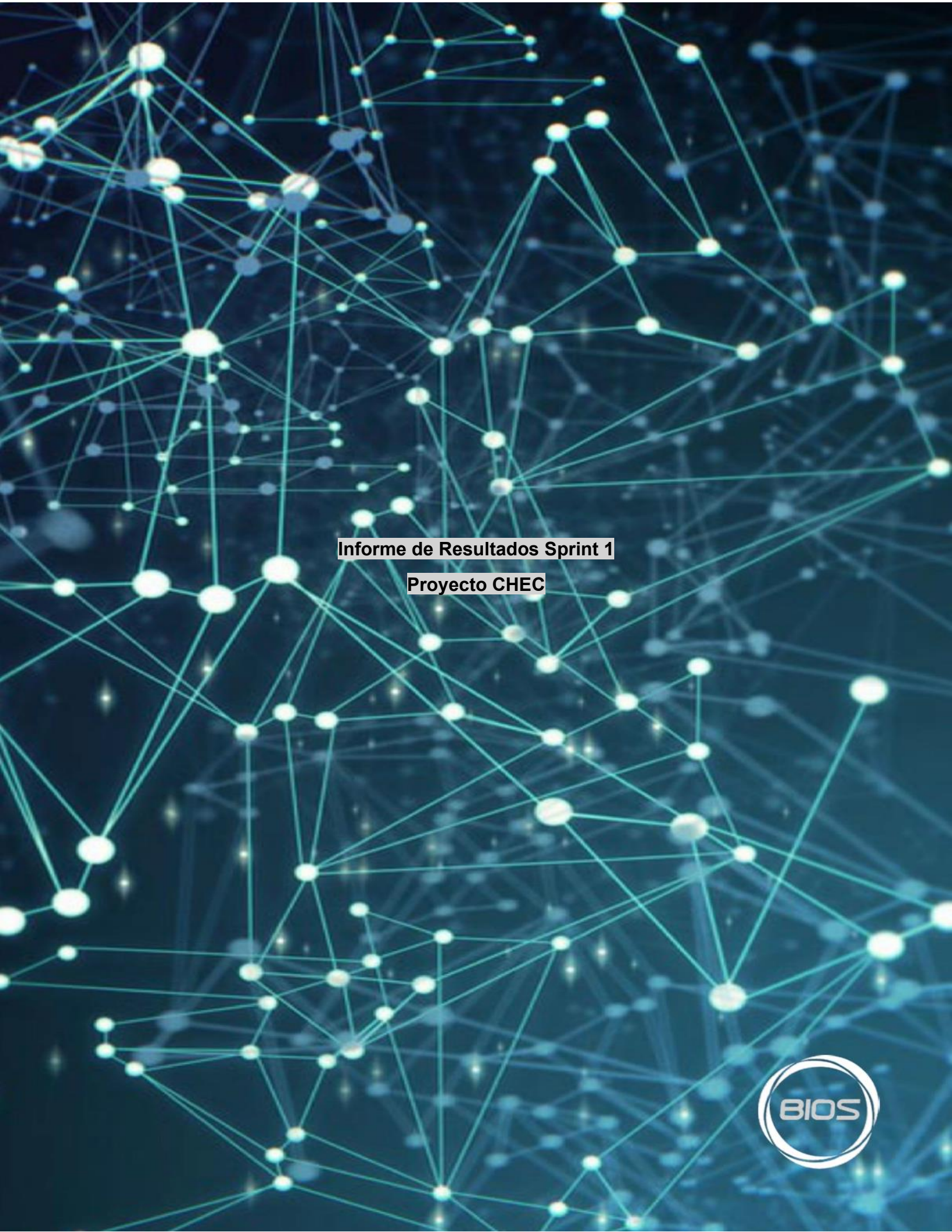
Informe de Resultados Sprint 1
Proyecto CHEC
BIOS

Introducción

El presente informe corresponde al primer 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, así como se ha realizado un análisis preliminar de las fuentes de información con el fin de definir la estrategia para la adquisición y unificación de estos.

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.
- Garantizar que el equipo de desarrollo tenga acceso pleno a los datos necesarios para el proyecto y que comprenda adecuadamente su relevancia y utilidad en el análisis de calidad de los datos. Esto incluye la identificación de posibles obstáculos o limitaciones en el acceso a los datos y su posterior resolución.

