Estudio de los Contactos Realizados al Sistema Único de Atención Ciudadana (SUACI)

Budziñski, Maia Ludmila[1]; Buttafuoco, Fernando[2]; Rodriguez, Joaquin[3]

[1]Facultad de Ciencias Naturales y Exactas, Universidad de Buenos Aires (UBA); [2]Facultad de Ingenieria, Instituto Tecnológico de Buenos Aires (ITBA;, [3]Facultad Regional Buenos Aires, Universidad Tecnológica Nacional (UTN)

Resumen

En este estudio, se utilizó el dataset 'Información de los contactos realizados al Sistema Único de Atención Ciudadana (SUACI). Se realizó un Análisis Exploratorio de Datos de la información proveniente de los contactos al SUACI desde el 2017 al 2019. Observamos que la mayoría de las variables analizadas mostraron una tendencia similar año a año. Partiendo de esta observación, decidimos corroborar si existe alguna relación entre categorías de consultas del año y ubicación geográfica, para lo cual decidimos realizar un experimento de *clustering*. La aplicación del modelo, a pesar de arrojar muy buenos resultados de clusterización, no permitió una clara distinción de los clusters a nivel geográfico.

Palabras Clave

Sistema Único de Atención Ciudadana, Cluster Analysis, Trámites, K-means.

1. INTRODUCCIÓN

Sistema Único de Atención Ciudadana (SUACI) es el sistema oficial del Gobierno de la Ciudad a través del cual los ciudadanos tienen la posibilidad de efectuar solicitudes, reclamos, denuncias, quejas o reportes respecto a distintos servicios de la ciudad. El siguiente trabajo presenta fundamentalmente dos objetivos: la exploración de datos de la información proveniente de los contactos al SUACI desde el 2017 al 2019 y la inferencia de propiedades y estructuras de la distribución de los datos del año 2019, buscando encontrar grupos de muestras que posean alta relación entre sí mediante Cluster Analysis utilizando el algoritmo K-means.

2. DESCRIPCIÓN DEL DATASET

En este trabajo, se utilizó el dataset 'Información de los contactos realizados al Sistema Único de Atención Ciudadana (SUACI)'. Los datos incluidos en este estudio, provienen del reservorio de datos abiertos del GCBA (https://data.buenosaires.gob.ar/dataset/sistema-unico-atencion-ciudadana). Se utilizaron las bases con los registros correspondientes a los años completos 2017 y 2018, y parcial 2019 (Enero – Agosto). Estas se encuentran en formato *.csv*, correspondiendo un archivo para cada año. En ellas se encuentran todos los contactos generados en la Ciudad de Buenos Aires. En total, para los 3 años, se realizaron aproximadamente 2,2 millones de contactos. En cada uno de estos, se relevan los siguientes conceptos: contacto, fecha de ingreso, prestación, categoria, sub-categoria, tipo de contacto, comuna, barrio, calle, altura, esquina próxima, latitud y longitud.

3. ANÁLISIS EXPLORATORIO DE DATOS

Antes de comenzar el análisis, los datasets fueron limpiados de datos nulos utilizando el siguiente criterio:

Eliminación de variable (feature): variable con más del 50% de registros nulos de datos nulos.

Eliminación de instancia (sample): siempre y cuando no implique eliminar más del 20% del total de instancias del dataset.

De un análisis preliminar de las variables: Canal, Género, Estado, Tipo De Prestación, Evolución Temporal Comunas, Categoría, Sub-Categoría, Concepto y Tiempo De Resolución., se derivaron las siguientes observaciones generales:

En cuanto al **CANAL** de contacto, se observa que los medios de comunicación utilizados por los usuarios para comunicarse con el SUACI son tres (vía Web, Aplicación móvil y telefónicamente al 147). Además, se evidencia un aumento año a año del uso de la aplicación móvil como medio de comunicación predominante.

Figura 1. Countplot de los Medios de Comunicación utilizado por los usuarios para contactarse con el SUACI separados por Año (2017-2019).

En cuanto al **GÉNERO** de los contactos, se observa una clasificación en masculino, femenino o desconocido (la cual puede ser atribuida a que el usuario no se siente representado con ninguna de las opciones ofrecida o bien se trata de un dato faltante). Al respecto observamos que la cantidad de contactos de hombres, en los últimos años, ha superado la de las mujeres.

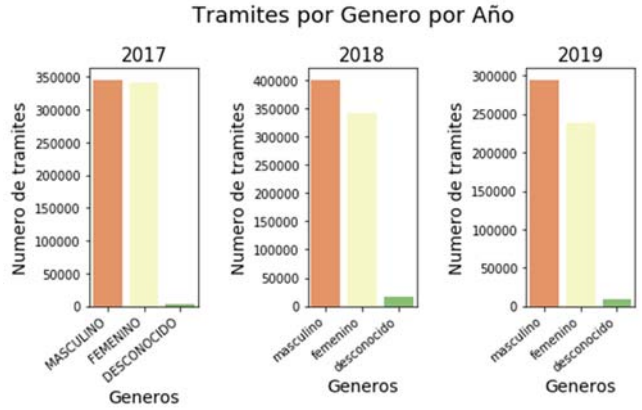

Figura 2. Countplot de Trámites registrados por el SUACI, separados por Género y por Año (2017-2019).

