UNIVERSIDAD NACIONAL DEL CALLAO

FACULTAD DE INGENIERIA ELECTRICA Y ELECTRONICA

UNIDAD DE INVESTIGACIÓN DE LA FACULTAD DE INGENIERIA ELECTRICA Y ELECTRONICA

INFORME FINAL PROYECTO DE INVESTIGACIÓN

"DISEÑO DE UN ALGORITMO PREDICTIVO PARA MONITOREO TEMPRANO DE REDES DE AGUA POTABLE EN LA CIUDAD DE LIMA, 2019"

AUTOR : Dr. Ing. SANTIAGO LINDER RUBIÑOS JIMENEZ

DOCENTE COLABORADOR : M.Sc. Ing. JUAN ANTONIO APESTEGUIA INFANTES

(Periodo de ejecución: del 01 de junio 2019 al 31 de mayo de 2020)

(Resolución de aprobación N° 608-2019-R)

Callao, 2020

DEDICATORIA

A Dios, por ser el que guía mis pasos cada día, a mi Madre e hijas, por su constante apoyo para que luche por mis objetivos.

AGRADECIMIENTO

Agradezco a Dios por permitirme seguir y lograr mis objetivos, a la Universidad Nacional del Callao por ser la que me enseñó a trazar metas y gracias a mi familia por apoyarme en cada decisión y proyecto.

ÍNIDCE

ÍNDICE **6**
TABLA DE CONTENIDO **8**
RESUMEN **10**
ABSTRACT **11**
INTRODUCCIÓN **12**

CAPITULO I: PLANTEAMIENTO DEL PROBLEMA **13**
1.1 Descripción de la realidad problemática 14
1.2 Formulación del problema 16
1.2.1 Problema general 16
1.2.2 Problema específicos 16
1.3 Objetivos 17
1.3.1 Objetivo general 17
1.3.2. Objetivos específicos 17
1.4 Limitantes de la investigación 18

CAPÍTULO II. MARCO TEÓRICO **19**
2.1 Antecedentes 19
2.2 Marco 24
2.2.1Teórico 24
2.2.2Conceptual 27
2.3 Definición de términos básicos 49

CAPÍTULO III. HIPÓTESIS Y VARIABLES **73**
3.1Hipótesis 73
3.1.1Hipótesis general 73
3.1.2Hipótesis específicas 73
3.2 Definición conceptual de variables 74
3.3 Operacionalización de variables 74

CAPÍTULO IV. DISEÑO METODOLÓGICO 75
4.1 Tipo y diseño de investigación 75
4.2 Método de investigación 75
4.3 Población y muestra 132
4.4 Lugar de estudio y periodo desarrollado 133

CAPÍTULO V. RESULTADOS 135
5.1 Resultados descriptivos 135

CAPÍTULO VI. DISCUSIÓN DE RESULTADOS 147
6.1 Contrastación y demostración de la hipótesis con los resultados 147

CONCLUSIONES 154
RECOMENDACIONES 155
REFERENCIAS BIBLIOGRÁFICAS 156
ANEXO: 159
Anexo: Matriz De Consistencia 159

TABLA DE CONTENIDOS: TABLAS, CUADROS, DIAGRAMAS Y FIGURAS

- **TABLAS**

Tabla Nº 1 Unidad de medidas de un Manómetro 37
Tabla Nº 2 Distrito Con Más Emergencias Reportadas 2016-2019 145
Tabla Nº 3 Tipo De Fugas En Emergencia Reportadas 2016-2019..... 146
Tabla Nº 4 Distrito Con Más Emergencias Reportadas 2017-2020 148
Tabla Nº 5 Tipo De Fugas En Emergencia Más Reportadas 2017-2020..... 149

- **CUADROS**

Cuadro Nº 1 Países con más emisiones de C02 27
Cuadro Nº 2 Correlador Soundsens Especificaciones 48
Cuadro Nº 3 Tuberías comerciales para agua potable y sus diámetros…..... 60

- **FIGURAS**

Figura Nº 1Unidad Móvil de Detección de Fugas 34
Figura Nº 2Instalación de Tubería Matriz 35
Figura Nº 3Manómetro de Glicerina para Fluidos presurizados 36
Figura Nº 4Manómetro de Glicerina especificaciones técnicas..... 38
Figura Nº 5 Geófono de piso mecánico, muy similar a un estetoscopio 39
Figura Nº 6 Geófono electrónico LMIC 40
Figura Nº 7Geófono electrónico XMIC 41
Figura Nº 8Correcto uso de los Geófonos electrónicos XMIC y Lmic 42
Figura Nº 9Funcionamiento Correlador Palmer 43
Figura Nº 10 Correlador Palmer + ... 44
Figura Nº 11 Correlador Palmer + en funcionamiento 45
Figura Nº 12 Correlador AquaScan en funcionamiento 46
Figura Nº 13 Correlador Soundsens 47
Figura Nº 14 Software Correlador Soundsens 49
Figura Nº 15 Detalle de un recibo de facturación de agua 52
Figura Nº 16 Detalle de los 50 Distritos de la Ciudad de Lima 53
Figura Nº 17 Detalle de un NIS ubicado en el plano geo referenciado 54
Figura Nº 18 Detalle del distrito de Lima y sus Sectores 55
Figura Nº 19 Corte Transversal de un sistema de conexión domiciliaria 63
Figura Nº 20 Software de Gestión de Incidencias operativas SGIO 64
Figura Nº 21 Base de Datos de Fugas 2014 – 2019 65
Figura Nº 22 Instalación del Software R 78
Figura Nº 23 Instalación del Librerías del Software R 78
Figura Nº 24 Servidores Disponibles Librerías Software R 79
Figura Nº 25 Instalación del Packages Software R 79
Figura Nº 26 Llamar a una Librería en Software R 80
Figura Nº 27 Cambio de Ruta de Trabajo 81
Figura Nº 28 Selección de Carpeta llamada Proyecto 82

Figura Nº 29 Datos en Excel exportados de la Base de Datos 85
Figura Nº 30 Datos de Excel exportados a CSV..... 86
Figura Nº 31 Cargar archivo fugas.csv en Memoria 87
Figura Nº 32 Archivo fugas.csv cargado en R 88
Figura Nº 33 Archivo CSV usando la librería tidyverse 89
Figura Nº 34 Tabla Con Datos Fecha 97
Figura Nº 35 Tabla Con Datos Fecha Limpios 98
Figura Nº 36 Tabla Como Data.Frame 99
Figura Nº 37 Levels Tipofuga 100
Figura Nº 38 Levels Distrito 101
Figura Nº 39 Matriz De Dispersion De Nuestra Tabla 103
Figura Nº 40 Grafica Del Modelo Predictivo 105
Figura Nº 41 Datos Obtenidos Modelo Del Tipo Polilineal 107
Figura Nº 42 Datos Obtenidos Modelo Del Tipo Barras 109
Figura Nº 43 Datos Obtenidos Modelo Del Tipo Barras 111
Figura Nº 44 Datos Obtenidos Modelo Del Tipo Barras 113
Figura Nº 45 Datos Obtenidos Modelo Del Tipo Polilineal 115
Figura Nº 46 Datos Obtenidos Modelo Del Tipo Barras 117
Figura Nº 47 Datos Obtenidos Modelo Del Tipo Barras 119
Figura Nº 48 Datos Obtenidos Modelo Del Tipo Barras 121
Figura Nº 49 Datos Obtenidos Modelo Del Tipo Polilineal 123
Figura Nº 50 Datos Obtenidos Modelo Del Tipo Barras 125
Figura Nº 51 Datos Obtenidos Modelo Del Tipo Barras 127
Figura Nº 52 Datos Obtenidos Modelo Del Tipo Barras 129
Figura Nº 53 Modelo Del Tipo Barras : Distritos & Fecha 130
Figura Nº 54 Modelo Del Tipo Barras: Distritos & Tipofugas 131
Figura Nº 55 Modelo Del Tipo Barras : Tipofugas & Fecha 131
Figura Nº 56 Modelo Del Tipo Barras : Tipofugas & Distrito 2014-2019 136
Figura Nº 57 Modelo Del Tipo Barras : Tipofugas & Distrito 2014-2019 138
Figura Nº 58 Modelo Del Tipo Barras Ordenado Por Tipofugas & Distrito 139
Figura Nº 59 Modelo Del Tipo Barras ordenado 2014-2019 142
Figura Nº 60 Modelo Del Tipo Barras : Distrito 2014-2019 143
Figura Nº 61 Modelo Del Tipo Barras : Distrito 2020 150
Figura Nº 62 Modelo Del Tipo Barras : Distrito 2020 151
Figura Nº 63 Modelo Del Tipo Barras : Tipofugas 2020 152
Figura Nº 64 Modelo Del Tipo Barras : Tipofugas 2020 153

RESUMEN

En la actualidad, la cobertura de agua potable en el país aún es insuficiente, sobre todo en las zonas rurales del país; de manera similar este problema se presenta en las zonas periurbanas de Lima.

Es por ello, que la presente investigación plantea una propuesta de diseño del sistema de abastecimiento de agua potable para la zona de estudio, para lo cual se toma en consideración que el caudal requerido será brindado por SEDAPAL.

Sedapal tiene que velar por la calidad del servicio ofrecido, sin embargo, esto se ve perjudicado sustancialmente por las fugas de agua potable de emergencias , (del medidor a la red) las cuales podrían ser manejadas de una mejor manera si se tuviera información referente a proyecciones e información estadística al respecto .

El “Diseño De Un Algoritmo Predictivo Para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019” podría facilitar dicha labor ya que se analizara la información con la que ya se cuenta para lograr un mejor desempeño del servicio mejorando los tiempos de respuesta a las emergencias y controlando y monitoreando el abastecimiento de agua potable en la ciudad de Lima.

En ese sentido, el método utilizado plantea una propuesta de mejora en el proceso de diseño del sistema de abastecimiento de agua potable.

ABSTRAC

At present, potable water coverage in the country is still insufficient, especially in rural areas of the country; similarly, this problem occurs in the peri-urban areas of Lima.

That is why the present investigation raises a proposal for the design of the drinking water supply system for the study area, for which it is taken into consideration that the required flow will be provided by SEDAPAL.

Sedapal has to ensure the quality of the service offered, however, this is substantially impaired by emergency drinking water leaks, (from the meter to the network) which could be managed in a better way if information regarding projections and statistical information in this regard.

The "Design of a Predictive Algorithm for Early Monitoring of Potable Water Networks in the City of Lima, 2019" could facilitate such work since the information that is already available will be analyzed to achieve a better performance of the service, improving the times of response to emergencies and controlling and monitoring the drinking water supply in the city of Lima.

In this sense, the method used raises a proposal for improvement in the design process of the drinking water supply system

INTRODUCCIÓN

La analítica predictiva es una forma de análisis avanzado que utiliza datos nuevos e históricos para predecir la actividad futura, el comportamiento y las tendencias. Implica la aplicación de técnicas de análisis estadístico, consultas analíticas y algoritmos automáticos de aprendizaje automático a conjuntos de datos para crear modelos predictivos que sitúen un valor numérico o puntuación en la probabilidad de que ocurra un evento particular.

Actualmente el problema del desabastecimiento de agua potable en la gran Lima, es debido en gran medida a las pérdidas de agua no facturada debido a fugas en tuberías rotas y/o desgastadas ya sea por el tipo de material utilizado, por la antigüedad, por la ubicación, etc.

El desarrollo de un algoritmo predictivo puede permitir en la medida de lo posible, la detección temprana de fugas de agua por Distritos en las redes de agua potable de la ciudad de Lima en el año 2019.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

El problema del desabastecimiento de agua potable para el consumo humano es un desafío para las futuras generaciones, la población a nivel mundial aumenta y en contraste los recursos hídricos son más escasos año con año.

Sólo muy poca agua es utilizada para el consumo del hombre, ya que: el 90 % es agua de mar y tiene sal, el 2 % es hielo y está en los polos, y sólo el 1 % de toda el agua del planeta es dulce, encontrándose en ríos, lagos y mantos subterráneos. Además, el agua tal como se encuentra en la naturaleza, para ser utilizada sin riesgo para el consumo humano requiere ser tratada, para eliminar las partículas y organismos que pueden ser dañinos para la salud. Y finalmente debe ser distribuida a través de tuberías hasta las casas, para que se pueda consumir sin ningún problema ni riesgo alguno.

En el caso peruano las empresas de distribución de agua son las encargadas del proceso de potabilización de agua, en el caso específico de la ciudad de Lima, la empresa prestadora del servicio es la estatal SEDAPAL desde hace 55 años.

Sedapal tiene que velar por la calidad del servicio ofrecido, sin embargo, esto se ve perjudicado sustancialmente por las fugas de agua potable de emergencias , (del medidor a la red) las cuales podrían ser manejadas de una mejor manera si se tuviera información referente a proyecciones e información estadística al respecto .

El “Diseño De Un Algoritmo Predictivo Para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019” podría facilitar dicha labor ya que se analizara la información con la que ya se cuenta para lograr un mejor desempeño del servicio mejorando los tiempos de respuesta a las emergencias y controlando y monitoreando el abastecimiento de agua potable en la ciudad de Lima.

1.1 Descripción de la realidad problemática

SEDAPAL como empresa prestadora del servicio de distribución de agua potable es la responsable de velar por el abastecimiento del líquido elemento en la ciudad de Lima, los usuarios por su parte son responsables de mantener sus fábricas, universidades, colegios, oficinas, domicilios, cisternas de agua, etc. sin desperfectos en sus tuberías, conexiones y accesorios para evitar el despilfarro de agua innecesariamente y evitar así los montos elevados en la facturación de los recibos a fin de mes y el posible desabastecimiento generado por dichos desperfectos.

Son más de 14000 km de tuberías de agua potable distribuidas en la gran Lima las cuales por su longitud, por su antigüedad y múltiples factores más; son vulnerables a sufrir desperfectos llámese averías las cuales implican perdidas

sustanciales de agua potable las cuales afectan el normal abastecimiento por parte de SEDAPAL , estas averías pueden según su magnitud de fuga de agua : pequeñas , medianas , grandes y a veces enormes cuando las tuberías afectadas son tuberías matrices o principales ocasionando cortes en el suministro de agua en distritos completos.

DISEÑO DE UN ALGORITMO PREDICTIVO PARA MONITOREO TEMPRANO DE REDES DE AGUA POTABLE EN LA CIUDAD DE LIMA, 2019” se toma como parte de una estrategia de mejoramiento del proceso de detección de fugas de emergencia no visibles en la ciudad de Lima.

En el sistema de distribución de agua potable externo al usuario existe una gran cantidad de situaciones que pueden llevas al desabastecimiento, en términos generales se pueden presentar: emergencias por roturas de tuberías, fallas temporales, permanentes, fallas operacionales y problemas de índole natural: huaycos, lluvias, sequias, etc.

Es por eso que viendo esta problemática con respecto al abastecimiento de agua potable se plantea el diseño de un sistema de control predictivo que permitirá anticipar este proceso de fugas de emergencia haciendo mas eficiente dicha labor.

1.2 Formulación del problema

1.2.1 Problema General

El reconocimiento o formulación del problema objeto de investigación, primordialmente se refiere a la selección general del problema enmarcado en las dimensiones epistemológicas que gobiernan las ciencias de la ingeniería consecuentemente, del problema general ser desmembrados en problemas específicos a ser investigados con sus correspondientes sub objetivos y sub hipótesis de solución a dichos problemas.

La selección general del problema objeto de investigación, se justifica desmembrar o fragmentar de forma lógica mental en problemas específicos coherentes a las dimensiones que gobiernan las Ciencias de la Ingeniería Electrónica, correspondientes en:

- Programas de Investigación
- Sub Programas de Investigación y en
- Líneas de Investigación

Para la selección general del problema objeto de investigación, es necesario formular la siguiente interrogante:

¿Será posible utilizar un Desarrollar un Algoritmo Predictivo en el **Software R** para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019"?

1.2.2. Problemas específicos

Para mayor fidelidad sobre la detección de fugas no visibles, es indispensable desmembrar coherentemente del principal problema objeto de investigación, en sub problemas específicos formulando las siguientes interrogantes:

• ¿Se podrá Desarrollar un algoritmo predictivo desarrollado en el Software R para el monitoreo temprano de las fugas de agua potable en la ciudad de Lima 2019?

• ¿Se podrá determinar con anticipación los factores que afectan la detección de fugas no visibles en las redes de agua potable de la ciudad de Lima a través de este algoritmo?

• ¿Influirá el crecimiento de la población en las proyecciones estimadas por el algoritmo?

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un Algoritmo Predictivo en el Software R para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019.

1.3.2 Objetivo Específicos

Para definir los objetivos específicos del problema objeto de investigación, es necesario desmembrar del objetivo principal, en los siguientes objetivos específicos:

• Emplear un algoritmo predictivo

• Determinar los factores que afectan el servicio de emergencia de detección de fugas no visibles en las redes de agua potable de la ciudad de Lima a través del estudio de la información con la que se cuenta.

• Determinar si el crecimiento de la población influye en la predicción del algoritmo

.

1.4 Limitantes de la investigación

• Desarrollar un Algoritmo Predictivo en el Software R para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019 se justifica porque permite analizar tendencias en el servicio de emergencia y optimizar recursos y enfocarlos en procesos de prevención sobre posible fugas de agua potable que ocurran en el futuro.

• La actual evolución de la tecnología permite que se puedan diseñar sistemas de este tipo y ayudar a la optimización del servicio.

• Socialmente se justifica, toda vez que el desarrollo de esta tecnología permitirá mejoras en el servicio de detección de fugas no visibles y su aplicación a mayor escala.

• Este desarrollo permitirá su implementación para muchos otros campos de la industria en los cuales se necesita algoritmos predictivos.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes

- **Internacional**

Fuentes-Mariles,Rodríguez-Vázquez y Palma-Nava en Estimación y localización de fugas en una red de tuberías de agua potable usando algoritmos genéticos (Ingeniería Investigación y Tecnología, Vol. XII, Núm. 2, 2011) propone un método experimental para la estimación de pérdidas por fugas de agua.

> *"(…) Las fugas de agua potable en las redes de distribución producen desperdicio de agua, reducen la eficiencia de las redes y generan una pérdida económica a los organismos operadores del sistema de distribución de este tipo de líquido. La detección de fugas en una red de tuberías es complicada, ya que en su mayoría no se encuentran visibles. Para reducirlas es necesario contar con procedimientos e instrumentos especiales para localizarlas y eliminarlas. En este artículo se expone un método para detectar fugas a lo largo de las tuberías. Con él se determina el caudal de las mismas y la posición donde ocurren en las tuberías de una red cerrada. Se considera que la red funciona en flujo permanente a presión y que sólo existe una fuga por cada tubería de la red. El procedimiento propuesto se basa en las mediciones de la presión en la unión de las tuberías, el conocimiento de las características de la red y la estimación de las demandas de caudal. Se emplea la teoría de la computación evolutiva, en particular, un algoritmo genético simple, como mecanismo de búsqueda de la solución óptima. Este proceso es iterativo hasta disminuir el error entre las cargas medidas y las calculadas con este método.*
>
> *Se incluye un ejemplo a partir de las mediciones obtenidas en laboratorio para mostrar la bondad del método y la forma de aplicarlo."*

Luego del análisis de resultados de esta investigación se llega a la conclusión que:

> *Con base en los resultados obtenidos en este trabajo y en otros similares, se confirma que los algoritmos genéticos son herramientas robustas para esta clase de problemas de optimación, y que los valores obtenidos tienen una adecuada precisión, ya que el porcentaje de error fue bajo.*

Por otro lado, aun cuando estos resultados son alentadores, será necesario aplicar el procedimiento propuesto en redes de mayor tamaño para comprobar su eficacia, y determinar pesos o factores que permitan elegir y acotar los espacios de solución, implementando el funcionamiento hidráulico de la red, consideración muy importante para el caso de redes grandes y complicadas.

Ramírez-Quintana y Piris-Ruano en Una aplicación de minería de datos para el análisis de la propiedad de terminación de SRTs (Ramírez-Quintana y Piris-Ruano 2015) plantea la posibilidad de predecir la propiedad de terminación de los sistemas de reescritura de términos.

"(…) La minería de datos juega a día de hoy un papel importante en muchos campos de la ciencia debido a la gran cantidad de información con la que se trabaja diariamente y la que se puede deducir de la misma. Las posibilidades que ofrece la minería de datos son muy diversas, siendo además su ámbito de aplicación muy extenso. Uno de los posibles campos y del que versa este trabajo, es la predicción de propiedades de sistemas software.

Para ser más concretos, en este proyecto hemos planteado la posibilidad de predecir la propiedad de terminación de los sistemas de reescritura de términos. Para ello se ha hecho uso de la base de datos de resultados de la Termination Competition, competición que se celebra anualmente desde 2006 y que trata sobre la demostración de la terminación de sistemas de reescritura. En esta base de datos se almacenan en forma de registros los resultados obtenidos por herramientas de terminación que participan en la competición, registros tales como el resultado de la demostración, tiempo empleado para obtener la respuesta y la traza de la demostración. Al día de hoy existen multitud de actividades de la vida cotidiana que constatan como la tecnología ha mejorado nuestra calidad de vida y junto a ella nuestra forma de ver las cosas que la componen. Vivimos en una sociedad masificada tecnológicamente, donde los objetos más simples se han convertido en pequeños sistemas informáticos conectados con un ente más grande y complejo. Con el tiempo hemos adquirido unas rutinas que no conciben la posibilidad de un funcionamiento incorrecto de los sistemas informáticos. Esta forma de ver la tecnología, dependiendo siempre del contexto puede conllevar desde un simple reinicio del sistema en cuestión, hasta perder grandes sumas de dinero y en el peor de los casos (entornos críticos) podemos llegar a hablar incluso de la pérdida de vidas. Con esto no se intenta decir que nadie revisa los programas o sistemas, ni mucho menos, es más de no ser revisados estaríamos hablando bajo otro punto de vista, pero a lo mejor sí que es cierto que no son revisados con la exhaustividad necesaria para asegurar sin ninguna duda su correcto funcionamiento a lo largo del tiempo y en cualquier contexto. Con motivo de poder dotar de una confianza bien fundada a los programas es necesario analizarlos

con la exhaustividad antes mencionada. Existe un área de la informática encargada de este aspecto, es decir del desarrollo de métodos y técnicas de análisis de software. En general, hay dos enfoques posibles: el análisis estático y el análisis dinámico. La principal diferencia entre ambos es que el análisis dinámico necesita la ejecución de los programas para analizarlos y el estático no."

Luego del análisis de resultados de esta investigación se llega a la conclusión que:

Este trabajo responde a la idea expuesta al inicio de esta memoria sobre la necesidad de examinar cualquier tipo de software con la mejor de las lupas con el único fin de analizar sus propiedades. Además, este trabajo se enmarca dentro del ámbito del análisis y verificación de código, pero se diferencia de otros estudios existentes en que tiene en cuenta el valor añadido de la información.

En concreto, nos hemos planteado estudiar la propiedad de terminación de sistemas de reescritura de términos a partir de los datos de la Termination Competition aplicando técnicas de minería de datos, para ello hemos realizado lo siguiente:

- *Extraer los datos y prepararlos para la aplicación de técnica de minería de datos.*
- *Definir el conjunto de propiedades con las que se representaran los TRS.*
- *Implementar varios experimentos usando diferentes configuraciones (supervisado, semisupervisado) intentando tener en cuenta las características intrínsecas del problema (clases desbalanceadas).*

Los mejores resultados se han obtenido en el ´ultimo experimento (enfoque semisupervisado), usando las clases 'terminante' y 'no terminante' y balanceando las clases.

Nacido este proyecto de la idea de obtener un sistema de ayuda para alguna de las herramientas de demostración, los resultados obtenidos dan pie a pensar que es posible la determinación de la terminación de los programas con simples propiedades de los mismos haciendo uso de técnicas de minería de datos. Este estudio aporta a la investigación de la terminación de los programas unos primeros resultados de la experimentación haciendo uso de técnicas de minería de datos. Además, este aporta una completa metodología de trabajo para este contexto de aplicación, siendo el código generado reutilizable y adaptable abriendo un amplio abanico de posibilidades de estudio.

- **Nacional**

Grandez-Marquez en Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales (Grandez-Marquez-2017) encuentra reglas que determinen el patrón de consumo de clientes en una distribuidora de suplementos nutricionales, aplicando técnicas de minería de datos.

> *"(…) Granular la información nos permite conocer aspectos no analizados que pueden incidir en el desempeño de una organización.*
>
> *Explorar los registros contenidos en la base de datos de la empresa con técnicas de minería de datos, nos facilita comprender reglas que determinan patrones de consumo y tendencias que siguen los datos, lo que a su vez genera un conocimiento; el cual nos permitirá asociar productos que tengan mayor rotación con aquellos que no la tienen, generando de esta forma venta cruzada que resulte en el incremento de ingresos.*
>
> *Se centra en aplicar Minería de Datos en un negocio de venta de suplementos nutricionales (SN), dado el crecimiento explosivo en la distribución de estos productos en Lima y que según investigaciones en el año 2011 cada persona destinó 180 soles mensuales en suplementos y que para el 2016 se destinó hasta 270 soles mensuales en su consumo."*

Luego del análisis de resultados de esta investigación se llega a la conclusión que:

> *La minería de datos es un área de conocimiento que nos ayuda a tomar decisiones en base a información del propio negocio.*
>
> *En el algoritmo de asociación podemos ver que se forman reglas con cierta probabilidad de ocurrencia, así como también con una importancia, estas reglas lo que hacen es encontrar patrones sólidos dentro de la base de datos que relacione las características de los consumidores con los productos que suelen consumir, mientras tanto en el algoritmo neuronal se indica que tan determinante es una variable en la compra de un determinado producto y en el algoritmo Clúster se agrupa las variables de acuerdo al valor que puede tomar y asigna a cada una un nivel de ocurrencia.*

La base de datos debe de ser analizada y refinada antes de aplicar técnicas de minería de datos o mejor aún antes de iniciar un proyecto de Data Mining se debe realizar un planeamiento en la recolección de datos.

Es importante emplear el número adecuado de registros que sean significativos, ya que los algoritmos que utiliza minería de datos deben de detectar tendencias para mostrar resultados.

Apesteguia-Infantes y Huarcaya-Gonzales en Modelamiento inteligente para la detección de fugas no visibles en las redes de agua potable de la ciudad de Lima (Apesteguia-Infantes yHuarcaya-Gonzales 2017) realiza un estudio de cómo se realiza la detección de fugas de agua potable en la ciudad de Lima y la automatización de dicho proceso usando redes neuronales para reducir los falsos positivos.

"(…) se presenta como un tema de investigación interdisciplinario inédito el cual propone como premisa el modelamiento usando redes neuronales para la detección de fugas de agua potable en las redes de la ciudad de Lima, esto permitirá hacer una detección y clasificación de las fugas en tiempo real y con un margen de error menor con respecto al error humano.

Para identificar la relación que existe entre las variables que intervienen en la detección de fugas es necesario su modelamiento.

Sedapal tiene que velar por la calidad del servicio ofrecido, sin embargo, esto se ve perjudicado sustancialmente por las fugas de agua potable no facturadas, (del medidor a la red) las cuales podrían ser detectadas más fácilmente con el modelamiento inteligente para la detección de las fugas no visibles de agua potable para con ello reducir la cantidad de agua no facturada debido a las perdidas en tuberías y lograr mejorar el abastecimiento de agua potable en la ciudad de Lima.

Luego del análisis de resultados de esta investigación se llega a la conclusión que:

El modelamiento inteligente usando redes neuronales permitió detectar las fugas no visibles en las redes de agua potable hasta en un 30% más efectivamente que el método tradicional de detección de fugas con personal calificado.

Instaurando el Geófono Digital Smart en la detección de fugas no visibles en las redes de agua potable se logró reducir los porcentajes de agua no facturada al aumentar los equipos de detección de fugas no visibles en las redes de agua potable.

El ahorro pudo ser cuantificado ya que se cuenta con una base de datos para contrastar el agua producida versus el agua facturada por parte de la empresa prestadora del servicio.

2.2 Marco

El Marco filosófico de la presente investigación es de suma importancia, por tratarse de la detección de fugas no visibles en las redes de agua potable, el agua recurso natural sin el cual no existiría la vida; amparándose en tres ejes filosóficos como son: lo ontológico sobre la concepción del ser humano generando el compromiso de asumir fundamentalmente el mejoramiento de la calidad de vida de las personas. De igual manera, en lo metodológico aplicando nuevos métodos, técnicas y/o estrategias para la solución total de los problemas y en lo epistemológico que trata la parte doctrinaria que toda investigación tiene y que se concibe en el "cómo debe de ser" la solución al problema objeto de investigación.

2.2.1 Teóricos

La ontología como parte de la filosofía trata sobre el "universo del ser" valorar el recurso en este caso que se desea preservar que es el agua potable evitando su desperdicio detectando a tiempo las fugas tanto a nivel comercial y residencial (Onto = ser, ente. Logo = estudio, ciencia, teoría), y como la principal responsabilidad del investigador científico es conocer, entender y sistematizar el problema objeto de estudio en este caso la detección de fugas no visibles de agua usando como herramienta las redes neuronales.

- **Fundamentación metodológica**

En la presente investigación la fundamentación metodológica, se refiere al "universo del hacer", sobre el proceso de la detección de fugas no visibles usando las redes neuronales como herramienta para hacer este proceso mas eficiente y menos propenso a los errores evitando el desperdicio de agua potable no facturada.

El concepto de Metodología hace referencia al plan de investigación que permite cumplir ciertos objetivos en el marco de una ciencia, por lo tanto, puede entenderse a la metodología como el conjunto de procedimientos que determinan una investigación de tipo científico o marcan el rumbo de una exposición doctrinal. El vocablo Metodología es generado a partir de tres palabras griegas: meta (más allá), odós (camino) y logos (estudio), sin embargo, es importante la distinción entre el método (nombre que recibe cada plan seleccionado para alcanzar un objetivo) y la metodología (rama que estudia el método).

- **Fundamentación Epistemológica**

El sustantivo epistemología o gnoseología está compuesto por la unión de dos palabras griegas, episteme que se refiere al "conocimiento o ciencia" y logos como "discurso", concluyentemente, el fundamento epistemológico concierne al "universo del conocer" y la parte doctrinaria de esta investigación fundamentalmente experimental aplicada, radica en que los interesados en este caso la población en general deben saber el "cómo debe de ser" que se cuiden nuestros recursos naturales en este caso el agua potable con un sistema de detección de fugas eficiente que permita reducir estas en bien de todos.

Cuidar el medio ambiente es necesario, pues lo que ocurra en él afecta a todos los seres vivos que conviven en un mismo sistema.

Hablar del medio ambiente es un tema extenso, pues comprende – inicialmente – a todo lo que rodea a los seres vivos, y también implica

variables que condicionan de forma importante a la sociedad y la cultura de los seres humanos, por ende, marca muchas veces el rumbo a seguir de las próximas generaciones. La palabra medio ambiente proviene del latin “medius” y “ambiens”, las que en conjunto se enfocan hacia el concepto de “lo que está a ambos lados”.

El cuidado del medio ambiente es tan importante, que se necesita de una preocupación exhaustiva por los detalles, debido a que es imposible conseguir un entorno adecuado, y el buen funcionamiento de los factores que le modifican, si no se cuidan las variables que son capaces de alterarlos. Un ejemplo del tema es el que se refiere a la contaminación de los rios y mares convirtiéndose en una problemática mundial, por que estamos acabando con las fuentes naturales de agua para consumo humano, dentro de 50 años las guerras ya no serán por petróleo sino por el control de las reservas de agua potable del planeta.

- **Cuidado ambiental y política ambientalista**

El índice de Desempeño Ambiental 2010 (Eviromental Performance Index), realizado por Yale Center for Enviroment Law & Policy de la Universidad de Yale, en combinación al trabajo de la Universidad de Columbia (Center for International Earth Science Information Network), y la Comisión Europea en conjunto al World Economic Forum, clasificaron a un total de 163 países según el desempeño sobre distintas mediciones que eran categorizadas según: salud ambiental, calidad del aire, manejo de recursos, cambio climático, agricultura, biodiversidad, entre otros.

En este tipo de índices de medición se puede observar cómo la política ambientalista de los distintos países del mundo afecta al medio, influye en el desarrollo de la sociedad y la cultura, y decide en gran parte el futuro de las próximas generaciones.

"China es el país que más contamina el mundo en términos absolutos y Estados Unidos el que más contamina en términos relativos. ¿Qué hay de

las energías limpias como la solar, geotérmica, eólica y demás? La respuesta es tan cruda como cruel: cuando el petróleo es barato, es barato ensuciar. El boom de las energías alternativas se disparó con un barril por encima de los 100 dólares. Sólo tendremos un mundo limpio cuando el petróleo sea caro, y convenga explorar alternativas más limpias. Si el petróleo baja de precio, tendremos un mundo cada vez más sucio. Así de sencillo".

Cuadro N°1

Países con más emisiones de CO2

	Emisiones de CO2 (kilotones)		
	World	33.615.389	
1	China	8.286.892	24,7%
2	United States	5.433.057	16,2%
3	India	2.008.823	6,0%
4	Russian Federation	1.740.776	5,2%
5	Japan	1.170.715	3,5%
6	Germany	745.384	2,2%
7	Iran, Islamic Rep.	571.612	1,7%
8	Korea, Rep.	567.567	1,7%
9	Canada	499.137	1,5%
10	United Kingdom	493.505	1,5%
11	Saudi Arabia	464.481	1,4%
12	South Africa	460.124	1,4%
13	Mexico	443.674	1,3%
14	Indonesia	433.989	1,3%
15	Brazil	419.754	1,2%
16	Italy	406.307	1,2%
17	Australia	373.081	1,1%
18	France	361.273	1,1%
19	Poland	317.254	0,9%
20	Ukraine	304.805	0,9%

Fuente: Reporte Banco Mundial 2013

2.2.2 Conceptuales

El agua es un recurso limitado e insustituible que es clave para el bienestar humano y solo funciona como recurso renovable si está bien gestionado. Hoy en día, más de 1.700 millones de personas viven en cuencas fluviales en las que su uso supera la recarga natural, una tendencia que indica que dos tercios de la población mundial podría vivir en países con escasez de agua para 2025. El agua puede suponer un serio desafío para el desarrollo sostenible pero,

gestionada de manera eficiente y equitativa, el agua puede jugar un papel facilitador clave en el fortalecimiento de la resiliencia de los sistemas sociales, económicos y ambientales a la luz de unos cambios rápidos e imprevisibles.

- **¿Qué es el "desarrollo sostenible"?**

El desarrollo sostenible se popularizó de manera explícita y contextualizada por la Comisión Brundtland en el documento "Nuestro Futuro Común" donde se define como "el desarrollo que satisface las necesidades del presente sin comprometer la capacidad de las generaciones futuras para atender sus propias necesidades" (ONU, 1987). La Comisión Brundtland se centró en tres pilares del bienestar humano: las condiciones económicas, sociopolíticas y ecológicas/ambientales. Este concepto básico fue desarrollado como apoyo a la implementación de medidas sólidas dirigidas a impulsar el desarrollo económico y social, en particular para las personas de los países en vías de desarrollo y, al mismo tiempo, garantizar que la integridad del medio ambiente se mantenga para las generaciones futuras.

Por otra parte, estas cifras globales ocultan grandes disparidades entre las naciones y las regiones, entre los ricos y los pobres, entre las poblaciones rurales y las urbanas, así como entre los grupos desfavorecidos y la población en general.

Actualmente no existe una meta mundial para mejorar la higiene, a pesar de ser una de las intervenciones de salud pública individuales más rentables.

- **Desarrollo sostenible y agua**

La agricultura es, con diferencia, el mayor consumidor de agua a nivel mundial, representando el 70% de las extracciones de agua en todo el mundo, aunque esta cifra varía considerablemente entre países. La agricultura de secano es el sistema de producción agrícola predominante en todo el mundo y su productividad actual es, en promedio, un poco más de la mitad del potencial a obtener sobre una gestión agrícola óptima. Para 2050, la

agricultura tendrá que producir un 60% más de alimentos a nivel mundial y un 100% más en los países en vías de desarrollo.

La industria y la energía juntas representan el 20% de la demanda de agua. Los países más desarrollados tienen una proporción mucho mayor de extracciones de agua dulce para la industria que los países menos desarrollados, donde predomina la agricultura. El equilibrio entre los requisitos de sostenibilidad frente a la visión convencional de la producción industrial en masa crea una serie de interrogantes para la industria. A gran escala, la globalización y la forma de extender los beneficios de la industrialización a todo el mundo equitativamente y sin impactos insostenibles sobre el agua y otros recursos naturales es la cuestión clave.

El sector doméstico representa el 10% del uso total de agua. Y, en todo el mundo, se estima que 748 millones de personas siguen sin tener acceso a una fuente mejorada de agua y que 2.500 millones siguen sin acceso a unos servicios de saneamiento mejorados.

- **Recursos Hídricos**

Lo primero que hay que definir antes de entrar en la definición de recursos hídricos es conocer su origen etimológico:

-Recursos es una palabra cuya raíz es el latín “recursus”, que viene a hacer referencia a hacer uso de los medios o bienes de los que dispone alguien para acometer algo en concreto.

