Universidad del Desarrollo
Facultad de Ingeniería

AJUSTE DE LA DISTRIBUCIÓN PROBABILÍSTICA Y CONSTRUCCIÓN DE DATA SINTÉTICA DE XDR´S

POR: ALBERTO DÍAZ – ELIZABETH VÁSQUEZ

Capstone Project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESORES GUÍA:

LORETO BRAVO – LEONARDO FERRES

DICIEMBRE 2022

SANTIAGO

AGRADECIMIENTOS

A nuestras familias y parejas, agradecemos su tolerancia, paciencia infinita y apoyo incondicional durante este proceso. También queremos agradecer a nuestros hijos por ceder parte de su tiempo para ayudarnos a alcanzar nuestro objetivo.

ÍNDICE

Resumen

El patrón temporal de registros de usos de datos entrega información de utilidad para las compañías de telecomunicaciones (TELCO), donde se detecta tiempo critico de tráfico de datos de sus clientes. Este es un indicador que permite mejorar la calidad del servicio junto con detectar los tramos críticos en los que se recomienda optimizar el servicio, dado que, fallas eventuales en el periodo de mayor uso de datos traería reclamos a la compañía, lo que afectaría a los índices de satisfacción.

Por ello nos embarcamos en este proyecto que tiene como motivación caracterizar y describir las series de XDRs en distintos tiempos comparando comportamientos de distribuciones, y al mismo tiempo, distinguir el tramo de horario de mayor actividad mediante las conexiones de sus clientes capturadas por hits en las antenas, y dada la información conocida, predecir un día de actividad a partir de 3 días conocidos. Con ello se podrá establecer y reconocer futuros puntos críticos, información valiosa para preparar medidas de contingencia previstas ante el alza de conexiones.

1. Introducción

Este proyecto se basa en caracterizar y obtener el ajuste de distribución de probabilidades a una serie de registros de uso de datos de clientes TELCO en registros de detalle extendido, XDRs.

Se pronosticará la frecuencia de ocurrencia de la magnitud del fenómeno detectando el periodo de tiempo critico de tráfico de datos, y a modo de proximidad, el patrón temporal de registros de uso de datos nos proporcionaría el tramo horario de mayor actividad, todo esto a consecuencia de las conexiones de sus clientes capturada por Hits de las antenas de la empresa TELCO a nivel nacional, que es donde se origina la información.

Esta información resulta valiosa en cuanto permite aplicar mejoras a la planificación vial y la promoción de actividades económicas dentro de la región.

Por ello en el desarrollo de este estudio se busca conocer el comportamiento de la distribución de probabilidades, que permite inferir con confianza, y al mismo tiempo predecir, un día de actividad mediante la información de XDRs.

2. Trabajos Relacionados

Los estudios relacionados con el patrón temporal de los datos XDRs pueden incluirse en investigaciones sobre cómo cambian estos datos a lo largo del tiempo, cómo se puede predecir o modelar esos cambios y cómo se pueden utilizar los datos XDRs para analizar y comprender fenómenos temporales. Por ejemplo, se ha examinado el problema de la pronosticabilidad de la movilidad a gran escala en un estudio titulado "Límites de Previsibilidad en la Movilidad Humana" (CHAOMING SONG, 2010), en el que se concluyó que, entre otras determinantes, los patrones de movilidad son predecibles a corto plazo, pero la previsibilidad es limitada a largo plazo debido a la complejidad del comportamiento humano.

- En el artículo: “Sensing Urban Patterns with Antenna Mappings: The Case of Santiago, Chile” (José García, 2016), se describe como se pueden utilizar mapas de antenas para comprender los patrones de movilidad humana en una ciudad. Para ello, se utilizaron datos de telefonía móvil y GPS para comprender cómo se desplazan las personas por la ciudad. Algunos de los resultados mostraron que las personas tienden a moverse más dentro de sus barrios de origen que a otras partes de la ciudad.
- Otro estudio realizado que a trabajado patrones de datos XDRs ha sido el articulo “Toward Finding Latent Cities with Non-Negative Matrix Factorization”, (Denis Parra, 2018). En este estudio, se utilizaron datos de la red de telefonía móvil para inferir patrones automáticos de viaje utilizando factorización de matriz no negativa. El método

se puso en práctica en una gran ciudad y se estableció que las movilidades revelan estructuras latentes e interpretables, entre otros resultados.

En resumen, los estudios relacionados con el patrón temporal de los datos XDRs se centran en el análisis de cómo cambian estos datos a lo largo del tiempo y en cómo se pueden utilizar para entender fenómenos temporales, por ello en este proyecto final, se busca utilizar estos estudios como base para estudiar la distribución de probabilidad de los datos XDR según el día que fueron recolectados, con el objetivo de comparar el tráfico entre días y determinar el tramo horario con mayor actividad y con la información recolectada y analizada predecir el día de actividad siguiente a lo observado, de tal forma de entender el comportamiento de los datos XDRs con tres días de observación.

2.1. Enfoque Investigativo

En este desarrollo, a diferencia de otros estudios revisados, se busca construir data sintética de la distribución de frecuencias de XDRs para encontrar un patrón que sea útil para la inferencia de la movilidad durante 24 horas. Además, se analiza la serie temporal a nivel de segundos para un día de frecuencias de XDRs a nivel nacional. Esto representa un aporte científico novedoso. Se espera que este estudio sirva como anclaje para correlacionar con el flujo de movilidad en 24 horas. Si se encuentra una alta correlación y estudios de distribuciones entre las variables de frecuencias de XDRs y flujo de movilidad, se podría inferir que el patrón de XDRs en un día es similar al flujo de viajes interurbanos.

3. Hipótesis y Objetivos

3.1. Hipótesis

La distribución de probabilidades del tráfico de datos XDRs del 1 de enero del 2021, es igual a la distribución de probabilidades de tres días del flujo de actividad de XDRs de marzo 2022 a nivel nacional.