En cuanto al **ESTADO** de los trámites, El dataset no otorga gran información sobre la resolución de los contactos; las únicas medidas relacionadas a ello, son el estado (abierto o cerrado) y las fechas de inicio y fin del contacto. El que aún hayan contactos abiertos para 2017 y 2018 sugiere poca confiabilidad en este dato. De todas formas no hemos profundizado análisis en este tema (quedará para futuros análisis)

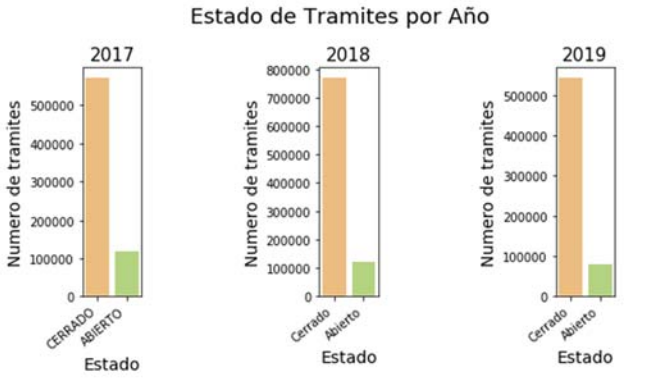

Figura 3. Countplot de los Trámites registrados por el SUACI, separados por el Estado de los mismos y por Año (2017-2019).

En cuanto al **TIPO DE PRESTACIÓN** de los trámites de los tres años analizados, vemos que predominan las solicitudes (pedido o reclamo al GCBA sobre una materia de su competencia y por la cual debe realizar una acción específica), seguidos por servicios (solicitud o consulta en base a un servicio prestado en forma directa por el GCBA) y denuncias (pedido de inspección o corroboración de una infracción a la normativa vigente, realizada por un tercero) y, en última instancia, se encuentran las quejas (agradecimiento, sugerencia, felicitación o queja relacionado con algún servicio o proceso prestado por el GCBA).

Esta clasificación resulta confusa, dado que un contacto podría indistintamente clasificarse como prestación u otra. Por ejemplo “RESTOS DE OBRA EN VEREDA QUE IMPIDEN EL PASO” está clasificada como SOLICITUD y “RETIRO DE ESCOMBROS” como servicio.

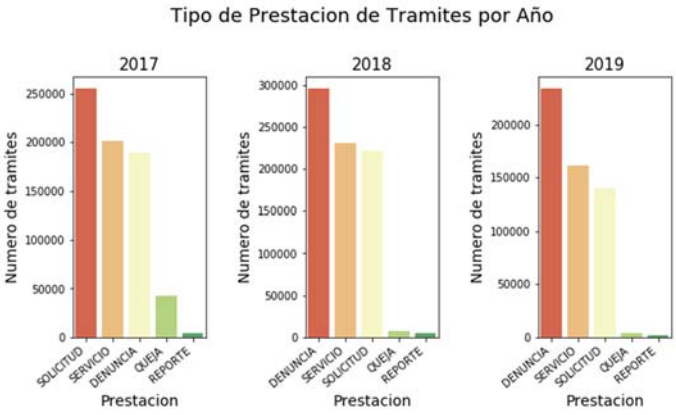

Figura 4. Countplot de los Trámites registrados por el SUACI, separados por Tipo de Prestación de los mismos y por Año (2017-2019).

En cuanto a la **EVOLUCIÓN TEMPORAL** de los trámites, en primer lugar se evidencia un aumento interanual en todos los meses. Esto podría deberse al notable crecimiento del uso de la aplicación móvil. Sumado a esto, en febrero se observa una caída de los contactos y un posterior repunte en los meses siguientes.

.

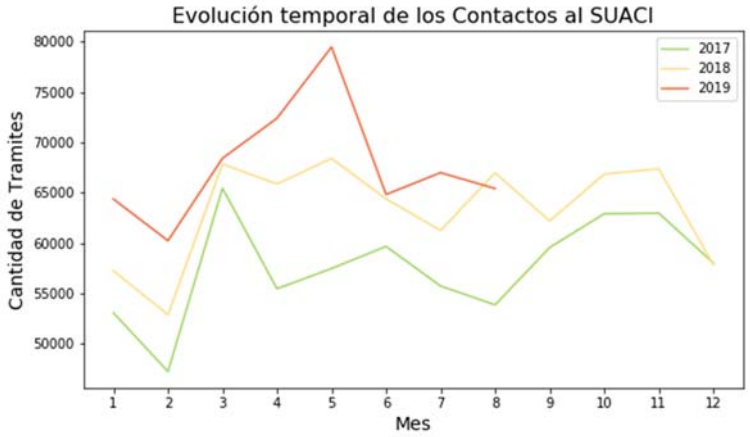

Figura 5. Cantidad de Trámites registrados por el SUACI en cada Mes, separados por Año (2017-2019).

Respecto a los trámites por **COMUNAS**, el análisis evidencia en primera instancia un comportamiento muy estable a lo largo de los 3 años. Adicionalmente, se observa que las 5 comunas con mayor número de contactos son las ubicadas en la zona norte de CABA.

.