Descripción y Acceso a las Fuentes de Datos

Para este proyecto se suministraron accesos iniciales por SQL Server mediante runas, conectado por la vpn interna de chec, donde se dieron permisos de visualización para dos fuentes de datos :

- **DM_OPERACION :** Esta base de datos contiene la vista V_MEDIDAS_ANALOGAS, la cual contiene información acerca de los registros de las fuentes de medida de los dos sistemas SCADA de interés en este proyecto, estos, se encargan de registrar los datos de las medidas

correspondientes a los 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 vista:

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 (V_MEDIDAS_ANALOGAS_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_OPERACION , tambien se dio un acceso a la vista V_EVENTOS_SCADA, la cual contiene información como su nombre indica , de los eventos asociados a los circuitos de las mediciones captadas por el SCADA, como alarmas, mediciones erroneas, sistemas apagados entre otros, los datos registrados en esta tabla no esta referenciada bajo un periodo de muestreo fijo, sino que es captada en el momento en que ocurre el suceso. Las columnas de interes se ven reflejada en la siguiente tabla:

Variable	Descripción
EVE_CIRCUITO	Código identificador de cada uno de los circuitos donde se genero el evento.

EVE_FECHA	Fecha y hora en la cual se dio el registro del evento para el correspondiente circuito.
EVE_ESTADO	Corresponde a una referencia de la alerta, es decir da una corta descripción del motivo por el cual se genero el evento.
SCADA	Indica en que sistema escada fue realizada la medición (no fue incluida en el estudio preliminar, pero puede ser de interés mas adelante).

Tabla 1.2 (V_EVENTOS_SCADA VARIABLES)

- **DM_RI_HANA :** Esta base de datos contiene la tabla tmp_EXTRACCION_PRIME, la cual contiene información acerca de los registros de las fuentes de medida del sistema PRIME de interés en este proyecto, este, se encarga de registrar los datos de las medidas correspondientes a los circuitos que conectan con los canales secundarios de los transformadores, con un periodo de muestreo de una hora para cada uno de los circuitos, mas sin embargo , existen algunos circuitos donde el periodo de muestreo es de 15 minutos. 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
Fecha	Fecha y hora en la cual se dio el registro de las mediciones para el correspondiente circuito
Valor_I2H_A , Valor_I2H_B , Valor_I2H_C	Corriente en cada una de las fases del circuito en Amperios (A)
Valor_V2h_A , Valor_V2h_B , Valor_V2h_C	Voltajes en cada una de las fases del circuito en Voltios (V)
Valor_A	Corresponde a la potencia activa del circuito en kiloVatios (kW)
Valor_R	Corresponde a la potencia reactiva del circuito en kiloVatios (kW)
Valor_Ar	Corresponde a la potencia activa del circuito en kiloVatios (kW) en dirección inversa (Esto no sucede en todos los circuitos)

Valor_Rr	Corresponde a la potencia reactiva del circuito en kiloVatios (kW) en dirección inversa (Esto no sucede en todos los circuitos)
Codigo	Corresponde a la frontera del circuito

Tabla 1.3 (tmp_EXTRACCION_PRIME VARIABLES)

Es de destacar que, si bien se tienen diferentes fuentes de datos, el modelo de imputación que se diseñara en este proyecto, solo actuara como modelo interpolador para los datos de los sistemas SCADA.

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 [91.859.961]**

Variable	Cantidad perdidos	% Perdidos
CIRCUITO	0	0%
TIEMPO_AJUSTADO	0	0%
IA	5.653.462	6,154%
IB	6.636.847	7,262%
IC	6.071.560	7,262%
VA	8.605.513	9,368%
VB	9.007.577	9,805%
VC	9.332.313	10,159%
P	8.299.262	9,034%
Q	8.238.363	8,968%
	Promedio	8,496%

Tabla 2.1 (V_MEDIDAS_ANALOGAS DATOS PERDIDOS)

Como se puede ver tanto el porcentaje de datos perdidos para cada una de las variables como su promedio, esta por debajo del 10%, por lo que se puede concluir que en el sistema SCADA, 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.