Un recurso es una materia prima o un bien que dispone de una utilidad en pos de un objetivo. Por lo general se trata de algo que satisface una necesidad o que permite la subsistencia. Hídrico, por su parte, es aquello que está vinculado al agua.

Los recursos hídricos son los cuerpos de agua que existen en el planeta, desde los océanos hasta los ríos pasando por los lagos, los arroyos y las

lagunas. Estos recursos deben preservarse y utilizarse de forma racional ya que son indispensables para la existencia de la vida.

El problema es que, aunque en su mayoría son recursos renovables, la sobreexplotación y la contaminación que provocan diversas actividades humanas hacen que los recursos hídricos estén en riesgo. Su capacidad de regeneración muchas veces no resulta suficiente ante el ritmo de uso.

Una de las grandes dificultades que enfrenta la Humanidad es la falta de agua dulce. Más del 97% del agua de la Tierra es agua salada, cuyo aprovechamiento es complejo. Por eso el agua dulce, que se utiliza para el consumo humano y un sinfín de actividades, es tan importante.

En concreto, las estimaciones llevadas a cabo vienen a establecer que el 100 % del agua total del planeta se distribuye de la siguiente manera: 97,47 % de agua salina, 2,53 % de agua dulce, 1,76 % de glaciares y capas polares, 0,76 % de agua subterránea y 0,01 % de lagos, ríos y atmósfera.

Uno de los principales problemas que enfrenta el mundo en un futuro cercano es la disminución del suministro de agua hasta en un 40 % para el año 2030, por lo que es importante mejorar considerablemente la gestión de este recurso.

Según datos del Ministerio del Ambiente, en la vertiente amazónica reside el 26 % de la población, que cuenta con el 97,7 % de agua, mientras que en la vertiente del Pacífico reside el 70 % de la población, que cuenta tan solo con el 1,8 % de agua.

Los recursos y servicios relacionados con el agua son esenciales para el logro de la sostenibilidad global, puesto que estos ayudan al crecimiento económico, la reducción de la pobreza y la sostenibilidad ambiental.

- **Políticas de Manejo del Agua Potable**

Lima es la segunda ciudad en el mundo afincada en un desierto después del Cairo en Egipto, por lo que los recursos hídricos que poseemos son insuficientes y lo serán más en los próximos años.

Es necesario por ello hacer un uso adecuado de las fuentes de agua (en nuestro caso de las fuentes de agua potable de la ciudad de Lima)

El Decreto Supremo N° 006-2015-MINAGRI y 013-2015-MINAGRI aprobaron la Política y Estrategia Nacional de Recurso Hídricos y el plan Nacional de Recursos Hídricos respectivamente.

En el año 1981 el gobierno de Fernando Belaúnde Terry fusionó las tres Empresas de Saneamiento de Lima, Arequipa y Trujillo y la fusionó en una sola empresa estatal matriz: el Servicio Nacional de Abastecimiento de Agua y Alcantarillado (SENAPA). El SENAPA estaba conformado por 15 empresas filiales y 10 unidades operativas distribuidas a lo largo del país. SEDAPAL en Lima era la más grande de estas empresas filiales estatales. Sin embargo, 200 ciudades (20%) quedaron afuera del SENAPA y administraron sus propios servicios.

- **Sedapal**

El Servicio de Agua Potable y Alcantarillado de Lima - Sedapal S.A. (SEDAPAL), es una empresa estatal peruana creada en 1981.

Brinda prestaciones de agua potable y alcantarillado al sector urbano de la ciudad de Lima. SEDAPAL gestiona el abastecimiento de agua potable del área metropolitana de Lima y Callao.

El agua que suministra está tratada en La Atarjea en El Agustino que abastece a más de 9 millones de habitantes de Lima.

- **Objetivos de Sedapal**

  - El objetivo de Sedapal es la prestación de los servicios de saneamiento como agua potable y alcantarillado sanitario.
  - Ejecuta la política del sector en la operación, mantenimiento, control y desarrollo de los servicios básicos, con funciones específicas en aspectos de normatividad, planeamiento, programación,
  - Elaboración de proyectos, financiación, ejecución de obras, asesoría y asistencia técnica.
  - Además, puede dedicarse a otras actividades afines, vinculadas, conexas y/o complementarias a su objeto social.

- **Servicios de Saneamiento**

  - Instalación y reubicación de conexiones domiciliarias
  - Revisión y aprobación de proyectos
  - Supervisión de obras
  - Empalmes a la red existente
  - Cierre y reapertura de conexiones
  - Inspecciones en las redes primarias, detección de fugas
  - Factibilidad de servicios para habilitaciones urbanas
  - Otros que determine la Superintendencia

- **Servicio de Detección de Fugas no visibles en redes primarias (ECRF)**

La finalidad pública del presente servicio está relacionada con el objetivo de primer nivel “Disminuir el Agua No Facturada” y el objetivo de segundo nivel “Reducir los volúmenes de pérdidas de agua potable”; lo que permitirá reducir las pérdidas de agua y brindar un mejor servicio.

El presente servicio lo ejecuta El Equipo Control y Reducción de Fugas (ECRF) de la Gerencia de Producción y Distribución Primaria de SEDAPAL.

Para la realización de estas labores se cuentan con 5 unidades móviles las cuales se desplazan por la ciudad de Lima, realizando labores diversas como:

- Inspección de Redes Primarias de Agua Potable según programación establecida con anticipación.
- Detección de fugas imprevistas en calles y avenidas.
- Detección de conexiones clandestinas de agua.
- *Monitoreo de la presión de agua (psi) por cuadrantes y/o zonas.

Para la realización de las actividades diarias, las unidades cuentan con instrumental especializado tanto digital como analógico.

Este equipo se utiliza para las labores de detección de fugas de agua, obstrucción de tuberías, tuberías clandestinas y así como para controlar la presión de agua en la red de tuberías.

El personal de cada unidad ha recibido entrenamiento especial para el manejo y la operación de estos equipos (por lo general para el servicio se cuenta con personal con una experiencia no menor a 5 años dado el tipo de labor que se realiza)

Cada unidad móvil cuenta con 5 personas (Técnico, Operario especializado, Chofer y 2 operarios).

Figura N° 1

Unidad Móvil Detección de Fugas

Fuente: propia del autor

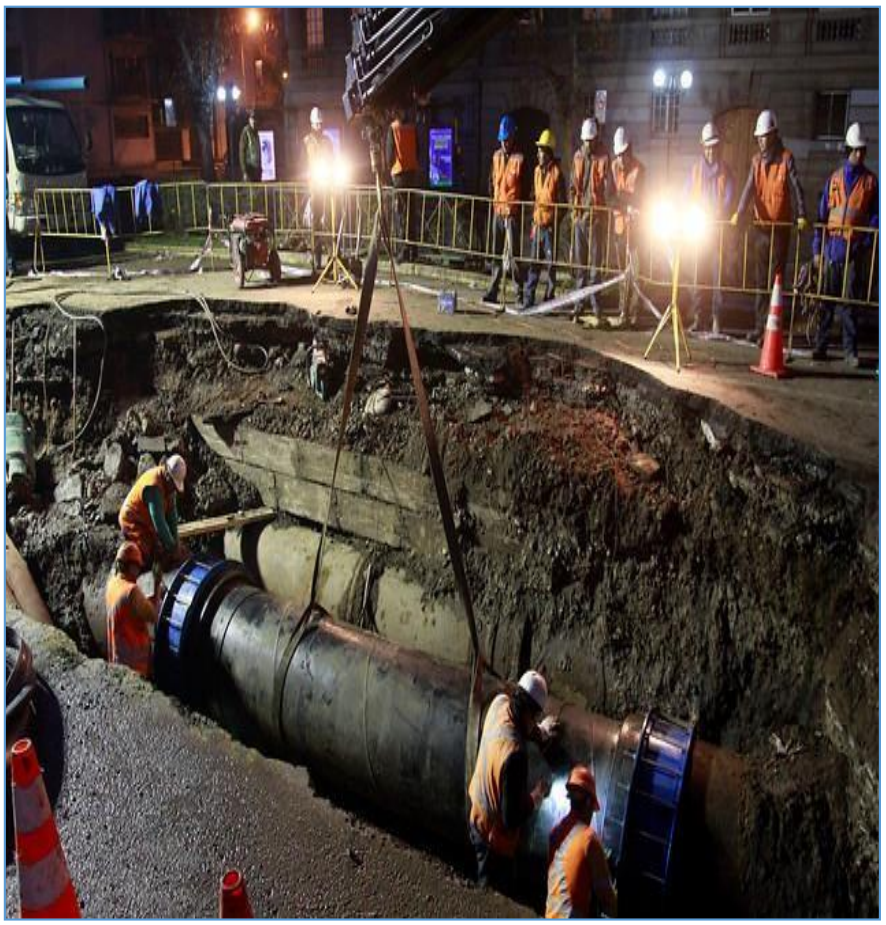
- **Equipos utilizados por el (ECRF)**

Para realizar su labor el ECRF cuenta con equipos electrónicos portátiles de última generación los cuales les permiten realizar su labor en forma eficiente.

Antiguamente para reparar una tubería matriz con fuga se tenía que excavar toda la cuadra ya que no se contaba con equipos para detectar la zona de la tubería averiada, esto repercutía en problemas de logística (cerrar las calles aledañas, redirigir el tráfico de vehículos, cortar el suministro de agua de toda la zona, etc.). Las reparaciones podían durar días y hasta incluso semanas.

Figura N° 2

Instalación de Tubería Matriz



Fuente: propia del autor

Es por ello importante resaltar lo importante que es contar con estos equipos en la actualidad para poder así reparar las tuberías en cuestión de horas, gracias a la localización exacta del tramo de tubería dañada.

a. Manómetro

El manómetro (del gr. μανός, ligero y μέτρον, medida) es un instrumento de medición para la presión de fluidos contenidos en recipientes cerrados. Se distinguen dos tipos de manómetros, según se empleen para medir la presión de líquidos o de gases.

Figura N° 3

Manómetro de Glicerina para Fluidos presurizados

Fuente: propia del autor

Las unidades que más se utilizan en manómetros para fluidos PSI, Mpa y Bar.

- PSI = Libras por pulgada cuadrada
- Mpa = Megapascal
- Bar = 1 Atmosfera = 1000000 de barias

Tabla N° 1 Unidad de medidas de un Manómetro

1 Bar	0.1 Megapascal	14.5038 psi

Fuente: propia del autor

Este tipo de instrumentos funcionan de la misma forma que un manómetro convencional, pero con la diferencia de que poseen glicerina.

La función de la glicerina es proteger el mecanismo interno del manómetro; este relleno brinda estabilidad a la aguja indicadora, una vez que el instrumento ha sido instalando en zonas de vibraciones.

Cuando un manómetro no posee glicerina, este puede no funcionar de la forma correcta, provocando que se atore la aguja y obteniendo mediciones alteradas.

Asimismo, es importante señalar que si el manómetro posee una filtración de aceite en la caratula, es mejor cambiarlo, ya que esta fisura, también, provocaría que se logre una medición errónea.

Los manómetros de glicerina pueden ser rellenados con glicerina para baja temperatura o silicona, cualesquiera de las dos opciones brindarán la protección adecuada al equipo.

Los manómetros con glicerina son realmente útiles en aplicaciones donde intervenga:

- Aire
- Agua
- Aceite
- Fluidos compatibles con la conexión y mecanismo
- Sistemas hidráulicos
- Sistemas oleo hidráulicos
- Turbinas

• Motores

• Uso industria (donde haya vibración y golpes

Existen diferentes aplicaciones y modelos de manómetros, los utilizados en (ECRF) son del tipo de metal con relleno de glicerina y se utilizan para las labores de toma de presión en los medidores domésticos, la presión para la ciudad de Lima oscila entre 20 y 65 psi, el rango mínimo de presión que se necesita para poder realizar trabajos de detección de fugas es de 10 psi, debajo de ese valor no se puede realizar trabajos de detección de fugas.

Una hoja de datos técnicos del manómetro utilizado por el ECRF

Figura N° 4

Manómetro de Glicerina especificaciones técnicas

Manómetro con muelle tubular
Modelo 213.53, Líquido de relleno, Caja acero inoxidable

Hoja técnica WIKA PM 02.12

otras homologaciones
véase página 2

Aplicaciones

- Para puntos de medida con elevadas cargas dinámicas y vibraciones
- Para medios gaseosos, líquidos, no viscosos y no cristalizantes, compatibles con aleaciones de cobre
- Hidráulica
- Compresores, industria naval

Características

- Resistente contra vibraciones y golpes
- Construcción de extrema robustez
- DN 63 y 100 con homologación Germanischer Lloyd y Gost
- Rangos de indicación hasta 0 ... 1.000 bar

Manómetro con muelle tubular, modelo 213.53.100, conexión inferior

Descripción

Versión
EN 837-1

Diámetro en mm
50, 63, 100

Clase de exactitud
DN 50, 63: 1,6
DN100: 1,0

Rangos de indicación
DN 50: 0 ... 1 a 0 ... 400 bar
DN 63, 100: 0 ... 0,6 a 0 ... 1.000 bar
así como todas las gamas correspondientes para presión negativa y sobrepresión negativa y positiva

Carga de presión máxima

DN 50, 63:	Carga estática:	3/4 x valor final de escala
	Carga dinámica:	2/3 x valor final de escala
	Puntual:	valor final de escala
DN 100:	Carga estática:	valor final de escala
	Carga dinámica:	0,9 x valor final de escala
	Puntual:	1,3 x valor final de escala

Temperatura admisible
Ambiente: -20 ... +60 °C
Medio: máx. +60 °C

Influencia de temperatura
En caso de desviación de la temperatura de referencia en el sistema de medición (+20 °C): máx. ±0,4 %/10 K de la gama de indicación

Tipo de protección
IP 65 según EN 60529 / IEC 60529

Fuente: propia del autor

b. Geófono

El Geófono, es un detector de pérdidas de fugas de agua localiza con precisión fugas subterráneas. Elimina costos de excavaciones y reduce tiempos en la localización de fugas.

Geófonos Mecánicos

Los geófonos son mecánicos y electrónicos, operan con el principio de sismógrafo. Son extremadamente sensibles. Un operador experimentado de geófonos puede incluso determinar el tamaño de una fuga con gran precisión.

Muchas empresas distribuidoras de aguas y plantas industriales están utilizando con éxito este instrumento para detectar fugas de aguas subterráneas.

El principio de funcionamiento se basa en que los “auriculares” son colocados en el piso, siendo éstos muy sensibles. Estos recogen el sonido de las vibraciones, estos son amplificados por el instrumento y se transmitirán al auricular del operador.

El más leve goteo se puede escuchar y se puede rastrear este moviendo los “auriculares de piso” hasta encontrar el punto de máximo sonido.

Figura N° 5

Geófono de piso mecánico, muy similar a un estetoscopio

Fuente: propia del autor

Geófonos Electrónicos

Se cuenta con Geófonos Electrónicos para la escucha de las fugas en las tuberías y son de 2 tipos:

a. Geófono Lmic

La sencillez en el manejo del LMIC®, así como la doble funcionalidad de varilla de escucha para válvulas, acometidas y terrenos blandos, y el sensor de escucha para terrenos duros, permiten a este equipo obtener importantes prestaciones en detección de fugas con una muy interesante relación calidad-precio.

El LMIC® está provisto de un sensor de amplificación que posibilita detectar ruidos de fuga que el oído humano no es capaz de escuchar. Los auriculares disponen de control de volumen independiente y se conectan directamente al módulo sensor, que incorpora la electrónica y la batería.

Presionando un simple botón, el usuario activa el sistema para su escucha. Una vez se deja de presionar, el circuito de alimentación se desactiva.

Figura N° 6

Geófono electrónico LMIC

Fuente: propia del autor

b. Geófono Xmic

El Xmic es un geófono, muy avanzado en tecnología, diseñado para amplificar el ruido que genera el agua que escapa de las tuberías, en situaciones de fuga. Al identificar la posición del ruido de fuga más agudo, estaremos en presencia de la fuga propiamente tal.

Este geófono está compuesto por un módulo amplificador, muy liviano, con un cargador de baterías integrado, un set de audífonos con calidad HI-FI y un sensor de superficie para la escucha del suelo.

El Xmic es un geófono muy avanzado y de fácil uso. Posee lo último en tecnología de amplificación acústica, entregando excelente calidad en sonido, mientras una gran cantidad de características propias del Xmic ayudan en la localización eficaz y precisa de las fugas en el terreno.

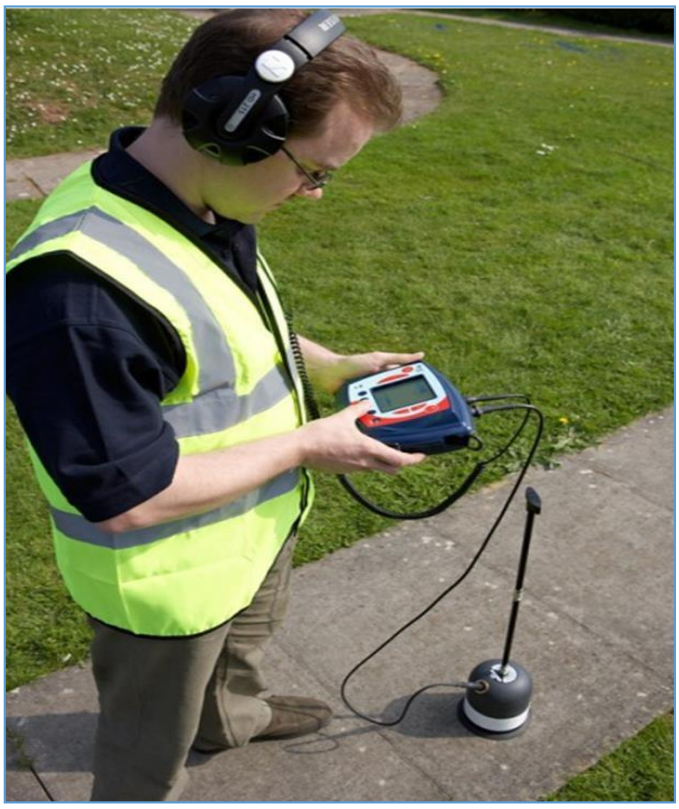
Figura N°7

Geófono electrónico XMIC

Fuente: Cortesía PALMER INC

Figura N°8

Correcto uso de los Geófonos electrónicos XMIC y Lmic



Fuente: Cortesía PALMER INC

Correlador

Los Correladores son poderosos dispositivos electrónicos de localización de fugas en tuberías a presión, donde la ubicación aproximada de la fuga se desconoce y las distancias son relativamente altas. Dos (o más) sensores se colocan en contacto con la tubería a ambos lados de la fuga sospechada. Esos sensores registran y transmiten el sonido por radio a la unidad de procesamiento. Algoritmos matemáticos se utilizan para determinar la ubicación exacta de ciertos perfiles de ruido (por ejemplo, silbido de una fuga) en la tubería, mediante la correlación de los ruidos que llegan a los dos sensores y midiendo la diferencia del viaje del sonido en la tubería desde la fuga hacia cada sensor.

Existen Correladores que hacen la correlación en tiempo real y en otros modelos es necesario descargar la información a una pc y recién de ahí hacer la correlación

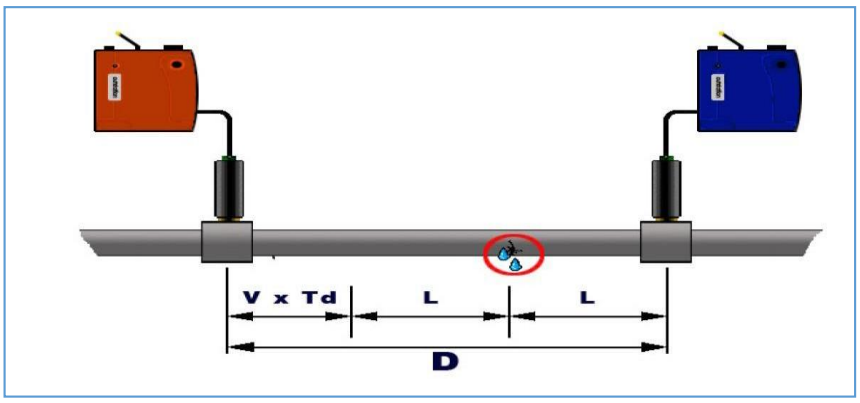
EL principio de la correlación está definido por L = ½ (D – V x Td)

V= Velocidad del sonido dentro de la tubería

Td = Diferencia de la velocidad del sonido entre uno y otro sensor

Figura N° 9

Funcionamiento Correlador Palmer



Fuente: Cortesía PALMER INC

a. Correlador electrónico Palmer Micro Call+ (Correlación en tiempo real)

Correlador Palmer (1)

Radios Transmisoras de Señal (2,3 y 4)

Audífonos Estéreo HI-FI

Sensores de Audio (6 y 7)

Figura N° 10

Correlador Palmer +

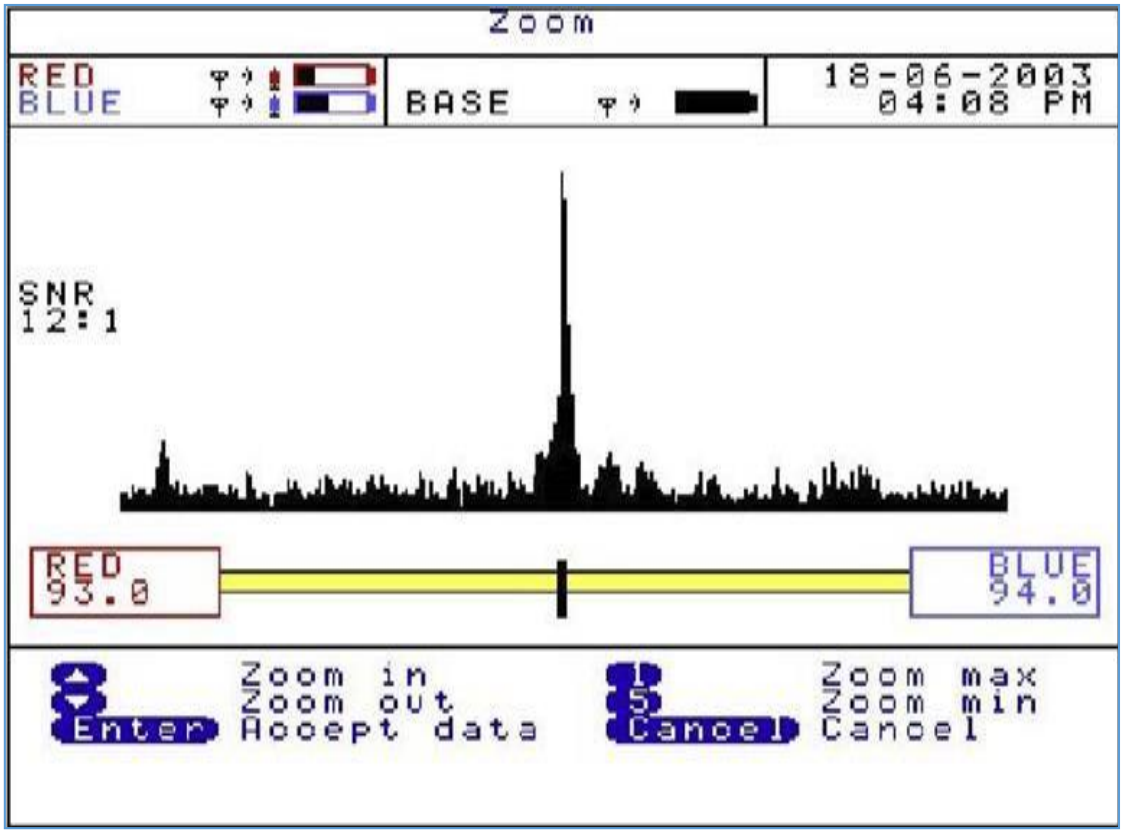
Fuente: Cortesía PALMER INC

Se utiliza para detectar fugas de agua No visibles haciendo uso de sus 2 o 3 sensores, instalados intercaladamente en las conexiones domiciliaras.

Se conectan a la unidad central vía ondas de radio (FM) y esta da la posición aproximada de la tubería rota.

Figura N° 11

Correlador Palmer + en funcionamiento



Fuente: Cortesía PALMER INC

b. EL Correlador electrónico Aquascan 610 (Correlación en tiempo real)

- Correlador AquaScan (1)
- Radios Transmisoras de Señal (2 y 3)
- Audífonos Estéreo Bluetooth
- Sensores de Audio (2 y 3)

Figura N° 12

Correlador AquaScan en funcionamiento

Fuente: Gutermann INC

c. Correlador electrónico Multipunto Soundsens (No es en tiempo real)

- Correlador Soundsens (1)
- Radios Transmisoras de Señal (2)
- Audífonos Estéreo HI-FI
- Computadora
- Se obtiene una mejor respuesta por que se utilizan más de 3 sensores

Figura N° 13

Correlador Soundsens

Fuente: propia del autor

Soundsens, ha sido diseñado con lo último en software y hardware para entregar los más altos resultados, versatilidad y rapidez al utilizarse en campo.

Los operadores pueden programar y descargar los datos del equipo, sin la necesidad de un computador. La unidad es capaz de almacenar semanas de datos antes de ser descargados éstos a un PC, facilitando las operaciones en campo. Los sensores se descargan a través de señales infra-rojas, por lo que no hay necesidad de conexión directa física a través de cables. Todo ésto hace que la instalación y descarga de datos, sea mucho más sencilla.

Mas maletas pueden ser interconectadas, para permitir la descarga o programación simultanea de todos los equipos existentes. Ésto, combinado con una conexión USB para comunicación con el computador personal (PC), hace al proceso de detección de fugas algo completamente rápido y eficiente.

Cuadro N° 2

Correlador Soundsens Especificaciones

Sensor de Entrada		Acelerómetro Interno de Alto Rendimiento para recoger todos los ruidos audibles de las tuberías
	Anclaje del logger / Sensor	Poderoso magneto de anclaje para instalación de los loggers en tuberías / válvulas
Características del Registrador	Memoria	Grabación de 650,000 lecturas (memoria expandible a 1,35 millones de lecturas de acuerdo a requerimiento) Registros individuales pueden ser pre-programados en series de 32 registros individuales.
	Frecuencia de muestreo	Ajustable por el usuario
	Inicio Retardado	Comienzo a cualquier hora del día o posterior a un período de tiempo ajustado, por ejemplo, para correlaciones nocturnas sin supervisión.
	ID del Logger	Número de ID programado de fábrica Además, se puede introducir otro ID para facilitar el reconocimiento por parte de los operadores
	Reloj	Incorporado, de 24 horas, en tiempo real con asignación de fecha
	Software	Compatible con el software de análisis de Radcom, Sound Sens
Comunicaciones	Maleta a PC	A través de conexión USB para descarga rápida de datos
	Maleta a Loggers	Comunicación simultánea desde los loggers a la maleta, vía infra-rojo
	Maleta a Maleta	Vía cable RS232 de 9 pin
Equipo	Dimensiones	Logger avanzado de correlación, con dimensiones reducidas: 160 Alto (incluído magneto) x 55Diam. mm (6.3A x2.10D")
	Construcción	Logger: Carcasa de aluminio, cubierta con pintura spray en polvo Maleta de Transporte: Estructura resistente con revestimiento de aluminio
	Peso	Cada logger correlador: 0.7 kg (1.54 lb)
	Maleta de transporte	Versiones de 2 - 4 loggers o de 6 - 8 loggers (Dos maletas pueden interconectarse)
	Temp. de Trabajo	-10 a +50°C (14 a +122°F)
	Protección Ambiental	Individual Correlator Pods: IP68 submersible
	Energía	Loggers de correlación tienen baterías de ion-litio, operativas hasta por 5 años en condiciones normales de trabajo. Maleta de Transporte contiene baterías de NiCad que deben ser cargadas aproximadamente una vez por mes o menos. Junto con el kit, se suministra un cable de conexión a la red eléctrica (transformador 110v/220v incluido en el producto).

Fuente: propia del autor

Con una vida típica de baterías de 5 años, pantalla LCD con luz de fondo y una simple interfaz, Soundsens está listo para ser usado como Correlador portátil diurno o Correlador fijo durante la noche (ideal para sectores con fugas difíciles de encontrar).

Figura N° 14

Software Correlador Soundsens

Fuente: propia del autor

2.3 Definición de términos básicos

2.3.1.- Fugas Programadas

El servicio de detección de fugas no visibles de fugas programadas como su nombre lo indica es el que se encarga de realizar una programación previamente establecida de cuadrantes y sectores

dentro de la ciudad de Lima y Callao, en esta programación se hace un recorrido o barrido de calles tratando de detectar fugas no visibles para su posterior notificación y/o reparación.

Dicha programación una vez culminada, en los 50 distritos de la ciudad de Lima, el área metropolitana de Lima se distribuye sobre 50 distritos, que son parte integrante de la Provincia de Lima (43 distritos) sumada a la Provincia Constitucional del Callao (7 distritos), se repite en forma de ciclos variando el orden de los distritos cada vez, por lo general se repite los ciclos cada 2 años aproximadamente.

2.3.2.- Fugas Emergencia

El servicio de detección de fugas no visibles de fugas de emergencia como su nombre lo indica son reportadas por las mismas Zonales de Sedapal (oficinas de centro de servicio y cobro ubicadas estratégicamente en varios distritos de la capital).

La información reportada por las zonales es recolectada por diversos medios: llamadas telefónicas, inspecciones realizadas por Sedapal, denuncias presenciales, rotura de pistas por paso de vehículos pesados, etc.

Estas fugas Emergencia son atendidas por las unidades móviles de acuerdo a su importancia y/o urgencia y se determina si efectivamente existe una fuga de agua o no luego de lo cual se procede hace su informe respectivo catalogando y clasificando la fuga y su informe es

ingresado a la base de datos del sistema para su posterior reparación y/o archivamiento en caso la fuga no haya sido detectada.

FILTRADO Y DISCRIMINACION DE DATOS

No todos los campos son necesarios para hacer el diseño del algoritmo por lo tanto se eliminarán aquellos campos que no sean relevantes para nuestro estudio tales como nombre dirección, código interno y códigos de fecha duplicados, etc.

Detallaremos los campos que se han conservado para su posterior análisis y estudio.

NUMERO DE FUGAS

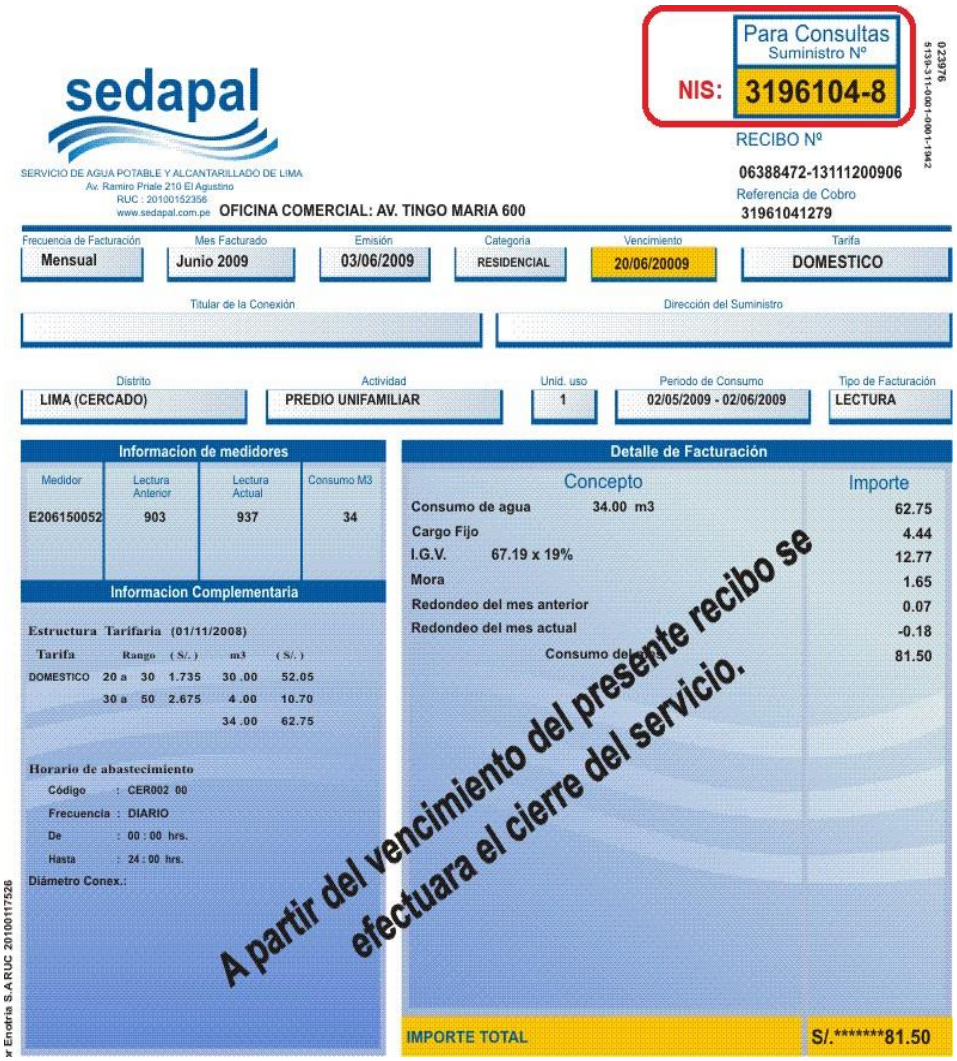
La cantidad de fugas registradas en el sistema en estos años incluyen las fugas detectadas y las que no eran fugas también para el archivo.

NIS

Número de Identificación del Suministro (NIS), es un código de nueve dígitos,  e invariables que identifican el suministro y dos variables que identifican el contrato vigente. Este se encuentra en todos los recibos emitidos por la empresa  adora del servicio y es único.

Figura N° 15

Detalle de un recibo de facturación de agua y el número de NIS



Fuente: propia del autor

FECHA

Fecha de la detección de la fuga

DISTRITO

Donde se realizó la auscultación de la fuga o fugas

Figura N° 16

Detalle de los 50 Distritos de la Ciudad de Lima y Callao

Fuente: propia del autor

URBANIZACION

Donde se realizó la auscultación de la fuga o fugas

Figura N° 17

Detalle de un NIS ubicado en el plano geo referenciado

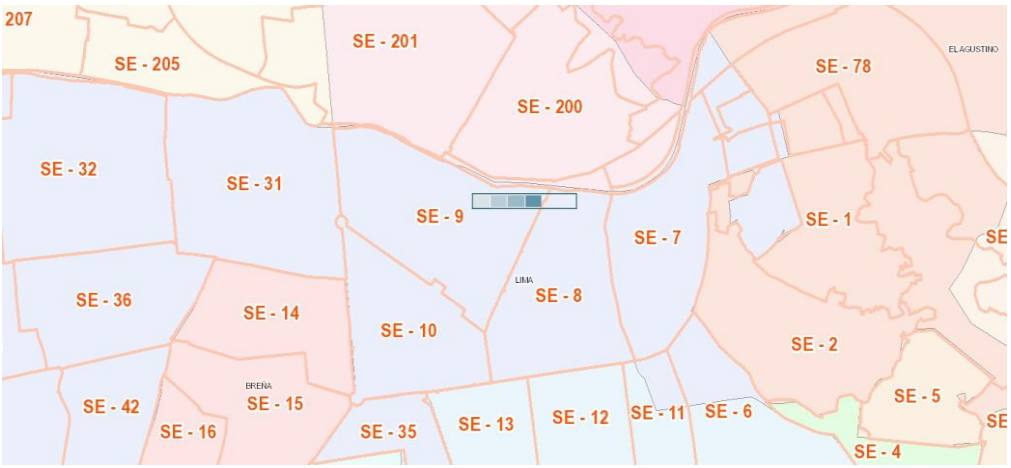
Fuente: propia del autor

CODIGO DE SECTOR

Como ya se indico anteriormente los 50 distritos de Lima están subdivididos a su vez en sectores para que sea más rápida la ubicación de un suministro o NIS para ello se cuenta con una codificación propia.

Figura N° 18

Detalle del distrito de Lima y sus Sectores



Fuente: propia del autor

El proyecto de sectorización de Lima y Callao tuvo en cuenta los siguientes parámetros:

- El Área debe ser menor a 3 Km2
- Las Presiones deben estar entre 15 PSI y 50 PSI.
- Utilizar de preferencia las avenidas como límite de sector, en especial aquellas donde el tendido de tubería es por ambos lados de la calle.
- Definir anillos o circuitos conformados por tuberías de gran capacidad.

- Se debe evitar en lo posible dejar puntos muertos en la red, considerando redes secundarias complementarias que los anule.

- Las tuberías mayores e iguales a 6” de diámetro que crucen el límite de un sector deben cerrarse por medio de una válvula, mientras que las tuberías menores o iguales a 4” deberán ser cortadas o taponadas. Asimismo, se dejará disponible por lo menos un pase de emergencia.

