CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TRABAJO FINAL INTEGRADOR

ANEXO – I: Análisis Exploratorio de Datos (EDA)

TITULO

Pronóstico del Salto Hidráulico de una planta de generación hidroeléctrica, utilizando algoritmos de Aprendizaje Automático

Nombre y Apellido del Alumno/a: José Luis Beltramone

Título de grado o posgrado (último): Especialista en Gestión de las Telecomunicaciones

Director; Mg. Ing. Gustavo Denicolay

Lugar y Fecha: La Plata, junio de 2022

Tabla de Contenido

Lista de Ilustraciones

Introducción

En este anexo se desarrolla el análisis exploratorio de datos, completo.

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Objetivo específico: Entender el rol de las variables que componen el conjunto de datos

Como plataforma de gráficos se utilizó la librería 'Plotly' (Plotly Python Graphing Library) en conjunto con Cufflinks (GitHub - santosjorge/cufflinks: Productivity Tools for Plotly + Pandas) para obtener gráficos interactivos que permitan una exploración más sencilla y rápida (Create Stunning Plots on Pandas Dataframes in One Line of Code | Towards Data Science).

Para cada variable se muestra un gráfico en el tiempo, para poder observar tendencias, estacionalidades y variaciones, un histograma para poder observar la distribución y centro de los valores y sus frecuencias y un diagrama de caja (boxplot) para observar, además, los valores atípicos.

Variables que intervienen

Cotas de embalse y de restitución:

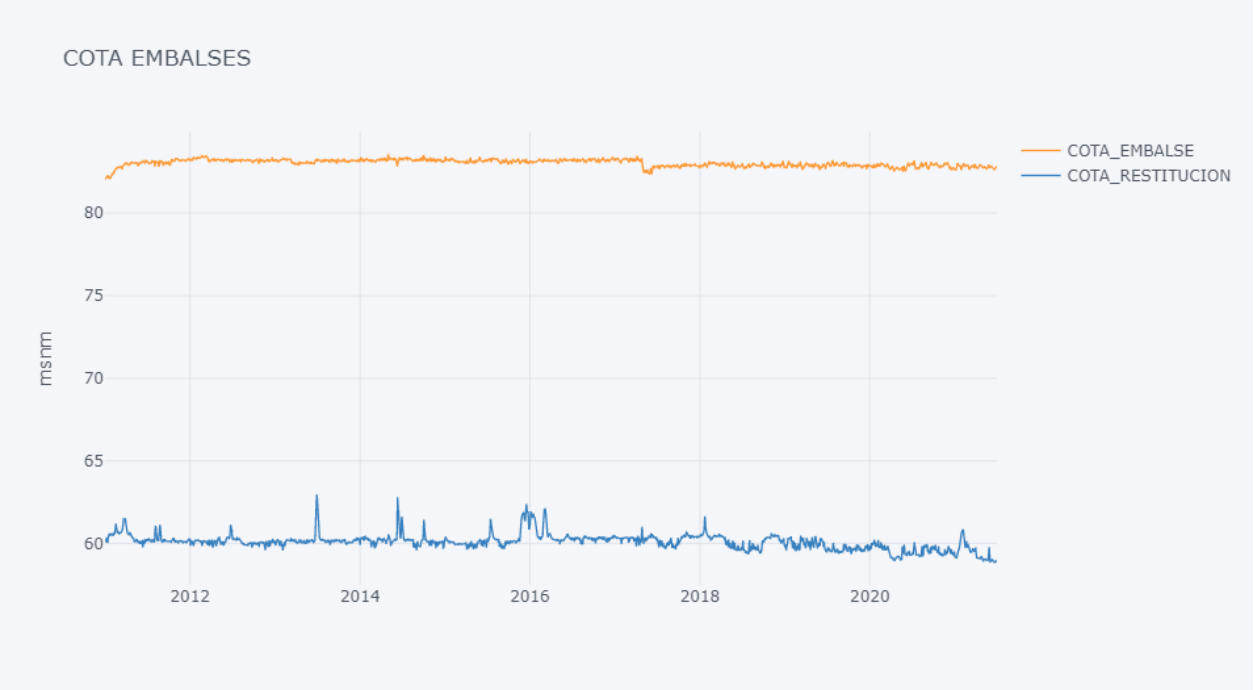

Ilustración 1: Cotas de embalse y de restitución

Se observa que ambas cotas tienen poca variación relativa a lo largo de la muestra de diez años. El cambio en la cota de embalse en abril de 2017 fue por un ajuste de la cota de embalse de trabajo. Las pequeñas variaciones en la cota de restitución son para mantener la de embalse con mayor regularidad (vertido).

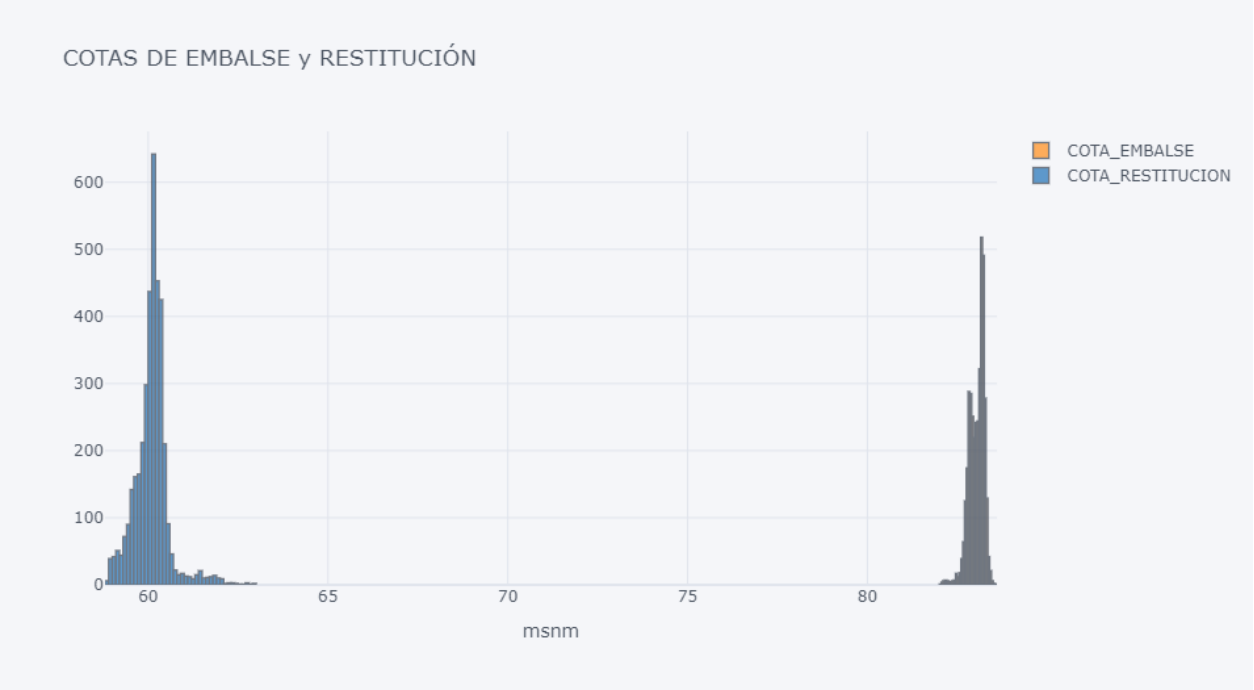

Ilustración 2: Cotas de embalse y restitución, histogramas

La cota de embalse se centra alrededor de los 83 metros sobre el nivel del mar y la de embalse de los 60 msnm.

Ilustración 3: Cotas de embalse y restitución, box-plots

Este comportamiento de las cotas se corresponde con las de una central de paso o pasada, tal cual se referenció en párrafo anterior.