- **PRIME [61.377.718]**

Variable	Cantidad perdidos	% Perdidos
Circuito	600.156	0.977%
Fecha	0	0%
Valor_I2H_A	40.351.651	65,743%
Valor_I2H_B	40.351.773	65,743%
Valor_I2H_C	40.672.923	66,26%
Valor_V2h_A	40.109.700	65,394%
Valor_V2h_B	40.137.13	65,394%
Valor_V2h_C	40.490.691	65,969%
Valor_A	1.080.821	1,7%
Valor_R	1.080.821	1,8%
Valor_Ar	15.021.461	24,47%
Valor_Rr	18.000.749	29,33%
	Promedio	37,73%

Tabla 2.2 (tmp.EXTRACCION_PRIME DATOS PERDIDOS)

Como se puede ver en la tabla anterior, para la mayoria de las variables de interés, mas del 50% de los datos estan perdidos, por lo que a priori se podria concluir, que apoyarse del sistema PRIME para calcular datos perdidos en el sistema SCADA, no es una opción muy viable; mas sin embargo en el sprint 2 se hara un analisis de las respectivas series de tiempo del prime, para determinar la viabilidad del uso de estos datos, pues por ejemplo puede que la perdida de estos datos se deba a un periodo en el cual el sistema haya estado apagado o en mantenimiento.

- **EVENTOS_SCADA [43.954.444]**

Variable	Cantidad perdidos	% Perdidos
EVE_CIRCUITO	6098	0,013%
EVE_FECHA	0	0%
EVE_ESTADO	74	0,000168%
SCADA	0	0,003
	Promedio	0,003%

Tabla 2.3 (V_EVENTOS_SCADA DATOS PERDIDOS)

Como se puede ver en la tabla 2.3 la perdida de datos en las variables de interés, es casi nula, por lo cual esta fuente de datos, podria ser de gran utilidad, a la hora de realizar el etiquetado dentro del modelo de machine learning para la detección de datos anomalos y sobrecargas en los circuitos en el sistema SCADA.

Posteriormente se hizo un analisis para determinar que circuitos estan tanto en el sistema SCADA como en el sistema PRIME, para de esta manera saber, a que circuitos se le pueden imputar datos perdidos en el sistema SCADA, 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 la cantidad de circuitos tanto a nivel general como en el piloto proporcionado para el ejercicio con el PRIME y el SCADA:

Analisis	SCADA	PRIME	Cantidad en común
GENERAL	1008	207	189
PILOTO	25	11	11

Tabla 3.1 (Circuitos PRIME-SCADA)

A partir de la tabla anterior se puede concluir que , con una cantidad muy baja de circuitos (189) , se podran imputar datos perdidos del sistema SCADA, haciendo calculos a partir del PRIME, y de las relaciones de corrientes y voltajes, de las respectivas fronteras.

Luego, se procedio a hacer el mismo analisis, pero entre el sistema SCADA y EVENTOS_SCADA, obteniendo los siguientes datos:

Analisis	SCADA	EVENTOS_SCADA	Cantidad en común
GENERAL	1008	1690	943

Tabla 3.2 (Circuitos EVENTOS_SCADA-SCADA)

Teniendo en cuenta los datos de la tabla anterior se puede concluir, que definitivamente, la fuente de datos EVENTOS_SCADA, no presenta inconvenientes para el etiquetado de comportamientos anomalos o de sobrecarga, en los circuitos del SCADA, puesto que la mayoria de estos presentan registros en la tabla de eventos. Además , se genera la siguiente duda: ¿Por qué hay mas circuitos en EVENTOS_SCADA que en SCADA?

Luego, sabiendo que el sistema SCADA y el sistema PRIME tienen distintas bases de tiempo, se hace necesario corroborar que los circuitos que estan presenten en ambos sistemas, tengan datos durante el mismo rango de tiempo; mas sin embargo a la hora de comprobar esto, se pudo ver que estos circuitos no tienen datos durante el mismo rango de tiempo en ambos sistemas. A continuación se muestran graficos de barras que muestran la cantidad de veces que las series de tiempo, de los circuitos comienzan antes en un sistema que en el otro y en cual sistema tienen datos mas recientes.