3.2. Objetivo

Determinar la distribución de probabilidad asociada al patrón de Hits XDRs según día.

3.3. Objetivos específicos

- Comparar distribución de tráfico de datos entre días para conocer comportamiento.
- Concluir el tramo horario de mayor actividad de XDRs.
- Predecir un día de movilidad a partir de 3 días anteriores a nivel de segundos para la frecuencia de hits XDRs

4. Datos y Metodología

4.1. Datos

Se trabaja con bases de datos entregados por Universidad Del Desarrollo, UDD, Definida como Human Mobility XDR (eXtended Detail Records) Registro de detalle extendido, los cuales contienen las transacciones de tráfico para todos los usuarios de la red de Movistar.

Se utilizaron dos conjuntos de datos en distintos tiempos.
En un principio se trabajó con muestra proporcionada por la UDD cuya fecha indicada fue el 2022-01-01 un inicio de año viernes, el cual correspondería a un feriado y fin de semana largo a esta se definió como:

- Base 1: Muestra de datos correspondientes al primero de enero 2021, de la cuál es la base para explotar para conocer distribuciones y comportamiento de movilidad. El Conjunto de campos analizados se definen como:
  - Phone_ID: Número de identificación codificado
  - Timestamp: Fecha y hora
  - Lat: Latitud
  - Lon: Longitud

Por otra parte, y dado la capacidad de cómputo de nuestras maquinas (32 gigas de RAM, GEFORCE RTX 2700, 1 tera SDD de disco duro y 8 núcleos) tomamos una muestra de 30 archivos Parquet, los cuales posee información de XDRs de 152 millones de registros a nivel nacional), la cual se definió como:

- Base 2: Está compuesta por tres días de marzo 2022; martes, miércoles y jueves. Este set de datos principalmente se trabaja para corroborar hipótesis. Respecto al viernes primero de enero del 2022.

El conjunto de datos contiene:

- Dispositivo: Número de identificación codificado
- Datatime: Fecha y hora
- Lan: Latitud

4.2.Metodología

Con el fin de extraer conocimientos de la ingente cantidad de datos XDRs de muestra entregada por la UDD, se aplicó el concepto de Minería de datos. como una fase más de un proceso de descubrimiento de conocimiento en base de datos conocido como KDD (Knowledge Discovery in Databases)

En el proceso KDD se llevó a cabo la extracción automatizada de conocimiento dado el gran volumen de datos, el cual es de naturaleza iterativa, por lo tanto, es aplicable tantas veces como sea necesario hasta obtener la información necesaria.

Para trabajar el Patrón XDR, se utilizarán el proceso Knowledge Discovery in Databases, lo cual se traduce en la siguiente metodología.

1. Los datos son recolectados desde los repositorios de la universidad que contienen información de los patrones de XDRs los cuales contienen las transacciones de tráfico en una hora y ubicación determinada. Los archivos fueron trabajados mediante librería Dask (DASK, 2014-2018), la cual permite paralelizar archivos con grandes volúmenes de datos gracias a la segmentación de los cores de la RAM del computador, dicha librería está conformada para trabajar con pandas dataframe paralelizables.
2. Preprocesamiento y tratamiento de los datos recolectados.

   Se trabajo en procesar ambos conjuntos de datos detallados en sección [3.1], Se extrajeron de forma aleatoria 30 archivos de tipo parquet mediante secuencia de instrucciones, bucle, concatenándose para obtener el conjunto de datos a procesar, donde se solicitó solo el año 2022.

Con ello se mantuvieron los tres días de estudio de marzo 2022. Paso siguiente se procedió a la limpieza de datos donde se grafico la serie y no se observaban comportamientos anómalos, sin embargo al observar las latitudes y longitudes se registraron nulos, por lo cual se procedió a eliminar tales registros del conjunto de datos.

3. Se procede ajustar distribución a base del primero de enero del 2021 como la de tres días de marzo desde el martes 1 al jueves 3 de dicho periodo. Se presentan varias librerías que colaboran en ajustar distribuciones, en un principio se trabajará con la librería Scipy.stasts (Sphinx, 2008-2022), con la cual se podrá comparar múltiples distribuciones con la finalidad de identificar cuál de ellas se ajustan mejor al comportamiento de los datos.

4. Se procede a reconstruir prediciendo un día de hits de frecuencias XDRS mediante el modelo facebook prophet (Phophet, s.f.); el cual captura la estacionalidad intradía, lo anterior es concluyente para determinar el tramo horario de mayor actividad de hits dada la descomposición de la serie temporal que el algoritmo de Prophet permite realizar.

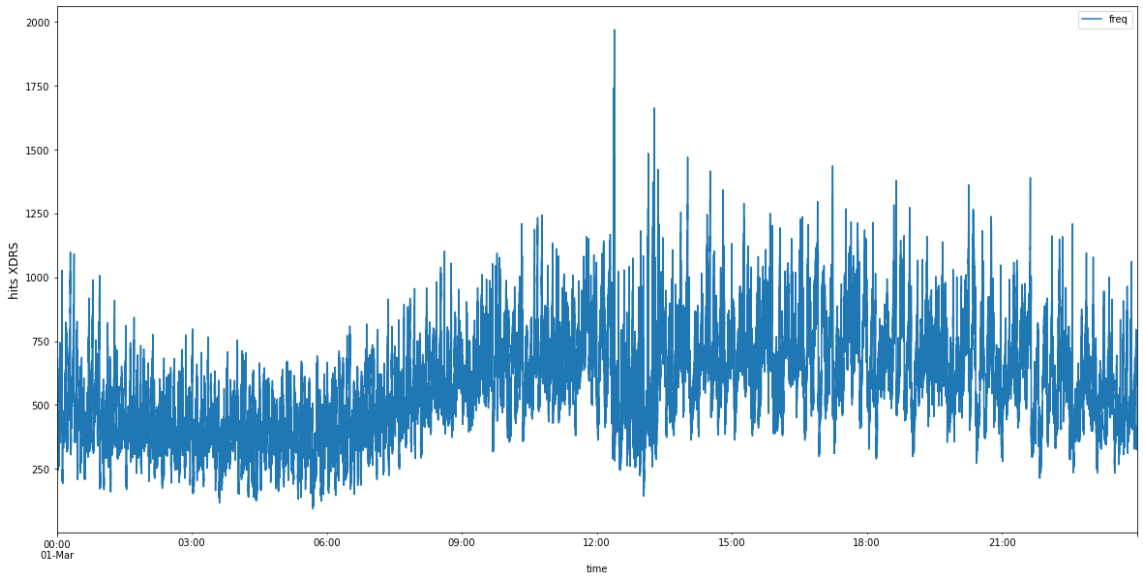
Con este desarrollo se logra concluir el horario de mayor actividad XDRs.