Caudales: Afluente, Turbinado y Vertidos

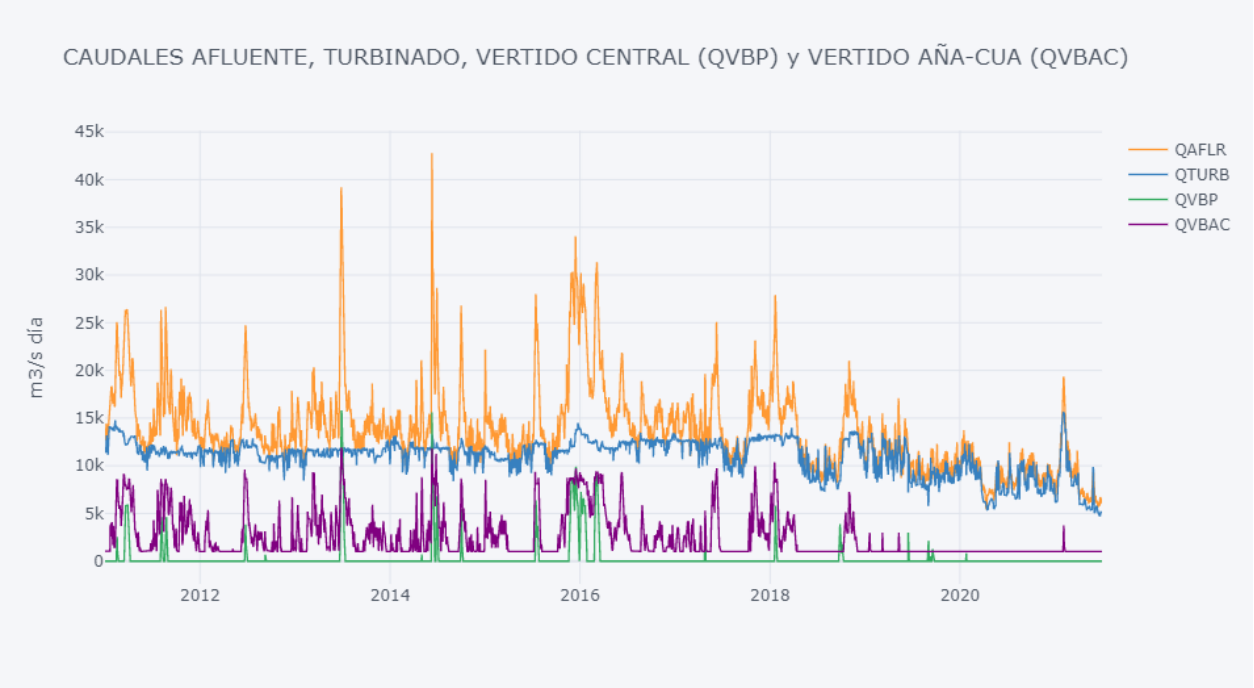

Ilustración 4: Caudales Afluente, Turbinado y Vertido

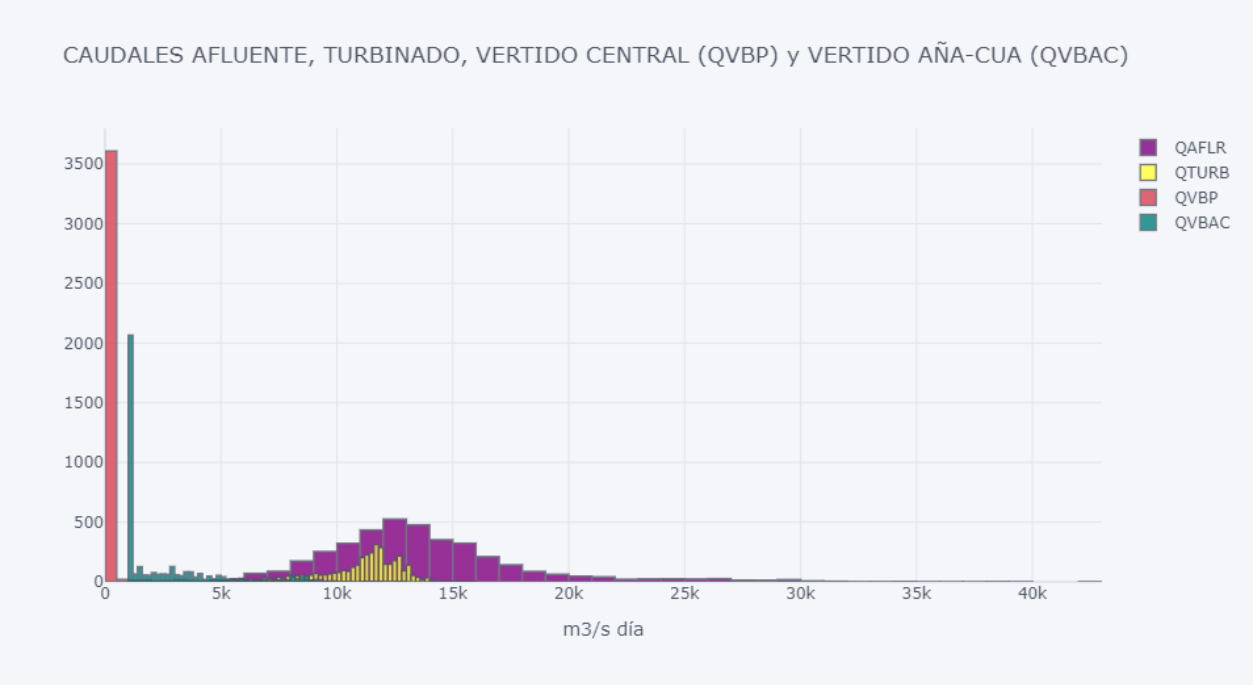

Ilustración 5: Caudales Afluente, Turbinado y Vertido, histogramas

Puede verse en estos dos gráficos la relación entre los caudades afluente y los turbinados y vertidos y corroborar una vez más que, por su diseño (planta de paso), la central no almacena agua más allá de sus cotas de diseño máximas y mínimas, de embalse y de restitución.

En consecuencia, el caudal afluente al embalse se turbina para la generación de energía o se vierte, sin pasar por las turbinas. Esto resulta en el mantenimiento de las cotas de ambos embalses. En otras palabras: Caudal Afluente = Caudal Turbinado + Caudales Vertidos en todo momento.

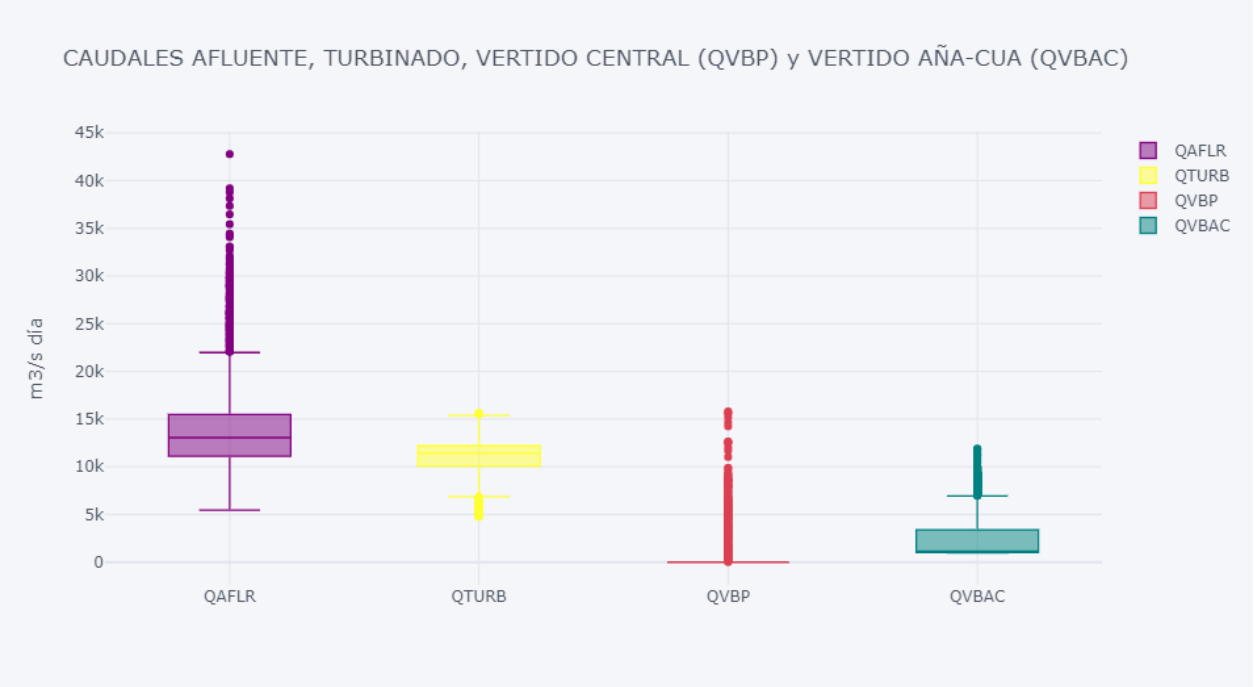

Ilustración 6: caudales Afluente, Turbinado y Vertido, box-plots

Se observa también que el vertedero de Aña-Cuá (QVBAC) funciona regularmente y con bajos niveles de caudal, mientras que el de la central (QVBP) solo eventualmente y con grandes niveles de caudal.

