Análisis de grandes volúmenes de datos de telecomunicaciones con aplicación a la detección de usuarios poco frecuentes

Gisela Villanueva - Raul Marusca
Coordinador: Juan D. Gonzalez

24 de octubre de 2023

Resumen Ejecutivo

Por medio de rutinas computacionales, se analizaron más de 7 millones y medio de comunicaciones. El objetivo es reconocer números telefónicos de dispositivos móviles que resulten infrecuentes estadísticamente, para una fecha y zona de interés determinadas. La lista encontrada se ofrece en las conclusiones de estr trabajo, soportadas por los cuadros 1.3,1.4,1.5. El presente documento describe algunos detalles del procedimiento. Se debe prestar especial atención a que la zona geográfica de los llamados, presuntamente podría no corresponder al lugar indicado y se está a la espera de confirmación. No obstante, los programas desarrollados pueden ser naturalmente reutilizados de manera directa para otros archivos con formato similar al analizado.

Resumen

Se requiere identificar un número telefónico celular que haya operado en las cercanías de una dirección en un determinado periodo de tiempo. Para hacerlo se lleva a cabo el análisis programático de grandes volúmenes de datos provistos por una operadora de telefonía celular que tiene cobertura en la zona. Este reporte preliminar presenta como resultado la obtención de una lista de candidatos que son compatibles con el requisito solicitado, específicamente, que la llamada sea poco frecuente en la zona. Estos reportes cubre una ventana de tiempo mayor a la buscada de manera de encontrar cuales son los números que frecuentan la zona y usar esa información para centrar el estudio en los números poco frecuentes.

1.1. Análisis

1.1.1. Conjunto de Datos

Los datos en su versión más cruda corresponde a 10??? planillas de cálculo similares excels en formato ods. Conteniendo numero de registros totales de mas de 7 millones y medio. El volumen de la información es demasiado grande como para hacer analisis por medio de la observación manual de los archivos y se desarrollo la funcionalidad necesaria para su lectura por medio de rutinas computacionales en `Python` y `R` que son dos lenguajes de programación muy utilizados en la Ciencia de Datos.

Poner Screenshots de la tabla e indicar el significado de los campos (breve) Explicar sobre todo el término celda (investigar un poco acerca de la regularidad geométrica) si tienen forma rectacngular o no definida geometricamente...:

1.1.2. Períodos analizados

El hecho ocurrió el día 7 de septiembre de 2023 aproximadamente a las 20:45. Los reportes suministrados por la compañía de comunicaciones abarcan desde el día 7 de agosto de 2023 a las 0 horas, hasta el día 13 de septiembre del mismo año a las 23:59:59. Se consideran dos posibles *ventanas* de tiempo centradas en el momento del hecho: una de 40 minutos (desde las 20:25 hasta las 21:05) y otra de 10 minutos (de las 20:40 a las 20:50)

1.1.3. Tipos de comunicación

El reporte comprende dos tipos de comunicaciones de acuerdo al sentido de la misma:

- Llamadas SALIENTES
  Son las comunicaciones realizadas desde un dispositivo que se halla conectado a la celda reportada.
- Llamadas ENTRANTES
  Son las comunicaciones recibidas por un dispositivo conectado a la celda reportada.

1.2. Destinos de la comunicación

Las comunicaciones pueden ser desde o hacia:

- Una comunicación de datos, por ejemplo cuando el usuario navega en internet.

- Un teléfono de la misma compañía.
- Un teléfono de otra compañía.
- Un servicio de atención al cliente.
- Números gratuitos.
- Un teléfono en una provincia argentina.
- Un teléfono en otro país.
- Un servicio (como WhatsApp Free) que se brinda de forma gratuita.
- Un teléfono de línea.

1.2.1. Cantidad de llamados

En el total del período reportado se realizaron un total de 6.986.719 comunicaciones salientes y 402.632 entrantes. En el total del período reportado se contabilizan 69.414 números diferentes realizando comunicaciones salientes. Operando cada uno de ellos entre 1 y 79.226 comunicaciones por dispositivo.

