



Universidad Cenfotec

Maestría en Tecnología de Bases de Datos

Documento Final del proyecto de Investigación Aplicada 2

Evaluación de la información recolectada por medio de modelos de minería de datos, para mejorar el uso de los recursos del acueducto del cantón de Belén

Johnnatan Chaves Calvo

Diciembre 2017

DECLARATORIA DE DERECHOS DE AUTOR

Yo, Johnnatan Chaves Calvo, número de identificación 1-1040-0040, estudiante de la Universidad Cenfotec, de la carrera Maestría en Tecnologías de Bases de Datos, autorizo a la Universidad Cenfotec, a sus estudiantes activos y cualquier ente gubernamental que utilice el presente documento para ser consultado única y exclusivamente con fines académicos y con el espíritu de la búsqueda beneficio de las comunidades de nuestro país.

Johnnatan Chaves Calvo

DEDICATORIA

A mis padres, por mostrarme que con el ejemplo del trabajo duro y disciplina se pueden superar y conquistar los retos que se presenten.

A mi esposa e hijos por ser siempre mi inspiración y brindarme el tiempo requerido para completar los logros que ahora disfrutamos juntos.

TABLA DE CONTENIDO

Resumen Ejecutivo	xiv
Capítulo 1. Introducción	1
1.1 Generalidades.....	1
1.2 Antecedentes del problema.....	1
1.3 Definición y descripción del problema.....	1
Áreas de recarga	3
Pozos	3
Manantiales	3
1.4 Justificación	4
1.5 Viabilidad.....	5
1.5.1 Punto de vista técnico.....	5
1.5.2 Punto de vista operativo	5
1.5.3 Punto de vista económico.....	5
1.6 Objetivos	8
1.6.1 Objetivo general	8
1.6.2 Objetivos específicos.....	8
1.7 Alcances y limitaciones.....	9
1.7.1 Alcances	9

1.7.2 Limitaciones.....	10
1.8 Marco de referencia organizacional y socioeconómico.....	10
1.8.1 Historia.....	10
1.8.2 Tipo de negocio y mercado meta.....	11
1.8.3 Misión, visión y valores.....	12
1.8.4 Políticas institucionales.....	13
1.9 Estado de la cuestión.....	15
1.9.1 Formulación de la pregunta.....	15
1.9.2 Calidad y amplitud de la pregunta.....	15
1.9.3 Selección de fuentes.....	18
1.9.4 Evaluación después de las fuentes seleccionadas.....	20
1.9.5 Verificación de referencias.....	20
1.9.6 Selección de los estudios.....	20
1.9.7 Ejecución de selección en los sitios web de información.....	22
Capítulo 2. Marco teórico o conceptual.....	32
2.1 CRISP-DM.....	33
Comprensión del negocio.....	33
Comprensión de datos.....	34
Preparación de datos.....	34
Modelado.....	34

Evaluación.....	34
Despliegue.....	35
2.2 Clustering.....	35
2.3 Algoritmo Kmeans.....	36
2.4 Series de tiempo.....	37
Capítulo 3. Marco metodológico.....	39
3.1 Tipo de investigación.....	39
3.2 Alcance investigativo.....	39
3.3 Enfoque.....	40
3.4 Diseño.....	40
3.5 Población y muestreo.....	41
3.6 Instrumentos de recolección de datos.....	42
3.7 Técnicas de análisis de la información.....	42
Fase I. Comprensión del negocio.....	43
Fase II. Estudio y comprensión de los datos.....	43
Fase III. Análisis de los datos y selección de características.....	44
Fase IV. Modelado.....	44
Fase V. Evaluación de resultados.....	44
Fase VI. Despliegue y puesto en producción.....	45
3.8 Estrategia de desarrollo de la propuesta.....	46

Capítulo 4. Análisis del diagnóstico	47
Capítulo 5. Propuesta de solución	51
5.1 Carga de datos.....	51
5.2 Optimización en las consultas.....	52
5.3 Exploración de los datos.	54
5.3.1 Promedio consumo en el año 2015	55
5.3.2 Grandes consumidores 2014 vs 2015.....	56
5.3.3 Cantidad de consumidores por tipo de paja	57
5.4 Clustering.....	58
5.4.1 Segmentación Belén.....	58
5.4.2 Segmentación acueducto de Zamora.....	64
5.4.3 Segmentación acueducto de Mangos	68
5.4.4 Segmentación acueducto de Cariari	72
5.5 Series de tiempo.....	76
5.5.1 Generación de las series	77
5.5.2 Explorando las series de tiempo.....	78
5.5.2 Analizando las series de tiempo de todo Belén.....	81
5.5.3 Generando predicciones en la serie de tiempo de Belén.....	85
5.6 Analizando la serie de tiempo del acueducto de Zamora	96
5.6.1 Generando predicciones en la serie de tiempo de Zamora.....	101

5.6.2 Selección del modelo predictivo del acueducto de Zamora	103
5.6.3 Modelo predictivo en el acueducto de Zamora para el año 2017.....	104
5.7 Analizando la serie de tiempo del acueducto de Mangos	105
5.7.1 Generando predicciones en la serie de tiempo de Mangos	110
5.7.2 Selección del modelo predictivo del acueducto de Mangos	112
5.7.3 Modelo predictivo en el acueducto de Mangos para el año 2017	114
5.8 Analizando la serie de tiempo del acueducto de Cariari.....	115
5.8.1 Generando predicciones en la serie de tiempo de Cariari	120
5.8.2 Selección del modelo predictivo del acueducto de Cariari	122
5.8.3 Modelo predictivo en el acueducto de Cariari para el año 2017	123
Capítulo 6. Conclusiones y recomendaciones	125
6.1 Conclusión sobre la selección los datos requeridos para la construcción de modelos de minería de datos.	125
6.2 Conclusión sobre aplicar el modelo de minería de datos de series de tiempo en la evaluación del uso y abastecimiento del recurso hídrico.....	125
6.3 Conclusiones del aplicar modelos de minería k-means en el agrupamiento de las comunidades	127
6.4 Entrega a los encargados del acueducto de Belén los resultados del proyecto, así como las recomendaciones en la optimización de los recursos	130
6.5 Recomendaciones	131
Capítulo 7. Reflexiones finales.....	132

Capítulo 8. Trabajos a futuro.....	134
Glosario	136
Referencias	138
Anexo A.....	141
Anexo B.....	142
Anexo C.....	143
Anexo D.....	144

ÍNDICE DE FIGURAS

Ilustración 1 Mapa conceptual del proyecto.....	32
Ilustración 2 Metodología CRISP-DM.....	33
Ilustración 3 Algoritmo Kmeans	36
Ilustración 4 Series de Tiempo	37
Ilustración 5 Diseño exploratorio del proyecto	41
Ilustración 6 Fases del modelo CRISP-DM	43
Ilustración 7 Calificación gestión municipal 2016.....	48
Ilustración 8 Archivo muestra	49
Ilustración 9 Carga de datos	51
Ilustración 10 Salvando los datos	52
Ilustración 11 Índices generados	53
Ilustración 12 Vistas creadas para las consultas.....	53
Ilustración 13 Total de consumidores por tipo	54
Ilustración 14 Consumo promedio por acueducto.....	55
Ilustración 15 Grandes consumidores del 2014.....	56
Ilustración 16 Grades consumidores del 2015.....	56
Ilustración 17 Consumidores por tipo	57
Ilustración 18 Selección datos para segmentación	59
Ilustración 19 Gráfico de los K para Belén	60
Ilustración 20 Gráfico clúster de Belén	62
Ilustración 21 Gráfico radar clusters Bélen	63

Ilustración 22 K=5 acueducto Zamora	65
<i>Ilustración 23 K=5 clúster Zamora.....</i>	<i>66</i>
Ilustración 24 Gráfico radar Zamora	67
Ilustración 25 K=5 del acueducto Mangos	69
Ilustración 26 clúster Mangos.....	70
Ilustración 27 Gráfico radar Mangos.....	71
Ilustración 28 K=5 del acueducto Cariari.....	73
Ilustración 29 clúster Cariari	74
Ilustración 30 Gráfico radar Cariari.....	75
Ilustración 31 Series de tiempo generadas	77
Ilustración 32 Uso de vista viseriesacueductos	78
Ilustración 33 Consumo M3 mensual en Belén.....	80
Ilustración 34 Comportamiento de los datos de la serie de tiempo en Belén.	82
Ilustración 35 Análisis serie de tiempo Belén	83
Ilustración 36 Frecuencia de mayor consumo	85
Ilustración 37 Resultado autoarima	89
Ilustración 38 Gráfico radar modelos series de tiempo Belén.....	92
Ilustración 39 Resultados de los modelos predictivos en Belén.....	93
Ilustración 40 Gráfico predicción consumo 6 meses de 2017 en Belén.....	96
Ilustración 41 Selección de los datos del acueducto de Zamora	97
Ilustración 42 Gráfico consumo acueducto Zamora.....	98
Ilustración 43 Comportamiento de los datos de la serie de tiempo en Zamora.....	99
Ilustración 44 Análisis serie de tiempo Zamora	99
Ilustración 45 Frecuencia de mayor consumo en Zamora	100

Ilustración 46 Gráfico radar modelos series de tiempo acueducto de Zamora.....	103
Ilustración 47 Validando los modelos predictivos para el acueducto de Zamora	104
Ilustración 48 Gráfico predicción consumo 6 meses de 2017 en Zamora.....	105
Ilustración 49 Gráfico consumo acueducto Mangos	107
Ilustración 50 Comportamiento de los datos de la serie de tiempo en Mangos.	108
Ilustración 51 análisis serie de tiempo Mangos	109
Ilustración 52 Frecuencia de mayor consumo en Mangos	110
Ilustración 53 Gráfico radar modelos series de tiempo acueducto de Mangos	113
Ilustración 54 Validando los modelos predictivos para el acueducto de Mangos	114
Ilustración 55 Gráfico predicción consumo 6 meses de 2017 en Mangos	115
Ilustración 56 Selección de los datos del acueducto de Cariari.....	116
Ilustración 57 Gráfico consumo acueducto Cariari	117
Ilustración 58 Comportamiento de los datos de la serie de tiempo en Cariari.	117
<i>Ilustración 59 análisis serie de tiempo Cariari.....</i>	<i>118</i>
Ilustración 60 Frecuencia de mayor consumo en Cariari	119
Ilustración 61 Gráfico radar modelos series de tiempo acueducto de Cariari	122
Ilustración 62 Validando los modelos predictivos para el acueducto de Cariari.....	123
Ilustración 63 Gráfico predicción consumo 6 meses de 2017 en Cariari	124
Ilustración 64 Resultado Modelos Predictivos Belén.....	126
Ilustración 65 Comparación entre las segmentaciones.....	128
Ilustración 66 Rangos consumo.....	129
Ilustración 67 Equivalente grandes consumidores hogares.....	130

ÍNDICE DE TABLAS

Tabla 1 Costos del proyecto	7
Tabla 2 Selección de estudios.....	22
Tabla 3 Estudios seleccionados.	24
Tabla 4 Formato extracción de los datos del estudio.....	25
Tabla 5 Información primer Estudio	25
Tabla 6 Información segundo Estudio.....	27
Tabla 7 Información tercer estudio.....	28
Tabla 8 Resumen de estudios relevantes	29
Tabla 9 Comparación de los estudios y métodos de minería.	31
Tabla 10 Características de los clúster	61
Tabla 11 Consumidores clúster #2 Belén.....	64
Tabla 12 Características de los clúster de Zamora	¡Error! Marcador no definido.
Tabla 13 Grandes consumidores del acueducto de Zamora	68
Tabla 14 Características clúster Mangos	70
Tabla 15 Grandes Consumidores Mangos.....	72
Tabla 16 Características clúster Cariari.....	74
Tabla 17 Grandes consumidores Cariari	76
Tabla 18 Evaluación de los modelos.	127

RESUMEN EJECUTIVO

Aunque la humanidad reconoció hace mucho tiempo su dependencia del agua y la ha catalogado como un recurso imprescindible, es hasta hace poco que se ha sufrido las consecuencias del uso desmedido de este recurso. A esto se debe sumar el impacto del cambio climático, el cual limita su abastecimiento.

Por esta razón, es urgente y necesaria una gestión responsable, enfocada en brindar la debida protección del agua, tomando en cuenta que es un recurso natural esencial para nuestra generación y las generaciones futuras. Sin agua, no puede haber vida.

La contaminación del agua y su escasez plantean amenazas para la salud humana y la calidad de vida, pero su incidencia ecológica es más general. La falta de agua de buena calidad perjudica al medio acuático, húmedo y terrestre. Somete a una presión mayor a la flora y la fauna, que padecen las repercusiones de la urbanización y el cambio climático.

Es por esto que el cantón de Belén posee un departamento asignado y responsable del acueducto municipal el cual se encarga de suplir el recurso hídrico a todo el cantón. Belén cuenta con un área territorial de 11.81 km², divididos en 3 distritos: San Antonio, La Ribera y La Asunción.

Sus límites geográficos son: al Este con los cantones de Heredia y Flores, al Norte y Oeste con el cantón de Alajuela y al Sur con los cantones de San José, Escazú y Santa Ana.

El cantón de Belén tiene una población aproximada de 22.530 habitantes, ubicados en 5.201 viviendas ocupadas.

El proyecto busca utilizar técnicas de minería de datos¹ para investigar cuáles modelos se pueden implementar para optimizar el uso y almacenaje del recurso hídrico. Se utilizará la información que facilitarán los personeros de la municipalidad del cantón de Belén sobre de almacenamiento y consumo del agua.

Con estos datos y modelos de minería se busca ser más eficientes en la proyección y en el abastecimiento necesario para época seca, así como prever el requerimiento de agua para el incremento de la población en el cantón.

¹ La minería de datos posee técnicas para detectar información valiosa en grandes conjuntos de datos.

Capítulo 1. Introducción

1.1 Generalidades

La administración del acueducto municipal del cantón de Belén posee procesos técnicos y administrativos que involucran la preservación del ambiente, así como la construcción e implementación de nuevas obras para la Comunidad Belemita, cuyo objetivo es producir agua de calidad y brindar un servicio de abastecimiento de excelencia a la población de Belén.

1.2 Antecedentes del problema

El cambio climático, así como el crecimiento acelerado en de la industria en este cantón, hacen que el proceso de producción del agua potable sea una tarea cada vez más compleja. La construcción de pozos y de tanques para el almacenamiento del agua requiere del uso métodos más precisos para garantizar la eficiencia en almacenaje recolección, así como la inversión necesaria en infraestructura para no dimensionar erróneamente los requerimientos de cada comunidad del cantón.

1.3 Definición y descripción del problema

Desde que se estableció el acuerdo de ley 1634 que define el uso y manejo del Agua Potable en el año 1953, en el artículo 5 se indica lo siguiente:

Las Municipalidades tendrán a su cargo la administración plena de los sistemas de abastecimiento de aguas potables que están bajo su competencia (Autoridad Reguladora de los servicios Públicos, 2017).

Con lo que se decretó que los municipios del país tienen la responsabilidad de velar por la entrega de este recurso en servicios del agua potable, así como el saneamiento del valioso líquido. Las municipalidades han tenido que implementar diferentes planes para regular y administrar de manera correcta este recurso.

El cambio climático ha generado un desbalance en el comportamiento de las lluvias durante el año lo cual ha forzado al racionamiento y corte por largos lapsos, lo que afecta a escuelas, centros médicos y comunidades de todo el país.

Es por esto que, en el 2017, en el comunicado de la Presidencia de la República del 20 de febrero, se hizo énfasis en los problemas de desabastecimiento que posiblemente se presentarán entre marzo y julio del presente año. Por esta razón, Acueductos y Alcantarillados (AyA) pretende contrarrestar el déficit del agua en varias de las regiones afectadas.

Cabe destacar que el sistema hidrológico del cantón de Belén pertenece a la cuenca del Río Grande de Tárcoles, perteneciente a la vertiente del Pacífico, además viven cerca de 1.600.000 habitantes, (el 50% de la población del país) y se ubica el 80% de la industria, el comercio y los servicios.

Aunque el Cantón de Belén cuenta con diferentes áreas de recarga, así como manantiales y pozos para la extracción y almacenamiento del agua en distintas zonas del cantón, todos estos son afectados por la falta de caudal que se pueda presentar en las siguientes fuentes.

Áreas de recarga

1. Áreas Favorables para la Recarga: se ubica principalmente en el distrito de la Ribera, en el extremo este. Nótese que en esta zona está el manantial de Ojo de Agua.
2. Áreas Moderadas Poco Favorables: comprenden la mayor cantidad de terreno del distrito La Ribera y una pequeña sección noreste del distrito la Asunción. Para el caso de San Antonio, no se presenta terreno en esta categoría.
3. Áreas desfavorables para la recarga: abarca casi en su totalidad el distrito de San Antonio, salvo al noroeste, mientras que en la Asunción abarca casi todo el distrito salvo al noreste. Para el distrito de la Ribera, se presenta una pequeña sección al sur (Laboratorio de Análisis Ambiental, 2017).

Pozos

Los pozos de extracción de agua en el cantón de Belén se encuentran ubicados en la zona industrial abarcando los distritos de la Ribera y Asunción.

El distrito de San Antonio se identificó como la zona que cuenta con menor cantidad de pozos para el abastecimiento y uso de los consumidores.

Manantiales

Belén cuenta con cinco manantiales en su territorio, dos de ellos se ubican en el distrito de la Ribera. En el caso de San Antonio el único es el de Puente de Mulas. Por último, está la Asunción con dos manantiales, el primero San Antonio ubicado al noroeste, mientras en la parte central cercana al convento está el de La Gruta.

1.4 Justificación

El uso del presupuesto municipal, así como emplear recursos públicos para la construcción de obras para el beneficio de los ciudadanos para el almacenamiento del agua o bien distribución del recurso hídrico, deben cumplir con requisitos bastante rígidos y a su vez es necesario que beneficien de la forma esperada a los vecinos del cantón.

Para este proyecto se utilizaron los datos recopilados por los responsables del acueducto de la municipalidad de Belén, la información que se facilitó contiene los datos relacionados al consumo del agua en las diferentes zonas de Belén. Los datos fueron utilizados con el fin de generar modelos de minería de datos mediante la metodología CRISP-DM². Se aplicaron los métodos predictivos de series de tiempo³ para proyectar el consumo para periodos de tiempo determinados, a partir de los datos almacenados y su frecuencia de lectura.

El uso de estos modelos permitirá tener bases para invertir de forma adecuada el recurso económico administrado por la municipalidad, con lo que se implementarán las obras requeridas en las zonas que realmente las necesitan. Así no se duplicará infraestructura que requiere espacio y mantenimiento, con lo que se reducirán costos y se brindará el servicio de forma eficiente.

Adicionalmente, el set de datos fue utilizado para aplicar modelos de minería de datos de segmentación como K-Means⁴. Se logró segmentar los datos relacionados según los requerimientos de consumo en los distritos y sus consumidores por su comportamiento en el

²Cross-industry standard process for data mining

³Conjunto de datos agrupados cronológicamente para su análisis en predicción y pronóstico.

⁴Método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos

consumo del agua. Con esto se pudo obtener una segmentación, la cual permite estimar y clasificar los requerimientos futuros en el crecimiento de cada uno de los distritos.

1.5 Viabilidad

1.5.1 Punto de vista técnico

La municipalidad de Belén cuenta con herramientas para la extracción de información de sus zonas de recarga, así como manantiales y pozos. Además, tiene suficientes recursos para cubrir las necesidades de almacenamiento físico y lógico de información, así como para cubrir las necesidades de implementación del proyecto de minería de datos.

1.5.2 Punto de vista operativo

Para la realización de este proyecto se tendrá acceso a los datos por medio de copias digitales, por lo que no es necesario el acceso a redes y otros recursos físicos de la institución. Adicionalmente, se trabajará con una persona de la municipalidad, la cual proporcionará la información requerida para llevar a cabo este proyecto.

1.5.3 Punto de vista económico

En el caso del nivel económico, no se considera necesario realizar una inversión monetaria debido a que este proyecto no será facturado a la municipalidad.

Para desarrollar un proyecto de esta magnitud se requiere una serie de cualidades y destrezas profesionales con habilidades y funciones específicas en el proyecto.

Project manager

Como en todo proyecto es ideal contar con un líder que permita y facilite la interacción entre los diferentes recursos y etapas del mismo.

Aparte de anticipar posibles retos y llevar un control del avance de cada tarea el administrador del proyecto se enfoca en el éxito apoyándose en este caso en la metodología SCRUM⁵.

DBA: las funciones ejecutadas por el DBA⁶ en este proyecto están relacionadas con

- Implementación del motor de base de datos.
- Generación de las tablas.
- Roles de Seguridad.
- Creación de Vistas para facilitar las consultas de los datos.
- Optimizar las consultas mediante el uso de índices.

Analytics consultant

El rol del consultor de Analytics⁷ está enfocado en aplicar modelos y métodos, en este caso con el Lenguaje R, sobre los datos que se cargaron previamente las bases de datos. Además, debe entender los requerimientos del negocio para el modelado de los datos y validar sus resultados, debe ser capaz de demostrar la validez y modificar lo necesario luego de consultarlo con el experto del negocio para mejorar la certeza de sus proyecciones.

⁵ Metodología ágil para el manejo de proyectos.

⁶ Administrador de Base de Datos.

⁷ Uso intensivo de datos, estadística y análisis cuantitativo, modelos predictivos y explicativos y gestión basada en hechos para dar soporte al proceso de toma de decisiones

Tableau professional

Resumir los resultados de forma visual requiere de técnicas y herramientas esenciales, las cuales permitirán mostrar los resultados de todo el proyecto en *dahsboards*⁸. La selección de gráficos debe ser cuidadosa para no causar confusión, además de resaltar los datos más relevantes.

Tabla 1 Costos del proyecto

Rol	Costo por hora	Tarea	Tiempo del proyecto	Horas	Total
DBA	\$ 45,00	Preparación de los datos	60%	\$180,00	\$ 8 100,00
Project Manager	\$ 50,00	Reuniones	5%	\$ 15,00	\$ 750,00
Analytics Consultant	\$ 65,00	Modelado y evaluación	25%	\$ 75,00	\$ 4 875,00
Tableau Professional	\$ 55,00	Visualización	10%	\$ 30,00	\$ 1 650,00
			100%		\$15 375,00

Hardware

Con respecto al hardware⁹ para esta propuesta se utilizarán las máquinas de cómputo con las que cuenta el equipo investigador.

Software

El *software*¹⁰ seleccionado en su mayor parte es *software* libre:

Base de datos MySQL se seleccionó esta base de datos para manipular y almacenar los datos facilitados para el proyecto.

⁸ Tableros Gráficos de Resumen.

⁹ Conjunto de elementos físicos o materiales que constituyen una computadora.

¹⁰ Conjunto de programas y rutinas que permiten a la computadora realizar determinadas tareas.

RStudio¹¹ está interfaz de programación del lenguaje de programación R cuenta con distintas librerías las cuales fueron utilizadas para aplicar métodos específicos de minería de datos.

Al utilizar este *software* se logra disminuir costos, además de evitar realizar algún tipo de licitación para la compra de máquinas o *software*.

1.6 Objetivos

Se selecciona la taxonomía de Bloom, debido a que sirve para organizar de manera clara y simple actividades de formación y propósitos, objetivos o metas de aprendizaje relacionados con el desarrollo de habilidades.

1.6.1 Objetivo general

Evaluar la información recolectada con el uso de modelos de minería de datos aplicando la metodología CRISP-DM, como una opción viable para mejorar el uso de los recursos del acueducto del cantón de Belén.

1.6.2 Objetivos específicos

- Escoger los datos requeridos para la construcción de modelos de minería de datos.
- Aplicar el modelo de minería de datos de series de tiempo en la evaluación del uso y abastecimiento del recurso hídrico.

¹¹ Interfaz de Programación del lenguaje R.

- Aplicar modelos de minería k-means en la evaluación y agrupamiento de las comunidades por características similares para el mejoramiento de los resultados de los modelos.
- Analizar los resultados de los modelos, así como las recomendaciones que se pueden hacer para un mejor uso de los recursos hídricos.
- Proponer a los encargados del acueducto de Belén el uso de los modelos evaluados para su consideración.

1.7 Alcances y limitaciones

1.7.1 Alcances

- Los modelos de minería seleccionados se confeccionarán con los datos suministrados por los responsables del acueducto de la Municipalidad de Belén.
- Se realizará un análisis de las variables en los datos, así como la evaluación del impacto de las variables más significativas para crear el modelo predictivo.
- Uso de la metodología CRISP-DM para la generación de los modelos.
- Creación los modelos de minería con mejor desempeño para que la municipalidad pueda hacer uso de ellos utilizando el lenguaje R.
- Por último, se brindará una presentación a los encargados del acueducto municipal con el análisis de los resultados del uso de los modelos y recomendación para futuros usos por parte de los encargados en la municipalidad.

1.7.2 Limitaciones

- Los resultados que se entregan como modelos brindan una probabilidad estadística en caso de materializar su uso.
- La implementación futura y uso de nuevos datos será responsabilidad de los personeros de la municipalidad de Belén.
- Se utilizarán únicamente datos limpios y en formato digital.
- No se darán recomendaciones para la construcción de obras o adquisición de equipo.

1.8 Marco de referencia organizacional y socioeconómico

1.8.1 Historia

En 1907 se iniciaron las gestiones para lograr el cantonato de Belén, un grupo de vecinos fueron los encargados de promover, dicha gestión, apuntando como principal motivo el desarrollo, económico de San Antonio de Belén. Esto gracias a que por este lugar se presentaba el comercio por el ferrocarril al Atlántico y al Pacífico.

Según el artículo 1 del código municipal, se puede afirmar que un municipio está constituido por un conjunto de personas vecinas residentes en un mismo cantón, que promueven y administran sus propios intereses, por medio del gobierno municipal.

Es por esto que oficialmente se realizó la oficialmente la primera sesión municipal, el 15 de julio de 1915.

Belén es el cantón número 7 de la provincia de Heredia. La provincia forma parte de la Gran Área Metropolitana, 25% de la superficie del cantón es área habitacional. Tiene una

superficie territorial de 12.15 km² y contiene 3 distritos. Entre sus límites están: al norte y oeste con el cantón de Alajuela, al este con Flores y Heredia y al sur con San José, Escazú y Santa Ana. Belén fue fundado el 6 de junio de 1907. Su cabecera es el distrito de San Antonio de Belén (Belén, s. f.).

1.8.2 Tipo de negocio y mercado meta

Las municipalidades son personas jurídicas a las que se les atribuyen derechos y obligaciones para el cumplimiento de la administración de los intereses y servicios locales.

Sus facultades emanan de la Constitución y de las normas legales para dirigir el gobierno municipal según los intereses locales.

- Dictar los reglamentos autónomos de organización y de servicio, así como cualquier otra disposición que autorice el ordenamiento jurídico.
- Acordar sus presupuestos y ejecutarlos.
- Administrar y prestar los servicios públicos municipales.
- Aprobar las tasas, los precios y las contribuciones municipales, así como proponer los proyectos de tarifas de impuestos municipales.
- Percibir y administrar, en su carácter de administración tributaria, los tributos y demás ingresos municipales.
- Concertar, con personas o entidades nacionales o extranjeras, pactos, convenios o contratos necesarios para el cumplimiento de sus funciones.

- Convocar al municipio a consultas populares, para los fines establecidos en esta Ley y su Reglamento.
- Promover un desarrollo local participativo e inclusivo, que contemple la diversidad de las necesidades y los intereses de la población.
- Impulsar políticas públicas locales para la promoción de los derechos y la ciudadanía de las mujeres, en favor de la igualdad y la equidad de género.

1.8.3 Misión, visión y valores

Misión

Somos una institución autónoma territorial que promueve el desarrollo integral y equitativo, administra servicios de manera innovadora, eficiente y oportuna, con el propósito de contribuir al bienestar de sus habitantes.

Visión

Ser una institución que, mediante un desarrollo integral, equitativo y equilibrado, garantice el bienestar de sus habitantes (Municipalidad de Belén, s. f.).

Valores

Trabajo en equipo: fomentar una cultura participativa e integradora de esfuerzos donde el resultado es el producto del aporte de todos.

