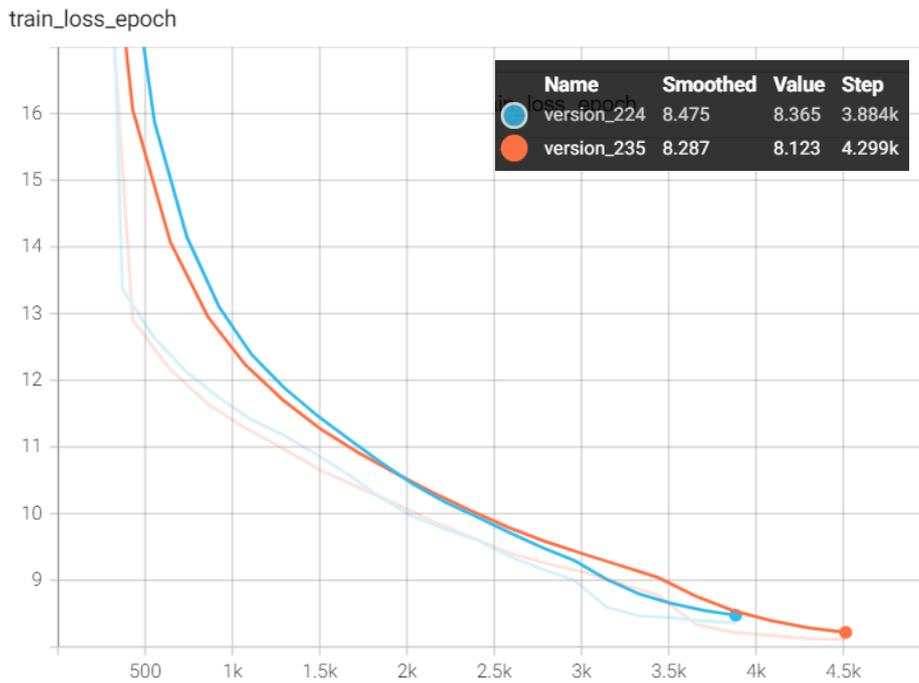


Figura 21. Variación de la métrica pérdida (*train_loss*) para los experimentos números 224 (entrenado solo usando el conjunto de entrenamiento) y 235 (entrenado con todo el set de datos).

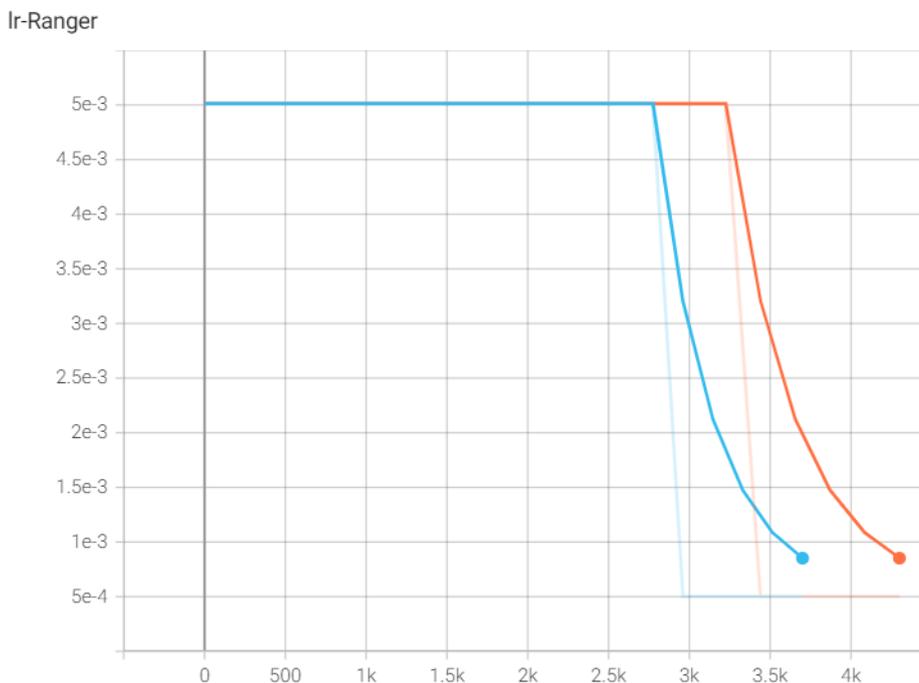


Figura 22. Programa de tasa de aprendizaje para los experimentos números 224 (entrenado solo usando el conjunto de entrenamiento) y 235 (entrenado con todo el set de datos).