- Los sectores, de ser posible, respetarán los límites de separación de zonas de presión.

En el caso por ejemplo del Distrito de Lima Cercado este cuenta con 10 Sectores enumerados SE-1 hasta SE-10

OBJETIVOS DE LA SECTORIZACION

Los objetivos que persigue la sectorización son:

- Permitir controlar, en un área definida, parámetros importantes para el buen funcionamiento del Sistema de Distribución de Agua Potable. Estos parámetros son: caudal de ingreso al sector y presión en la red (que debe ser entre de 15 a 50 psi)

- Permitir la aplicación de una justa política de racionamiento de agua, en épocas de escasez, mediante la correcta utilización de fuentes superficiales y subterráneas, en lo que se denomina uso conjuntivo.
- Determinar la cantidad de agua no Facturada , obtenida como la diferencia del volumen de agua que ingresa al sector y el volumen facturado, obtenido a través de la micro medición.
- Permitir el aislamiento de un sector con respecto al resto del sistema a fin de realizar trabajos de mantenimiento y reparación por problemas de emergencia en una zona definida de la red de agua. Con ello se reducirá las molestias a los usuarios por falta de agua, pasando una gran área del Sistema de Distribución afectada hacia un pequeño sector en el futuro.

TIPO DE FUGA

- **La fuga en Caja**

  Se refiere a las fugas de agua que se suscitan en el mismo buzón de la conexión domiciliaria (medidor de agua) y las llaves de paso, llamadas llaves telescópicas, son las más comunes y las más fáciles de detectar.

- **La fuga en Línea**

  Se refiere a la tubería de PVC que normalmente es de ½ pulgada que sale desde el medidor de agua y se prolonga hasta la tubería matriz y puede tener una longitud variable que llegan hasta los 12 metros en algunos casos y la cual se encuentra a una

profundidad de hasta 2 metros. Su sonido es característico, pero no son visibles a simple vista y dependerá de la presión del agua y del tipo de terreno donde se encuentra la tubería para que sea fácil su escucha.

- **La fuga en Corporation**

  Se refiere al tramo de la tubería matriz generalmente de 4 a 32 pulgadas donde va colocada una abrazadera que servirá de conexión entre esta y la tubería de ½ pulgada de PVC.

- **La fuga en Tubería**

  Se refiere a la tubería matriz generalmente de 4 a 32 pulgadas que alimenta a las conexiones domiciliarias y que puede estar construida de PVC , FOFO , Hierro Ductil , etc .

- **Valvula de Red**

  Se refiere a las válvulas colocadas en las tuberías principales para controlar y regular el paso y la presión de agua de un sector a otro

- **Valvula Grifo / CL**

  Se refiere a los Hidrantes colocados en las esquinas para apagar los incendios Un hidrante de incendio o boca de incendio es una toma de agua diseñada para proporcionar un caudal considerable en caso de incendio. El agua puede obtenerla de la red urbana de abastecimiento o de un depósito, mediante una bomba

DIAMETRO DE LA TUBERIA

Se refiere al diámetro de la tubería en donde se presenta la fuga no visible, la cual puede varias desde los 25 mm^2 (1/2 pulgada) hasta las 1600 mm^2 (32 pulgadas).

IPO DE TUBERIA

Se refiere al material del cual está fabricada la tubería de agua y puede influir en el tiempo de vida de esta por el material y la marca.

Se clasifican en :

- PVC
- HIERO DUCTIL
- FIERRO FUNDIDO
- ASBESTO CEMENTO

PRESION

La presión del agua es un factor muy importante porque incidirá en la cantidad de agua desperdiciada por la fuga en un lapso de tiempo ya que a mayor presión de agua, mayor será el volumen de agua no facturada.

Las presiones en la ciudad de Lima no son uniformes registrándose valores que van desde 0 PSI hasta los 110 PSI en algunos casos.

Cuadro N° 3

Tuberías comerciales para agua potable y sus diámetros interior y exterior

Diámetro de la Tubería (in)	Diámetro Exterior (in)	Pared Mínima (in)	Diámetro Interior Promedio (in)	Peso de la Tubería (lbs/ft)	Presión Máxima de Agua a 73 °F (psi)
¼	0.540	0.119	0.288	0.110	1130
⅜	0.675	0.126	0.407	0.153	920
½	0.840	0.147	0.528	0.225	850
¾	1.050	0.154	0.724	0.305	690
1	1.315	0.179	0.935	0.450	630
1 ¼	1.660	0.191	1.256	0.621	520
1 ½	1.900	0.200	1.476	0.754	470
2	2.375	0.218	1.913	1.043	400
2 ½	2.875	0.276	2.289	1.594	420
3	3.500	0.300	2.864	2.132	370
4	4.500	0.337	3.786	3.116	320
6	6.625	0.432	5.709	5.951	280
8	8.625	0.500	7.565	9.040	250
10	10.750	0.593	9.492	13.413	230
12	12.750	0.687	11.294	18.440	230
14	14.000	0.750	12.410	22.119	220
16	16.000	0.843	14.214	28.424	220

Fuente: propia del autor

CAUDAL (LT/DIA)

Es la cantidad de agua no facturada que se pierde por una fuga no visible y dependerá de diversos factores, tales como la presión de agua, el diámetro de la tubería, el tipo de tubería, etc. y se calcula en función de la cantidad de litros por día perdidos por dicha fuga los cuales son por lo general aproximados.

VEHICULO

Se refiere a la unidad móvil encargada de realizar la detección de fugas identificada con un código único.

FUGAS DETECTADAS

La cantidad de fugas detectadas por inspección pueden variar en numero.

OBSERVACIONES

Algunas observaciones o comentarios respecto a la fuga o fuas encontradas.

MODELO PREDICTIVO

El modelo predictivo es un modelo de datos, basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a un determinado evento. El modelo predictivo utiliza estadísticas para predecir los resultados.

Según nuestra Hipótesis planteada **“El desarrollo de un algoritmo predictivo el cual permitirá la detección temprana de fugas de agua en las redes de agua potable de la ciudad de Lima, 2019”**

Para poder llevar a cabo dicha labor necesitamos primero recopilar la información con la cual desarrollar nuestro modelo.

Para eso como indicamos el equipo ECRF es el encargado de cumplir 2 objetivos: el objetivo de primer nivel “Disminuir el Agua No Facturada” y el objetivo de segundo nivel “Reducir los volúmenes de pérdidas de agua potable”; lo que permitirá reducir las pérdidas de agua y brindar un mejor servicio.

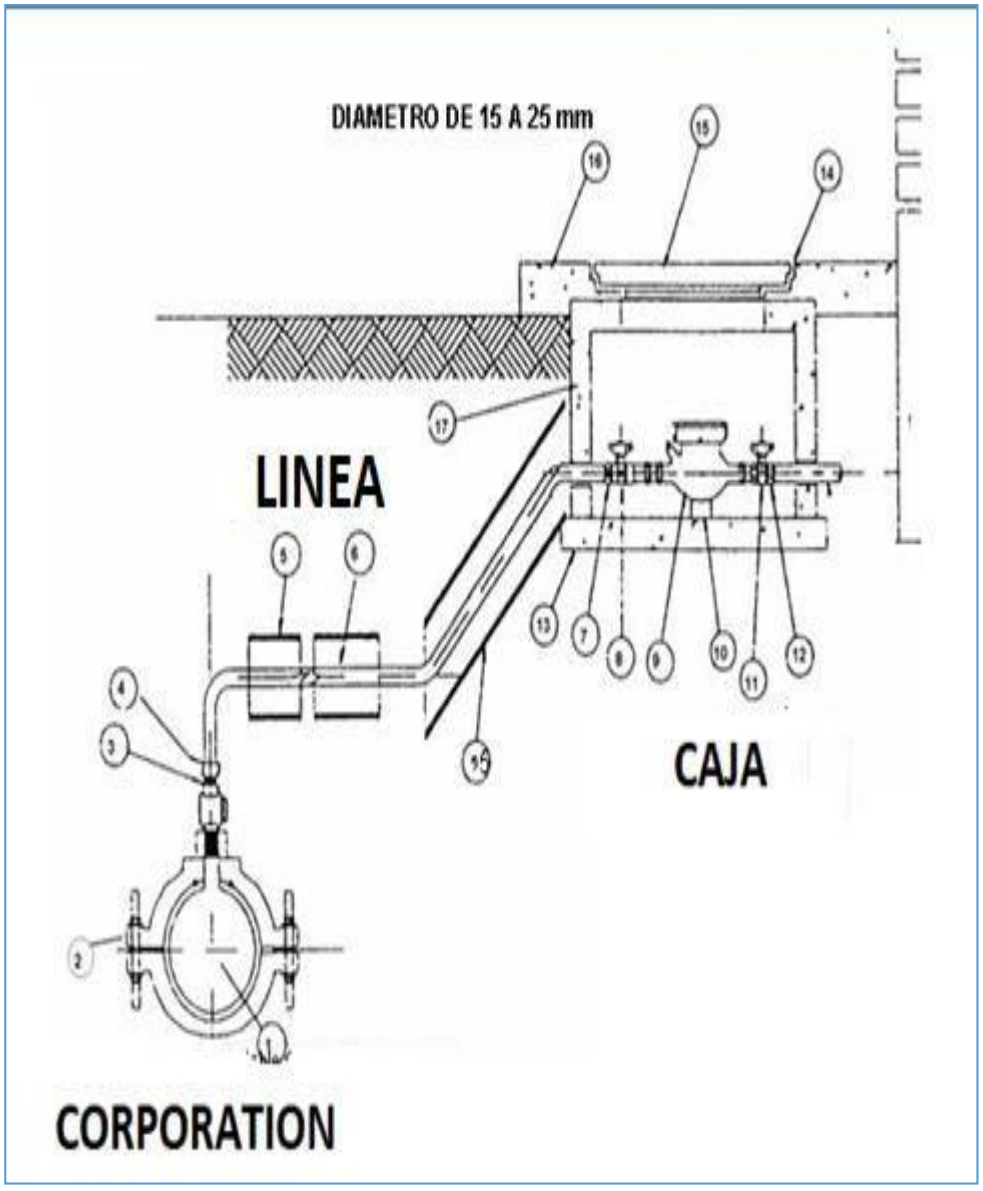
Los Distritos son los 50 que se encuentran en la Ciudad de Lima incluyendo La Provincia Constitucional del Callao.

Los tipos de Fugas que se pueden encontrar son de los siguientes tipos:

- "Corporation
- ","Tuberia",
- "Caja A/Med",
- "Caja D/Med",
- "Linea A/Caja",
- "Linea D/Caja",
- "Valvula Red",
- "Valvula Grifo C/L",
- "Medidor"

Figura Nº 19

Corte Transversal de un sistema de conexión domiciliaria de agua



Fuente: Propia del autor

Para ello luego de detectada la fuga se hace un reporte con fecha y hora , el cual contiene información con respecto a la fuga encontrada el caudal

de la fuga , el tipo de esta , la presión , así como la dirección completa , el distrito y el número de suministro del usuario llamado **NIS**.

Figura Nº 20

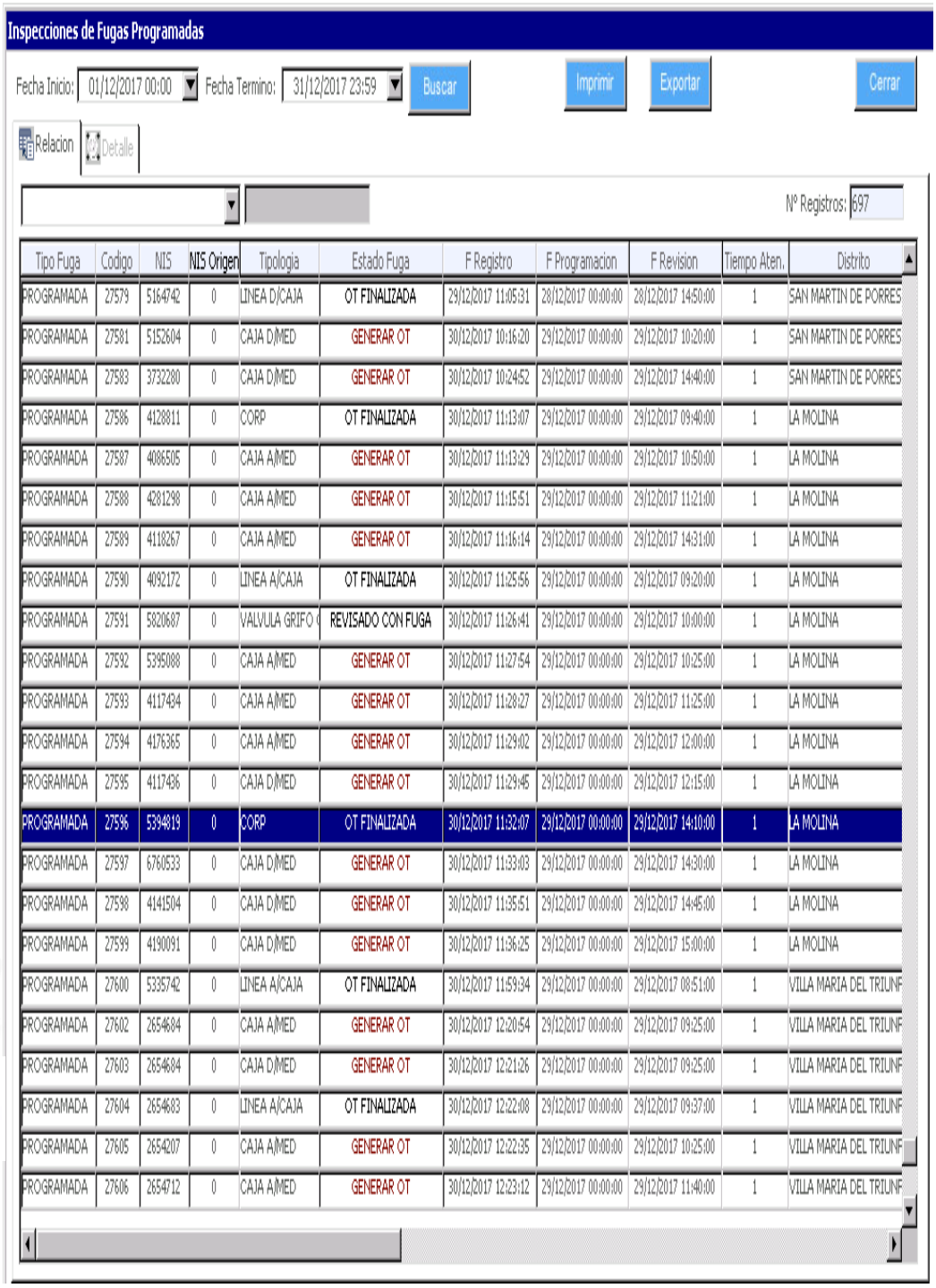
Software de Gestión de Incidencias operativas SGIO

Fuente: Sedapal

Toda esa información que le llamares **BASE DE DATOS DE FUGAS** que contiene la información histórica, es la que usaremos para el desarrollo de nuestro algoritmo.

Figura Nº 21

Base de Datos de Fugas 2014 - 2019



Inspecciones de Fugas Programadas

Fecha Inicio: 01/12/2017 00:00 Fecha Termino: 31/12/2017 23:59 Buscar Imprimir Exportar Cerrar

Relacion Detalle

Nº Registros: 697

Tipo Fuga	Codigo	NIS	NIS Origen	Tipologia	Estado Fuga	F Registro	F Programacion	F Revision	Tiempo Aten.	Distrito
PROGRAMADA	27579	5164742	0	LINEA D/CAJA	OT FINALIZADA	29/12/2017 11:05:31	28/12/2017 00:00:00	28/12/2017 14:50:00	1	SAN MARTIN DE PORRES
PROGRAMADA	27581	5152604	0	CAJA D/MED	GENERAR OT	30/12/2017 10:16:20	29/12/2017 00:00:00	29/12/2017 10:20:00	1	SAN MARTIN DE PORRES
PROGRAMADA	27583	3732280	0	CAJA D/MED	GENERAR OT	30/12/2017 10:24:52	29/12/2017 00:00:00	29/12/2017 14:40:00	1	SAN MARTIN DE PORRES
PROGRAMADA	27586	4128811	0	CORP	OT FINALIZADA	30/12/2017 11:13:07	29/12/2017 00:00:00	29/12/2017 09:40:00	1	LA MOLINA
PROGRAMADA	27587	4086505	0	CAJA A/MED	GENERAR OT	30/12/2017 11:13:29	29/12/2017 00:00:00	29/12/2017 10:50:00	1	LA MOLINA
PROGRAMADA	27588	4281298	0	CAJA A/MED	GENERAR OT	30/12/2017 11:15:51	29/12/2017 00:00:00	29/12/2017 11:21:00	1	LA MOLINA
PROGRAMADA	27589	4118267	0	CAJA A/MED	GENERAR OT	30/12/2017 11:16:14	29/12/2017 00:00:00	29/12/2017 14:31:00	1	LA MOLINA
PROGRAMADA	27590	4092172	0	LINEA A/CAJA	OT FINALIZADA	30/12/2017 11:25:56	29/12/2017 00:00:00	29/12/2017 09:20:00	1	LA MOLINA
PROGRAMADA	27591	5820687	0	VALVULA GRIFO (	REVISADO CON FUGA	30/12/2017 11:26:41	29/12/2017 00:00:00	29/12/2017 10:00:00	1	LA MOLINA
PROGRAMADA	27592	5395088	0	CAJA A/MED	GENERAR OT	30/12/2017 11:27:54	29/12/2017 00:00:00	29/12/2017 10:25:00	1	LA MOLINA
PROGRAMADA	27593	4117434	0	CAJA A/MED	GENERAR OT	30/12/2017 11:28:27	29/12/2017 00:00:00	29/12/2017 11:25:00	1	LA MOLINA
PROGRAMADA	27594	4176365	0	CAJA A/MED	GENERAR OT	30/12/2017 11:29:02	29/12/2017 00:00:00	29/12/2017 12:00:00	1	LA MOLINA
PROGRAMADA	27595	4117436	0	CAJA D/MED	GENERAR OT	30/12/2017 11:29:45	29/12/2017 00:00:00	29/12/2017 12:15:00	1	LA MOLINA
PROGRAMADA	27596	5394819	0	CORP	OT FINALIZADA	30/12/2017 11:32:07	29/12/2017 00:00:00	29/12/2017 14:10:00	1	LA MOLINA
PROGRAMADA	27597	6760533	0	CAJA D/MED	GENERAR OT	30/12/2017 11:33:03	29/12/2017 00:00:00	29/12/2017 14:30:00	1	LA MOLINA
PROGRAMADA	27598	4141504	0	CAJA D/MED	GENERAR OT	30/12/2017 11:35:51	29/12/2017 00:00:00	29/12/2017 14:45:00	1	LA MOLINA
PROGRAMADA	27599	4190091	0	CAJA D/MED	GENERAR OT	30/12/2017 11:36:25	29/12/2017 00:00:00	29/12/2017 15:00:00	1	LA MOLINA
PROGRAMADA	27600	5335742	0	LINEA A/CAJA	OT FINALIZADA	30/12/2017 11:59:34	29/12/2017 00:00:00	29/12/2017 08:51:00	1	VILLA MARIA DEL TRIUNF
PROGRAMADA	27602	2654684	0	CAJA A/MED	GENERAR OT	30/12/2017 12:20:54	29/12/2017 00:00:00	29/12/2017 09:25:00	1	VILLA MARIA DEL TRIUNF
PROGRAMADA	27603	2654684	0	CAJA D/MED	GENERAR OT	30/12/2017 12:21:26	29/12/2017 00:00:00	29/12/2017 09:25:00	1	VILLA MARIA DEL TRIUNF
PROGRAMADA	27604	2654683	0	LINEA A/CAJA	OT FINALIZADA	30/12/2017 12:22:08	29/12/2017 00:00:00	29/12/2017 09:37:00	1	VILLA MARIA DEL TRIUNF
PROGRAMADA	27605	2654207	0	CAJA A/MED	GENERAR OT	30/12/2017 12:22:35	29/12/2017 00:00:00	29/12/2017 10:25:00	1	VILLA MARIA DEL TRIUNF
PROGRAMADA	27606	2654712	0	CAJA A/MED	GENERAR OT	30/12/2017 12:23:12	29/12/2017 00:00:00	29/12/2017 11:40:00	1	VILLA MARIA DEL TRIUNF

Fuente: Sedapal

ANÁLISIS PREDICTIVO Y MACHINE LEARNING

Esta forma de analytics, la prediciva, sigue siendo user driven, es decir, implica la interacción humana, necesita de la guía de un experto que:

- Confirme hipótesis.
- Determine los requisitos de los datos.
- Establezca prioridades.

Un paso más allá del análisis predictivo se encuentra el machine learning, aprendizaje automático, que para muchos es el núcleo de donde parte análisis predictivo. Sin embargo, en este caso el impulsor no es la mente humana y su expertise sino que son los propios datos. En base a ellos, y solamente a ellos, se generan hipótesis, se profundiza en la información y se obtienen predicciones individuales.

El machine learning requiere muchísima menor preparación de los datos, y también menos supuestos. Además está totalmente orientada a resultados, algo que facilita su monitorización en un entorno de negocio.

Es habitual la confusión entre analítica predictiva y machine learning, pero también es frecuente suponerla equivalente a un pronóstico, cuando, en realidad análisis predictivo es mucho más que el simple forecasting.

El análisis predictivo es algo completamente distinto, más allá de la previsión normalizada que se centra en asignar una puntuación de predicción para cada cliente o cualquier otro elemento organizativo que se busque valorar. Por el contrario, la previsión ofrece estimaciones

agregadas globales, tales como el número total de las adquisiciones del trimestre entrante o los beneficios que se espera generar en un periodo de un año.

Mediante técnicas de forecasting se puede, por ejemplo, estimar el número total de automóviles que se venderán en una determinada región, mientras que gracias al análisis predictivo es posible profundizar en esta información para conocer las características de los clientes individuales más propensos a comprar un coche.

Si análisis predictivo se consideraba una rama del aprendizaje automático, el pronóstico es, sin duda, un componente de cualquier modelo predictivo. Ambos elementos esenciales para inspirar el cambio con dosis de realidad. De ahí el poder de Business Analytics, que se refiere a la exploración de los datos históricos de muchos sistemas de origen a través de análisis estadísticos, análisis cuantitativo, minería de datos, modelado predictivo y otras técnicas de análisis predictivo que, de un modo u otro, hacen posible identificar tendencias y comprender la información que puede impulsar el cambio.

MACHINE LEARNING

El machine learning, conocido en español como aprendizaje automático o aprendizaje de máquina, nació como una idea ambiciosa de la IA en la década de los 60. Para ser más exactos, fue una subdisciplina de la IA, producto de las ciencias de la computación y las neurociencias.

Lo que esta rama pretendía estudiar era el reconocimiento de patrones (en los procesos de ingeniería, matemáticas, computación, etc.) y el aprendizaje por parte de las computadoras. En los albores de la IA, los investigadores estaban ávidos por encontrar una forma en la cual las computadoras pudieran aprender únicamente basándose en datos.

Sucedió con el paso de los años que el machine learning comenzó a enfocarse en diferentes asuntos, tales como el razonamiento probabilístico, investigación basada en la estadística, recuperación de información, y continuó profundizando cada vez más en el reconocimiento de patrones (todos estos asuntos aplicados a procesos de ingeniería, matemáticas, computación y otros campos relacionados con objetos físicos o abstractos).

Esto ocasionó que en los 90 se separara de la IA para convertirse en una disciplina por sí sola, aunque muchos puristas aún la consideran como parte de la IA. Ahora, el principal objetivo del machine learning es abordar y resolver problemas prácticos en donde se aplique cualquiera de las disciplinas numéricas antes mencionadas.

El propósito del machine learning es que las personas y las máquinas trabajen de la mano, al éstas ser capaces de aprender como un humano lo haría. Precisamente esto es lo que hacen los algoritmos, permiten que las máquinas ejecuten tareas, tanto generales como específicas.

Si bien al principio sus funciones eran básicas y se limitaban a filtrar emails, hoy en día puede hacer cosas tan complejas como predicciones

de tráfico en intersecciones muy transitadas, detectar cáncer, mapear sitios para generar proyectos de construcción en tiempo real, e incluso, definir la compatibilidad entre dos personas.

El principal objetivo de todo aprendiz (learner) es desarrollar la capacidad de generalizar y asociar. Cuando traducimos esto a una máquina o computadora, significa que éstas deberían poder desempeñarse con precisión y exactitud, tanto en tareas familiares, como en actividades nuevas o imprevistas.

¿Y cómo es posible esto? Haciendo que repliquen las facultades cognitivas del ser humano, formando modelos que “generalicen” la información que se les presenta para realizar sus predicciones. Y el ingrediente clave en toda esta cuestión son los datos.

En realidad, el origen y el formato de los datos no es tan relevante, dado que el machine learning es capaz de asimilar una amplia gama de éstos, lo que se conoce como big data, pero éste no los percibe como datos, sino como una enorme lista de ejemplos prácticos.

Podríamos decir que sus algoritmos se dividen principalmente en tres grandes categorías: supervised learning (aprendizaje supervisado), unsupervised learning (aprendizaje no supervisado) y reinforcement learning (aprendizaje por refuerzo). A continuación, detallaremos las diferencias entre éstas.

Supervised learning

Depende de datos previamente etiquetados, como podría ser el que una computadora logré distinguir imágenes de coches, de las de aviones. Para esto, lo normal es que estas etiquetas o rótulos sean colocadas por seres humanos para asegurar la efectividad y calidad de los datos.

En otras palabras, son problemas que ya hemos resuelto, pero que seguirán surgiendo en un futuro. La idea es que las computadoras aprendan de una multitud de ejemplos, y a partir de ahí puedan hacer el resto de cálculos necesarios para que nosotros no tengamos que volver a ingresar ninguna información.

Ejemplos: reconocimiento de voz, detección de spam, reconocimiento de escritura, entre otros.

Unsupervised learning

En esta categoría lo que sucede es que al algoritmo se le despoja de cualquier etiqueta, de modo que no cuenta con ninguna indicación previa. En cambio, se le provee de una enorme cantidad de datos con las características propias de un objeto (aspectos o partes que conforman a un avión o a un coche, por ej.), para que pueda determinar qué es, a partir de la información recopilada.

Ejemplos: detectar morfología en oraciones, clasificar información, etc.

Reinforcement learning

En este caso particular, la base del aprendizaje es el refuerzo. La máquina es capaz de aprender con base a pruebas y errores en un número de diversas situaciones.

Aunque conoce los resultados desde el principio, no sabe cuáles son las mejores decisiones para llegar a obtenerlos. Lo que sucede es que el algoritmo progresivamente va asociando los patrones de éxito, para repetirlos una y otra vez hasta perfeccionarlos y volverse infalible.

Ejemplos: navegación de un vehículo en automático, toma de decisiones, etc.

Dado que el machine learning es un sistema basado en el procesamiento y análisis de datos que son traducidos a hallazgos, se puede aplicar a cualquier campo que cuente con bases de datos lo suficientemente grandes. De momento, algunos de sus usos más populares y desarrollados son:

- Clasificación de secuencias de DNA
- Predicciones económicas y fluctuaciones en el mercado bursátil
- Mapeos y modelados 3D
- Detección de fraudes
- Diagnósticos médicos
- Buscadores en Internet
- Sistemas de reconocimiento de voz

- Optimización e implementación de campañas digitales publicitarias
- Sistemas de diagnostico de imágenes medicas
- Bolsa de Valores
- Economía
- Medicina

CAPÍTULO III: HIPÓTESIS Y VARIABLES

3.1 Hipótesis

3.1.1 Hipótesis general

El Algoritmo predictivo desarrollado en el Software R Permitirá el monitoreo temprano de las fugas de agua potable en la ciudad de Lima.

3.1.2 Hipótesis específica

- ✓ H1 El Algoritmo predictivo desarrollado en el Software R permitirá predecir y anticipar los tipos de fugas de agua potable a lo largo de los años.
- ✓ H2 El Algoritmo predictivo desarrollado en el Software R permitirá determinar los factores que afectan las proyecciones estimadas por el algoritmo

3.2 Definición conceptual de variables

3.2.1 Diseño de un Algoritmo Predictivo para Monitoreo Temprano de Redes de Agua Potable en la Ciudad de Lima, 2019" Indicador: Algoritmo Predictivo

Es la variable independiente (X) el algoritmo (programa) el cual permita de manera eficaz y eficiente predecir tempranamente el tipo de fugas que se presentaran con el tiempo.

3.2.2 Determinar con anticipación los factores que afectan la detección de fugas no visibles en las redes de agua potable de la ciudad de Lima a través de este algoritmo

Es la variable dependiente (Y) establece que al anticipar los factores se podrá reducir los tiempos de respuesta y los costos asociados a este servicio.

3.2.3 El crecimiento poblacional afectara los datos proporcionados por el algoritmo

Es la variable interviniente (Z) la cual nos indica que factores afectan las proyecciones estimadas por el algoritmo

3.3 Operacionalización de variables

Variable: X (independiente)

Algoritmo predictivo desarrollado en el Software R para el monitoreo temprano de las fugas de agua potable en la ciudad de Lima 2019 . **Indicador:** Big Data de los reporte de Fugas hasta la fecha.

Variable: Y (dependiente)

Establece que al anticipar los factores se podrá reducir los tiempos de respuesta y los costos asociados a este servicio.

Variable: Z (interviniente)

Factores afectan las proyecciones estimadas por el algoritmo

CAPÍTULO IV. DISEÑO METODOLÓGICO

4.1 Tipo y diseño de investigación

De acuerdo al problema objeto de estudio, esta investigación es predominantemente estadística, experimental tecnológica, aplicada o I+D, científica y transversal, las que se justifican:

- **Experimental tecnológica**. Porque se utilizara hardware y software para la implementación del algoritmo, de igual manera, es tecnológica porque para su instauración se utilizarán equipos electrónicos.
- **Aplicada (I+D)**. Con la Investigación más Desarrollo, se tiene como objetivo, evidenciar que no se ha realizado investigaciones de este tipo y por ende esta primera versión puede ser mejorado y/o optimizado con beneficio para la empresa prestadora del servicio y por ende para los consumidores.
- **Científica**. Porque se aplicarán conocimientos estadísticos y de Big Data usando el software *R* para el tratamiento de la informacion.
- **Transversal**. Toma este nombre porque el inicio y término de esta investigación es menor a un año o doce meses.

4.2 Método de Investigación

4.2 .1 Procesamiento del archivo excel con el software R

Se tienen en total 14 campos los cuales contienen la información más resaltante con respecto al histórico de las fugas detectadas las cuales utilizaremos para realizar un análisis y obtener patrones que nos permitan hacer más fácil el algoritmo predictivo.

Como se indicó se utilizará parte de la data para el estudio 2014 al 2019 y parte se usará como control 2020 para hacer las pruebas al algoritmo predictivo.

Como dato adicional solo consideraremos los caudales con valores iguales o mayores a 1000 litros por ser más relevantes para esta investigación, ya que perdidas a menores a 1000 litros no se pueden considerar como pérdidas significativas.

Se encontró que la fuga de mayor caudal fue una de 150000 litros por dia la cual estaba ubicada en el distrito de San Borja y que ocurrió el año 2017 registrada con NIS 2717633 en el sector SE-72

Se encontró que la fuga de menor caudal fue una de 1000 litros por día la cual estaba ubicada en el distrito del Agustino en el año 2018 registrado con NIS 4200016 en el sector SE-2.

31 fugas fueron registradas en el Callao y las restantes en Lima, se tienen en total 2858 registros de fugas registradas en ese periodo de tiempo con diferentes presiones y diferentes diámetros de tuberías, los cuales se tienen que filtrar por distrito y por caudales para realizar una mejor aproximación del modelo.

Las unidades móviles han ido aumentando a lo largo de los años para poder dar seguimiento a todas las fugas reportadas como emergencias en el año 2014 eran solo 3 unidades móviles y a la fecha ya son 6 unidades móviles que recorren los 50 distritos de la gran Lima.

Por eso la cantidad de fugas procesadas ha crecido en forma ascendente a lo largo de estos años ya que se cuentan con más unidades móviles para dicho fin.

Se tiene por año la cantidad de fugas atendidas:

- El año 2014 se atendieron 62 reportes de Fugas de emergencia
- El año 2015 se atendieron 77 reportes de Fugas de emergencia
- El año 2016 se atendieron 86 reportes de Fugas de emergencia
- El año 2017 se atendieron 563 reportes de Fugas de emergencia
- El año 2018 se atendieron 984 reportes de Fugas de emergencia
- El año 2019 se atendieron 1150 reportes de Fugas de emergencia

4.2.2 Carga de datos al software R

R actualmente en su versión 3.6.3 es un software libre que permite realizar análisis estadísticos y el más usado en la comunidad científica.

Este programa está disponible en la página web: ***http://www.r- project.org*** y consta de una aplicación central y de librerías de multitud de temas que se pueden instalar según necesidad. R es un programa de instrucciones, y por tanto, no resulta del todo "amigable" para los usuarios que no están acostumbrados a este tipo de manejo. Actualmente existe una interfaz que permite el manejo del programa R mediante una ventana de menús, este interfaz se llama RCommander.

Como primer paso debemos instalar el software R y cargar algunas librerías sin las cuales no es posible hacer el tratamiento de la información.

Figura Nº 22

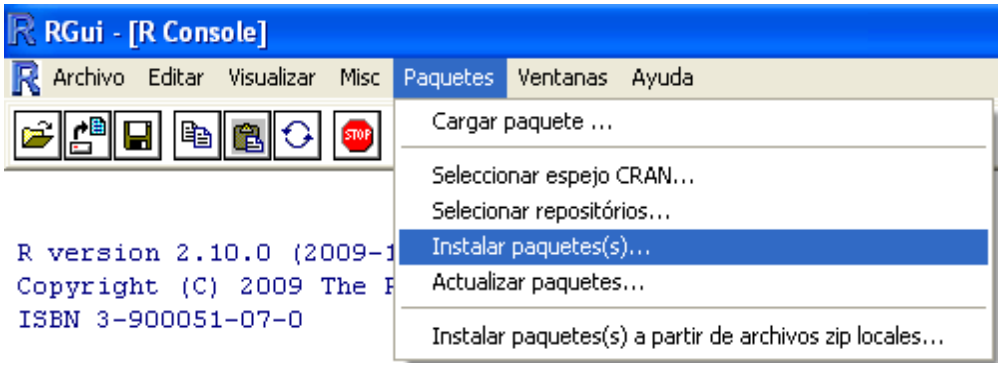
Instalación del Software R

Fuente: Propia del autor

Para instalar las librerías se debe ingresar a la opción Instalar paquetes:

Figura Nº 23

Instalación del Librerías del Software R

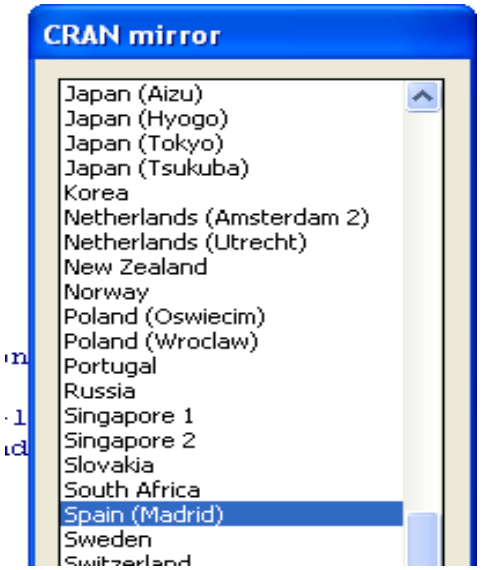


Fuente: Propia del autor

A continuación, seleccionamos la región del que queremos descargar el paquete y pulsamos OK:

Figura Nº 24

Servidores Disponibles Librerías Software R

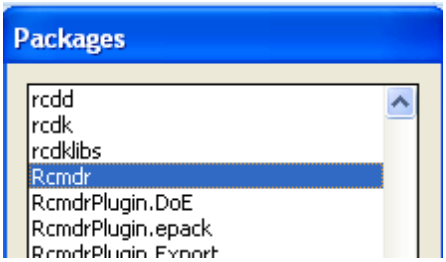


Fuente: Propia del autor

En la ventana Packages que aparece buscaremos desplazándonos con la barra lateral hacia abajo buscando las librerías que necesitamos para nuestro proyecto de investigación y le damos OK.

Figura Nº 25

Instalación del Packages Software R



Fuente: Propia del autor

Comenzará entonces la descarga e instalación de esta librería al terminar la descarga hay que instalarla con el comando **install** seguido del nombre de la librería asi por ejemplo para la librería dplyr, desde la línea de comandos.

