

Informe de Resultados Sprint 2
Proyecto CHEC



Introducción

El presente informe corresponde al segundo sprint del proyecto “Sistema inteligente de consolidación de datos oficiales a partir de las medidas análogas registradas por CHEC-EPM”, con el objetivo de generar modelamientos basados en información válida y confiable. Durante este sprint, se han establecido los criterios y las métricas necesarias para evaluar la calidad de los datos.

Objetivos

- Definir y documentar las métricas de calidad de los datos para cada una de las variables relevantes, asegurando que sean comprensibles y aplicables en el contexto del proyecto.
- Realizar un análisis preliminar de las fuentes de información disponibles, evaluando su confiabilidad y viabilidad para ser incluidas en los conjuntos de datos desde el año 2019 hasta el 2023, así como los datos actuales.
- Generar un informe detallado de calidad de los datos, que proporcione una visión integral de la confiabilidad de las fuentes de datos y la viabilidad de su uso en los modelamientos previstos, considerando tanto el período histórico como los datos más recientes.
- Construcción de un tablero de power BI en el cual se definan las respectivas métricas para calidad de los datos.

Descripción y Acceso a las Fuentes de Datos

Teniendo en cuenta que, en la finalización del primer sprint, se decidió cambiar las fuentes de datos de las cuales se extraen los datos para los sistemas SCADA (ABB y SURVALENT), se hizo necesario realizar nuevamente un perfilamiento de datos perdidos para estas nuevas fuentes de datos, tal y como se hizo en el primer sprint.

- **DM_OPERACION** : Esta base de datos contiene la tabla MEDIDAS_ANALOGAS, la cual contiene información acerca de los registros de las fuentes de medida del sistema SCADA ABB, el cual se encarga de registrar los datos de las medidas correspondientes a algunos circuitos que conectan con los canales primarios de los transformadores, con un periodo de muestreo de 15 minutos para cada uno de los circuitos. A continuación se presenta una breve descripción de cada una de las variables de interés contenidas en la respectiva tabla:

Variable	Descripción
CIRCUITO	Código identificador de cada uno de los circuitos del sistema
TIEMPO_AJUSTADO	Fecha y hora en la cual se dio el registro de las mediciones para el correspondiente circuito
IA , IB , IC	Corriente en cada una de las fases del circuito en Amperios (A)
VA , VB , VC	Voltajes en cada una de las fases del circuito en kiloVoltios (kV)
P	Corresponde a la potencia activa del circuito en MegaVatios (MW)
Q	Corresponde a la potencia reactiva del circuito en MegaVatios (MW)
SCADA	Indica en que sistema escada fue realizada la medición (no fue incluida en el estudio preliminar, pero puede ser de interés mas adelante)
FECHA_CARGA	Corresponde a la fecha en la cual fue subido el dato en el servidor (no fue incluida en el estudio preliminar, pero puede ser de interés mas adelante)

Tabla 1.1 (MEDIDAS_ANALOGAS_HORIZONTALES VARIABLES)

Las variables EAe, ERe, ERi y EAi no fueron de interes en este estudio; ya que no están relacionadas con el objetivo de este proyecto.

- Dentro de DM_OPERACIÓN, también se encuentra la tabla MEDIDAS_ANALOGAS_SURVALENT, la cual contiene información acerca de los registros de las fuentes de medida del sistema SCADA SURVALENT, el cual se encarga de registrar los datos de las medidas correspondientes a algunos circuitos que conectan con los canales primarios de los transformadores, con un periodo de muestreo de 15 minutos para cada uno de los circuitos. A continuación se presenta una breve descripción de cada una de las variables de interés contenidas en la respectiva tabla:

Variable	Descripción
CIRCUITO	Código identificador de cada uno de los circuitos del sistema

TIEMPO_AJUSTADO	Fecha y hora en la cual se dio el registro de las mediciones para el correspondiente circuito
IA , IB , IC	Corriente en cada una de las fases del circuito en Amperios (A)
VA , VB , VC	Voltajes en cada una de las fases del circuito en kiloVoltios (kV)
P	Corresponde a la potencia activa del circuito en MegaVatios (MW)
Q	Corresponde a la potencia reactiva del circuito en MegaVatios (MW)
SCADA	Indica en que sistema escada fue realizada la medición (no fue incluida en el estudio preliminar, pero puede ser de interés mas adelante)
FECHA_CARGA	Corresponde a la fecha en la cual fue subido el dato en el servidor (no fue incluida en el estudio preliminar, pero puede ser de interés mas adelante)

Tabla 1.2 (MEDIDAS_ANALOGAS_SURVALENT_H VARIABLES)

Las variables EAe, ERe, ERi y EAi no fueron de interes en este estudio; ya que no están relacionadas con el objetivo de este proyecto.

Observación: Es de destacar que en la actualidad, algunos circuitos se están mudando del sistema SCADA ABB al sistema SCADA SURVALENT y viceversa; por lo cual a la hora de desarrollar una estrategia de imputación de datos, se debe tener en cuenta que esto es dinámico.

Estudio Preliminar y Análisis de la Calidad de los Datos

Para iniciar el estudio, la idea base fue la revisión de la calidad de los datos, y comprobar si es viable generar un modelo de imputación para la reconstrucción de las series de tiempo, por lo tanto, el primer paso fue establecer el balance de datos perdidos que se tienen registrados en estos momentos con respecto a cada variable de interés.

- **SCADA ABB [63.029.296]**

Variable	Cantidad perdidos	% Perdidos
CIRCUITO	0	0%
TIEMPO_AJUSTADO	0	0%
IA	4.845.973	7,688%
IB	5.593.492	8,874%
IC	5.551.449	8,807%
VA	6.740.416	10,694%
VB	7.162.585	11,363%
VC	7.559.815	11,994%
P	7.124.025	11,302%
Q	7.091.349	11,250%
	Promedio	8,197%

Tabla 2.1 (MEDIDAS_ANALOGAS_HORIZONTALES DATOS PERDIDOS)

Como se puede ver tanto el porcentaje de datos perdidos para cada una de las variables como su promedio, está por debajo del 12%, por lo que se puede concluir que en el sistema SCADA ABB, hay datos suficientes para entrenar un modelo de machine learning que permita interpolar los datos perdidos en cada una de las series de tiempo de interés.

- **SCADA SURVALENT [30.191.692]**

Variable	Cantidad perdidos	% Perdidos
CIRCUITO	0	0%
TIEMPO_AJUSTADO	0	0%
IA	869.793	2,880%
IB	1.131.178	3,746%
IC	1.207.909	4,0007%
VA	1.953.444	6,470%
VB	1.931.577	6,397%

VC	1.862.818	6,169%
P	1.284.574	4,254%
Q	1.257.334	4,164%
	Promedio	3,808%

Tabla 2.2 (MEDIDAS_ANALOGAS_SURVALENT_H DATOS PERDIDOS)

. Como se puede ver tanto el porcentaje de datos perdidos para cada una de las variables como su promedio, está por debajo del 7%, por lo que se puede concluir que en el sistema SCADA SURVALENT, hay datos suficientes para entrenar un modelo de machine learning que permita interpolar los datos perdidos en cada una de las series de tiempo de interés.