5. Resultados

A continuación, se plasman los siguientes resultados, detallando graficas de series temporales para el martes 1 de marzo al jueves 3 de marzo del 2022 y la serie temporal del viernes 1 de enero 2021, además se ilustran las comparativas de dichos días en términos de distribuciones de densidades que determinen una distribución de probabilidad subyacente de los datos.

Se ilustrarán resultados de comparativas de distribuciones ajustadas a una normal previamente el correcto testeo para definir si procede al comportamiento de distribución normal, como además de aplicar el modelo Facebook prophet con el fin de reconstruir un día de XDRS y detectar de forma exploratoria los tramos horarios y sus volúmenes

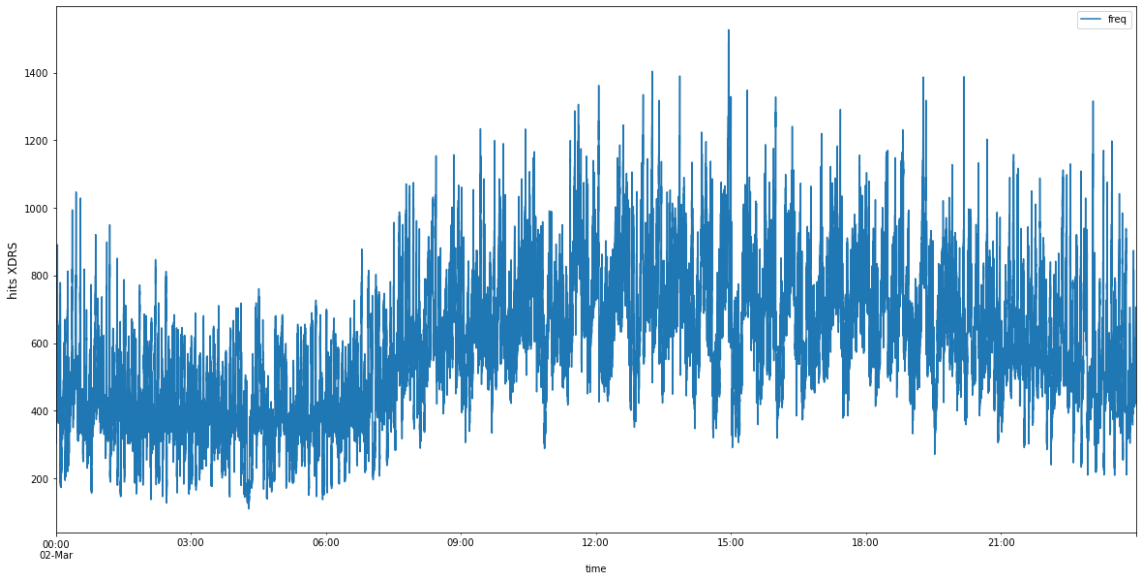
Gráfica 1: Distribución de Hits XDR martes 1 de marzo 2022



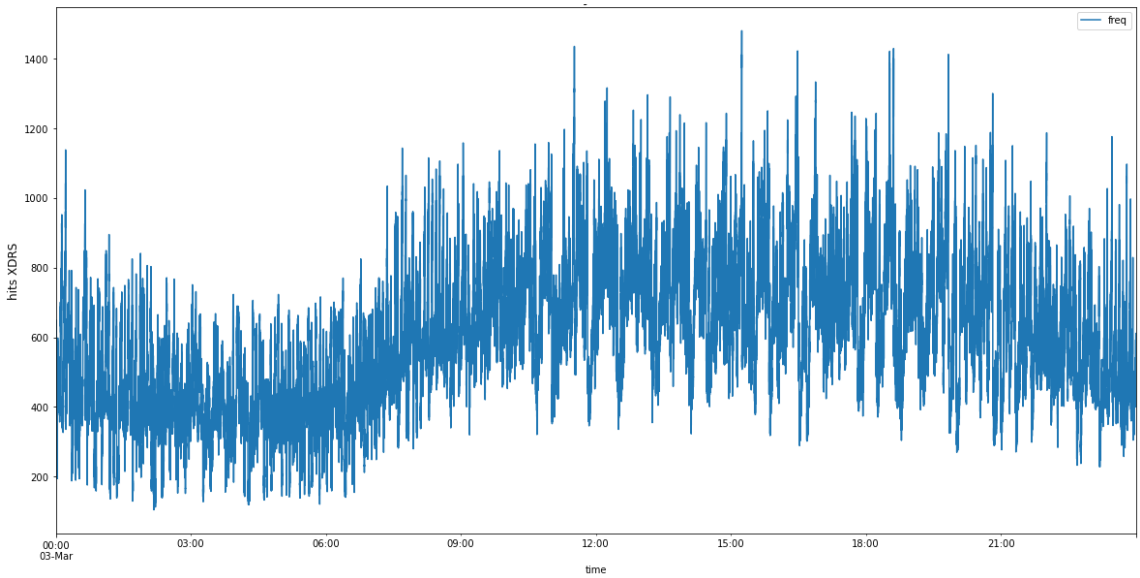
El gráfico 1, ilustra un comportamiento suave en términos de longitud de onda y a tendencia como además una alta variabilidad debido al fenómeno en estudio, frecuencias de hits

durante 24 horas, se aprecia un peak alto en el periodo cercano a medio día el cual se conocería como un dato atípico debido a una causa desconocida que logra producir dicha alza en hits.