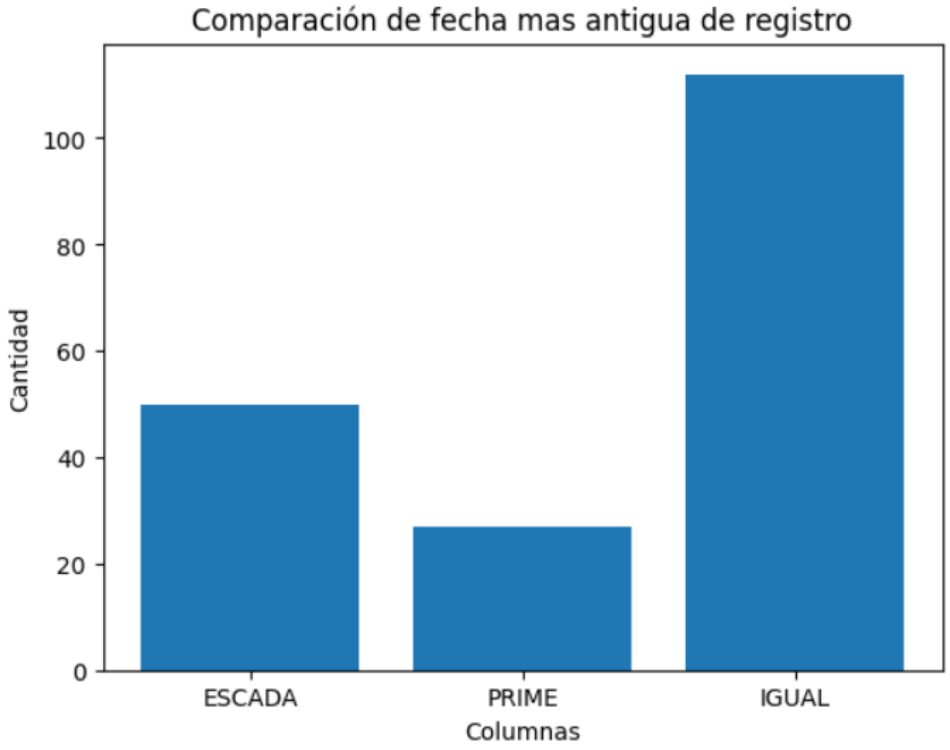

Tabla 4.1 (Registro mas antiguo)

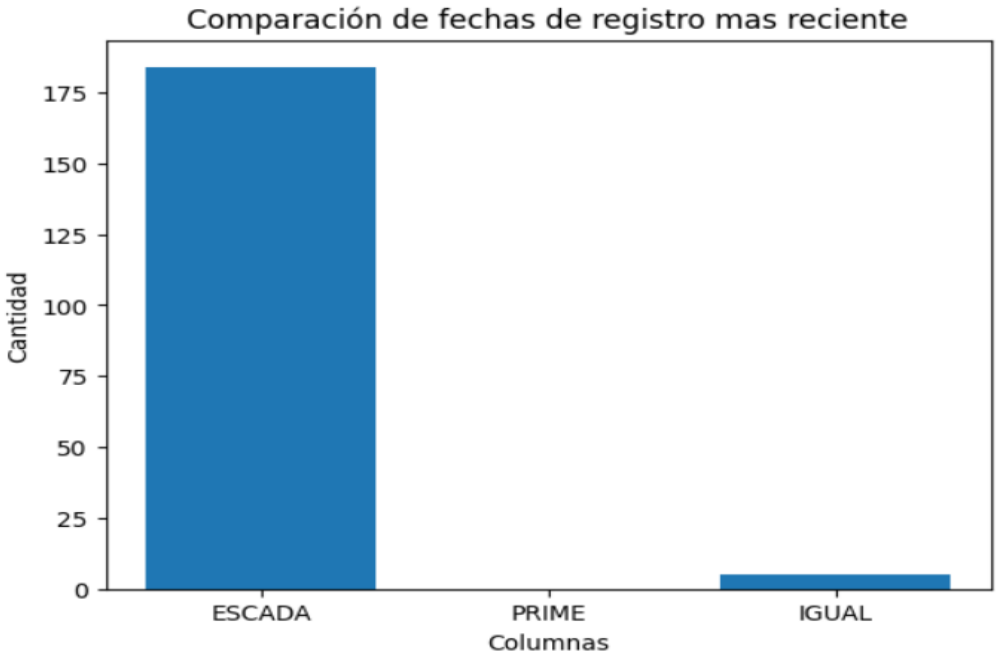

Tabla 4.2 (Registro mas reciente)

De los anteriores graficos, facilmente se puede concluir, que el sistema SCADA, tiene un periodo de actualización mas rapido que el del PRIME para la mayoria de los circuitos que estan en ambos sistemas; por tanto a la hora de realizar la imputación de datos en el sistema SCADA, esta imputación se debe hacer hasta la ultima fecha de actualización del sistema PRIME.