Una vez reentrenado, el modelo se puede aplicar a datos nuevos. No es necesario definir de antemano ninguna variable futura dado que en la definición del modelo no se han incluido covariables (numéricas o categóricas) conocidas en el futuro (a excepción del índice de tiempo).

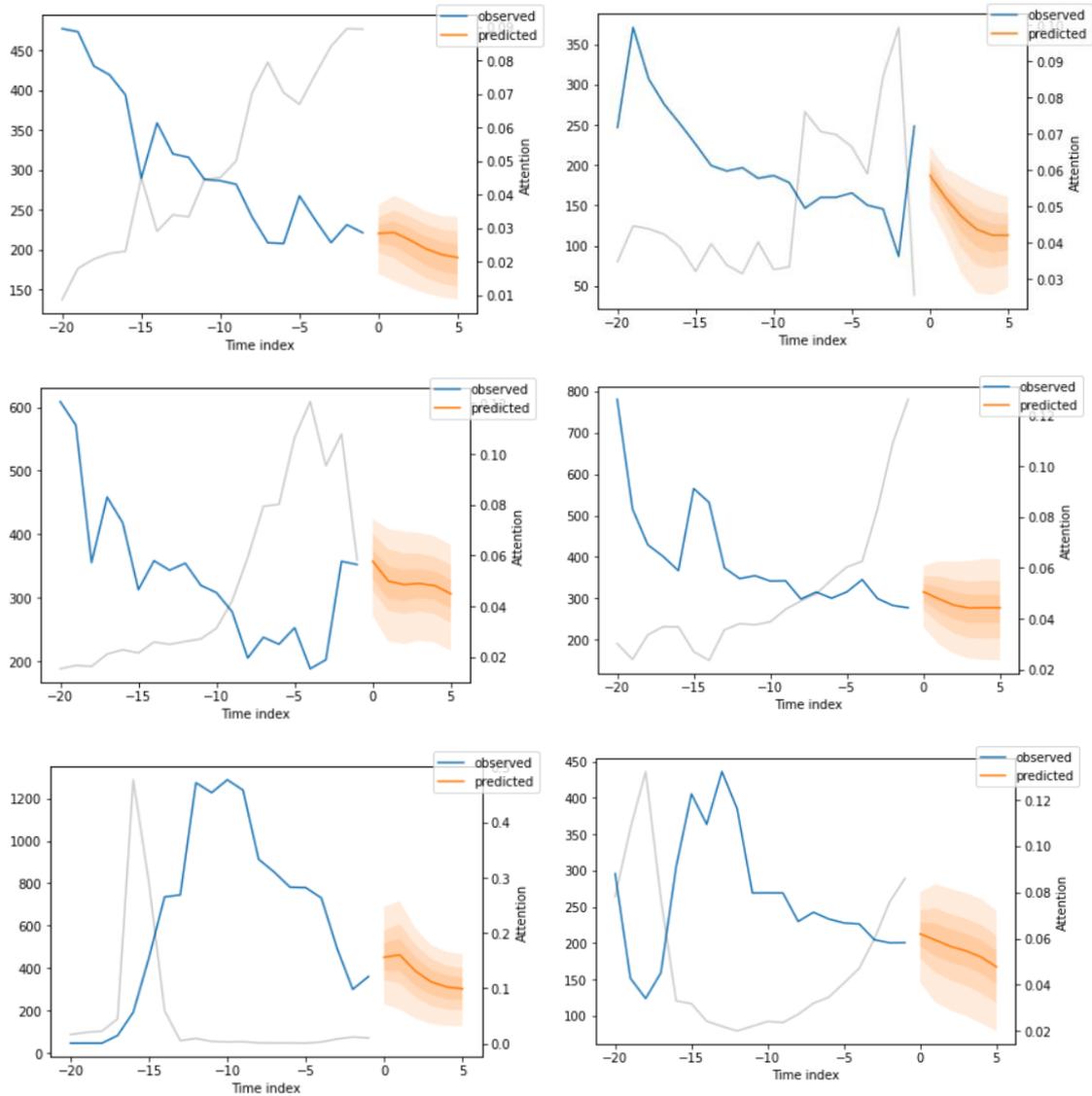


Figura 23. Ejemplos de producción futura (6 meses) de petróleo.

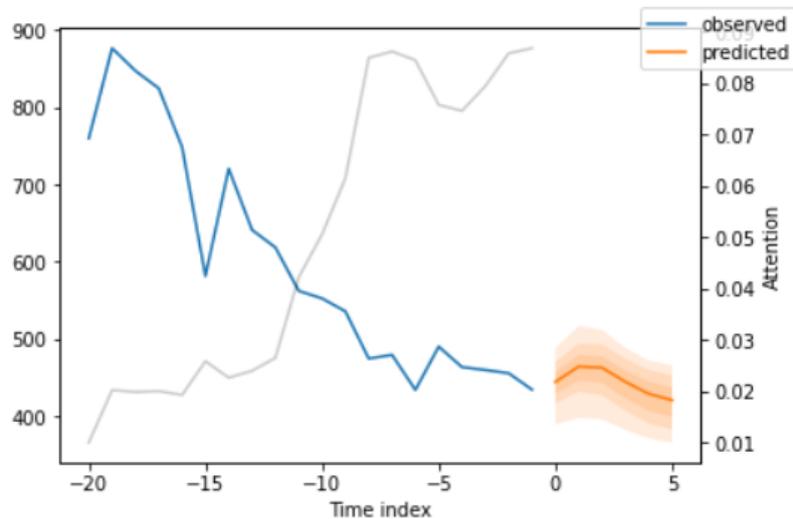


Figura 24. Predicción de valores futuros, pozo AF-5(h).

La **Figura 23** muestra algunos ejemplos de pronóstico de producción futura de petróleo (6 meses). En adición a esto, continuando con el ejemplo correspondiente al pozo AF-5(h), la **Figura 24** presenta la predicción de la producción para este pozo. La Tabla 10 lista los valores de producción (i.e., cuantiles) pronosticados. Se resaltan los cuantiles correspondientes a una probabilidad *P90* (*q2*, 90% de probabilidad de producir este valor), *P50* (*q4*, 50% de probabilidad de producir este valor) y *P10* (*q6*, 10% de probabilidad de producir este valor) dado que es la convención que utiliza en la industria petrolera para reportar el rango probable de producción.

TABLA 10. *Quantiles future prediction for well AF-5(h)*

Mes	q1 (0.02)	q2 (0.10)	q3 (0.25)	q4 (0.50)	q5 (0.75)	q6 (0.90)	q7 (0.98)
1	404.0803	432.5252	446.0333	463.1706	478.9948	491.5942	502.3402
2	405.0707	436.7318	453.1129	471.9417	488.8147	503.6898	519.2156
3	403.3322	433.6080	451.5269	470.1102	486.2979	501.1376	514.8983
4	396.7097	425.8762	445.2687	464.0139	479.8018	494.7725	507.0128
5	389.0611	417.4184	438.2770	457.5937	473.3562	488.8160	500.0141
6	389.0611	408.8358	430.9346	451.0272	467.0300	483.1436	493.8255

10. Discusión

En el presente estudio se usó un set de datos de la Secretaría de Gobierno de Energía, conteniendo la producción mensual de 275 pozos no-convencionales en la formación Vaca Muerta. Para el análisis, se usaron los últimos 20 meses para pronosticar los siguientes 6 meses de producción de petróleo.