Cuando se considera una ventana de tiempo de 40 minutos (20:25-21:05) se encontraron 5.893 comunicaciones salientes y 391 entrantes, mientras que en la ventana de tiempo más pequeña de 10 minutos (20:40-20:50) se realizaron 1.474 comunicaciones salientes y 112 entrantes.

En el total del período reportado se contabilizan 36.828 números diferentes realizando comunicaciones entrantes. Recibiendo cada uno de ellos entre 1 y 3856 llamados por dispositivo.

1.3. Determinación de las comunicaciones no frecuentes

Considerando la distribución tan extrema de llamadas por dispositivos, se nos plantea el inconveniente de definir qué es un *número NO frecuente*.
En el caso de los salientes podemos hacer expresar esto con la siguiente tabla:
Y en el caso de los entrantes

Cantidad de dispositivos	Cantidad de llamadas
16979	1
8396	2
4178	3
3127	4
32680	4 o menos
69414	Todos

Cuadro 1.1: Distribución comunicaciones salientes

Podemos notar que en el caso de las comunicaciones salientes, si consideramos que los dispositivos NO frecuentes son los que realizaron 4 o menos llamadas, ya tenemos casi la mitad de los dispositivos registrados en las celdas consideradas.

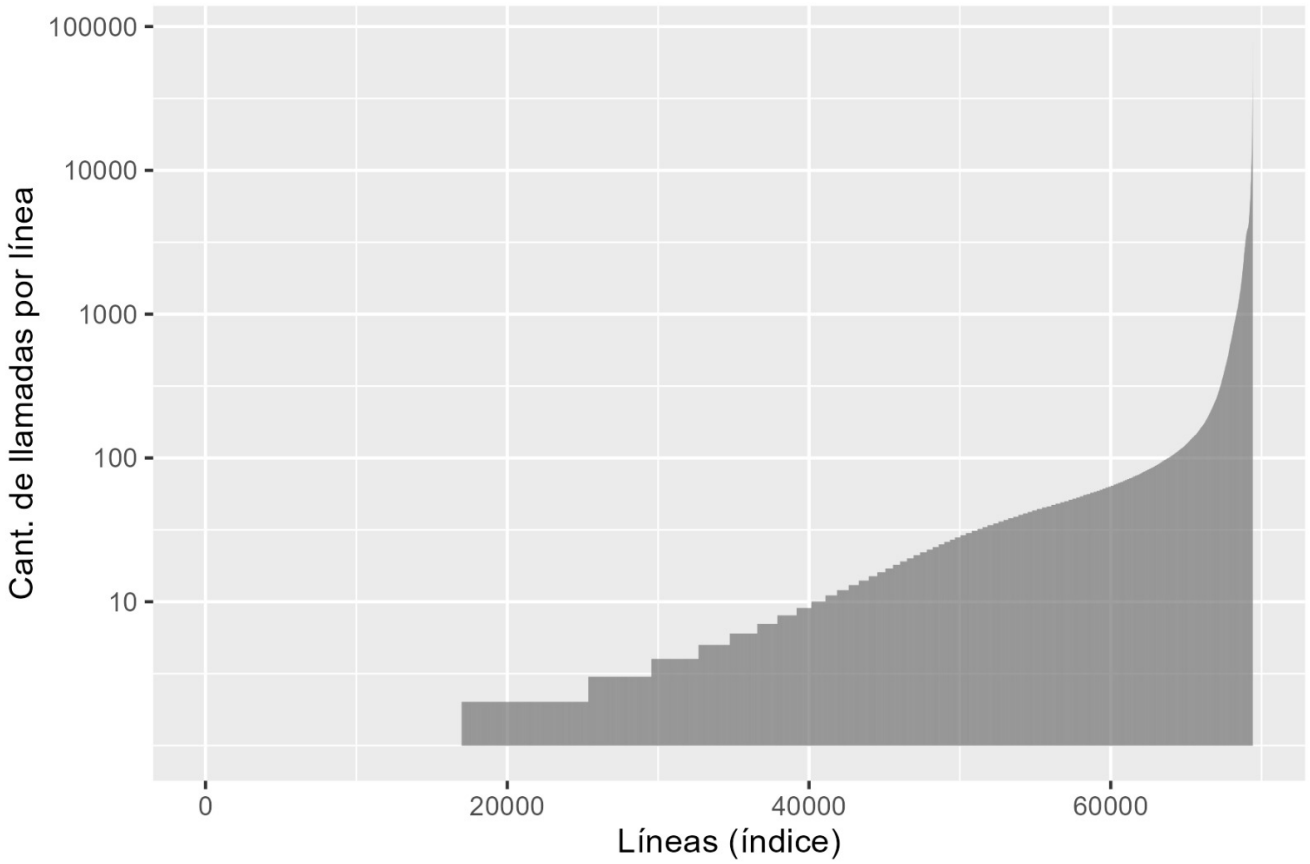

Figura 1.1: Distribución de comunicaciones salientes

Y lo mismo ocurre con los dispositivos entrantes: considerando los dispositivos que recibieron 3 o menos comunicaciones ya tenemos mas de la mitad del total.

Esto indica que hay pocos dispositivos que realizan o reciben una gran cantidad de comunicaciones y muchos dispositivos que se activan dentro de las celdas consideradas.
Por lo tanto, es muy probable que los dispositivos de interés estén dentro de los que operan **1 o 2** comunicaciones dentro de las celdas.