Actitud de servicio: ofrecer soluciones oportunas y eficaces a los usuarios internos y externos a la institución.

Honradez: ser íntegro y honesto en cada una de las actividades que realizamos y estar siempre dispuestos a rendir cuentas de nuestros actos.

Solidaridad: tener una actitud y disposición permanente orientada a las necesidades de la población.

Equidad: garantizar un trato justo y equilibrado en la gestión institucional.

Responsabilidad: cumplimiento de nuestros deberes y responsabilidad de forma oportuno y eficaz.

Lealtad: mantener una actitud de entrega de respecto a la institución completa cumpliendo fielmente las políticas, lineamientos, directrices, acuerdos de la Municipalidad de Belén en franca protección del bienestar de la comunidad Belemita.

Transparencia (Municipalidad de Belén, s. f.).

1.8.4 Políticas institucionales

Políticas, Premisas y Principios Fundamentales:

Se establece oficial y formalmente el seguimiento de los acuerdos, lineamientos, políticas, directrices, priorizaciones y planes ya aprobados por Concejo Municipal como base fundamental e integral.

Participación ciudadana: garantizar la participación propositiva y activa de la ciudadanía en el desarrollo del cantón de Belén.

Desarrollo humano: La acción institucional estará orientada prioritariamente a contribuir con una mejor calidad de vida de los habitantes del cantón y sostenibilidad.

Calidad: garantizar la excelencia de los servicios públicos y satisfacer las necesidades de los pobladores del cantón de Belén.

Desarrollo integral: el desarrollo del cantón se dará en armonía con el ambiente y el bienestar de la población.

Eficiencia administrativa: maximizar el uso de los recursos a través del incremento en la productividad y la racionalidad del gasto.

Igualdad de oportunidades: el quehacer institucional garantiza el acceso de toda la población en igualdad de condiciones y oportunidades, a los servicios que brinda la Municipalidad.

Información y comunicación: promover estrategias de información y comunicación tanto a lo interno como a lo externo, veraz y oportuna para la toma de decisiones, para crear opinión en la comunidad sobre el desarrollo de nuestro cantón y el quehacer municipal.

Transparencia: apertura a brindar de manera permanente el acceso necesario a la información y a la rendición de cuentas.

Desarrollo humano: promover un ambiente seguro que propicie además la salud mental y física, la promoción de oportunidades a partir del estímulo de nuevas formas de expresión, a través de la cultura, el arte, el deporte y la recreación, con la participación de la población.

Autonomía municipal: respetar y hacer respetar la autonomía institucional estableciendo estrategias que la fortalezcan.

Innovación: crear y desarrollar procesos continuos de mejora que permita llevar a la Municipalidad a la satisfacción de las necesidades actuales y futuras del cantón

Integración: participar activamente en el desarrollo de proyectos e iniciativas regionales, nacionales e internacionales.

Justicia tributaria: promover la razonabilidad de los tributos y la adecuada retribución de los ingresos.

Estabilidad financiera: las decisiones institucionales deben garantizar el equilibrio entre lo económico y el desarrollo cantonal (Municipalidad de Belén, s. f.).

1.9 Estado de la cuestión

En esta sección se presenta evidencia del proceso de revisión sistemática que se llevó a cabo utilizando la “plantilla” que presenta Biolchini. Esta estructura presenta una revisión sistemática genérica. A continuación, se presenta la selección de los estudios a los cuales se les aplican criterios de inclusión y exclusión para obtener una lista de la cual se revisa y extrae información sobre el tema de este proyecto.

1.9.1 Formulación de la pregunta

Se inició formulando la pregunta para la investigación de forma que se relacionara con el área de interés del proyecto, así como con el problema a tratar y sus principales características.

1.9.1.1 Foco de la pregunta

Se pretende identificar las iniciativas, experiencias en estudios que aplican el manejo del almacenamiento y distribución del agua utilizando técnicas de minería de datos.

El objetivo de esta evaluación previa es conocer la aplicabilidad de este tipo de iniciativas además de respaldar las técnicas que se utilizaran para el desarrollo y resultado exitoso de este proyecto.

1.9.2 Calidad y amplitud de la pregunta

1.9.2.1 Problema

Tal como se detalló en la introducción, se busca optimizar el uso del agua con los datos recopilados sobre el consumo del agua en las diferentes zonas de Belén. Esto con el fin de generar modelos de minería de datos para proyectar el consumo del agua en un periodo determinado. Además, se busca segmentar los distritos por su comportamiento en el consumo del agua y así poder estimar los requerimientos de agua relacionado con el crecimiento de la población.

1.9.2.2 Pregunta de la investigación

Se busca responder a la pregunta: ¿Cuáles iniciativas fueron efectuadas y se encuentran relacionadas con el estudio y análisis de técnicas de minería de datos? Adicionalmente se validarán los estudios que aplican modelos predictivos para el uso y almacenamiento de agua.

1.9.2.3 Palabras claves y sinónimos

Para ejecutar las búsquedas en los distintos sitios web y buscadores se realizó la compilación de palabras claves para obtener la información necesaria en los artículos explorados. A continuación, se detallan las palabras utilizadas para realizar las búsquedas.

- Data Mining
- Analysis
- K-means
- Reservoirs
- Forecast
- Clustering
- Water supply operation
- Cluster analysis
- Comparative methods
- Sustainable urban water systems
- Water budget Water use

- Time series

1.9.2.3.1 Intervención

La evaluación de investigaciones en minería de datos en el uso del agua y observar los algoritmos¹² usados, además de analizar el resultado y sus beneficios obtenidos.

1.9.2.4 Control

Hasta el momento, los estudios verificados utilizan datos propios según las necesidades de cada problema. Dichos datos no serán utilizados en este proyecto, puesto que se cuenta con un *set* de datos proporcionado por los encargados del acueducto de Belén, lo cuales cuentan con las características necesarias para su exploración y análisis.

1.9.2.5 Efecto

Se pretende identificar las iniciativas relacionadas con el problema propuesto en el proyecto, además de estimar los resultados e impacto en el problema con la hipótesis del investigador.

1.9.2.6 Medida de resultado

Se busca explorar y analizar distintas investigaciones relacionadas con el problema establecido, dentro de los resultados obtenidos serán verificados mediante la lectura y comprensión de las soluciones propuestas, obtenidas en los sitios de búsqueda con un respaldo serio y reconocido.

1.9.2.7 Población

¹² Conjunto ordenado de operaciones sistemáticas que permiten hacer un cálculo y hallar la solución de un tipo de problema específico.

Los resultados de las publicaciones obtenidas de los distintos repositorios¹³ relacionados a minería de datos, la predicción, uso, reserva y abastecimiento del agua.

1.9.2.8 Aplicación

Esta revisión sistemática está relacionada con el beneficio que se podrá brindar al proyecto con la retroalimentación de estudios similares, que permitan analizar los datos relacionados con el manejo del agua y aplicar el estudio a municipios y/o comunidades, por lo que investigaciones similares podrán brindar una idea de cómo emplear los algoritmos en el manejo del agua.

1.9.2.9 Diseño experimental

El análisis de las fuentes y aportes de profesionales tiene como fin validar las prácticas relacionadas con minería de datos y seleccionar los estudios que brinden modelos y métodos de minería para ampliar el espectro de uso, así como aplicabilidad de modelos de minería como series de tiempo y agrupación K medias.

1.9.3 Selección de fuentes

En esta sección se pretenden analizar las fuentes de información que permitirán obtener resultados óptimos, aplicando la revisión de los artículos relacionados con el proyecto.

1.9.3.1 Definición del criterio de la selección de fuentes

¹³ Sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos

Como criterio de selección se utilizarán motores de búsqueda en línea que permitan contar con de mecanismos de selección y evaluación, asimismo se aplicarán filtros de palabras claves y publicaciones sugeridas por expertos.

1.9.3.2 Lenguaje de los estudios

El lenguaje más utilizado por los profesionales y científicos en el mundo es el inglés, por lo que la búsqueda será efectuada en ese idioma, debido a los términos técnicos, el análisis sistemático se realizará en el idioma español.

1.9.3.3 Identificación de las fuentes

Las fuentes identificadas serán utilizadas, así como su método de selección y listas de fuentes de información consideradas en la cadena de búsqueda.

1.9.3.4 Método de la búsqueda de las fuentes

En este trabajo se explorarán metodologías de datos como CRISP-DM con la cual se contienen métodos de minería de datos que son compatibles con el estudio a realizar. Adicionalmente, se explorarán artículos de alta calidad relacionados con las necesidades de uso adecuado del agua en motores de búsqueda especializados.

1.9.3.5 Lista de las fuentes

La lista de fuentes de información obtenidas para la búsqueda es la siguiente:

- Journal Lake and Reservoir Management
- MDPI (Multidisciplinary Digital Publishing Institute)
- Science direct
- Taylor & Francis
- Research Gate

1.9.3.5.1 Cadena de búsqueda

Para lograr los mejores resultados se aplican criterios de búsqueda And y OR de la siguiente manera (Data Mining and water) OR (reservoirs or analysis) and K-means and water) OR (forecast and clustering and water) OR (Cluster analysis AND Comparative methods AND Sustainable urban water systems AND Water budget AND Water use)

1.9.4 Evaluación después de las fuentes seleccionadas

A priori, todas las fuentes listadas satisfacen el criterio de calidad por lo que los estudios seleccionados cumplen con los requisitos de calidad.

1.9.5 Verificación de referencias

Todas las fuentes utilizadas para la recolección de los estudios brindaron artículos de alta calidad y los artículos se analizaron con ayuda de expertos en el tema de minería en Science direct y MDPI.

1.9.6 Selección de los estudios

En el proceso de revisión se utilizará un modelo incremental e iterativo, donde se ejecutan búsquedas, extracción y análisis de la información mientras se verifica el contexto del estudio.



Ilustración 1 Tipo de búsqueda

Fuente: Caricari, 2014.

La ilustración 1 muestra cómo se filtra y refina la especificidad de los artículos y estudios seleccionados para el proyecto.

1.9.6.1 Procedimientos para la selección de estudios

La cadena de búsqueda será aplicada en las fuentes seleccionadas con el propósito de escoger los estudios relevantes, como parte de la selección se obtendrá el abstracto de los estudios seleccionados desde los motores de búsqueda para posteriormente leerlos, evaluarlos y según los criterios se aplicará la inclusión o exclusión. Adicionalmente, para refinar este conjunto de estudios, el texto completo de estos estudios se leerá y analizará.

1.9.6.2 Definición de criterios de inclusión y exclusión de estudios

Los criterios de inclusión utilizados en los estudios se enfocan en que los estudios deben presentar iniciativas en las cuales se utilizan y exploran datos para determinar y optimizar el abastecimiento del agua. Además, se presta mayor atención a los estudios que contengan el uso algoritmos predictivos para la optimización de los recursos hídricos.

Como criterio de exclusión, no se seleccionarán estudios basados en otros métodos que no utilizan datos o bien, no se encuentran categorizados dentro del ámbito de la minería de datos.

1.9.6.3 Definición de tipos de estudios

Todos los estudios relacionados con el tema de la investigación estarán relacionados con el análisis predictivo¹⁴, así como la segmentación de grupos de consumidores aplicando exploraciones de tipo cualitativa y cuantitativa a la información.

1.9.7 Ejecución de selección en los sitios web de información

Una vez planificado el proceso de revisión y selección, se procederá a realizar la revisión sistemática de las fuentes seleccionadas aplicando los criterios de búsqueda, palabras claves con sus respectivas cadenas de búsqueda y procedimientos especificados para la inclusión y exclusión de estudios.

1.9.7.1 Selección de estudios iniciales

Todas las publicaciones se validan por su aporte relacionado al tema adicionalmente se utilizarán los rangos de ubicación en pertenecientes ubicados del 1erQ y 2ndo Q según Scimago Journal & Country Rank.

A continuación, se muestra la lista de estudios seleccionados:

Tabla 2 Selección de estudios

¹⁴ Es la rama de minería de datos que tiene relación con la predicción de las probabilidades y tendencias futuras.

1	MDPI	Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water Ji, Y.; Lei, X.; Cai, S.; Wang, X. Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water 2016, 8, 599.
2	ISH Journal of Hydraulic Engineering	Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir. <i>S. Mohan, N. Ramsundram. (2013) Data-mining models for water resource applications. ISH Journal of Hydraulic Engineering 19:3, pages 211-218.</i>
3	Sustainable Cities and Society. Science Direct	Cluster analysis of urban water supply and demand Karen Noiva, John E. Fernández, James L. Wescoat, Cluster analysis of urban water supply and demand: Toward large-scale comparative sustainability planning, Sustainable Cities and Society, Volume 27, 2016, Pages 484-496, ISSN 2210-6707, http://dx.doi.org/10.1016/j.scs.2016.06.003 .
4		Water Reservoir Control with Data Mining
5	Research Gate	Application of Data Mining on Reservoir Wang, L., Wang, Z. and Tao, G. (2011). Application of Data Mining on Reservoir. <i>Advanced Materials Research</i> , 356-360, pp.2950-2953.
6	Urban Water, Science Direct	A data mining approach to modelling of water supply assets Babovic, V., Drécourt, J., Keijzer, M. and Friss Hansen, P. (2002). A data mining approach to modelling of water supply assets. <i>Urban Water</i> , 4(4), pp.401-414.
7	Research Gate	A novel water quality data analysis framework based on time-series data mining
8	Research Gate	Temporal data mining of uncertain water reservoir data

1.9.7.2 Evaluación de calidad de los estudios

Hasta este momento, solo tres de los estudios obtenidos reúnen todos los criterios de inclusión previamente definidos, además de contar con un respaldo de veracidad de las fuentes consultadas.

1.9.7.3 Revisión de selección

La selección de dichos estudios obedece al aporte que estos brindan al proyecto, tras analizar 6 estudios relacionados con la oportunidad de optimización del uso y distribución del

agua, únicamente tres reúnen los requerimientos y parámetros óptimos en la resolución del problema en cuestión.

Tabla 3 Estudios seleccionados.

1	MDPI	Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water Ji, Y.; Lei, X.; Cai, S.; Wang, X. Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water 2016, 8, 599.
2	ISH Journal of Hydraulic Engineering	Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir. S. Mohan, N. Ramsundram. (2013) Data-mining models for water resource applications. ISH Journal of Hydraulic Engineering 19:3, pages 211-218.
3	Sustainable Cities and Society. Science Direct	Cluster analysis of urban water supply and demand Karen Noiva, John E. Fernández, James L. Wescoat, Cluster analysis of urban water supply and demand: Toward large-scale comparative sustainability planning, Sustainable Cities and Society, Volume 27, 2016, Pages 484-496, ISSN 2210-6707.

1.9.7.4 Extracción de la información

En esta sección se aplica la extracción de los datos e información relevante de cada uno de los estudios seleccionados.

1.9.7.4.1 Definición del criterio de inclusión y exclusión de información

Para obtener la información relevante de los estudios se toma como base la importancia de los estudios en los cuales los aportes se realizaron con técnicas de minería de datos, como la estrategia para evaluar el uso y abastecimiento del agua.

1.9.7.4.2 Formularios de extracción de datos

La siguiente tabla número 4 es utilizada para documentar y extraer la información que fue realizada sobre cada estudio, se presenta información básica para ubicar el estudio, como el título o tema, la publicación y autor. Además, se presenta un detalle general y sobre las aportaciones relevantes. Finalmente se muestran varios aspectos que se consideraron importantes, a medida que se realizaba el análisis del estudio seleccionado.

Tabla 4 Formato extracción de los datos del estudio

Identificación	
Título	Titulo original del texto consultado
Publicación	<i>Nombre de la conferencia o publicación incluyendo la fecha y páginas.</i>
Autores	<i>Nombre de los autores.</i>
Resumen	Breve descripción del objetivo de la publicación
Aspectos que destacar	
Información relacionada con el título de mi investigación extraída de la publicación.	

1.9.7.4.3 Extracción de resultados objetivos y subjetivos

Esta sección se muestra la información extraída de cada estudio seleccionado utilizando el formulario mostrado previamente.

Tabla 5 Información primer estudio

Identificación	
Título	Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water

Publicación	MDPI. Water 2016, 8, 599.
Autores	Ji, Y.; Lei, X.; Cai, S.; Wang, X.
Descripción	
Resumen	Este estudio muestra dos algoritmos de minería de datos árbol de decisiones CART ¹⁵ y la red neural ¹⁶ artificial los cuales fueron estudiados para aplicarlos en la operación de suministro de agua.
Aspectos que destacar	
<p>Los resultados de clasificación del conjunto de datos de entrenamiento son mejores que los del conjunto de datos de prueba, y el efecto de clasificación de árboles de decisión CART ¹⁷ es mejor que el de redes neuronales.</p> <p>A través de una comparación de los resultados se indica que los modos que se distinguen por el CART pueden guiar la operación de suministro de agua del depósito de manera efectiva.</p> <p>Los crecientes conflictos entre la demanda y la oferta de agua han promovido una mayor necesidad de un desarrollo en el manejo eficiente y razonable de los recursos hídricos. En la actualidad, el objetivo de la operación de un embalse ha cambiado de objetivo único a objetivo múltiple, incluyendo riego agrícola, industria, abastecimiento urbano e incluso ecología de ríos. En consecuencia, es necesario llevar a cabo el análisis de la decisión del suministro de agua multi-objetivo para los embalses y construir un modelo conveniente de toma de decisiones de suministro de agua que sea práctico para que el personal administrativo.</p>	

¹⁵ Un árbol de decisión mapea observaciones sobre un objeto para generar conclusiones sobre el valor objetivo.

¹⁶ Es un modelo artificial inspirado en el comportamiento de las neuronas y conexiones del cerebro humano tratando de resolver un problema determinado.

¹⁷ Classification and Regression Trees.

Tabla 6 Información segundo estudio

Identificación	
Título	Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir.
Publicación	<i>Taylor & Francis. Lake and Reservoir Management 32:3, pages 209-224</i>
Autores	Adem Bayram, Meltem Kenanoğlu. (2016)
Descripción	
Resumen	<p>En este estudio, se utilizaron 4 conjuntos de datos con las cantidades mensuales de agua que se liberarán del yacimiento de riego de Eleviyan en Irán como insumos en un modelo de minería de datos.</p> <p>Para el uso de árboles de decisiones, los datos se compilan y las reglas se escriben usando el método "si-condicional", comenzando desde las raíces y terminando con las hojas del árbol.</p>
Aspectos que destacar	
<ul style="list-style-type: none"> • Este estudio encontró que el método de los árboles de decisión, podría ser utilizado para obtener reglas de operación mensuales comprensibles. Sin embargo, debido a la falta de datos, la formulación de reglas perfectas y definidas que abarcan tanto períodos secos como húmedos aún no es posible. • El aumento de la población, la disminución de las precipitaciones en el noroeste de Irán debido al cambio climático y el alto costo de las nuevas presas y sistemas de agua refuerza la necesidad de utilizar la optimización para mejorar el funcionamiento de las actuales instalaciones de agua. Se experimentaron problemas operativos significativos para satisfacer las demandas. • El uso óptimo de los recursos hídricos disponibles es cada vez más importante porque la demanda de agua está aumentando rápidamente. 	

Tabla 7 Información tercer estudio

Identificación	
Título	Cluster analysis of urban water supply and demand
Publicación	Sustainable Cities and Society, Volume 27, 2016, Pages 484-496
Autores	Karen Noiva, John E. Fernández, James L. Wescoat.
Descripción	
Resumen	<p>Se comparan datos de 142 ciudades urbanas para identificar similitudes y diferencias no anticipadas con respecto a sistemas de agua y problemas.</p> <p>Esta investigación se conduce con un análisis de agrupamiento jerárquico ¹⁸basado en dos variables principales tales como el presupuesto anual de agua climática para cada área urbana y el uso anual bruto de agua per cápita en cada ciudad</p>
Aspectos que destacar	
<p>Inicialmente se agruparon ciudades por su situación climatológica y eventualmente otras agrupaciones fueron encontradas tales como ciudades con mayor uso de agua que el presupuesto de agua acorde al clima. Al final logran agrupar las ciudades en seis categorías diferentes acordes de las variables seleccionadas.</p> <p>Es importante mencionar que esta investigación no hizo una distinción entre zonas residenciales o comerciales por lo cual los autores mencionan esto sería valiosa extensión de sus análisis.</p>	

1.9.7.5 Análisis de resultados

Tras finalizar la revisión de las fuentes, se identificó el conjunto de estudios primarios.

En esta sección se realizará la clasificación de los resultados obtenidos acorde a las áreas de

¹⁸ Método de análisis de grupos puntuales, el cual busca construir una jerarquía de grupos.

estudio relacionadas con la minería. Posteriormente se realizará una comparación formal de las principales propuestas mostrando las conclusiones obtenidas de esta comparación.

1.9.7.5.1 Estudios analizados

En la tabla #8 se muestra un resumen de los estudios analizados en cada fuente.

Tomando en cuenta el procedimiento de selección para estudios primarios se ejecutó la consulta en la fuente de datos seleccionada, se obtuvo como resultado un conjunto de estudios a los que posteriormente se les aplicó el criterio de inclusión, para obtener estudios relevantes. Seguidamente, sobre este conjunto se aplicó el criterio de exclusión y se obtuvo el conjunto de estudios primarios.

Una vez que se obtuvieron los estudios primarios de una fuente en concreto, se realizó el refinado, en el que se identificaron los estudios más importantes relacionados con los anteriores y que, debido a su importancia, se añadieron como estudios primarios representados en la columna derecha de la tabla.

Tabla 8 Resumen de estudios relevantes

Fuentes	Estudios	Relevantes	Exclusión	Primarios
MDPI	52	1	51	1
ISH Journal of Hydraulic Engineering. Taylor & Francis.	19	1	18	0
Sustainable Cities and	31	1	30	1

Society. Science Direct				
Research Gate	2	0	2	0
Urban Water, Science Direct	6	1	5	0
Science Direct	12	0	12	0
Lake and Reservoir Management. Taylor & Francis.	36	1	35	1

1.9.7.5.2 Presentación de resultados

Los resultados obtenidos fueron agrupados acorde a su aporte con el fin de clasificar las tendencias que están siguiendo los investigadores del área. Las áreas consideradas son:

- Estudios realizados con CRISP-DM.
- Estudios donde se utilizó el algoritmo de K-Means.
- Estudios donde se utilizó otro tipo de algoritmo de agrupación.
- Estudios donde se utilizó análisis de series de tiempo.
- Estudios donde se utilizó análisis de Arboles de Decisión.

Tabla 9 Comparación de los estudios y métodos de minería.

Estudio	CRISP-DM	K-Means	Other Clusters	Series de Tiempo	Arboles de Decisión
Application of a Classifier Based on Data Mining Techniques in Water Supply Operation. Water	0	0	0	1	0
Application of a data mining approach to derive operating rules for the Eleviyan irrigation reservoir.	0	0	1	0	1
Cluster analysis of urban water supply and demand	0	0	1	0	0

Capítulo 2. Marco teórico o conceptual

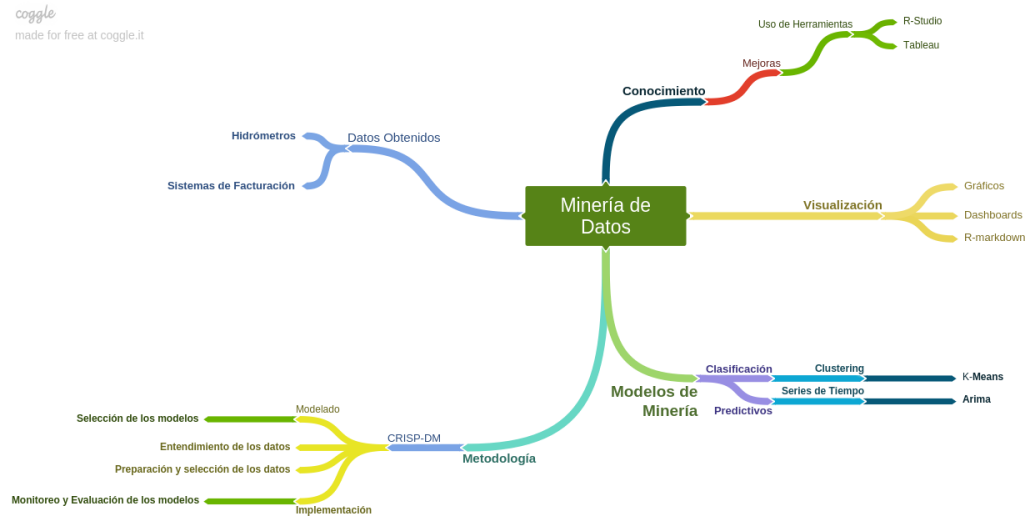


Ilustración 1 Mapa conceptual del proyecto

Fuente: Elaboración propia

Poder explorar los datos relacionados con el consumo del agua permitirá analizar patrones y comportamientos de consumo en el recurso hídrico. Si bien es cierto existen métodos operativos para distribuir y entregar el valioso recurso a cada consumidor, desde las fuentes generadoras hasta el hidrómetro, el control y consumo individual no debe ser representado por una factura o consumo mensual.

Debe imperar un control más allá de una facturación y esto puede ser posible con el uso de técnicas modernas donde el análisis de los datos, así como emplear técnicas predictivas que brindarán un panorama muy cercano a la realidad de lo que sucedió, de lo que está sucediendo y lo que pasará con las necesidades de agua en los próximos años en el cantón de Belén.

2.1 CRISP-DM

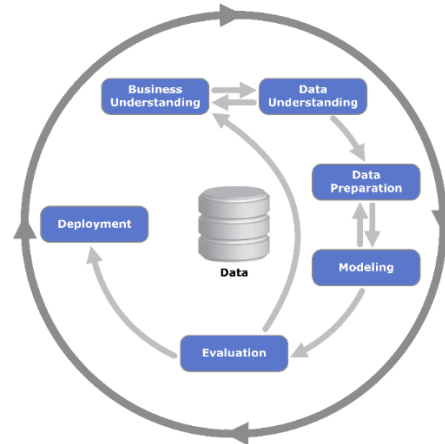


Ilustración 2 Metodología CRISP-DM

Fuente: Óscar Marbán, G. M. (2009, January)¹⁹

CRISP-DM es el acrónimo para Cross Industry Standard Process for Data Mining y es un modelo de proceso de minería de datos que detalla los enfoques o metodología que comúnmente usan los expertos (Marbán, 2009).

CRISP-DM tiene seis fases principales, las cuales se detallan a continuación y algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirían parcial o totalmente las fases anteriores:

Comprensión del negocio

Según Pete Chapman “Esta es la fase inicial se enfoca en el entendimiento de los objetivos y requerimientos del proyecto con una perspectiva empresarial” (NCR, 2000, p. 10),

¹⁹http://cdn.intechopen.com/pdfs/5937/InTechA_data_mining_amp_knowledge_discovery_process_model.pdf

después convertir este conocimiento en una definición del problema de minería de datos, y crear un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de datos

Empieza con una colección inicial de datos y procesos con actividades con la meta de entender mejor los datos, determinar la calidad de los problemas, para encontrar las primeras señales dentro de los datos e identificar temas interesantes para poder plantear la hipótesis de información oculta.

Preparación de datos

En esta fase se desarrollan todas las actividades para construir el conjunto de datos. Estas actividades se ejecutan en varias oportunidades sin ningún orden. Las tareas son de selección y transformación de tablas, registros y atributos y limpieza de datos para las herramientas de modelado.

Modelado

En la fase cuatro se escogen y ejecutan múltiples técnicas de modelado y se calibran los parámetros con el fin de tener los resultados óptimos. Existen varias técnicas que tienen requisitos específicos para la forma de los datos, por lo que algunas veces es necesario regresar a la fase anterior de preparación de datos.

Evaluación

En esta fase del proyecto se ha desarrollado un modelo (o modelos) que parece tener una buena calidad, desde un punto de vista de análisis de datos.

Despliegue

Según los requerimientos, puede ser un simple reporte o tan compleja como implementar un proceso de explotación de información que atraviese a toda la organización. Como afirma Marbán (2009) “el proceso de implementación no finaliza con el despliegue y puesta en marcha si no que se expone a un ciclo de mejora continua para implementar mejoras y durante su uso” (p. 6).