Aunque el set original de datos cuenta con alrededor de 65,000 registros de producción de más de 1,400 pozos en la cuenca, una vez aplicados los filtros relevantes han quedado sólo 12,390 registros para el análisis. Esto sin duda representa una limitación para la implementación de una solución basada en aprendizaje profundo (“Deep Learning”) la cual se favorece de la disponibilidad de datos para el aprendizaje de relaciones temporales entre las variables. Una opción para mejorar la cantidad de datos sería disponer de datos diarios de producción (en lugar de datos mensuales), o incluir más pozos en el análisis.

Un aspecto por enfatizar, concerniente a los datos utilizados, es la no disponibilidad de variables operativas, como por ejemplo presión en fondo o en superficie, o tamaño del orificio de

producción (“choke size”). Dado que estas variables están bajo control del operador del pozo (i.e., estrategia de drawdown), sería importante considerarlas para un próximo estudio, en especial como variables temporales conocidas en el futuro.

A pesar de las limitaciones del set de datos, la aplicación del modelo Temporal Fusion Transformer para el pronóstico de la producción futura de petróleo ha resultado en una mejora del 13% en el error absoluto promedio con respecto al modelo de referencia base. Los ejemplos que se incluyen en la Figura 14 muestran una predicción bastante cercana a los datos reales (lo cual se refleja en una baja pérdida), con un rango de probabilidad (cuantiles) acotado (i.e., baja incertidumbre). Las predicciones futuras de la Figura 23, por otra parte, aunque siguen la tendencia de los datos históricos, muestran una incertidumbre en la predicción un poco mayor. Esta incertidumbre podría reducirse si se incluyeran covariables (numéricas o categóricas) futuras “conocidas” en el modelo, indicando la variación en el tiempo de dichas covariables. Un ejemplo de dichas variables podría ser las mencionadas arriba: programa de presiones en superficie o fondo (drawdown) y tamaños planeados del orificio de producción.

Dado que no existe un componente importante de estacionalidad en la tasa de producción de petróleo, de los componentes responsables de capturar las relaciones temporales (i.e., capas de procesamiento local y auto-atención) el procesamiento local entre pasos de tiempo adyacentes parece tener un papel relevante. Esto es, resulta crítico estudiar las observaciones pasadas de la variable objetivo. El comentario anterior se verifica observando los resultados en la Figura 20. Se aprecia por ejemplo que, para valores en el pasado (codificador), valores pasados de la variable objetivo (e.g., *oil_month_bpd*, *oil_month_bpd_roll*, *log_oil_month_bpd*), son críticas dado que las predicciones son extrapolaciones de las observaciones pasadas.

En el caso de las variables estáticas, se observa gran importancia de las variables asociadas con las técnicas de completamiento de pozo (i.e., diseño del fracturamiento hidráulico), tales como “*avg_fluid_by_operator*”, “*number_of_stages*” y “*proppant_volume_Tn*”, así como de la longitud de la rama horizontal del pozo (“*horizontal_length*”). Dado que las variables “*operator*” y “*area*” aparecen también con alta importancia, se deben estudiar posibles diferencias significativas en las técnicas de completamiento empleadas por cada operador en cada área.

Una ventaja del modelo TFT con respecto a modelos tradicionales de aprendizaje automático es la capacidad de aprender “patrones” temporales pasados y persistentes a partir del análisis de patrones en el mecanismo de atención. En la Figura 19 se observa que el set de datos de producción de Vaca Muerta no exhibe un patrón estacional. Por el contrario, se observa un

dominio en importancia para los últimos meses de producción, es decir, un patrón de tendencia en decadencia hacia los meses más tempranos.

Un uso adicional de la “interpretabilidad” del modelo TFT es la identificación de regímenes o eventos que produzcan cambios significativos en dinámica temporal. Poder identificar dichos cambios de régimen ayuda en la comprensión del problema subyacente. Por ejemplo, el pico en el nivel de atención que se muestra en la Figura 19 podría estar asociado con un cambio en las condiciones de producción a nivel de yacimiento. Se requiere realizar análisis adicionales para verificar esta hipótesis.