Como se puede ver, En el sistema SCADA SURVALENT se pierden menos datos que en el sistema SCADA ABB; por lo que se puede concluir que la calidad de medida del primero es mejor.

Posteriormente se hizo un analisis para determinar que circuitos estan tanto en el sistema SCADA ABB como en el sistema PRIME, en el sistema SCADA SURVALENT como en el sistema PRIME y en el sistema SCADA ABB como en el sistema SCADA SURVALENT, para de esta manera saber, a que circuitos se le pueden imputar datos perdidos en el sistema SCADA ABB y/o SURVALENT, partiendo de los datos del PRIME, y de las respectivas relaciones de corrientes y voltajes de las fronteras de cada circuito. En la tabla que se muestra a continuación se tiene un resumen de este análisis:

Analisis	SCADA ABB	SCADA SURVALENT	PRIME	COMÚN ABB-PRIME	COMÚN SURVALENT- PRIME
GENERAL	708	481	207	189	52

Tabla 3.1 (Circuitos PRIME-SCADA)

Analisis	SCADA ABB	SCADA SURVALENT	COMÚN ABB- SURVALENT
GENERAL	708	481	180

Tabla 3.2 (Circuitos SCADA)

Como se puede ver en la tabla anterior, entre circuitos que están en SCADA y en SURVALENT, suman los 1008 circuitos que había en V_MEDIDAS_ANALOGAS (quitando los 180 duplicados que hay en ambas tablas). Mientras que los 189 circuitos entre SCADA ABB y PRIME se sigue manteniendo (pues ABB es el sistema más antiguo). Por otro lado, hay apenas 52 circuitos comunes entre PRIME y SURVALENT (pues no todos los circuitos del ABB se han pasado al SURVALENT).

Estudio sistemas Scada (Calidad de los datos Segundo Sprint)

Con las nuevas bases de datos suministradas para el sistema SCADA, y una vez finalizado el estudio general de las mismas, se entra a analizar lo que sería consistencia en los valores de ambos sistemas tanto ABB como SURVALENT, en esto se inicia observando la cantidad de circuitos presentes en ambos sistemas que ya se encuentra consagrada en la tabla anterior (3.2), por lo que se puede ver que hay 180 circuitos que permitirán analizar relaciones entre ambos sistemas y cual sería la ventaja de cada uno. Para esto se decidió empezar por ver los rangos de fechas en los cuales se encuentran registros de cada circuito en ABB y SURVALENT, consagrados en una tabla como sigue:

	Circuito	Fecha mínima abb	Fecha mínima survaley	Fecha maxima abb	Fecha maxima survaley
0	MLA23L12	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-11-27 17:00:00	2023-07-12 15:30:00
1	MLA23L13	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-11-27 17:00:00	2023-07-12 15:30:00
2	MLA23L14	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-11-27 17:00:00	2023-07-12 15:30:00
3	MLA23L15	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-11-27 17:00:00	2023-07-12 15:30:00
4	MLA30L14	2022-11-03 10:30:00	2021-01-01 00:00:00	2023-07-12 15:15:00	2023-07-12 15:30:00
5	MNA30L13	2020-03-16 13:45:00	2021-01-01 00:00:00	2021-04-19 15:00:00	2023-07-12 15:30:00
6	PRA23L12	2020-01-01 00:00:00	2021-01-01 00:00:00	2021-04-30 11:00:00	2023-07-12 15:30:00
7	PRA23L13	2020-01-01 00:00:00	2021-01-01 00:00:00	2021-04-30 11:00:00	2023-07-12 15:30:00
8	PUR40L19	2020-01-01 00:01:00	2021-01-01 00:00:00	2022-02-07 09:30:00	2023-07-12 15:30:00
9	VIR23L12	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-03-13 15:30:00	2023-07-12 15:30:00
10	VIR23L13	2020-01-01 00:00:00	2021-01-01 00:00:00	2020-03-13 15:30:00	2023-07-12 15:30:00
11	VIR23L14	2020-01-01 00:00:00	2021-03-29 16:45:00	2020-03-13 15:30:00	2023-07-12 15:30:00
12	VIR23T11	2020-01-01 00:00:00	2021-03-29 16:45:00	2020-03-13 15:30:00	2023-07-12 15:30:00
13	VIR30L15	2020-01-01 00:00:00	2021-03-29 16:45:00	2021-02-26 12:00:00	2023-07-12 15:30:00

Tabla 3.3 (Circuitos Rangos de fechas ABB-SURVALENT)

Teniendo los registros de estos rangos de fechas, se realizo un estudio para ver los circuitos que empezaban registro antes en SURVALENT o ABB, lo cual se ve en el siguiente diagrama:

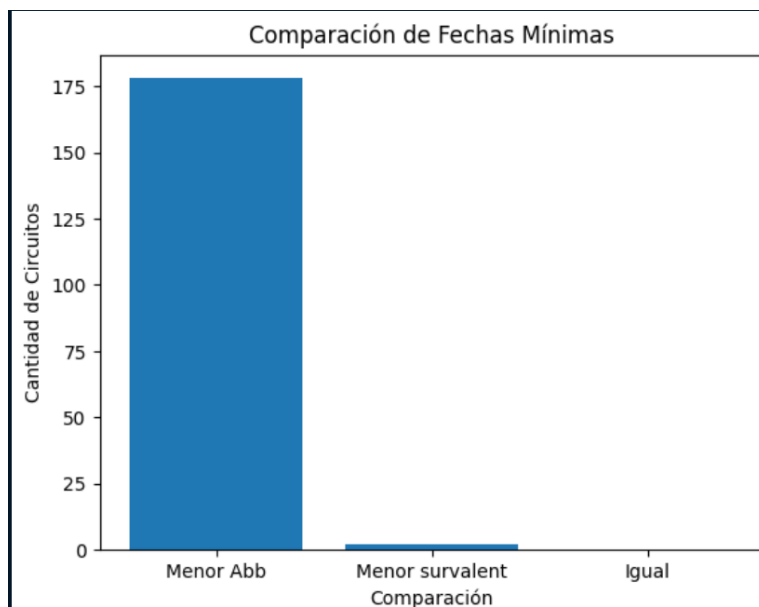


Tabla 3.4 (Fechas mínimas ABB-SURVALENT)

Aquí se puede observar, que el ABB es el sistema que mas tiempo lleva en la empresa, sin embargo hay unos pocos casos donde se encontró sistemas que iniciaban registros en el SURVALENT, definido en la siguiente tabla:

	SCADA ABB	SCADA SURVALENT	TOTAL
Cantidad	178	2	180

Tabla 3.5 (Circuitos SCADA)

Para el caso de los 2 circuitos que inician primero en el survalent podemos ver la información de los datos en la siguiente tabla:

CIRCUITO	ABB MINIMA	ABB MAXIMA	SURVALENT MINIMA	SURVALENT MAXIMA
MLA30L14	2022-11-03 10:30	2023-07-12 15:15	2021-01-01 00:00	2023-07-12 15:30
PUR40A18	2022-01-22 16:30	2022-02-07 09:30	2021-07-23 15:15	2023-07-12 15:30

Tabla 3.6 (SURVALENT - ABB)

Para el primer circuito podemos ver que a pesar de que se puede ver una migración de SURVALENT a ABB, presentan registros muy sincronizados para las fechas maximas, lo cual daría paso a interrogantes de porque se realizó esta migración y porque se mantienen registros en ambos. Para el segundo caso resulta ser una situación distinta puesto que a pesar de que se empezaron a registrar datos en ABB después de iniciar registros en el SURVALENT, se puede evidenciar que es el SURVALENT el que presenta registros más actualizados.