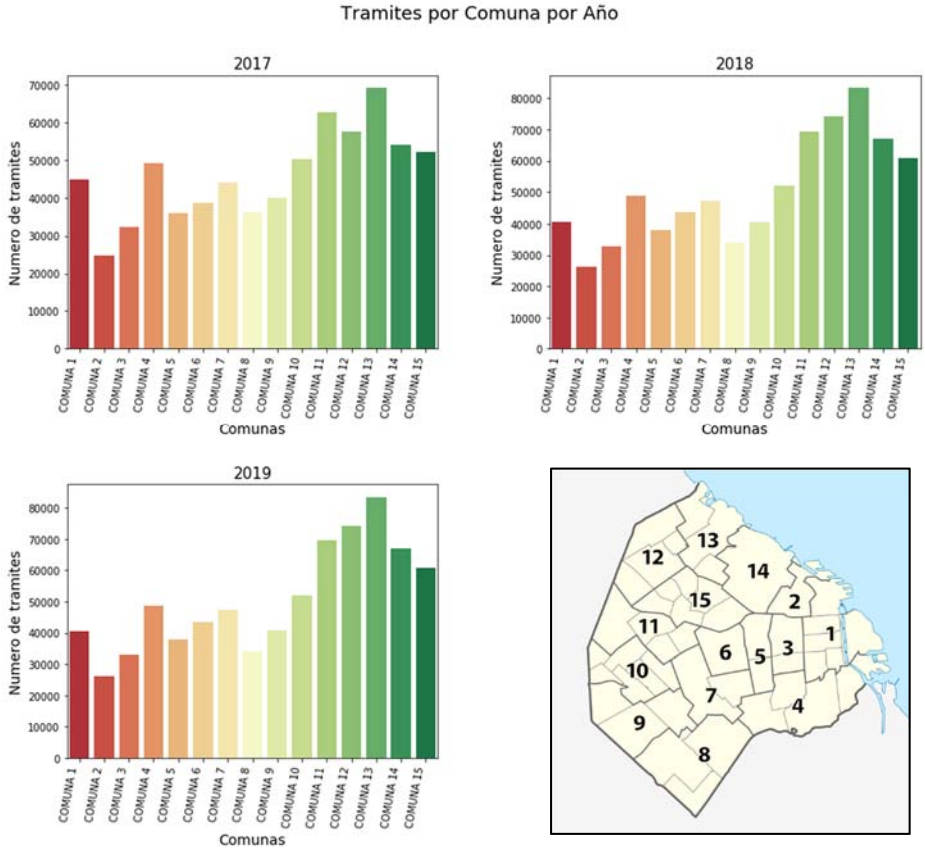

Figura 6. Countplot de los Trámites registrados por el SUACI, separados por Comuna y por Año (2017-2019) y mapa de Capital Federal de Buenos Aires, seccionado por Comunas como referencia.

Analizando en detalle los trámites por Categorías, Sub-Categorías y Conceptos, observamos que, sumando todos los contactos por categoría, existe una significativa concentración de contactos en las **CATEGORÍAS** “LIMPIEZA Y RECOLECCIÓN” y “TRANSITO”.

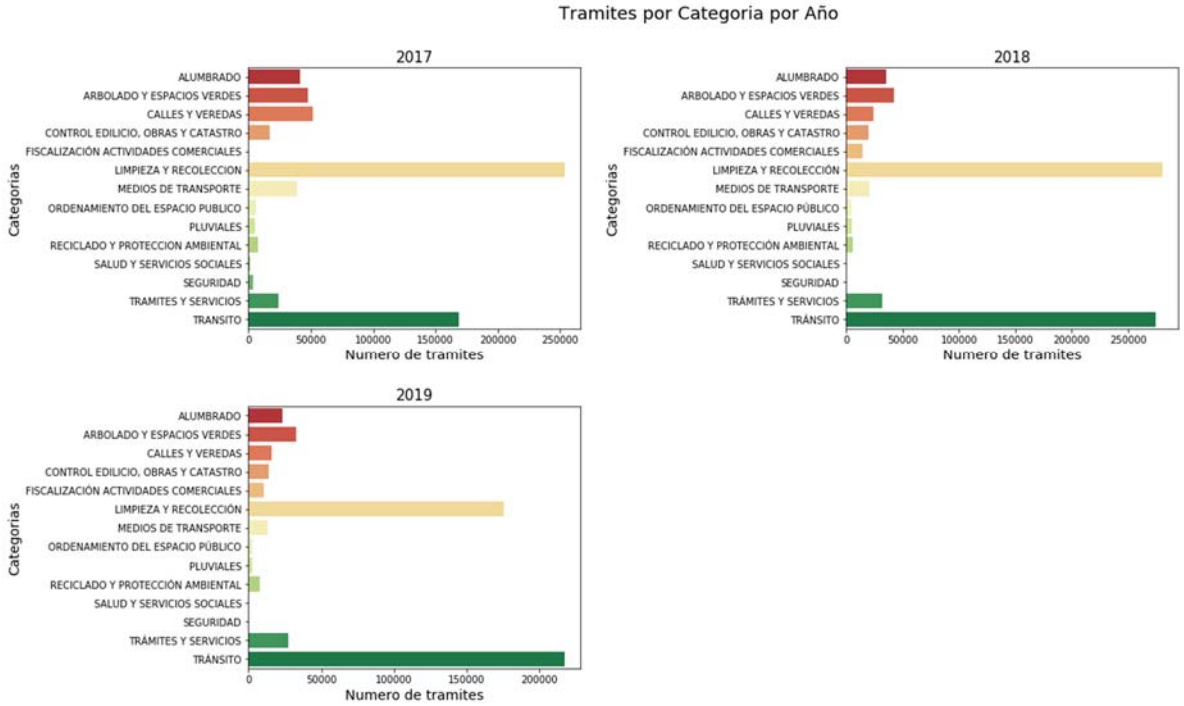

Figura 7. Countplot de los Trámites registrados por el SUACI, separados por Categoría y por Año (2017-2019)

Los contactos realizados al SUACI se catalogan en primer lugar en 213 conceptos, que es el nivel más bajo de categorización. Esos conceptos se agrupan en 58 subcategorías, las cuales a su vez pertenecen a 17 categorías diferentes, siendo éste último el orden superior.

Al tomar las Categorías predominantes, observamos que la mayor parte de los contactos se concentran en las **SUB-CATEGORÍAS** “RESIDUOS VOLUMINOSOS” Y “DENUNCIA VIAL” en los tres años analizados.