Energía generada

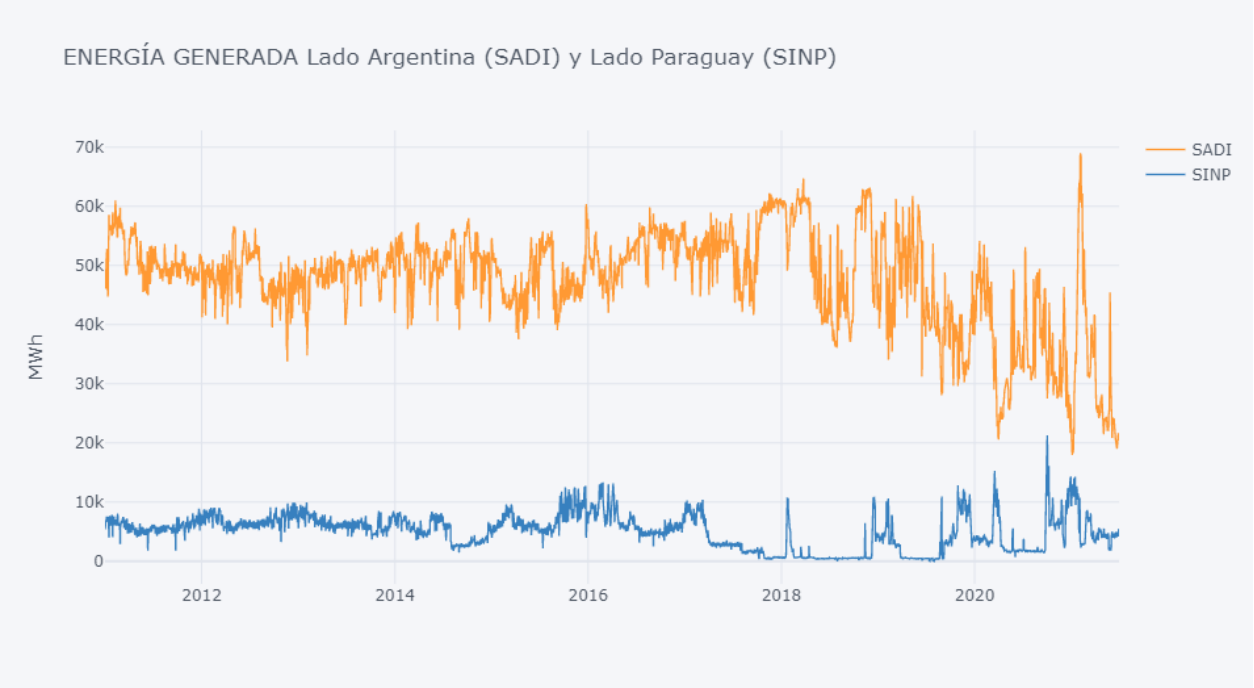

Ilustración 7: Energía generada

La CHY es una entidad binacional y por lo tanto genera energía, de manera ininterrumpida, para los sistemas interconectados de ambos países. Cuenta con 20 turbinas del tipo Kaplan, de las cuales 17 generan para el Sistema Argentino de Interconexión (SADI) y 3 para el Sistema Interconectado Nacional de Paraguay (SINP), con capacidad de ajustar esta relación en caso de necesidad, según la demanda de cada mercado eléctrico.