Ahora observando la tabla de fechas inicial (3.2) también se pudo ver que hay valores de fechas en el ABB y SURVALENT, que se solapan, es decir presentan registros en ambos, lo cual permite obtener un rango de tiempo, donde se puede estudiar las mediciones de ambos sistemas en igualdad de condiciones, por lo tanto, se pasa a filtrar la tabla de circuitos, ya que es un caso que no sucede en todos, ya que lo normal sería que al pasar de un sistema a otro más moderno como el SURVALENT, ya no llegarán registros del anterior, pero por el contrario, en

muchos circuitos (86), hay un rango de tiempo prudente donde se comparten mediciones, por ende la tarea consiste en estudiar el comportamiento en dicho margen de tiempo, reduciendolo al tiempo donde se dio el inicio de los registros del sistema mas nuevo (SURVALENT) a la ultima fecha donde se tienen registros del ABB (Claro esta siendo esta ultima mayor a la inicial del SURVALENT).

Tomando por ejemplo la variable IA, obteniamos las siguientes series temporales para algunos circuitos:

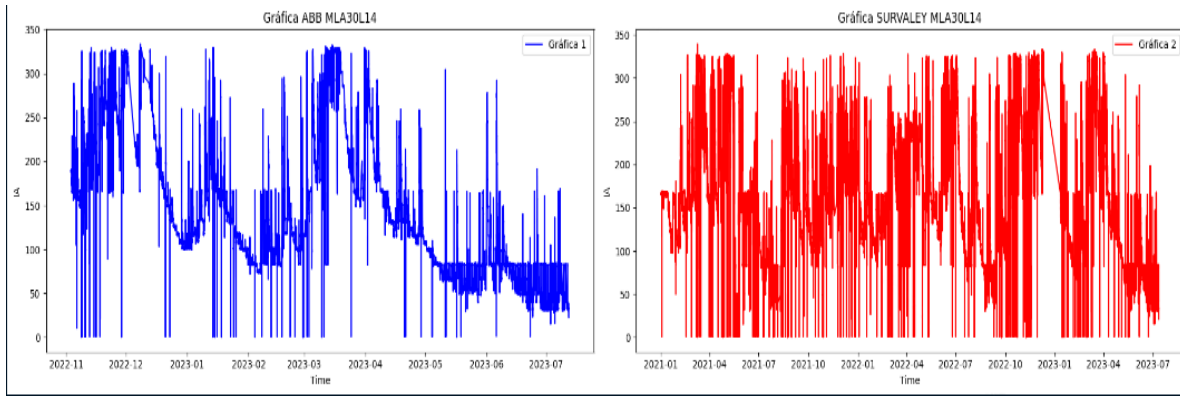


Imagen 1.1 (SURVALENT – ABB MLA30L14)

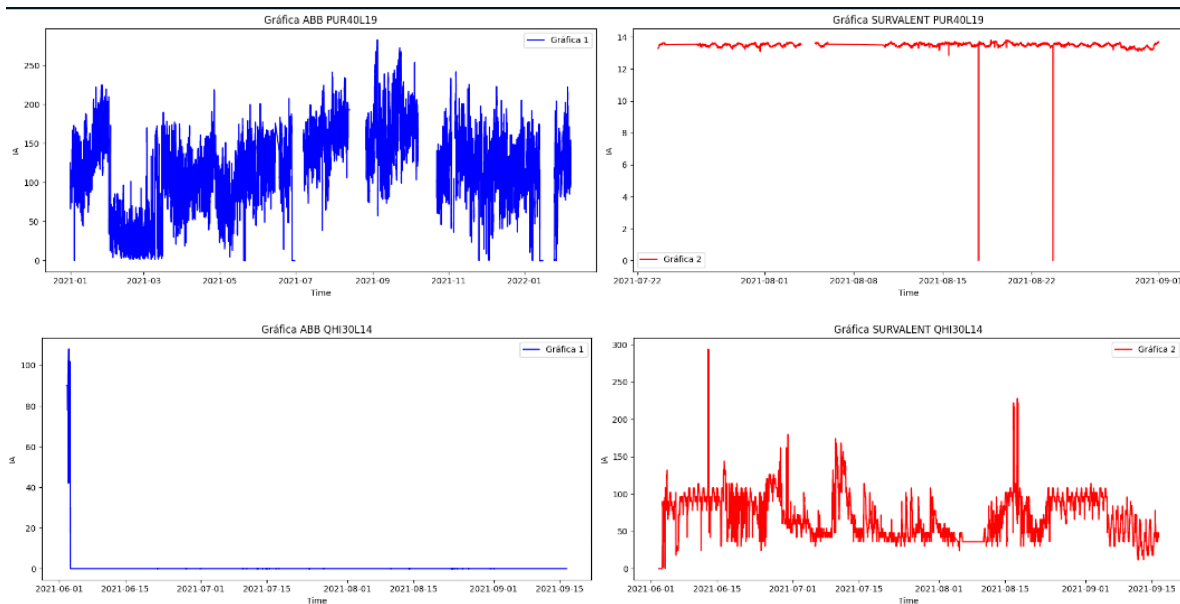


Imagen 1.2 (Series de tiempo IA ABB - SURVALENT)

Donde se puede evidenciar que a pesar de que se limito la serie de tiempo a un mismo rango, teniendo en cuenta que en estos circuitos para la mayoría iniciaron primero en el ABB y se tomo como fecha inicial la del SURVALENT, exceptuando el

caso del circuito MLA30L14, las bases de tiempo no estan sincronizadas y por otro lado hay datos que se pierden, puesto que los arreglos de tiempo no tienen los mismo tamaños. Por lo tanto para los 83 circuitos que pueden analizarse en ambos tenemos para este momento los siguientes diagramas:

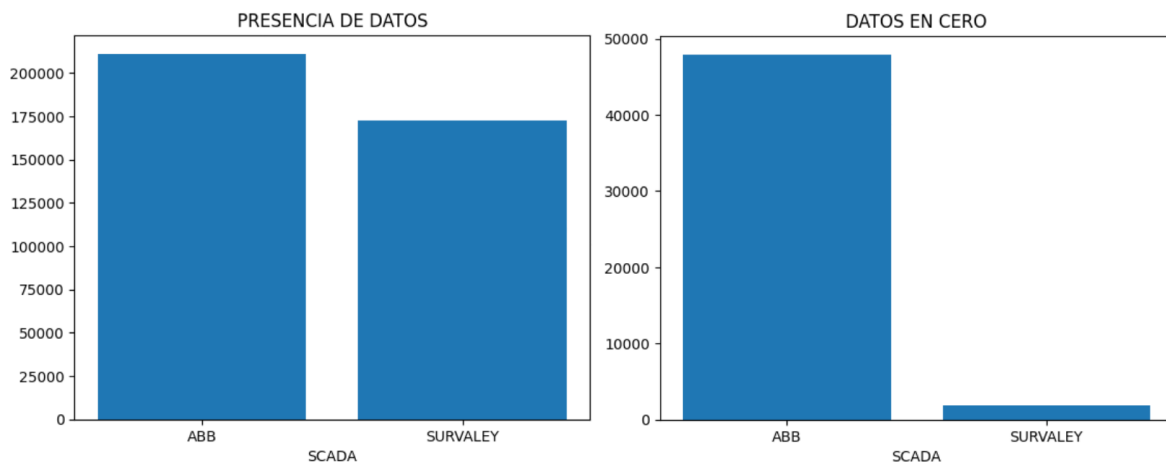
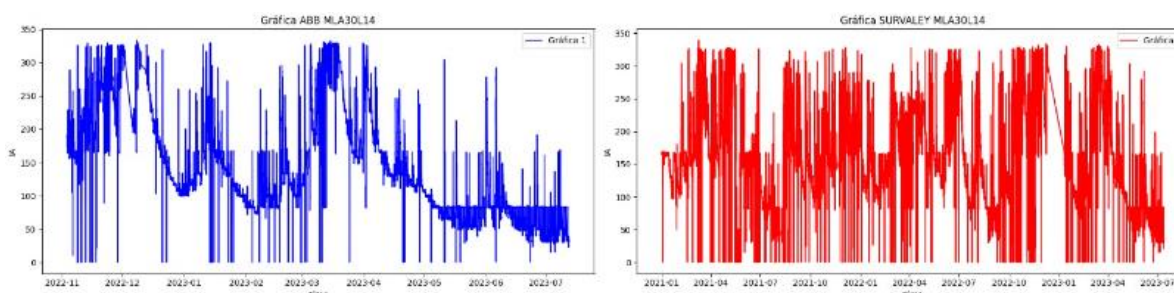
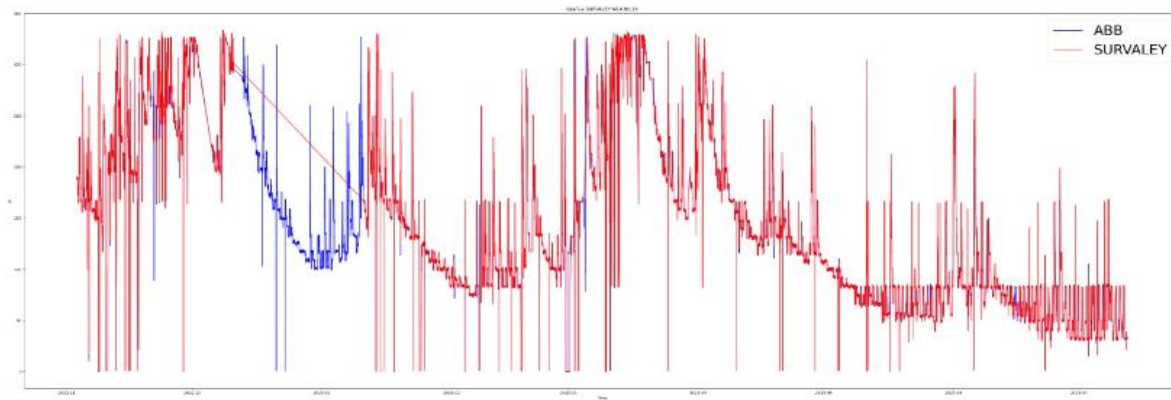


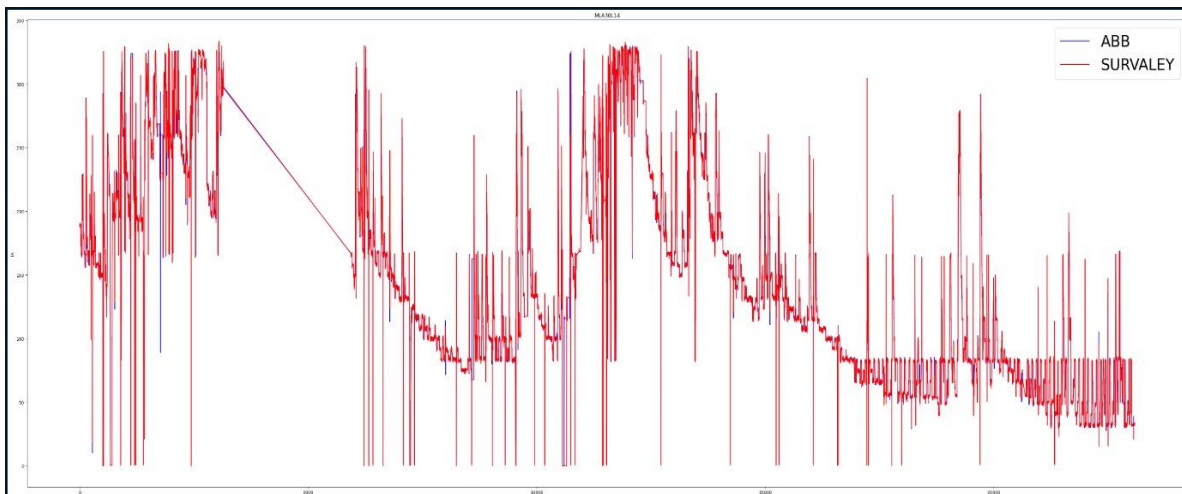
Imagen 1.3 (Diagramas de barras IA ABB - SURVALENT)

Ahora redefiniendo los rangos a medidas mas exactas, para poder superponer las series, tomando como ejemplo el caso MLA30L14:





Podemos ver que para el caso de MLA30L14 hay mayor presencia de datos en ABB, y con poca presencia de datos nulos, lo cual invita a dudar en la calidad de los nuevos sistemas, sin embargo faltaria estudiar los demas circuitos que veremos mas adelante, ahora solo tomando registros presentes en ambos sistemas en dicho rango de tiempo obtenemos:



Se pueden obtener medidas de correlación y mse para determinar relaciones y coincidencias entre los sistemas:

Correlación: 0.99
Error Cuadrático Medio (MSE): 67.39

Ahora este es el caso con mayor cantidad de registros en ambos sistemas es decir es aquel que nos proporciona mayor cantidad de información sobre la coherencia de los registros, mas sin embargo ya teniendo el ejercicio para este circuito podemos generalizarlo obteniendo algunas graficas como sigue identificando 3 casos:

- **Caso 1:** podemos observar una situación donde a pesar de que habíamos llegado a una conclusión parcial de la eficiencia en cuanto cantidad de registros para este análisis particular en el ABB, para la mayoría de los registros, parecen tender a valores constantes para este sistema, por ende resulta en una clara falla en la medición, mientras que en el SURVALENT, hay una coherencia mejor, puesto que los datos tienden a variar en la serie, por ende pueden aportar mejor información, ya después entraríamos a un análisis de atípicos y es la situación que más se presenta.
- **Caso 2:** Se presenta mayor aporte en el ABB, ya que el SURVALENT tiende a ser constante a pesar de que es un caso raro, significaría que hay una falencia en la medición en SURVALENT para dichos circuitos, y por ende no se podría usar como base para un modelo de imputación.
- **Caso 3:** Es uno de los pocos casos donde podemos ver una lógica de comportamiento al cambiar de sistema SCADA, es decir se denota un proceso en el cual la medición en SURVALENT es cero, mientras que en ABB si esta presente y llega un punto en el que empiezan a llegar registros al SURVALENT y ABB para ser constante, por lo cual surge una interrogante que es ¿Por qué en ABB al empezar registros en el SURVALENT no tienen valores nulos o cero, ya que la medición no sería relevante, ya teniendo el sistema moderno?, y además ¿Por qué no se detiene los registros en ABB una vez pasados al SURVALENT, exactamente que está midiendo?.