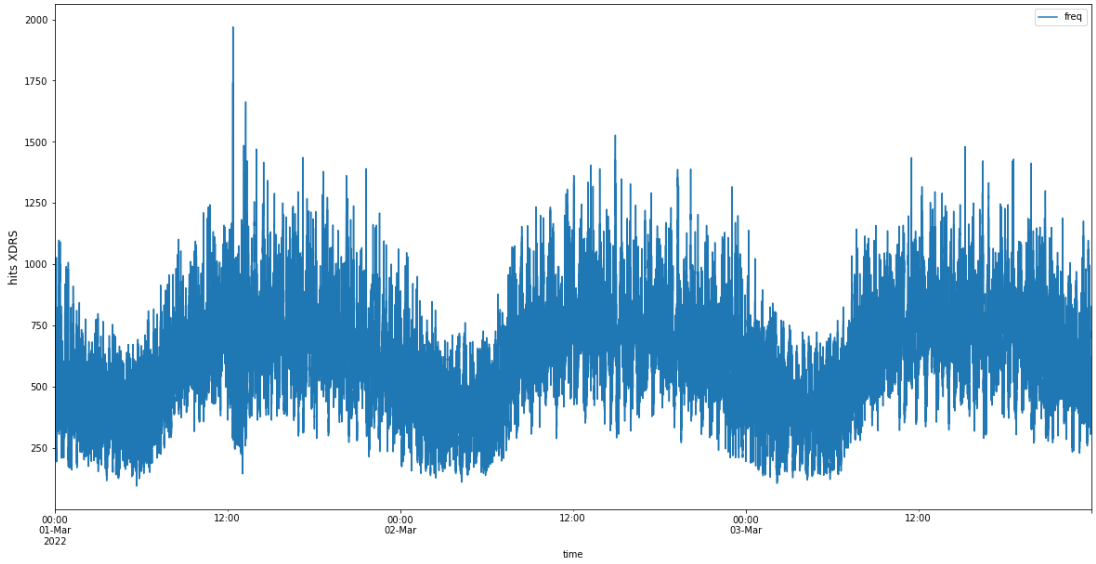
Gráfica 2: Distribución de Hits XDR miércoles 2 de marzo 2022 nivel nacional.



Gráfica 3: Distribución de Hits XDR jueves 3 de marzo 2022 nivel nacional.

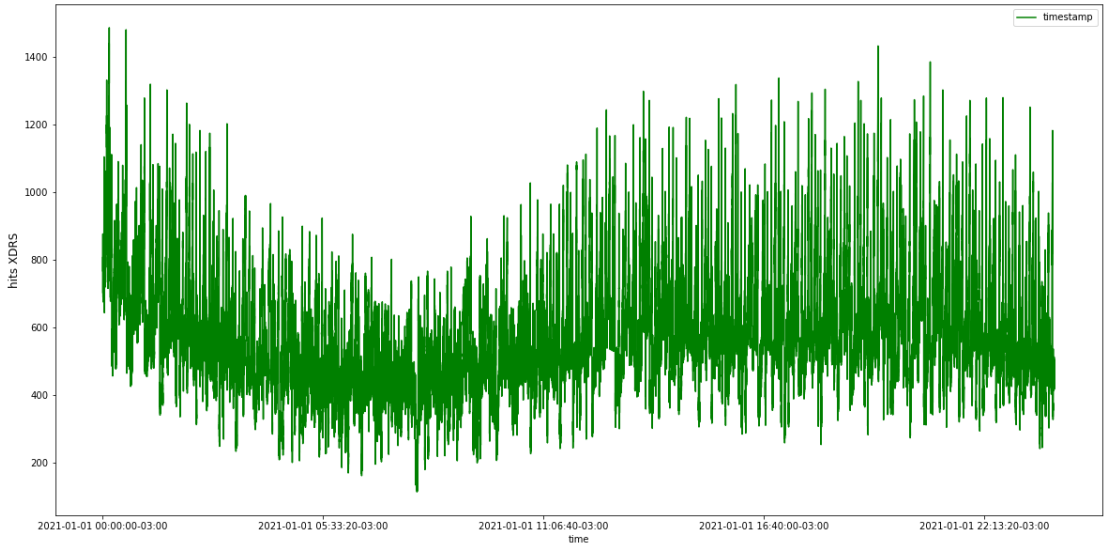


Gráfica 4: Distribución de Hits XDR del 1 al 3 de marzo 2022 nivel nacional



La grafica 4, se ilustra para observar continuidad en el comportamiento del ciclo de hits XDRS según día dado tramos horarios.

Gráfica 5: Distribución de Hits XDR viernes 1 del 2021 región metropolitana.



Se aprecia la tendencia en un alza de hits XDRS en el tramo cercano a medianoche hora cero, debido a inicio de año nuevo, por lo demás la variabilidad es mayor a simple vista en el tramo cercano a las 21 horas del día, se darán detalles precisos a medida se avanza en el desarrollos gráficos anteriores denotan un fenómeno de alta variabilidad y una tendencia en términos de onda periódica, lo cual en teoría se refleja el ciclo de hits XDRs que indicaría tramos de mayor flujo de hits durante el día, y podría apalancar información útil para determinar el instante de actividad con mayor significancia en un día.