2.2 Clustering

Según Wu (2012) “los clúster de datos contienen cualidades las cuales pueden funcionar como propiedades de los individuos las cuales podrán segmentar los grupos de individuos” (p. 2).

Esta es una técnica de minería de datos para análisis clúster o análisis de conglomerados (agrupaciones). Consiste en clasificar a los individuos o datos de un estudio creando grupos o clúster de elementos y que los datos o individuos que pertenezcan dentro de cada agrupación presenten cierto grado de homogeneidad, en base a los valores adoptados sobre un conjunto de variables.

Por ejemplo, puede clasificarse un grupo de consumidores de un producto según ciertas características personales y socioeconómicas (salario, sexo, nivel cultural, etc.) lo que proporcionará una segmentación de mercado. En base a este conocimiento, el vendedor se dirigirá a ellos con diferentes estrategias de mercadeo, que aprovechen mejor los recursos y el conocimiento.

2.3 Algoritmo Kmeans

Se puede afirmar que K-means es un proceso iterativo en el que se usan los resultados de la partición anterior para mejorar la siguiente.

A diferencia otros métodos de *clustering*, en este método es necesario especificar previamente la cantidad de grupos a formar y se trabaja directamente con la matriz de datos original en lugar de la matriz de distancias.

Esta característica hace que el método k-means sea idóneo para analizar un gran número de datos, debido a que no requiere gran capacidad de memoria.

El método k-means incluye un proceso de reasignación, debido a que un caso puede ser asignado a un cierto clúster en un determinado paso y luego puede ser reasignado en otro paso a otro clúster.

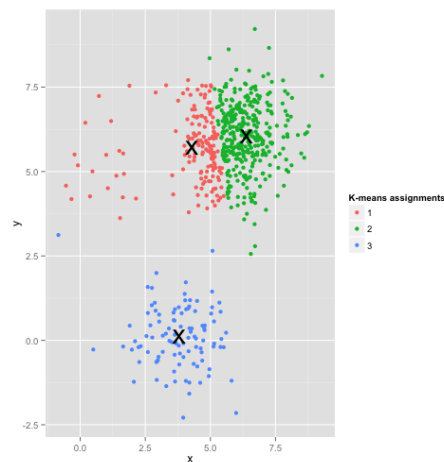


Ilustración 3 Algoritmo Kmeans

Fuente: r-bloggers.com

También es posible utilizar estimaciones previas de los centroides y con ellas generar una clasificación inicial. El lenguaje R dispone de varios algoritmos para el procedimiento de agrupación.

El mayor inconveniente de esta técnica es que, si la agrupación de los datos originales es muy distinta a los datos más recientes, podría arrojar un resultado muy distinto a la respuesta óptima.

2.4 Series de tiempo

Según el Shmueli (2016):

Indica que una serie de tiempo puede ser utilizada para optimizar reservas tanto en producción como en la demanda de un bien o servicio, por lo que el uso de estos datos adquieren un valor estratégico al utilizarlos con enfoque predictivo, gracias al apoyo de la tecnología actual se logra disminuir la incertidumbre relacionada en la toma de decisiones futuras basados comportamientos de los datos históricos (s. p.).

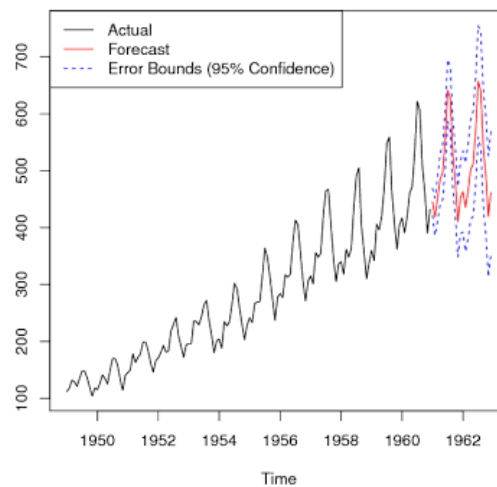


Ilustración 4 Series de Tiempo

Fuente: rdatamining.com

El término de series de tiempo es descrito como observaciones y registros que se obtienen en intervalos determinados y en una misma frecuencia, por ejemplo, diario, semanal, mensual. Un ejemplo de esto son las ventas que un almacén puede realizar semanalmente o las ventas totales de cada cuatrimestre de contratos de construcción.

Para analizar series de tiempo se utilizan métodos que permiten interpretar cada serie con el fin de predecir el comportamiento de la serie en momentos aun no observados, así sea en el futuro (extrapolación pronostica²⁰), o en el pasado (extrapolación retrógrada²¹) o en momentos intermedios (interpolación²²).

Entre los usos más comunes de las series de tiempo en los datos es su análisis para predicción y pronóstico (así como por ejemplo para predecir el clima o el comportamiento de la bolsa).

²⁰ Este tipo de pronóstico es el proceso de estimación en situaciones de incertidumbre.

²¹ El método para determinar el tiempo cero en un trazo, cuando el punto exacto de inicio no resulta obvio.

²² La obtención de nuevas lecturas partiendo del conocimiento de un conjunto discreto de lecturas.

Capítulo 3. Marco metodológico

3.1 Tipo de investigación

Para este proyecto se realizó una investigación de tipo aplicada, se pretendía optimizar el uso el abastecimiento del agua aplicando técnicas de minería de datos en los registros relacionados con el consumo de agua de los habitantes y comercios localizados en el cantón de Belén.

Con este estudio se buscaba beneficiar a todas las comunidades ubicadas en el municipio de Belén y se espera que los resultados puedan utilizarse posteriormente en distintas zonas del país.

En la actualidad, los encargados del acueducto y su equipo no utilizan modelos predictivos para la proyección de consumo, tampoco generan la agrupación de consumidores para identificar el impacto de grupos de consumidores con características similares en el consumo del acueducto de la zona.

3.2 Alcance investigativo

El alcance de este proyecto es de tipo exploratorio²³, ya que se trata de aplicar análisis de datos en un campo no muy explorado en nuestro país, el alcance de este proyecto está limitado a la búsqueda de patrones, análisis de los registros recolectados.

²³ Los estudios exploratorios aumentan el grado de familiaridad con fenómenos relativamente desconocidos

Los estudios exploratorios que se efectuaran en los datos tienen como objetivo identificar conceptos y variables que impactan en el consumo actual y futuro del abastecimiento del agua. Adicionalmente, se busca ser innovador con la generación de distintos enfoques de series de tiempo y agrupación para generar modelos predictivos aplicables a la realidad del municipio de Belén.

3.3 Enfoque

El enfoque que se aplicó en este proyecto fue de tipo mixto. El enfoque cualitativo fue aplicado para determinar el impacto de distintas variables y su correlación con los resultados del consumo, además de interpretar el sector donde se ubica el consumidor y tipo de consumidor con el objetivo de refinar los resultados.

Adicionalmente, el beneficio de utilizar el enfoque cuantitativo es validar y respaldar con los datos de consumo las hipótesis y patrones que se generan en la información posterior al análisis cualitativo.

3.4 Diseño

El diseño que se aplicó al enfoque del proyecto fue un diseño exploratorio secuencial, se realizaron análisis comparativos, cada uno cumplía con alguna de las etapas del diseño de este enfoque. Esto permitió avanzar con la siguiente, hasta finalmente interpretar los resultados.

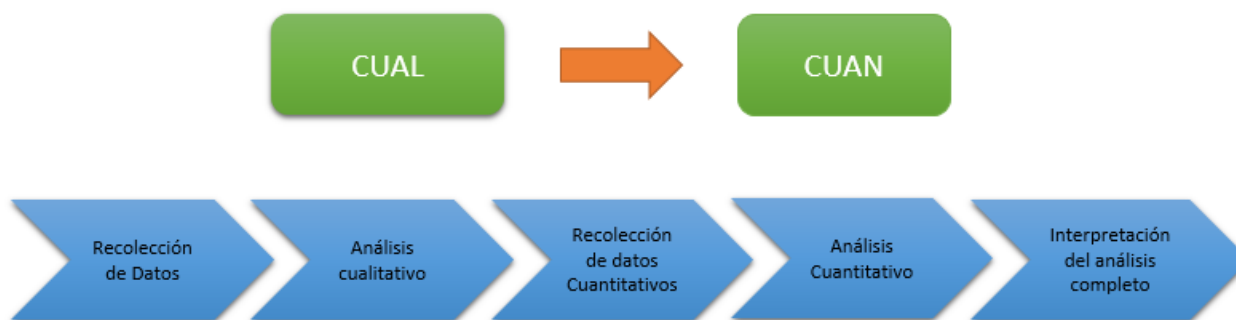


Ilustración 5 Diseño exploratorio del proyecto

Fuente: Elaboración propia

3.5 Población y muestreo

La población de los datos relacionados con el consumo contiene una media 7 mil registros anuales para este estudio se utilizaron los años 2014,2015 y 2016.

Dichos registros conforman la totalidad de hidrómetros asignados a cada uno de los consumidores de agua en el cantón de Belén.

Para la generación de los modelos se utilizó la funcionalidad `sample.split`²⁴ de R para poder segmentar los datos en dos secciones los cuales serán llamados entrenamiento con un total de 70 % de los datos y para prueba el 30 % de los datos restantes. Adicionalmente se utilizó la función `set.seed`²⁵ con un valor determinado para hacer repetible la segmentación de los datos.

²⁴ Función del programa R la cual divide los datos en dos o más grupos según será requerido.

²⁵ Función del programa R para ajustar los parámetros aleatorios en los computadores y el resultado sea el mismo.

3.6 Instrumentos de recolección de datos

Los datos utilizados en este proyecto se recolectaron de forma mensual por los personeros del acueducto del cantón de Belén quienes extraen la información del consumo de metros cúbicos de cada hidrómetro²⁶ (o medidores, como comúnmente se les conoce). Dichos dispositivos cuentan con un nivel de exactitud lo suficientemente preciso para indicar el consumo del abonado, dicha información se ingresó a un sistema informático del municipio.

Los datos recolectados se recibieron en cuatro archivos con formato Excel los cuales contienen toda la información necesaria para el presente estudio.

3.7 Técnicas de análisis de la información

En este proyecto se aplicó el estándar para la minería de datos denominada CRISP-DM (Cross Industry Standard Process for Data Mining) con el que se analizó la información recolectada en los sistemas informáticos del municipio de Belén.

Este modelo cubre todas las fases necesarias para explorar y analizar los datos que se utilizaron en este proyecto, así como la generación de tareas requeridas.

La metodología contempla el proceso de análisis de datos de forma estructurada en el desarrollo del proyecto, esto se logra estableciendo un contexto mucho más rico que influye en la elaboración de los modelos de minería de datos.

²⁶ Dispositivo de alta precisión utilizado en la medición de agua.

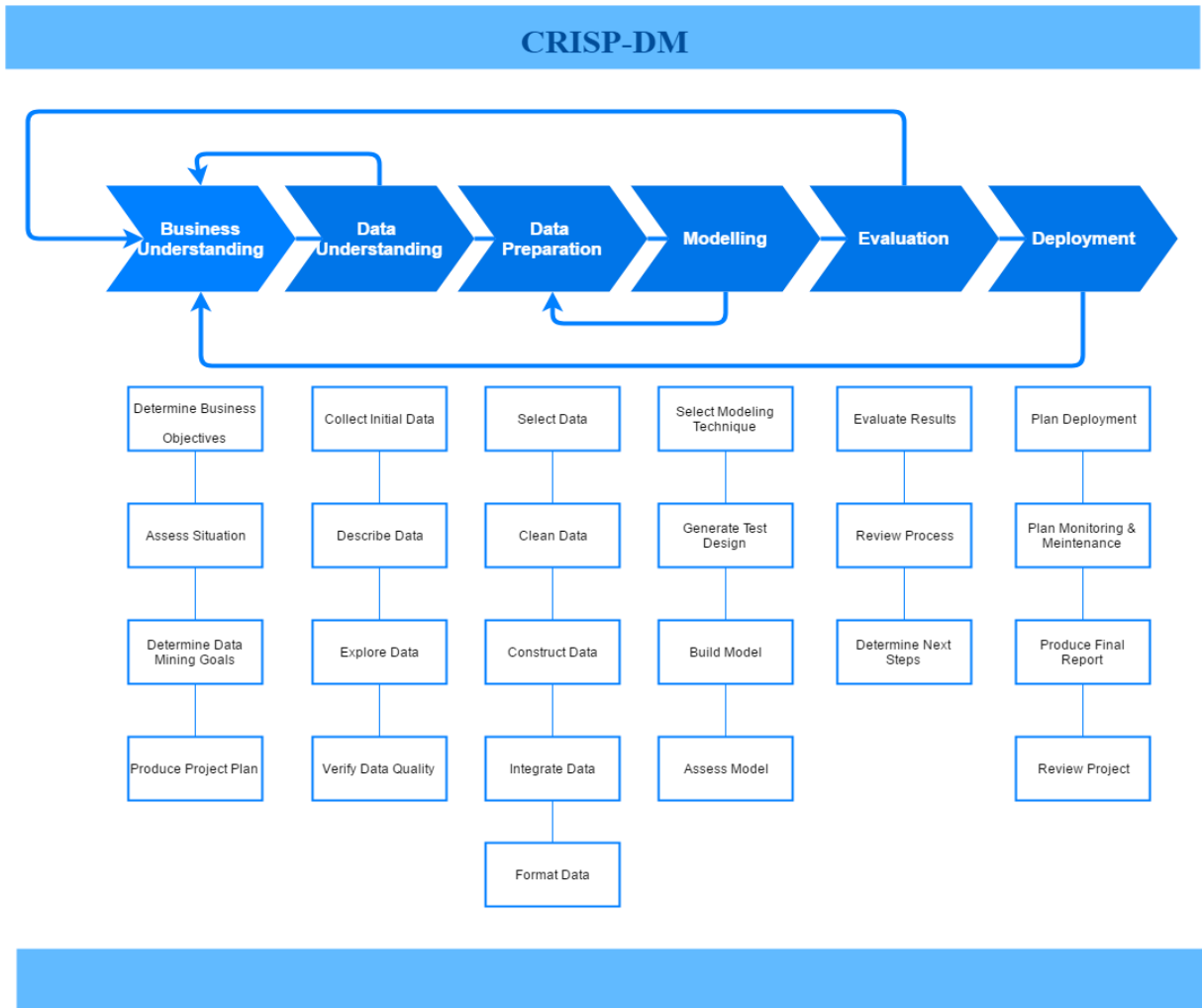


Ilustración 6 Fases del modelo CRISP-DM

Fuente: <http://jsndesign.co.uk/blog/data-mining-musical-instruments-identification/>

Fase I. Comprensión del negocio

La fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después, se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase IV. Modelado

En esta fase se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Fase V. Evaluación de resultados

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido

considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. Despliegue y puesto en producción

Generalmente, la creación del modelo no es el final del proyecto, incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y automatizada de un proceso de análisis de datos en la organización.

Adicionalmente se busca apoyar la exploración de los datos mediante distintas herramientas de *software*:

- MySQL
- R Studio
- Excel
- Tableau²⁷

²⁷ Software especializado para la visualización y manipulación de datos.

3.8 Estrategia de desarrollo de la propuesta

Para el desarrollo de este proyecto es esencial explorar los datos de forma exhaustiva, para lograr dicho objetivo es imprescindible utilizar el lenguaje de programación R con diferentes librerías y además de graficar los datos mediante el *software* TABLEAU. Con estas herramientas se podrán explorar los datos y manipularlos para explorar los posibles patrones que se contengan en los datos.

Posteriormente se utilizará la metodología CRISP-DM para avanzar en cada una de las etapas y modelar los datos actuales en distintos escenarios, para obtener un modelo que podrá optimizar el uso y almacenaje del agua en el cantón de Belén.

Capítulo 4. Análisis del diagnóstico

Día con día los proveedores de servicios trabajan en ofrecer servicios de alta calidad y eficiencia, en este caso la entrega del servicio de agua potable tiene una naturaleza de alta prioridad e importancia, por ende, los administradores de este recurso deben establecer prioridades en el manejo y una mejora continua en su distribución. Esto queda plasmado en el informe que evalúa el desempeño llamado índice de gestión municipal los resultados del año 2016 emitidos por la Contraloría General de la República en el informe DFOE-DL-IF-00007-2017²⁸ muestran que la municipalidad de Belén obtiene una calificación de 82,78 siendo la segunda mejor municipalidad de todo el país superada únicamente por la municipalidad de San Carlos la cual obtuvo una puntuación de 87,84.

²⁸ <https://sites.google.com/cgr.go.cr/igm/presentaci%C3%B3n-igm-2016>



Ilustración 7 Calificación gestión municipal 2016

Fuente: <https://sites.google.com/cgr.go.cr/igm/presentaci%C3%B3n-igm-2016>

La calificación obtenida muestra que en área relacionada con la gestión ambiental recibió la puntuación más baja dentro de los parámetros de evaluación lo que reafirma que hay importantes oportunidades de mejora tomando en cuenta que la administración del agua es una de ellas.

En la actualidad, a pesar de que los responsables del acueducto de Belén requieren el apoyo del departamento de informática para obtener la información relacionada a las lecturas y el consumo de los hidrómetros, se presentan retos importantes en cuanto a los tiempos en la obtención de los datos en forma expedita.

El departamento de informática de la municipalidad facilita archivos de Excel como el que se muestra en la Ilustración #7 el cual resulta bastante complejo de manipular.

	A	B	C	D	E	F	G	H	I	J	K	L	M
	MEDIDOR	CUENTA	CEDULA	NOMBRE	DIRECCION	FINCA	CATASTRO	LOC. CATASTRAL	RUTA	LOTE	TIPO PAJIA	TELEFONO	CONSUMO 01-2015
1													FACTURA
2	0000000540	51082	3109997254	CONDOMINIO HZ VIE 200 M ESTE DEL HOTT 003941 M000		409419211991	030601012	206	0696	ORD		84147476	702
3	0000000525	47311	3101083380	CREDIBANCO SOCIED DE ENTRADA A CALLI 167956 000		4029574195	0108107104	014	0686	ORD		81230605	663
4	0000000944	49171	3101079006	BANCO IMPROSA SO COSTADO OESTE DE 206743 000		4087308703	0306201003	502	0001	ORD		0	778
5	0000000632	02228	3102574288	S & R TRUSTEE COMF.COSTADO S. DE LA LI 056165 000		4031427478	0304901006	206	0840	DOM			611
6	0000000715	49155	3101260814	CONSTRUCTORA BRÉ 200 M ESTE INTERSEC 156131P 000		2049000482	0108303006	003	0115	DOM		88155567	314
7	0096007281	21891	3101186874	COMPONENTES INTE COMPONENTES INTE 162128 000		4053571498	0202001006	206	0395	REP		83116770	500
8	0000974880	10667	010906043	VENE GAS LILLOA CHZ 25 M NORTE DE ESTU 031232 002		4093281090	0107104013	009	2750	DOM		86552125	28
9	0000000088	10295	3101076364	MONTE POTRERILLO-FINCA MONTE POTRI 033829 000		4031252096	0108501002	002	0285	DOM			382
10	0009625824	21720	3101186874	COMPONENTES INTE COMPONENTES INTE 162128 000		4053571498	0202001006	206	0415	REP		83116770	500
11	00000004	10173	3101499764	INVERSIONES VASQ 100 M OESTE DE LA K				014	0678	REP		88129295	0
12	0000000597	45311	3101574038	ELIETH Y VILMA R & E 82 M SUR DE ABONO 149632 000		4020780694	0108205160	003	3433	IND		83800407	408
13	0000000623	44984	3101079006	BANCO IMPROSA SO 280 M NORTE DE ENT 216555 000		4089601903	0306101007	206	0740	IND		0	651
14	0000001075	12838	03010101460449	TRESCO C.E.B.S.A. CALLE POTRERILLOS 1				014	0625	DOM			501
15	0000000454	11226	3101291031	INMOBILIARIA Z F SC SAN ANTONIO LOTE 118335 000		4068476487	0107303002	501	5700	ORD			239
16	0000001549	11284	0401350056	MURILLO GONZALEZ RESIDENCIAL LA AMI 129329 001		4090645090	0108105075	004	0240	DOM		86884154	3028
17	0000002072	21156	3102487891	CENTRO DE RECREAC CLUB DE EMPLEADOS 005226 000		4012123822008	0205601002	014	0402	REP		86038967	0
18	0000000562	40545	3102487891	CENTRO DE RECREAC CLUB DE EMPLEADOS 005226 000		4012123822008	0205601002	014	0398	REP		86038967	0
19	0000000811	10295	3101076364	MONTE POTRERILLO-FINCA MONTE POTRI 033829 000		4031252096	0108501002	002	0280	DOM			234
20	0000974878	44181	3101214285	INVERSIONES AVENI 200 M NORTE DE CRE 117619 000		4078700688	0306201017	502	0076	REP			331
21	0000969831	10610	3101596193	JOSE Y ANA E HIJOS 542 M ESTE DE POLLIC 235792 000		4160511012	0106903030	008	5654	ORD			222
22	0000972081	47892	3101113946	LA CASITA DEL CHOC.FRENTE A TRANSFOR				003	3838	DOM			361
23	0000150410	10128	3101151439	ESCUELA SANTA MAI 180 M OESTE DE CALI 238358 000		4165409913	0107011040	006	3096	REP		83980561	398
24	0000000537	40566	3109401191	CONDOMINIO HORIZ.FRENTE A ESCUELA S 002086 M000		4037743997	0105702031	014	0406	DOM			1747
25	0000000096	12514	3101219735	INVERSIONES LA RAL 750 E. PRODUCTOS D				014	0658	DOM			355
26	0000002481	11394	0104078074	MOIRA BUSTAMANTE 180 M OESTE DEL CEP 018539 000		4048977598	0105702052	007	0760	REP			0
27	0000002712	11264	3101223839	INVERSIONES OSTR 150 ESTE DE LA NATH				004	0792	REP			285
28	0000004769	47186	3101291031	INMOBILIARIA Z F SC 180m SUR DE RESIDE 185976 000		4080290402	0308502002	503	0524	DOM			
29	0000973009	50409	3101226022	ALBIPLAS INTERNAC COSTADO NORTE DE 166402 000		4082325002	0306201025	206	0790	IND			
30	0000973674	51977	3101090966	VIMBERLY CI ARK.COSTADO OESTE DE ENTR 333647 000		4156370517	0108303030	504	5064	ORD		830371	

Ilustración 8 Archivo muestra

Fuente: Elaboración propia

Durante el proceso de exploración de las lecturas, el cual se ejecuta posteriormente, se realizan agrupaciones y sumatorias de forma manual para tratar de identificar comportamientos irregulares en el consumo de las comunidades.

Estas tareas manuales limitan y no cubren el espectro de análisis que se puede efectuar en los datos obtenidos.

Queda en evidencia la falta de fuentes de información específicas para los responsables del recurso hídrico, lo cual genera una alta dependencia del área de sistemas en la entrega de los datos en formatos crudos.

En este momento los encargados no cuentan ni utilizan métodos automatizados para poder generar las proyecciones relacionadas con el consumo del agua basándose en los datos recolectados.

El tratamiento que se le brinda a los datos en este momento es de forma reactiva y en muchos casos tardía, dado que durante la recolección de lecturas no se brinda una alerta o alarma a los usuarios en caso de tener un comportamiento inusual, por falta de controles automatizados.

El consumo que se realizan en cada uno de los acueductos no se analizan según el tamaño de la población, patrones de consumo o tipo de consumidores que reciben el servicio, lo que representa una oportunidad de utilizar los datos recopilados para obtener el conocimiento histórico necesario apoyados en los datos.

Capítulo 5. Propuesta de solución

Para cumplir con las expectativas del proyecto se plantea el uso de dos técnicas como la segmentación y el uso de serie de tiempo, el primer paso es segmentar a los consumidores por su comportamiento analizando los años 2014, 2015 y 2016. Esto se logra cargando los datos en el motor de base de datos MySQL.

5.1 Carga de datos

Carga de Datos en R

```
setwd("C:/Users/jchaves/OneDrive/Cursos/PROMiDAT/Proyecto/Datos/LISTOS")
#Carga del año 2015 #CSV

suppressWarnings(suppressMessages(library(openxlsx)))

consumo2014<- read.xlsx(xlsxFile = "Consumo-2014-acueducto-V1.xlsx" , sheet = 1, rowNames = T)
consumo2015<- read.xlsx(xlsxFile = "Consumo-2015-acueducto-V2.xlsx" , sheet = 1, rowNames = T)
consumo2016<- read.xlsx(xlsxFile = "Consumo-2016-acueducto-V1.xlsx" , sheet = 1, rowNames = T)
```

Ilustración 9 Carga de datos

Fuente: Elaboración propia

En la ilustración 8 se puede observar el código utilizado para la carga de los datos en Dataframes²⁹ de lenguaje de programación R.

Una vez cargados los datos se realiza una conexión con el motor de base de datos para guardar cada Dataframe como una tabla, estas se pueden utilizar posteriormente en cualquier momento, con lo que se logra agilizar las consultas de los datos.

²⁹ Estructura de datos almacenada en Memoria

Creando la conexión a MYSQL

```
suppressWarnings(suppressMessages(library(RODBC)))  
canal<-odbcConnect("R-MYSQL", uid = "ruser")
```

Codigo para salvar los datos en MYSQL

```
#salvamos los datos en la tabla de consumo  
sqlSave(canal, consumo2014 , tablename = "LECTURAS2014") #Salvando las lecturas mensuales en mysql  
sqlSave(canal, consumo2015 , tablename = "LECTURAS2015") #Salvando las lecturas mensuales en mysql  
sqlSave(canal, consumo2016 , tablename = "LECTURAS2016") #Salvando las lecturas mensuales en mysql  
  
#salvamos los datos en la tabla de consumo  
sqlSave(canal, acuaductos , tablename = "acuaductos") #Salvando datos relacionados con las fuentes d  
e agua en mysql
```

Ilustración 10 Salvando los datos

Fuente: Elaboración Propia

En la ilustración 9 se describe el código necesario para guardar los datos directamente en el servidor de base de datos, una vez que los datos están alojados en el servidor se logra visualizar la estructura de las tablas desde herramientas como MySQL WorkBench.

5.2 Optimización en las consultas

Para lograr una mejora y optimización en los tiempos de respuestas de las consultas y con el conocimiento de que cada tabla posee aproximadamente 7600 registros, se procedió con la generación de 4 índices para agilizar las consultas entre tablas utilizando el cambio llave de cada tabla llamado *rowname*, el cual contiene el código de hidrante.

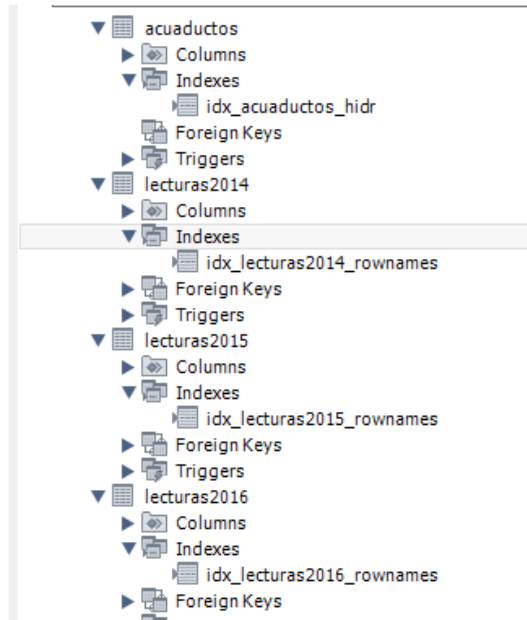


Ilustración 11 Índices generados

Fuente: Elaboración Propia

En este mismo proceso de optimización se elaboraron 4 vistas para utilizarlas en los diferentes reportes. Cada una de ellas permite obtener los datos de forma más sencilla dada la cantidad de columnas a utilizar.