Comunicaciones NO frecuentes en las ventanas temporales

Eliminando los dispositivos no frecuentes dentro de la ventana de $\pm$ 20 minutos, podemos observar que se realizan 18 comunicaciones salientes y 17 entrantes.
Y en la ventana de $\pm$ 5 minutos hay 5 salientes y 7 entrantes.

Cantidad de dispositivos	Cantidad de llamadas
10443	1
5111	2
2761	3
2143	4
18315	3 o menos
36828	Todos

Cuadro 1.2: Distribución comunicaciones entrantes

Fecha y hora	Dispositivo Emisor	Dispositivo Receptor	Tipo comunicación	Duración (s)
2023-09-07 20:26:33	1122747186	1156095002	LLAM. LOCAL	15
2023-09-07 20:27:50	1168441770	1155812801	LLAM. LOCAL	25
2023-09-07 20:29:03	1122363386	158759006	Conexión Móvil	336698679
2023-09-07 20:29:14	1168310005	841310004	Conexión Móvil	10996
2023-09-07 20:32:01	1134883024	2836713025	Conexión Móvil	26558134
2023-09-07 20:32:34	1127975952	1309311007	Conexión Móvil	42925688
2023-09-07 20:42:05	1131804484	348303885	Tráfico incluido	2130848
2023-09-07 20:42:14	1131804484	348303886	TRAFICO DATOS	3997
2023-09-07 20:44:12	1138863797	1964713016	Conexión Móvil	14236203
2023-09-07 20:49:02	1157345937	2747	ASIST.RET	24
2023-09-07 20:50:27	1137848153	2216611013	Conexión Móvil	122436479
2023-09-07 20:51:10	1144751313	1153119614	LLAM. LOCAL	12
2023-09-07 20:56:37	1162335309	159413025	Conexion Movil	9925536
2023-09-07 21:00:22	1170286582	81469005	Conexion Movil	484
2023-09-07 21:01:28	1161686108	1121684935	LLAM. LOCAL	82
2023-09-07 21:02:31	1170286582	1909412029	Conexion Movil	188615353
2023-09-07 21:03:29	1123488404	1137664349	LLAM. LOCAL	2
2023-09-07 21:04:56	1132626386	2633212006	Conexión Móvil	4578116