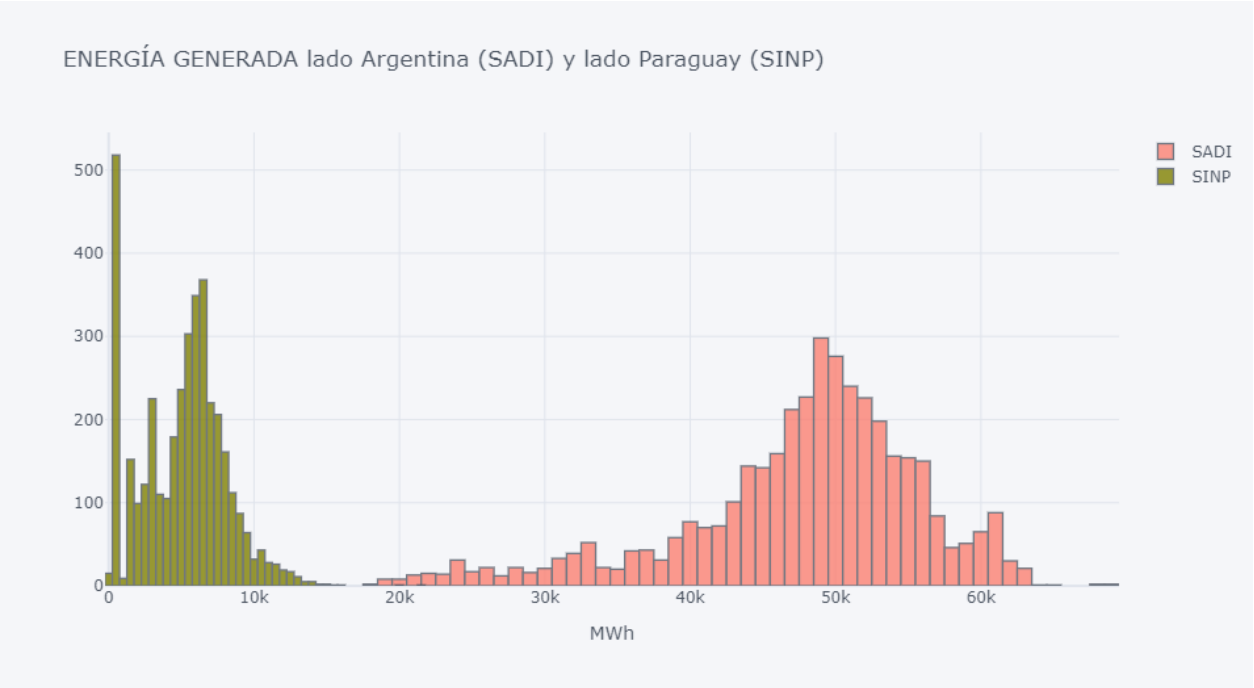

Ilustración 8: Energía generada, histograma

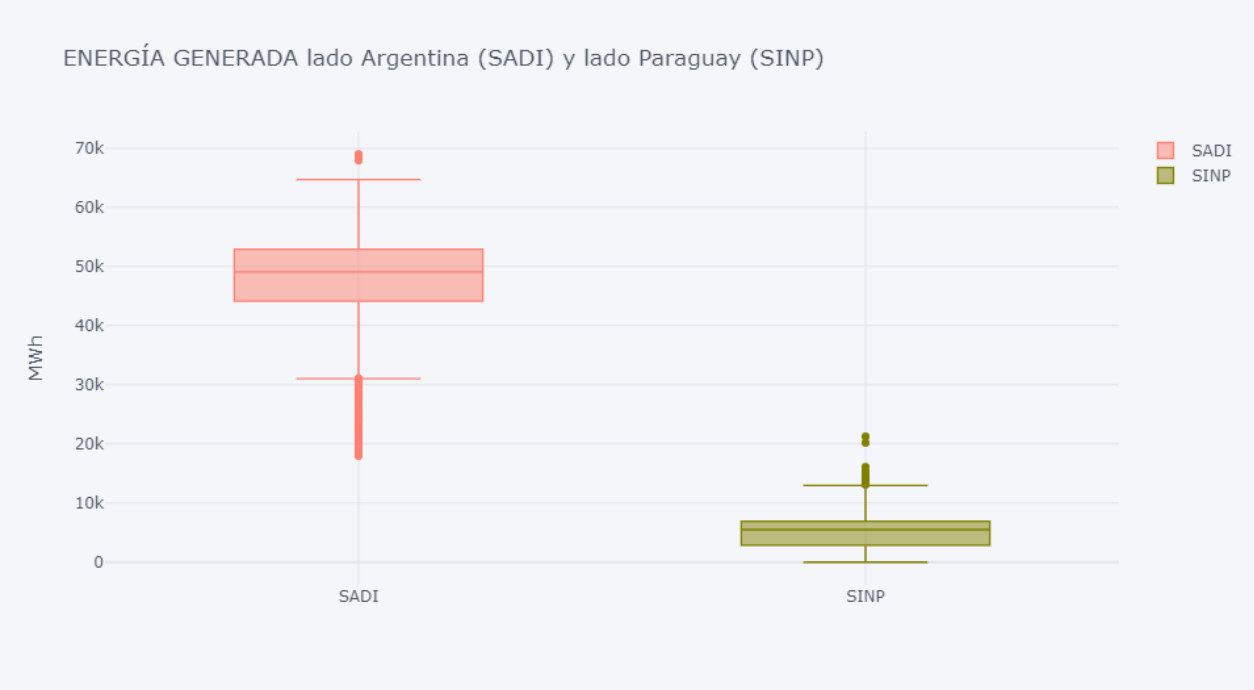

Ilustración 9: Energía generada, box-plots

Se observa una mayor dispersión en la generación hacia el SADI que hacia el SINP.

Análisis cruzado de variables

Se utilizará un diagrama de pares, del tipo 'pairs plot', que permitirá ver la interrelación entre las variables del conjunto de datos, a través de su correlación de a dos.

Tal como se vio anteriormente, los niveles de embalse y de restitución, por diseño, se mantienen prácticamente constantes en el tiempo, y por lo tanto, su impacto en la generación de la energía es mínima y, para facilitar el análisis gráfico, se excluirán del siguiente diagrama (notar que en el notebook correspondiente puede verse el diagrama completo, con esas dos variables incluidas).

Adicionalmente: se unificaron las variables de energía y caudales vertidos, a saber:

Energía = Energía SADI + Energía SINP

Caudal Vertido = Caudal Vertido Central (QVBP) + Caudal Vertido Aña-Cuá (QVAC)

dado que sus comportamientos respecto de la generación de la energía son equivalentes.

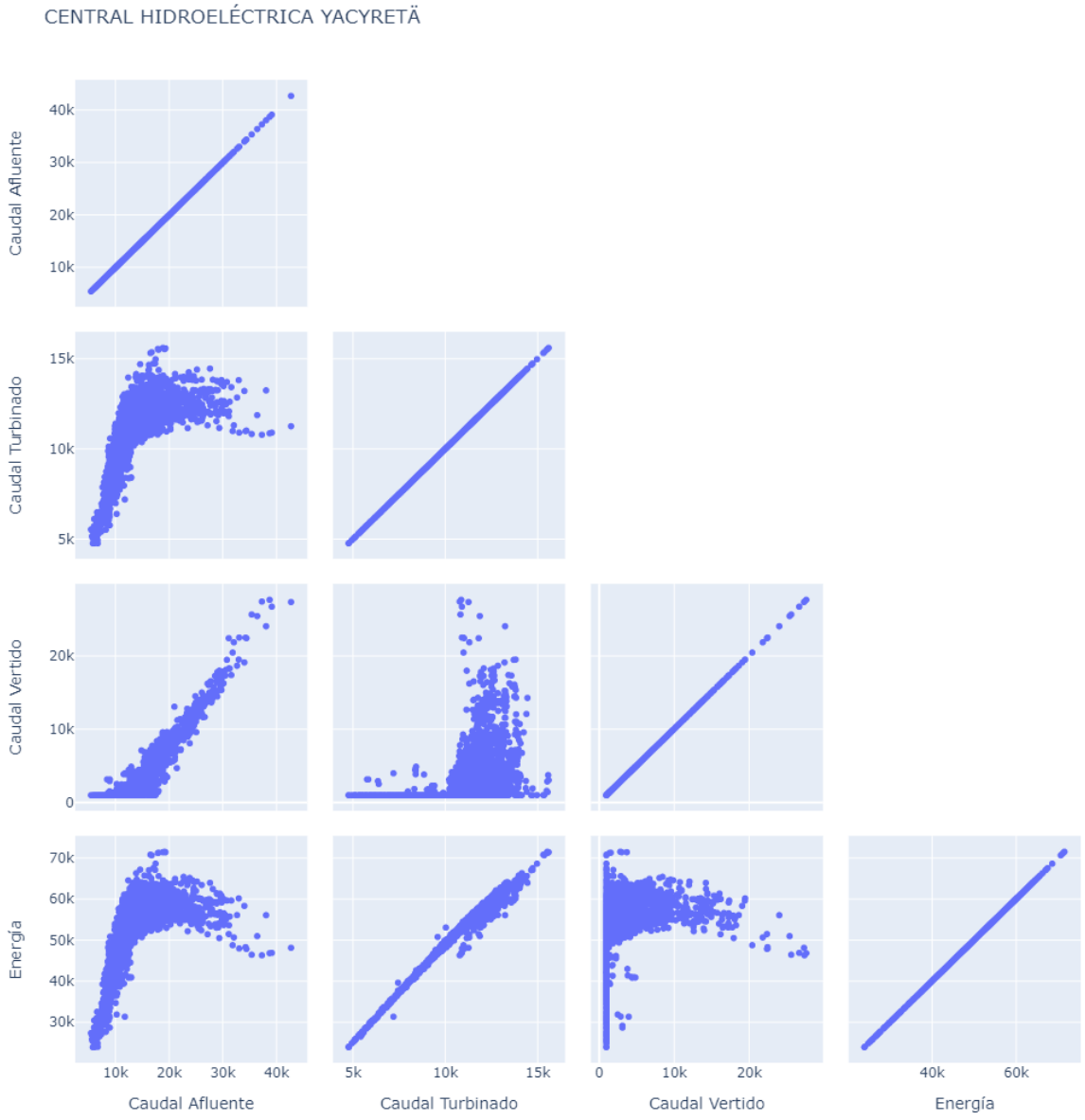

Ilustración 10: Diagrama de pair-plots de correlación

Se observa la fuerte correlación positiva entre el caudal turbinado y la energía generada, lo que es de esperarse, dado que, salvo variaciones en la eficiencia de las turbinas, esta relación es directamente proporcional.

Así mismo, la relación entre el caudal afluente y el caudal turbinado es también positiva, pero solo hasta el punto de máxima capacidad de generación de la central. Desde allí en adelante, el excedente de caudal afluente debe verterse para mantener el nivel de embalse.

También es interesante notar la relación entre caudal vertido y caudal turbinado. Aquí se ve que el vertido de caudal comienza cuando el caudal turbinado alcanza una saturación, sea ésta por capacidad de turbinado o por un límite fijado en dicho momento en la cantidad de energía a generar.

Se aprecia también una correlación positiva del caudal vertido y el caudal afluente, a partir del punto donde comienza a ser necesario el vertido de caudal, porque la central no puede, o no desea, generar más energía.

Nuevamente, hay que recordar que la CHY debe, por un tema de impacto ambiental en el ecosistema donde opera, mantener los niveles de embalse y restitución entre cotas máximas y mínimas establecidas, lo más constante en el tiempo posible. Por lo tanto, el caudal afluente que no se turbina, se debe verter.