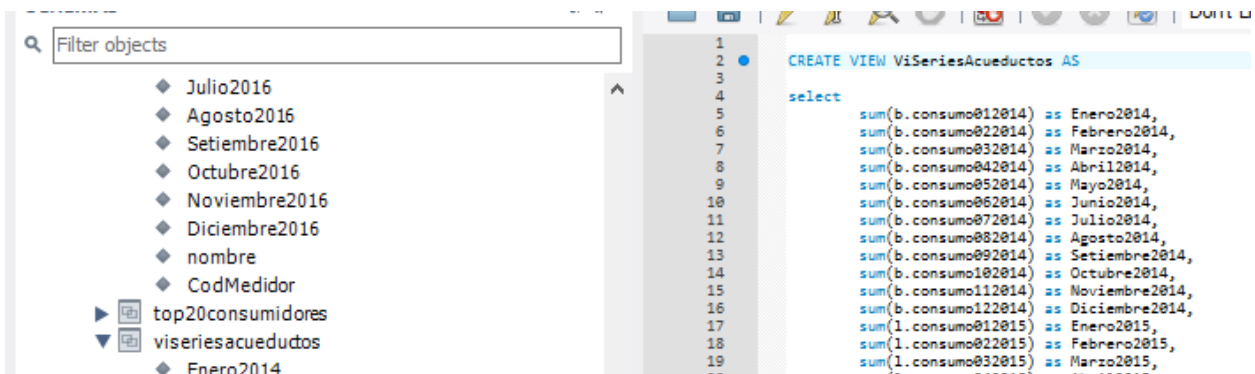


Ilustración 12 Vistas creadas para las consultas.

Fuente: Elaboración Propia

La ilustración 11 muestra el código utilizado para crear una vista, la cual obtiene las sumatorias del consumo mensual por cada acueducto. El resultado de dicha vista será utilizado posteriormente en los análisis de los datos, obteniendo beneficios propios de SQL.

5.3 Exploración de los datos.

Luego de que los datos fueron cargados, se inicia con la exploración por medio de gráficos en la herramienta Tableau utilizando la conexión por ODBC³⁰ para leer los datos directamente del Servidor en MySQL.

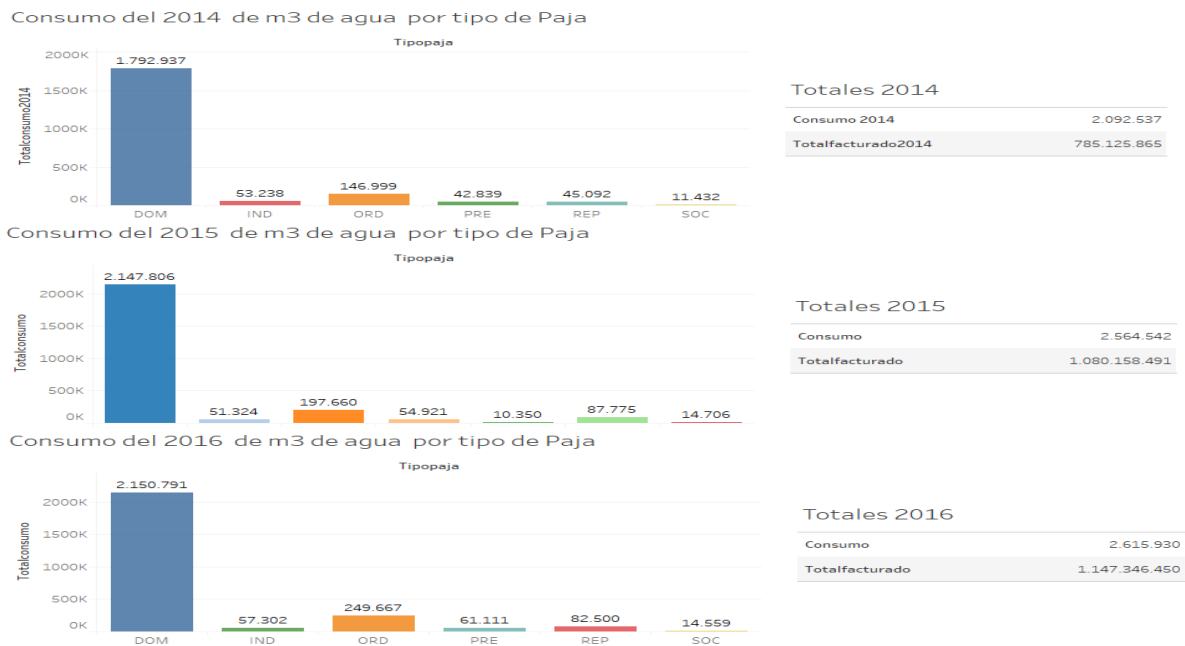


Ilustración 13 Total de consumidores por tipo

Fuente: Elaboración Propia

³⁰ Estándar de acceso a las bases de datos desarrollado por SQL Access Group

La ilustración muestra que la mayor cantidad de consumidores registrados en Belén es por tipo DOM, el cual representa domicilio seguido de ORD, el cual es ordinario ligado a Centros Corporativos o Condominios.

Está claro que el uso de herramientas que permiten visualizar los datos son de gran ayuda y eso se confirma, no solo con el despliegue de la información sino con el texto del matemático Theodosia Prodromou (2017), que señala:

La visualización de los datos brinda un mejor enfoque para comprender los problemas y las posibles soluciones. Tener solo datos no es suficiente, debemos permitir que las personas comprendan cuáles son estos datos y cómo utilizarlos. La visualización representada en forma de imágenes o gráficos permite explorar con facilidad los conceptos y las tendencias de un problema para los responsables de la toma de decisiones (s. p.).

5.3.1 Promedio consumo en el año 2015

La Ilustración 13 muestra el promedio de consumo en cada uno de los acueductos, podemos observar que los acueductos de Cariari, Mangos y Zamora tienen un consumo mucho más alto en comparación a los otros acueductos.

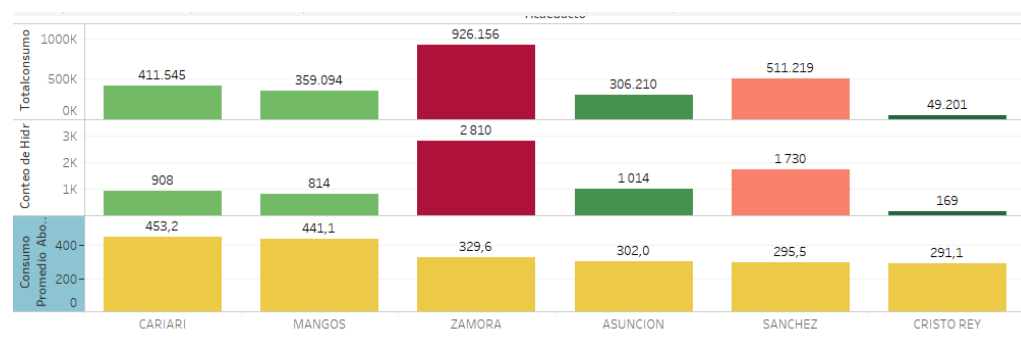


Ilustración 14 Consumo promedio por acueducto

Fuente: Elaboración Propia

5.3.2 Grandes consumidores 2014 vs 2015

Los siguientes gráficos nos ilustran el crecimiento en grandes consumidores de los acueductos de Belén, comparando los años 2014 y 2015.

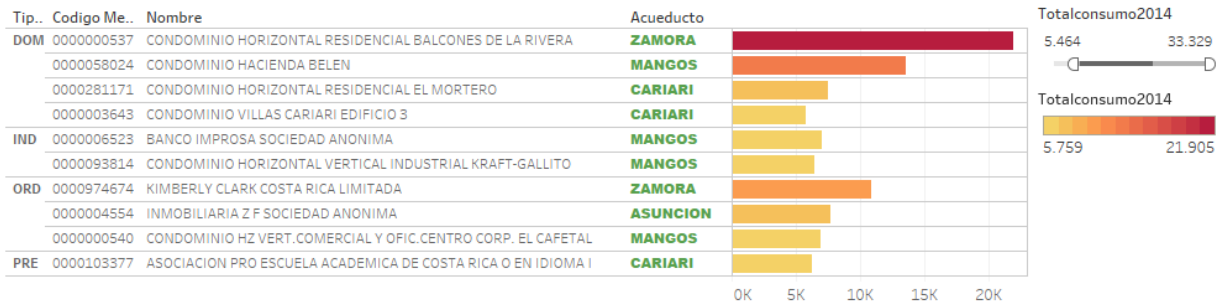


Ilustración 15 Grandes consumidores del 2014

Fuente: Elaboración Propia

Esta comparación permite tener una noción del incremento en el consumo en el lapso de un año en la comunidad, además de destacar el incremento de condominios que se abastece del acueducto de los Mangos.

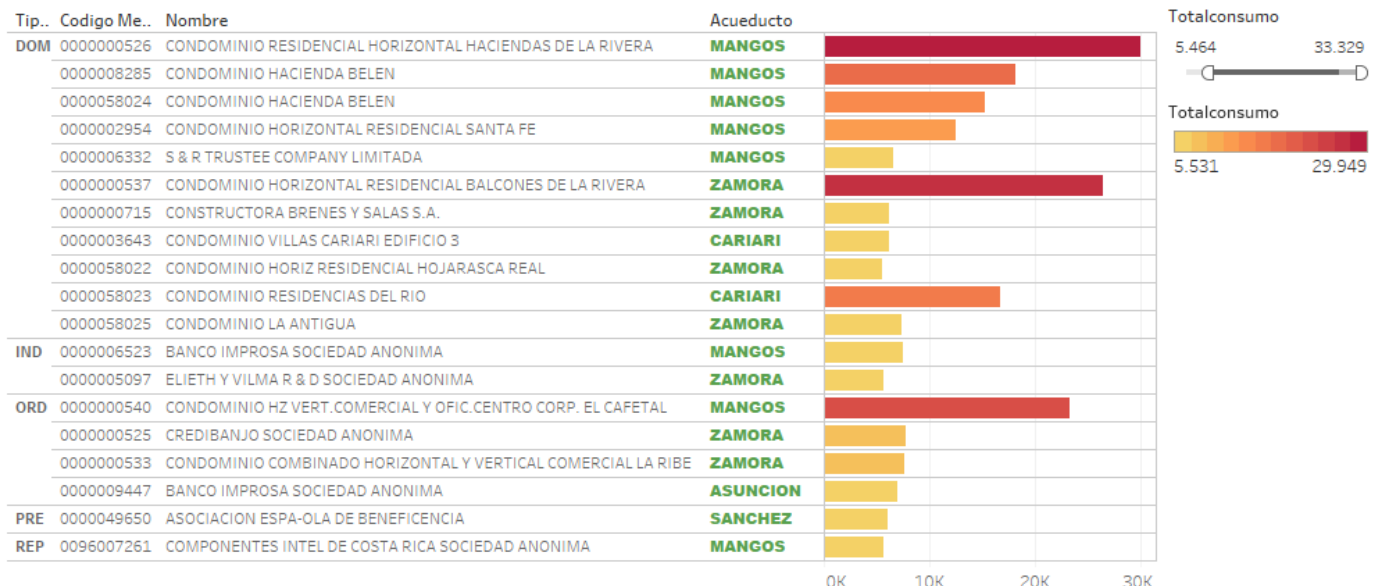


Ilustración 16 Grades consumidores del 2015

Fuente: Elaboración Propia

5.3.3 Cantidad de consumidores por tipo de paja

En la ilustración 16 se puede observar la cantidad de consumidores por tipo de paja, además de resaltar el tamaño el cual representa el consumo en relación a los demás consumidores de la misma categoría para el año 2015.

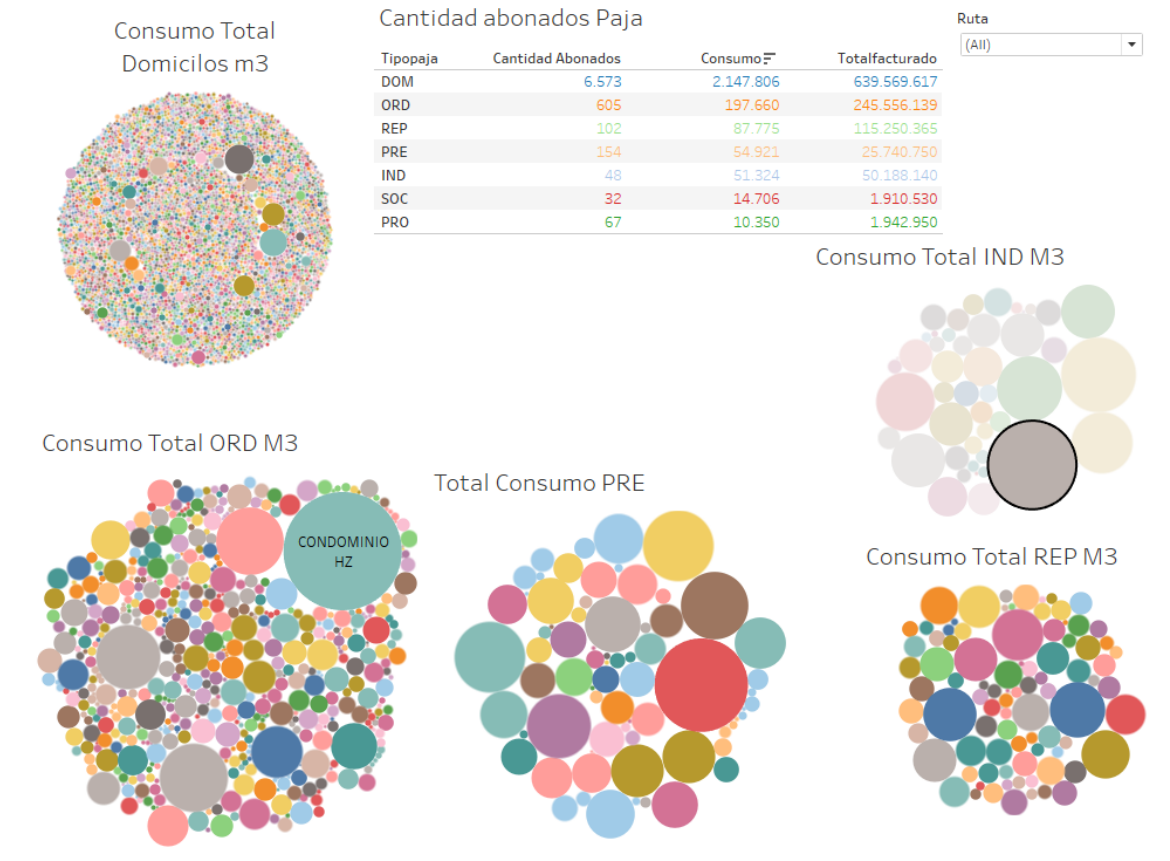


Ilustración 17 Consumidores por tipo

Fuente: Elaboración Propia

5.4 Clustering

Siguiendo los pasos recomendados por Siddhartha Bhattacharyya (2017)

- Se utilizaron los atributos de selección de los datos.
- Selección del algoritmo de creación para los clústeres
- Validación del clúster.
- Se realizó una interpretación de los resultados.

Luego de observar el comportamiento de los datos, se procedió con la creación de 4 C utilizando los datos del año 2015 el cual representa un crecimiento en la población, adicionalmente se realizará una segmentación utilizando todos los datos del 2015 y las otras tres basándose en las comunidades con el promedio más alto de consumo Cariari, Mangos y Zamora.

5.4.1 Segmentación Belén

Para poder una segmentación en los datos seleccionados del año 2015, los cuales suman 7213 registros, se utilizó el método del codo³¹ en el lenguaje de programación R. Este indicará la cantidad de grupos utilizando las distancias entre los centroides, entre mayor sea la distancia entre los centros mejor será el resultado de la clasificación de los consumidores.

³¹ Método utilizado para generar gráficamente la estimación para la creación de grupos en un set de datos determinado.

```

kBelen <-sqldf( "select
  Enero,
  Febrero,
  Marzo,
  Abril,
  Mayo,
  Junio,
  Julio,
  Agosto,
  Setiembre,
  Octubre,
  Noviembre,
  Diciembre,
  totalconsumo,
  TipoConsumo,
  Cacueducto

  from consumidoresActivos2015 order by totalconsumo desc ")

consumidores<-kBelen #CONSUMIDORES
kBelen$totalconsumo<-as.factor(kBelen$totalconsumo)

kBelen[is.na(kBelen)] <- 0
consumidores[is.na(consumidores)] <- 0

set.seed(660)

```

Ilustración 18 Selección datos para segmentación

Fuente: Elaboración Propia

Creacion de clústers con The elbow method

El método del codo es muy utilizado para determinar la cantidad de clúster óptimos que conforman un conjunto de datos. Este método se destaca por la precisión, pero llega a ser subjetivo ya que grafica la distancia de los centroides ya estables.

Entre mayor sea la distancia de los centroides, las distancias entre los individuos serán más claras por lo que la segmentación genera el valor esperado.

El siguiente código del lenguaje de programación R muestra las instrucciones utilizadas necesarias para generar el grafico del codo en las distancias entre los centroides del

clúster y el punto de equilibrio. Es importante tener en cuenta que utilizar un K^{32} muy alto significa segmentaciones que podrían llegar a ser complejas de entender dada la poca distancia entre los centros.

```
InerciaIC<-rep(0,30)
for(k in 1:30) {
  grupos<-kmeans(kBelen,k ,iter.max=300,nstart=100)
  InerciaIC[k]<-grupos$tot.withinss
}
```

El resultado de graficar las inercias es el siguiente, se puede observar que utilizando un $K=4$ es posible segmentar los consumidores del cantón de Belén de forma adecuada.

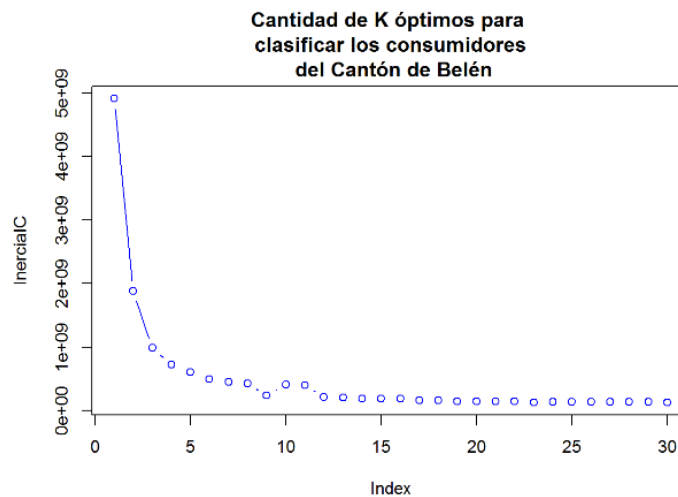


Ilustración 19 Gráfico de los K para Belén

Fuente: Elaboración Propia

Generando la segmentación de Belén.

El siguiente paso es aplicar el K en el modelo de segmentación de la siguiente manera

³² En el método K-means el K representa el número de grupos

```
grupos<-kmeans (kBelen,4,iter.max=300,nstart=100)
```

La variable grupos almacena los resultados de la segmentación utilizando 4 clúster.

Detalles de los grupos en Belén

La siguiente consulta SQL³³ retorna un resumen con las características de cada clúster generados para todos los datos del año 2015.

```
resultadoskBelen<-sqldf("select Grupo,count(*) as  
TotalRegistros,min(totalconsumo) as Minimo,max(totalconsumo) as Maximo,  
avg(totalconsumo) as Promedio from consumidores group by Grupo")
```

La tabla 10 muestra las características de cada uno del clúster generados con los datos, se pueden apreciar los rangos de cada uno de los grupos y su promedio de consumo en m3.

Tabla 10 Características de los clúster

Grupo	Total Registros	Mínimo	Máximo	Promedio
1	1487	441	2536	667,08
2	6	15227	29949	21624,33
3	5668	1	441	213,55
4	52	2600	12446	4469,03

Análisis de la agrupación

Se puede observar en el siguiente gráfico de la ilustración 24 la cantidad de individuos en cada grupo. Los dos grupos con mayor cantidad de individuos son el clúster 1 y 3.

³³ Solicitud de datos almacenados en un servidor mediante el lenguaje SQL

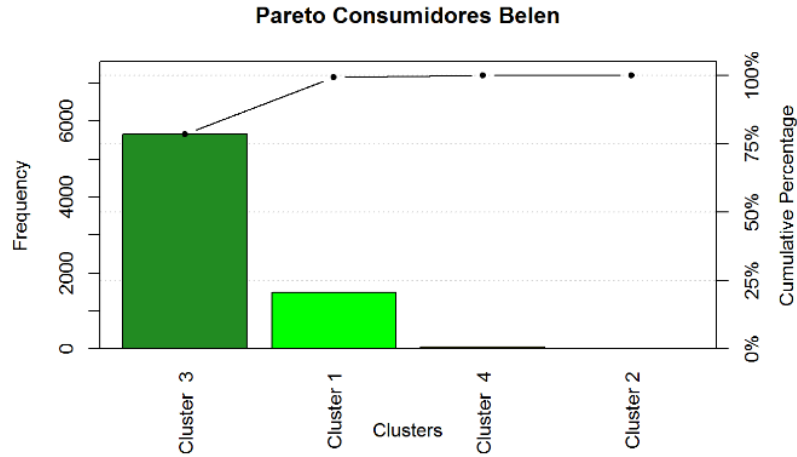


Ilustración 20 Gráfico clúster de Belén

Fuente: Elaboración Propia

Gráfico radar consumidores de Belén

Con el gráfico tipo Radar generado desde R con la ayuda del paquete **fmsb**³⁴ el cual contiene una función llamada `radarchart`³⁵, se puede graficar los datos del dataframe desplegando las características de los individuos en los en los 4 clúster generados por k-means.

³⁴ <https://www.rdocumentation.org/packages/fmsb/versions/0.6.1/topics/radarchart>

³⁵ Gráfico tipo Radar

**Comparación de clúster
de consumidores en el canton de Belen**

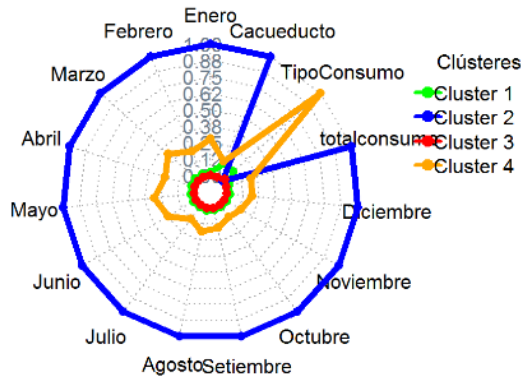


Ilustración 21 Gráfico radar clusters Bélen

Fuente: Elaboración Propia

Se puede observar un consumo alto en los individuos que pertenecen al clúster 2. Así mismo el clúster 4 muestra un consumo menor al clúster 2, pero con un consumo alto en meses como enero, marzo, mayo y junio. Finalmente, el clúster 1 y el clúster 3 tienen un consumo regular durante el año y se caracterizan por tener un consumo menor al clúster 2 y 4.

Código para generar el gráfico tipo radar.

El siguiente código permite generar los gráficos tipo radar en el lenguaje de programación R.

```
library(fmsb)

radarchart(centros,maxmin=TRUE,axistype=4,axislabcol="slategray4",
centerzero=FALSE,seg=8,cglcol="gray67",
pcol=c("green","blue","red","orange","black"),
plty=1,
plwd=5,
title="Comparación de clúster \n de consumidores en el canton de
Belén")
```

```

legenda <-legend(1.5,1, legend=c("Cluster 1","Cluster 2","Cluster 3","Cluster 4"),
  seg.len=-1.4,
  title="clústeres",
  pch=21,
  bty="n" ,lwd=3, y.intersp=1, horiz=FALSE,

  col=c("green","blue","red","orange","black"))

```

Representación del clúster en Belén

La Tabla 11 muestra los mayores consumidores del cantón de Belén, los cuales se sitúan en el clúster 2 compuesto por 6 registros.

Tabla 11 Consumidores clúster #2 Belén

Consumidor	TipoPaja	Acueducto	Consumo M3
## 1 CONDOMINIO RESIDENCIAL HORIZONTAL HACIENDAS DE LA RIVERA	DOM	MANGOS	29949
## 2 CONDOMINIO HORIZONTAL RESIDENCIAL BALCONES DE LA RIVERA	DOM	ZAMORA	26407
## 3 CONDOMINIO HZ VERT.COMERCIAL Y OFIC.CENTRO CORP. EL CAFETAL	ORD	MANGOS	23327
## 4 CONDOMINIO HACIENDA BELEN	DOM	MANGOS	18102
## 5 CONDOMINIO RESIDENCIAS DEL RIO	DOM	CARIARI	16734
## 6 CONDOMINIO HACIENDA BELEN	DOM	MANGOS	15227

5.4.2 Segmentación acueducto de Zamora.

Al observar la ilustración 13, destaca que el acueducto de Zamora presenta el tercer consumo promedio más alto en el cantón de Belén, por lo que se realizó la exploración de las características de consumidores que se abastecen por dicho acueducto, en este caso los datos seleccionados del año 2015 los cuales suman 2716 registros.

Calculo del K utilizando “The elbow method” en Zamora

Aplicando la misma técnica para obtener la cantidad de clúster el resultado del gráfico obtenido es aplicar un K=5 lo cual se puede corroborar con la ilustración 21.

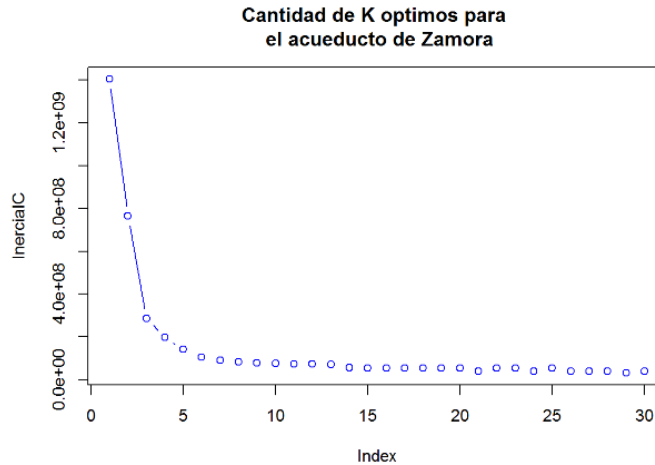


Ilustración 22 K=5 acueducto Zamora

Fuente: Elaboración Propia

Generando la segmentación de consumidores en Zamora.

El siguiente paso es aplicar el K en el modelo para generar segmentación de los datos relacionados con el acueducto de Zamora la siguiente manera:

```
gruposza<-kmeans(kZAMORA,5,iter.max=200,nstart=100)
```

La variable gruposza almacena los resultados de la segmentación utilizando 5 clúster.

Detalles de los grupos en el acueducto de Zamora

La siguiente consulta SQL retorna un resumen con las características de cada clúster generados para todos los datos del año 2015.

```
resultadoskzamora<-sqldf("select Grupo,count(*) as TotalRegistros,min(totalconsumo) as Minimo,max(totalconsumo) as Maximo, avg(totalconsumo) as Promedio from consumidoresza group by Grupo")
```


En la tabla 12 se pueden apreciar las características del clúster generados para el set de datos con los consumidores del acueducto de Zamora, se pueden observar los valores más importantes relacionados con la segmentación.

Tabla 12 Características de los clústeres de Zamora

Grupo	Total Registros	Mínimo	Máximo	Promedio
1	1	26 407	26 407	26 407
2	33	1 418	3 750	2 255
3	9	4 247	7 808	5 943
4	745	369	1 387	548
5	1 928	1	368	188

Análisis de la agrupación

Se puede observar en el siguiente grafico el cual representa la cantidad de individuos en cada grupo donde los dos grupos con mayor cantidad de individuos son el clúster 5 y 4.