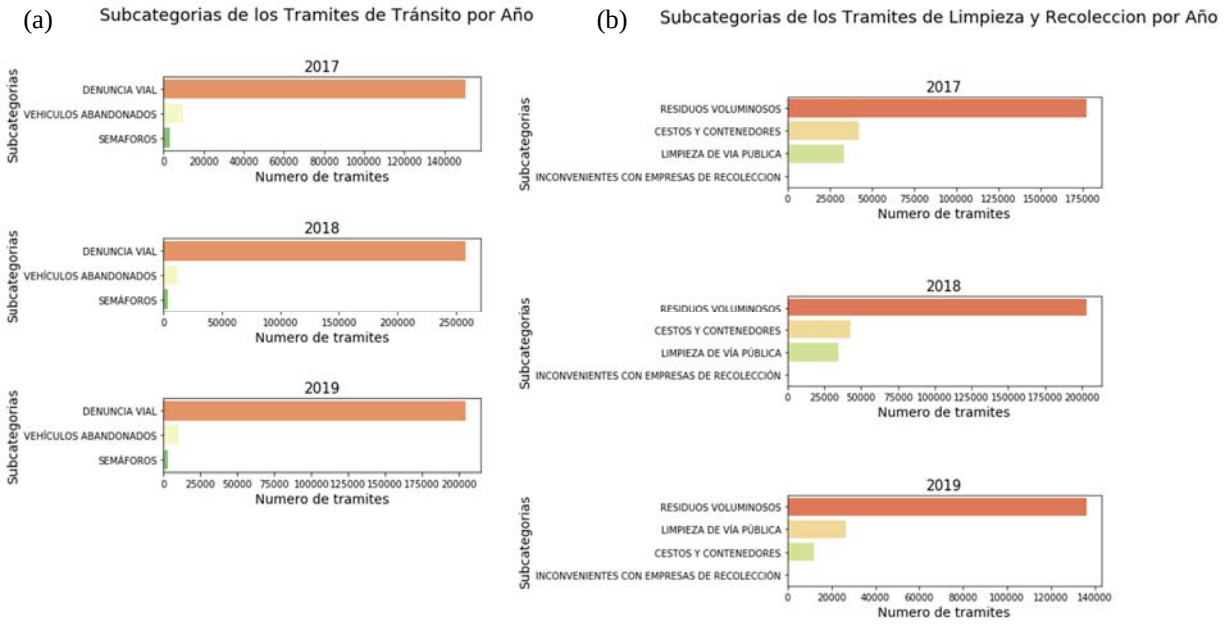

Figura 8. Countplot de los Trámites registrados por el SUACI, separados por Año (2017-2019) y (a) la Sub-Categoria Limpieza y Recolección o (b) la Sub-Categoria Transito.

Finalmente, al tomar sólo las principales Sub-categorías predominantes, observamos que año a año los trámites se concentran en los **CONCEPTOS** "RETIRO DE ESCOMBROS" y "VEHICULO MAL ESTACIONADO".

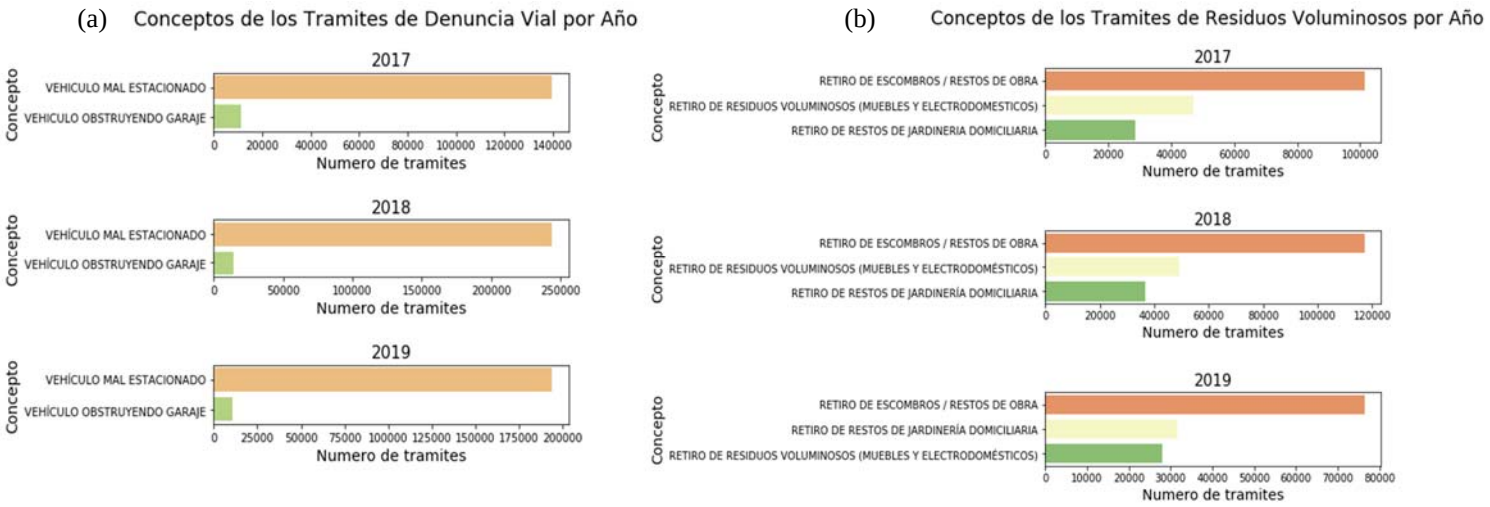

Figura 9. Countplot de los Trámites registrados por el SUACI, separados por Año (2017-2019) y (a) el Concepto Tramites de Denuncia Vial o (b) el Concepto Residuos Voluminosos.

Sugerimos para futuros análisis, segregar estos Conceptos del resto y analizar ambos grupos de trámites (con y sin los datos de los trámites de Conceptos de mayor predominancia) por separado, para poder analizar otros Conceptos a que quedan rezagados como consecuencia del gran peso que poseen los Conceptos principales, "RETIRO DE ESCOMBROS" y "VEHICULO MAL ESTACIONADO".

Para identificar los **TIEMPO DE RESOLUCIÓN** de los contactos, creamos una nueva feature a partir de la diferencia entre entre la fecha de fin de contacto y la fecha de inicio. Esto solo pudo realizarse para los datos de 2018 y 2019 ya que el dataset del 2017 no contaba con datos sobre la fecha de fin de contacto. Posteriormente, Para profundizar un poco en el análisis, realizamos un boxplot con los datos.