11. Conclusiones

En el presente estudio se ha evaluado la aplicabilidad de *Temporal Fusion Transformers*, un novedoso modelo de aprendizaje profundo, para el pronóstico de la producción de petróleo en múltiples horizontes de tiempo. El modelo permite utilizar en el análisis tanto variables estáticas (no dependientes del tiempo), como variables observadas (que han variado en el tiempo durante el pasado) y variables futuras conocidas de antemano.

El modelo se ha utilizado para la predicción de 6 meses de producción en el futuro de 275 pozos horizontales de petróleo en la formación no-convencional Vaca Muerta. El horizonte de predicción ha estado limitado por la cantidad y frecuencia de datos disponibles. Sin embargo, al comparar contra un modelo base, el modelo muestra un buen desempeño (bajo error).

Haciendo uso de las características de interpretabilidad del modelo TFT, ha sido posible identificar ciertas variables que tienen mayor importancia en la predicción de la producción de petróleo. Por ejemplo, el análisis muestra que variables de completamiento como el volumen de fluido y apuntalante inyectado, la longitud horizontal y el número de etapas de fractura son relevantes en las áreas más productivas.

El presente estudio ha servido como una prueba de concepto sobre la potencial aplicación de este modelo en el pronóstico de producción en la industria de gas y petróleo. Estudios adicionales, donde se corrijan las limitaciones mencionadas referente al set de datos, son sugeridos para confirmar las ventajas de aplicación de este método.

12. Referencias Bibliográficas

Abdullayeva, F., & Imamverdiyev, Y. (2019). Development of Oil Production Forecasting Method Based on Deep Learning. *Statistics, Optimization & Information Computing* 7.4: 826-839.

Ballin, P. R., Journel, A. G., & Aziz, K. (1992). Prediction of Uncertainty in Reservoir Performance Forecast. *Petroleum Society of Canada*. doi:10.2118/92-04-05.

Berneti, S. M., & Shahbazian, M. (2011). An imperialist competitive algorithm artificial neural network method to predict oil flow rate of the wells. *International journal of computer applications*, 26(10), 47-50.

GAO, H. H., QU, Z. Q., & Zhao, M. (2008). Experimental study on the factors affecting the productivity of fractured horizontal wells. *Journal of Southwest Petroleum University (Science & Technology Edition)*, 4, 020.

Hong, A., Bratvold, R. B., Lake, L. W., & Ruiz Maraggi, L. M. (2019). Integrating Model Uncertainty in Probabilistic Decline-Curve Analysis for Unconventional-Oil-Production Forecasting. *Society of Petroleum Engineers*. doi:10.2118/194503-PA.

Joshi, K. G., Awoleke, O. O., & Mohabbat, A. (2018). Uncertainty Quantification of Gas Production in the Barnett Shale Using Time Series Analysis. *Society of Petroleum Engineers*. doi:10.2118/190124-MS.

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*.

Liu, Y., Gong, C., Yang, L., & Chen, Y. (2020). DSTP-RNN: A dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. *Expert Systems with Applications*, 143, 113082.

Olominu, O., & Sulaimon, A. A. (2014). Application of Time Series Analysis to Predict Reservoir Production Performance. *Society of Petroleum Engineers*. doi:10.2118/172395-MS.

“PyTorch Forecasting Documentation” PyTorch Forecasting, <https://pytorch-forecasting.readthedocs.io/en/latest/index.html>. Visitado en Septiembre 2021.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971.

Rotondi, M., Nicotra, G., Godi, A., Contento, F. M., Blunt, M. J., & Christie, M. (2006). Hydrocarbon Production Forecast and Uncertainty Quantification: A Field Application. Society of Petroleum Engineers. doi:10.2118/102135-MS.

Sagheer, A., & Kotb, M. (2019). Time series Forecasting of Petroleum Production Using Deep LSTM Recurrent Networks. Neurocomputing, 323, 203-213.

Zhou, W., Gupta, S., Banerjee, R., Poe, B., Spath, J., & Thambynayagam, M. (2013). Production Forecasting and Analysis for Unconventional Resources. International Petroleum Technology Conference. doi:10.2523/IPTC-17176-MS.