5.1.Descriptivos

Tabla 1: Descriptivas por días

Día	Fecha	Segund.	Medía	Desv. Estándar	Min	25%	50%	75%	Max
Martes	01-03-2022	86.400	595,09	196,47	105	445	585	735	1480
Miércoles	02-03-2022	86.400	574,38	191,48	94	430	561	701	1969
Jueves	03-03-2022	86.400	591,11	196,28	110	442	580	726	1527
Viernes	01-01-2021	86.400	574,28	164.28	114	468	551	656	1487

Fuente: Elaboración propia

Tabla 2: Máximo y Mínimos Hits de frecuencias XDRs

Estadístico	Fecha hits		Frecuencia
Máximo	01-03-2022	12:23:15	1.969
Mínimo	01-03-2022	05:41:22	94
Mínimo	01-03-2022	05:41:24	94

Fuente: Elaboración propia

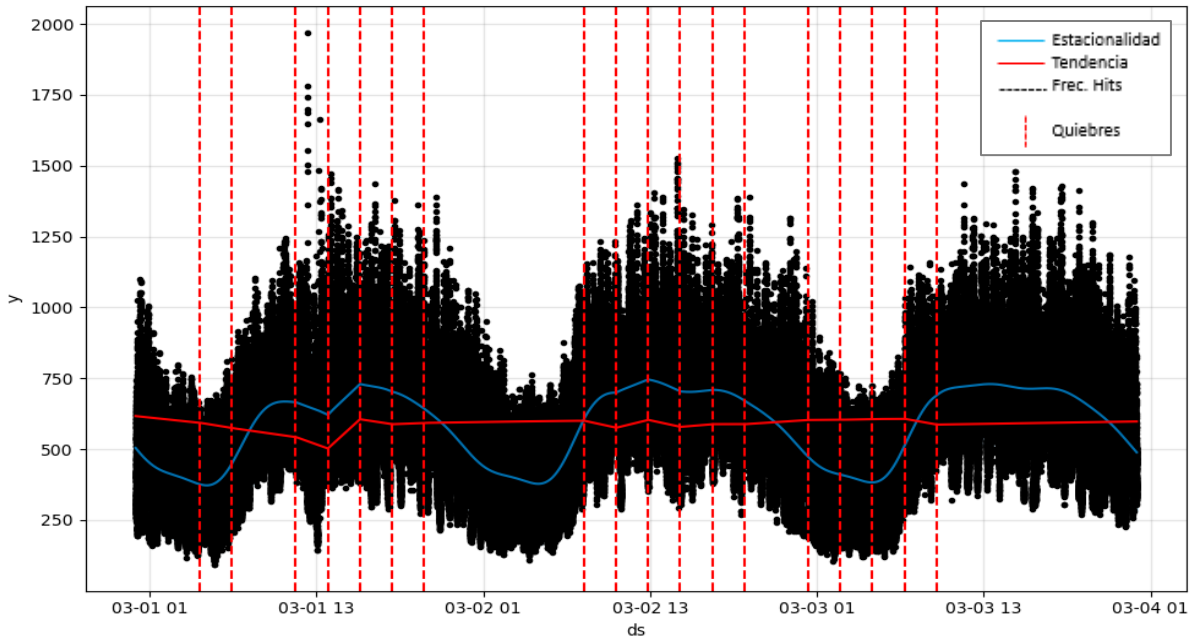
Tabla 3: Frecuencia de Medianas de Hits por día. (año 2022)

Tramo Horario	Jueves	Martes	Miércoles
00:00 a 06:00	398	402	391
07:00 a 11:00	480	450	464
12:00 a 16:00	730	641	723
17:00 a 20:00	694	692	694
21:00 a 23:00	633	630	627

Fuente: Elaboración propia

5.2.Estudio de Serie temporal martes 2 a jueves 3 de marzo 2022

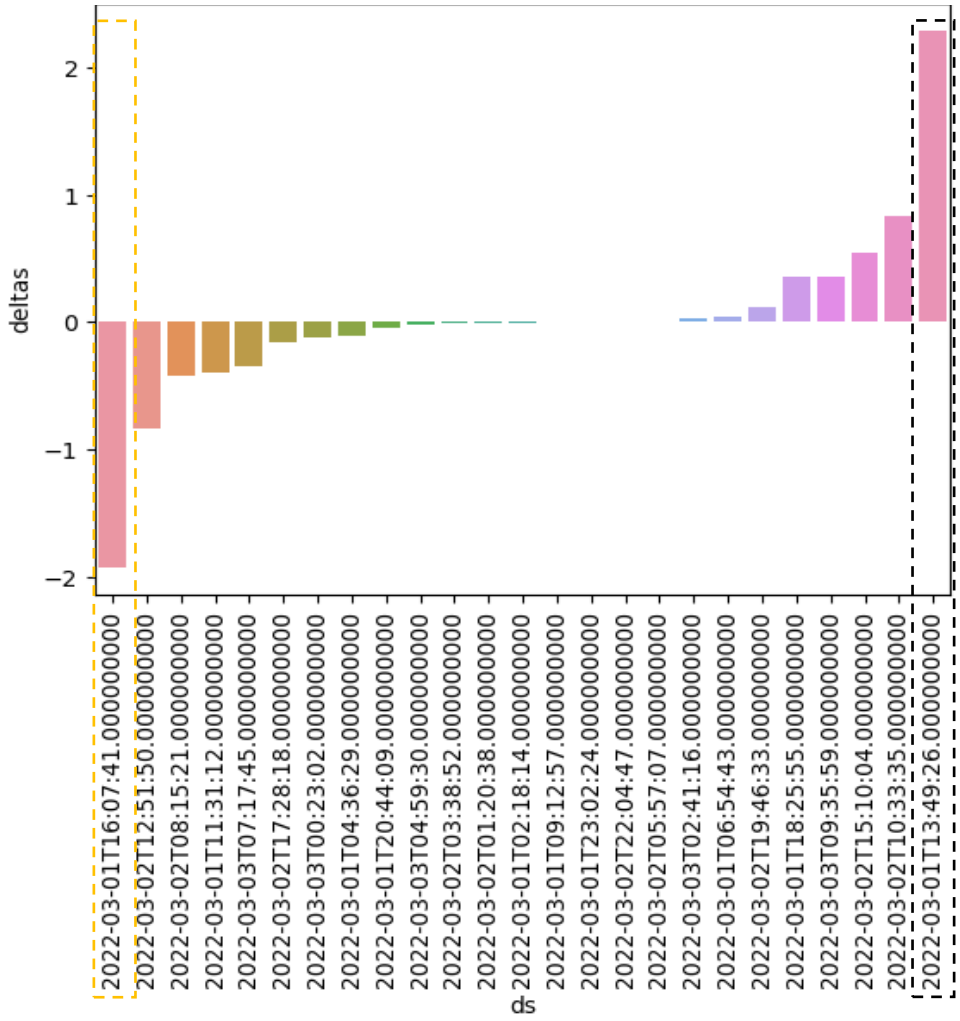
Gráfica 6: Descomposición por puntos de quiebre y tendencias.