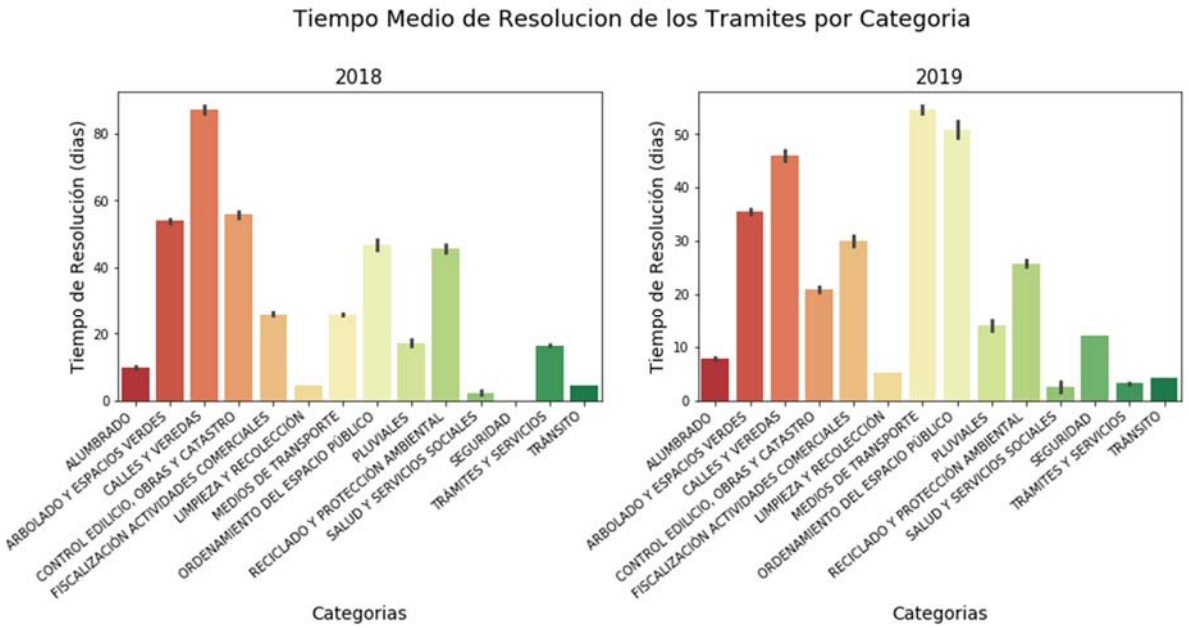

Figura10. Análisis de la media del Tiempo de Resolución de los trámites registrados por el SUACI, separados por Año (2018-2019).

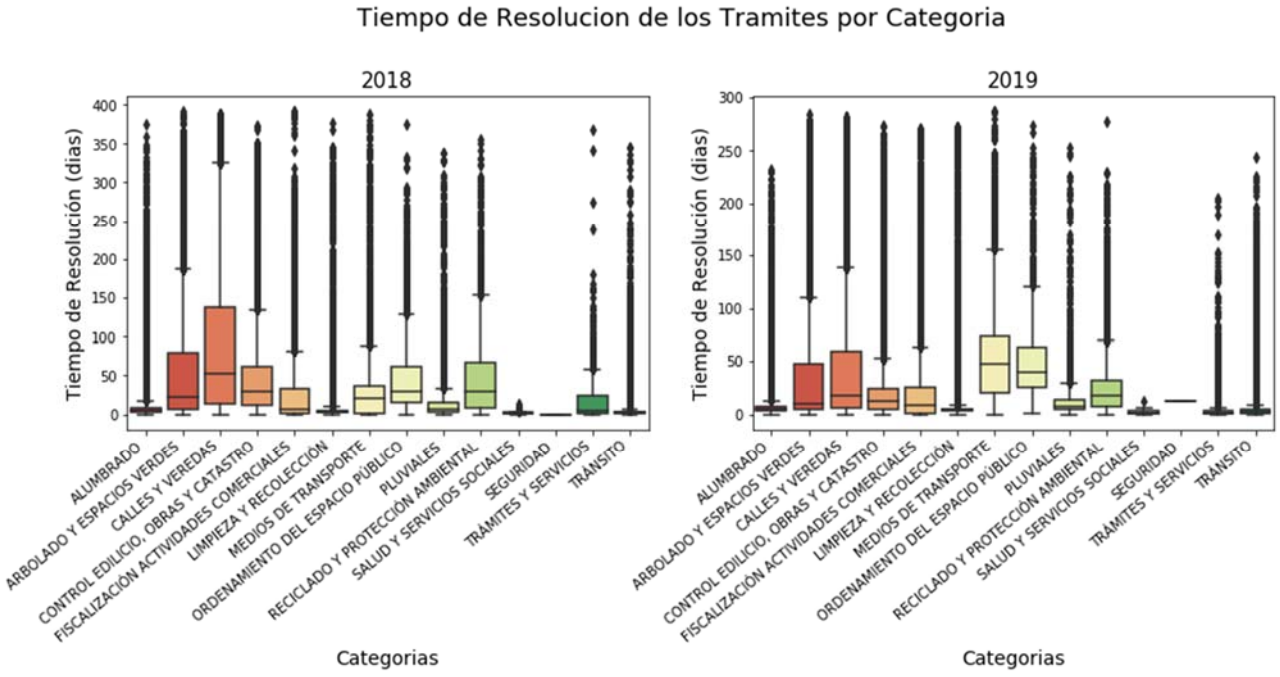

Figura 11. Boxplot del Tiempo de Resolución de los trámites registrados por el SUACI, separados por Año (2018-2019).

Dada la considerable presencia de outliers, no arribamos a ninguna conclusión. Queda para futuros análisis extraer dichos outliers y analizar por separado ambos grupos; sin embargo, queda claro que la media no resulta el parámetro estadístico idóneo para este caso.