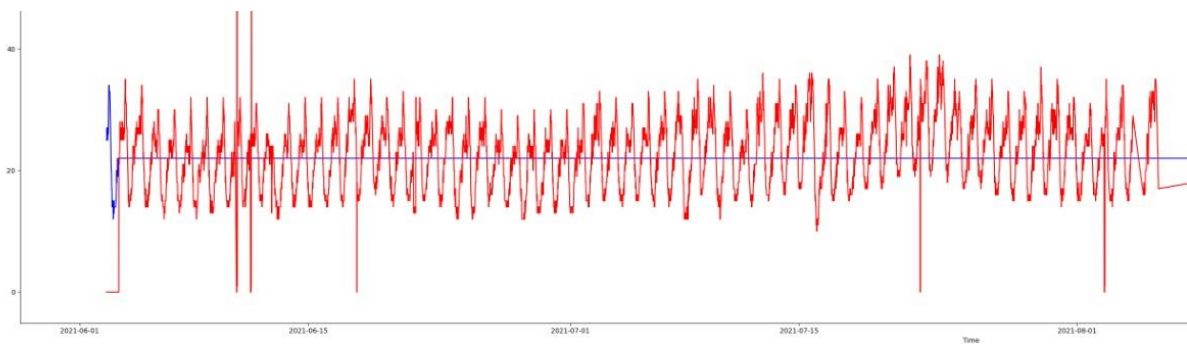


Imagen caso 1

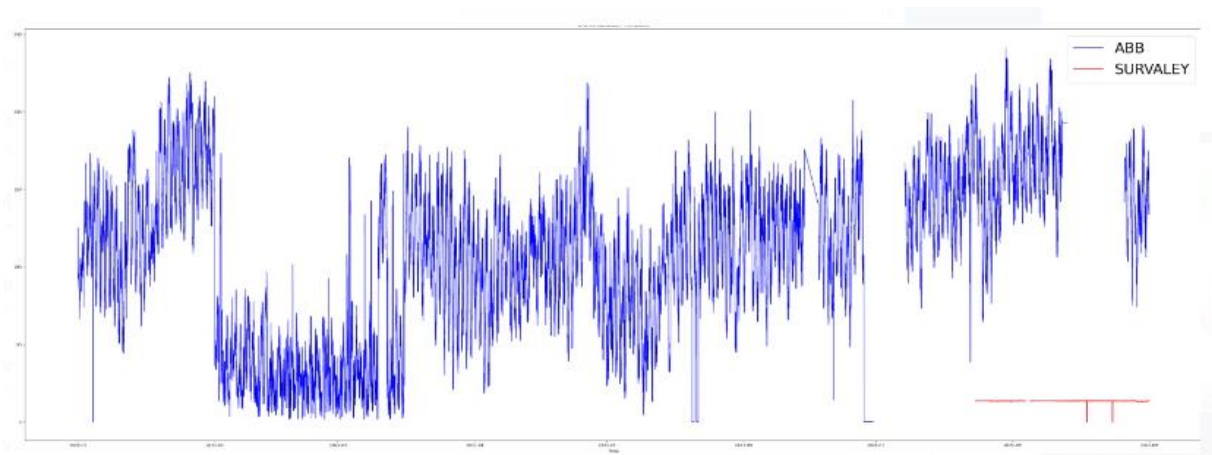


Imagen caso 2

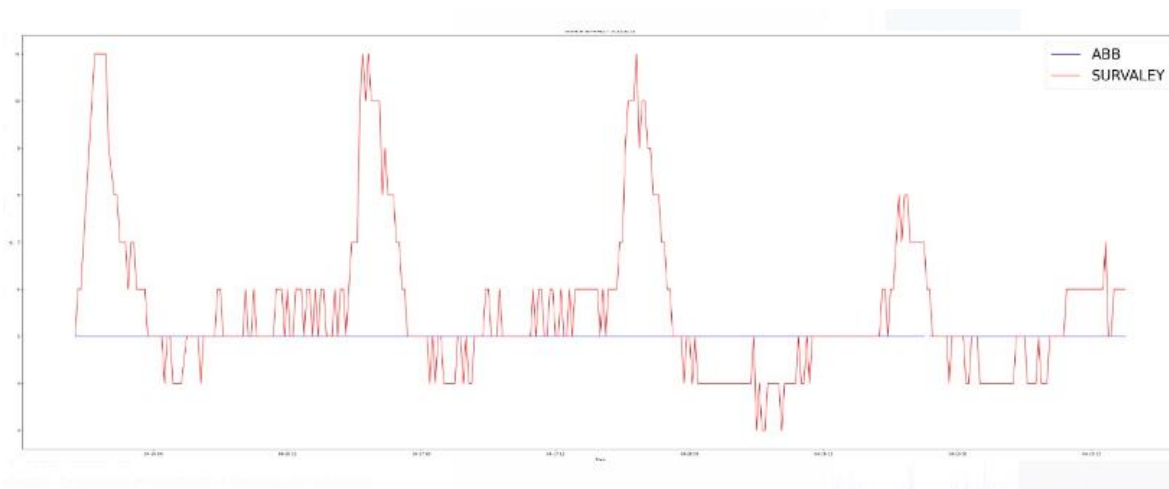


Imagen caso 1

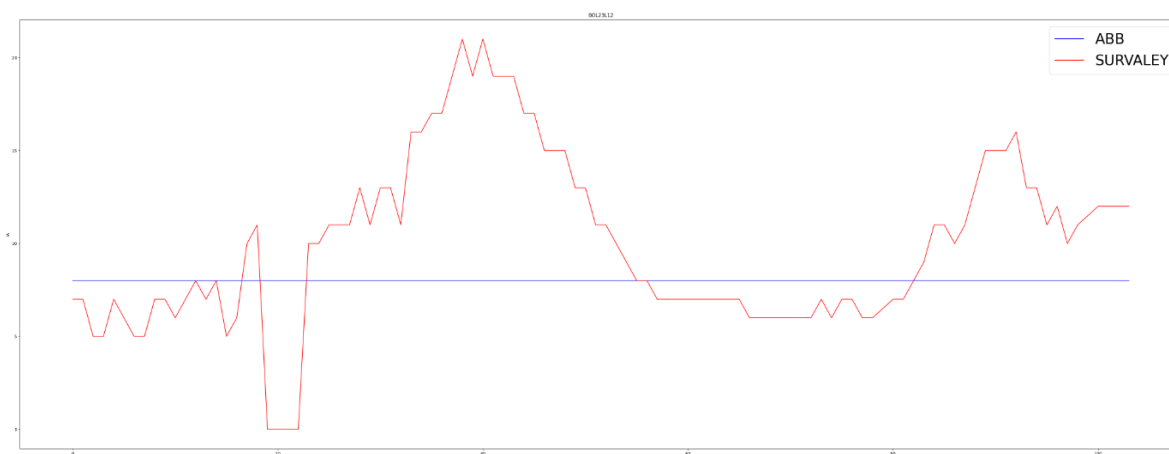


Imagen caso 1

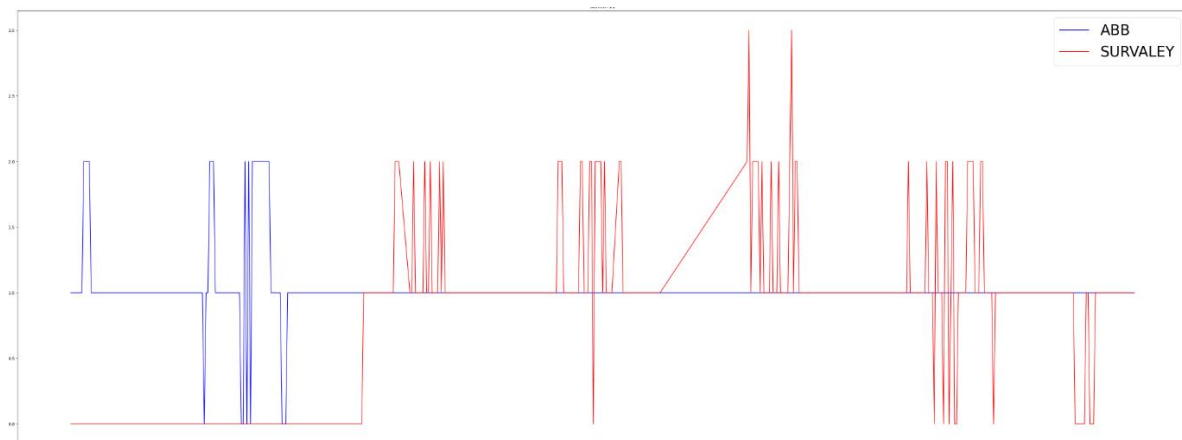


Imagen caso 3

Dashboard (Power bi)

Uno de los objetos del sprint consistía en desarrollar un Dashboard para el análisis de las series de tiempo de las variables de interés como son las corrientes, voltajes y potencias de los sistemas SCADA y PRIME para evaluar su relevancia a la hora de realizar los modelos de imputación planteados en el marco de este proyecto, además funciona como insumo para que el equipo CHEC se lo utilice para más adelante estudiar la calidad de sus datos, teniendo en cuenta esto, se desarrollo dicho tablero cuya estructura se explica a continuación:

Filtros generales:



Imagen 1.1 (Filtro de tiempo y nivel de tensión)



Imagen 1.2 (Filtros Voltajes regulatorios – Circuito -Pandemia)

Estipulados estos filtros generales, realizamos su debida explicación:

- Línea de tiempo: Como su nombre indica es un filtro temporal que permite segmentar por año, mes, semana y día.
- Franja horaria: Como en el filtro anterior no es posible segmentar la hora, se añadió este seleccionable para determinar 3 rangos de hora que corresponden a las 3 etapas del día (mañana, tarde y noche).
- Nivel de tensión: Debido a que los circuitos se encuentran segmentados en categorías y es relevante hacer un análisis para cada una de ellas tanto global como específica, este filtro permite acceder directamente a cada una de esas categorías.
- Circuito: Como su nombre indica es un seleccionable para cada uno de los circuitos registrados entre las tablas de interés, enfocándonos exclusivamente en el sistema SCADA, puesto que es el sistema al cual va enfocado el modelo de imputación.
- Vx Regulatorio: Este filtro que está tanto para cada una de las fases con respecto a los voltajes, es un filtro que permite separar los valores que están dentro de los rangos regulatorios de cada circuito y aquellos que no se encuentran dentro de estos.