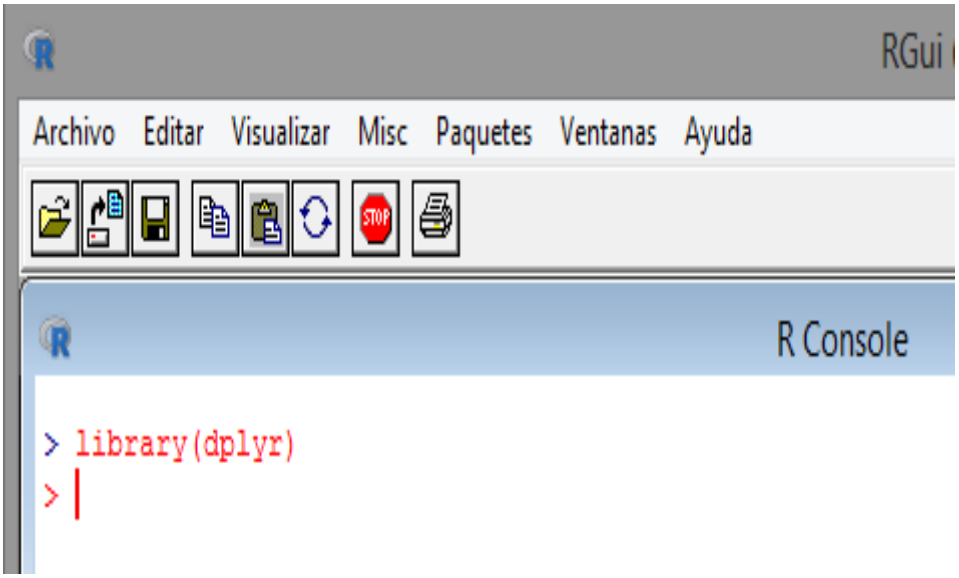
install.packages("dplyr")

Finalmente para poder usar la librería tenemos que llamarla cada vez que iniciemos una sesión en R con el comando library seguido del nombre de la librería.

library(dplyr)

Figura Nº 26

Llamar a una Librería en Software R



Fuente: Propia del autor

Las librerías son útiles para múltiples propósitos y facilitan mucho el tratamiento de la información y el consiguiente diseño algorítmico.

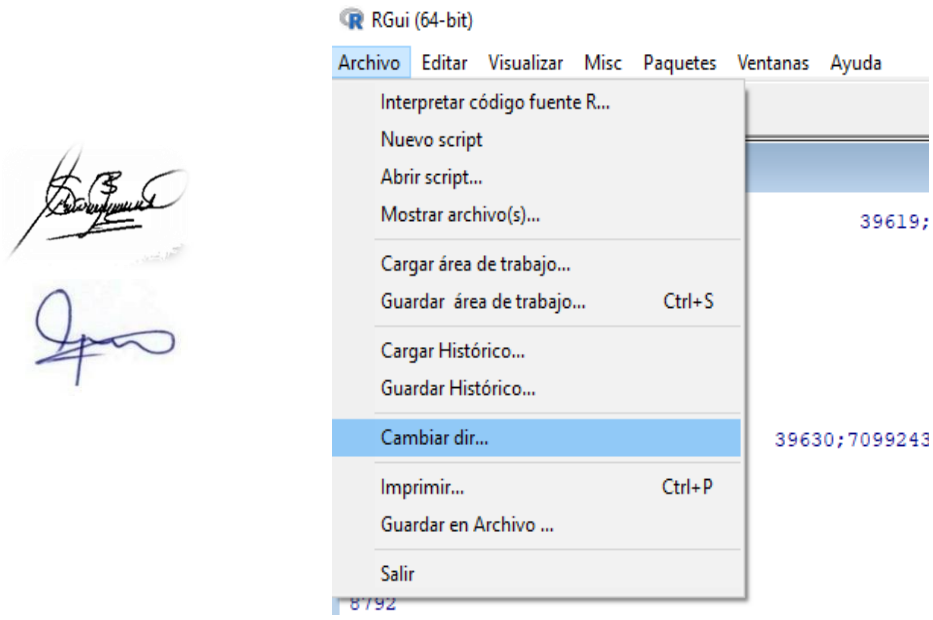
4.2.3 Librerias a utlizar software R

- library(tidyverse) Utilizado para Data Science)
- library(AppliedPredictiveModeling) (Modelos Predictivos)
- library(ggplot) (Para diseño Tablas)
- library(dplyr) (Para trabajar con Data.Frames)

Se debe seleccionar al inicio la ruta de trabajo para el proyecto con la opción ***Cambiar dir…*** ya que si no podría ocurrir errores con los archivos y/o librerías, por defecto se debe configurar la carpeta ***mis documentos*** o en su defecto otra que ya tengamos con los datos.

Figura Nº 27

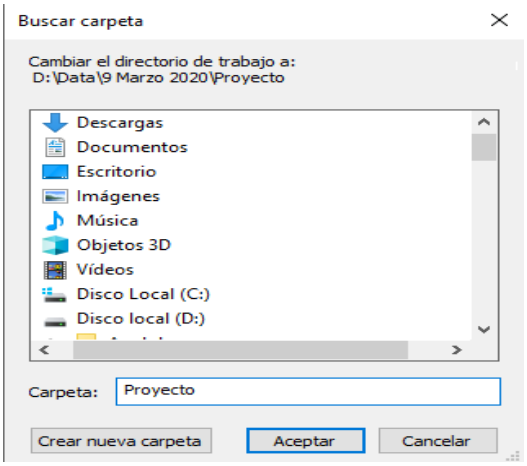
Cambio de Ruta de Trabajo



Fuente: Propia del autor

Figura Nº 28

Selección de Carpeta llamada Proyecto



Fuente: Propia del autor

Para empezar a trabajar tendremos que cargar la información donde se encuentran almacenados los datos a procesar en Software R acepta multitud de formatos entre los mas conocidos tenemos :

- CSV: csv
- EXCEL: .xls y .xlsx
- SPSS: .sav y .por
- STATA: .dta
- SAS: .sas
- R objects: .RData o .rda
- Serialized R objects: .rds
- JSON
- XML
- Bases de Datos , etc

4.2.4 Pre procesamiento de datos software R

Para ello como indicamos anteriormente, disponemos de una base de datos de fugas histórica, de la cual tomaremos el periodo comprendido entre los años 2014 - 2019 del cual tomaremos las fugas más resaltantes y con mayor pérdida para de ahí partir para generar nuestro modelo predictivo.

Como dato adicional solo consideraremos los caudales con valores iguales o mayores a 1000 litros por ser más relevantes para esta investigación, ya que perdidas a menores a 1000 litros no se pueden considerar como pérdidas significativas.

23 fugas fueron registradas en el Callao y las restantes en Lima, se tienen en total 2858 registros de fugas registradas en ese periodo de tiempo con diferentes presiones y diferentes diámetros de tuberías, los cuales se tienen que filtrar por distrito y por caudales para realizar una mejor aproximación del modelo.

Lo primero es hacer el filtrado de la información y centrarnos en un tipo de fuga ya que será más conveniente para el análisis predictivo, procedemos a exportar el archivo a Excel para que sea más fácil el filtrado de la información por tipo y por fecha y por tipo de fuga y caudal y los demás datos que sean importantes para el filtrado.

Como mencionamos las fugas dependen de donde están ubicadas en el tendido de la red de agua potable y dependerá de la presión del agua y del tipo de terreno donde se encuentra la tubería para que sea fácil su escucha.

Toda esa información que le llamaremos BASE DE DATOS DE FUGAS que contiene la información histórica, es la que usaremos para el desarrollo de nuestro algoritmo.

Los tipos de fuga pueden ser:

- CAJA A/MEDIDOR
- CAJA D/MEDIDOR
- LINEA A/CAJA
- LINEA A/CAJA
- CORPORATION
- TUBERIA
- VALVULA DE RED
- VALVULA GRIFO C/L

Lo primero es hacer el filtrado de la información y centrarnos en los campos más relevantes de la tabla para ejecutar análisis predictivo, procedemos a exportar el archivo a Excel para que sea más fácil el filtrado de la información por distrito , por fecha y por tipo de fuga y los demás datos que sean importantes para el filtrado.

La base de datos original fue extraida de una aplicación en entorno SQL los datos pueden ser exportados en Excel , una vez que tenemos los datos

exportados a Excel procederemos a hacer la discriminación de los datos para lo cual escogeremos los datos de las fugas llamadas de emergencia las cuales son incidencias en tuberías y redes primarias de un periodo comprendido entre los años 2012 al año 2019 es un periodo de 7 años y usaremos el año 2020 como dato de control para nuestro algoritmo.

Podríamos también de los datos directamente de Excel enviarlos a software R y desde ahí hacer la discriminación de datos pero para fines didácticos lo mostraremos desde Excel para visualizar los campos y algunos tipos de datos en ellos y su posterior conversión a CSV.

Figura Nº 29

Datos en Excel exportados de la Base de Datos

	A	B	C	D	E	F	G	H	I	J	K
1	Nº DE FUGA	NIS	FECHA	DISTRITO	URBANIZACION	CODIGO SECTOR	TIPO DE FUGA	DIAMETRO TUBERI	PRESION (PSI)	CAUDAL (	VEHICUL
1420	23265	2717633	20/07/2017 15:19	SAN BORJA	URB CHACARILLA DEL ESTANQUE	72	TUBERIA	600	15	150000	4072
1426	24532	2533457	9/09/2017 15:04	MIRAFLORES	URB BARBONCITO	55	TUBERIA	100	14	150000	4072
1430	30958	3045311	9/05/2018 12:04	LIMA (CERCADO)	CERCADO	9	TUBERIA	300	40	150000	4073
1433	33310	3821039	22/08/2018 12:09	RIMAC	P.J VILLA FATIMA	200	TUBERIA	100	55	150000	4073
1434	27408	3003841	21/12/2017 09:44	LIMA (CERCADO)	CERCADO	8	CORP	15	0	120000	4074
1436	7391	3573654	15/02/2015 18:32	COMAS	URB SAN FELIPE 2DA ET.	350	TUBERIA	150	10	100000	4075
1437	8710	4201997	21/05/2015 07:50	EL AGUSTINO	P.J ELAGUSTINO 1 ZONA	2	TUBERIA	110	40	100000	771
1439	14305	3186841	13/04/2016 16:46	LA VICTORIA	URB SANTA CATALINA	18	TUBERIA	100	36	100000	771
1442	22380	3530577	8/06/2017 09:46	SAN MARTIN DE PORRES	URB PERU	205	TUBERIA	250	6	100000	4071
1443	22600	2565919	19/06/2017 10:40	CHORRILLOS	URB CHORRILLOS	75	TUBERIA	150	20	100000	771
1447	22879	6059692	1/07/2017 14:24	CARABAYLLO	A.H SAN GRABIEL	357	TUBERIA	60	60	100000	4072
1451	23163	3633227	15/07/2017 11:01	PUENTE PIEDRA	CERCADO	386	TUBERIA	200	42	100000	4071
1461	27912	2596083	13/01/2018 15:24	SANTIAGO DE SURCO	URB DERBY, EL	297	TUBERIA	100	23	100000	4072
1462	30144	2742059	10/04/2018 09:27	SAN ISIDRO	URB CHACARILLA DE SANTA CRUZ	52	TUBERIA	150	26	100000	4073
1463	32894	3726145	10/08/2018 08:27	SAN MARTIN DE PORRES	URB PALAO 1RA ET.	206	TUBERIA	200	38	100000	4074
1470	32966	5880178	10/08/2018 15:49	CARABAYLLO	A.H VILLA ESPERANZA AMPL. DEL	357	TUBERIA	100	80	100000	4074
1472	34473	3800631	16/10/2018 09:40	RIMAC	A.H MUNICIPAL 3	203	CORP	15	10	100000	4074
1512	21526	2611609	29/04/2017 13:45	SAN BORJA	URB SAN BORJA	71	TUBERIA	100	20	90000	4072
1524	25618	6433321	21/10/2017 13:47	PUENTE PIEDRA	A.H TIWINZA	389	TUBERIA	100	50	85000	836
1547	2815	5008163	9/07/2014 15:54	PUCUSANA	ASOC ISLA GALAPAGOS	541	TUBERIA	75	40	80000	
1548	2960	2801997	6/08/2014 14:02	VILLA MARIA DEL TRIUNFO	P.J GALVEZ, POETA JOSE	315	CORP	15	20	80000	4071
1553	9514	2543813	7/07/2015 07:20	BARRANCO	CERCADO	74	TUBERIA	350	28	80000	771
1555	13333	2534961	15/02/2016 09:11	SAN ISIDRO	URB SAN GABRIEL	50	TUBERIA	100	20	80000	4071
1579	13742	3204606	10/03/2016 13:04	PUEBLO LIBRE	CERC CERCADO	30	TUBERIA	200	20	80000	771
1580	15276	5115567	9/06/2016 06:45	SAN MARTIN DE PORRES	A.H SEÑOR DE LOS MILAGROS	206	TUBERIA	90	65	80000	4073
1581	17888	6797976	11/10/2016 16:55	SAN MIGUEL	URB MARANGA 1RA ET.	45	TUBERIA	100	6	80000	771
1598	18719	2901250	18/11/2016 14:07	VILLA MARIA DEL TRIUNFO	P.J GALVEZ, POETA JOSE	315	TUBERIA	100	40	80000	4074
1602	25738	5288420	27/10/2017 09:26	SAN MARTIN DE PORRES	A.H NUEVO AMANECER	206	TUBERIA	160	47	80000	4072
1608	27644	5087320	4/01/2018 10:10	EL AGUSTINO	A.H TERRAZAS DE CATALINA HUAN	1	CORP	15	40	80000	4073
1612	27962	5767593	16/01/2018 08:28	COMAS	A.H PRIMAVERA Y MANCO INCA 11	339	TUBERIA	90	40	80000	4072
1625	29001	3719844	24/02/2018 12:25	COMAS	A.H COLLIQUE	349	TUBERIA	75	40	80000	4074
1627	29052	3150544	27/02/2018 10:46	MAGDALENA	URB PRIMAVERA	48	CORP	15	30	80000	836
1635	29419	5155776	13/03/2018 13:51	LOS OLIVOS	P.J JULIO, 28 DE	80	TUBERIA	75	9	80000	841
1643	29979	2590977	3/04/2018 10:20	SAN ISIDRO	URB CORPAC	53	TUBERIA	100	28	80000	4073
1654	30206	3590244	12/04/2018 10:03	LOS OLIVOS	URB PARQUE EL NARANJAL 1RA E	83	LINEA A/CAJA	15	17	80000	4072
1673	31474	5370507	1/06/2018 09:34	CARABAYLLO	A.H FLOR II ET. SECT. I Y II, LA	357	TUBERIA	110	100	80000	4073
1675	31735	3813933	14/06/2018 11:16	RIMAC	A.H ZEVALLOS, HORACIO	203	CORP	15	30	80000	836
1676	33958	5185092	19/09/2018 10:58	LA MOLINA	URB CAPILLA, LA	199	CORP	15	15	80000	4074
1680	34750	2888821	26/10/2018 09:44	SAN BORJA	A.H SAN JUAN MASIAS	20	TUBERIA	250	30	80000	4072
1695	35968	3130907	17/12/2018 17:59	PUEBLO LIBRE	CERC CERCADO	30	TUBERIA	100	37	80000	4072

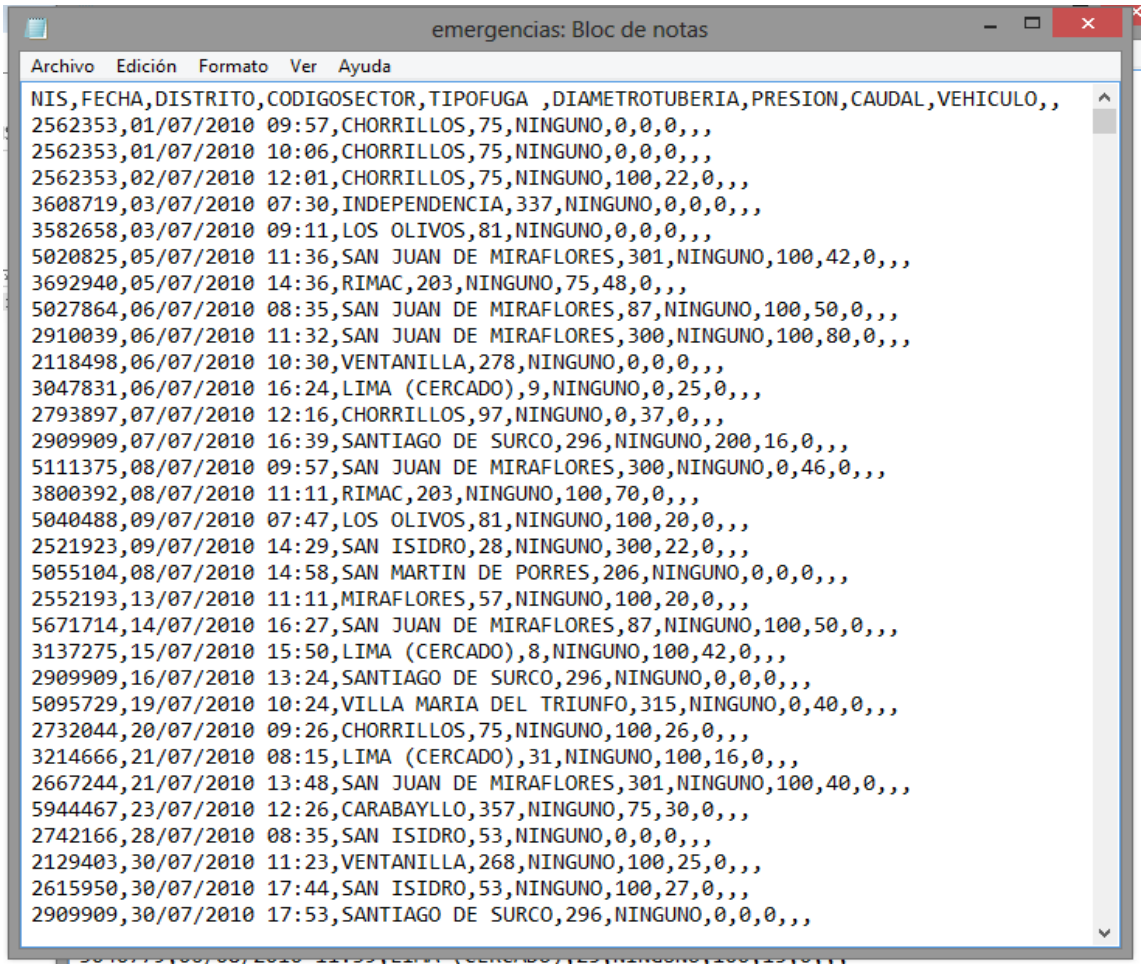
Fuente: Propia del autor

La Base de datos cuenta con 67 columnas y 8226 filas de los cuales varios son vacíos y/o duplicados los cuales procederemos a eliminar y quedarnos con solo los campos más importantes.

Una vez terminado exportamos el archivo a formato CSV separado por comas quedando de la siguiente manera.

Figura Nº 30

Datos de Excel exportados a CSV



Fuente: Propia del autor

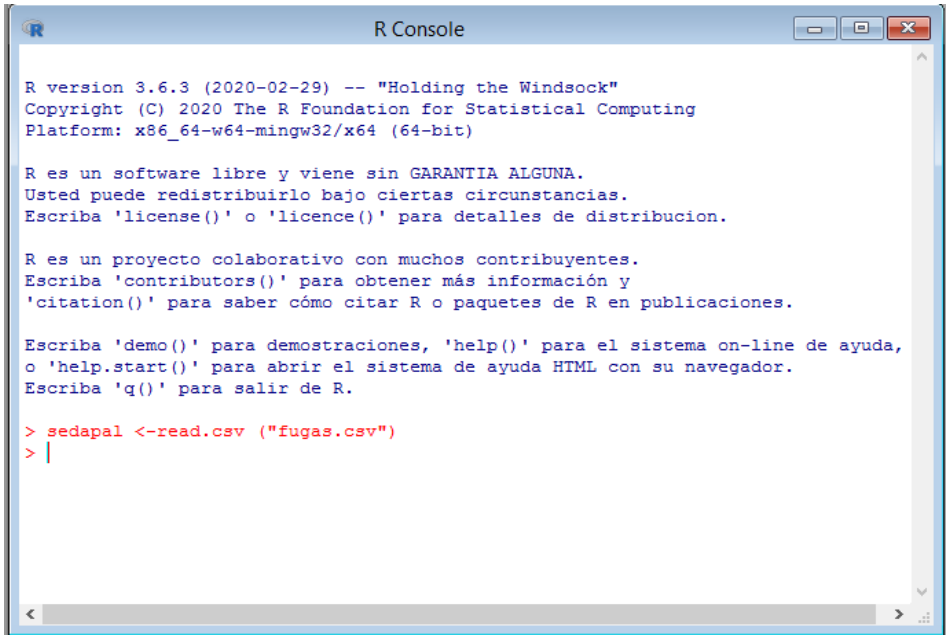
Luego para cargar el archivo ejecutamos el comando ***read.csv*** **sedapal <-read.csv ("fugas.csv")**

El formato se escribe así porque primero le asignamos un nombre al archivo fugas.csv para que sea cargado en memoria en un arreglo de datos, en este caso lo llamaremos sedapal y le damos **enter** y no debe aparecer ningún mensaje de error.

Figura Nº 31

Cargar archivo fugas.csv en Memoria

+

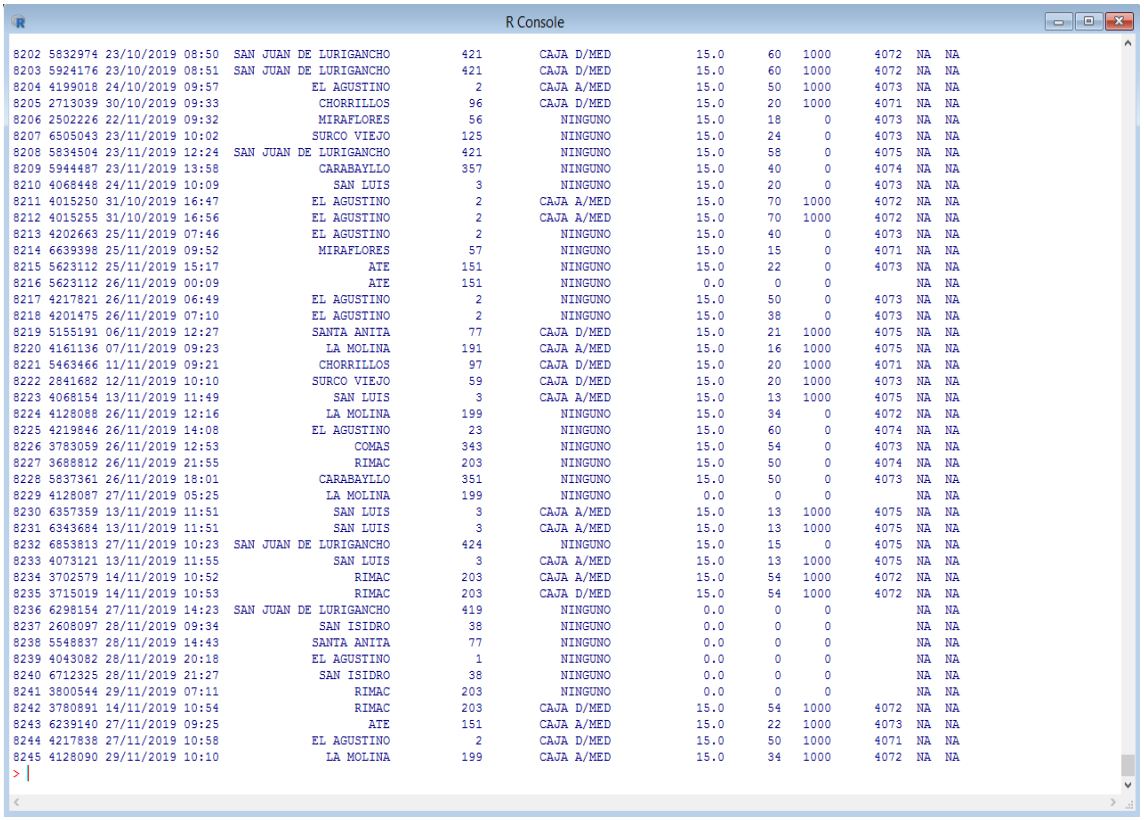


Fuente: Propia del autor

Finalmente escribimos el nombre con el cual hemos designado a nuestro archivo en este caso **sedapal** y le damos **enter**, dando el sgt. resultado.

Figura Nº 32

Archivo fugas.csv cargado en R



Fuente: Propia del autor

Otra forma de llamar al archivo es con el comando ***read_csv*** pero necesitamos de la librería ***library(tidyverse)*** lo cual no muestra de otra forma los datos en pantalla, con información sobre los campos y el tipo de dato.

library(tidyverse)

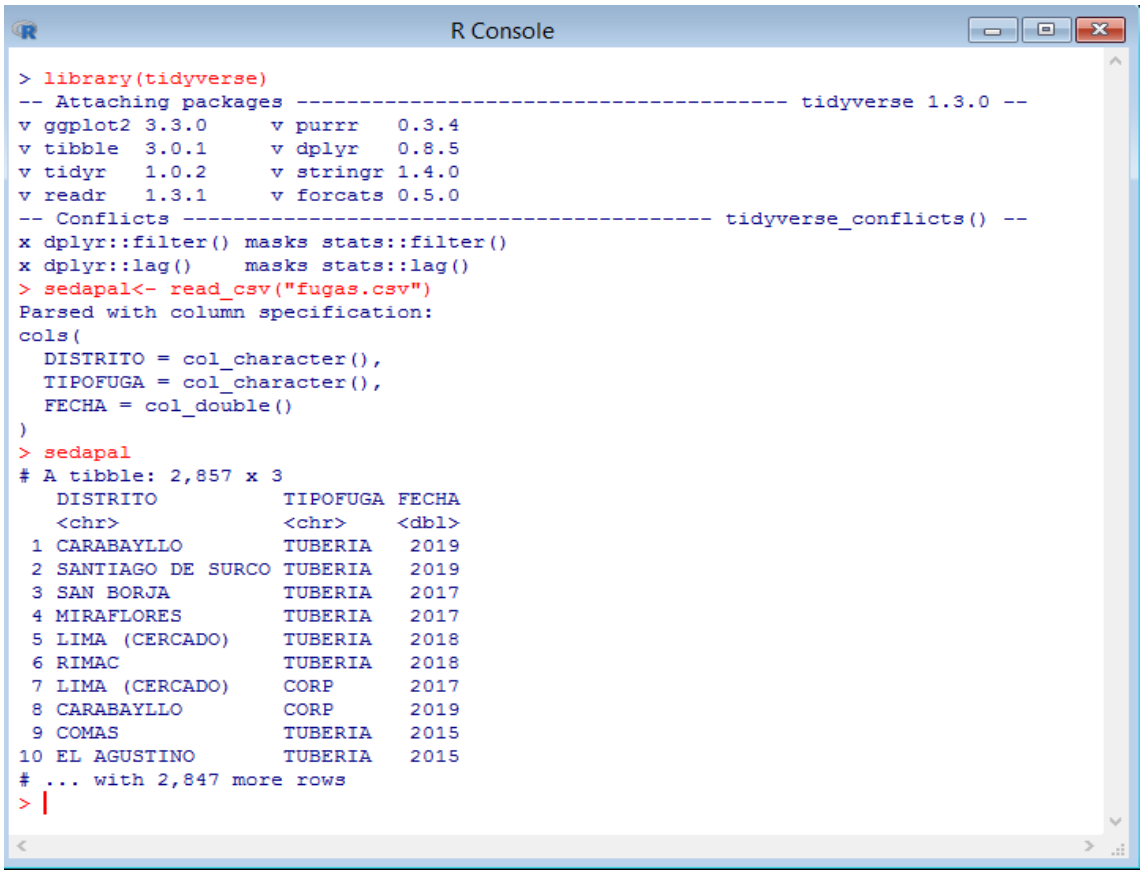
sedapal<- read_csv("fugas.csv")

Esto es muy útil para conocer cómo hacer el tratamiento de la información y su posterior análisis en detalle.

Finalmente escribimos el nombre con el cual hemos designado a nuestro archivo en este caso **sedapal** y le damos **enter**, dando el sgt resultado.

Figura Nº 33

Archivo CSV usando la librería tidyverse



Fuente: Propia del autor

4.2.5 Uso del comando SPLIT

Supóngase ahora que se desea tener la misma información, pero clasificada por años. Con este propósito se puede emplear la función **split(),** que recibe básicamente dos argumentos principales: el objeto a ser clasificado -típicamente,

un data frame o un vector-, y una serie de datos que servirán para la clasificación -típicamente, un factor o un vector numérico o de cadenas de caracteres-. Este último argumento debe tener la misma longitud que una columna del data frame, dado como primer argumento, o que el vector dado como argumento. Según el caso, y no es necesario que este argumento sea parte del objeto dado como primer argumento. Si, como primer argumento de la función

a) Uso del comando split por vehiculo

Supóngase ahora que se desea tener la misma información, pero clasificada por vehículos. Con este propósito se puede emplear la función split(), que recibe básicamente dos argumentos principales: el objeto a ser clasificado -típicamente, un data frame o un vector-, y una serie de datos que servirán para la clasificación -típicamente, un factor o un vector numérico o de cadenas de caracteres-. Este último argumento debe tener la misma longitud que una columna del data frame, dado como primer argumento, o que el vector dado como argumento. Según el caso, y no es necesario que este argumento sea parte del objeto dado como primer argumento. Si, como primer argumento de la función pasamos una porción de la tabla contenida en el data frame justamente leído -sólo las columnas 3 a la 5 del data frame-, y como segundo argumento pasamos la columna correspondiente a los VEHICULOS, la función operaría como se muestra y en este caso el resultado de la función, en el caso del ejemplo, es una lista de tablas, cada una correspondiente a cada uno de los años registrados en la columna sedapal$VEHICULO, como se muestra a continuación:

sedapal_vehiculo <- split(sedapal[, 1:3], sedapal$DISTRITO)

sedapal_vehiculo

al ejecutar el comando podemos clasificar los datos en este caso por vehiculo :

$`4073`

DISTRITO	.SECTOR	TIPOFUGA
1604 SAN MARTIN DE PORRES	208	CORP
2299 SAN MARTIN DE PORRES	205	NINGUNO

$`4074`

DISTRITO	SECTOR	TIPOFUGA
1401 SANTIAGO DE SURCO	123	TUBERIA

b) Uso del comando split por codigo del sector

Supóngase ahora que se desea tener la misma información, pero clasificada por CODIGO.SECTOR. Con este propósito se puede emplear la función split(), que recibe básicamente dos argumentos principales: el objeto a ser clasificado -típicamente, un data frame o un vector-, y una serie de datos que servirán para la clasificación -típicamente, un factor o un vector numérico o de cadenas de caracteres-. Este último argumento debe tener la misma longitud que una columna del data frame, dado como primer argumento, o que el vector dado como argumento. según el caso, y no es necesario que este argumento sea parte del objeto dado como primer argumento. Si, como primer argumento de la función pasamos una porción de la tabla contenida en el data frame justamente leído -sólo las columnas 1 a la 8 del data frame-, y como segundo argumento

pasamos la columna correspondiente a los CÓDIGO DE SECTOR , la función operaría como se muestra y en este caso el resultado de la función, en el caso del ejemplo, es una lista de tablas, cada una correspondiente a cada uno de los años registrados en la columna sedapal$CODIGO.SECTOR, como se muestra a continuación:

sedapal_sector <- split(sedapal[, 1:3], sedapal$CODIGO.SECTOR)

sedapal_sector

$`467`

	FECHA	DISTRITO	TIPOFUGA	DIAMETRO	PRESION	CAUDAL
1400	21/03/2014 15:04	LURIN	NINGUNO	0	40	0
1422	4/04/2014 10:58	LURIN	NINGUNO	0	0	0

$`501`

	FECHA	DISTRITO	TIPOFUGA	DIAMETRO	PRESION	CAUDAL
521	9/04/2012 11:41	PUNTA HERMOSA	NINGUNO	0	0	0
532	16/04/2012 11:47	PUNTA HERMOSA	NINGUNO	50	15	0

c) Uso del comando split por caudal

Supóngase ahora que se desea tener la misma información, pero clasificada por CAUDAL. Con este propósito se puede emplear la función split(), que recibe básicamente dos argumentos principales: el objeto a ser clasificado -típicamente, un data frame o un vector-, y una serie de datos que servirán para la clasificación típicamente, un factor o un vector numérico o de cadenas de caracteres-. Este

último argumento debe tener la misma longitud que una columna del data frame, dado como primer argumento, o que el vector dado como argumento. según el caso, y no es necesario que este argumento sea parte del objeto dado como primer argumento. Si, como primer argumento de la función pasamos una porción de la tabla contenida en el data frame justamente leído -sólo las columnas 1 a la 8 del data frame-, y como segundo argumento pasamos la columna correspondiente a los CAUDAL , la función operaría como se muestra y en este caso el resultado de la función, en el caso del ejemplo, es una lista de tablas, cada una correspondiente a cada uno de los años registrados en la columna sedapal$CAUDAL, como se muestra a continuación:

sedapal_caudal <- split(sedapal[, 1:3], sedapal$CAUDAL)

sedapal_caudal

`150000`

NIS FECHA DISTRITO CODIGO.SECTOR TIPOFUGA DIAMETROTUBERIA PRESION CAUDAL

1419 2717633 20/07/2017 15:19 SAN BORJA 72 TUBERIA 600 15 150000

1425 2533457 9/09/2017 15:04 MIRAFLORES 55 TUBERIA 100 14 150000

$`180000`

NIS FECHA DISTRITO CODIGO.SECTOR TIPOFUGA DIAMETROTUBERIA PRESION CAUDAL

1401 5968591 17/05/2019 10:56 SANTIAGO DE SURCO 123 TUBERIA 100 8 180000

$`200000`

NIS FECHA DISTRITO CODIGO.SECTOR TIPOFUGA DIAMETROTUBERIA PRESION CAUDAL

1399 5183985 11/10/2019 11:10 CARABAYLLO 357 TUBERIA 100 20 200000

El comando SPLIT es uno de los comandos mas útiles del R por que nos permite ver en forma ordenada lo cual es útil para agrupar los datos y hacer más fácilmente el análisis respectivo de los datos.

d) Uso del comando tapply

Finalmente, en este tipo de funciones está la función tapply(), que efectúa el mismo tipo de operaciones de clasificación y operación, pero actúa sobre un vector y no sobre un data frame. Cada una de las columnas del data frame leído, es o bien un vector numérico, o un factor. Debido a la organización de la tabla anterior, para obtener mismo el resultado que las funciones usadas anteriormente, by() y aggregate(), la función tapply() tiene que sacar provecho de que el segundo argumento, que sirve para clasificar los datos, puede ser una lista de uno o mas objetos, interpretables como factores, y que en el caso del ejemplo estará constituida por las columnas correspondiente a los TIPOSDEFUGA y a los CAUDALES. Así, el resultado deseado se obtiene como sigue:

sedapal <- read.csv("FUGAS.csv")

sedapal

> (rr <- tapply(sedapal$DISTRITO, list(sedapal$TIPOFUGA, sedapal$CAUDAL), mean, na.rm = T))

Así, el resultado deseado se obtiene como sigue:

0 15 32 50 100 200 500 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000

12000 15000 16000

CAJA A/MED 140.6667 NA NA 275.2 2 NA 88.3 132.39394 150.10323 108.83929 194.7500 187.5185 343.0000 165.4000 92.5000 NA NA NA 203.0000 NA

CAJA D/MED NA NA NA NA NA 188.5 79.2 115.42647 137.66364 149.90244 186.6364 181.0625 NA 177.5000 332.0000 NA 178.0000 NA NA NA

CORP 132.0000 NA NA NA NA NA NA 145.00000 132.00000 155.00000 128.5556 148.5338 159.4348 165.3448 148.3786 203 153.5420 138 142.6364 18

LINEA A/CAJA 2.0000 NA NA 203.0 NA NA 9.5 147.38462 174.17647 98.43333 139.3750 161.8431 103.0000 138.4615 127.9333 NA 139.3750 NA 120.5000 NA

LINEA D/CAJA NA NA NA NA NA NA NA 98.18182 137.32353 145.20000 271.7778 138.2692 272.5000 188.2500 243.8571 NA 167.0000 NA NA NA

MEDIDOR NA NA NA NA NA NA NA NA 90.00000 NA NA 90.0000 NA NA NA NA NA NA NA NA

NINGUNO 173.4030 27 NA NA NA NA NA NA 228.00000 203.00000 NA NA NA 2.0000 NA NA 223.5000 NA NA NA

Se están evaluando y comparando varios parámetros a la vez utilizando el comando TAPPLY para obtener los datos agrupados de una mejor manera en este caso con respecto al SECTOR junto con los valores de TIPODE FUGA y CAUDAL respectivamente.

Como se puede ver por lo expuesto, mucho del trabajo que otros lenguajes resuelven con ciclos, decisiones y secuencias de instrucciones, en R, alternativamente, se puede hacer por medio de este elegante conjunto de funciones de clasificación, transformación y agregación de datos, lo que junto con las operaciones para manipular porciones de los datos estructurados, hacen del lenguaje una poderosa herramienta para el procesamiento de información.