Tomando en conjunto el análisis de **COMUNAS** (Figura 6) y **CATEGORIAS** (Figura 7), observamos que tanto la cantidad de contactos por comuna, como la cantidad de contactos por categoría, mantienen año a año una misma conducta; la proporción de contactos de cada categoría con respecto a las otras es idéntica en 2017, 2018 y 2019. De igual manera ocurre con las comunas. Partiendo de esta observación, decidimos corroborar si existe alguna relación entre CATEGORIAS y ubicación geográfica, para lo cual decidimos realizar un experimento de *clustering*.

4. MATERIALES Y MÉTODOS

Para el ejercicio de clusterización nos propusimos trabajar en la identificación de zonas en la Ciudad Autónoma de Buenos Aires con problemáticas similares, definidas a partir de los contactos realizados al SUACI.

Utilizamos a este efecto la variable "categoría" y buscamos: determinar el número de clusters óptimo para el algoritmo K-means, calculado a partir del Silhouette Score, y encontrar el mínimo valor de *k* tal que verifique un Silhouette Score por encima de un umbral de corte determinado. Esto último, con el objeto de obtener una categorización eficiente de zonas de la ciudad con problemáticas similares, dentro de un número de clusters que resulte aún inferior a la división en 15 comunas o 17 categorías

4.1 Pre-procesamiento de los datos

4.1.1 Selección de variables

Considerando el volúmen de datos que se tienen, tomamos únicamente el dataset con los contactos realizados en el año calendario 2019. Tomamos esta decisión a partir de la necesidad de reducir la cantidad de samples por el elevado costo computacional de los algoritmos involucrados. Esto redujo el dataset a aproximadamente 620mil muestras.

Utilizamos la variable "categoria" que aporta una categorización descriptiva del tipo de contacto, con un nivel de apertura adecuado. A efectos del problema propuesto no nos interesa el canal a través del cual se realizó el contacto ni el género del denunciante, así como la fecha en que se registró el mismo, en tanto tomamos el horizonte temporal de todos los contactos realizados durante el año calendario 2019.

Definimos entonces un nuevo dataset con las variables estrictamente importantes para nuestro ejercicio, comprendidas por:

- Las variables dummies obtenidas a partir de la variable categórica "categoría"
- El par latitud-longitud, para que el algoritmo le asigne también un peso a la proximidad geográfica de los contactos a la hora de clusterizar

La dimensión final del dataset a *clusterizar* es (626101, 19)

4.1.2 Estandarización

Las coordenadas latitud y longitud presentan problemas de continuidad (por ejemplo, en el caso del antemeridiano de Greenwich, donde la longitud toma valores de +180 y -180 a ambos lados, representando un salto gigante en la distancia para dos puntos próximos entre sí). Restringiendo el dominio a la Ciudad de Buenos Aires, no tenemos este problema, y por otro lado, al expresarse los valores en decimales, la diferencia máxima de los valores en este dominio es inferior a la unidad. Por tanto, no hay diferencias en el orden de magnitud de la diferencia entre las variables y, en principio, no sería necesario escalar nuestras variables

De todos modos, decidimos realizar un escalado de nuestras variables y guardarlo en un dataset separado del original a fines comparativos.

4.1.3 Principal Component Analysis (PCA)

Realizamos un PCA a partir del dataset original para reducir la dimensionalidad de la matriz y poder trabajar con un número reducido de componentes principales que representaran la mayor variabilidad posible.

4.2 Modelo de Aprendizaje

Utilizamos K-means para realizar un Cluster Analysis de los datos originales, estandarizados y las dos componentes principales que obtuvimos de realizar el PCA sobre los datos originales. K-means es un algoritmo de clasificación no supervisada que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

Inicialización: una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.

Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.

Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso. El algoritmo *k-means* resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster. Los objetos se representan con vectores reales de d dimensiones ($x_1,x_2,\ldots,x_n$) y el algoritmo *k-means* construye k grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo $S=\{S_1,S_2,\ldots,S_k\}$ a su centroide. El problema se puede formular de la siguiente forma:

$$E(\mu_i) = \min_{s} \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \|x_j - \mu_i\|^2 \quad (1)$$

Se obtienen k grupos o clusters con su correspondiente centroide μ_i.

En cada actualización de los centroides, desde el punto de vista matemático, imponemos la condición necesaria de extremo a la función $E(\mu_i)$ que, para la función cuadrática (1) es:

$$\frac{\partial E}{\partial_{\mu_i}} = 0 \Longrightarrow \mu_i{}^{(t+1)} = \frac{1}{|\, S_i^{(t)} \,|} \sum_{x_j \epsilon S_i^{(t)}} x_j$$

y se toma el promedio de los elementos de cada grupo como nuevo centroide.

Para medir la calidad de los clusters se computó el Silhouette Index (S) utilizando la distancia media intra-cluster (a) y la distancia media inter-cluster cercano (b) para cada muestra.

$$S_{(i)} = \frac{(b_{(i)} - a_{(i)})}{max\{(b_{(i)}, a_{(i)})\}}$$

Dónde,

$a_{(i)}$ es la diferencia promedio de un objeto con respecto a todos los demás objetos en el mismo grupo

$b_{(i)}$ es la diferencia promedio del objeto con todos los objetos en el grupo más cercano.