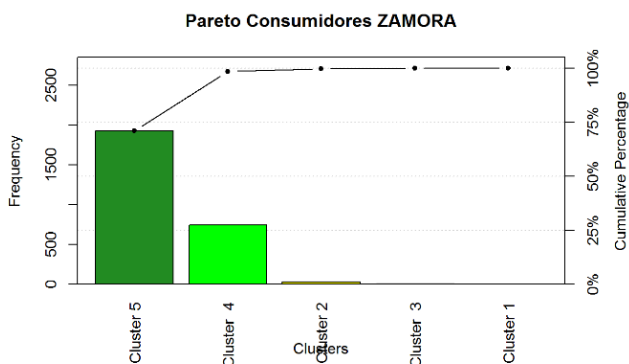


Ilustración 23 K=5 clúster Zamora

Fuente: Elaboración Propia

Gráfico Radar consumidores de Zamora

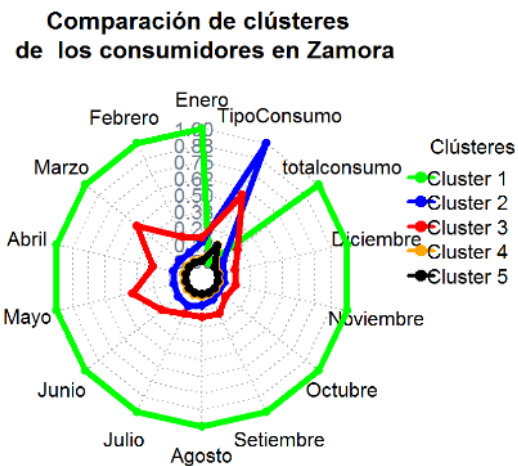


Ilustración 24 Gráfico radar Zamora

Fuente: Elaboración Propia

- El primer clúster cuenta con un único registro que muestra un consumo de 26407 m³ anuales siendo el grupo que posee el individuo que genera más consumo durante este año en el acueducto de Zamora.
- En el segundo clúster se pueden observar 9 registros con un consumo que va de 4247 m³ a 7808 m³. Este grupo de consumidores siguen un comportamiento de consumo alto, está compuesto por condominios habitacionales y consumo industrial en ZAMORA.
- El tercer clúster contiene 33 consumidores contiene un rango de 1418 m³ hasta 3750 m³.
- El cuarto clúster contiene 745 individuos que muestran un consumo alto su rango de consumo oscila entre 369 m³ y 1387 m³.

- El quinto clúster contiene 1928 individuos que muestran más bajo y regular su rango de consumo oscila entre 1 m³ y 368 m³ este consumo se realiza en su mayoría por hogares de la zona.

Análisis de los mayores consumidores en los clústeres de Zamora

La Tabla 13 muestra los mayores consumidores del acueducto de Zamora se sitúan en el clúster 1 y con 3.

Tabla 13 Grandes consumidores del acueducto de Zamora

Consumidores Zamora	Tipo Paja	Consumo M3	clúster
## 1 CONDOMINIO HORIZONTAL RESIDENCIAL BALCONES DE LA RIVERA	DOM	26407	1
## 2 CREDIBANJO SOCIEDAD ANONIMA	ORD	7808	3
## 3 CONDOMINIO COMBINADO HORIZONTAL Y VERTICAL COMERCIAL LA RIBE	ORD	7618	3
## 4 CONDOMINIO LA ANTIGUA	DOM	7393	3
## 5 CONSTRUCTORA BRENES Y SALAS S.A.	DOM	6131	3
## 6 ELIETH Y VILMA R & D SOCIEDAD ANONIMA	IND	5621	3

5.4.3 Segmentación acueducto de Mangos

Como se mostró anteriormente en la ilustración 13 el acueducto de Mangos presenta el segundo consumo promedio más alto en el cantón de Belén. En este caso los datos seleccionados son del año 2015, los cuales suman 794 registros, aunque el número de consumidores es menor se nota un alto consumo en esta zona.

Calculo del K utilizando “The elbow method” en Mangos

Aplicando la misma técnica para obtener la cantidad de clúster el resultado del gráfico obtenido es aplicar un K=5, lo cual se puede corroborar con la ilustración 24.

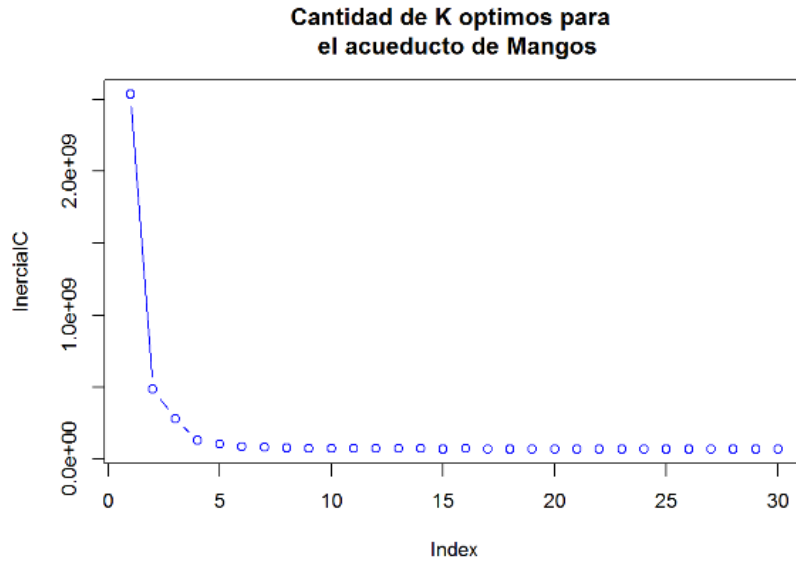


Ilustración 25 K=5 del acueducto Mangos

Fuente: Elaboración Propia

Generando la segmentación de consumidores en Mangos.

El modelo para la segmentación se generó de la siguiente manera.

```
gruposma<-kmeans (kMANGOS, 5, iter.max=200, nstart=100)
```

La variable gruposma almacena los resultados de la segmentación utilizando 5 clúster.

Detalles de los grupos en el acueducto de Mangos

El siguiente query retorna un resumen con las características de cada clúster generados para todos los datos del año 2015.

```
resultadoskmangos<-sqldf("select Grupo, count(*) as TotalRegistros, min(totalconsumo) as Minimo, max(totalconsumo) as Maximo, avg(totalconsumo) as Promedio from consumidoresma group by Grupo")
```

La tabla 14 detalla las características principales de cada uno de los grupos obtenidos con la segmentación de los 5 grupos en el acueducto de Mangos.

Tabla 14 Características clúster Mangos

Grupo	Total Registros	Mínimo	Máximo	Promedio
1	2	23 327	29 949	26 638
2	6	3 440	7 446	5 437
3	654	1	443	214
4	129	447	2 899	678
5	3	12 446	18 102	15 258

Análisis de la agrupación

Se puede observar en el siguiente gráfico, el cual representa la cantidad de individuos en cada grupo. Los dos grupos con mayor cantidad de individuos son el clúster 3 y 4.

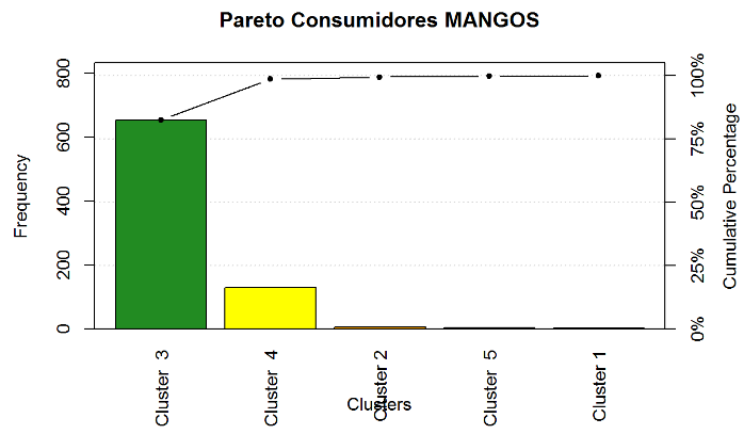


Ilustración 26 clúster Mangos

Fuente: Elaboración Propia

Gráfico radar consumidores de Mangos

- El primer clúster cuenta con dos registros que mantienen un consumo de 23327 m³ hasta 29949 m³ anuales.
- En el segundo clúster es posible observar 6 registros con un consumo que va de 3440 m³ a 7446 m³ este grupo muestra alto requerimiento de agua de la zona de los Mangos.

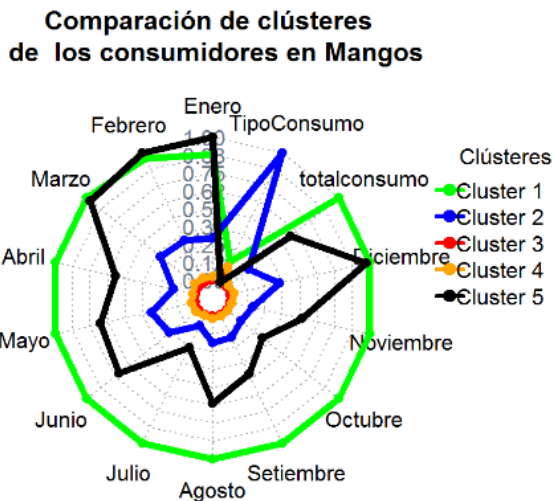


Ilustración 27 Gráfico radar Mangos

Fuente: Elaboración Propia

- El tercer clúster contiene 654 individuos que muestran un consumo bajo, el rango de consumo oscila entre 1 m³ y 443 m³.
- El Cuarto clúster contiene 129 individuos que muestran un consumo alto su rango de consumo oscila entre 447 m³ y 2899 m³.

- El Quinto clúster contiene 3 individuos que muestran un consumo alto su rango de consumo oscila entre 12446 m3 y 18102 m3.

Análisis de los mayores consumidores en los clústeres de Mangos

Como se nos muestra en la Tabla 15 los mayores consumidores del acueducto de Mangos se sitúan en el clúster 1, 5 y 2.

Tabla 15 Grandes consumidores Mangos

Consumidores Mangos	Tipo Paja	Consumo M3	clúster
## 1 CONDOMINIO RESIDENCIAL HORIZONTAL HACIENDAS DE LA RIVERA	DOM	29949	1
## 2 CONDOMINIO HZ VERT.COMERCIAL Y OFIC.CENTRO CORP. EL CAFETAL	ORD	23327	1
## 3 CONDOMINIO HACIENDA BELEN	DOM	18102	5
## 4 CONDOMINIO HACIENDA BELEN	DOM	15227	5
## 5 CONDOMINIO HORIZONTAL RESIDENCIAL SANTA FE	DOM	12446	5
## 6 BANCO IMPROSA SOCIEDAD ANONIMA	IND	7446	2

5.4.4 Segmentación acueducto de Cariari

Como se mostró, en la ilustración 13 el acueducto de Cariari presenta el promedio más alto de consumo en el cantón de Belén, para este grupo de datos del año 2015 los cuales suman 876 registros.

Calculo del K utilizando “The elbow method” en Cariari

Aplicando la misma técnica que en los acueductos anteriores con el fin de obtener la cantidad de clúster óptimos el resultado que se obtiene es el de aplicar un K=5 lo cual se puede corroborar con la ilustración 27.

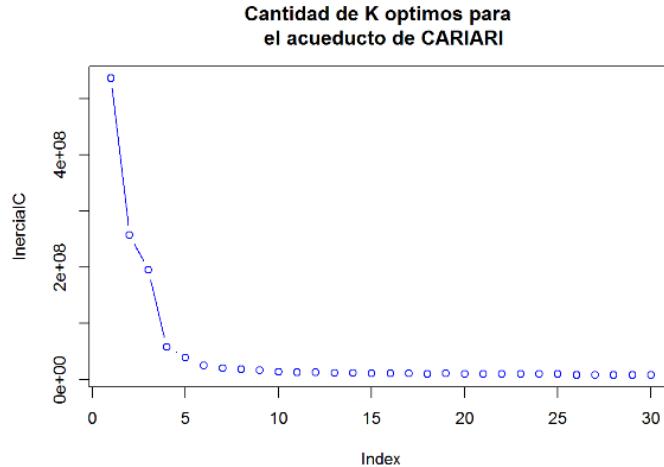


Ilustración 28 K=5 del acueducto Cariari

Fuente: Elaboración Propia

Generando la segmentación de consumidores en Cariari.

El siguiente paso es aplicar el K obtenido el modelo para la segmentación de la siguiente manera.

```
gruposca<-kmeans(kCARIARI,5, iter.max=200,nstart=100)
```

La variable gruposca almacena los resultados de la segmentación utilizando 5 clúster.

Detalles de los grupos en el acueducto de Cariari

El siguiente query retorna un resumen con las características de cada clúster generados para todos los datos del año 2015.

```
resultadoskCARIARI<-sqldf("select Grupo,count(*) as TotalRegistros,min(totalconsumo) as Minimo,max(totalconsumo) as Maximo, avg(totalconsumo) as Promedio from consumidoresca group by Grupo")
```


La tabla 16 detalla las características principales de cada uno de los grupos obtenidos con la segmentación de los 5 grupos en el acueducto de Mangos.

Tabla 16 Características clúster Cariari

Grupo	Total Registros	Mínimo	Máximo	Promedio
1	1	16 734	16 734	16 734
2	77	830	2 413	1 119
3	352	363	819	533
4	9	2 771	6 205	4 174
5	437	1	361	191

Análisis de la agrupación

Según el siguiente gráfico, el cual representa la cantidad de individuos en cada grupo, los dos grupos con mayor cantidad de individuos son el clúster el 5, 3 y 2.

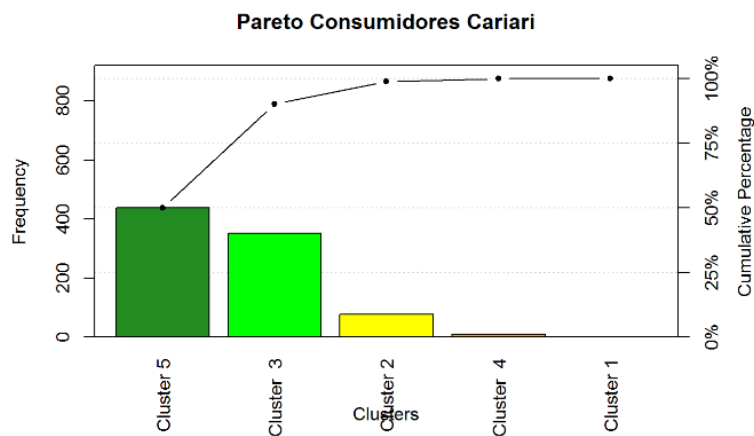


Ilustración 29 clúster Cariari

Fuente: Elaboración Propia

Gráfico Radar consumidores de Mangos

- El primer clúster cuenta con un registro que muestran un consumo de 16734 m³ anuales.
- En el segundo clúster se puede observar 77 registros con un consumo que va de 830 m³ a 2413 m³ en este grupo.

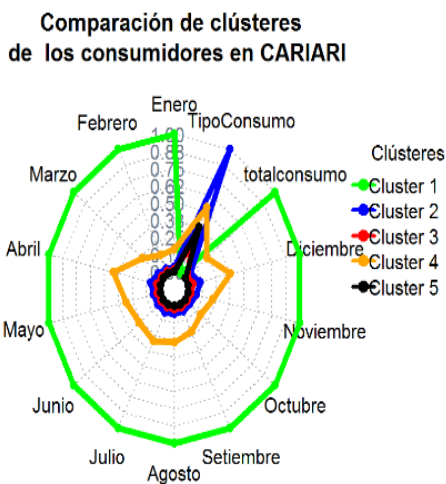


Ilustración 30 Gráfico radar Cariari

Fuente: Elaboración Propia

- El tercer clúster contiene 352 individuos que muestran un consumo bajo su rango de consumo oscila entre 363 m³ y 819 m³.
- El cuarto clúster contiene 9 individuos que muestran un consumo alto su rango de consumo es de 2771 m³ hasta 6205 anuales.

- El quinto clúster contiene 437 individuos muestran un consumo bajo su rango de consumo es de 1 m3 y 361 m3 anuales.

Análisis de los mayores consumidores en los clústeres de Cariari

Como se nos muestra en la Tabla 17 los 6 mayores consumidores del acueducto de Cariari se sitúan en el clúster 1 y 4.

Tabla 17 Grandes consumidores Cariari

Consumidores Cariari	Tipo Paja	Consumo M3	clúster
## 1 CONDOMINIO RESIDENCIAS DEL RIO	DOM	16734	1
## 2 CONDOMINIO VILLAS CARIARI EDIFICIO 3	DOM	6205	4
## 3 CONDOMINIO HORIZONTAL RESIDENCIAL EL MORTERO	DOM	5375	4
## 4 CONDOMINIO LA JOLLA	DOM	4824	4
## 5 CONDOMINIO CENTRO COMERCIAL CARIARI	DOM	4475	4
## 6 CONDOMINIO VILLAS CARIARI EDIFICIO 3	DOM	4277	4

5.5 Series de tiempo

Las técnicas de series de tiempo son utilizadas para analizar una secuencia de datos o valores los cuales son medidos en determinados momentos y ordenados cronológicamente.

Con el paso del tiempo el uso de *software* y métodos específicos aplicados para facilitar la interpretación y extracción de datos siguen aportando mejoras importantes, por lo que con el uso de estas herramientas se podrán utilizar estas series de tiempo, lo que permitirá extrapolar o interpolar los datos para predecir el requerimiento de agua de los consumidores de los distintos acueductos de Belén.

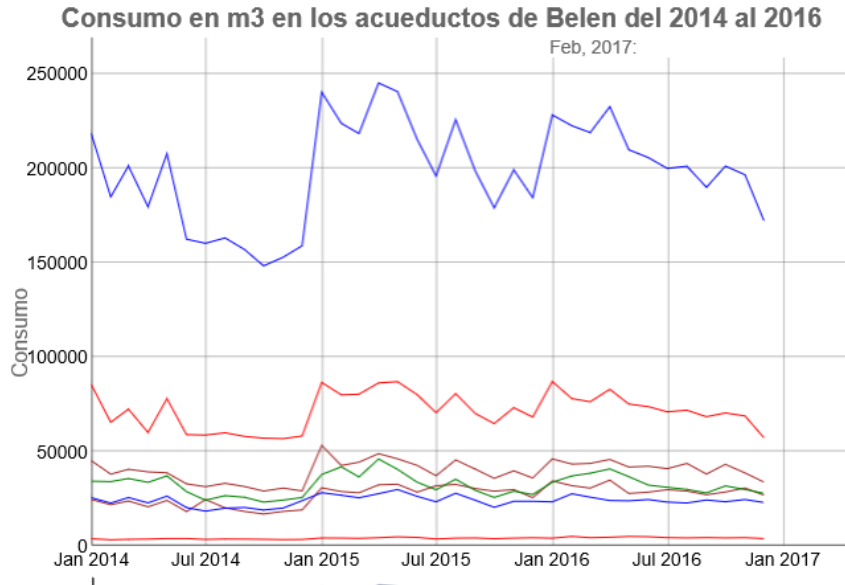


Ilustración 31 Series de tiempo generadas

Fuente: Elaboración Propia

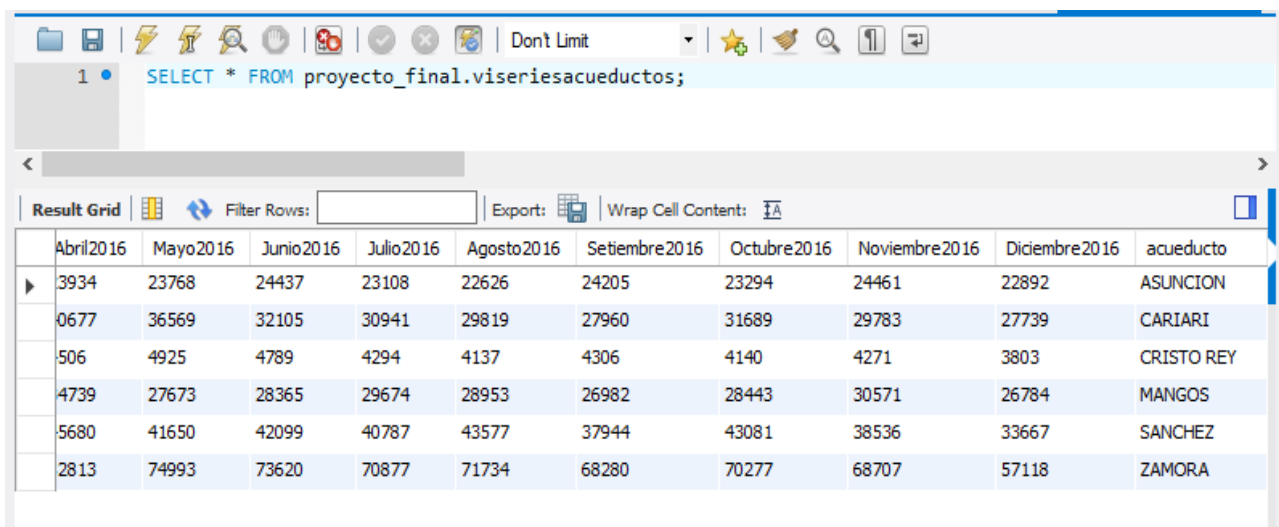
La ilustración 30 detalla los datos recopilados en todo Belén sobre las lecturas de consumo mensuales por acueducto de los años 2014, 2015 y 2016, estos datos están generados mediante una frecuencia mensual esto permitirá realizar el análisis de las series temporales.

Por ende, es fundamental utilizar métodos precisos para crear un pronóstico del consumo del agua, con lo se podrá anticipar y tomar las medidas necesarias para optimizar el almacenaje y distribución del agua.

5.5.1 Generación de las series

Para trabajar con los datos en el formato requerido fue necesario resumir los registros relacionados con los consumos de agua de forma mensual por cada uno de los acueductos, para acceder a dichos datos es indispensable utilizar dos de las vistas generadas en el motor de base de datos.

Dichas vistas extraen la información de forma agrupada y sumariada por cada uno de los acueductos tal y como se muestra en la ilustración 31.



The screenshot shows a database query tool interface. At the top, there is a toolbar with various icons. Below the toolbar, a SQL query is entered in a text area: `SELECT * FROM proyecto_final.viseriesacueductos;`. Below the query, there is a 'Result Grid' section. The grid has a header row with columns for months from April 2016 to December 2016, and a final column for 'acueducto'. The data rows show numerical values for each month and the name of the aqueduct.

	Abril2016	Mayo2016	Junio2016	Julio2016	Agosto2016	Setiembre2016	Octubre2016	Noviembre2016	Diciembre2016	acueducto
▶	3934	23768	24437	23108	22626	24205	23294	24461	22892	ASUNCION
	0677	36569	32105	30941	29819	27960	31689	29783	27739	CARIARI
	506	4925	4789	4294	4137	4306	4140	4271	3803	CRISTO REY
	4739	27673	28365	29674	28953	26982	28443	30571	26784	MANGOS
	5680	41650	42099	40787	43577	37944	43081	38536	33667	SANCHEZ
	2813	74993	73620	70877	71734	68280	70277	68707	57118	ZAMORA

Ilustración 32 Uso de vista viseriesacueductos

Fuente: Elaboración Propia

5.5.2 Explorando las series de tiempo

Una vez obtenidos los datos por medio de las vistas programadas en el motor de base de datos MySQL se podrán graficar de la siguiente manera en el lenguaje de programación R.

Selección de los datos

Las siguientes líneas de código permiten seleccionar los datos con el uso de la vista y posteriormente convertirlos en una serie de tiempo.

```
Serietotal <-sqlQuery(canal, "SELECT * FROM proyecto_final.vseriestotal;")
Serietotal<-t(Serietotal) #creacion de la serie de tiempo.
```

```
Serietotal<-as.data.frame(Serietotal)
```

La siguiente línea de código brindará el formato de colores en el gráfico.

```
cbPalette <- c("#D55E00", "#009E73", "#0072B2", "#D55E00", "#009E73",  
"#0072B2", "#D55E00", "#009E73", "#0072B2",  
"#D55E00", "#009E73", "#0072B2", "#D55E00", "#009E73",  
"#0072B2", "#D55E00", "#009E73", "#0072B2",  
"#D55E00", "#009E73", "#0072B2", "#D55E00", "#009E73",  
"#0072B2", "#D55E00", "#009E73", "#0072B2",  
"#D55E00", "#009E73", "#0072B2", "#D55E00", "#009E73",  
"#0072B2", "#D55E00", "#009E73", "#0072B2")
```

Generación del gráfico

```
ggplot(Serietotal, aes(x=Mes, y=Consumo)) + geom_bar(stat="identity" ,  
fill=cbPalette) + xlab("Año") + ylab("Promedio") + ggtitle("Consumo Mensual m3  
por mes para todo Belén") + scale_fill_brewer(palette="Set1") +theme(axis.text.x =  
element_text(angle = 90, hjust = 1)) +theme(axis.text.x = element_text(angle = 90,  
hjust = 1))
```

El resultado del código es el siguiente.

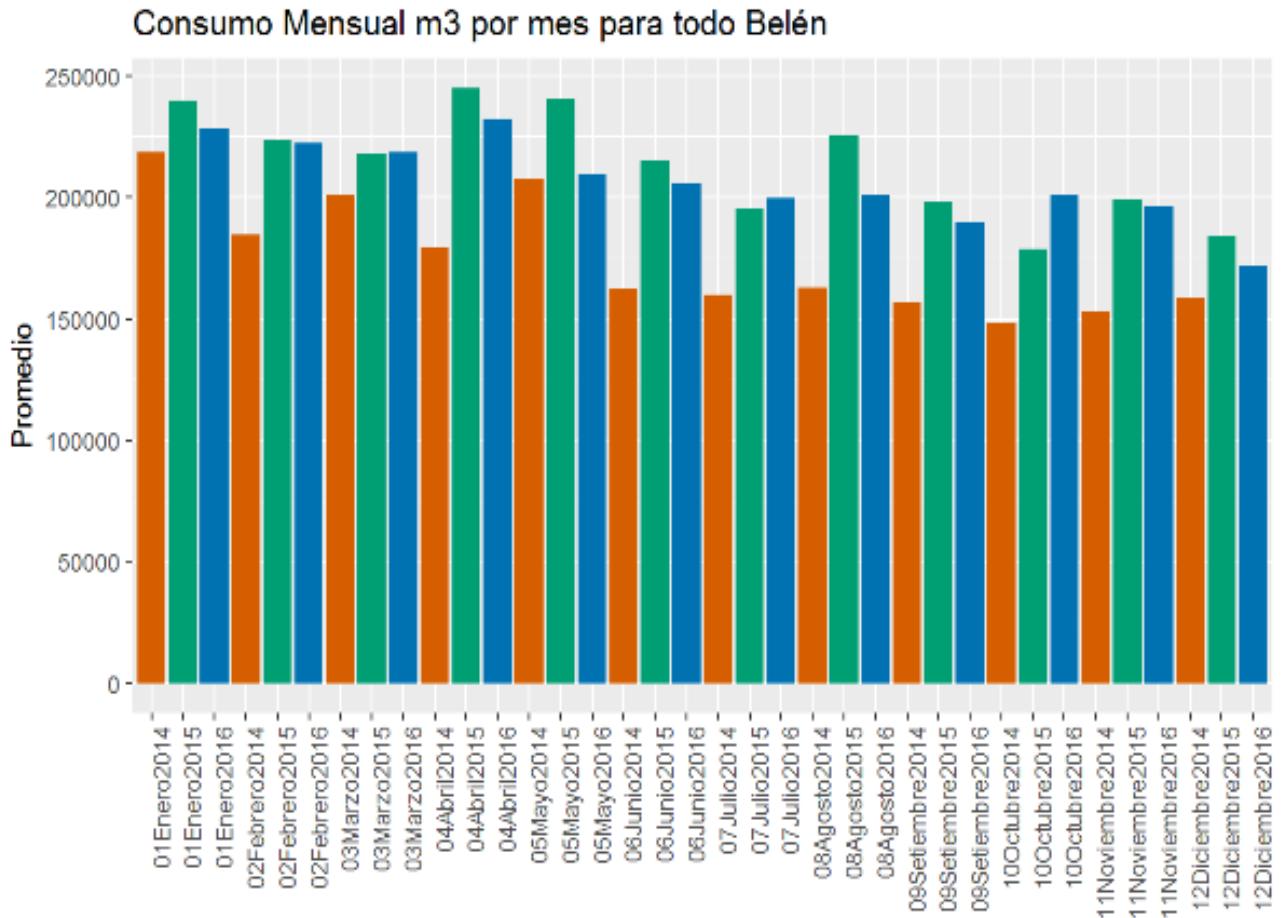


Ilustración 33 Consumo M3 mensual en Belén

Fuente: Elaboración Propia

La ilustración 32 muestra la existencia de un crecimiento en el consumo de agua en comparación de cada uno de los meses desde año el 2014. Se presentó un crecimiento importante en el 2015 y hubo una disminución en el año 2016.