Análisis de la serie de tiempo de interés (SALTO Hidráulico)

Dado que la capacidad de generar energía de la CHY depende de la diferencia entre las cotas de embalse y de restitución y que ésta es la variable de pronóstico para la gestión eficiente de la planta, se define una nueva variable en el conjunto de datos igual a la diferencia entre ambas cotas, llamada 'salto hidráulico', con nomenclatura 'SALTO' dentro del dataset:

SALTO = COTA_EMBALSE – COTA_RESTITUCION,

siendo ésta la variable focal en el desarrollo del trabajo.

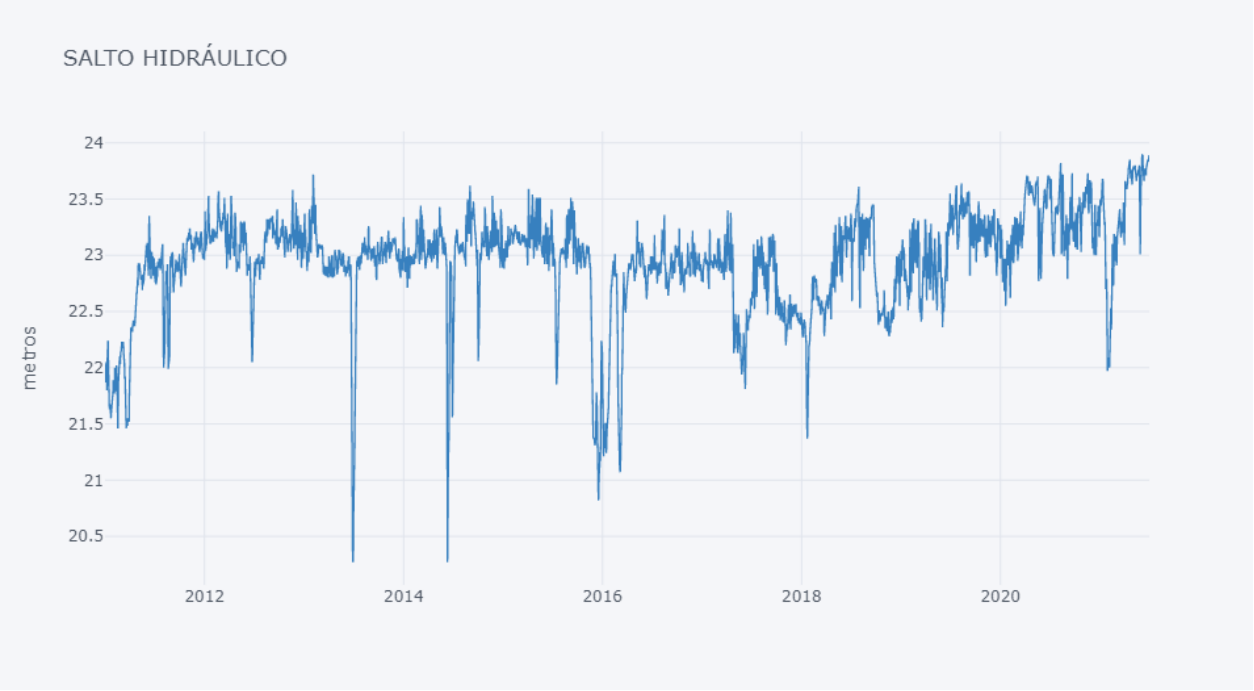

Ilustración 11: Salto Hidráulico

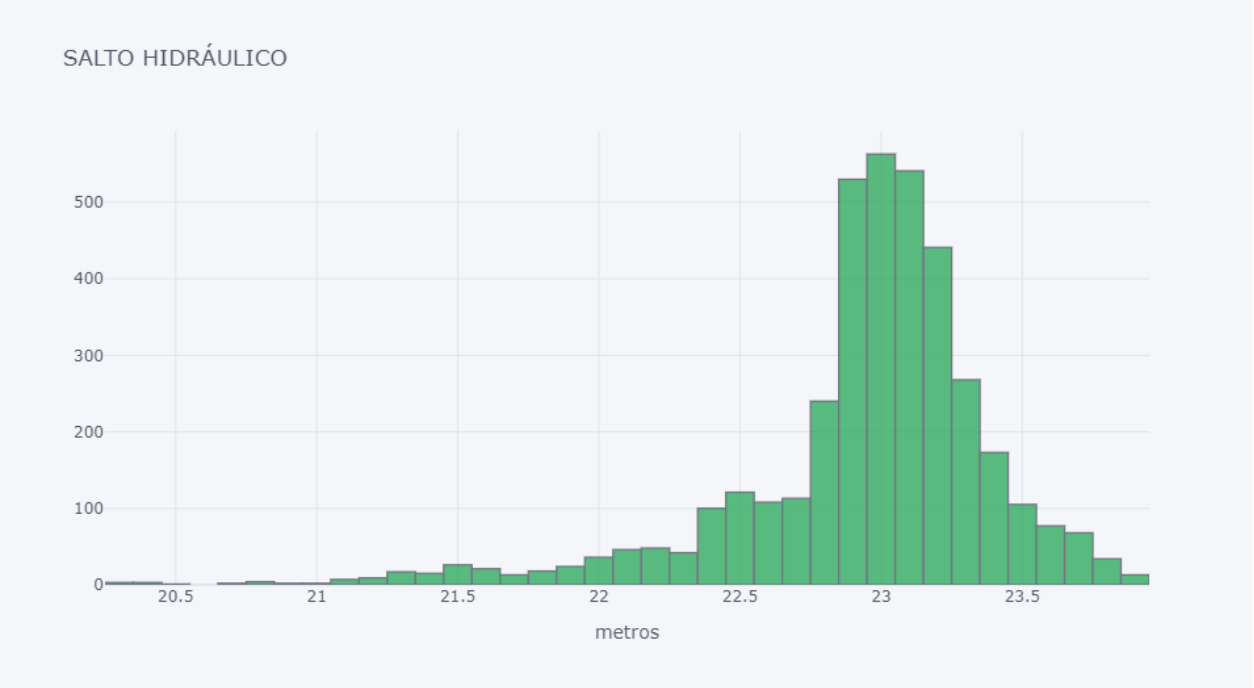

Ilustración 12: Salto Hidráulico, Histograma

Ilustración 13: Salto Hidráulico, box-plot

Puede verse de los diagramas anteriores, que el valor del salto hidráulico está alrededor de los 23 metros (correspondiente a la diferencia de los niveles de embalse y de restitución), con cuartiles, Q1: 22.89 metros, Q2: 23.01 metros y Q3: 23.19 metros y que no experimenta variaciones estacionales importantes a lo largo de los diez años de la muestra.

Los valores que aparecen más allá de los bigotes inferior y superior del diagrama de boxplot fueron validados con el especialista de dominio, justificados y tomados como verdaderos.

Adicionalmente puede verse en el histograma que los valores más allá de los bigotes (atípicos) son muy poco frecuentes comparados con el central (~15 vs. 663).