Hemos visto que R es versátil para el manejo y procesamiento de la información y la forma de representarla, nuestra base de datos de las fugas obtenidas a lo largo de varios años, va ser exportada y procesada filtrando los datos más relevantes para con ellos hacer el diseño del algoritmo. Nos interesan 3 campos los más importantes para poder parametrizar mejor los datos y estos son:

- Distrito
- Tipo de Fuga
- Fecha

Una vez que importamos datos a R es conveniente limpiarlos, esto implica almacenarlos de una manera consistente que nos permita enfocarnos en responder preguntas de los datos en lugar de estar luchando con los datos. Entonces, datos limpios son datos que facilitan las tareas del análisis de datos:

Manipulación: Manipulación de variables como agregar, filtrar, reordenar, transformar.

Visualización: Resúmenes de datos usando gráficas, análisis exploratorio, o presentación de resultados.

Modelación: Ajustar modelos es sencillo si los datos están en la forma correcta.

Los principios de datos limpios proveen una manera estándar de organizar la información:

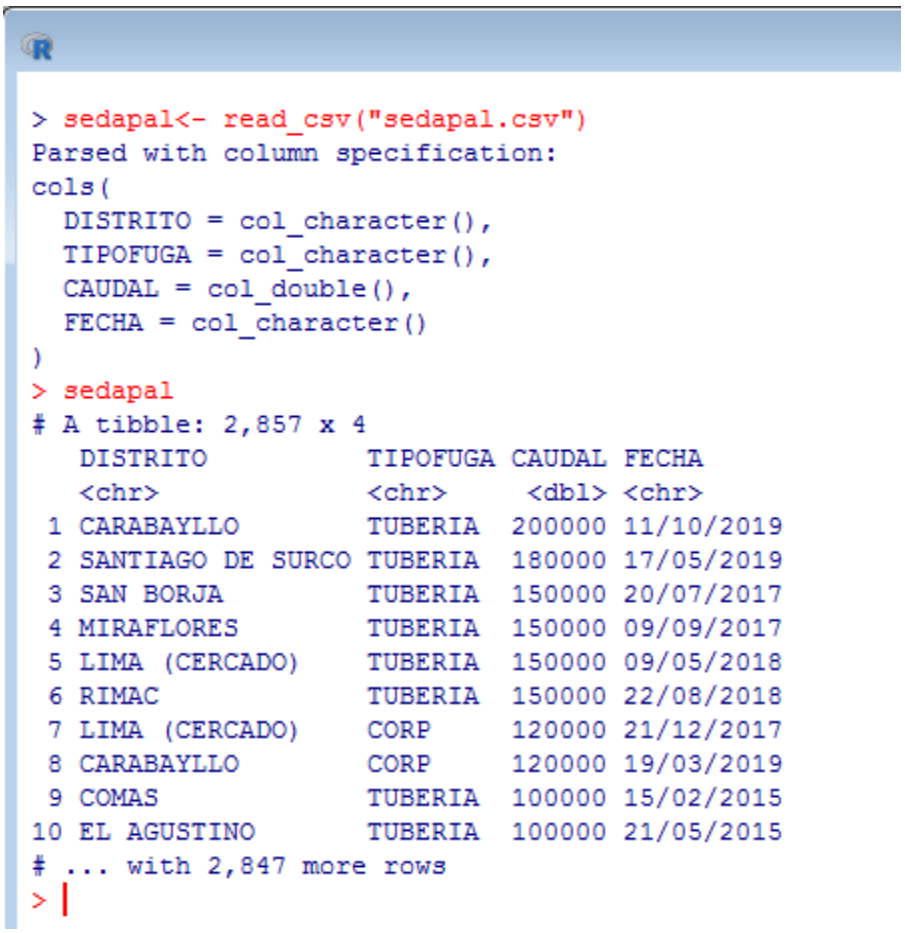
- Cada variable forma una columna.
- Cada observación forma un renglón.
- Cada tipo de unidad observacional forma una tabla.

Vale la pena notar que los principios de los datos limpios se pueden ver como teoría de algebra relacional para estadísticos, estos principios equivalen a la tercera forma normal de Codd con enfoque en una sola tabla de datos en lugar de muchas conectadas en bases de datos relacionales.

Una vez que importamos datos a R es conveniente limpiarlos, y como vemos el dato FECHA sería más conveniente que solo estuviera como año (2019) y no como formato con día y mes, ejemplo el día (11/10/2019).

Figura Nº 34

TABLA CON DATOS FECHA



```
> sedapal<- read_csv("sedapal.csv")
Parsed with column specification:
cols(
  DISTRITO = col_character(),
  TIPOFUGA = col_character(),
  CAUDAL = col_double(),
  FECHA = col_character()
)
> sedapal
# A tibble: 2,857 x 4
   DISTRITO          TIPOFUGA CAUDAL FECHA
   <chr>             <chr>     <dbl> <chr>
 1 CARABAYLLO        TUBERIA  200000 11/10/2019
 2 SANTIAGO DE SURCO TUBERIA  180000 17/05/2019
 3 SAN BORJA         TUBERIA  150000 20/07/2017
 4 MIRAFLORES        TUBERIA  150000 09/09/2017
 5 LIMA (CERCADO)    TUBERIA  150000 09/05/2018
 6 RIMAC             TUBERIA  150000 22/08/2018
 7 LIMA (CERCADO)    CORP     120000 21/12/2017
 8 CARABAYLLO        CORP     120000 19/03/2019
 9 COMAS             TUBERIA  100000 15/02/2015
10 EL AGUSTINO       TUBERIA  100000 21/05/2015
# ... with 2,847 more rows
>
```

Fuente: Propia del autor

```
sedapal <- fugas %>%
  mutate(
    FECHA = parse_number(week),
    date = date.entered * (week - 1),
    rank = as.numeric(rank)
    ) %>%
    select(-date.entered)
```

Figura Nº 35

TABLA CON DATOS FECHA LIMPIOS

```
> sedapal<- read_csv("fugas.csv")
Parsed with column specification:
cols(
  DISTRITO = col_character(),
  TIPOFUGA = col_character(),
  CAUDAL = col_double(),
  FECHA = col_double()
)
> sedapal
# A tibble: 2,857 x 4
   DISTRITO          TIPOFUGA CAUDAL FECHA
   <chr>             <chr>     <dbl> <dbl>
 1 CARABAYLLO        TUBERIA  200000  2019
 2 SANTIAGO DE SURCO TUBERIA  180000  2019
 3 SAN BORJA         TUBERIA  150000  2017
 4 MIRAFLORES        TUBERIA  150000  2017
 5 LIMA (CERCADO)    TUBERIA  150000  2018
 6 RIMAC             TUBERIA  150000  2018
 7 LIMA (CERCADO)    CORP     120000  2017
 8 CARABAYLLO        CORP     120000  2019
 9 COMAS             TUBERIA  100000  2015
10 EL AGUSTINO       TUBERIA  100000  2015
# ... with 2,847 more rows
>
```

Fuente: Propia del autor

e) Conversion de un archivo csv a data.frame

Un ***data.frame*** es una lista, cuyos componentes pueden ser vectores, matrices o factores, con la única salvedad de que las longitudes, o número

de renglones, en el caso de matrices, deben coincidir en todos los componentes.

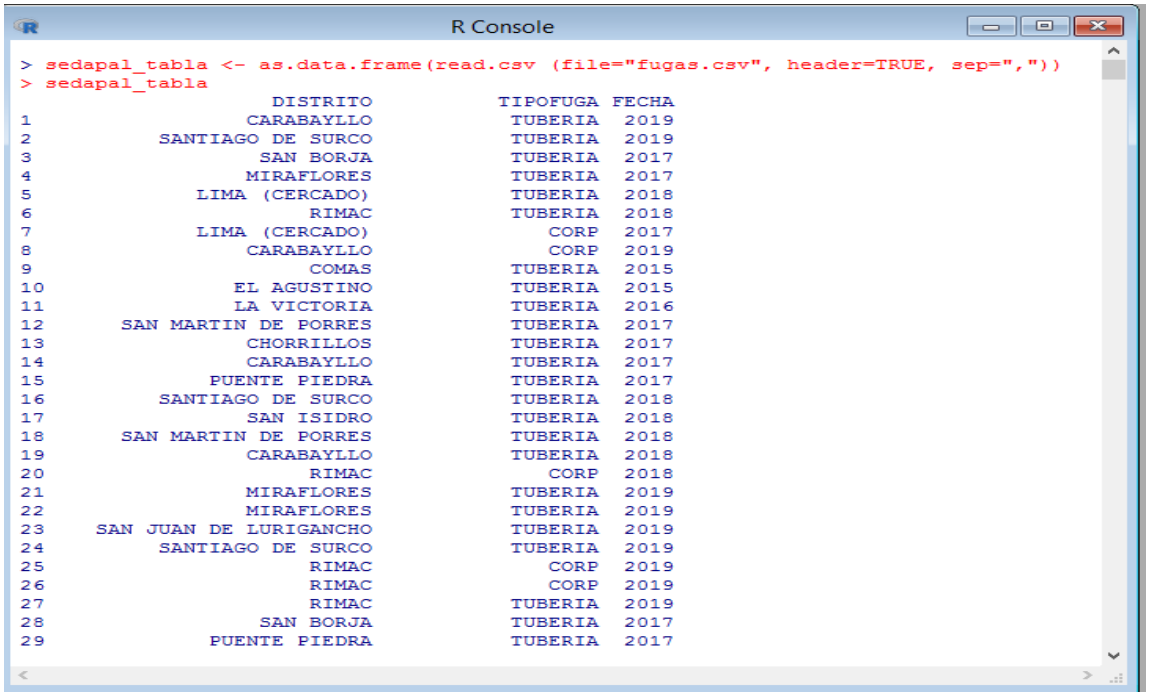
La apariencia de un ***data.frame*** es la de una tabla y una forma de crearlos es mediante la función data.frame() .

Para nuestro caso puntual es recomendable trabajar los datos como ***data.frame*** para poder ordenarlos y averiguar los niveles y demás información.

sedapal_tabla <- as.data.frame(read.csv (file="fugas.csv", header=TRUE, sep=","))

Figura Nº 36

TABLA COMO DATA.FRAME



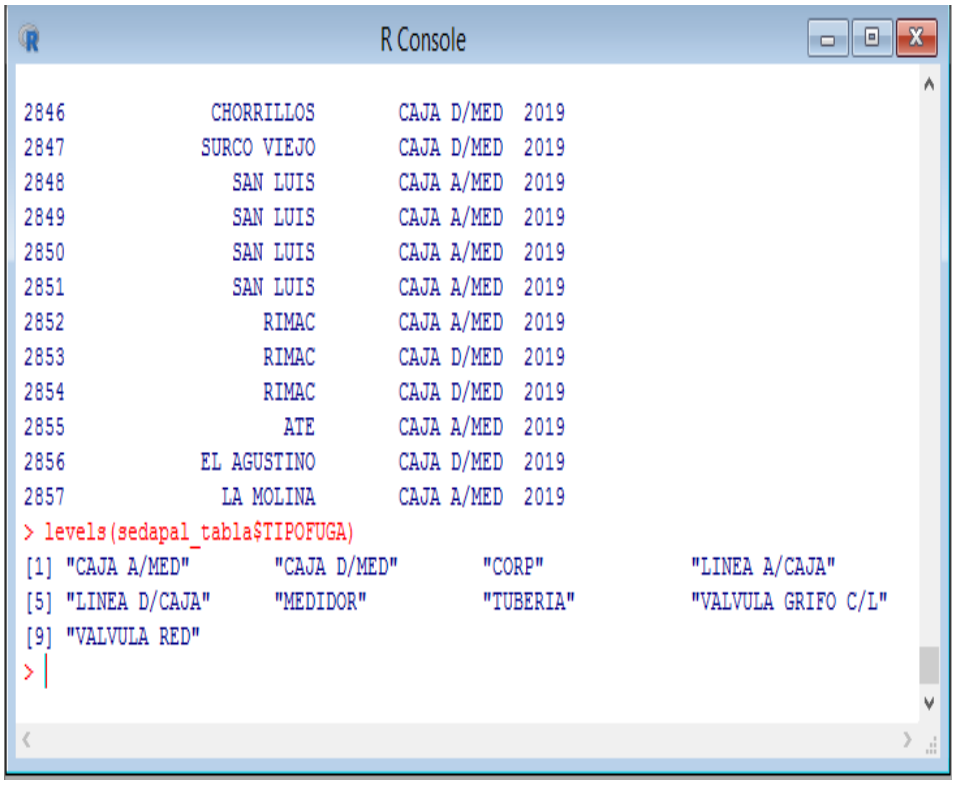
Fuente: Propia del autor

Ya tenemos nuestro CSV como un ***data.frame*** y podemos observar lo que son los niveles (valores) de un campo ***CHAR*** como TIPOFUGA.

levels(sedapal_tabla$TIPOFUGA)

Figura Nº 37

LEVELS TIPOFUGA



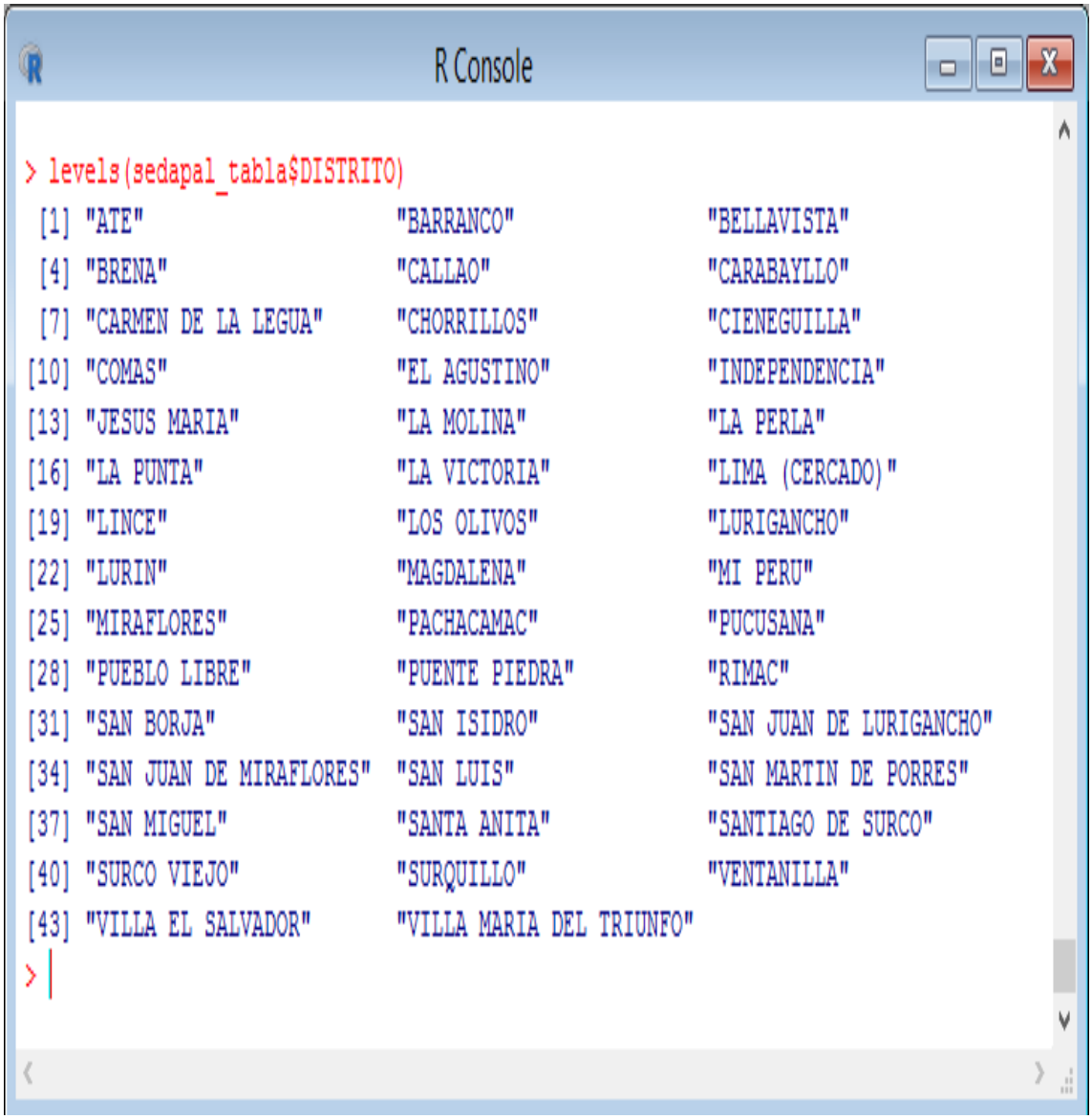
Fuente: Propia del autor

Por ejemplo para el campo DISTRITO

levels(sedapal_tabla$DISTRITO)

Figura Nº 38

LEVELS TIPOFUGA



Fuente: Propia del autor

f) Tratamiento de los datos usando GGPLOT

MATRIZ DE GRÁFICOS DE DISPERSIÓN

Una matriz de gráficos de dispersión es una cuadrícula (o matriz) de gráficos de dispersión que se utiliza para visualizar relaciones bivariante entre combinaciones de variables. Cada gráfico de dispersión de la matriz muestra la relación entre un par de variables, lo que permite explorar muchas relaciones en un solo gráfico.

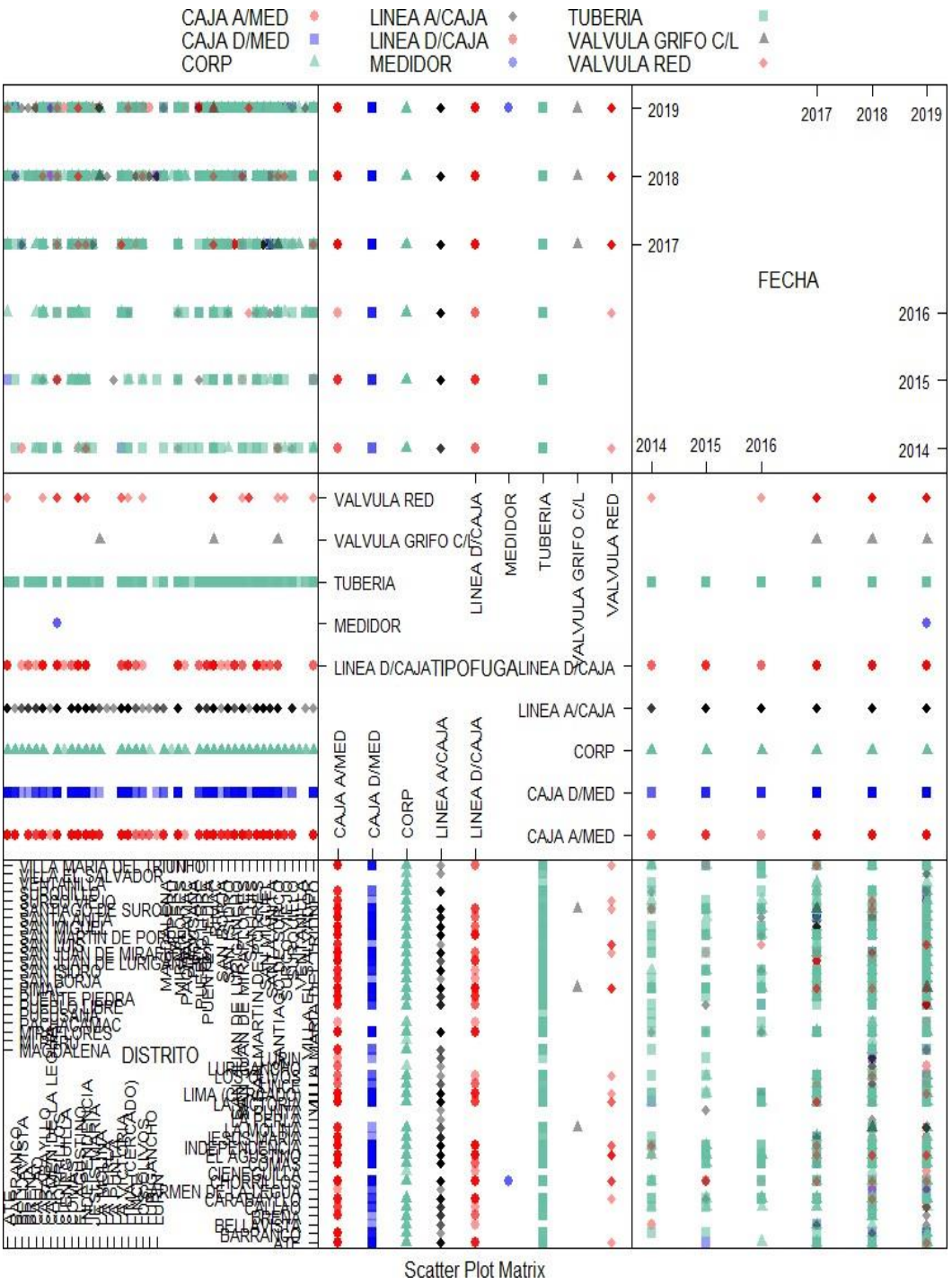
Una matriz de gráficos de dispersión se compone de una cuadrícula de pequeños gráficos y un gráfico de vista previa más grande que muestra un gráfico pequeño seleccionado con mayor detalle. Además, se puede agregar a la matriz un histograma que muestra la distribución de cada variable numérica.

Ingresamos el código en R para visualizar la matriz digitando el siguiente algoritmo en la ventana de comandos.

```
library(AppliedPredictiveModeling)

transparentTheme(trans = .4)

library(caret)

featurePlot(x =  sedapal_tabla[, 1:3],

        y = sedapal_tabla$TIPOFUGA,

        plot = "pairs",

        ## Add a key at the top

        auto.key = list(columns = 3))
```

Figura Nº 39

MATRIZ DE DISPERSION DE NUESTRA TABLA



Fuente: Propia del autor

Como se puede apreciar tenemos ya un primer acercamiento a una tendencia utilizando la matriz de dispersión entre 3 variables como son TIPOFUGA, DISTRITO y FECHA ya que se puede apreciar una concentración de datos para el caso de TIPOFUGA en 2 de sus variables TUBERIA y CORP notándose un incremento de estas en el 2019, como vemos el software R permite agrupar en este caso datos CUALITATIVOS y NUMERICOS y mostrarnos la tendencia de un modelo predictivo a partir de las gráficas obtenidas.

Nuestro modelo predictivo debe indicarnos de la cantidad de datos dispersos y sin relación varios parámetros a saber tales como:

- Distritos con más incidencia de emergencias reportadas
- Tipos de Fugas más incidencia de emergencias reportadas
- Variación de los datos en función de una fecha especifica
- Relacionar estas variables en un solo cuadro
- Relacionar las variables entre ellas

Para ellos haremos uso de la información con la que contamos para hallar una relación y un modelo predictivo, teniendo en cuenta que estos datos no son necesariamente concluyentes y que a mayor cantidad de datos se obtienen modelos y tendencias más precisos.

Luego estos datos procederemos a compararlos con los datos de control que no hemos procesado del periodo 2020 para ver su eficacia.

Cabe aclarar que basamos nuestra investigación en datos reales obtenidos en un periodo de 7 años y con algunas variaciones en la tomas de muestras debido a que se utilizaron cada vez más personal para ello (vehículos) y que esto en el futuro puede seguir aumentando.

Como se indicó, el servicio inicio con 4 vehículos y los primeros registros no son datos completamente homogéneos y luego entre los años 2017, 2018 y 2019 ya con 6 vehículos los datos se hicieron más fiables y aumentaron en numero la cantidad de muestras, en una forma dramática como se aprecia en los resultados.

En los últimos años estamos viviendo una gran explosión de aplicaciones y servicios que giran alrededor del big data y los sistemas predictivos. Estamos rodeados de sistemas que, apoyándose en algoritmos, son capaces de procesar grandes volúmenes de información, hasta el punto de realizar predicciones o, directamente, indicarnos qué tenemos que hacer ante determinada situación.

Lo que se intenta con esta investigación es a partir de los datos obtenidos profundizar en modelos predictivos aplicados a no solo este caso en particular sino a muchos otros, donde se tengan información la cual se requiere analizar y encontrar un patrón que permita su modelamiento matemático.

Figura Nº 40

GRAFICA DEL MODELO PREDICTIVO

Fuente: Propia del autor

Tenemos una gran cantidad de datos por analizar desde diversos frentes ya que contamos con 3 variables 2 de ellas son datos cualitativos y uno cuantitativo, es por ello que se le da un enfoque tratando de encontrar la relación que mejor defina la tendencia de la información con la que contamos.

Nuestro algoritmo se aplica al tratamiento de la información y mostrarla en forma gráfica para su mejor comprensión y predicción de la tendencia, haciendo uso de las herramientas y librerías que el software R nos brinda.

PROCESAMIENTO DE LA INFORMACION MEDIANTE ALGORITMOS

- **RELACION ENTRE VARIABLES DISTRITO Y FECHA**

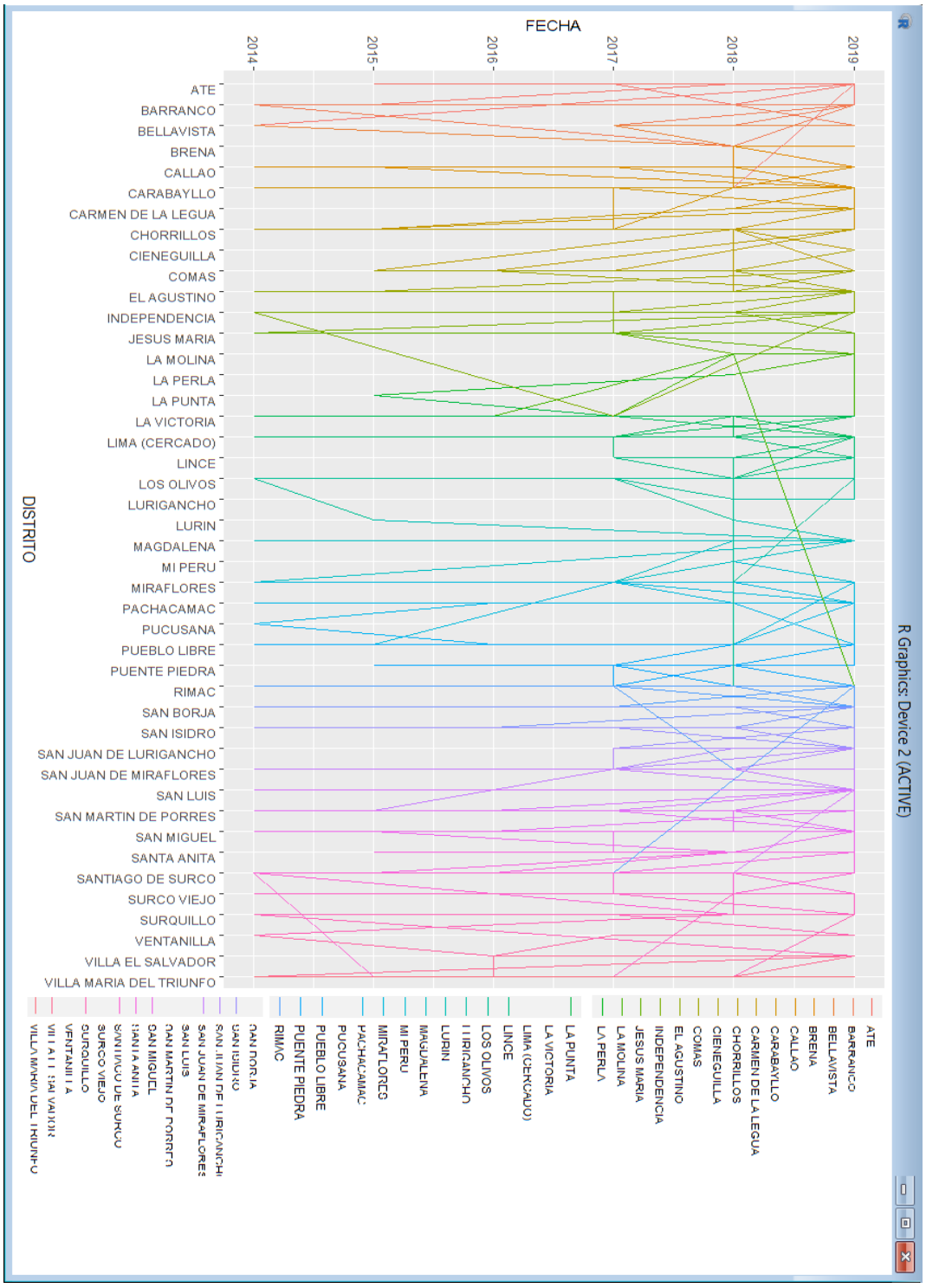
```
ggplot(sedapal, aes(x = DISTRITO, y = FECHA, color =
DISTRITO, group = TIPOFUGA)) + geom_line() +
  theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido no aporta mucha información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo polilínea el cual no es por lo visto el más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar .

Como observamos tenemos en el eje **Y** los años desde el 2014 al 2019 y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

Figura Nº 41

DATOS OBTENIDOS MODELO DEL TIPO POLILINEAL



Fuente: Propia del autor

DATOS OBTENIDOS MODELO DEL TIPO POLILINEAL

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(data=sedapal_tabla, aes(x=reorder
(FECHA,DISTRITO), y=1, fill=DISTRITO)) +
   geom_bar(stat="identity", position="stack")
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

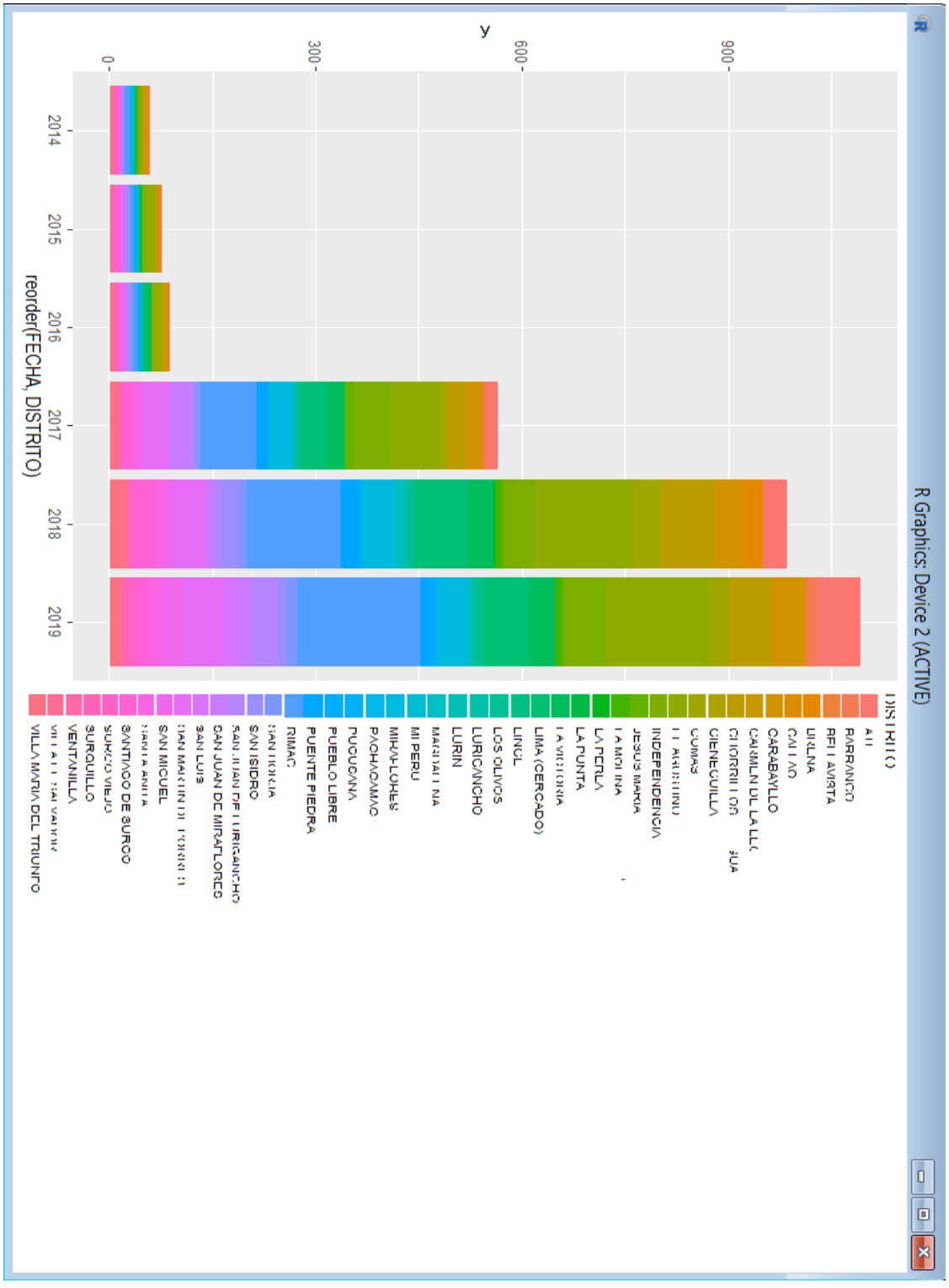
Como apreciamos tenemos aumento en las fugas en los años desde el 2017 al 2019 y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe modelar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Hacemos uso del comando **reorder** para ordenar los datos y también estamos utilizando colores para diferenciar los distritos, sin embargo no logramos aun información concluyente.

Figura Nº 42

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(sedapal, aes(x = DISTRITO, y = 1, group =
DISTRITO)) +
facet_wrap(~ FECHA, nrow = 1) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

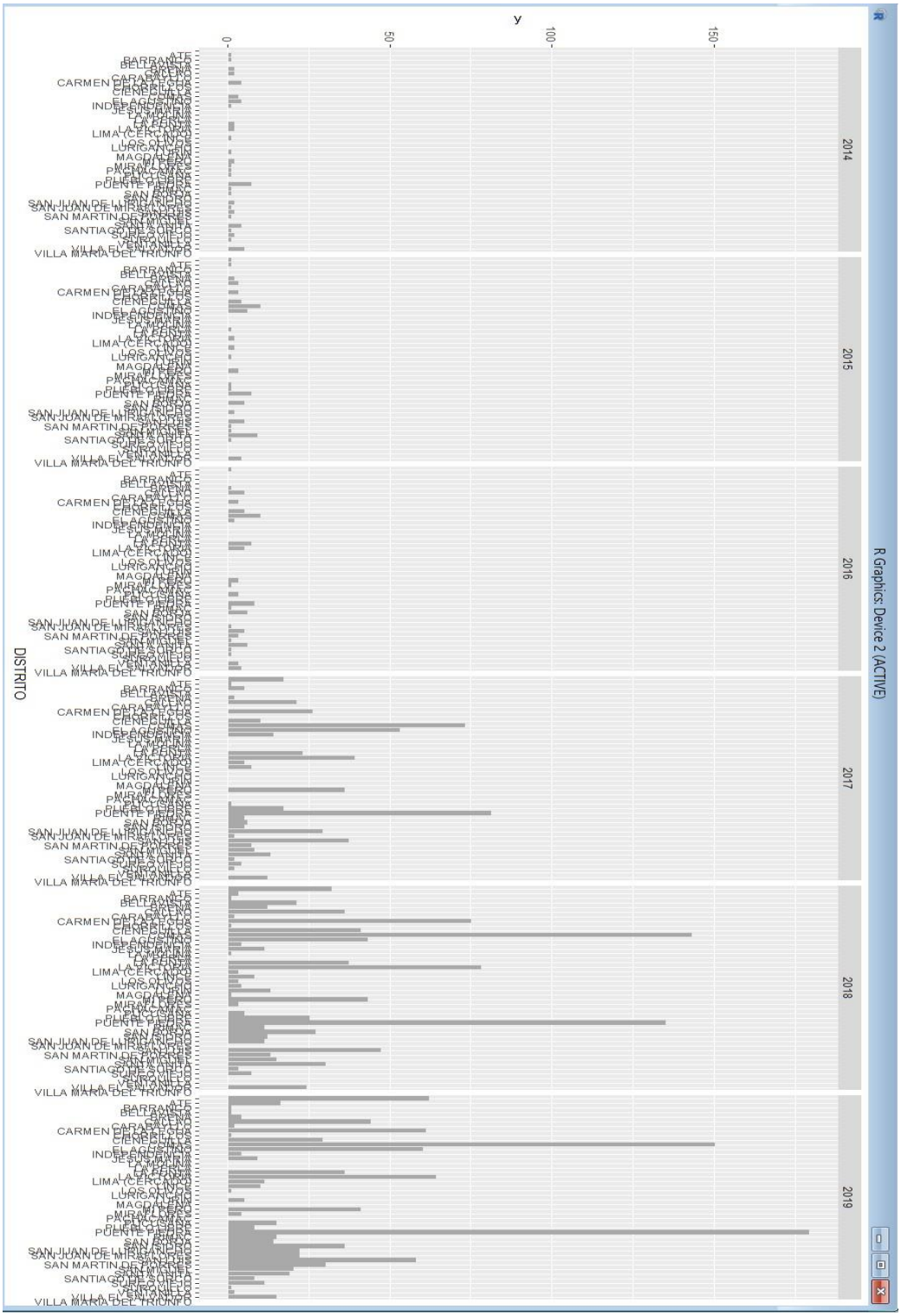
Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos aumento en las fugas en los años desde el 2017 al 2019 y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe modelar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Figura Nº 43

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
# FUGAS POR DISTRITO Y POR TIPO DE FUGA

ggplot(sedapal, aes(x = DISTRITO, y = 1, group =

DISTRITO)) +

facet_wrap(~ FECHA, nrow = 3) +

geom_bar(stat = "identity", fill = "darkgray") +

theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

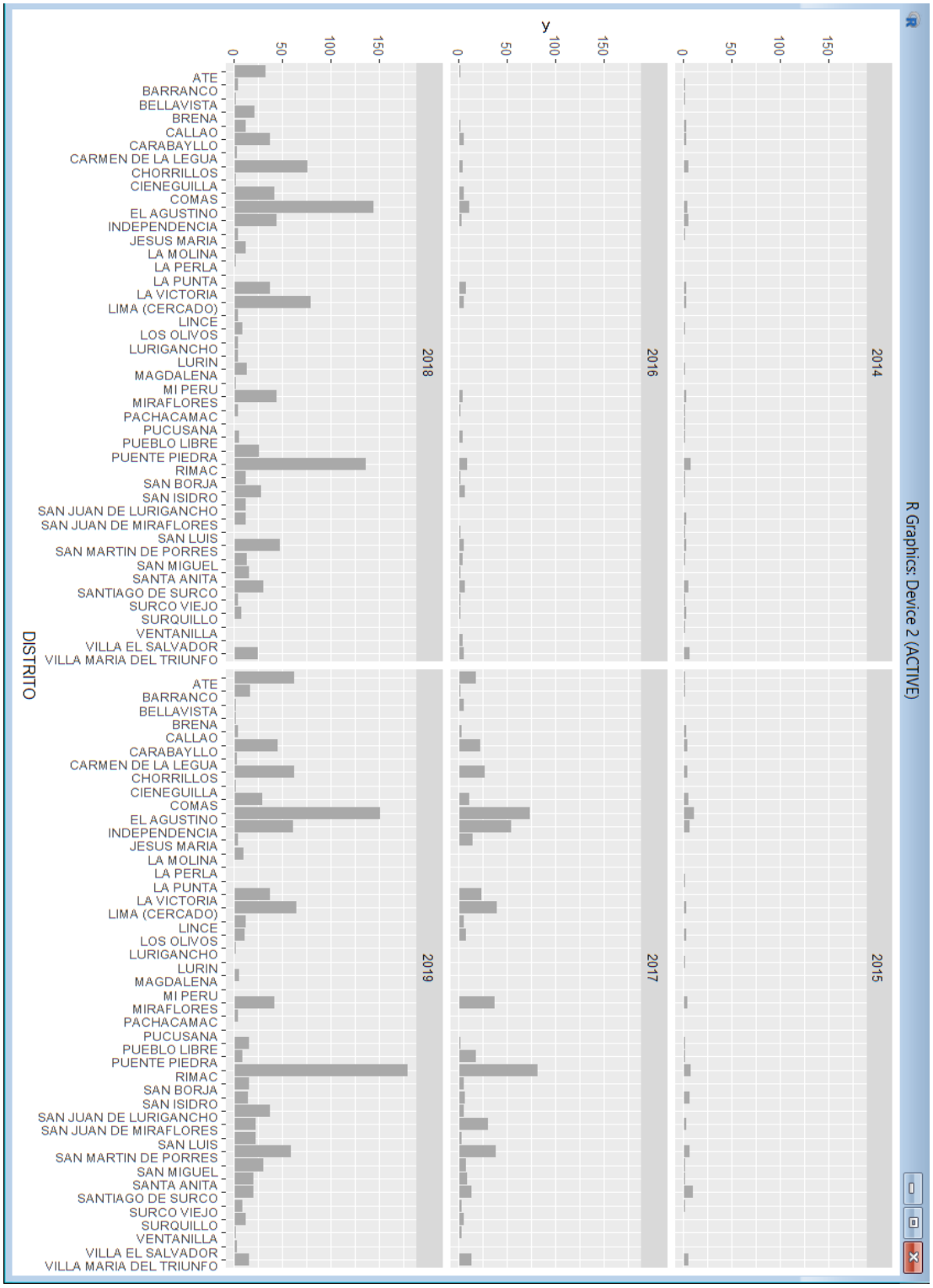
Al ejecutar nuestro algoritmo modificado, como podemos apreciar este tipo de grafico obtenido da mucha más información al mostrarla en formato de columnas superpuestas mostrando distritos y los años, hemos aplicado el l tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos aumento en las fugas en los años desde el 2017 al 2019 y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible y se nota la tendencia al crecimiento de las fugas según el año y también ya es posible apreciar que distritos son los que más servicios de fugas de emergencia presentan.

Figura Nº 44

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Tenemos una gran cantidad de datos por analizar desde diversos frentes ya que contamos con 3 variables 2 de ellas son datos cualitativos y uno cuantitativo, es por ello que se le da un enfoque tratando de encontrar la relación que mejor defina la tendencia de la información con la que contamos.

Nuestro algoritmo se aplica al tratamiento de la información y mostrarla en forma gráfica para su mejor comprensión y predicción de la tendencia, haciendo uso de las herramientas y librerías que el software R nos brinda.

- **RELACION ENTRE VARIABLES DISTRITO Y TIPOFUGA**

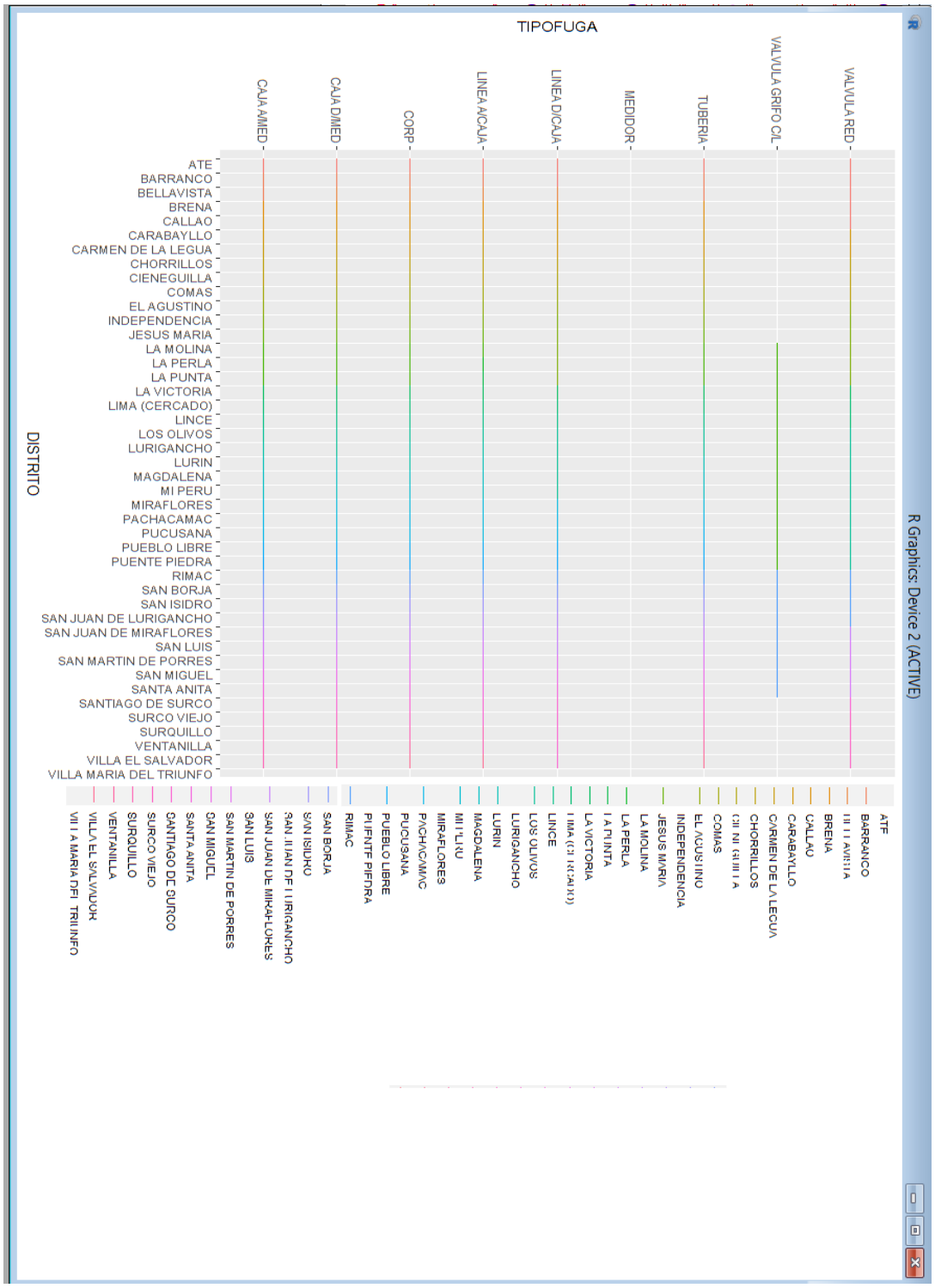
```
ggplot(sedapal, aes(x = DISTRITO, y = TIPOFUGA, color =
DISTRITO, group = TIPOFUGA)) + geom_line() +
  theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido no aporta mucha información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo polilínea el cual no es por lo visto el más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar .

Como observamos tenemos en el eje **Y** los tipos de fugas y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

Figura Nº 45

DATOS OBTENIDOS MODELO DEL TIPO POLILINEAL



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(data=sedapal_tabla, aes(x=reorder
(TIPOFUGA,DISTRITO), y=1, fill=DISTRITO)) +
   geom_bar(stat="identity", position="stack")
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

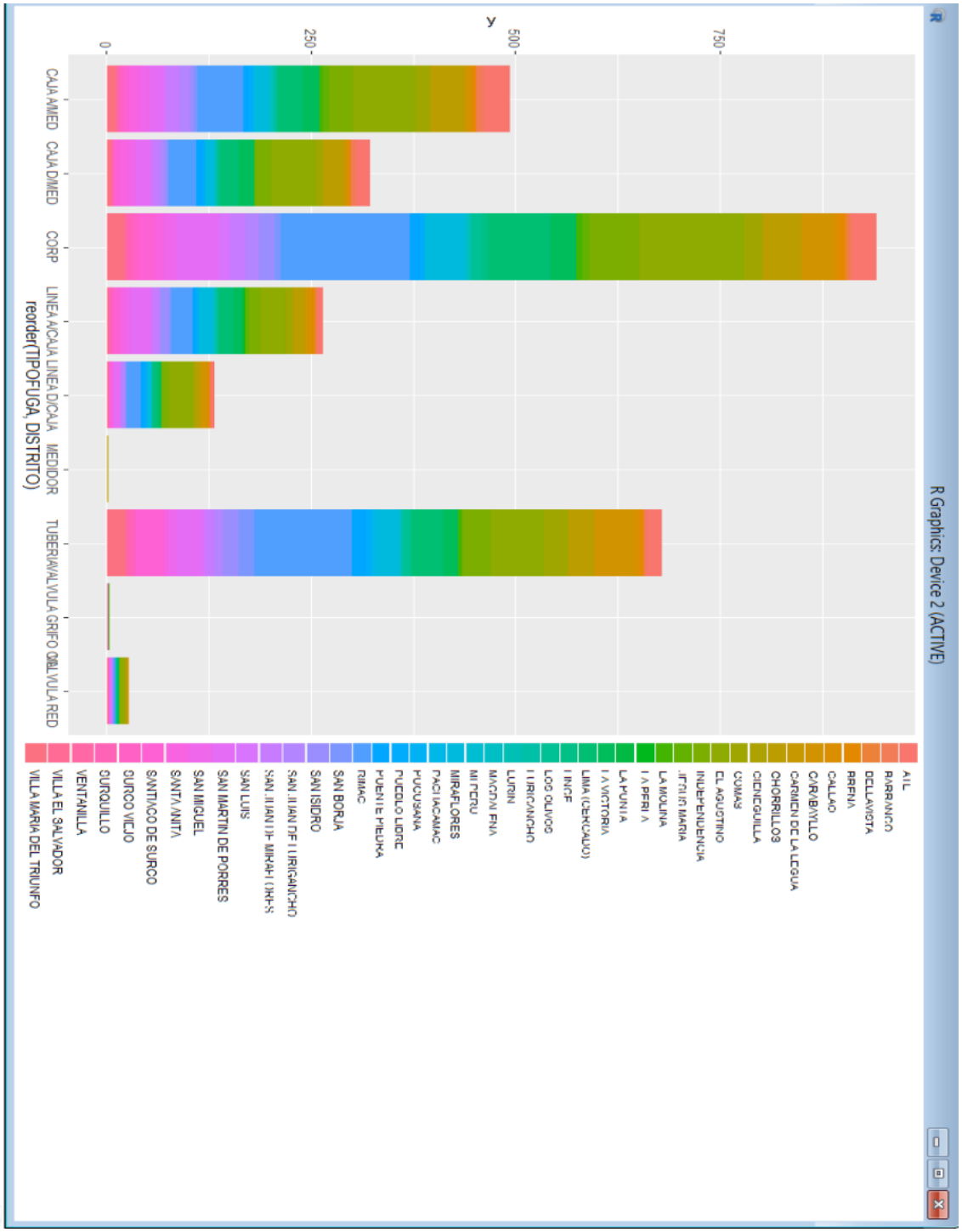
Como apreciamos tenemos los distintos tipos de fugas encontradas en las emergencias y distritos en el eje **X** las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe detallar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Hacemos uso del comando **reorder** para ordenar los datos y también estamos utilizando colores para diferenciar los distritos, sin embargo no logramos aun información concluyente.

Figura Nº 46

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(sedapal, aes(x = DISTRITO, y = 1, group =
DISTRITO)) +
facet_wrap(~ TIPOFUGA, nrow = 1) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

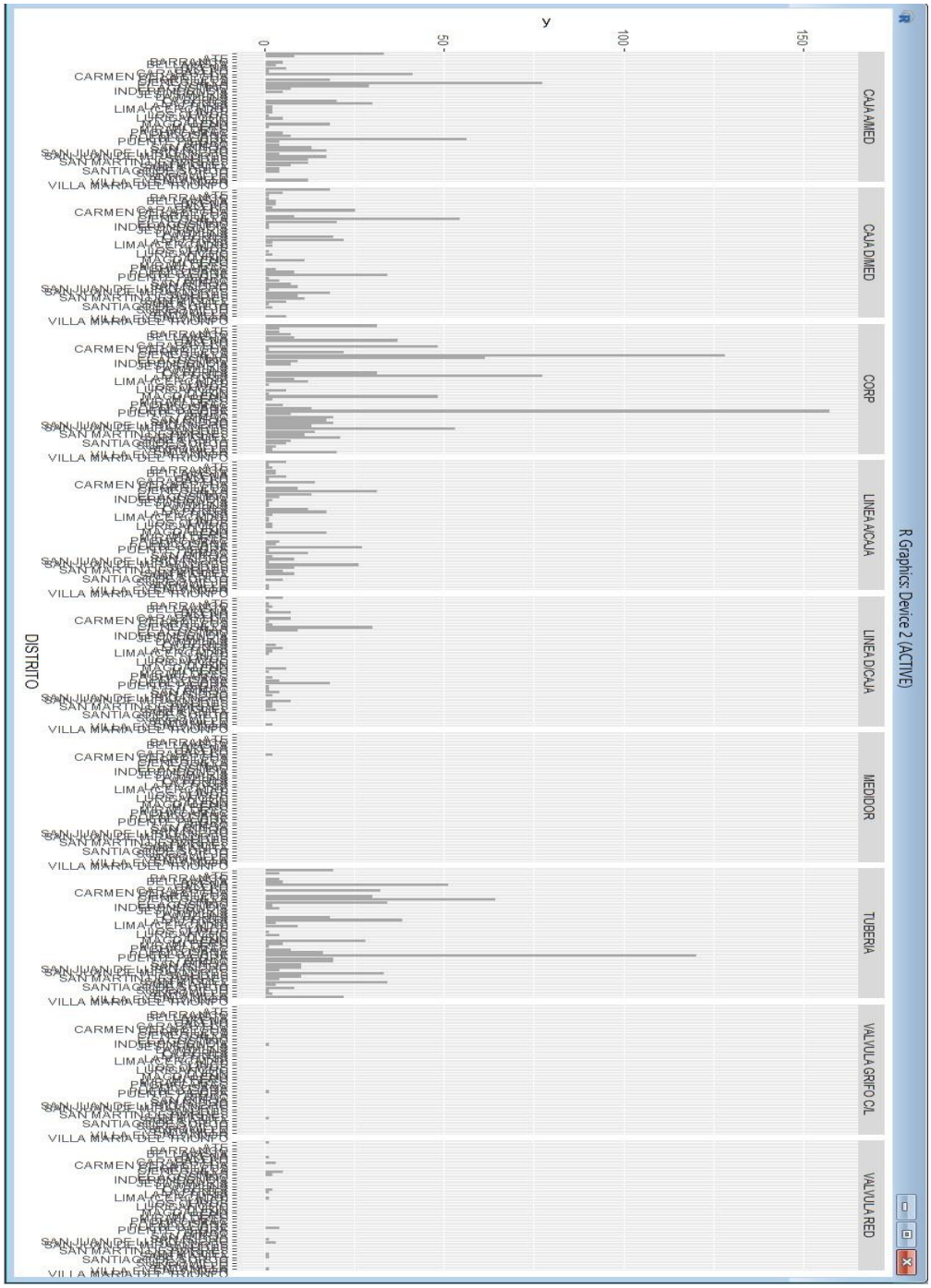
Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos los distintos tipos de fugas encontradas en las emergencias y distritos en el eje **X** las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe detallar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Figura Nº 47

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(sedapal, aes(x = DISTRITO, y = 1, group =
DISTRITO)) +
facet_wrap(~ TIPOFUGA, nrow = 3) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

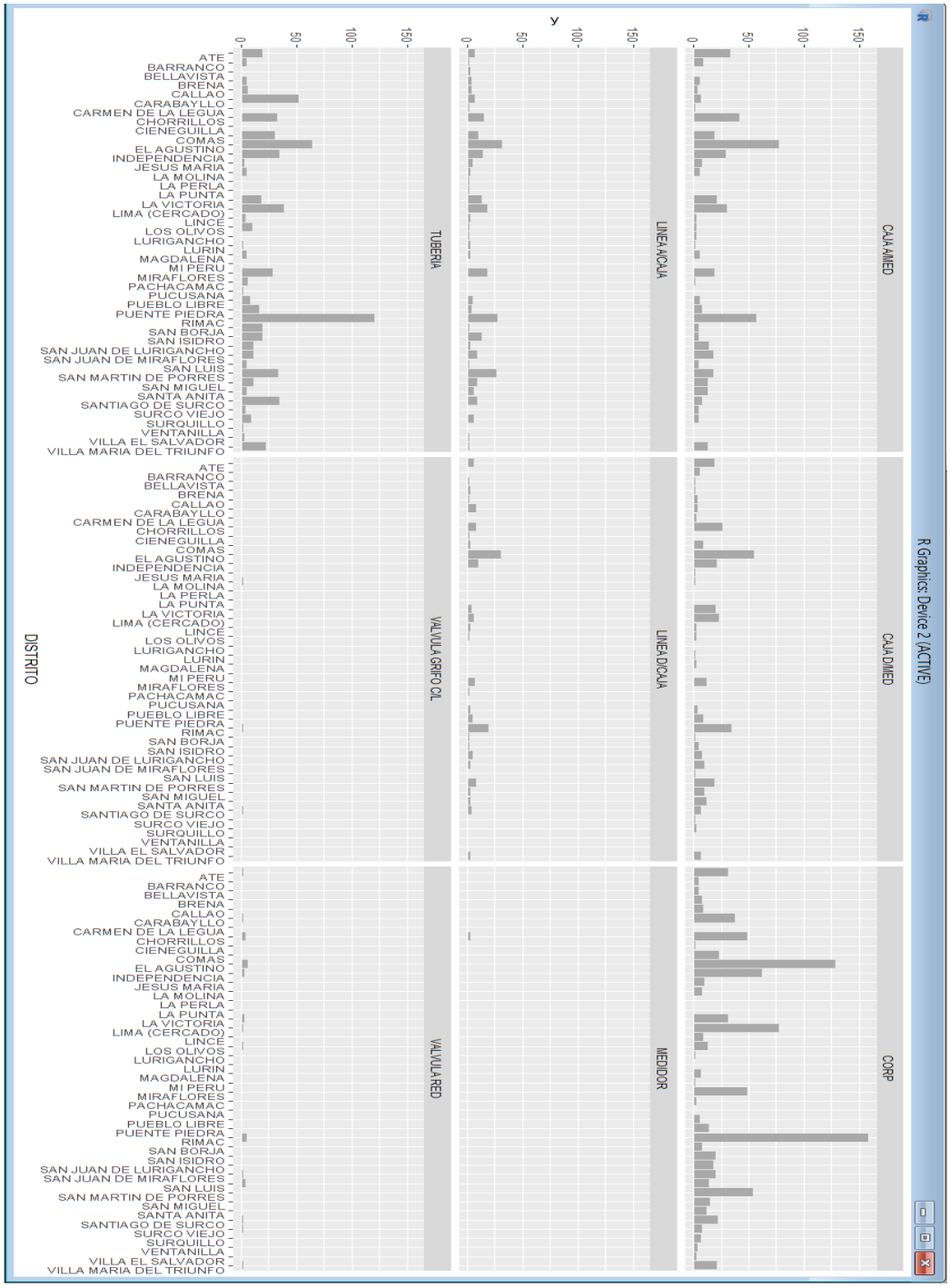
Al ejecutar nuestro algoritmo modificado, como podemos apreciar este tipo de grafico obtenido da mucha más información al mostrarla en formato de columnas superpuestas mostrando distritos y los tipos de fugas, hemos aplicado el l tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos fugas con mucha incidencia y algunos tipos son casi inexistentes y en el eje **X** la relación de distritos en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible y se nota la tendencia al crecimiento de las fugas y también ya es posible apreciar que distritos son los que más servicios de fugas de emergencia presentan.

Figura Nº 48

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Tenemos una gran cantidad de datos por analizar desde diversos frentes ya que contamos con 3 variables 2 de ellas son datos cualitativos y uno cuantitativo, es por ello que se le da un enfoque tratando de encontrar la relación que mejor defina la tendencia de la información con la que contamos.

Nuestro algoritmo se aplica al tratamiento de la información y mostrarla en forma gráfica para su mejor comprensión y predicción de la tendencia, haciendo uso de las herramientas y librerías que el software R nos brinda.

- **RELACION ENTRE VARIABLES FECHA Y TIPOFUGA**

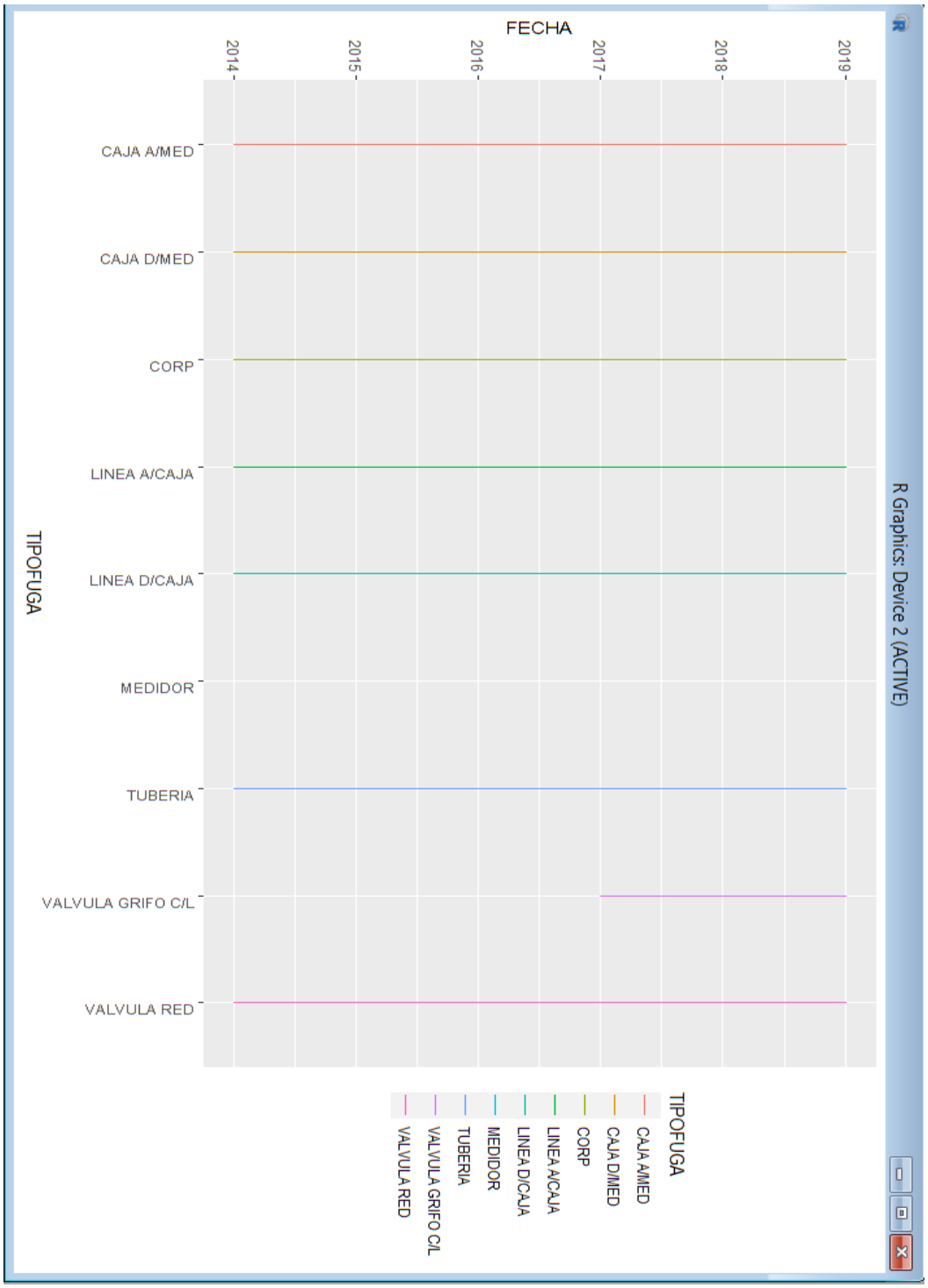
```
ggplot(sedapal, aes(x = TIPOFUGA, y = FECHA, color =
TIPOFUGA, group = TIPOFUGA)) + geom_line() +
  theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido no aporta mucha información sobre las tendencias de las fugas en función de los tipos de fugas y de los años, hemos aplicado un tipo de grafico del tipo polilínea el cual no es por lo visto el más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar .

Como observamos tenemos en el eje **Y** los tipos años y en el eje **X** los tipos de fugas los cuales se presentaron de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

Figura Nº 49

DATOS OBTENIDOS MODELO DEL TIPO POLILINEAL



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(data=sedapal_tabla, aes(x=reorder
(TIPOFUGA,FECHA), y=1, fill=FECHA)) +
   geom_bar(stat="identity", position="stack")
```

Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

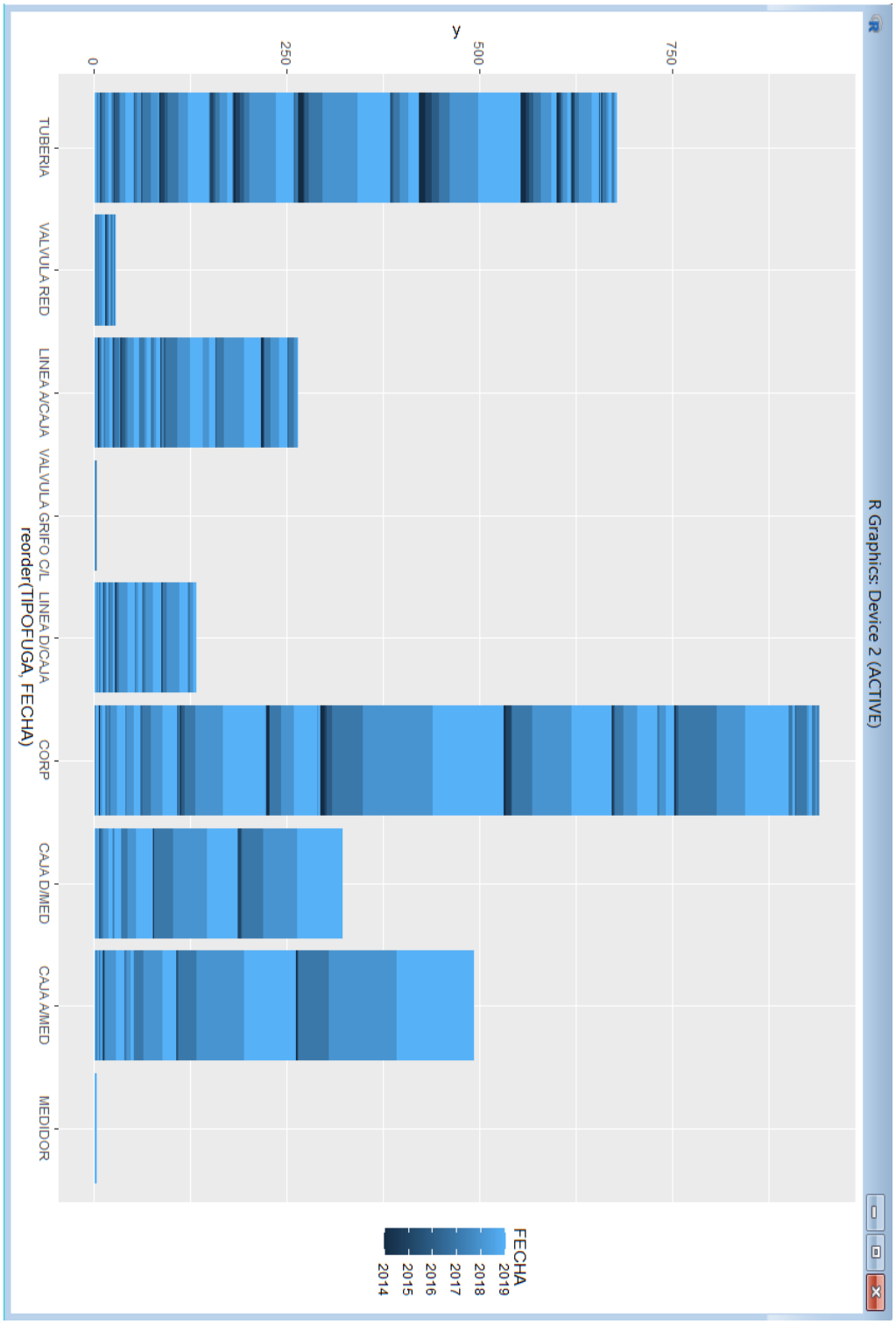
Como apreciamos tenemos los distintos tipos de fugas encontradas en las emergencias en el eje **X** las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe detallar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Hacemos uso del comando **reorder** para ordenar los datos y también estamos utilizando colores para diferenciar los distritos, sin embargo no logramos aun información concluyente.

Figura Nº 50

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(sedapal, aes(x = FECHA, y = 1, group =
FECHA)) +
facet_wrap(~ TIPOFUGA, nrow = 1) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

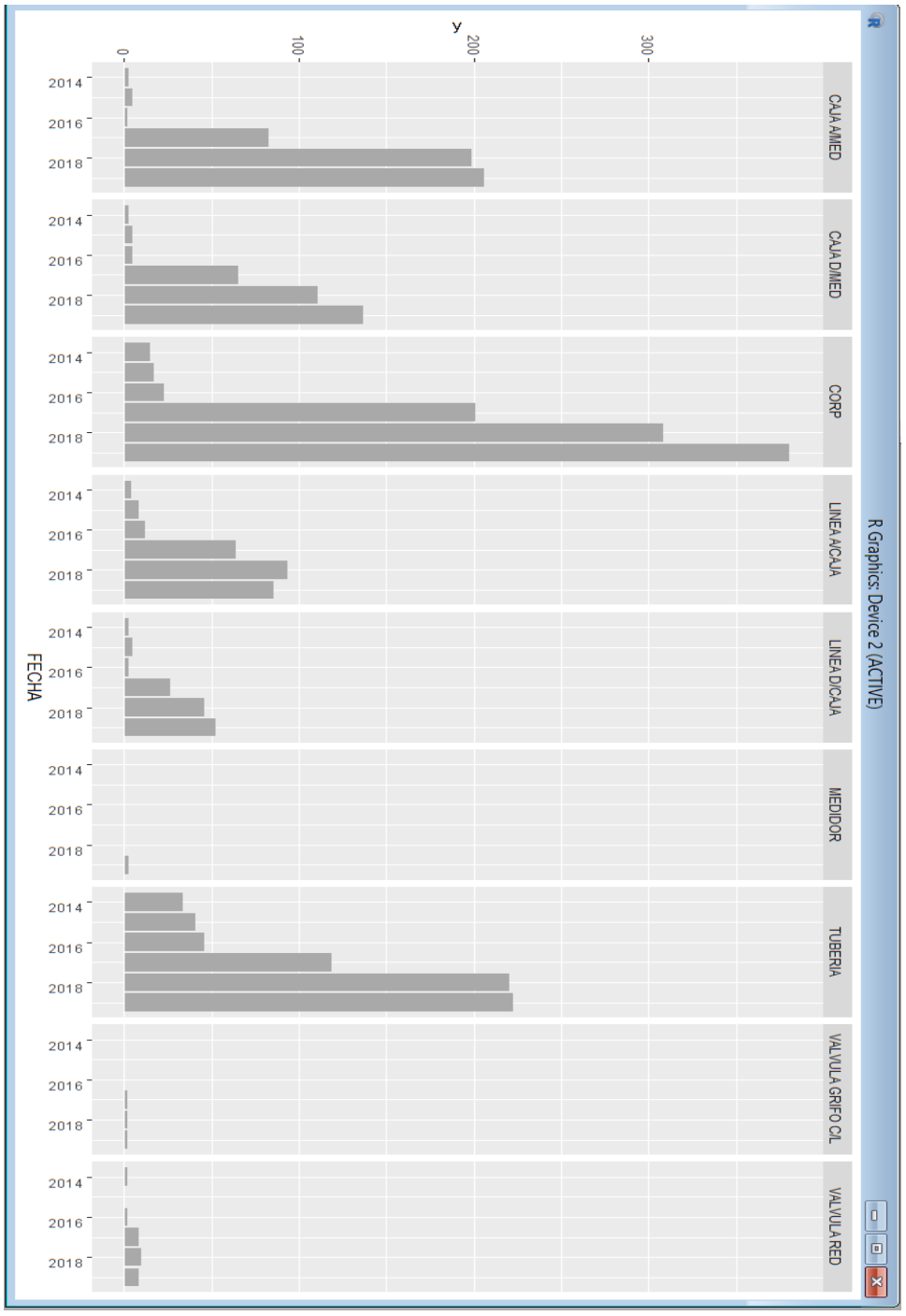
Al ejecutar nuestro algoritmo, como podemos apreciar este tipo de dato obtenido aporta algo de información sobre las tendencias de las fugas en función de los distritos y de los años, hemos aplicado un tipo de grafico del tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos los distintos tipos de fugas encontradas en las emergencias y distritos en el eje **X** las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible pero se debe detallar aún más para encontrar la tendencia en los datos obtenidos para utilizarlo como modelo.