5. RESULTADOS

Aplicamos el algoritmo de K-means sobre 3 bases de datos diferentes:

1. Sobre el dataset original, sin estandarizar
2. Sobre el dataset estandarizado, con media 0 y desvío estándar 1
3. Sobre los dos primeros componentes principales que resultan de aplicar el PCA sobre el dataset original

Al respecto, se realizaron sucesivas clusterizaciones iterando el número de clusters en el rango 2-17, en busca del valor de k que maximiza el Silhouette Score

El costo computacional del algoritmo que determina el Silhouette Score es muy elevado, resultando inviable calcular el mismo sobre la totalidad de los puntos en nuestro dataset para cada una de las iteraciones. Por tanto, optamos por calcular el Silhouette sobre una muestra equivalente al 25% de las samples.

k	**original**	**estándar**	**PCA**
2	0.502577	0.149954	0.694812
3	0.677233	0.276740	0.953700
4	0.741617	0.268479	0.962623
5	0.801972	0.376448	0.945650
6	0.840586	0.370748	0.957260
7	0.871688	0.449204	0.972521
8	0.894982	0.458170	0.978152
9	0.915882	0.520183	0.983389
10	0.930210	0.503786	0.983617
11	0.939197	0.528888	0.984984
12	0.945906	0.559020	0.919577
13	0.950885	0.557645	0.917206
14	0.953608	0.572797	0.751852
15	0.955420	0.574903	0.579532
16	0.955636	0.582280	0.575680
17	0.793940	0.582908	0.553312

Tabla 1. Resultados del Silhouette Score para el número de clusters k, sobre el dataset original, estandarizado o los dos primeros componentes principales que resultan de aplicar el PCA sobre el dataset original

Observamos que los resultados obtenidos sobre el dataset original son consistentemente superiores a los obtenidos sobre la base estandarizada, para todo valor de *k*. De la misma manera, los valores arrojados por el modelo aplicado sobre los componentes principales demostraron ser incluso superiores a los obtenidos sobre el dataset original hasta alcanzar el valor de *k*=11.

En los dos primeros casos, el valor de *k* que maximiza el Silhouette Score es muy similar (16 en el caso del dataset original y 17 para el dataset estandarizado).

Decidimos continuar el análisis sobre la base de componentes principales, por ser la que presentó los mejores resultados (valor de *k*=11). El análisis de componentes principales arrojó resultados contundentes, con las primeras dos componentes explicando un 60% de la variabilidad de los datos del dataset original.

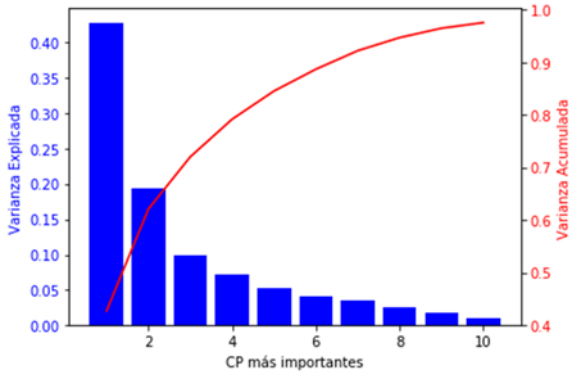

Figura 12. Porcentaje de varianza del datase original explicado por las primeras n componentes principal

Para profundizar nuestro análisis, estudiamos la clusterización sobre los componentes principales para los casos de k=11 (resultado óptimo), y k=3 (correspondiente al menor número de k que verifica Silhouette Score > 0.95).

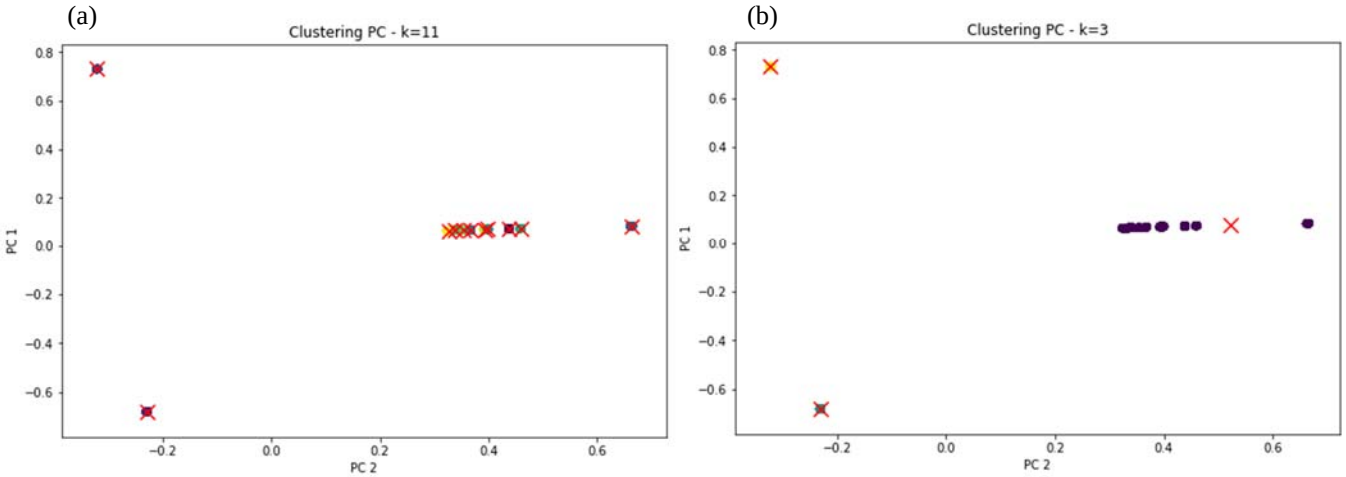

Figura 13. Proyección de los datos sobre el plano definido por las 2 primeras componentes principales. Se aprecia la agrupación de los datos en clusters y la ubicación de los respectivos centroides para (a) *k*=11 o (b) *k*=3

Al partir de la misma matriz de datos, la dispersión de los datos en el plano formado por las dos componentes principales es la misma en ambos casos. Lo que varía es el número de clusters, y por tanto la ubicación de los centroides, identificados con una x. Hay dos conjuntos de datos muy separados del resto, mientras que el resto pareciera ubicarse sobre una recta, próximos entre sí. Para *k*=3, se asigna un cluster propio a los conjuntos de datos que presentan mayor separación del resto (sobre la izquierda del gráfico), y agrupa al resto en un cluster. Al incrementar el número de clusters, se realizan divisiones adicionales sobre éste último conjunto.