El gráfico 6, presenta la descomposición de la serie temporal de hits XDRs diarios mostrando tendencia constante a largo plazo y componente estacional cíclica. Adicional a

ello se presenta quiebres denotadas por las líneas verticales rojas a causa de factores desconocidos.

Gráfica 7: Puntos de Quiebre en Tendencia de Serie Temporal XDRs



Gráfica 7, corresponde a puntos de quiebre en la tendencia de la serie temporal, la cual es detectada de forma automática por el modelo Facebook prophet, se aprecia que los valores negativos corresponden a tendencias a la baja en sus respectivos timestamp, mientras que

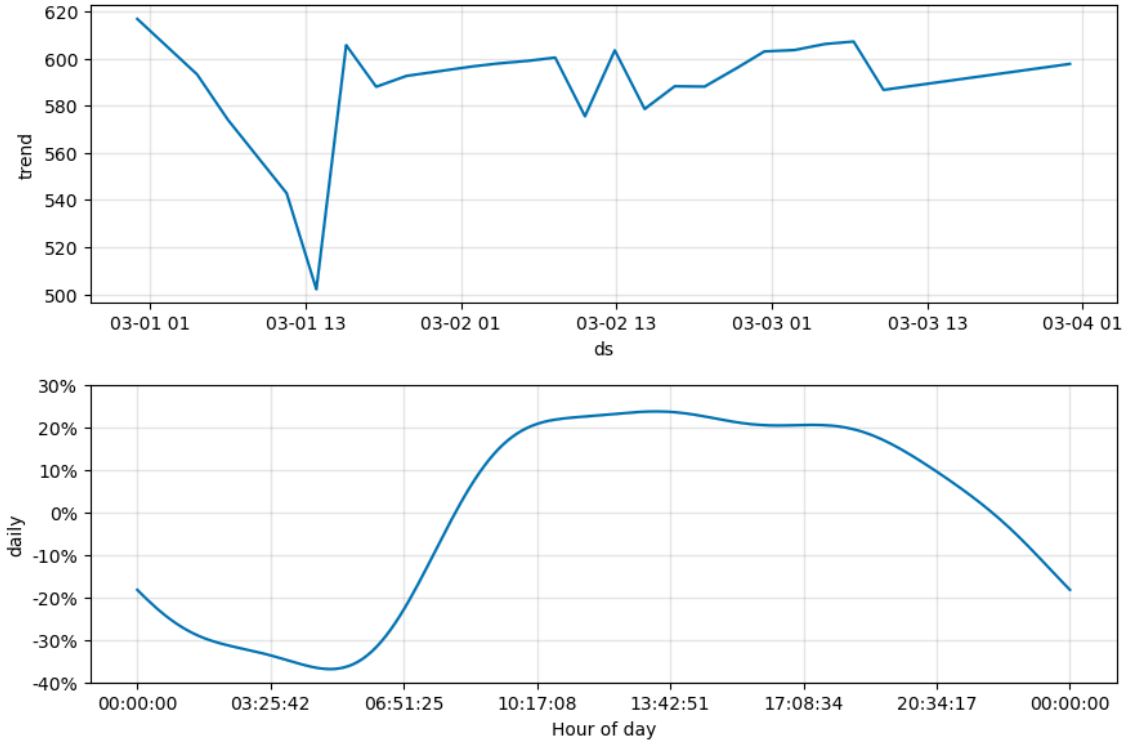
valores positivos corresponden a valores por sobre la tendencia y/o alzas en hits según su respectivo tiempo.

Se valora que durante los tres días estudiados de marzo 2022, el punto de quiebre de mayor hits corresponde al horario de las 13:49 un día martes 1 de marzo 2022, mientras que el día con menor actividad de hits se dio el mismo martes a las 16:07 horas.

5.3.Descomposición serie temporal martes 2 a jueves 3 de marzo 2022.

En este apartado se busca, la detección del tramo horario de mayor flujo según estacionalidad.