- Pandemia: Como su nombre indica, es un filtro para poder ubicar el tablero antes, durante y después de la pandemia en caso de querer hacer un análisis exploratorio de dichas épocas.

Sistema SCADA (Página 1)

La pagina 1 de este tablero va enfocada únicamente al sistema SCADA, donde partiendo de las 2 bases de datos horizontalizadas que corresponden a los 2 tipos ABB y SURVALENT, se realizó un análisis de distribuciones, datos perdidos, anómalos y coherencia de los datos, detallando la siguiente estructura:



Imagen 1.3 (Corrientes SCADA)

En este apartado, se muestran las series de tiempo para los 2 sistemas, tanto ABB y SURVALENT, de las corrientes en las 3 fases, puesto que hay sistemas, que como vimos antes están presentes en ambos y nos permiten ayudar a terminar de estudiar el análisis anterior a la explicación del Dashboard, de aquí podemos desprender información del comportamiento de la media y desviación de los datos, que básicamente son medidas que podemos comparar para el tema de detección de atípicos, y además complementar con el análisis de distribuciones de los datos, que también es mostrado dentro de este Dashboard, debido a que dependiendo de la distribución de los datos podemos aplicar diferentes técnicas exploratorias para el preprocesamiento de los datos al momento de crear un modelo de imputación, concluyendo este apartado con un análisis de datos perdidos y con poca información como serian los datos en cero que es un estudio del primer sprint.



Imagen 1.4(Voltajes ABB y SURVALENT)

Continuando tenemos el análisis de voltajes, donde sumado al estudio descriptivo de las series de tiempo realizada para las corrientes, es decir el comportamiento de las distribuciones, aquí analizamos la coincidencia de los valores de voltaje de cada circuito, con el rango regulatorio establecido para dichos circuitos, para detección de datos atípicos, donde hayamos un porcentaje (%) con de dichos anómalos con respecto al total, y un promedio diario de obtención de atípicos.