Adicionalmente, se observa un mayor consumo en los primeros cuatro meses de cada año lo cual calza con la época seca en Costa Rica.

5.5.2 Analizando las series de tiempo de todo Belén

Una vez que los datos se encuentran en el formato de serie de tiempo se deben realizar distintos análisis para asegurar que son óptimos para la generación de predicciones. Es importante tener claro que entre mayor información se tenga más alta es la probabilidad de obtener un resultado muy aproximado a la realidad en el corto plazo.

Con el siguiente código se analiza la serie de tiempo.

```
hist(diff(SerieTodoBelen),prob=T,col="green", main = "Test de Normalidad
del consumo de m3 de agua \n durante los años del 2014 al 2016 en Belén")

lines(density(diff(SerieTodoBelen)),lwd=2, col="RED")
mu<-mean(diff(SerieTodoBelen))
sigma<-sd(diff(SerieTodoBelen))
x<-seq(-60000,60000,length=1000)
y<-dnorm(x,mu,sigma)
lines(x,y,lwd=3,col="black")
```

Lo que se busca es confirmar que los datos se comporten como la línea negra, la cual indica si hay un comportamiento normal en los datos.

Para comprobar la validez del comportamiento de la serie se utiliza la función de la campana de Gauss para identificar el comportamiento de los datos.

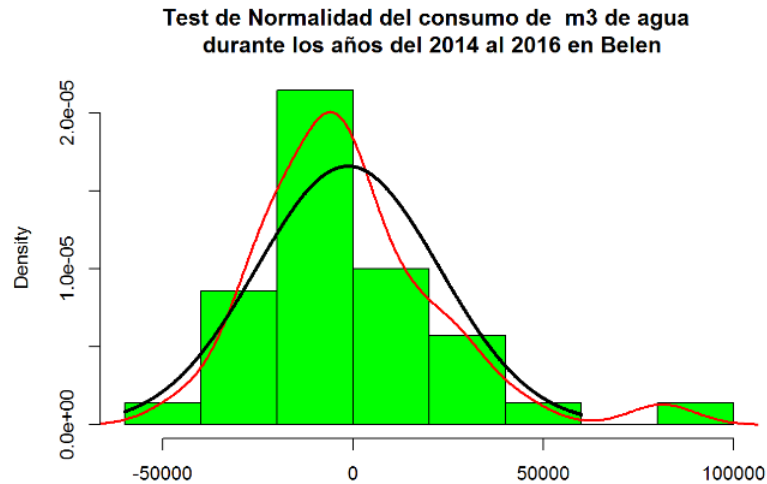


Ilustración 34 Comportamiento de los datos de la serie de tiempo en Belén.

Fuente: Elaboración Propia

La ilustración 33 muestra la serie de tiempo de Belén la cual sigue la curva normal³⁶ en su comportamiento, lo cual permitirá utilizar los datos para predicciones.

Las siguientes comprobaciones que se aplicaron a los datos están enfocadas a validar el comportamiento de los datos, periodicidad y tendencia de los mismos esto lo logramos con el siguiente código.

```
plot(stl(SerieTodoBelen,s.window="periodic"), main = "Descomposición de la serie de tiempo \n consumo de m3 de agua \n durante del 2014 al 2016 en Belén")
```

³⁶ La curva normal sirve para modelar fenómenos sociales y naturales de forma aproximada a la realidad

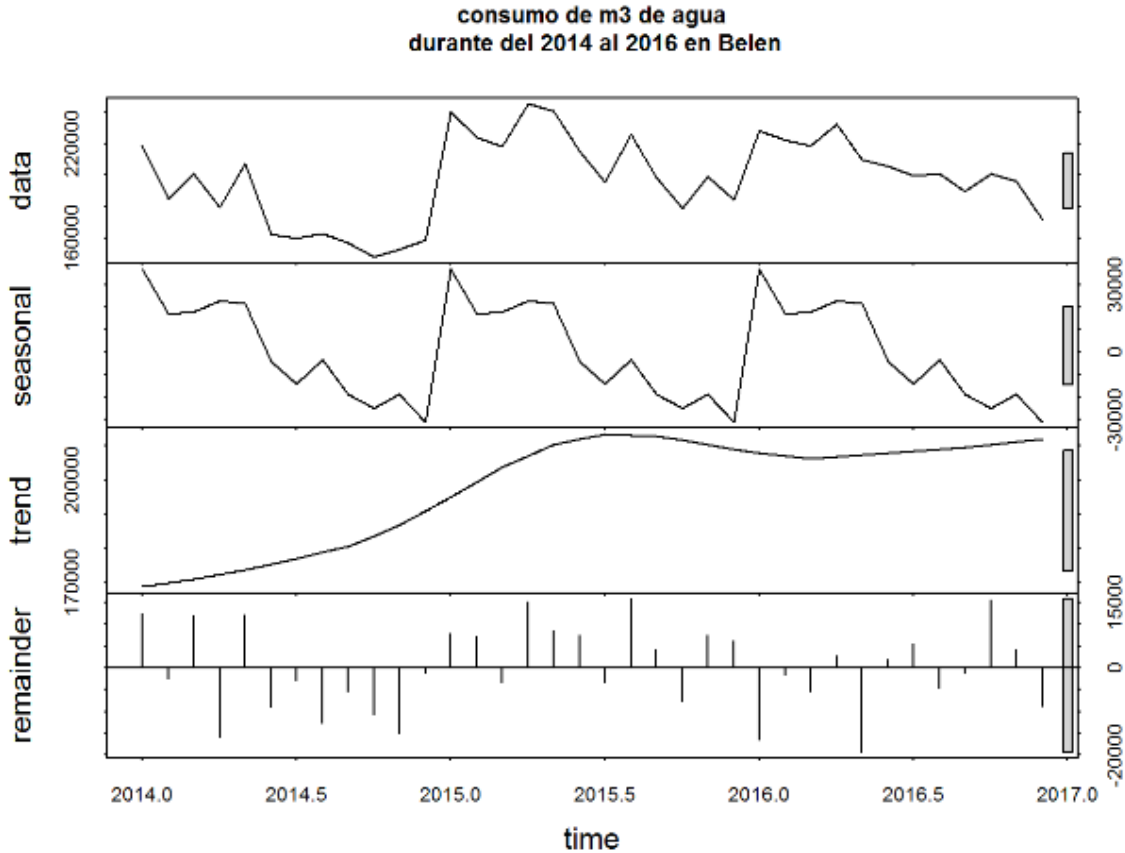


Ilustración 35 Análisis serie de tiempo Belén

Fuente: Elaboración Propia

La ilustración 34 revela que los datos de la serie de tiempo de Belén tienen un comportamiento similar en los años 2014, 2015 y 2016. Una vez más se observa que al inicio de cada año existe un mayor consumo y posteriormente el mismo empieza a disminuir.

En el apartado de trend³⁷ muestra la tendencia y se observa que con el paso del tiempo los datos muestran un crecimiento lo cual se apega a la realidad dado el crecimiento en la población y comercio en el cantón de Belén.

³⁷ Tendencia de los valores contenidos en los datos.

Finalmente se realizó un análisis de la periodicidad de mayor consumo en la serie de tiempo, se utilizó el resultado del espectro de los datos con el siguiente código del lenguaje R.

```
res<-spec.pgram(SerieTodoBelen, log = "no" , plot = F)
pos<-order(res$spec, res$freq, decreasing = TRUE)
```

El resultado obtenido es un vector³⁸ de datos organizado por la frecuencia de mayor consumo en todo el set de datos.

En este caso al graficar los resultados con el siguiente código.

```
dygraph(res, ylab="Spectrum", xlab = "Frecuencia" , main="Consumo m3 Belén
en los años 2014 al 2016") %>%
dyEvent(max1, "Maximo 1", labelLoc = "bottom" , color="red" ) %>%
dyEvent(max2, "Maximo 2", labelLoc = "bottom", color="green") %>%
dyEvent(max3, "Maximo 3", labelLoc = "bottom") %>%
dyRangeSelector()
```

³⁸ Se compone de un arreglo con un determinado número de elementos ordenados.

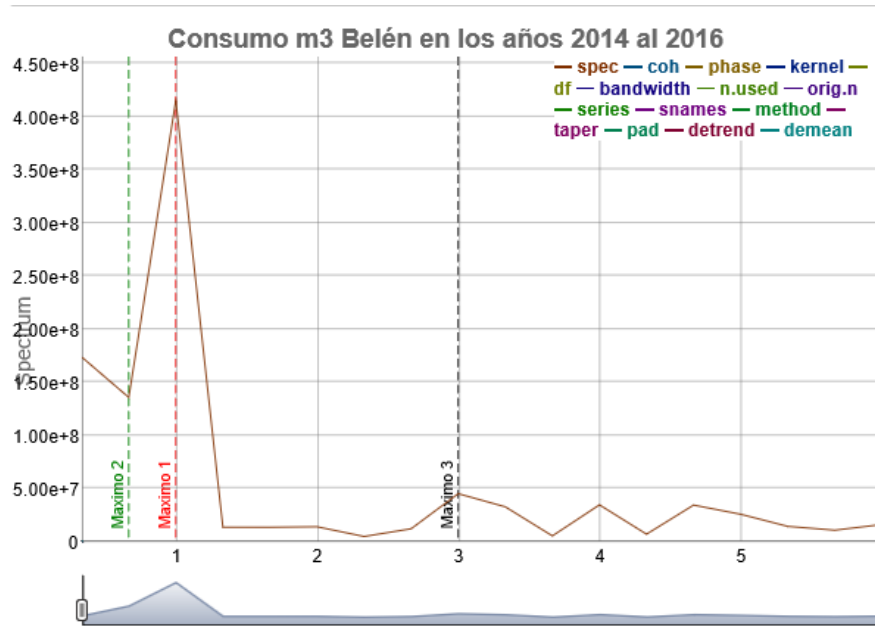


Ilustración 36 Frecuencia de mayor consumo

Fuente: Elaboración Propia

El resultado indica que los picos de mayor consumo en esta serie se dan cada 12, 18 y 4 meses respectivamente.

Este valor de mensual se obtiene dividiendo la frecuencia de la serie que en este caso es 12 por ser mensual entre el valor obtenido de cada máximo.

5.5.3 Generando predicciones en la serie de tiempo de Belén

Se corroboró que los datos contenidos en la serie de tiempo de Belén son óptimos para realizar predicciones del comportamiento de los consumidores al corto plazo. Para esto se generaron cuatro modelos distintos de predicción utilizando 30 de los meses en el set de datos para entrenamiento y los restantes 6 meses serán utilizados como prueba. Con el fin de validar la proyección de consumo se seleccionaron dos algoritmos predictivos a Box Jenkins y Holt Winters.

El Método de Box Jenkins desarrollado en 1970 utiliza funciones auto regresivas en una función construida con el nombre ARIMA por sus siglas, este método puede utilizar una o más partes de la función para generar un modelo con alta precisión para una serie.

La selección del modelo para este estudio se basa en los buenos resultados en las predicciones generadas y como lo indica el profesor Rafael de Arce es:

Una vez identificado y estimado el modelo ARIMA, se plantea su utilización para conseguir la mejor predicción de los valores a futuro de una serie a partir de su propia historia. El primer interrogante que surge se referirá a la determinación del PREDICTOR ÓPTIMO para este fin (s. f., p. 27).

Lo que significa que una vez calibrado el modelo ARIMA su efectividad es sumamente alta.

El segundo método utilizado para realizar las predicciones es el método de Holt Winters, dicha técnica fue elaborada entre los años 1957 por el doctor Charles C Holt. Peter R. Winters, en 1965, mejoró el algoritmo contemplando la estacionalidad de las series para el cálculo predictivo, la mayor ventaja de este método radica tener la ventaja de adaptar el modelo conforme nueva información se agrega (Trubetskoy, 2016).

Subdivisión de los datos

El criterio de selección el tamaño del set de datos para la generación de los modelos está basado en el estándar de minería de datos donde se recomienda utilizar 80% de entrenamiento y el restante 20% de prueba estos sets de datos se componen de la siguiente manera.

Set de datos entrenamiento contiene lecturas mensuales de 30 meses iniciando en enero del 2014 hasta junio del 2016.

Set de datos prueba contiene las lecturas mensuales de Julio del 2016 hasta diciembre del 2016.

El siguiente código del lenguaje R permite subdividir la información de la serie de tiempo permitiéndonos ahora contar con dos sets de datos uno de aprendizaje y otro de prueba.

```
STOTAL<-ts(STOTAL,start=c(2014,1),freq=12) #Convierte los datos a tipo
sBelen.aprende<-STOTAL[1:30]

sBelen.aprende<-ts(sBelen.aprende,start=c(2014,1),freq=12) #Convierte los
datos a tipo
sBelen.test<-STOTAL[31:36]
```

5.5.3.1 Creación de los modelos

Para la confección los modelos en el lenguaje de Programación R se utilizaron los sets de datos generados anteriormente para entrenar y validar cuál modelo permite obtener el menor error generado en la predicción de los meses de julio a diciembre del 2016, esto aprovechando que contamos actualmente con los datos reales.

Como se indicó anteriormente, se realizó la construcción de 4 modelos predictivos, 3 de estos fueron contruidos con el modelo Holt Winters y uno con el modelo Box Jenkins.

Comprobando la efectividad del modelo

La efectividad de cada modelo fue evaluada con la siguiente función programada con el nombre **ECM** la cual valida la diferencia entre el resultado de las predicciones de cada

modelo y lo compara con los datos de prueba, por lo que el error cuadrático medio representa la cantidad de M3 de agua en que falló el modelo, tomando en cuenta que brinde el modelo con el menor error será el más adecuado para las próximas predicciones. Lo anterior se comprueba con el estudio desarrollado por Pilar González en el Instituto de Estadística de España, en el que valida el uso de error cuadrático medio para comprobar con datos obtenidos actualmente a la efectividad de la predicción (González, 2010).

Código fuente de la validación el Error cuadrático Medio.

```
# Error Cuadratico Medio)
ECM<-function(Pred,Real) {
  N<-length(Real)
  ss<-sum((Real-Pred)^2)
  return((1/N)*ss)
}
```

Primer modelo predictivo

El primer modelo permite al algoritmo de Holt Winters seleccionar los parámetros de calibración que posee la librería de R Forecast³⁹.

Dicha librería permite auto generar parámetros de calibración para cumplir con una predicción deseada, por lo que en este caso el primer modelo recibe los datos del set de datos y un periodo de tiempo a predecir en nuestro caso es de 6 Meses.

Código fuente del primer modelo

```
mod1<-HoltWinters(sBelen.aprende)
#generamos el modelo para predecir 6 periodos
res1<-predict(mod1,n.ahead=6)
```

³⁹ Librería especializada del lenguaje R <https://cran.r-project.org/web/packages/forecast/forecast.pdf>

```
#Almacena el valor de error para validar el modelo
ecm1<-sqrt(ECM(res1,sBelen.test))
```

Segundo modelo Predictivo

El segundo modelo predictivo se generó luego de modificar de forma manual los valores Alpha, Beta y Gama para obtener un resultado optimizado sobre los resultados del primer modelo de HoltWinters.

```
mod2<-HoltWinters(sBelen.aprende,alpha=0.75,beta=1,gamma=1)
#generamos el modelo para predecir 6 periodos
res2<-predict(mod2,n.ahead=6)
#Almacena el valor de error para validar el modelo
ecm2<-sqrt(ECM(res2,sBelen.test))
```

Tercer modelo predictivo

Este tercer modelo utiliza la librería forecast y la función auto.arima de Box y Jenkins la cual genera parámetros que permiten calibrar el modelo para poder realizar predicciones.

```
> auto.arima(sBelen.aprende)
Series: sBelen.aprende
ARIMA(1,1,0)(0,1,0)[12]

Coefficients:
      ar1
      -0.558
s.e.    0.197

sigma^2 estimated as 393738157: log likelihood=-192.02
AIC=388.04  AICc=388.89  BIC=389.7
>
```

Ilustración 37 Resultado autoarima

Fuente: Elaboración Propia

El siguiente código optimiza el modelo arima para almacenarlo posteriormente en la variable fit y utilizarla para las predicciones posteriores.

```
fit<-arima(sBelen.aprende,order=c(1,1,0),seasonal=list(order=c(0,1,0),period=12))
LH.pred<-predict(fit,n.ahead=6)

#Almacena el valor de error para validar el modelo
ecm3<-sqrt(ECM(LH.pred$pred,sBelen.test))
```

Cuarto modelo predictivo

Para generar el cuarto y último modelo se utiliza siguiente función programada llamada “calibrar” programada en R la cual genera de forma automática los parámetros alpha , beta y gamma dentro de un ciclo el cual validan que se obtenga el menor error cuadrático medio de la función ECM.

La función recibe los sets de datos de aprendizaje y prueba, posteriormente compara el resultado del error cuadrático medio y lo almacena mientras se van modificando los valores alpha beta y gama. Finalmente la función guardará los parámetros que brindan el mejor rendimiento y retorna los mejores valores para la predicción.

A continuación, se muestra el código de la función calibrar:

```
calibrar<-function(serie.aprendizaje,serie.testing) {
  error.c<-Inf
  alpha.i<-0.1 # alpha no puede ser cero
  while(alpha.i<=1) {
    beta.i<-0
    while(beta.i<=1) {
      gamma.i<-0
      while(gamma.i<=1) {
        mod.i<-
        holtWinters(serie.aprendizaje,alpha=alpha.i,beta=beta.i,gamma=gamma.i)
        res.i<-predict(mod.i,n.ahead=length(serie.testing))
        error.i<-sqrt(ECM(res.i,serie.testing))
        if(error.i<error.c) {
          error.c<-error.i
```

```

        mod.c<-mod.i
    }
    gamma.i<-gamma.i+0.1
}
beta.i<-beta.i+0.1
}
alpha.i<-alpha.i+0.1
}
return(mod.c)
}

```

Realizando la predicción con la función del cuarto modelo

```

modelo<-calibrar(sBelen.aprende, sBelen.test)
res.c<-predict(modelo,n.ahead=length(sBelen.test))

#Almacena el valor de error para validar el modelo
ecm4<-sqrt(ECM(res.c,sBelen.test))

```

5.5.3.1 Selección del mejor modelo predictivo de series de tiempo para el acueducto

de Belén

Para la selección del mejor modelo se diseñó un gráfico tipo radar el cual se compone de los datos recolectados sobre los errores de cada modelo, estos resultados se agrupan en la variable errores con el fin de graficarlos.

```

errores<-rbind(err1,err2,err3,err4)
rownames(errores)<-c("Errores Holt-Winters Modelo 1","Errores Holt-Winters Modelo 2","Errores Box-Jenkins","Errores Holt-Winters Modelo Calibrado")
colnames(errores)<-c("Error Relativo","PFA","PTFA","Error Cuadratico Medio")
errores<-as.data.frame(errores)
maximos<-apply(errores,2,max)
minimos<-apply(errores,2,min)
errores<-rbind(minimos,errores)
errores<-rbind(maximos,errores)
errores

```

Finalmente, el gráfico utiliza los errores y los representa por medio de color por cada uno de los modelos.

```
radarchart(errores,maxmin=TRUE,axistype=4,axislabcol="slategray4",
centerzero=FALSE,seg=8,cglcol="gray67",
pcol=c("green","blue","red","Magenta"),
plty=1,
plwd=3,
title="Comparación de Errores Series Belén")
```

```
legenda <- legend(1.5,1, legend=c("H-Winters Auto","H-Winters Manual","Arima B-
J","H-Winters FIT"),
seg.len=-1.4,
title="Errores",
pch=21,
bty="n" ,lwd=3, y.intersp=1, horiz=FALSE,
col=c("green","blue","red","Magenta"))
```

Comparación de Errores Series Belén

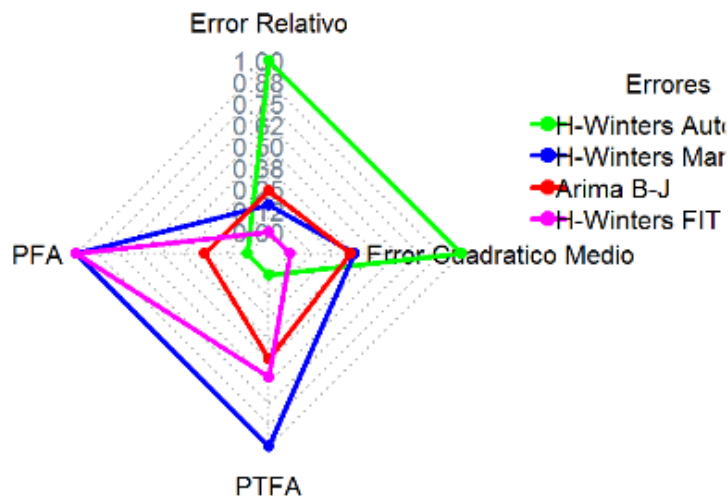


Ilustración 38 Gráfico radar modelos series de tiempo Belén

Fuente: Elaboración Propia

En el gráfico de la ilustración 37 se observa que los modelos que presentan el menor error cuadrático medio, los mejores modelos en este caso son H-Wilters Fit, Arima y Finalmente H-Winters Manual.

En el siguiente gráfico, en la ilustración 38, se puede observar la proyección del modelo HW Manual el cual realiza una proyección muy alta del consumo real a partir de

octubre, en el caso de HW Fit da una proyección abaja a partir de setiembre y finalmente el modelo de Arima está bastante por debajo del consumo si se compara con los datos originales.

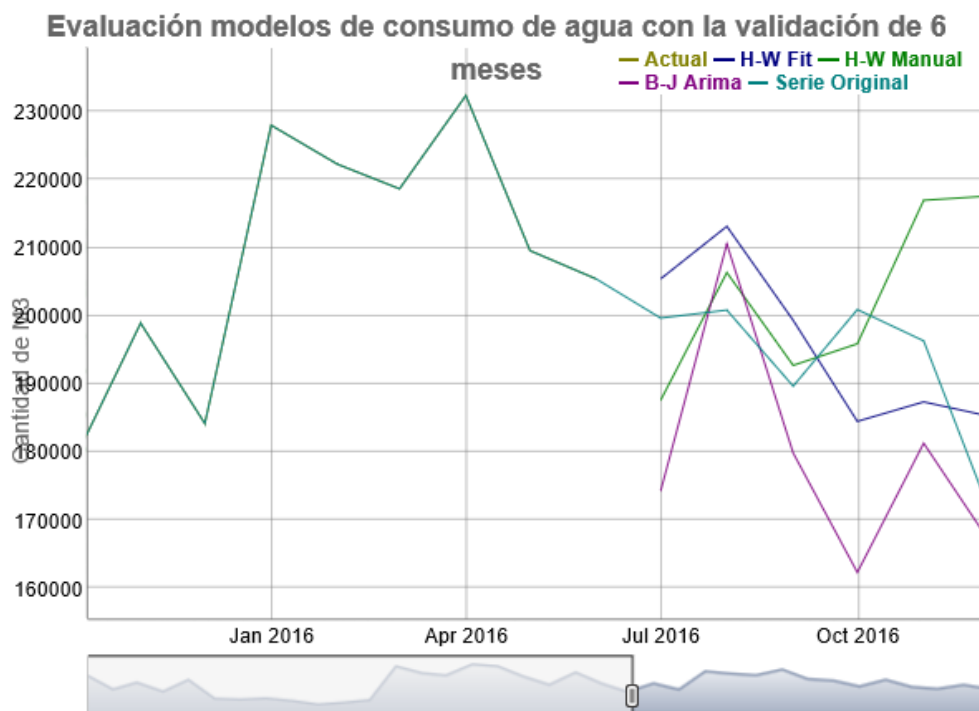


Ilustración 39 Resultados de los modelos predictivos en Belén

Fuente: Elaboración Propia

El uso del modelo para la predicción está relacionado con el fin de cada modelo, debido que el resultado es una probabilidad en la que muchas veces se desea tener un superávit en la disponibilidad del agua para el consumo de los habitantes, por lo que en este caso el mejor modelo podría llegar a ser el HW-Manual.

Para generar el siguiente gráfico de la ilustración 38 se utiliza el siguiente código fuente.

Se consolidan los datos de aprendizaje, y la data set de entrenamiento, además de los resultados generados por cada modelo dichos datos se almacenan en la variable **allce**

```
allce <- cbind(sBelen.aprende, res.c, SerieTodoBelen , res2, res.arima)
```

Utilizando la librería dygraph de R se puede obtener interacción con el archivo generado y observar el comportamiento de los datos.

Adicionalmente se grafica cada serie de tiempo para comparar los resultados obtenidos con la serie de datos original.

```
dygraph(allce, main="Evaluación modelos de consumo de agua \n con la
validación de 6 meses", ylab="Cantidad de M3") %>%
  dySeries("sBelen.aprende", label = "Actual") %>%
  dySeries(c("res.c"), label = "H-W Fit") %>%
  dySeries(c("res2"), label = "H-W Manual") %>%
  dySeries(c("res.arima"), label = "B-J Arima") %>%
  dySeries(c("SerieTodoBelen"), label = "Serie Original") %>%
  dyRangeSelector()
```

5.5.3.2 Modelo predictivo en Belén para el año 2017.

Una vez analizados los comportamientos de cada uno de los modelos a utilizar y sus parámetros de calibración obtenidos en las pruebas para la proyección de los meses de enero a junio del 2017.

Para la proyección de meses futuros se utilizó un único set de datos el cual contiene todas las 36 lecturas mensuales. Se generarán 3 modelos para elegir cual sería el mejor o realizar un mix de los 3 resultados.

El siguiente código permite generar la proyección utilizando los parámetros generados en el ejercicio de selección.

#Modelo Arima

```
auto.arima(SerieTodoBelen)

fit<-
arima(SerieTodoBelen, order=c(2,0,0), seasonal=list(order=c(1,0,0), period=12)
)
LH.pred<-predict(fit, n.ahead=6)
preds<-LH.pred$pred
#Modelo HoltWinters Manual

mod3<-HoltWinters(SerieTodoBelen, alpha=0.75, beta=1, gamma=1)
res3<-predict(mod3, n.ahead=6)

#Modelo HoltWinters Fit
mod2<-HoltWinters(SerieTodoBelen, alpha=0.1, beta=0.2, gamma=0.2)
#generamos el modelo para predecir 6 periodos
res2<-predict(mod2, n.ahead=6)
```

Después de generar los modelos almacenaremos los resultados en un DataFrame para poder graficarlo utilizando nuevamente Dygraph

```
todas.seriesbe<-
cbind(Consumo=xts(SerieTodoBelen, order.by=per_1), Arima=xts(preds, order.by=per_2),
conexionfechas=xts(conexionfechas, order.by=per_3), conexionfechasbp=xts(
conexionfechasbp, order.by=per_3), HWFit=xts(res2, order.by=per_2), conexionf
echasbp2=xts(conexionfechasbp2, order.by=per_3), HWManual=xts(res3, order.by=p
er_2) )

colnames(todas.seriesbe)<-
c("Consumo", "Arima", "ConexionFechas", "ConexionFechasbp", "HW-
FIT", "ConexionFechas2", "HW-Manual")

dygraph(todas.seriesbe, main="Consumos m3 del año 2014 al 2016 \n con 6
meses de proyección en Belén", ylab="Consumo m3")%>%

  dyRangeSelector(height = 20, strokeColor = "")%>%
  dyOptions(axisLineColor = "navy",
            gridLineColor = "lightblue")%>%
  dyRangeSelector()
```

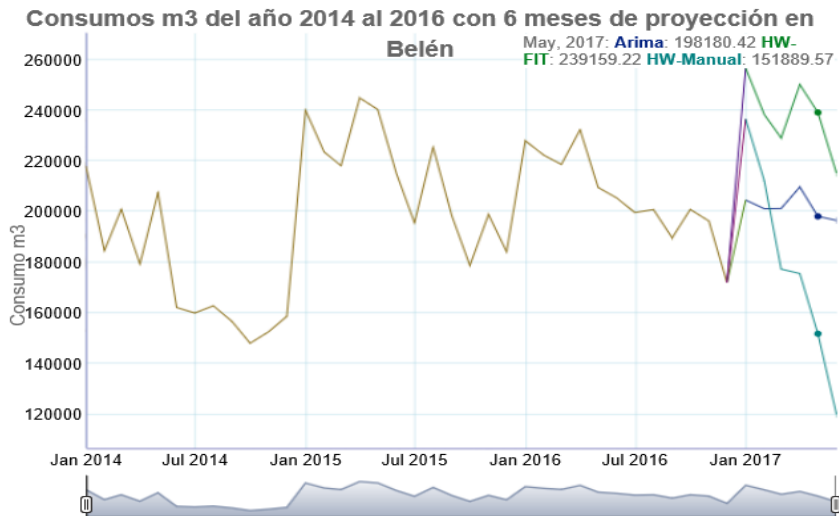


Ilustración 40 Gráfico predicción consumo 6 meses de 2017 en Belén

Fuente: Elaboración Propia

El resultado del gráfico muestra en la ilustración 39 se observa que ahora el modelo HW-Fit tiene un comportamiento más alineado con el consumo de años anteriores, por su parte el modelo Arima es muy conservador.