Gráfica 8: Tendencia y Estacionalidad

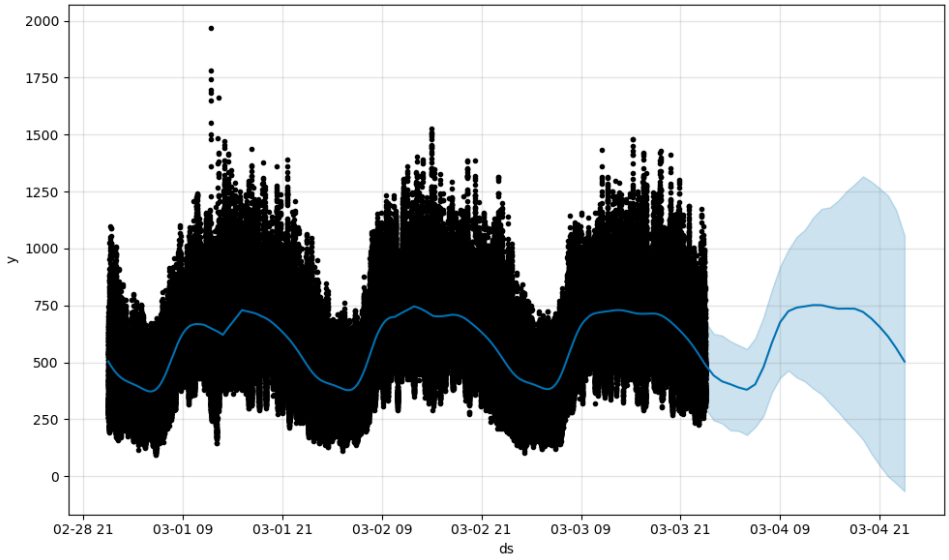


La gráfica 8, ilustra las componentes de tendencia y estacionalidad de la serie temporal de los 3 días analizados del año 2022, en este gráfico (daily) se logra dar respuesta a la interrogante definida en los objetivos de conocer el tramo de mayor actividad de hits durante un día. Se aprecia que el tramo de mayor incidencia corresponde a partir de las 6:51 hrs del día cuyo crecimiento llega en un 20% de aumento a las 10:17 horas, donde dicho 20% se mantiene relativamente constante, luego a partir de las 20:30 horas se aprecia un decrecimiento en la actividad de hits la cual llega al –20% a las 00:00 horas, teniendo el mínimo de actividad cerca de las 03:30 horas.

5.4.Predicción de un día de actividad XDRs.

A partir del martes 1 de marzo al jueves 3 de marzo 2022 se reconstruye el día siguiente mediante la predicción utilizando el modelo Facebook Prophet (Phophet, s.f.).

Gráfica 9: Predicción de viernes 4 de marzo 2022.



Gráfica 10: Tendencia y Estacionalidad de día Predicho

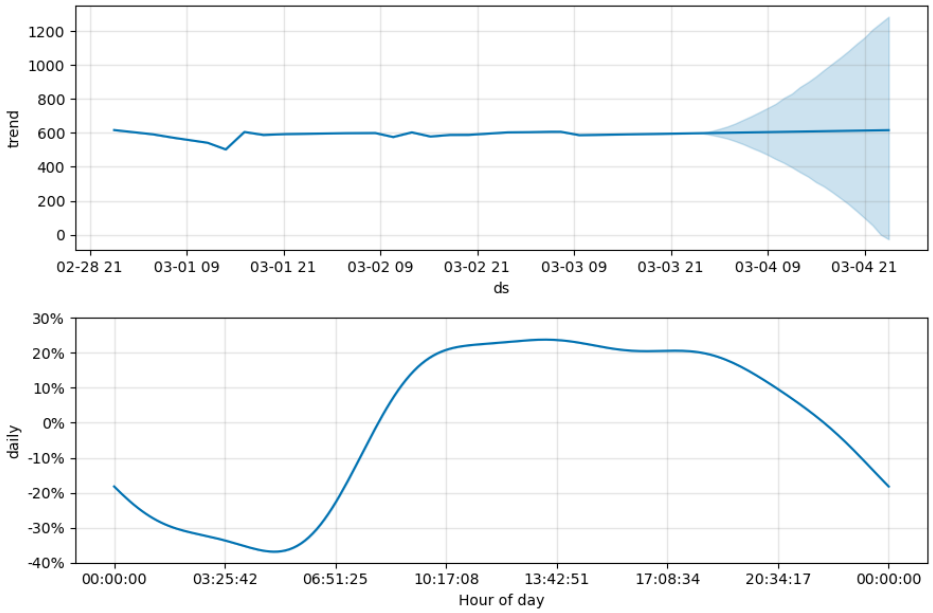


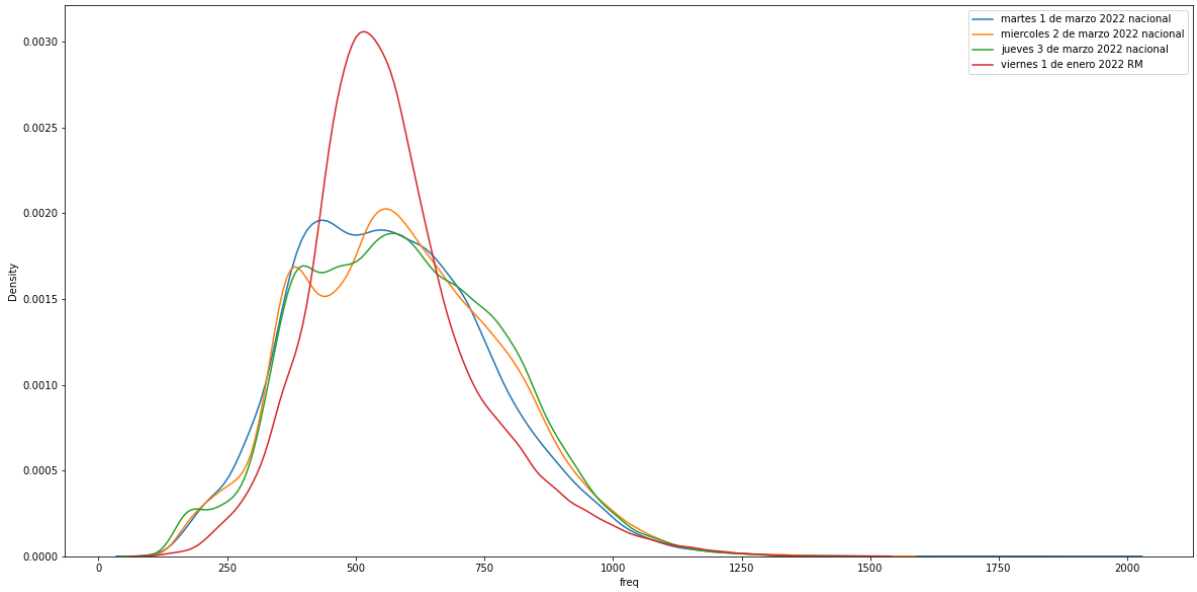
Tabla 4: Resumen de Predicciones

Timestamp	Tendencia	Estacionalidad	predicción
04-03-2022 19:59:00	612	0.129936	692
04-03-2022 20:59:00	613	0.069331	656
04-03-2022 21:59:00	614	-0.000357	614
04-03-2022 22:59:00	615	-0.086287	562
04-03-2022 23:59:00	616	-0.182154	503

Se destaca que la predicción está compuesta por tendencias y estacionalidades como otros factores que el modelo considera internamente. Tabla 4, refleja la predicción cada una hora a partir de las 20:00 horas hasta las 23:59 del viernes 4 de marzo 2022.