Cuadro 1.3: SALIENTES: Dispositivos no frecuentes en ventana de 20 minutos

Fecha y hora	Dispositivo Emisor	Dispositivo Receptor	Tipo comunicación	Duración (s)
2023-09-07 20:42:05	1131804484	348303885	Tráfico incluido	2130848
2023-09-07 20:42:14	1131804484	348303886	TRAFICO DATOS	3997
2023-09-07 20:44:12	1138863797	1964713016	Conexión Móvil	14236203
2023-09-07 20:49:02	1157345937	2747	ASIST.RET	24
2023-09-07 20:50:27	1137848153	2216611013	Conexión Móvil	122436479

Cuadro 1.4: SALIENTES: Dispositivos no frecuentes en ventana de 5 minutos

Fecha y hora	Dispositivo Emisor	Dispositivo Receptor	Tipo comunicación
2023-09-07 20:25:08	1123841860	1150569808	48
2023-09-07 20:31:39	1160363074	1169948365	57
2023-09-07 20:33:08	1127224501	1165212800	65
2023-09-07 20:33:27	1155082345	1140633838	52
2023-09-07 20:34:28	1136542554	1126793382	24
2023-09-07 20:34:30	1136175661	1138700444	2
2023-09-07 20:40:55	1156294118	1164678323	365
2023-09-07 20:42:00	1133243077	1167686403	1
2023-09-07 20:42:29	1160191720	1169546648	43
2023-09-07 20:44:07	1160286644	1165654808	83
2023-09-07 20:47:38	1164165900	3813023189	5116
2023-09-07 20:48:45	1131910645	1158168696	148
2023-09-07 20:49:44	1157662720	1161190634	164
2023-09-07 20:52:42	1161234760	2234239322	7
2023-09-07 20:54:39	1132441220	1136313481	19
2023-09-07 20:57:39	1126860252	1170361319	71
2023-09-07 20:58:14	1139252329	1130921657	126

Cuadro 1.5: ENTRANTES: Dispositivos no frecuentes en ventana de 20 minutos

Fecha y hora	Dispositivo Emisor	Dispositivo Receptor	Tipo comunicación
2023-09-07 20:40:55	1156294118	1164678323	365
2023-09-07 20:42:00	1133243077	1167686403	1
2023-09-07 20:42:29	1160191720	1169546648	43
2023-09-07 20:44:07	1160286644	1165654808	83
2023-09-07 20:47:38	1164165900	3813023189	5116
2023-09-07 20:48:45	1131910645	1158168696	148
2023-09-07 20:49:44	1157662720	1161190634	164

Cuadro 1.6: ENTRANTES: Dispositivos no frecuentes en ventana de 5 minutos

Conclusión

A partir de la información suministrada por la compañía proveedora de servicios de telefonía celular, se procede a determinar qué números de dispositivo no son de operatoria frecuente en la zona.
Obtenidos esos dispositivos, se comprueba cuáles de ellos fueron operados de forma saliente o entrante durante unas ventanas de tiempo alrededor de la hora del hecho.
De esas comunicaciones se descartan las que son repetidas y las que son a un servicio de números cortos de la compañía celular (buzón de voz), llegándose a una lista de 32 identificadores de dispositivo que serían de interés.

Los números de interés son:

- 1157662720
- 1132441220
- 1164165900
- 1127975952
- 1123488404
- 1168441770
- 1136175661
- 1122747186
- 1160286644
- 1138863797
- 1127224501
- 1160363074
- 1131804484
- 1123841860
- 1133243077
- 1161234760
- 1162335309
- 1134883024
- 1144751313
- 1132626386
- 1137848153

- 1136542554
- 1161686108
- 1126860252
- 1156294118
- 1160191720
- 1155082345
- 1139252329
- 1168310005
- 1170286582
- 1131910645
- 1122363386

Los números en rojo son los correspondientes a la ventana de $\pm$ 5 minutos.