Posteriormente, sabiendo que no habian datos nulos tanto en el PRIME como en el SCADA en cuanto a las fechas de registro de los datos, esto lleva a pensar que aunque hubieran datos perdidos en cualquiera de las variables de interes, el registro siempre se generaba en las bases de datos con su respectivo periodo de muestreo; mas sin embargo, se hacia necesario realizar un ejercicio de comprobación con los circuitos piloto que estuvieran en ambos sistemas, para esto , se cada uno de los circuitos tanto en el PRIME como en el SCADA, a un mismo rango de tiempo y además , se tomaron los registros unicamente de las horas exactas , obteniendo los siguientes resultados:

Circuito	REGISTROS PRIME	REGISTROS SCADA	REGISTROS esperados
AMA30T15	76642	29825	30721
BEL23L12	26999	24242	27470
BEL23L14	145184	28172	30726
BQE23L12	61582	28333	30070
BQE23L13	30304	28925	30665
ROS23L12	101430	30142	30665
ROS23L13	153416	30142	30665
ROS23L14	162658	30184	31307
ROS23L15	408931	30143	30666
ROS23L16	422912	30142	30665
ROS23L19	30324	30133	30662

Tabla 4.3 (Analisis de registros)

Como se puede observar en la tabla anterior, en el sistema SCADA, para todos los circuitos siempre hay menor cantidad de registros que los esperados, lo cual quiere decir que hay registros que no se están cargando a la base de datos, por lo cual, se debe evaluar la posibilidad de que estos registros también puedan ser imputados por el modelo de imputación de datos a desarrollar.

Por otro lado, es importante notar que para el caso del sistema PRIME, para la mayoría de los circuitos hay más registros que los esperados, debiéndose esto a que hay circuitos que tienen mas de una frontera. Además, para algunos circuitos hay menos registros que los esperados, lo cual indica que hay registros que no se están cargando en la respectiva base de datos; lo cual es otra limitante a la hora de imputar datos en el SCADA a partir del sistema PRIME.

Propuestas Segundo Sprint

Teniendo en cuenta que el segundo sprint está enfocado principalmente en continuar con un análisis en cuanto a la calidad de las fuentes de datos, se propone la siguiente lista de actividades para el mismo:

- Volver a realizar el análisis de calidad de datos realizado en el primer sprint, pero ahora enfocado hacia las nuevas fuentes de datos.
- Realizar análisis de calidad de los datos, teniendo en cuenta análisis nodal y consistencia de los datos entre sistema SCADA y sistema PRIME.

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:

- Mirar que porcentaje de los voltajes, tanto del sistema SCADA como del sistema PRIME, están dentro de su rango nominal, y así mismo que porcentaje de esos son anómalos.
- 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, interconexiones y demás.
- Realizar un análisis detallado de las series de tiempo, con respecto a la pandemia (Antes, durante y después), para determinar su influencia en el

modelo, es decir si es posible incluirla en los datos de entreno, o representan valores anómalos.

- 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 codigo del circuito y/o antes de entrenar un modelo se puede realizar un proceso de agrupación de los circuitos (Quizás entrenar varios modelos).
- Definir reglas acerca de que se puede imputar por reglas impuestas y que se puede imputar mediante un modelo de machine learning.
- Hacer análisis de las series de tiempo por rangos.
- Mirar que por cada uno de los circuitos haya una cantidad considerable de datos para así saber si vale la pena entrenar el modelo con los datos de este circuito; pues de lo contrario, no será posible imputarlos con un modelo de machine learning.

Peticiones equipo CHEC

- Actualizar la tabla voltajes nominales para todos los circuitos.
- Actualizar archivo de interconexiones para todos los circuitos.
- Actualizar tabla de homologación fronteras-circuitos o se definirá otra llave para hacer esta relación.
- Fronteras o llaves principales para cada circuito.
- Solicitar el puerto en el cual se ejecuta la VPN de chec.
- Reactivar usuario de spinedaq.