5.6 Analizando la serie de tiempo del acueducto de Zamora

Para la obtener de la información relacionada con a la serie de tiempo de Zamora se utilizó la vista de SQL generada previamente en la base de datos MySQL. En este caso se aplicó la cláusula where como filtro por en la vista sobre el campo llamado “acueducto” seleccionando el de Zamora.

	Enero2016	Febrero2016	Marzo2016	Abril2016	Mayo2016	Junio2016	Julio2016	Agosto2016	Setiembre2016	Octubre2016	Noviembre2016	Diciembre2016	acueducto
15	86875	77878	76194	82813	74993	73620	70877	71734	68280	70277	68707	57118	ZAMORA

Ilustración 41 Selección de los datos del acueducto de Zamora

Fuente: Elaboración Propia

El siguiente código fuente de R muestra cómo se realiza la selección de los datos dentro del lenguaje de programación R, además de ilustrar como guardar los resultados en un **DataFrame**.

```

Serietotalza <-sqlQuery(canal, "select * from
proyecto_final.viseriesacueductos where acueducto = 'ZAMORA';")

#removemos la última columna que trae el nombre del acueducto.
Serietotalza<-Serietotalza[-c(37)]

Serietotalza<-t(Serietotalza) #ordenando los datos para proceder con la
creacion de la serie de tiempo.
Serietotalza<-as.data.frame(Serietotalza)

colnames(Serietotalza)<-c("Consumo")

Serietotalza$Mes<-vectormes

```

La ilustración 41 muestra el comportamiento de los consumidores del acueducto de Zamora. Entre sus características se destaca que mantienen un consumo alto en los primeros meses del año, lo que refuerza el comportamiento de todo el cantón. Además de mostrar un incremento considerable entre los años 2014 y 2016.

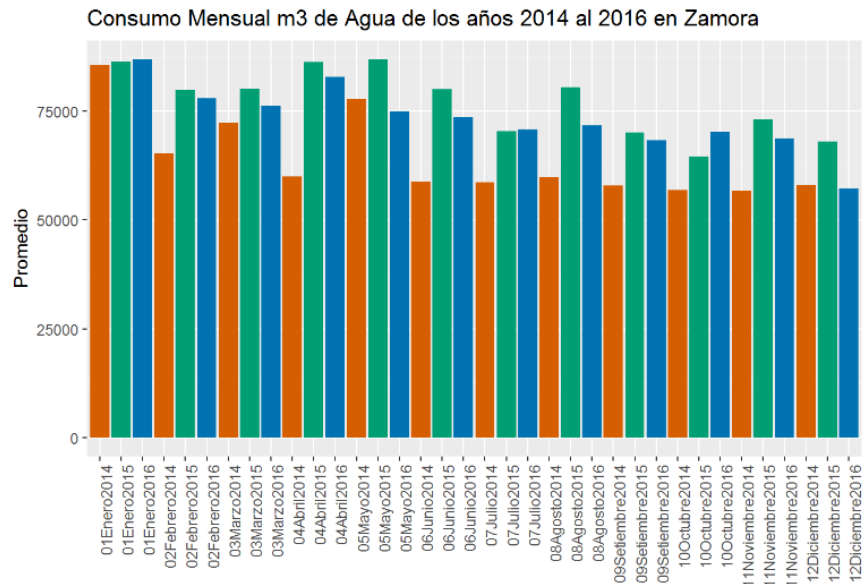


Ilustración 42 Gráfico consumo acueducto Zamora

Fuente: Elaboración Propia

El proceso de análisis de la serie valida el comportamiento de los datos los cuales se alienan con la línea azul la cual indica si hay un comportamiento normal en los datos de la serie de tiempo.

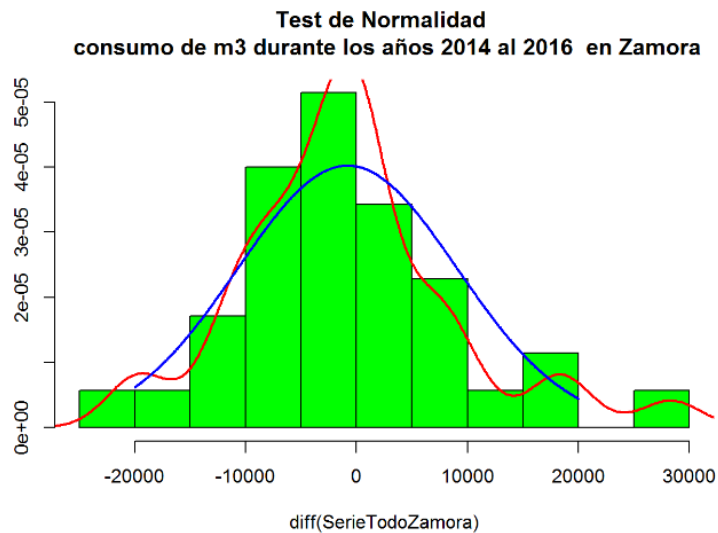


Ilustración 43 Comportamiento de los datos de la serie de tiempo en Zamora.

Fuente: Elaboración Propia

La ilustración 43 confirma que la serie de tiempo de Zamora sigue con un comportamiento normal lo cual permite utilizar los datos para predicciones.

Seguidamente, se aplicó la validación sobre el comportamiento, periodicidad y tendencia de la serie esto lo logramos con el siguiente código.

```
plot(stl(SerieTodoZamora,s.window="periodic"), main = "Descomposición de la serie de tiempo \n consumo de m3 durante los años del 2014 al 2016 en Zamora")
```

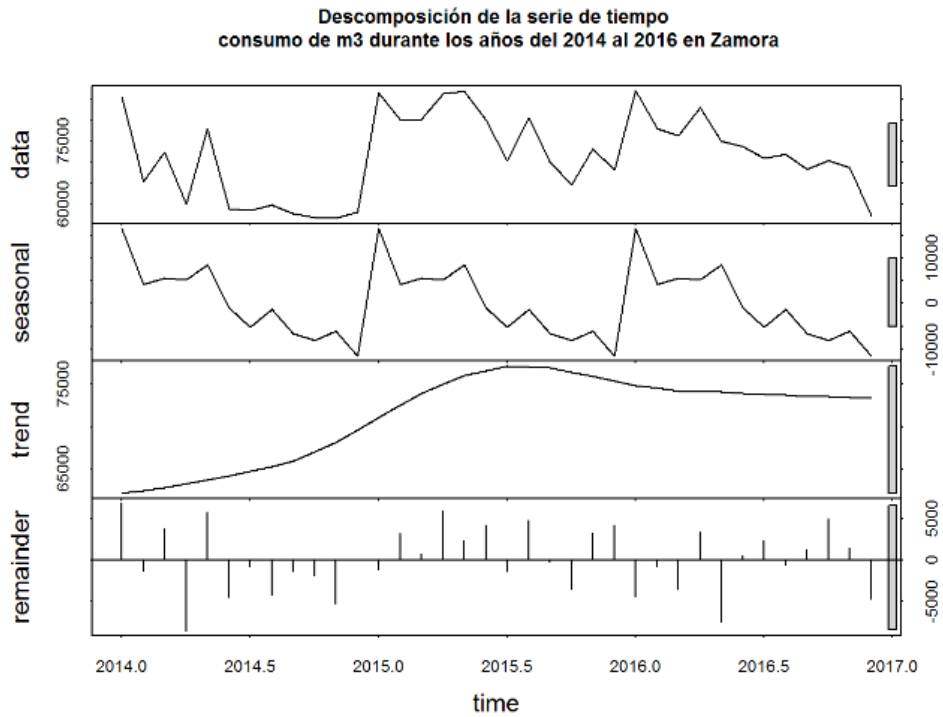


Ilustración 44 Análisis serie de tiempo Zamora

Fuente: Elaboración Propia

La ilustración 43 muestra un comportamiento similar en los tres periodos que se están analizando los cuales cubren los años 2014, 2015 y 2016, en los que al principio de cada año existe un mayor consumo y posteriormente el mismo empezó a disminuir.

En el apartado de trend muestra la tendencia de un incremento, lo cual se apega a la realidad dado el crecimiento en la población y comercio en el cantón de Belén, aunque en este acueducto hay una leve disminución en el año 2016.

Finalmente se realizó un análisis de la periodicidad de mayor consumo en la serie de Zamora, este caso al graficar los resultados podemos observar el grafico en la ilustración 44.



Ilustración 45 Frecuencia de mayor consumo en Zamora

Fuente: elaboración propia.

El resultado indica que los picos del mayor consumo en la serie de Zamora se presentan cada 12, 18 y 3 meses respectivamente.

Este valor se obtiene dividiendo la frecuencia que en este caso es 12 por ser mensual entre el valor obtenido de cada máximo.

5.6.1 Generando predicciones en la serie de tiempo de Zamora.

Para la generación de las proyecciones en este acueducto se realizó una segmentación de los datos en dos grupos con una distribución de 70 % para entrenamiento y 30 % para prueba.

Como se indicó anteriormente se generaron 4 modelos predictivos los se componen por 3 construidos con el modelo Holt Winters y uno con el modelo Box y Jenkins.

Comprobando la efectividad del modelo

La efectividad de cada modelo para la serie de tiempo de Zamora se evaluó con la función con el nombre “ER” la cual valida la diferencia entre el resultado de las predicciones de cada modelo y lo compara con los datos de prueba, por lo que el error cuadrático medio representa la cantidad de M3 de agua en que falló el modelo predictivo. Tomando en cuenta este valor, el modelo que presente el menor error será el más adecuado para las próximas predicciones.

Primer modelo predictivo

El primer modelo permite al algoritmo de Holt Winters seleccionar los parámetros de calibración que posee la librería de R Forecast.

```
mod1<-HoltWinters(szamora.aprende)
#generamos el modelo para predecir 6 periodos
res1<-predict(mod1,n.ahead=6)
```

Segundo modelo predictivo

En el segundo modelo predictivo se construyó modificando los valores Alpha, Beta y Gama para obtener un resultado forzado a valores tratando de optimizar los resultados del primer modelo Holt Winters.

```
mod2<-HoltWinters (szamora.aprende, alpha=0.3,beta=1,gamma=1)
#generamos el modelo para predecir 6 periodos
res2<-predict (mod2,n.ahead=6)
```

Tercer modelo predictivo

Este tercer modelo utiliza la librería forecast y la función auto.arima la cual genera parámetros que permiten calibrar el modelo para poder realizar predicciones.

```
fit<-
arima (szamora.aprende, order=c (1,1,0) , seasonal=list (order=c (0,1,0) ,period=12
))
LH.pred<-predict (fit,n.ahead=6)
res.arima <- LH.pred$pred
```

Cuarto modelo predictivo

El cuarto modelo utiliza la función **calibrar** la cual genera de forma automática los parámetros Alpha, Beta y Gamma comprobando dentro de un ciclo los resultados brindan el menor error cuadrático medio de la función ECM.

```
modelo<-calibrar (szamora.aprende, szamora.test)
res.c<-predict (modelo,n.ahead=length (szamora.test) )
er4<-ER (res.c, szamora.test)
```

5.6.2 Selección del modelo predictivo del acueducto de Zamora

La ilustración 48 permite observar la efectividad de los modelos al visualizar de una forma sencilla los mismos en el gráfico tipo radar.

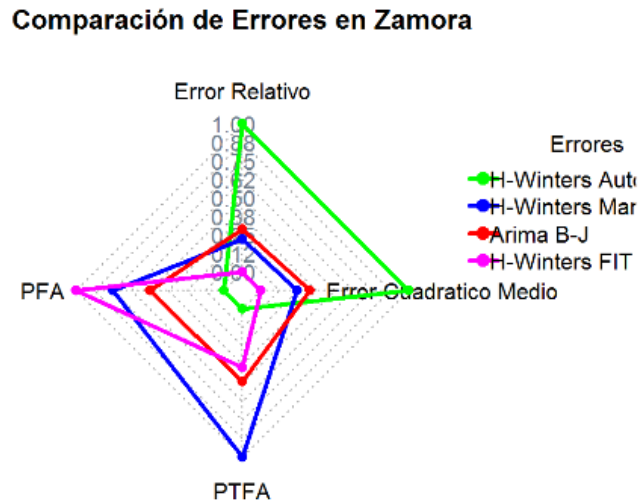


Ilustración 46 Gráfico radar modelos series de tiempo acueducto de Zamora

Fuente: Elaboración Propia

En el gráfico se observa que los modelos que presentan el menor error en el Acueducto de Zamora son H-Winters Fit, Arima y Finalmente H-Winters Manual.

Para entender mejor los modelos se utilizó el gráfico de serie de tiempo, se observa en la ilustración 46 el comportamiento de cada modelo predictivo con el set de datos de prueba.

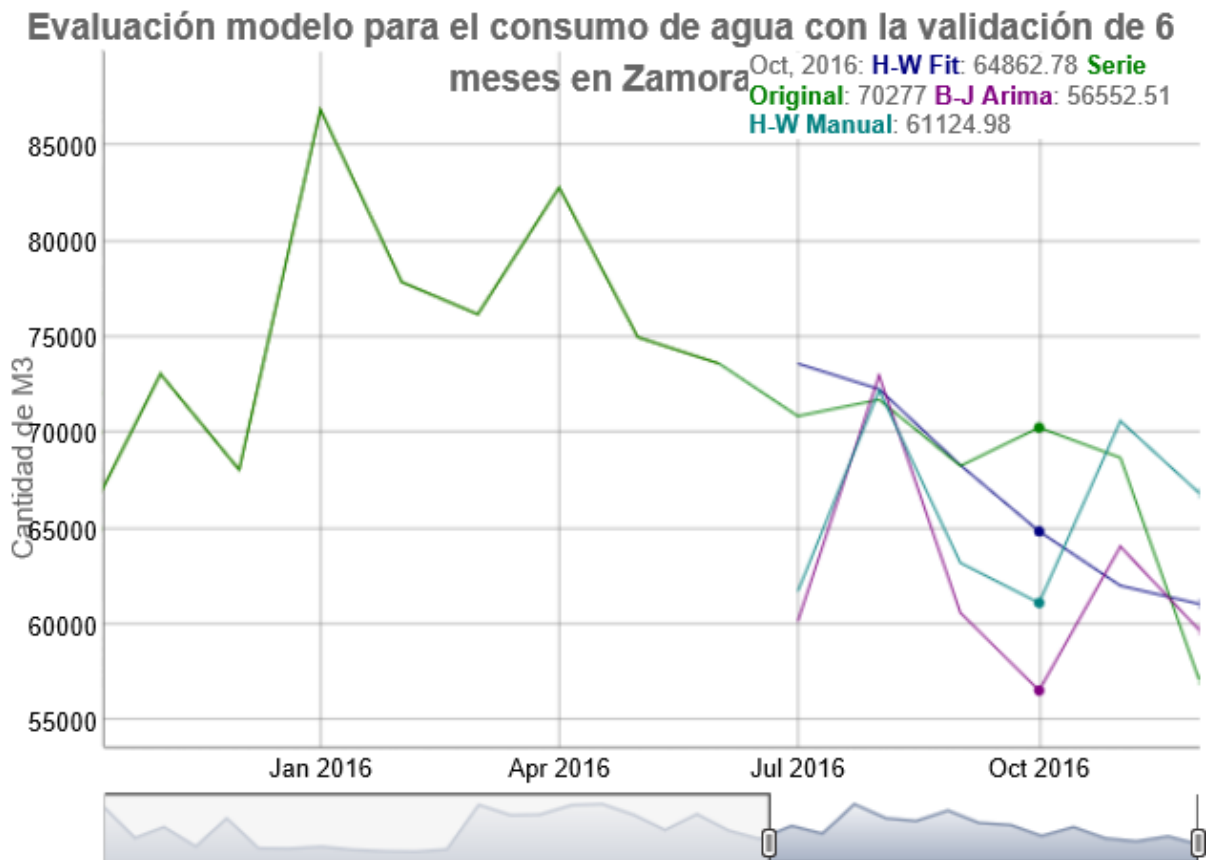


Ilustración 47 Validando los modelos predictivos para el acueducto de Zamora

Fuente: Elaboración Propia

La predicción del modelo HW Manual realiza una proyección aproximada del consumo real en los meses de agosto, septiembre, noviembre y diciembre el resto de los modelos pronostican menor consumo al real, posteriormente los modelos pronostican un incremento, lo cual se alinea con la realidad.

5.6.3 Modelo predictivo en el acueducto de Zamora para el año 2017

Luego de analizar el comportamiento de cada uno de los modelos que se utilizaron y aplicando parámetros de calibración obtenidos en las pruebas se procedió a proyectar los meses de enero a junio del 2017.

Para la predicción se utilizó un único set de datos con todas las 36 lecturas mensuales para generar los 3 modelos para elegir cuál sería el mejor o realizar un mix de los 3 resultados.

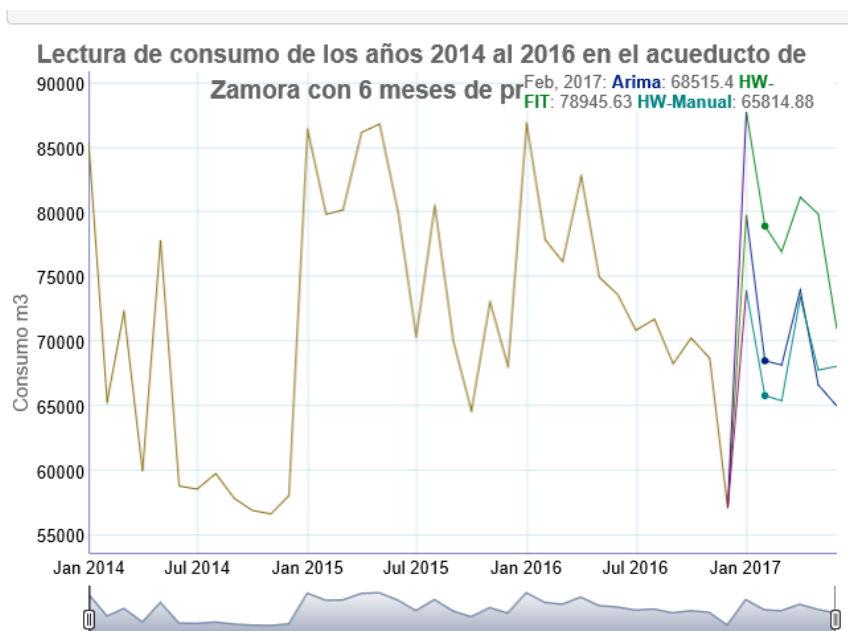


Ilustración 48 Gráfico predicción consumo 6 meses de 2017 en Zamora

Fuente: Elaboración Propia

El resultado del gráfico se muestra en la ilustración 47 en la que se observa que ahora el modelo HW-Fit tiene un comportamiento más alineado con el consumo de los años 2015 y 2016. Por otra parte, el modelo Arima y HW-Manual están por debajo del consumo futuro, es importante recalcar que a partir del año 2015 existe un incremento importante en el acueducto de Zamora por lo que el modelo HW-Fit sigue una línea apegada al 2016.

5.7 Analizando la serie de tiempo del acueducto de Mangos

Al igual que con los análisis del acueducto anterior se utilizó la vista generada en el motor de base de datos MySQL para la consulta de información relacionada con las series de

tiempo de Mangos, en este caso se utilizó la cláusula WHERE como filtro por el campo de la Base Datos llamado acueducto seleccionando el de Mangos.

El siguiente código fuente de R nos muestra cómo se realiza la selección de los datos dentro del lenguaje de programación R, además de ilustrar como salvar los resultados en un DataFrame.

```
Serietotalma <-sqlQuery(canal, "select * from
proyecto_final.viseriasacueductos where acueducto = 'MANGOS';")

#removemos la última columna que trae el nombre del acueducto.
Serietotalma<-Serietotalma[-c(37)]

Serietotalma<-t(Serietotalma) #ordenando los datos para proceder con la
creacion de la serie de tiempo.
Serietotalma<-as.data.frame(Serietotalma)
```

Luego de almacenar los datos en el DataFrame Serietotalma en su formato de serie de tiempo, se procedió a graficar la información para apreciar su comportamiento en la ilustración 48. Este gráfico muestra el comportamiento de los consumidores del acueducto de Zamora, en el que se mantiene un consumo alto en los primeros meses del año, lo que refuerza el comportamiento de todo el cantón.

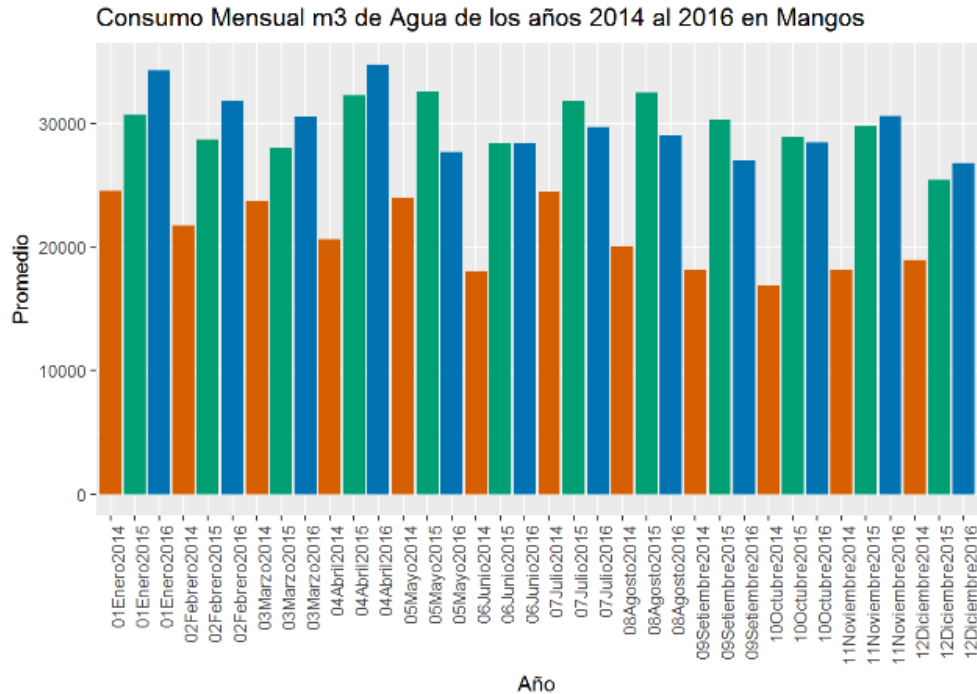


Ilustración 49 Gráfico consumo acueducto Mangos

Fuente: Elaboración Propia

Además de mostrar un incremento considerable en los años 2015 y 2016, esto se da por el gran crecimiento en la zona en condominios y en edificios de centros corporativos en la zona que se abastece por el acueducto de Mangos.

El proceso de análisis de la serie validó el comportamiento de los datos los cuales se alienan con la línea azul la cual indica un comportamiento normal en los datos de la serie de tiempo, en este caso fue la serie que hasta el momento tiene la mayor semejanza con la curva normal.

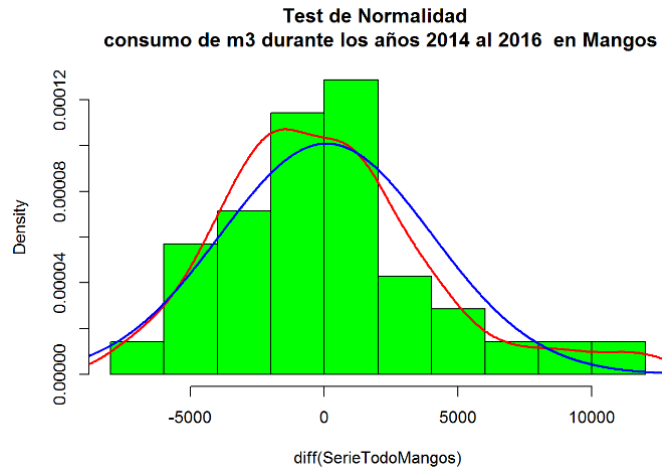


Ilustración 50 Comportamiento de los datos de la serie de tiempo en Mangos.

Fuente: Elaboración Propia

La siguiente comprobación que se aplicó a los datos fue analizar los datos, periodicidad y tendencia esto lo logramos con el siguiente código.

```
plot(stl(SerieTodoMangos,s.window="periodic"), main = "Descomposición de la
serie de tiempo \n consumo de m3 durante los años 2014 al 2016 en Mangos")
```

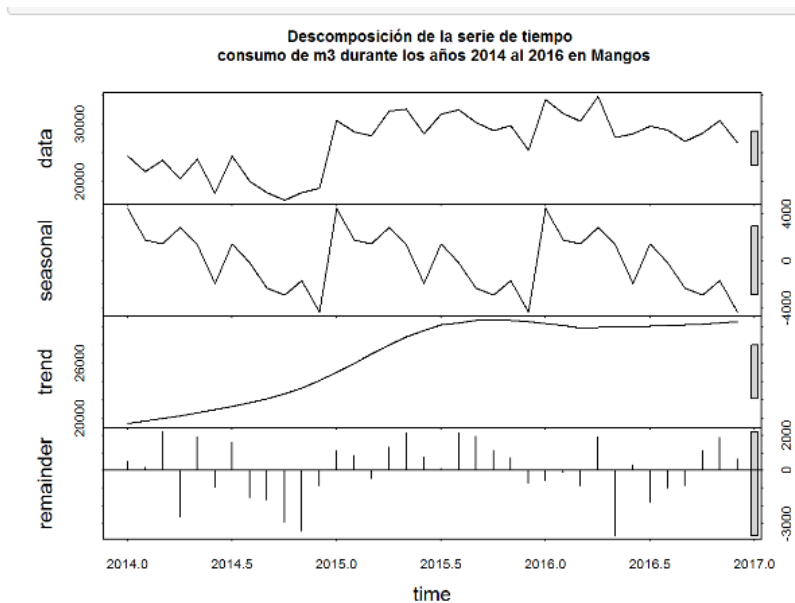


Ilustración 51 análisis serie de tiempo Mangos

Fuente: Elaboración Propia

Se puede observar en la ilustración 50 un comportamiento muy distinto luego del año 2014, en los años 2015 y 2016 se presentó un mayor incremento en el consumo, en comparación a las otras series, esta mantiene la tendencia de un consumo alto. Al inicio de cada año existe un mayor consumo y posteriormente el mismo empieza a disminuir.

En el apartado trend se observa que con el paso del tiempo los datos cuentan con un crecimiento lo cual se apega a la realidad dado el crecimiento en la población y comercio en el cantón de Belén, aunque en este acueducto hay un crecimiento importante en los dos últimos años.

Finalmente se realizó un análisis de la periodicidad de mayor consumo en la serie de Mangos, en este caso al graficar los resultados es posible observar el gráfico en la ilustración 51.