5.5. Funciones de densidad

Gráfica 11: Función de densidad por día XDRs.



Se aprecian las distribuciones de características simétricas, no obstante, se refleja de forma consistente la distribución del día viernes 1 de enero 2022 la cual corresponde a la región metropolitana, con características de distribución de hits homogénea con menor varianza que el resto de las distribuciones diarias del resto del país, dicha distribución del viernes posee un coeficiente de variación [1] de 28% siendo esta la distribución de mayor homogeneidad, debido a que es solo una región en particular.

[1] El Coeficiente de Variación es una medida de dispersión que permite el análisis de las desviaciones de los datos con respecto a la media y al mismo tiempo las dispersiones que tienen los datos dispersos entre sí ecuación: $cv = ds/\bar{x}$.

Es interesante mencionar que en las distribuciones diarias (martes a jueves 2022) se reflejan comportamientos bimodales[2] en la distribución de frecuencias de los hits XDRS, lo anterior indicaría que existe un comportamiento que, de alguna forma afecta a los hits en un determinado periodo de tiempo, ya sean problemas con antenas, caídas en el sistema entre otros.

5.6.Pruebas de ajustes de distribución

A continuación, se someten las distribuciones a pruebas de normalidad mediante la aplicación de prueba tales como la prueba de Shapiro-Wilks.

Este plantea la hipótesis nula que la muestra proviene de una distribución normal, es aquí donde elegimos un nivel de significancia del 5% y para su contraste se presenta la hipótesis alternativa que sostiene que la distribución no es normal.

Por lo tanto, tenemos que:

$$Ho:\ La\ distribucion\ es\ nomal\ \ X \sim N(\mu, \sigma^2)$$

H1: La distribución no es normal

- Distribución densidad hits XDRS martes 1 de marzo 2022

```
Estadísticos=0.988, p=0.000
La muestra parece Gaussiana o Normal (no se rechaza la hipótesis nula
H0)
```

[2] Corresponde a una distribución de probabilidad continua con dos diferentes modos, reflejándose dos picos que corresponde a máximos locales en la función de densidad de probabilidades.

- Distribución densidad hits XDRS miércoles 2 de marzo 2022

```
Estadísticos =0.991, p=0.000
La muestra parece Gaussiana o Normal (no se rechaza la hipótesis nula
H0)
```

- Distribución densidad hits XDRS jueves 3 de marzo 2022

```
Estadísticos =0.993, p=0.000
La muestra parece Gaussiana o Normal (no se rechaza la hipótesis nula
H0)
```

- Distribución densidad hits XDRS viernes 1 de enero 2022

```
Estadísticos =0.962, p=0.000
La muestra parece Gaussiana o Normal (no se rechaza la hipótesis nula
H0)
```

Se logra concluir que la distribución de frecuencias de hits XDRs posee un comportamiento ajustado a un fenómeno probabilístico de origen gaussiano.

6. Conclusiones

Mediante los análisis ilustrados en esta etapa que el flujo de mayor actividad de frecuencia de hits XDRS corresponde a partir de las 7 de la mañana hasta las 17 horas relativamente, con un crecimiento paulatino de un 20% en cuanto a su estacionalidad, no obstante, se tramificaron horarios que valoran ciertos periodos del día según cortes de hora valorando percentiles, denotando estadísticas descriptivas dentro de dichos cortes y se encuentra que el tramo de las 12 horas a las 16 horas posee una mayor mediana de frecuencia que el resto, se destaca que dicho tramo solo es superior en decimales al tramo de horario de 17 a 20 horas, conformándose en estos tramos la mayor actividad de hits XDRS a nivel nacional, siendo estos tramos similares al comportamiento de la RM.

Por último, se prueban las distribuciones a ajustes de una normal mediante pruebas de bondad de ajuste resultando todas las distribuciones normales.

Concluyendo finalmente con la estimación del día viernes 4 de marzo del 2022, a partir de los 3 días anteriores estudiados, lográndose una predicción cada una hora. Lo cual ayuda a establecer conocimiento de la actividad diaria del patrón XDRs, y con ello se logra establecer el horario para futuros estudios reconocer la zonas que tienen mayor actividad según XDRs.

7. Bibliografía

(1) Canavos, G. (s.f.). Probabilidad y Estadistica, Aplicaciones y Metodos. En G. Canavos.

(2) Chaiming Song, Z. Q.-L. (2010). *Límites de Previsibilidad en la Movilidad Humana.*

(3) *DASK*. (2014-2018). Obtenido de https://docs.dask.org/en/stable/dataframe.html

(4) Denis Parra, E. G.-G. (2018). Toward Finding Latent Cities with Non-Negative Matrix Factorization

(5) Espejo Miranda, F. P. (2009). Inferencia Estadística. En F. P. Espejo Miranda, *Inferencia Estadísitca Teoria Y problemas* (pág. 206). Universidad de cádiz.

(6) José García, E. G.-G. (2016). *Sensing Urban Patterns with Antenna Mappings: The Case of Santiago, Chile †.* Chile, Santiago.

(7) *Phophet*. (s.f.). Obtenido de https://facebook.github.io/prophet/

(8) Rodrigo, J. A. (2021). *https://www.cienciadedatos.net*. Obtenido de https://www.cienciadedatos.net.

(9) *Sensing Urban Patterns with antenna Mappings.* (s.f.).

(10) Sphinx. (2008-2022). *The SciPy community.* Obtenido de Statistical functions (scipy.stats): https://docs.scipy.org/doc/scipy/reference/stats.html#module-scipy.stats