Se incluye a continuación esta variable en el análisis de pares anterior, resultando el siguiente diagrama:

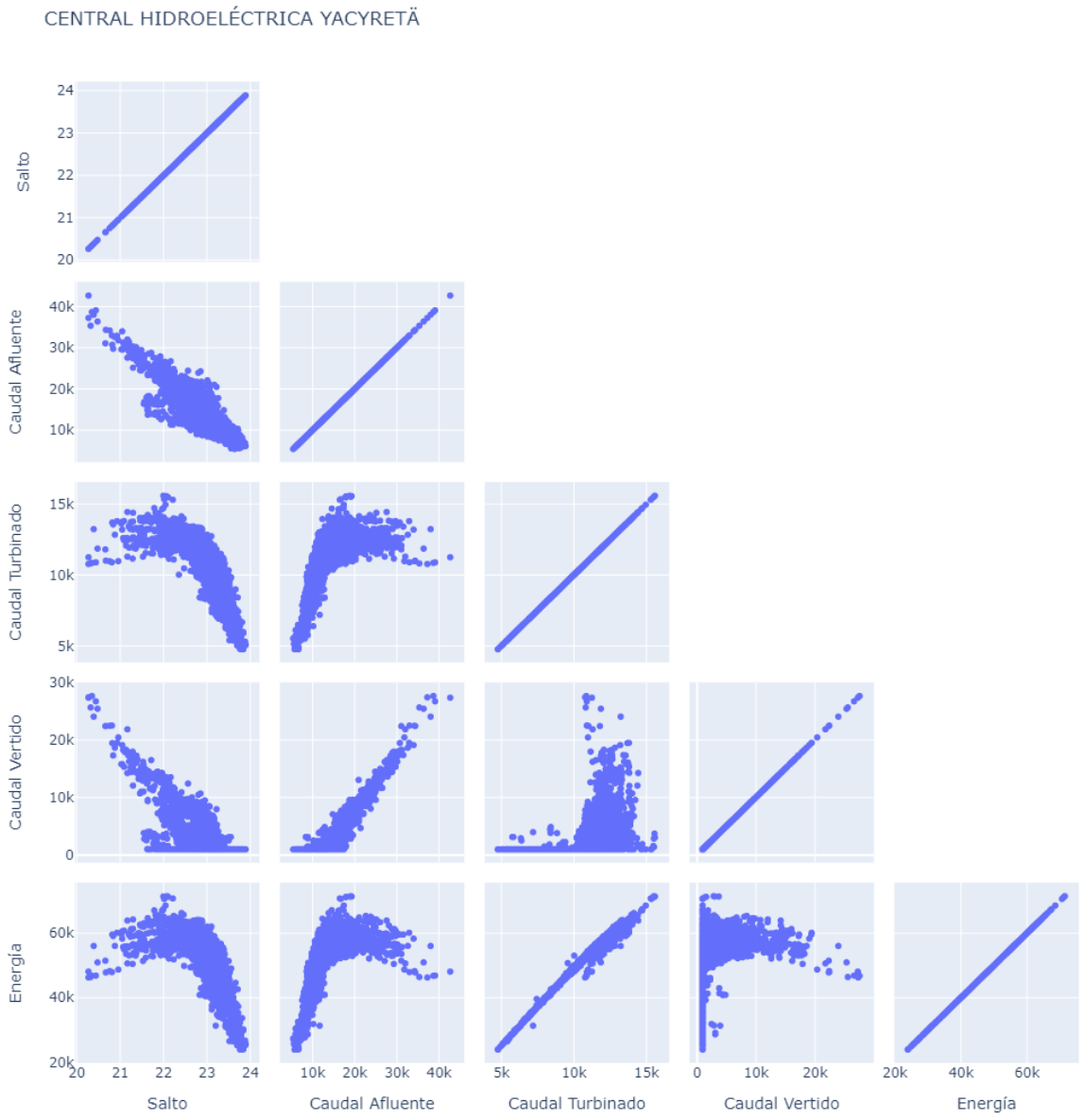

Ilustración 14: Diagrama pair-plot de correlación, incluyendo SALTO

Y uno más específico mostrando la interacción entre el salto, la energía generada y el caudal vertido.

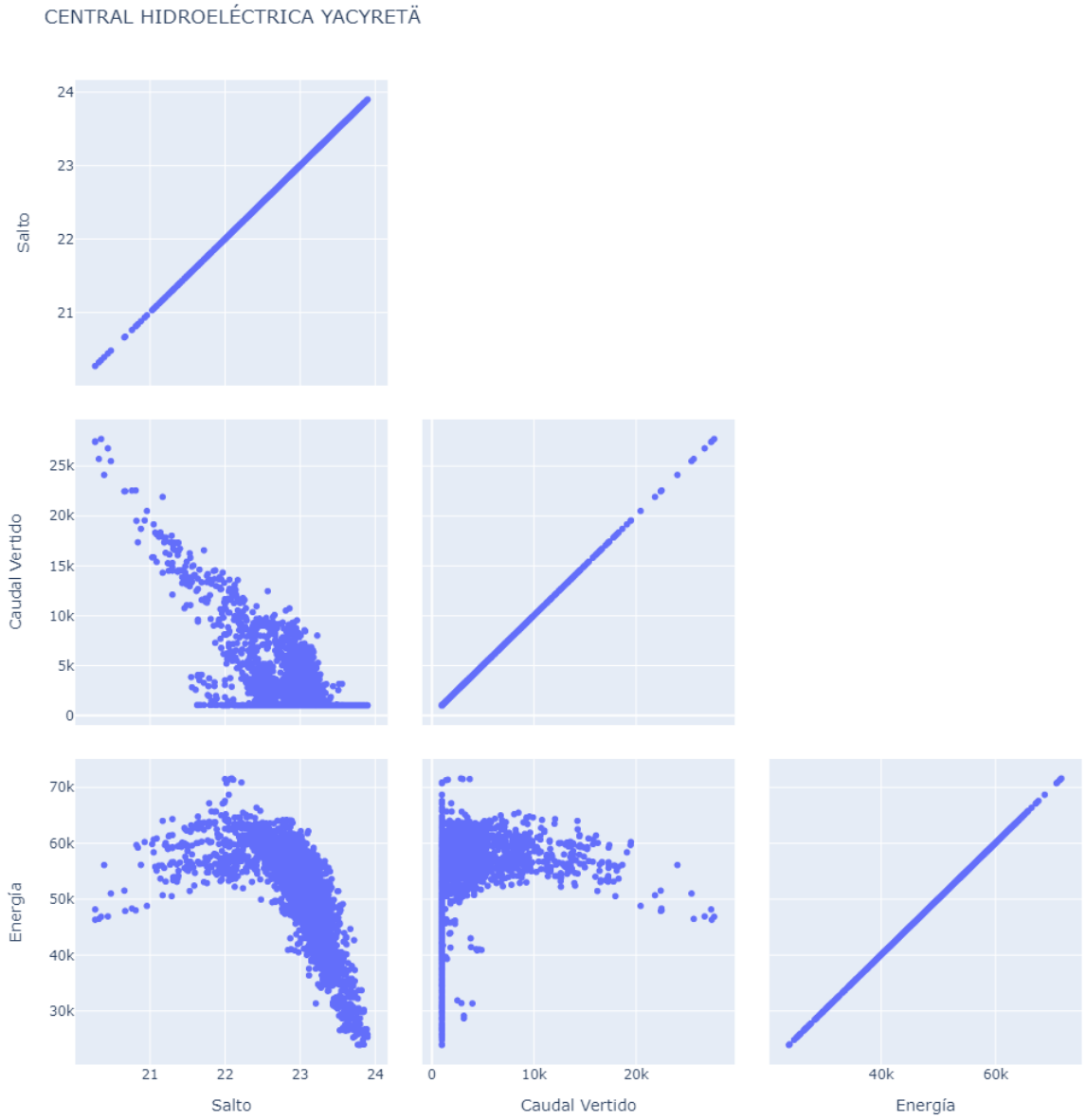

Ilustración 15: Diagrama pairs-plot SALTO, Caudal Vertido y Energía Generada

Notar que dadas las limitaciones (ya mencionadas anteriormente) en la capacidad máxima de generación y en la variación de los niveles de embalse y restitución (cotas máximas y mínimas), la relación entre el salto hidráulico y la energía generada es positiva para bajos niveles del salto, pero negativa para valores altos (gráfico inferior izquierdo).

Verificación de comportamiento 'White Noise' y 'Random Walk'

Antes de comenzar con los esfuerzos de pronóstico, es conveniente verificar si la serie de tiempo de interés, salto hidráulico (SALTO) es 'ruido blanco' (white noise) o se comporta como un 'paseo aleatorio' (random walk). Si así fuera, las chances de pronosticar con exactitud futuros valores serían muy bajas

cuando no nulas (How to Detect Random Walk and White Noise in Time Series Forecasting | by Bex T. | Towards Data Science).