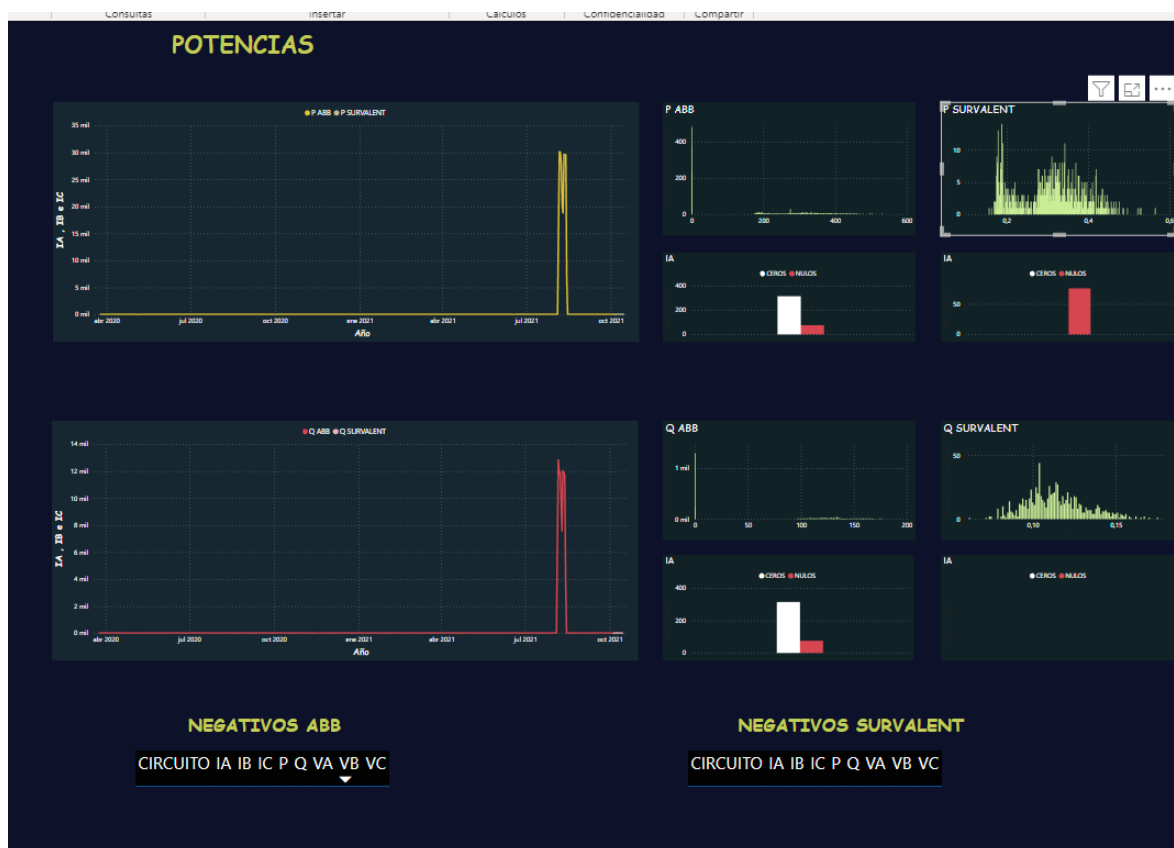


Imagen 1.5 (Potencias)

Finalmente obtenemos un estudio de las potencias de la misma manera que las variables de interés anteriores, junto a una tabla resumen para poder también estudiar valores negativos presentes en cada una de ellas y asociarlo a su circuito determinado.

PRIME (Página 2)

La segunda página es un análisis similar al realizado con el SCADA esto se puede ver reflejado en la distribución del tablero que se ve en la siguiente imagen:



Imagen 1.6 (Prime -Voltajes -Corrientes)

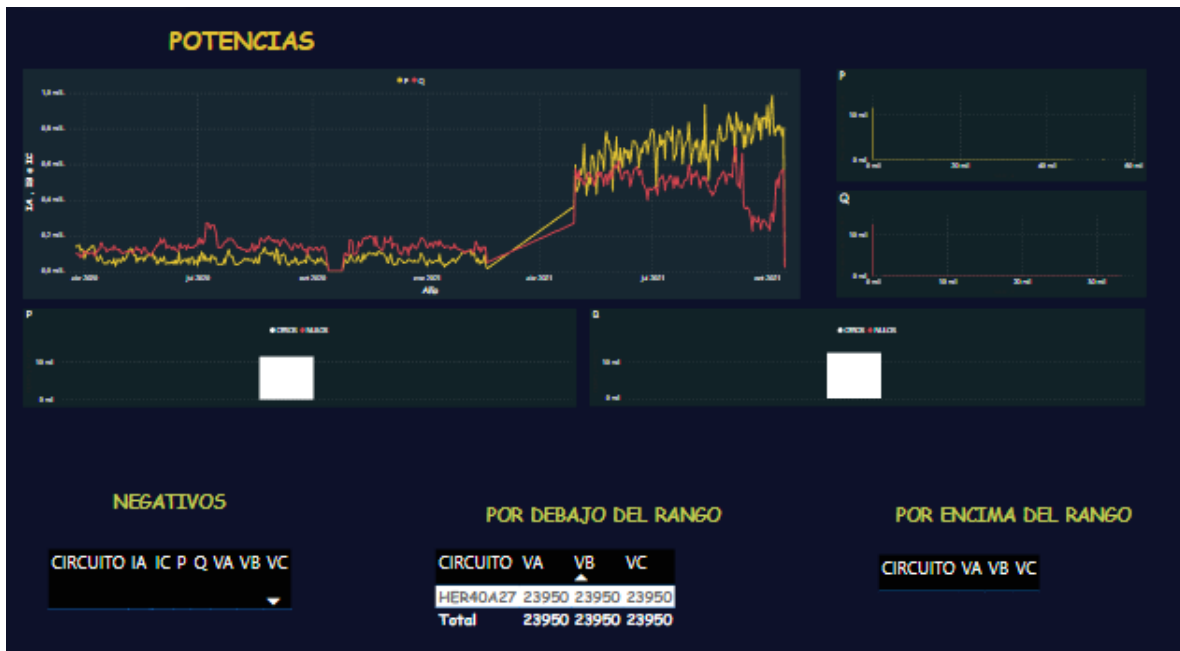


Imagen 1.7 (Prime -Voltajes -Corrientes)

Análisis y conclusiones:

- Corrientes y potencias: durante la construcción del Dashboard de Chec, analizando las distribuciones de los datos para variables como corrientes y potencias, para varios circuitos, se pudo observar que una gran cantidad de ellos, poseen comportamientos gaussianos o combinaciones de estas, lo cual es una gran ventaja a la hora de estudiar una serie tiempo, ya que facilita por ejemplo la detección de datos atípicos (datos entre el percentil 0-5 y el percentil 95-100).

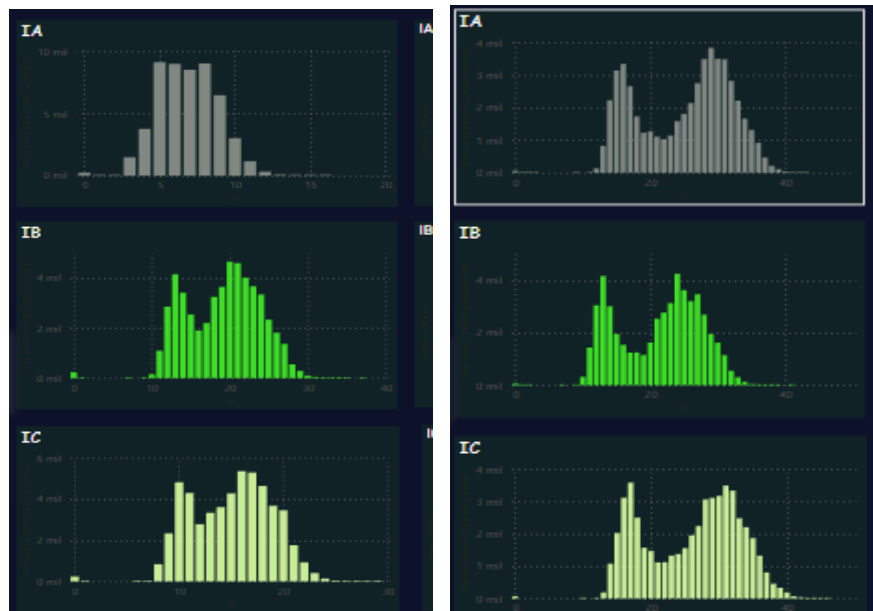


Imagen 1.8 (Corrientes - distribuciones)

Esto es relevante para variables diferentes de los voltajes en las 3 fases, debido a que al menos para estos tenemos una guía de datos atípicos utilizando, los rangos regulatorios según el código del circuito, por lo tanto, en las corrientes y potencias, si resulta una buena opción, realizar un test de gaussividad para poder aprovechar dichas características en un análisis posterior, como base tendríamos el test SHAPIRO WILK , el cual nos proporciona un valor de p-value que es un descriptor, el cual en el caso de que dicho valor, este por encima del valor de significancia (0.05) podemos asumir la serie de tiempo como gaussiana, (**Dichas pruebas aún se están corriendo en el Clúster de BIOS**).