Figura Nº 51

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



Fuente: Propia del autor

Como vemos tenemos que darle un enfoque diferente al análisis empezando por como ordenar las variables y como graficarlas de una mejor manera.

```
ggplot(sedapal, aes(x = FECHA, y = 1, group =
FECHA)) +
facet_wrap(~ TIPOFUGA, nrow = 3) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

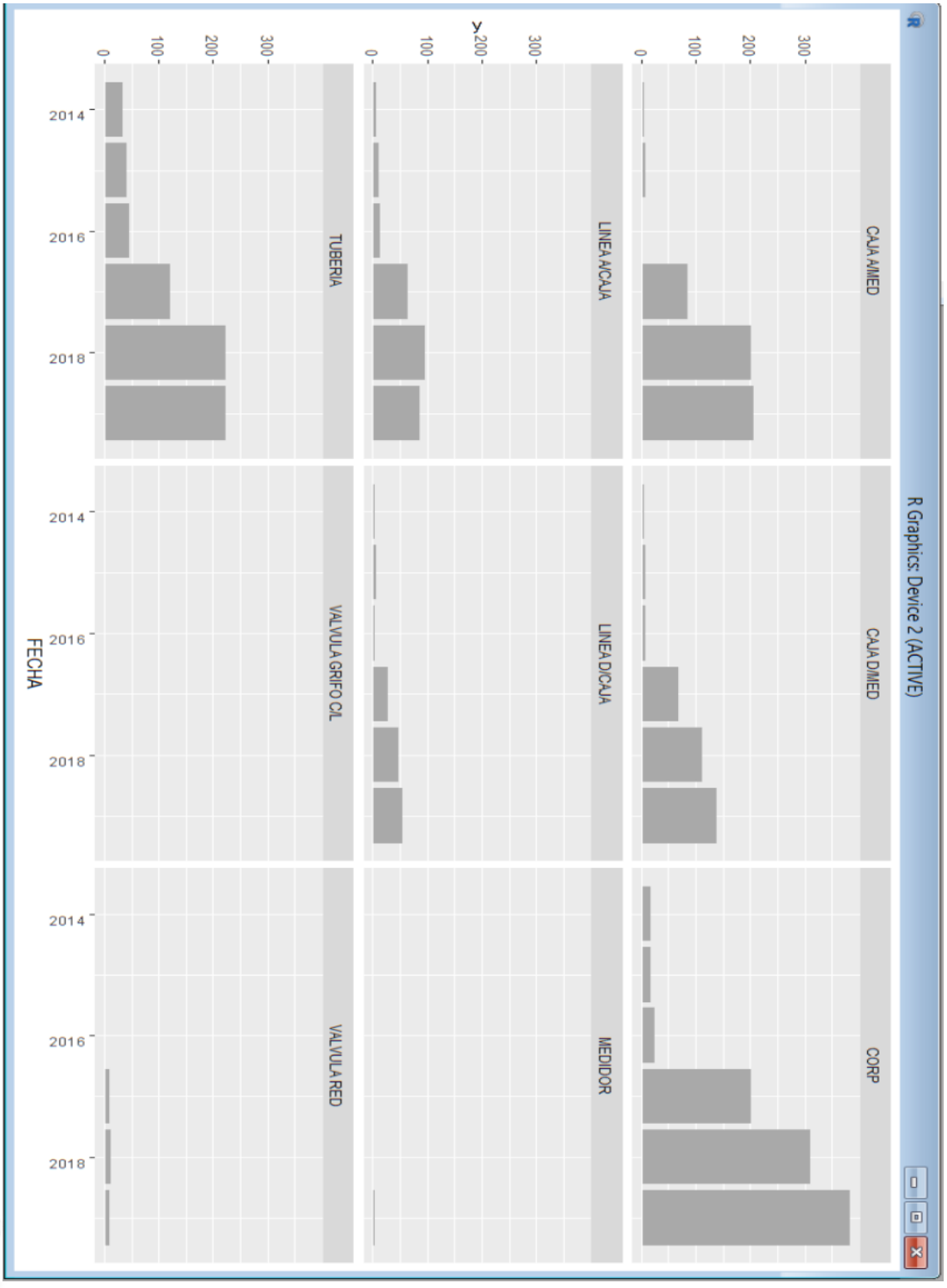
Al ejecutar nuestro algoritmo modificado, como podemos apreciar este tipo de grafico obtenido da mucha más información al mostrarla en formato de columnas superpuestas mostrando años y los tipos de fugas, hemos aplicado el tipo barras el cual es por lo visto más adecuado para nuestro modelo en cuestión, es muy importante tener en cuenta el tipo de grafico a utilizar, así como también el tipo de datos que vamos a analizar.

Como apreciamos tenemos fugas con mucha incidencia y algunos tipos son casi inexistentes y en el eje **X** la relación de años en los cuales se presentaron las fugas de agua potable de emergencia (tubería, corporation, fuga en caja, etc.) las cuales forman parte de esta base de datos 2014 -2019.

La información así presentada es un poco más legible y se nota la tendencia al crecimiento de las fugas y también ya es posible apreciar que distritos son los que más servicios de fugas de emergencia presentan.

Figura Nº 52

DATOS OBTENIDOS MODELO DEL TIPO BARRAS



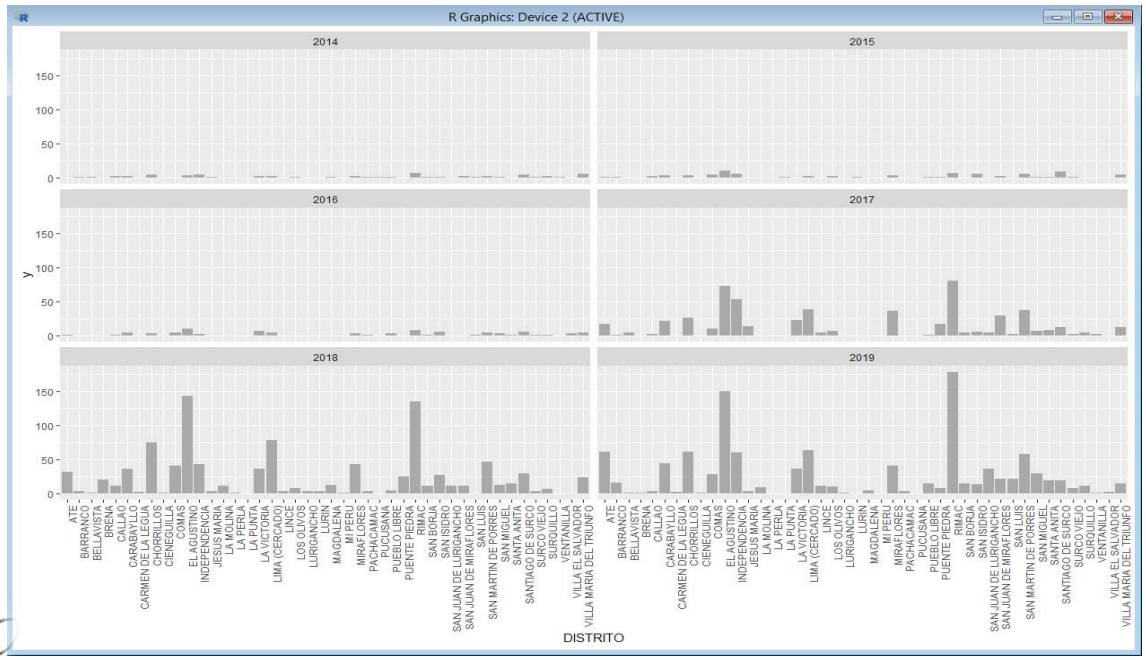
Fuente: Propia del autor

La información así presentada es un poco más legible y se nota la tendencia al crecimiento de las fugas y también ya es posible apreciar que distritos y en que años se dieron más estas emergencia, por lo que se utilizara este algoritmo como fuente para hacer la proyección estadística general.

```
ggplot(sedapal, aes(x = CAMPO1, y = 1, group =
CAMPO1)) +
facet_wrap(~ CAMPO2, nrow = 3) +
geom_bar(stat = "identity", fill = "darkgray") +
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Figura Nº 53

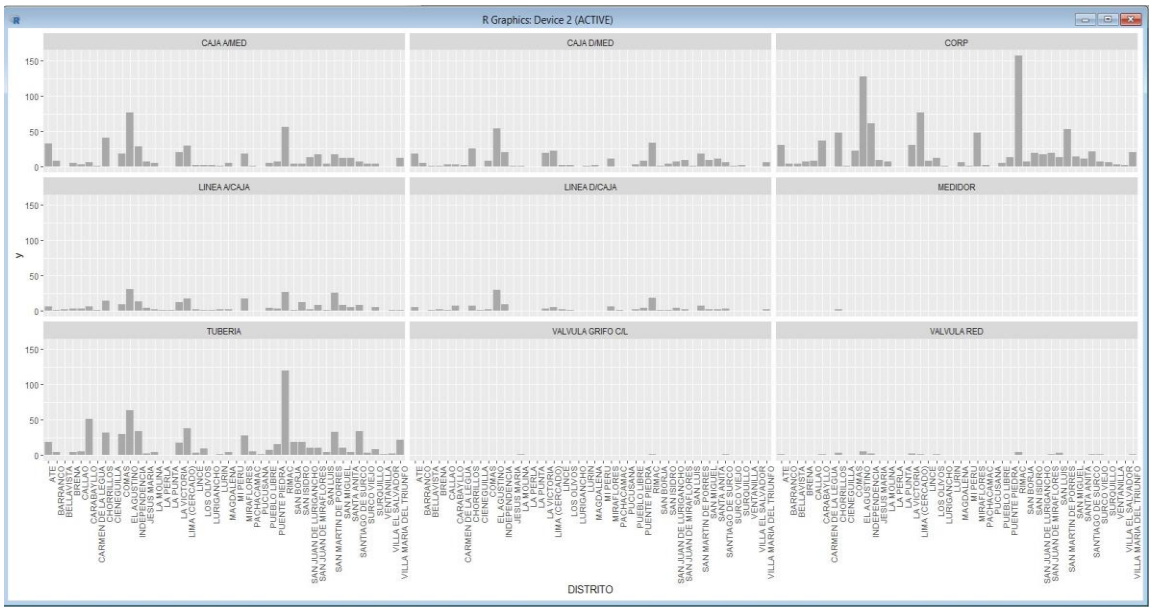
MODELO DEL TIPO BARRAS : DISTRITOS & FECHA



Fuente: Propia del autor

Figura Nº 54

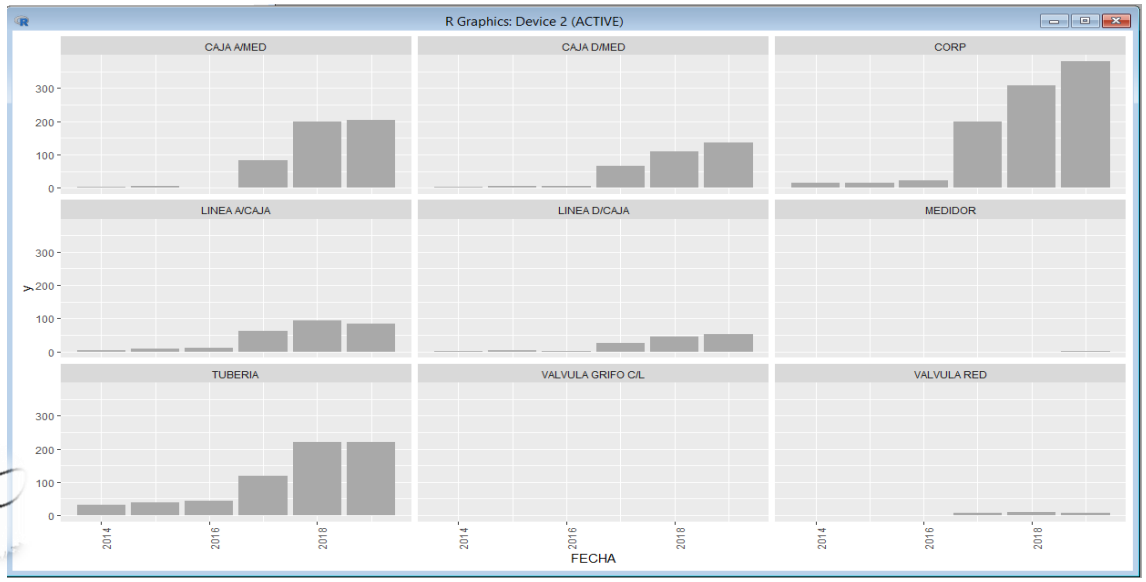
MODELO DEL TIPO BARRAS: DISTRITOS & TIPOFUGAS



Fuente: Propia del autor

Figura Nº 55

MODELO DEL TIPO BARRAS : TIPOFUGAS & FECHA



Fuente: Propia del autor

4.3. Población y Muestra

En esta investigación predominantemente tecnológica, transversal, intitulada: “DISEÑO DE UN ALGORITMO PREDICTIVO PARA MONITOREO TEMPRANO DE REDES DE AGUA POTABLE EN LA CIUDAD DE LIMA, 2019” es importante delimitar sus alcances para su análisis correspondiente.

4.3.1 Unidad de análisis

La delimitación de la Unidad de Análisis del problema objeto de investigación, para diseño de un algoritmo predictivo para monitoreo temprano de redes de agua potable en la ciudad de Lima, se ha determinado geográficamente como “unidad de análisis” a la ciudad de Lima en el periodo comprendido entre enero del 2014 a diciembre del 2019.

4.3.2 Población

En esta investigación se considera como población a las fugas de agua potable detectadas ya sean del tipo: Industriales, Comerciales y Residenciales, ubicados en la ciudad de Lima.

De esta gran muestra que son aproximadamente 8700 fugas detectadas e ingresadas al sistema SGIOC se procesaron algo más de 2500 fugas bajo la premisa de una presión de agua constante de entre 30 y 40 libras aproximadamente.

4.3.3 Tamaño de la Muestra

Para determinar probabilísticamente el “Tamaño de la Muestra” es necesario definir la característica principal de la población constituido por las fugas de agua potable detectadas en ese periodo de tiempo, que corresponde a una población “finita”, de (N = 2500) que es menor igual a el tamaño de la muestra de 8700 (N ≤ 8700).

4.3.4 Prueba piloto

Debido a que no existen antecedentes de estudio correspondiente a esta investigación para determinar el tamaño de la muestra, se ha optado por descargar los datos obtenidos durante ese periodo de tiempo 5 año y procesarlos para verificar que porcentaje de las fugas seleccionadas fueron correctamente procesadas y cuales no y los tipos de fugas.

El procesamiento arrojo que solo el 60% de las fugas fue detectado y catalogado correctamente en ese periodo de tiempo.

4.4. Lugar de estudio y periodo desarrollado

4.4.1 Algoritmo Predictivo usando Software Estadístico R

Para ello, ha sido necesario desarrollar las siguientes actividades:

- Justificar el uso del software R
- Procesar la información de la base de datos SGIOC
- Seleccionar la muestra a utilizar del periodo 2014 al 2019.
- Implementar el algoritmo predictivo para el tratamiento de la información

- Finalmente hacer pruebas en campo con ambos sistemas y ver la tasa de eficiencia de ambos.

4.4.2 La información obtenida por el algoritmo repercutirá en la mejora del servicio

Se implementó usando una laptop i5 con el software Excel y Software R además de la data que se encuentran en la base de datos SGIO de Sedapal.

4.4.3 El crecimiento de la población con servicio de agua potable influirá en los resultados obtenidos.

4.4.4 Procesamiento estadístico y análisis de datos

El procedimiento estadístico a ser aplicado en esta investigación para explicar, demostrar y verificar lo planteado en la hipótesis, consistirá primeramente en las lecturas de la data obtenida llámese las fugas propiamente dichas obtenidas por el software de gestión de Fugas SGIOC durante el periodo comprendido entre enero del 2014 a diciembre del 2019

CAPÍTULO V: RESULTADOS

5.1 Resultados descriptivos

Como se puede apreciar en la tabla 17 entre los años 2014 -2016 no se presentaron muchas fugas en los Distritos de la Gran Lima.

Como se puede apreciar en la tabla 18 los tipos de fugas : medidor , válvula grifo cl y valvula grifo no se detectaron o fueron casi nulos en los los Distritos.

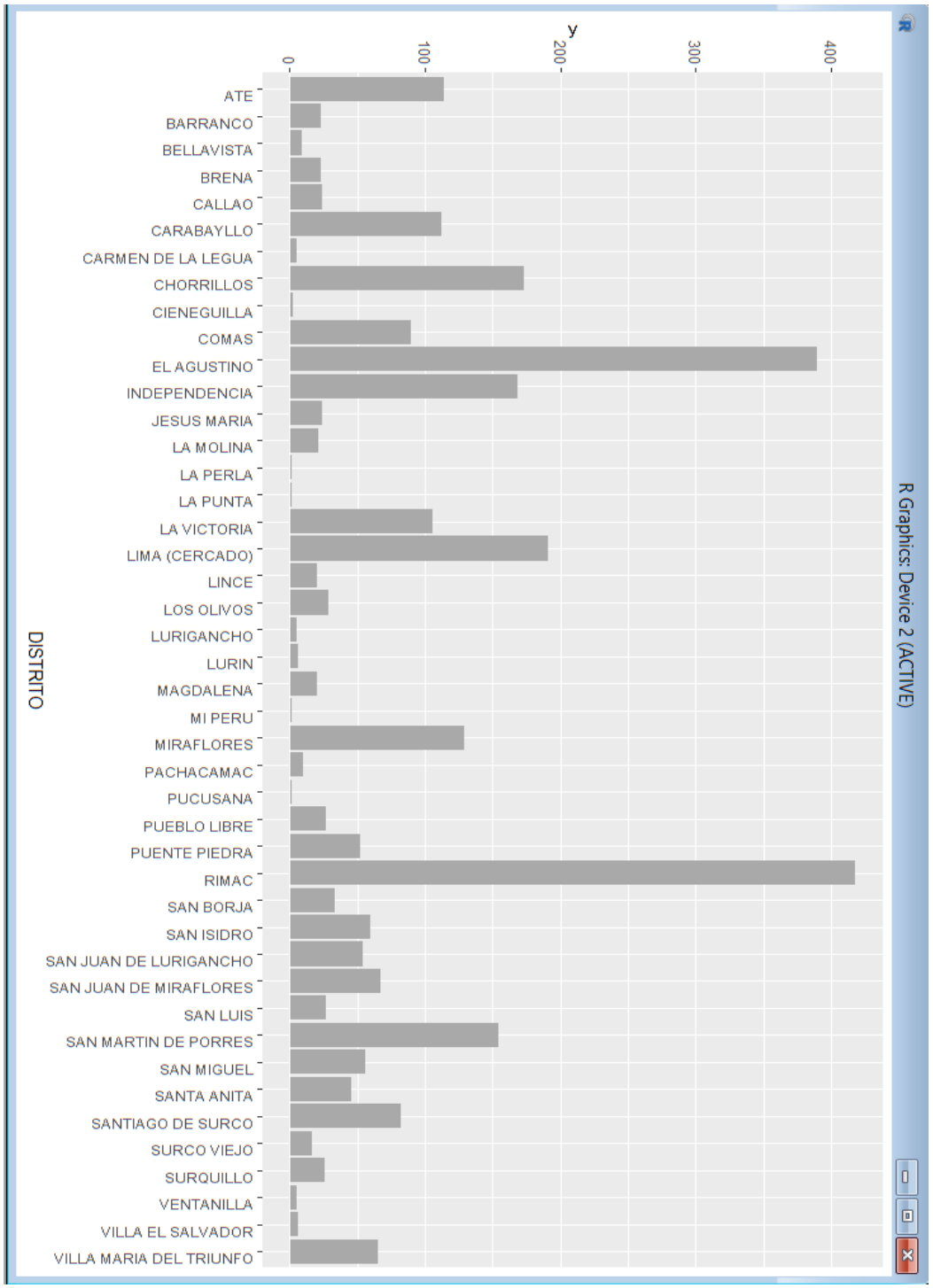
Como se puede apreciar en la tabla 19 los tipos de fugas : medidor , válvula grifo cl y valvula grifo no se detectaron o fueros casi nulos en el periodo 2014 - 2016. Tenemos un mejor modelo de tratamiento de los datos y se puede generalizar aún más para estimar las proyecciones. Es por eso que agruparemos la variable FECHA para tener datos globales de los 7 años, entonces modificamos el algoritmo y obviamos el campo FECHA y nos centramos en las variables TIPOFUGA en función de la variable DISTRITO para tener el consolidado de las proyecciones.

Modificamos nuestro algoritmo en función de la cantidad de TIPOFUGA encontradas y lo ejecutamos en la venta de comandos del Software R.

```
ggplot(data=sedapal_tabla, aes(x=DISTRITO, y = 1, )) +
   geom_bar(stat="identity",fill = "darkgray",
position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Figura Nº 56

MODELO DEL TIPO BARRAS : TIPOFUGAS & DISTRITO 2014-2019



Fuente: Propia del autor

Tenemos un mejor modelo de tratamiento de los datos y se puede generalizar aún más para estimar las proyecciones. Es por eso que agruparemos la variable FECHA para tener datos globales de los 7 años, entonces modificamos el algoritmo y obviamos el campo FECHA y nos centramos en las variables TIPOFUGA en función de la variable DISTRITO para tener el consolidado de las proyecciones.

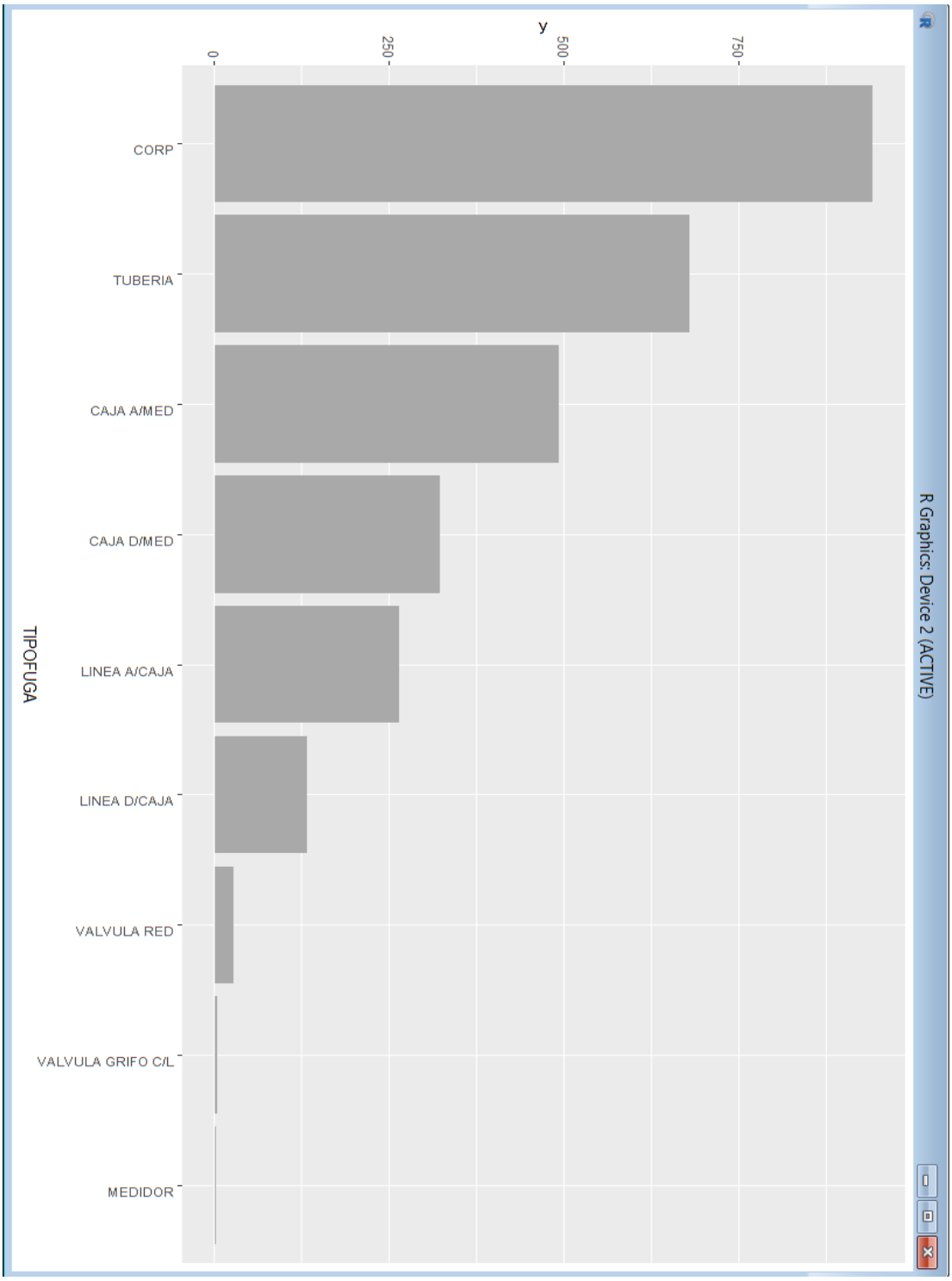
Luego procedemos a ordenar nuestra variable TIPOFUGA haciendo uso del comando **factor**

```
sedapal_tabla$TIPOFUGA = factor(sedapal_tabla$TIPOFUGA,
levels=c("CORP","TUBERIA", "CAJA A/MED", "CAJA D/MED", "LINEA A/CAJA",
"LINEA D/CAJA", "VALVULA RED", "VALVULA GRIFO C/L", "MEDIDOR" ))

levels(sedapal_tabla$TIPOFUGA)

 "CORP"          "TUBERIA"       "CAJA A/MED"    "CAJA D/MED"
"LINEA A/CAJA"  "LINEA D/CAJA"  "VALVULA RED"   "VALVULA GRIFO
C/L" "MEDIDOR"

ggplot(data=sedapal_tabla, aes(x=DISTRITO, y = 1, )) +
   geom_bar(stat="identity",fill = "darkgray",
position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

Como vemos ya tenemos ordenados los tipos de fugas TIPOFUGA por su número a los largo del periodo 2014-2019 desde CORP hasta MEDIDOR.

Figura Nº 57

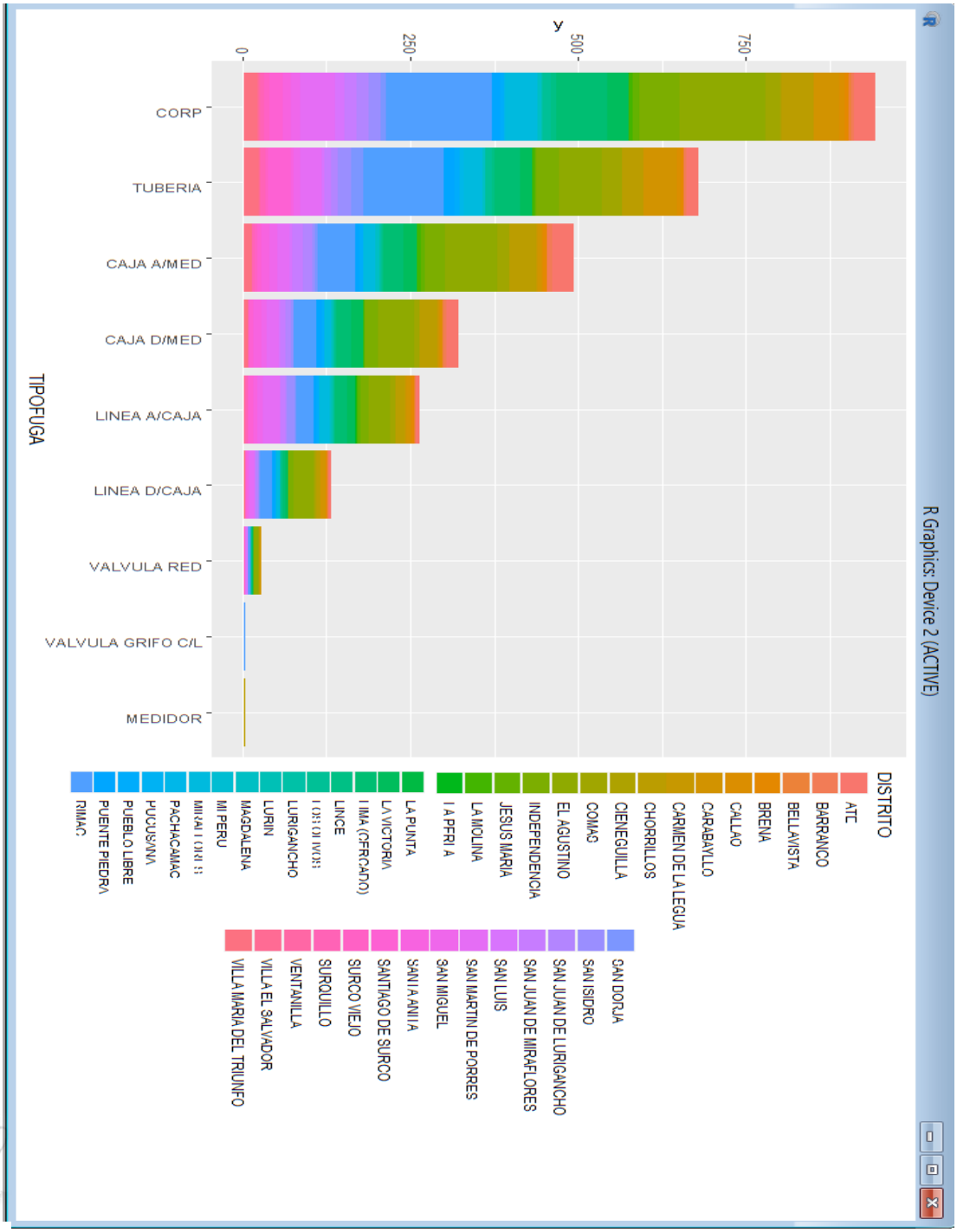
MODELO DEL TIPO BARRAS ORDENADO POR TIPOFUGAS 2014-2019



Fuente: Propia del autor

Figura Nº 58

MODELO DEL TIPO BARRAS ORDENADO POR TIPOFUGAS & DISTRITO 2014-2019



Fuente: Propia del autor

Tenemos un mejor modelo de tratamiento de los datos y se puede generalizar aún más para estimar las proyecciones. Es por eso que agruparemos la variable FECHA para tener datos globales de los 7 años, entonces modificamos el algoritmo y obviamos el campo FECHA y nos centramos en las variables DISTRITO en función de la variable TIPOFUGA para tener el consolidado de las proyecciones.

Luego procedemos a ordenar nuestra variable DISTRITO haciendo uso del comando **factor**

```
sedapal_tabla$DISTRITO = factor(sedapal_tabla$DISTRITO,
levels=c(
"RIMAC","EL AGUSTINO","LIMA (CERCADO)","CHORRILLOS",
"INDEPENDENCIA","SAN MARTIN DE PORRES","MIRAFLORES",
"ATE","CARABAYLLO","LA VICTORIA","COMAS", "SANTIAGO DE SURCO",
"SAN JUAN DE MIRAFLORES","VILLA MARIA DEL TRIUNFO","SAN
ISIDRO","SAN MIGUEL","SAN JUAN DE LURIGANCHO","PUENTE
PIEDRA","SANTA ANITA","SAN BORJA",
"LOS OLIVOS","PUEBLO LIBRE","SAN LUIS","SURQUILLO",
"CALLAO","BRENA","BARRANCO","JESUS MARIA","LA MOLINA",
"MAGDALENA","LINCE","SURCO VIEJO","PACHACAMAC",
"BELLAVISTA","LURIN","VILLA EL SALVADOR","VENTANILLA",
"LURIGANCHO","CARMEN DE LA LEGUA","CIENEGUILLA",      "MI
PERU","PUCUSANA","LA PERLA","LA PUNTA"
))
```

ggplot(data=sedapal_tabla, aes(x=DISTRITO, y = 1,)) +

geom_bar(stat="identity",fill = "darkgray",

position="stack")+

theme(axis.text.x = element_text(angle = 90, hjust = 1))

levels(sedapal_tabla$DISTRITO)

"RIMAC" "EL AGUSTINO" "LIMA (CERCADO)" "CHORRILLOS" "INDEPENDENCIA" "SAN MARTIN DE PORRES" "MIRAFLORES"

"ATE" "CARABAYLLO" "LA VICTORIA" "COMAS" "SANTIAGO DE SURCO" "SAN JUAN DE MIRAFLORES" "VILLA MARIA DEL TRIUNFO"

“SAN ISIDRO" "SAN MIGUEL" "SAN JUAN DE LURIGANCHO" "PUENTE PIEDRA" "SANTA ANITA" "SAN BORJA" "LOS OLIVOS"

"PUEBLO LIBRE" "SAN LUIS" "SURQUILLO" "CALLAO" "BRENA" "BARRANCO" "JESUS MARIA"

"LA MOLINA" "MAGDALENA" "LINCE" "SURCO VIEJO" "PACHACAMAC" "BELLAVISTA" "LURIN"

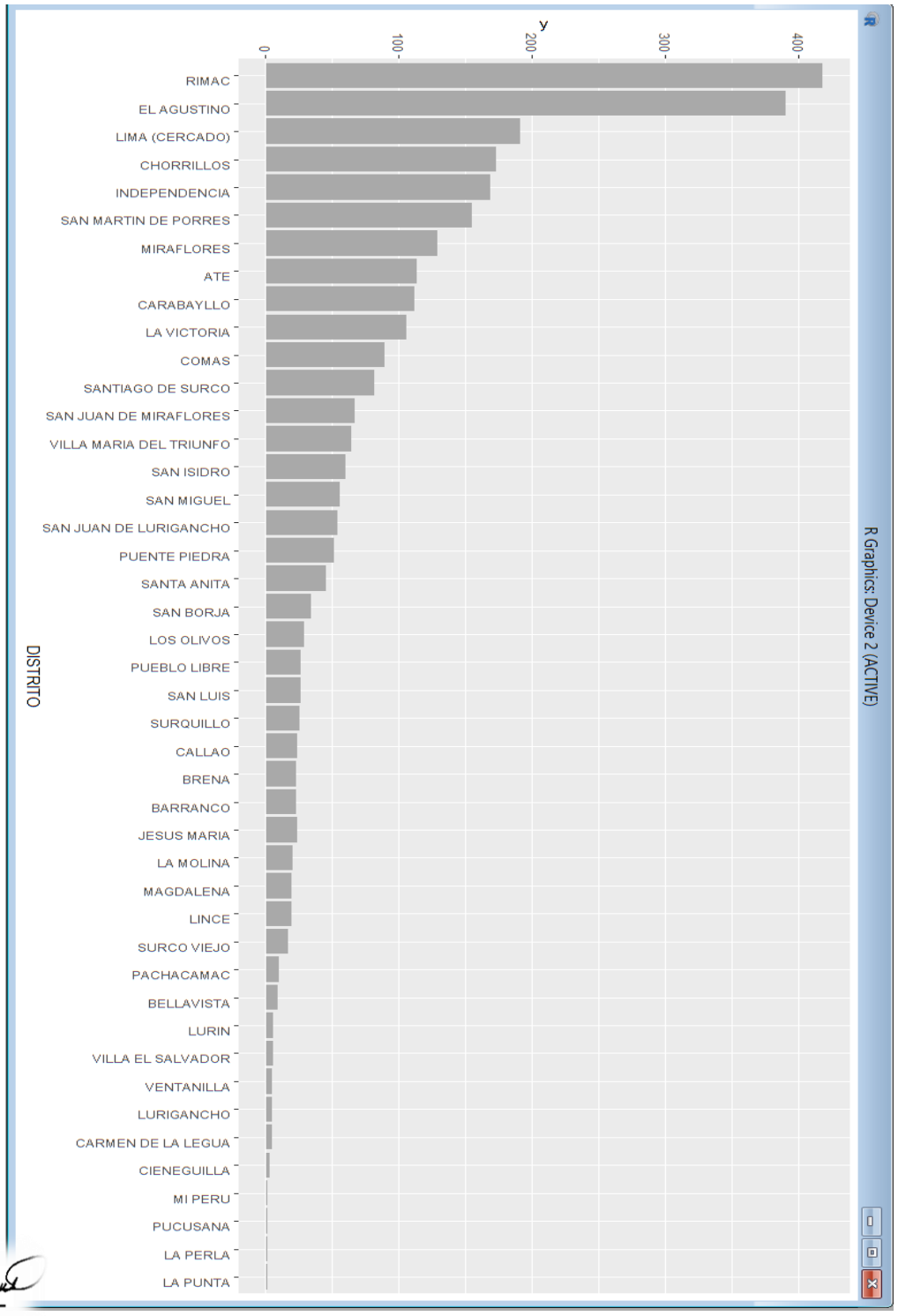
"VILLA EL SALVADOR" "VENTANILLA" "LURIGANCHO" "CARMEN DE LA LEGUA" "CIENEGUILLA" "MI PERU" "PUCUSANA"

"LA PERLA" "LA PUNTA"

Como vemos ya tenemos ordenados los distritos DISTRITO por su número de fugas detectadas a los largo del periodo 2014-2019 desde RIMAC hasta LA PUNTA (44 distritos).