1.1.1. Ruido blanco

Una serie numérica ruido blanco se caracteriza por:

- Tener media cero
- Varianza / desviación estándar constante en el tiempo
- Autocorrelación cero con todas sus versiones retrasadas ('lags')

Del gráfico en el tiempo del encabezado 13 se claramente que la media de la serie no es nula y que la varianza no es constante en el tiempo.

Respecto de la autocorrelación, en el siguiente gráfico se ve que los valores de la serie son fuertemente dependientes de los valores anteriores.

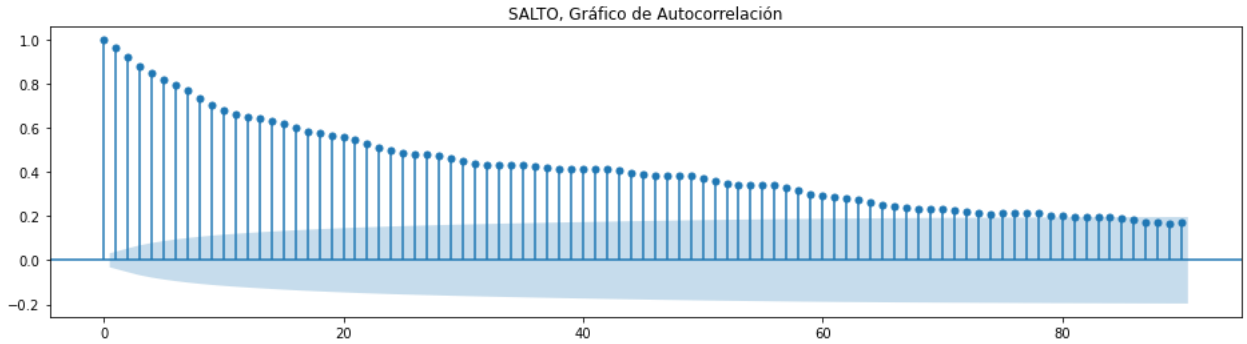

Ilustración 16: Diagrama de Autocorrelación de SALTO

Por lo tanto, se concluye que la serie temporal SALTO, no es un ruido blanco.

1.1.2. Paseo aleatorio (random – walk)

Al contrario del ruido blanco las series 'random walk' tienen media distinta de cero, varianza variable en el tiempo y valores de autocorrelación estadísticamente significativos, ya que los estados actuales resultan de la suma del estado anterior con un valor aleatorio:

Vt = Vt-1 + ruido blanco.

Por esta razón, diferenciar la serie aislaría el agregado de ruido blanco cancelaría el efecto en cada paso y si el resultado obtenido es efectivamente ruido blanco, entonces podríamos asegurar que estamos tratando con un 'paseo aleatorio'.

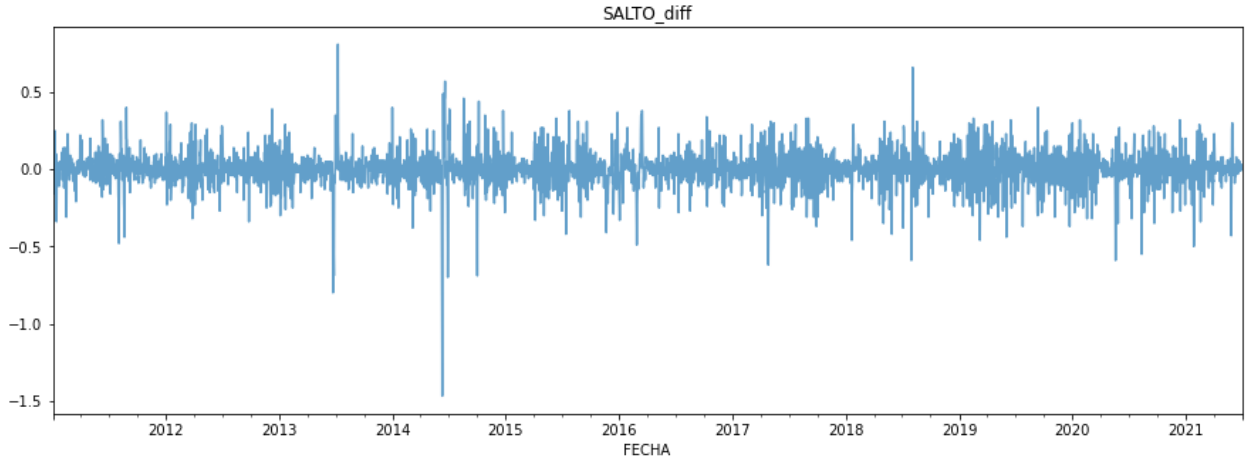

Ilustración 17: SALTO diferenciado (1 vez)

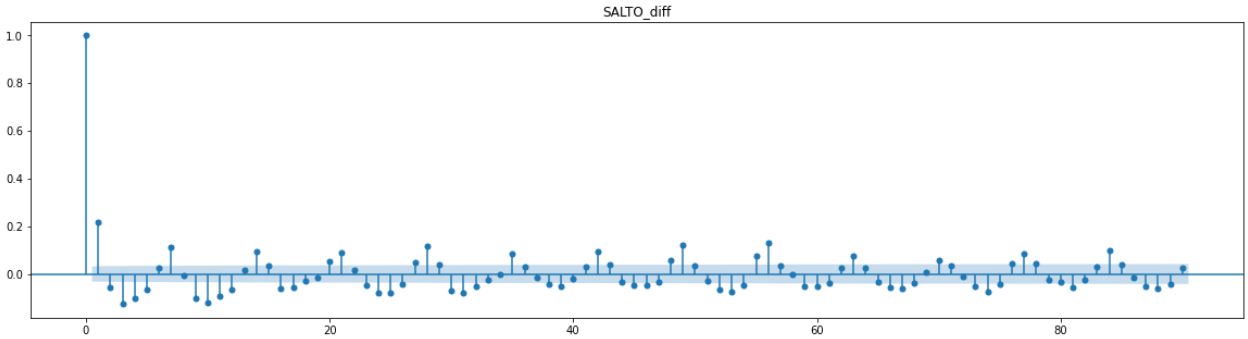

Ilustración 18: Diagrama de autocorrelación de SALTO diferenciado

Vemos que ahora la significación estadística de los valores de autocorrelación obtenidos es mínima y quedan dudas sobre el resultado.

En consecuencia, vamos a ensayar una prueba estadística para la determinación, test de Dickey-Fuller (1979 by Dicker D. A. and Fuller W. A), de la librería 'statsmodels', donde la hipótesis nula, H0, asume que la serie SALTO es un 'random walk' y la alternativa, H1, que no lo es.

El resultado es:

ADF Statistic: -5.602245800978897

p-value: 1.442260950188245e-05

El valor de p-value es mucho menor que el umbral de desición de 0.05, por lo tanto rechazamos la hipótesis nula HO y concluimos que SALTO no es un paseo aleatorio ('random-walk').

Concluimos en esta sección que SALTO no es ruido blanco ni paseo aleatorio.

Descomposición de la serie temporal SALTO

Continuando con el análisis vamos a descomponer la serie salto en un modelo aditivo de tendencia, estacionalidad y ruido. Esto nos permitirá, más adelante, utilizar mejores métodos econométricos para el pronóstico de valores futuros.