Si añadimos las etiquetas otorgadas por el algoritmo K-means al dataset original, y realizamos una consulta agrupando por cluster y categoría, obtenemos las siguientes tablas:

label03	categoria	
0	ALUMBRADO	24350
	ARBOLADO Y ESPACIOS VERDES	47818
	CALLES Y VEREDAS	47715
	CEMENTERIOS	32
	CONTROL EDILICIO, OBRAS Y CATASTRO	16487
	FISCALIZACIÓN ACTIVIDADES COMERCIALES	11018
	MEDIOS DE TRANSPORTE	17549
	ORDENAMIENTO DEL ESPACIO PÚBLICO	4607
	OTRAS	1022
	PLUVIALES	3366
	RECICLADO Y PROTECCIÓN AMBIENTAL	7984
	SALUD Y SERVICIOS SOCIALES	128
	SEGURIDAD	3806
	SUGERENCIAS Y LIBRO DE QUEJAS	1180
	TRÁMITES Y SERVICIOS	27469
1	TRÁNSITO	224426
2	LIMPIEZA Y RECOLECCIÓN	187144

Tabla 2. Agrupación de categorías por cluster, para k=3

label11	categoria	
0	ALUMBRADO	24350
1	TRÁNSITO	224426
2	LIMPIEZA Y RECOLECCIÓN	187144
3	ARBOLADO Y ESPACIOS VERDES	47818
	CALLES Y VEREDAS	47715
4	FISCALIZACIÓN ACTIVIDADES COMERCIALES	11018
5	MEDIOS DE TRANSPORTE	17549
6	TRÁMITES Y SERVICIOS	27469
7	ORDENAMIENTO DEL ESPACIO PÚBLICO	4607
	PLUVIALES	3366
	SEGURIDAD	3806
8	RECICLADO Y PROTECCIÓN AMBIENTAL	7984
9	CONTROL EDILICIO, OBRAS Y CATASTRO	16487
10	CEMENTERIOS	32
	OTRAS	1022
	SALUD Y SERVICIOS SOCIALES	128
	SUGERENCIAS Y LIBRO DE QUEJAS	1180

Tabla 3. Agrupación de categorías por cluster, para k=11

Vemos que el algoritmo K-means clusterizó los contactos de acuerdo a su categoría, sin tener en consideración la posición geográfica de los mismos. A la hora de agrupar categorías en clusters, se agruparon aquellos con menor número de contactos, actuando los centroides como un baricentro de los mismos.

Para finalizar, ploteamos la clusterización para k=3 utilizando las coordenadas de longitud y latitud buscando estudiar una correlación entre la agrupación de los clúster y el origen geográfico de las consultas.

Verificamos una distribución dispersa de los contactos a lo largo y ancho de toda la ciudad. La agrupación definida de los clusters en el plano de los componentes principales no se corresponde con una agrupación distintiva en regiones geográficas.

Figura 14. Ubicación según longitud y latitud de los contactos realizados al SUACI, identificados por cluster de pertenencia

6. DISCUSIÓN Y CONCLUSIONES

En primer lugar, el análisis exploratorio de datos fue de utilidad para relevar aquellas variables confiables que pudieran explicar con claridad el fenómeno analizado. Para ello fue de gran relevancia analizar los años 2017, 2018 y 2019 de manera segregada. Por un lado las variables **ESTADO DEL CONTACTO** (ver figura 3) y **TIPO DE PRESTACIÓN** (Figura 4) parecieron no arrojar información dada la poca confiabilidad de las variables. Por el contrario al observar las variables **CANAL DE CONTACTO** (ver figura 1) y **TRÁMITES POR GÉNERO** (ver figura 2) se evidenciaron variaciones interanuales coherentes que sugieren confiabilidad de los datos.

En segundo lugar, observando la cantidad de contactos por **COMUNA** y por **CATEGORIA**, concluimos en que mantienen año a año una misma conducta. La proporción de contactos de cada CATEGORIA con respecto a la otra es igual en 2017, como 2018 y 2019 (ver figura 7). Lo mismo ocurría con las COMUNAS (ver figura 6). Esto además de sugerir confiabilidad, nos sugirió la aplicación del modelo de *clustering.*

Quedan para futuras investigaciones ahondar el análisis de las variables **TIEMPO DE RESOLUCIÓN** segregando los outliers (ver figura 11), **ESTADO DEL CONTACTO** analizando los contactos abiertos vs cerrados (ver figura 3) y focalizar el análisis en aquellos **CONCEPTOS** con mayor cantidad de contactos (ver figura 9).

La aplicación del modelo de *clustering,* a pesar de arrojar muy buenos resultados de clusterización, no permitió una clara distinción de los clusters a nivel geográfico. Exceptuando la región Noreste de CABA, donde se evidencia una mayor proporción del cluster N° 1, el resto de las regiones de CABA presenta distribución uniforme de los clusters, concluyendo en que no es posible identificar zonas con clusters definidos. Una de las problemáticas a las que nos enfrentamos fue el alto costo computacional del algoritmo que determina el Silhouette Score en sucesivas iteraciones de *k*, lo cual nos llevó a utilizar solo el 25% de las samples para esta intancia. En el futuro, nos gustaría aplicar otros modelos de aprendizaje tanto supervizado como no supervizado sobre los datos del SUACI ya que nos parecio un dataset muy rico en información de .relevancia para la ciudad

7. REFERENCIAS

The Elements of. Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. February 2009. Trevor Hastie · Robert Tibshirani.Jain,

Jain, A.K. (2010) Data Clustering: 50 Years beyond K-Means. Pattern Recognition Letters, 31, 651-666. http://dx.doi.org/10.1016/j.patrec.2009.09.011

Chaturvedi, Nikhil & Rajavat, Anand. (2013). An Improvement in K-mean Clustering Algorithm Using Better Time and Accuracy. International Journal of Programming Languages and Applications. 3. 13-19. 10.5121/ijpla.2013.3402.

Shinde, Sachin V. and Bharat A. Tidke. “Improved K-means Algorithm for Searching Research Papers.” (2014).