Figura Nº 59

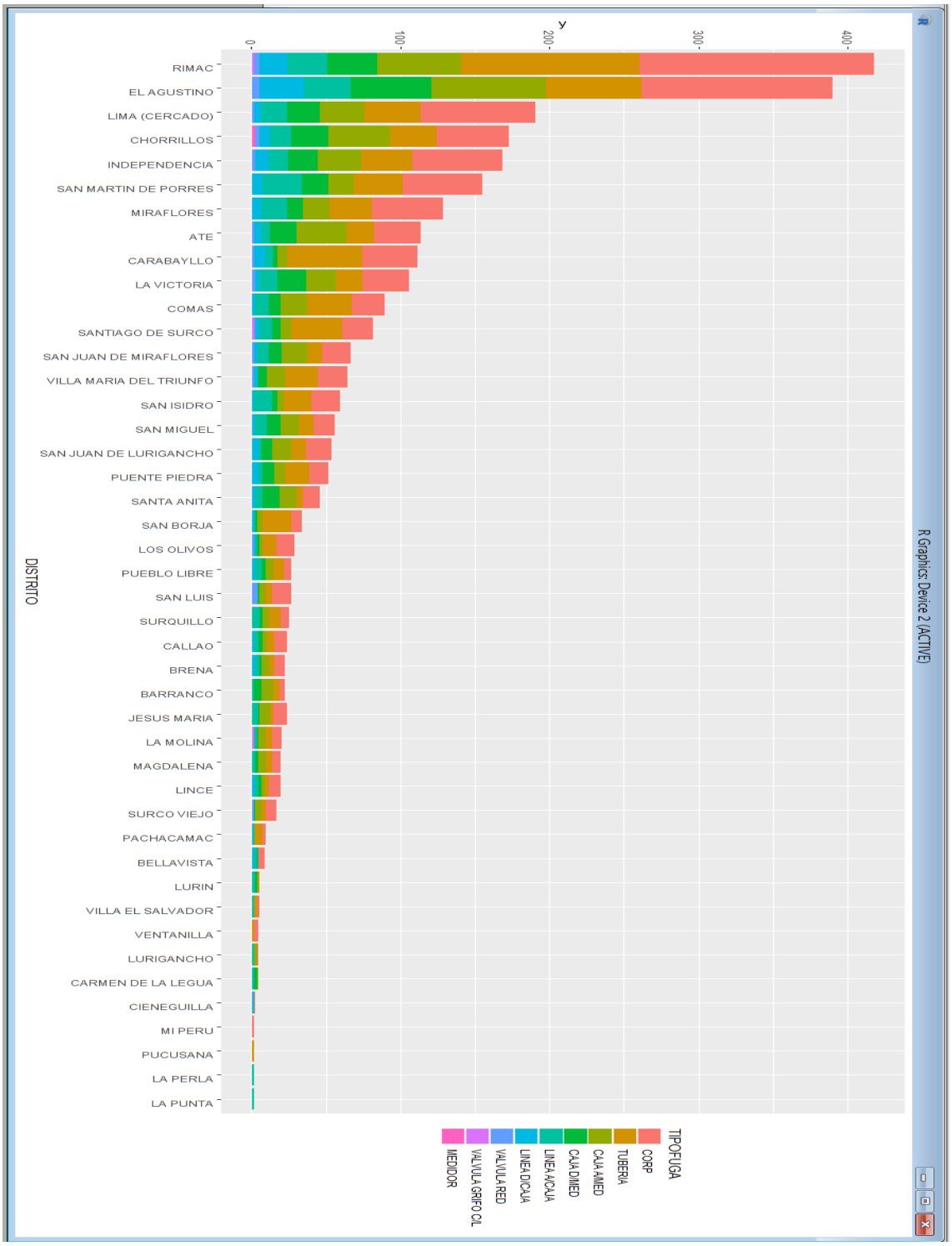
MODELO DEL TIPO BARRAS ORDENADO POR DISTRITO 2014-2019



Fuente: Propia del autor

Figura Nº 60

MODELO DEL TIPO BARRAS ORDENADO POR DISTRITO & TIPOFUGA 2014-2019



Fuente: Propia del autor

Resultados

Tenemos la información de los 44 distritos y de los diferentes tipos de fugas a los largo del periodo 2014-2019 y al haberla procesado con nuestro algoritmo obtuvimos los siguientes resultados luego de aplicar el algoritmo.

Los Distritos con más Apoyos de Emergencia en el periodo 2014-2019 fueron

- RIMAC
- EL AGUSTINO
- LIMA (CERCADO)
- CHORRILLOS
- INDEPENDENCIA
- SAN MARTIN DE PORRES
- MIRAFLORES
- ATE
- CARABAYLLO
- LA VICTORIA

La Fugas de Agua Potable que más se encontraron en el periodo 2014-2019 fueron:

- CORP
- TUBERIA
- CAJA A/MED
- CAJA D/MED
- LINEA A/CAJA

Como hemos observado el servicio de detección fugas depende en gran medida de la cantidad de unidades móviles que hay para el servicio, así como usuarios del servicio de agua potable (usuarios con medidor). En los últimos 3 años, 6 unidades móviles han sido destinadas para dicho fin por lo que su número se mantiene constante, no siendo así el caso de los usuarios del servicio, se estima que solo un 94.5% de las viviendas cuentan con medidor de agua[(1)] , es por eso que las emergencias en el servicio de detección de fugas de agua potable subirá aún más los próximos años al aumentar la cantidad de usuarios con medidor de agua en sus hogares, así se mantengan contante el número de unidades.

Podemos ver el resumen de distritos por emergencias reportadas, figurando el Rímac, El Agustino y Lima Cercado como los distritos con más fugas reportadas a lo largo de esos 4 últimos años.

Tabla N° 2

DISTRITO CON MÁS EMERGENCIAS REPORTADAS 2016-2019

DISTRITOS CON MAS EMERGENCIAS REPORTADAS			
2016	**2017**	**2018**	**2019**
RIMAC	**RIMAC**	**EL AGUSTINO**	**RIMAC**
EL AGUSTINO	**EL AGUSTINO**	**RIMAC**	**EL AGUSTINO**
LIMA CERCADO	**INDEPENDENCIA**	**LIMA CERCADO**	**LIMA CERCADO**
LA VICTORIA	**LIMA CERCADO**	**CHORRILLOS**	**INDEPENDENCIA**
CHORRILLOS	**SAN MARTIN**	**INDEPENDENCIA**	**CHORRILLOS**

[1] FUENTE : SEDAPAL, 2018

Fuente: Propia del autor

Podemos ver el resumen de tipo de fugas por emergencias reportadas figurando como la fuga más reportada la de rotura de corporación, tubería y caja a/ medidor como las más recurrentes en los últimos 4 años.

Tabla N° 3

TIPO DE FUGAS EN EMERGENCIA MÁS REPORTADAS 2016-2019

TIPOS DE FUGAS DE EMERGENCIAS REPORTADAS			
2016	**2017**	**2018**	**2019**
TUBERIA	**CORPORATION**	**CORPORATION**	**CORPORATION**
CORPORATION	**TUBERIA**	**TUBERIA**	**TUBERIA**
CAJA A/ MEDIDOR	**CAJA A/ MEDIDOR**	**CAJA A/ MEDIDOR**	**CAJA A/ MEDIDOR**
CAJA D/MEDIDOR	**CAJA D/MEDIDOR**	**CAJA D/MEDIDOR**	**CAJA D/MEDIDOR**
LINEA A/CAJA	**LINEA A/CAJA**	**LINEA A/CAJA**	**LINEA A/CAJA**

Fuente: Propia del autor

CAPÍTULO VI: DISCUSIÓN Y RESULTADOS

6.1 Contrastación y demostración de la hipótesis con los resultados

Tenemos los resultados de nuestro algoritmo predictivo y como indicamos en un principio, tenemos información la cual no hemos utilizado para este proyecto que guardamos como información de control con la cual procederemos a hacer un modelo con el periodo de esta información que data de Enero-Marzo del 2020.

Lo primero es hacer el filtrado de la información y centrarnos en los campos más relevantes de la tabla para ejecutar análisis predictivo, procedemos a exportar el archivo a Excel para que sea más fácil el filtrado de la información por distrito, por fecha y por tipo de fuga y los demás datos que sean importantes para el filtrado.

La información procesada data de 01/01/2020 al 15/03/2020 correspondiente a los 3 primeros meses del año.

Luego de correr nuestro modelo obtuvimos los siguientes resultados en donde se nota cierta variación en los Distritos del 3ro al 5to que puede deberse a que es solo información concerniente a los 3 primeros meses del 2020, así como a factores debido a que luego de la inspección no se encontró fuga (sin reporte de fuga) por lo que debe recabarse más información en un futuro.

Tabla N° 4

DISTRITO CON MÁS EMERGENCIAS REPORTADAS 2017-2020

DISTRITOS CON MAS EMERGENCIAS REPORTADAS			
2017	**2018**	**2019**	**2020**
RIMAC	**EL AGUSTINO**	**RIMAC**	**RIMAC**
EL AGUSTINO	**RIMAC**	**EL AGUSTINO**	**EL AGUSTINO**
INDEPENDENCIA	**LIMA CERCADO**	**LIMA CERCADO**	**SAN JUAN DE LURIGANCHO**
LIMA CERCADO	**CHORRILLOS**	**INDEPENDENCIA**	**SAN LUIS**
SAN MARTIN	**INDEPENDENCIA**	**CHORRILLOS**	**CHORRILLOS**

Fuente: Propia del autor

Luego de correr nuestro modelo obtuvimos los siguientes resultados en donde se nota que no hay variación con respecto a los tipos de fugas reportados en los años anteriores por lo que el algoritmo de tendencia predictiva es válido en este caso.

Tabla N° 5

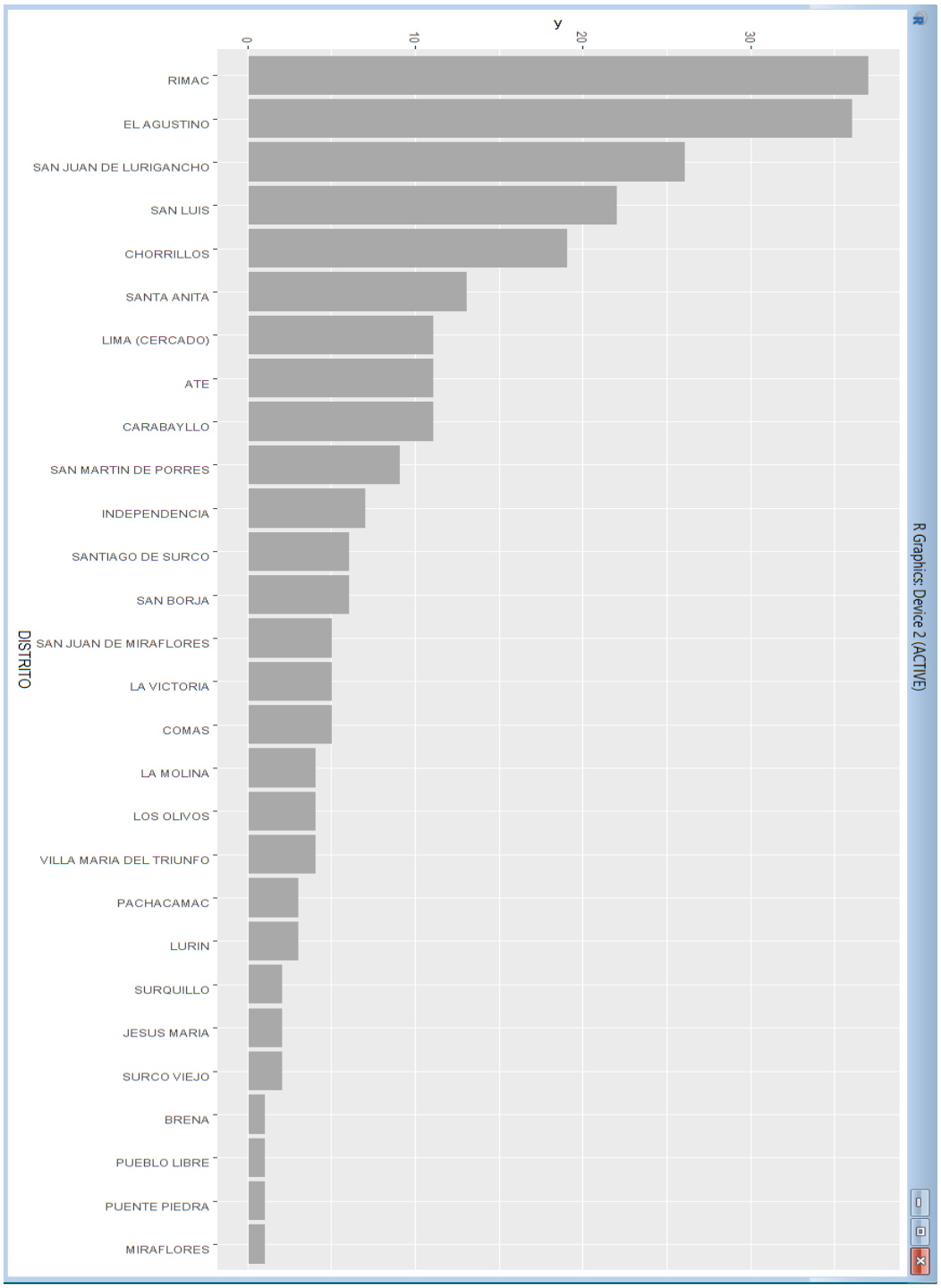
TIPO DE FUGAS EN EMERGENCIA MÁS REPORTADAS 2017-2020

TIPOS DE FUGAS DE EMERGENCIAS REPORTADAS			
2017	**2018**	**2019**	**2020**
TUBERIA	**CORPORATION**	**CORPORATION**	**CORPORATION**
CORPORATION	**TUBERIA**	**TUBERIA**	**TUBERIA**
CAJA A/ MEDIDOR	**CAJA A/ MEDIDOR**	**CAJA A/ MEDIDOR**	**CAJA A/ MEDIDOR**
CAJA D/MEDIDOR	**CAJA D/MEDIDOR**	**CAJA D/MEDIDOR**	**CAJA D/MEDIDOR**
LINEA A/CAJA	**LINEA A/CAJA**	**LINEA A/CAJA**	**LINEA A/CAJA**

Fuente: Propia del autor

Figura Nº 61

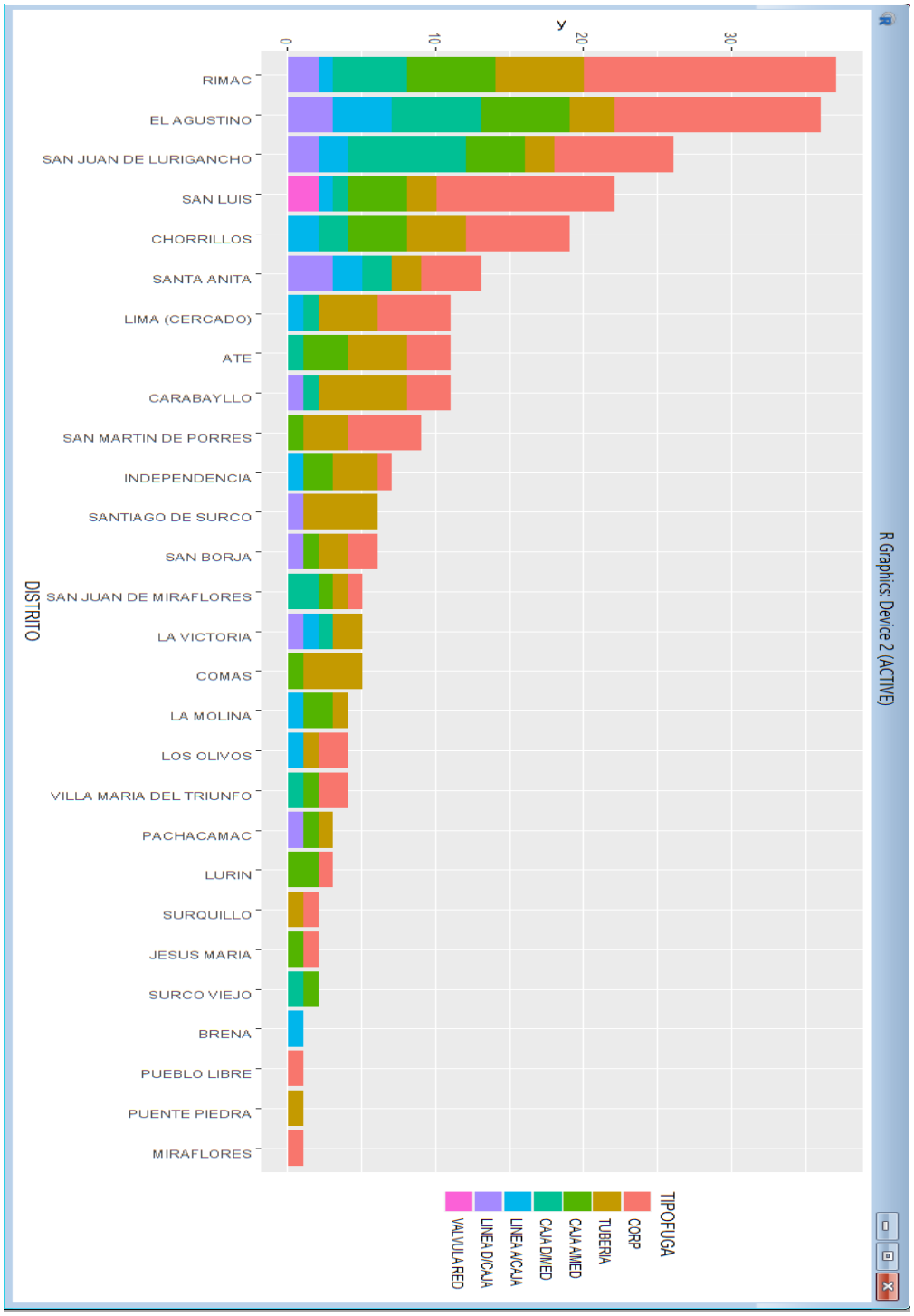
MODELO DEL TIPO BARRAS ORDENADO POR DISTRITO 2020



Fuente: Propia del autor

Figura Nº 62

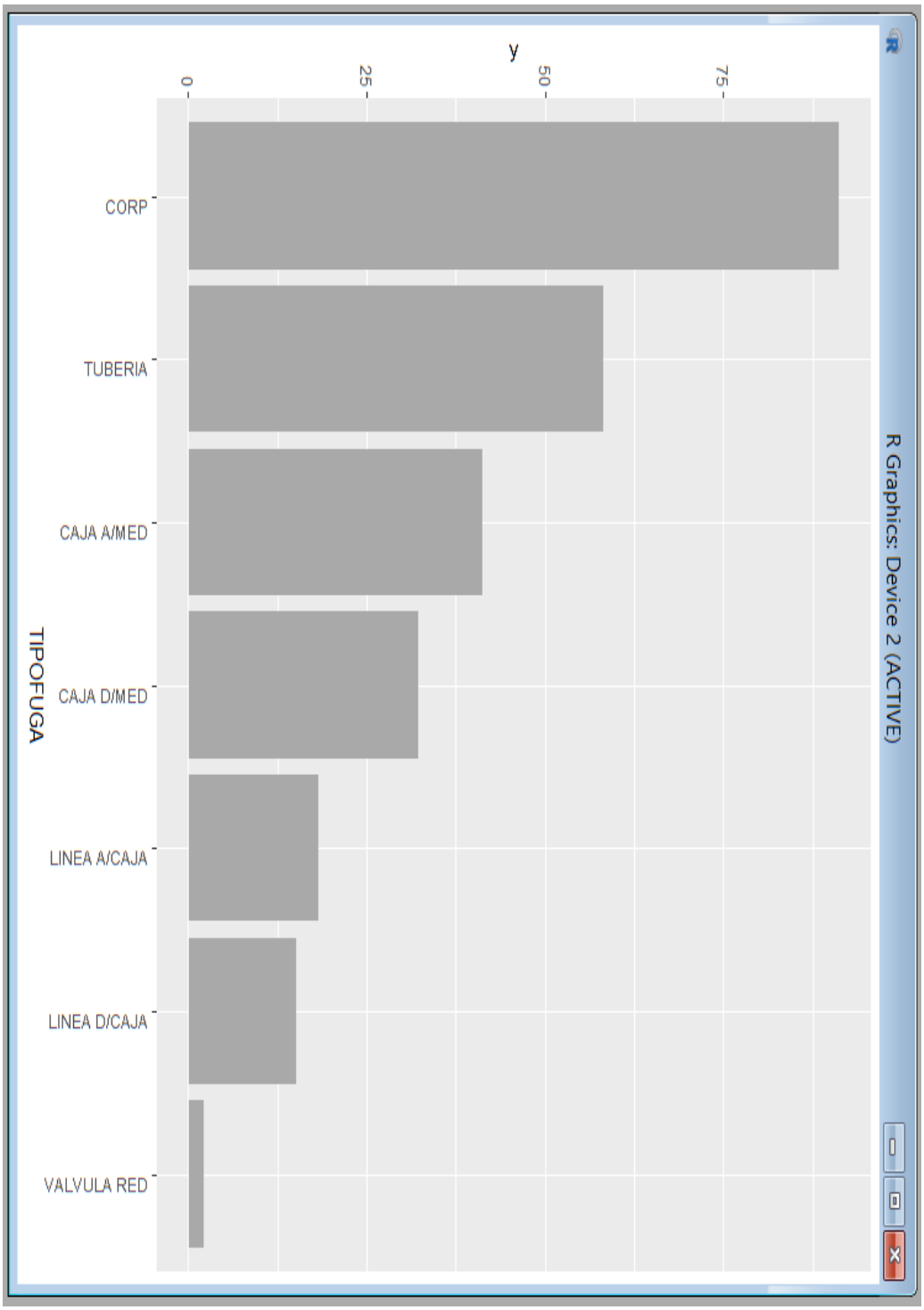
MODELO DEL TIPO BARRAS ORDENADO POR DISTRITO & TIPOFUGA 2020



Fuente: Propia del autor

Figura Nº 63

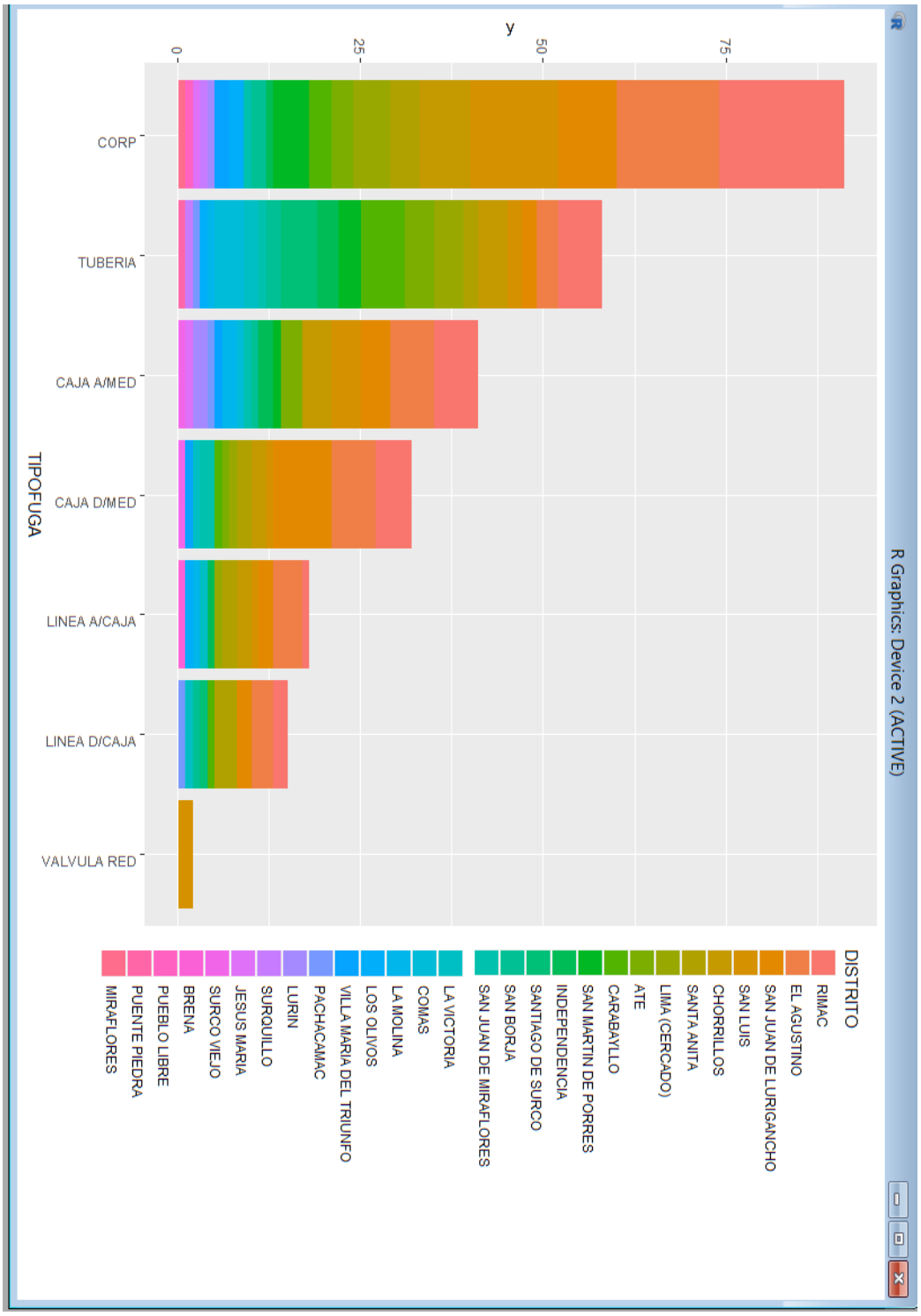
MODELO DEL TIPO BARRAS ORDENADO POR TIPOFUGAS 2020



Fuente: Propia del autor

Figura Nº 64

MODELO DEL TIPO BARRAS ORDENADO POR TIPOFUGA & DISTRITO 2020



Fuente: Propia del autor

CONCLUSIONES

- Se desarrolló un Algoritmo Predictivo en el Software R para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima, 2019 obteniendo resultados prometedores en lo que es la predicción de fugas.
- Se comprobó que un es posible analizando Big Data y árbol de decisiones generar un Algoritmo que pueda predecir en que Distritos se producirán nuevas fugas de agua potable por emergencia y de que tipos serán.
- En esta época en que el agua es un recurso básico para la supervivencia de la raza humana predecir donde se producirán fugas de agua ayudara a su control y reparación inmediata permitiendo reducir el alto porcentaje que bordea el 50% de agua no facturada en la actualidad.
- Sedapal utilizo la información obtenida como un mapa de monitoreo del servicio y que esto permitió organizar y optimizar de una mejor manera el servicio de detección de fugas.
- Este estudio podría mejorar si se conocieran más datos necesarios para las proyecciones, como por ejemplo la cantidad de medidores de agua instalados por año en la ciudad de Lima para estimar la curva de crecimiento del servicio.
- Este tipo de algoritmos predictivos necesitan mucha data con la cual trabajar para obtener resultados mas confiables.

RECOMENDACIONES

Epistemológicamente para el futuro es preciso encomendar se siga investigando sobre este tema para futuras mejoras.

- Desarrollo de un Algoritmo Predictivo en el Software R para Monitoreo Temprano De Redes De Agua Potable En La Ciudad De Lima.
- .El procesamiento de Big Data para estimaciones y proyecciones es un campo que está en constante evolución por lo que esta investigación deberá continuar para mejorar y superar los porcentajes de éxito obtenido.
- Promover el uso de este tipo de herramientas de Big Data por parte de las empresas para poder hacer estimaciones de sus productos y servicios a fin de optimizarlos y producir un ahorro porcentual.
- El ahorro porcentual dependerá en gran medida del compromiso de la empresa prestadora del servicio de agua potable para instaurar este sistema de forma que beneficie a toda la población en Lima y posteriormente replicar esto a todo el Perú.
- Determinar qué factores adicionales podrían encontrase, que afectan este tipo de servicio de detección de fugas de emergencia y buscar su solución.

REFERENCIAS BIBLIOGRÁFICAS

- Andya, A., Y Macy, R., (1996). Pattern Recognition With Neural Networks In C++. Crc Press, Boca Raton, Florida. Usa.

- Apesteguia Infantes y Huarcaya Gonzales (2017) Modelamiento inteligente para la detección de fugas no visibles en las redes de agua potable de la ciudad de Lima (Apesteguia Infantes y Huarcaya Gonzales) Universidad Nacional del Callao, Perú.

- Baldominos Gómez Alejandro (2017), Procesamiento Análisis Inteligente De Big Data, España

- Eric Siegel (2013), Analítica Predictiva: Predecir El Futuro Utilizando Big Data. Anaya Mulltimedia, España

- Espino Timón (2017) Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. Universidad de Cataluña, España.

- Esteban Moro (2013) Big Data Y Análisis Predictivo. Universidad Carlos III, España Vazquez Gomez (2018) Análisis y Diseño De Algoritmos. Red Tercer Milenio, España. Mataix, C. (1982). Mecánica De Fluidos Y Máquinas Hidráulicas. Ediciones Del Castillo, España.

- Fuentes Marfiles, Rodríguez Vázquez y Palma Nava (2011) Estimación y localización de fugas en una red de tuberías de agua potable usando algoritmos genéticos. Ingeniería Investigación y Tecnología. Vol. XII, Núm. 2, Mexico.

- Grandez Marquez (2017) Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales (Grandez Marquez) Universidad San Ignacio de Loyola , Perú.

- Hilera J, Martínez V. 1995. Redes Neuronales Artificiales: Fundamentos, Modelos Y Aplicaciones. Madrid: Addison-Wesley Iberoamericana. Ra-Ma.

- John D. Kelleher, Brian Mac Namee Y Aoife D'arcy (2010) Fundamentals Of Machine Learning For Predictive Data Analytics: Algorithms, Worked Examples, And Case Studies" Mit .

- Laudon, K., Y Laudon, J., (2000). Management Information Systems, Organization And Technology. Cuarta Edición, Editorial Prentice Hall Hispanoamericana.

- Peter Norvig (1992) Paradigms Of Artificial Intelligence Programming: Case Studies In Common Lisp. Morgan Kaufmann Publishers, Estados Unidos

- Ramírez Quintana y Piris Ruano (2015) Una aplicación de minería de datos para el análisis de la propiedad de terminación de SRTs (Ramírez Quintana y Piris Ruano) Universidad Politécnica de Valencia , España.

- Raschka, Sebastian (2015) Python Machine Learning . Pack Open Source

- Rumelhart, D., Et Al., (1986). Learning Internal Representations By Error Propagation. The Mit Press, Cambridge, Masachusetts. Estados Unidos.

- Saldarriaga, J. (1998). Hidráulica De Tuberías, Abastecimiento De Agua, Redes, Riegos. Alfaomega, Colombia.

- Sánchez E, Alanis A. (2006). Redes Neuronales:Conceptos Fundamentales Y Aplicaciones A Control Automático. Madrid. Prentice-Hall.

- Sellés, M., (2001). Optimización De Una Batería De Pruebas Mediante Una Red Neuronal Artificial. Editorial Omega.

- Zapata, S. E. (2009). Fugas En Redes De Distribución De Agua Potable. Universidad Nacional Autónoma De México, Mexico

- Página Web Sedapal, Programa de Educación Sanitaria y Ambiental http://www.sedapal.com.pe/educacion-sanitaria1 consultada en junio del 2019

- Página Web Sedapal, Control de Fugas y Laboratorio Móvil de Fugas No Visibles http://www.sedapal.com.pe/laboratorio-movil consultada en agosto del 2019

- Página Web Sedapal , Plan de Sectorización Consultada en Octubre del 2019 http://www.sedapal.com.pe/c/document_library/get_file?uuid=685267cf-f5dd-4d93-bffc- 8bef04c5b2a6&groupid=10154

- Página Web Sedapal, Plan Maestro de los Sistemas de Agua y Alcantarillado http://www.sedapal.com.pe/contenido/gdi_pmo/anexos/anexo%20e%20planos.pdf Software de análisis estadístico R https://cran.r-project.org/bin/windows/base/ Página Web Sedapal, ECRF http://www.sedapal.com.pe/contenido/cp-0009-2007-jbic-b.pdf y consultada en Julio 2019

- Página Web Empresa Acciona , Centro de Control del AguaConsultada en Junio del 2019 https://www.acciona-agua.com/es/salaprensa/a-fondo/2019/abril/cecoa-gestion- inteligente- agua/

- Página Web WaterWorld Big Value for Big Data in Water Consultada en Octubre del 2019 https://www.waterworld.com/technologies/amr-ami/article/16227142/big-value-for-big- data- in-water

MATRIZ DE CONSISTENCIA

VARIABLE	DEFINICION CONCEPTUAL	DEFINICION OPERACIONAL	DIMENSION	INDICADORES	INSTRUMENTOS
VARIABLE INDEPENDIENTE **X**	**DESARROLLO DE UN ALGORITMO PREDICTIVO**	El modelo predictivo es un modelo de datos, basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a un determinado proceso en este caso la detección de fugas de agua potable de servicio de emergencia	Fugas en redes de agua potable	Cantidad de reporte de fugas	Software Data histórica
			Análisis Predictivo	Desarrollo de el algoritmo predictivo en R	Software Data histórica
VARIABLE DEPENDIENTE **Y**	**DETECCION DE FUGAS DE AGUA POTABLE**	Saber en qué Distritos y que tipos de fugas se han registrado	Ahorro de Agua Potable	Cantidad de Unidades Móviles	Software Data histórica
			Registro de Fugas	Reporte Histórico de solicitudes	Crecimiento de la población

ALGORITMO DESARROLLADO EN SOFTWARE R

```
install.packages("dplyr")
library(dplyr)
install.packages("tidyverse")
library(tidyverse)
install.packages("AppliedPredictiveModeling")
library(AppliedPredictiveModeling)
sedapal<- read_csv("fugas.csv")

sedapal <- fugas %>%
  mutate(
    FECHA = parse_number(week),
    date = date.entered + 7 * (week - 1),
    rank = as.numeric(rank)
    ) %>%
    select(-date.entered)

sedapal_tabla <- as.data.frame(read.csv(file="fugas.csv", header=TRUE,
sep=","))

ggplot(sedapal, aes(x = DISTRITO, y = FECHA, color = DISTRITO,
group = TIPOFUGA)) +
ggplot(data=sedapal_tabla, aes(x=reorder(TIPOFUGA,-TIPOFUGA),
y=1, fill=TIPOFUGA)) +
    geom_bar(stat="identity", position="stack")+
  facet_wrap(~ FECHA, nrow = 9) +
  geom_bar(stat = "identity", fill = "darkgray") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
ggplot(sedapal, aes(x = DISTRITO, y = 1, group =
DISTRITO)) +
facet_wrap(~ FECHA, nrow = 3) +
geom_bar(stat = "identity", fill = "darkgray") +
```

```
theme(axis.text.x = element_text(angle = 90, hjust =
1))

sedapal_tabla$TIPOFUGA = factor(sedapal_tabla$TIPOFUGA,
levels=c("CORP","TUBERIA", "CAJA A/MED", "CAJA D/MED", "LINEA
A/CAJA", "LINEA D/CAJA", "VALVULA RED", "VALVULA GRIFO C/L",
"MEDIDOR" ))

sedapal_tabla$DISTRITO = factor(sedapal_tabla$DISTRITO, levels=c(
"RIMAC","EL AGUSTINO","LIMA (CERCADO)","CHORRILLOS",
"INDEPENDENCIA","SAN MARTIN DE PORRES","MIRAFLORES",
"ATE","CARABAYLLO","LA VICTORIA","COMAS", "SANTIAGO DE
SURCO", "SAN JUAN DE MIRAFLORES","VILLA MARIA DEL
TRIUNFO","SAN ISIDRO","SAN MIGUEL","SAN JUAN DE
LURIGANCHO","PUENTE PIEDRA","SANTA ANITA","SAN BORJA",
"LOS OLIVOS","PUEBLO LIBRE","SAN LUIS","SURQUILLO","JESUS
MARIA","CALLAO","BRENA","BARRANCO","LA MOLINA",
"MAGDALENA","LINCE","SURCO VIEJO","PACHACAMAC",
"BELLAVISTA","LURIN","VILLA EL SALVADOR","VENTANILLA",
"LURIGANCHO","CARMEN DE LA LEGUA","CIENEGUILLA", "MI
PERU","PUCUSANA","LA PERLA","LA PUNTA"
))

ggplot(data=sedapal_tabla, aes(x=DISTRITO, y = 1, )) +
   geom_bar(stat="identity",fill = "darkgray", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))

ggplot(sedapal_tabla, aes(x=TIPOFUGA, y = 1, )) +
   geom_bar(stat="identity",fill = "darkgray", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

```
ggplot(data=sedapal_tabla, aes(x=reorder(TIPOFUGA,-TIPOFUGA),
y=1, fill=TIPOFUGA)) +
   geom_bar(stat="identity", position="stack")

ggplot(data=sedapal_tabla, aes(x=reorder(FECHA,DISTRITO), y=1,
fill=DISTRITO)) +
   geom_bar(stat="identity", position="stack")

ggplot(data=sedapal_tabla, aes(x=reorder(FECHA,-TIPOFUGA), y=1,
fill=DISTRITO)) +
   geom_bar(stat="identity", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))

ggplot(data=sedapal_tabla, aes(x=reorder(DISTRITO,-DISTRITO), y=1,))
+
   geom_bar(stat="identity", fill = "darkgray",position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))

ggplot(sedapal, aes(x=DISTRITO, y = 1, )) +
geom_bar(stat="identity",fill = "darkgray", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(data=sedapal_tabla, aes(x=DISTRITO, y=1, fill=TIPOFUGA)) +
   geom_bar(stat="identity", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
1))
```

```
ggplot(data=sedapal_tabla, aes(x=TIPOFUGA, y=1, fill=DISTRITO)) +
   geom_bar(stat="identity", position="stack")+
theme(axis.text.x = element_text(angle = 90, hjust =
```

```
1))

# MATRIZ DE DISPERSION
library(AppliedPredictiveModeling)
transparentTheme(trans = .4)

library(caret)

featurePlot(x = sedapal_tabla[, 1:3],
            y = sedapal_tabla$TIPOFUGA,
            plot = "pairs",
            ## Add a key at the top
            auto.key = list(columns = 3))
```