- PRIME: estudiando el sistema prime, desde el sprint 1 determinamos ciertas falencias, como la presenta de datos nulos, en la cual se pudo determinar que tanto para los voltajes como las corrientes en todas sus fases el % de datos perdidos es mayor al **65%** , además en el Dashboard construido, se puede observar que la mayoría de voltajes en este sistema, están por fuera del limite regulatorio, lo cual no presenta mucha coherencia con el balance de datos por fuera del limite regulatorio en los sistemas SCADA.

PROMEDIO DIARIO FUERA DE RANGO VA	PROMEDIO DIARIO FUERA DE RANGO VB	PROMEDIO DIARIO FUERA DE RANGO VC
23,56	23,56	23,56

Imagen 1.9 (Tabla promedio diario fuera de rango)

- Valores negativos: Teniendo en cuenta que en el funcionamiento de los sistemas, las únicas variables que pueden presentar valores negativos, son

la potencia activa y la potencia reactiva (solo en niveles de tensión 3 y 4), se procedió, a hacer un análisis que pudiera corroborar que esto siempre es así, mas sin embargo como se puede ver en la siguiente tabla, también hay otras variables que presentan valores negativos y además, los valores negativos de potencia activa y reactiva, no solo se dan en niveles de tensión 3 y 4.

NEGATIVOS SURVALENT								
CIRCUITO	IA	IB	IC	P	Q	VA	VB	VC
2AG14R01						12		
2AG15R01				1		1		
2AM12R01				165		75		
2AM14R01				26712		26731		
2AM2IR21				305		299		
2AM3CH81				1885		1891		
2AM4IN21				252		252		
Total				2 391871		429234 561		32

Imagen 2.1 (Tabla valores negativos)

- Como se puede ver el desarrollo de este Dashboard, está enfocado al análisis de la calidad del dato para cada uno de los sistemas, principalmente en el estudio de datos atípicos, mas sin embargo, para cada uno de estos datos atípicos, es necesario aclarar por parte del equipo de CHEC, según su conocimiento acerca del sistema, si estos datos atípicos se dan por fallos en los circuitos, o si se dan por errores de medida; pues saber esto, permitirá descartar o no estos datos atípicos, a la hora de modelar el sistema de imputación de datos.

Propuestas Tercer Sprint

Teniendo en cuenta que el tercer sprint está basado principalmente, en un análisis exploratorio de las fuentes de datos, así como comenzar a generar propuestas para la imputación de estos en el sistema SCADA, se propone la siguiente lista de actividades para el mismo:

- Realizar análisis de cruce de datos entre sistemas SCADA ABB, SURVALENT y sistema PRIME (Para esto se hace necesario la definición de la llave que permita relacionar estos sistemas).
- Realizar análisis flujo de potencias (Para esto se hace necesario definir secciones de trabajo con el equipo CHEC para contextualizar al equipo BIOS en que consiste este análisis).
- Realizar análisis Nodal (Para esto se hace necesario definir secciones de trabajo con el equipo CHEC para contextualizar al equipo BIOS en que consiste este análisis).
- Realizar análisis de las series de tiempo , con respecto a la pandemia (antes, durante y después), apoyándose del tablero de power BI construido en el sprint 2, con el fin de determinar su posible influencia en el modelo, es decir, si es posible incluirla en los datos de entreno, o representan valores muy desviados.
- Realizar un análisis acerca de como se ven afectadas las series de tiempo, en los momentos en los cuales se registra un evento en el SCADA.
- Realizar un análisis detallado acerca del comportamiento de cada uno de los circuitos (Distribuciones), así como las relaciones que puedan existir entre los mismos; esto con el fin de determinar si a la entrada del modelo debe haber una variable que indique el código del circuito y/o antes de entrenar un modelo se puede realizar un proceso de agrupación de los circuitos (Quizás entrenar varios modelos).
- Hacer análisis de las series de tiempo por rangos apoyándose del tablero de Power BI construido en el sprint 2.
- Realizar análisis a cerca del diagrama de flujo del proceso de imputación de datos en SCADA, teniendo en cuenta los siguientes pasos: Reglas impuestas, modelo de machine learning, flujo de potencias, análisis nodal, análisis de series de tiempo, interconexiones y demás. (Definir reglas).

Dudas y Peticiones equipo CHEC

- Tabla con llaves definidas para relacionar sistema PRIME con sistemas SCADA.
- Definir secciones de trabajo con el equipo CHEC.
- En SCADA ABB están los siguientes códigos de circuitos, a los cuales no se les puede inferir el voltaje nominal; por tanto, se hace necesario una tabla con los voltajes nominales para estos circuitos: SYS_NETD, DOR_SEQ, B_ALARM, ATAS_LOG, ATAS_TST, AZESZ1K1, AZESZ1K2, AZESZ2K1, AZESZ2K2, AZESZ3K1, AZESZ3K2, AZESZ4K1, AZESZ4K2.
- En SCADA SURVALENT están los siguientes códigos de circuitos, a los cuales no se les puede inferir el voltaje nominal; por tanto, se hace necesario una tabla con los voltajes nominales para estos circuitos: PUR_SSAA, RTU_VIR, MM1AS3S1, CH3CH4S1, AS2MM1S1, AS3MM1S1, CH3MM3S1, CH6MM1S1, CH6MM4S1, MM4MM1S1, MM6MM4S1, CH6PS3S1, CH4S1A3, CH6S1A4, MM1S1A1, MM3S184, MM4S1A4, MM6S1A2, MM1S2A1, MM6S2A4.
- Tabla con voltaje nominal para los siguientes niveles de tensión: 10,25,32,70.
- ¿En SCADA ABB y PRIME solo hay circuitos de cabecera, mientras que en SURVALENT hay de cabecera y de red, esto a que se debe?, puede cambiar?
- 2C también es 13.2Kv para los circuitos de cabecera?
- ¿PRIME tiene la columna Nivel_De_Tension, pero esta no coincide con el nivel de tensión del código del circuito, cual es el nivel de tensión correcto?
- Cuenta de Power BI para publicar el Dashboard.
- Diccionario calidad del dato para SCADA SURVALENT.
- Se pasará script para crear vistas SCADA ABB y SURVALENT horizontalizadas, teniendo en cuenta la columna de calidad del dato. Una vez creadas estas vistas, se incorporarán en el Power BI, para tener en

cuenta en el análisis únicamente los registros con código de calidad del dato válido.

- Reactivar usuario de spinedaq.
- Agendar Reunión con Experto en TI CHEC para resolver dudas acerca de formato de solicitud de conexión VPN desde el clúster de BIOS.