Para ello usaremos la función 'sm.tsa.seasonal_decompose' de la librería 'statsmodel' (https://www.statsmodels.org/)

Obtenemos:

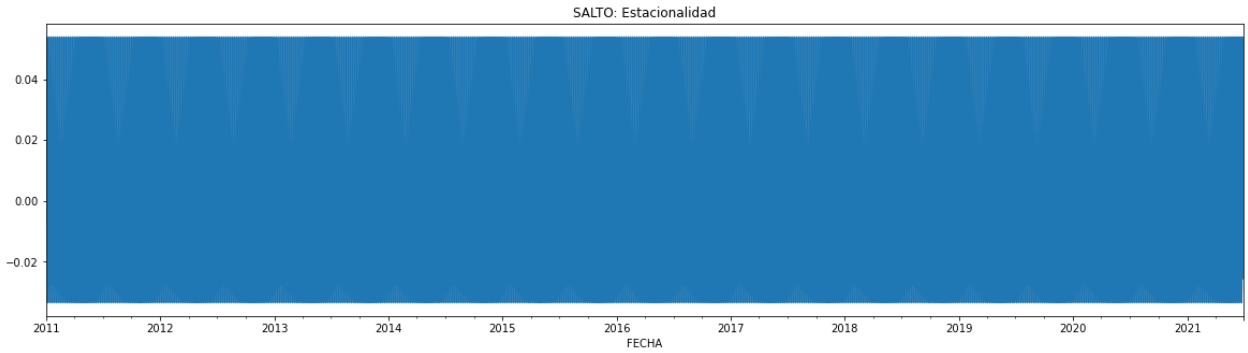

Ilustración 19: Estacionalidad de SALTO

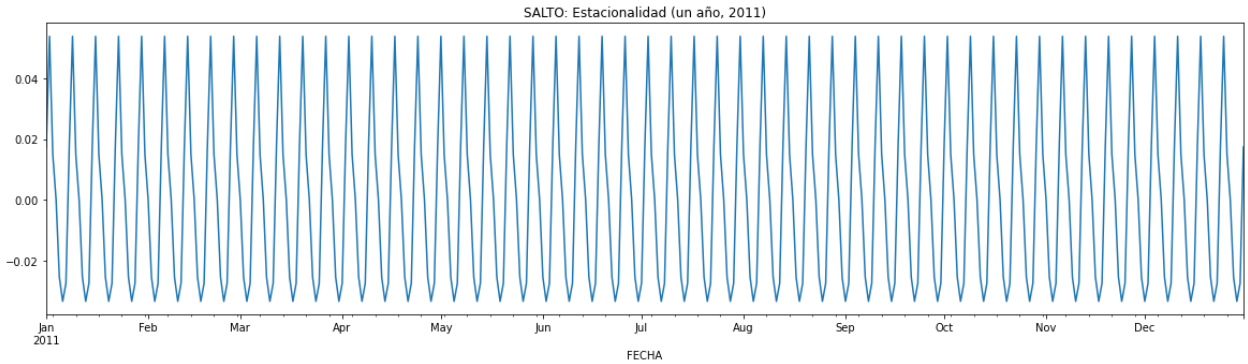

Ilustración 20: Estacionalidad de un mes de SALTO

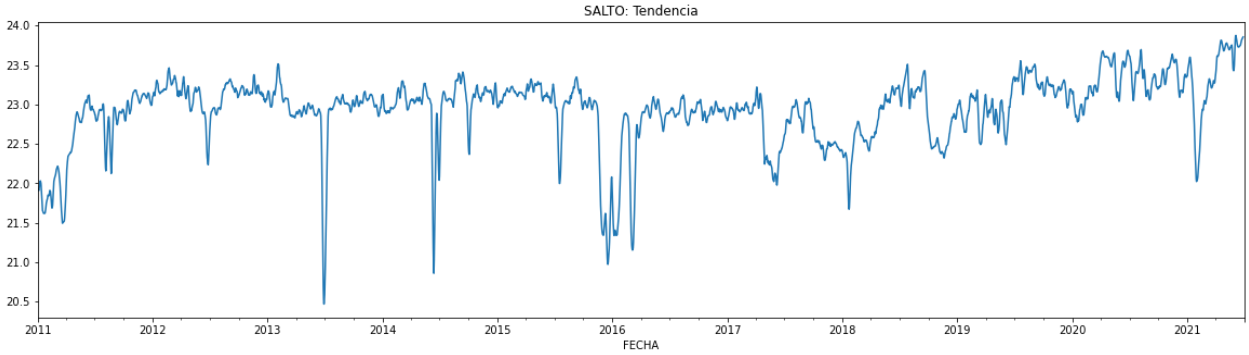

Ilustración 21: Tendencia de SALTO

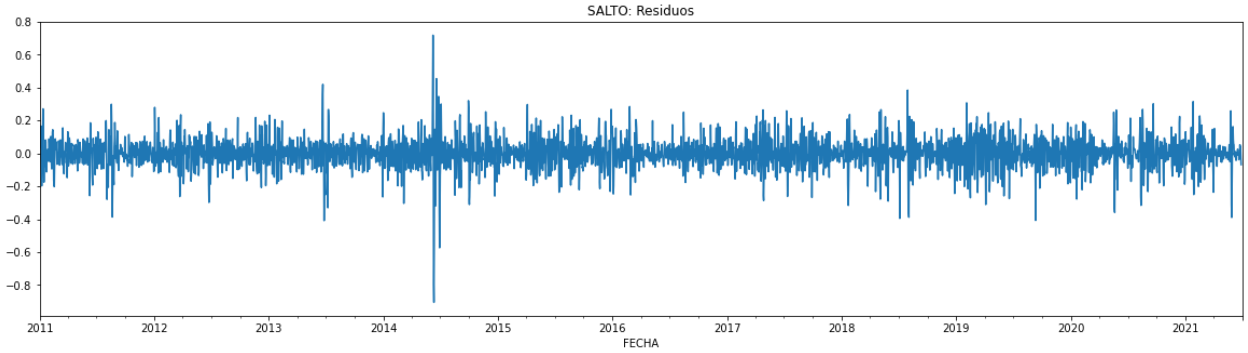

Ilustración 22: Residuos de SALTO

Estacionalidad: Se observa en el gráfico anual que la serie SALTO posee una marcada estacionalidad semanal. Esto ocurre a lo largo de toda su extensión, pero es difícil evidenciarlo en el gráfico completo ya que la proximidad de las líneas de la curva lo transforman en un bloque ininteligible. Para gráficos de más períodos referirse al cuaderno en cuestión.

Tendencia: Por las condiciones propias del fenómeno que le da lugar a la serie SALTO, no existe una tendencia sostenida a lo largo de toda la serie, sino por algunos períodos breves. Si existiera, significaría que los niveles de embalse o restitución varían constantemente en una misma dirección lo que provocaría que se seque o se desborde la presa.

Residuos: Se ve una resultante con media cero. Graficamos la autocorrelación de los residuos para volver a verificar la condición de paseo aleatorio y vemos que, aunque muy cerca de cumplirla, hay retrasos ('lags') que tienen autocorrelación estadísticamente significativa (se ve claramente en los primeros cuatro), que nos invita a abordar el proceso de predicción de valores futuros.

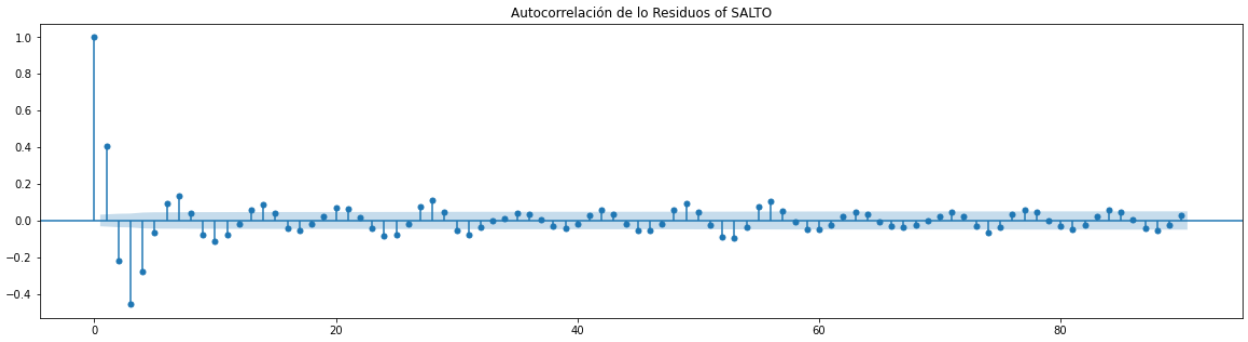

Ilustración 23: Autocorrelación de los residuos de SALTO