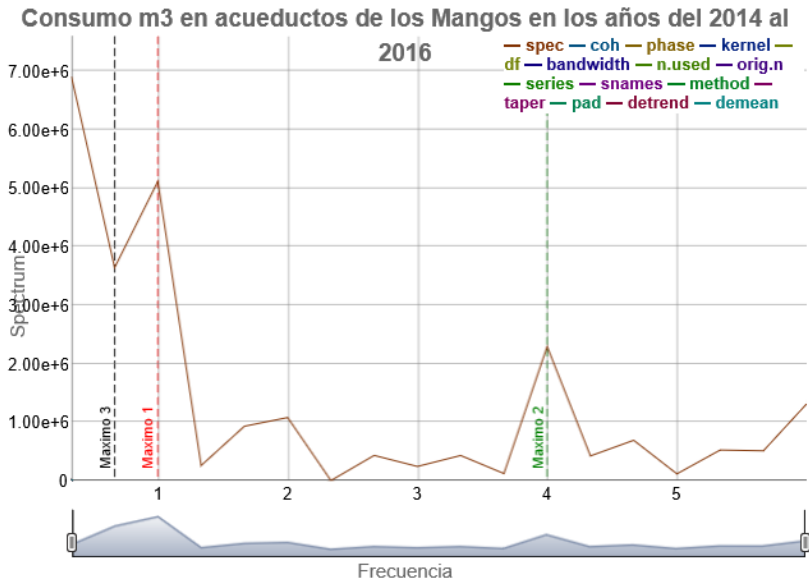


Ilustración 52 Frecuencia de mayor consumo en Mangos

Fuente: Elaboración Propia

El resultado indica que los picos de mayor consumo se dan cada 12, 3 y 18 meses respectivamente.

5.7.1 Generando predicciones en la serie de tiempo de Mangos

Para la generación de las proyecciones en este acueducto se utilizaron los datos en dos grupos con una distribución de 70 % para entrenamiento y 30 % para prueba.

Con estos datos se realizó la construcción de 4 modelos predictivos los cuales están compuesto por 3 con el algoritmo Holt Winters y uno Box y Jenkins

Comprobando la efectividad del modelo

La efectividad de cada modelo se evaluó con la función ECM la cual valida la diferencia entre el resultado de las predicciones de cada modelo y lo compara con los datos de prueba, por lo que el error cuadrático medio representa la cantidad de M3 de agua en los que fallo el modelo predictivo, tomando en cuenta este valor del modelo que brinde el menor error será el más adecuado para las próximas predicciones.

Primer modelo predictivo

El primer modelo permite al algoritmo de Holt Winters seleccionar los parámetros de calibración que posee la librería de R Forecast.

```
mod1<-HoltWinters(smangos.aprende)
#generamos el modelo para predecir 6 periodos
res1<-predict(mod1,n.ahead=6)
```

Segundo modelo predictivo

El segundo modelo predictivo se generó a partir de modificación los valores Alpha,, Beta y Gama del primer modelo, esto con el fin de generar una calibración manual y optimizar los resultados del modelo Holt Winters.

```
mod2<-HoltWinters(smangos.aprende,alpha=1,beta=0.5,gamma=0.5)
#generamos el modelo para predecir 6 periodos
res2<-predict(mod2,n.ahead=6)
```

Tercer modelo predictivo

Este tercer modelo utilizó la librería forecast y la función auto.arima, la cual auto genera parámetros que permiten calibrar el modelo para realizar predicciones.

```
fit<-arima(smangos.aprende,order=c(1,1,0))
LH.pred<-predict(fit,n.ahead=6)
res.arima <- LH.pred$pred
```

Cuarto modelo predictivo

De forma similar a los otros acueductos, para generar este cuarto modelo predictivo se utilizó la función “Calibrar” la cual genera de forma automática los parámetros Alpha, Beta y Gamma que brindan el menor error cuadrático medio esto se hace validando por medio de un ciclo y utilizando la función ECM para tener como resultado los parámetros de calibración los cuales se almacenan y utilizan en la proyección del consumo.

```
modelo<-calibrar(smangos.aprende,smangos.test)
res.c<-predict(modelo,n.ahead=length(smangos.test))
er4<-ER(res.c,smangos.test)
```

5.7.2 Selección del modelo predictivo del acueducto de Mangos

La ilustración 55 permite observar la efectividad de los modelos al visualizar de una forma sencilla el menor error cuadrático medio de cada modelo utilizando el grafico de Radar.

Comparación de Errores en Mangos

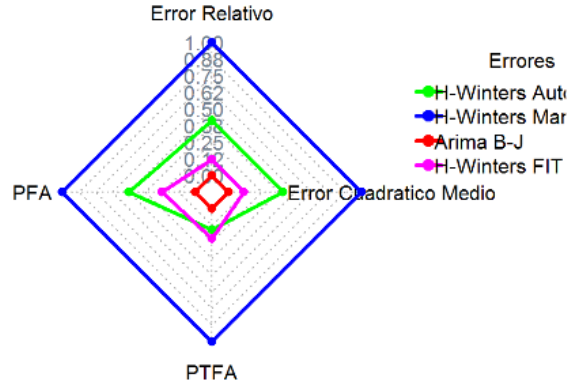


Ilustración 53 Gráfico radar modelos series de tiempo acueducto de Mangos

Fuente: Elaboración Propia

En el gráfico se observa que los modelos que presentan el menor error en el Acueducto de Zamora son Arima , H-Wilters Fit y Finalmente H-Winters Automático.

Para ilustrar de mejor manera estos modelos se utiliza el gráfico de serie de tiempo en la ilustración 57, en este se observa el comportamiento de cada modelo predictivo con el set de datos de prueba.

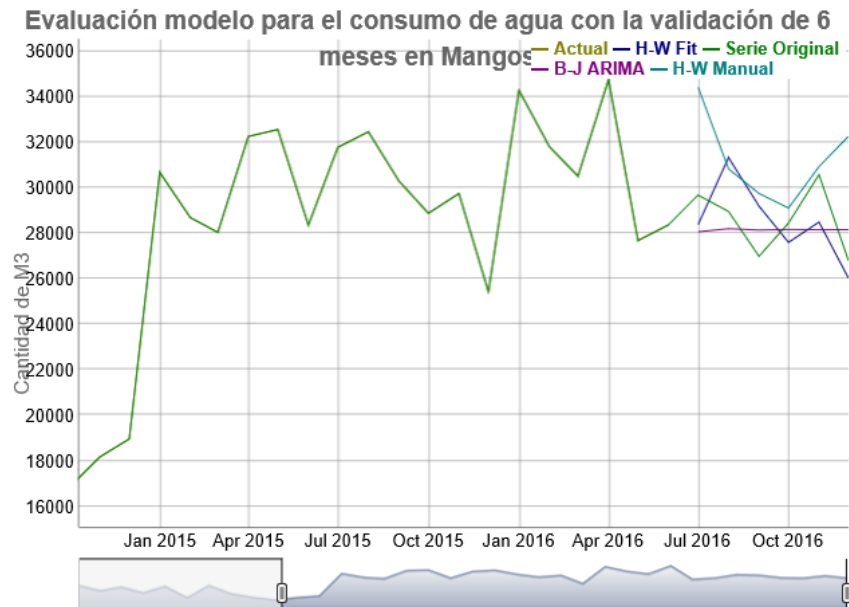


Ilustración 54 Validando los modelos predictivos para el acueducto de Mangos

Fuente: Elaboración Propia

Los resultados permiten observar que la proyección del modelo HW Fit realiza la proyección más aproximada del consumo real en los meses de setiembre, octubre y diciembre, el modelo de HW Manual sobredimensiona el consumo y el Modelo Arima queda por debajo del consumo real a partir del mes de setiembre.

5.7.3 Modelo predictivo en el acueducto de Mangos para el año 2017

Luego del análisis efectuado a los modelos con el mejor rendimiento predictivo, se realizó la proyección del consumo para los meses de enero a junio del 2017.

Para dicha proyección, se utilizó un único set de datos con todas las lecturas mensuales de consumo para generar los 3 modelos y así elegir cuál es el que brinda mejores resultados.

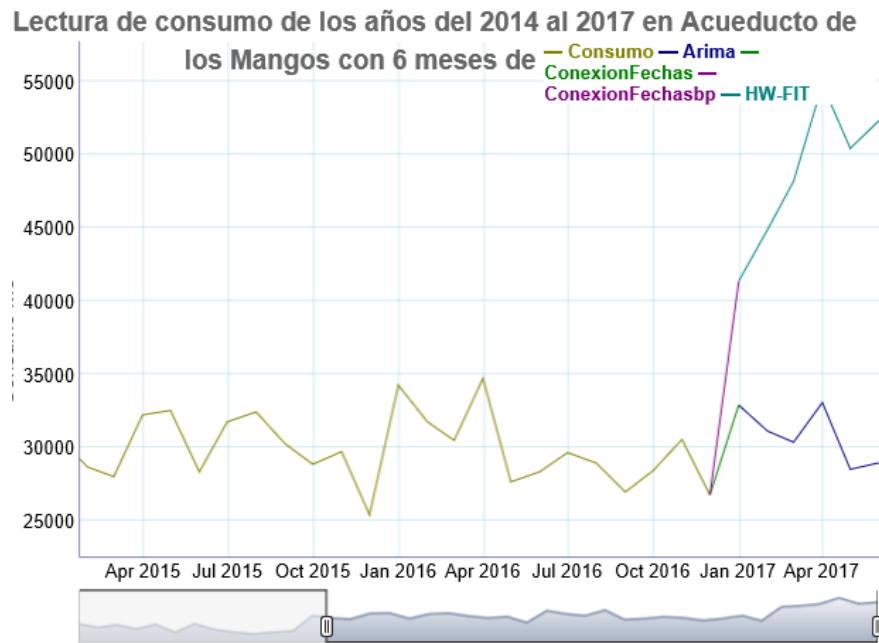


Ilustración 55 Gráfico predicción consumo 6 meses de 2017 en Mangos

Fuente: Elaboración Propia

En el resultado del gráfico de la ilustración 54 es posible observar que ahora el modelo HW-Fit tiene un comportamiento desproporcionado, si se compara con el consumo de los años 2014 al 2016, por otra parte el modelo Arima parece estar un poco más alineado con el consumo histórico del acueducto de Mangos.

5.8 Analizando la serie de tiempo del acueducto de Cariari

Para obtener la información relacionada con la serie de tiempo de Cariari se utilizó la vista de SQL, generada previamente en la base de datos MySQL. En este caso se utilizó la cláusula where como filtro en la vista sobre el campo llamado “acueducto”, seleccionando el de Cariari.

	2015	Enero2016	Febrero2016	Marzo2016	Abril2016	Mayo2016	Junio2016	Julio2016	Agosto2016	Setiembre2016	Octubre2016	Noviembre2016	Diciembre2016	acueducto
		33635	36935	38391	40677	36569	32105	30941	29819	27960	31689	29783	27739	CARIARI

Ilustración 56 Selección de los datos del acueducto de Cariari

Fuente: Elaboración Propia

El siguiente código fuente de R muestra cómo se realiza la selección de los datos dentro del lenguaje de programación R, además de ilustrar como salvar los resultados en un DataFrame.

```

Serietotalca <-sqlQuery(canal, "select * from proyecto_final.viseriesacueductos
where acueducto = 'CARIARI';")

#removemos la ultima columna que trae el nombre del acueducto.
Serietotalca<-Serietotalca[-c(37)]
Serietotalca<-t(Serietotalca)
#ordenando los datos para proceder con la creación de la serie de tiempo.
Serietotalca<-as.data.frame(Serietotalca)

```

La ilustración 55 muestra el comportamiento de los consumidores del acueducto de Cariari, entre sus características se destaca que en enero, abril y mayo hubo un alto consumo. Adicionalmente, en el año 2015 se muestra un consumo alto en comparación con los años 2014 y 2016.

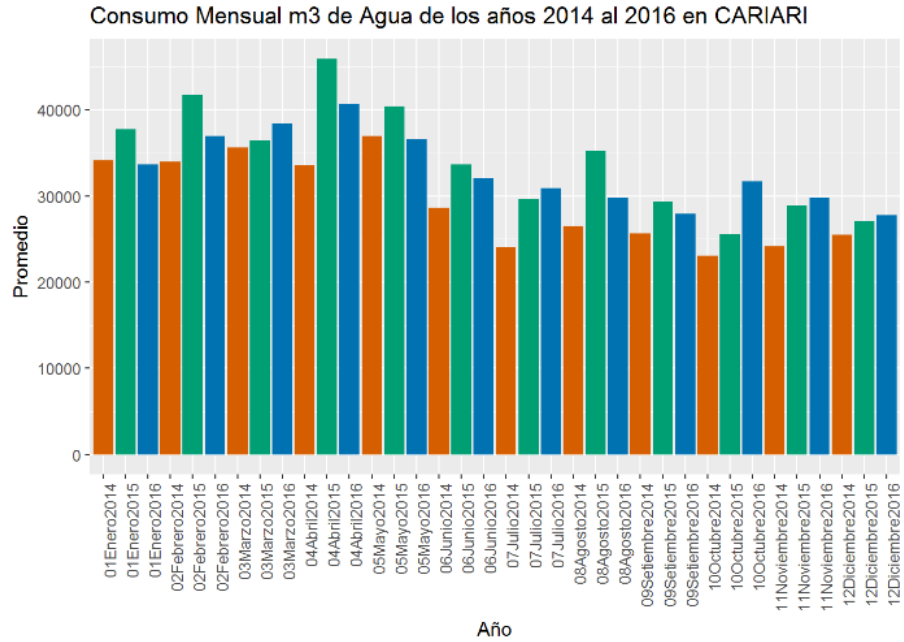


Ilustración 57 Gráfico consumo acueducto Cariari

Fuente: Elaboración Propia

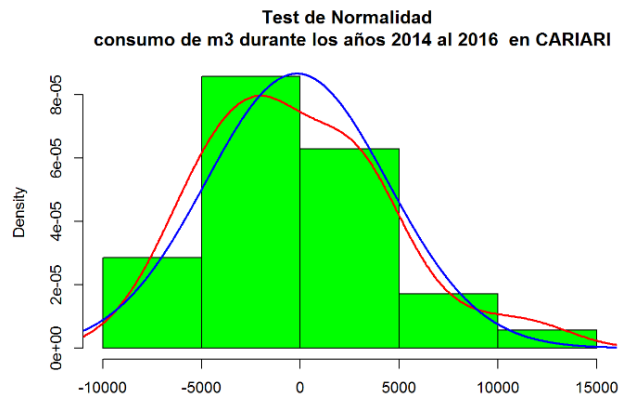


Ilustración 58 Comportamiento de los datos de la serie de tiempo en Cariari.

Fuente: Elaboración Propia

El proceso de análisis de la serie valida el comportamiento de los datos, los cuales se alinean con la línea azul, la cual indica si hay un comportamiento normal en los datos de la serie de tiempo.

La ilustración 57 confirma que la serie de tiempo de Cariari sigue con un comportamiento normal lo cual permite utilizar los datos para realizar predicciones.

Seguidamente se aplicó la validación sobre el comportamiento, periodicidad y tendencia de la serie esto lo logramos con el siguiente código.

```
plot(stl(SerieTodoCariari,s.window="periodic"), main = "Descomposición de la serie de tiempo \n consumo de m3 durante los años del 2014 al 2016 en Cariari")
```

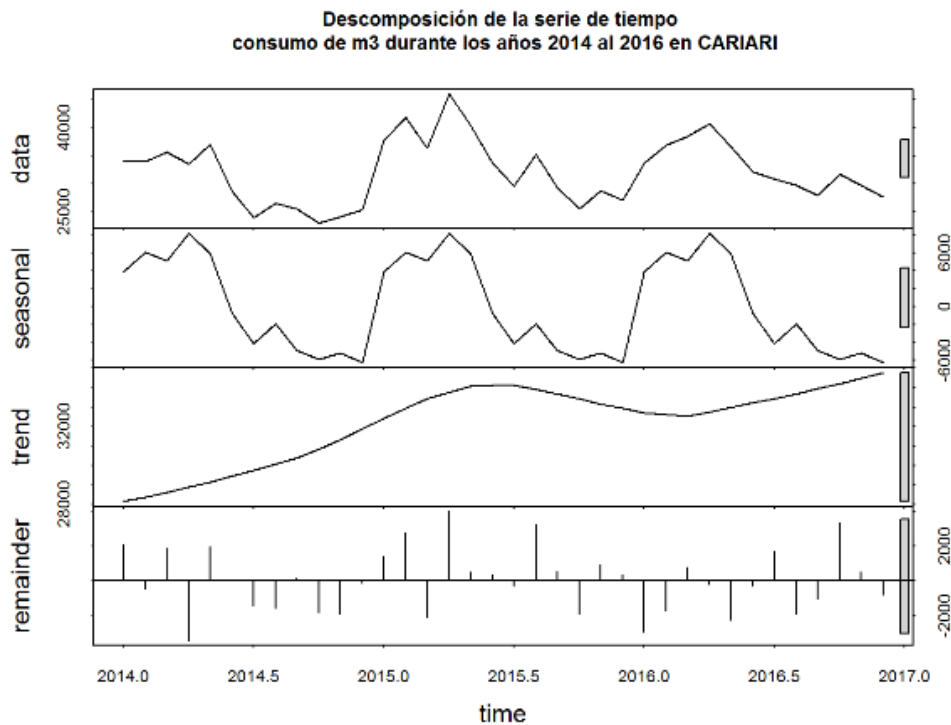


Ilustración 59 Análisis serie de tiempo Cariari

Fuente: Elaboración Propia

Se puede observar en la ilustración 58 que existe un incremento importante en el consumo de agua en el año 2015, posteriormente en el año 2016 se observa una disminución importante.

En el apartado de trend se detalla una tendencia de un incremento, lo cual se apega a la realidad, debido al crecimiento en la población y comercio en el cantón de Belén, aunque en este acueducto hay una leve disminución en el año 2016.

Finalmente se realizó un análisis de la periodicidad de mayor consumo en la serie de Cariari. En este caso, al graficar los resultados se puede observar el gráfico en la ilustración 59.

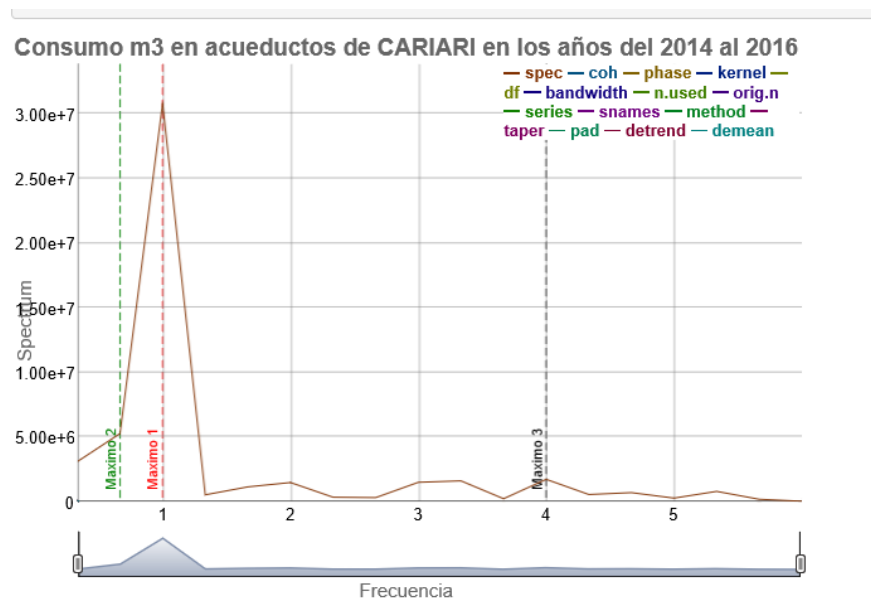


Ilustración 60 Frecuencia de mayor consumo en Cariari

Fuente: Elaboración Propia

El resultado indica que los picos del mayor consumo en la serie de Cariari se presentan cada 12, 18 y 3 meses respectivamente.

Este valor se obtiene dividiendo la frecuencia que en este caso es 12, por ser mensual, entre el valor obtenido de cada máximo.

5.8.1 Generando predicciones en la serie de tiempo de Cariari

Como se realizó con los otros acueductos y para este en particular se generaron 4 modelos predictivos los se componen por 3 construidos con el modelo Holtwinters y uno con el modelo Box y Jenkinx

Comprobando la efectividad del modelo

La efectividad de cada modelo para la serie de tiempo de Cariari se evaluó con la función con el nombre ECM, la cual valida la diferencia entre el resultado de las predicciones de cada modelo y lo compara con los datos de prueba, por lo que el error cuadrático medio representa la cantidad de M3 de agua en que falló el modelo predictivo. Tomando en cuenta este valor el modelo que brinde el menor error será el más adecuado para las próximas predicciones.

Primer modelo predictivo

El primer modelo permite al algoritmo de Holt Winter seleccionar los parámetros de calibración que posee la librería de R Forecast.

```
mod1<-HoltWinters(scariari.aprende)
#generamos el modelo para predecir 6 periodos
res1<-predict(mod1,n.ahead=6)
```

Segundo modelo predictivo

En el segundo modelo predictivo se construyó modificando los valores Alpha, Beta y Gama para obtener un resultado forzado a valores tratando de optimizar los resultados del primer modelo HoltWinters.

```
mod2<-HoltWinters(scariari.aprende,alpha=1,beta=0.5,gamma=0)
#generamos el modelo para predecir 6 periodos
res2<-predict(mod2,n.ahead=6)
```

Tercer modelo predictivo

Este tercer modelo utiliza la librería forecast y la función auto.arima, la cual genera parámetros que permiten calibrar el modelo para poder realizar predicciones.

```
fit<-arima(scariari.aprende,order=c(1,0,0))
LH.pred<-predict(fit,n.ahead=6)
res.arima <- LH.pred$pred
```

Cuarto modelo predictivo

El cuarto modelo utiliza la función “calibrar” la cual genera de forma automática los parámetros Alpha, Beta y Gamma comprobando dentro de un ciclo los resultados brinden el menor error cuadrático medio de la función ECM.

```
modelo<-calibrar(scariari.aprende,scariari.test)
res.c<-predict(modelo,n.ahead=length(scariari.test))
er4<-ER(res.c,scariari.test)
```


5.8.2 Selección del modelo predictivo del acueducto de Cariari

La ilustración 64 permite observar la efectividad de los modelos al visualizar de una forma sencilla los mismos en el gráfico tipo radar.

Comparación de Errores en CARIARI

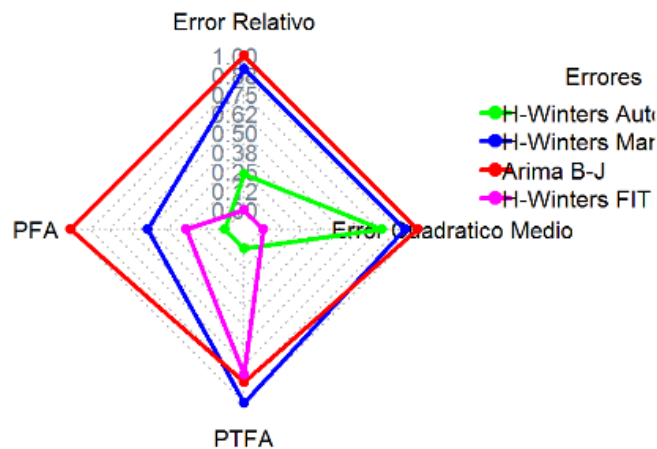


Ilustración 61 Gráfico radar modelos series de tiempo acueducto de Cariari

Fuente: Elaboración Propia

En el gráfico se observa que los modelos que presentan el menor error en el Acueducto de Cariari son H-Winters Fit, H-Winters Automático y Finalmente H-Winters Manual.

Para entender mejor los modelos se utilizó el gráfico de serie de tiempo, se observa en la ilustración 66 el comportamiento de cada modelo predictivo con el set de datos de prueba.

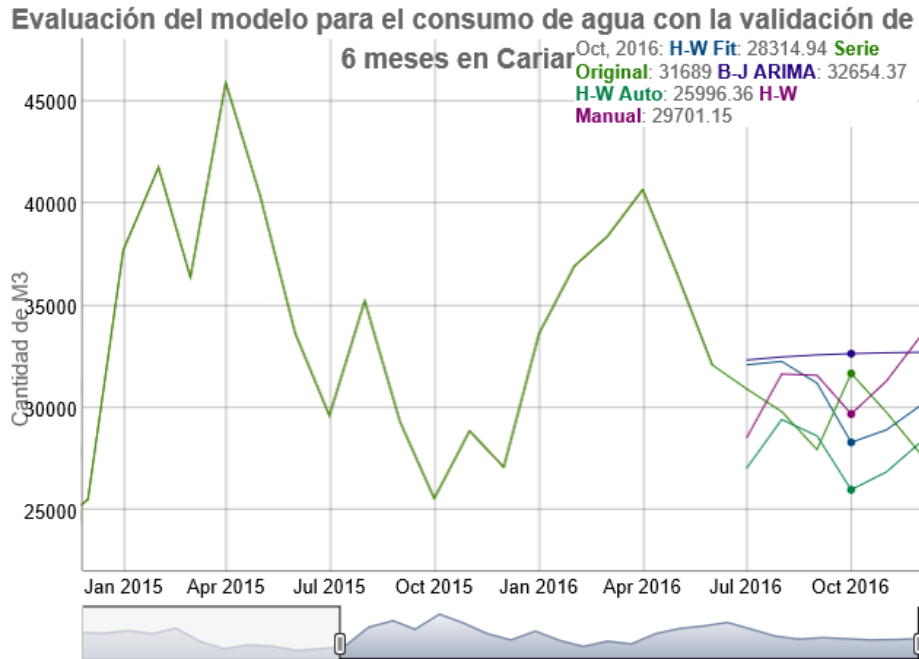


Ilustración 62 Validando los modelos predictivos para el acueducto de Cariari

Fuente: Elaboración Propia

Los resultados de las predicciones en el gráfico de radar son muy distintos por lo que se realizó un análisis más detallado de cada modelo iniciando por el modelo H-W Manual el cual creó una proyección con tendencia alta, el modelo arima no generó mayor variación entre meses y tiene una tendencia estática pero alta, el modelo H-W Auto tiene una proyección baja para los meses a partir de octubre. Finalmente, en el modelo HW Fit se acercó a la realidad, pero en el mes de octubre proyecta un consumo menor al real.

5.8.3 Modelo predictivo en el acueducto de Cariari para el año 2017

Luego de analizar el comportamiento de cada modelo obtenido en las pruebas, se procedió a proyectar de los meses de enero a junio del 2017 con los modelos HW FIT , HW Auto y Arima.

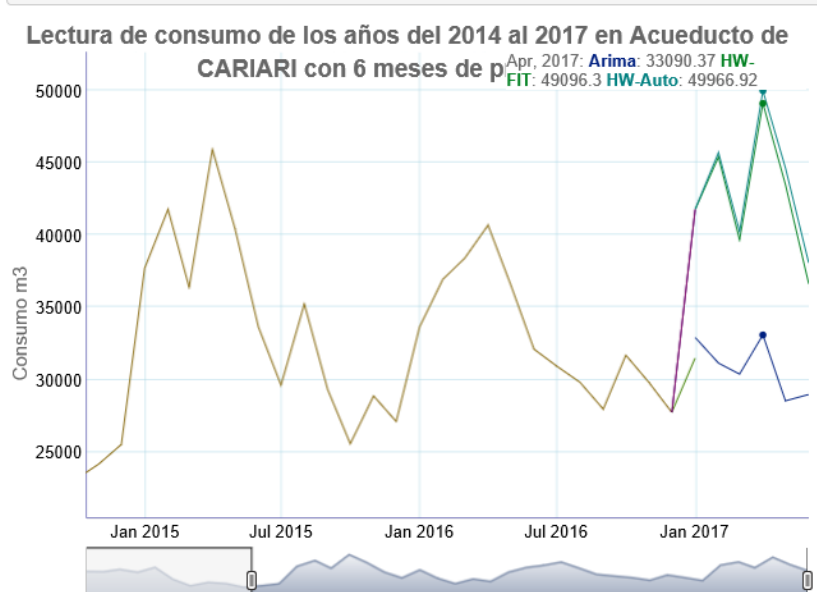


Ilustración 63 Gráfico predicción consumo 6 meses de 2017 en Cariari

Fuente: Elaboración Propia

La ilustración 62 permite observar que el modelo HW-Fit y HW-Manual tienen un comportamiento más alineado con el consumo del año 2015, por otra parte el modelo Arima está por debajo del consumo futuro,

Capítulo 6. Conclusiones y recomendaciones

6.1 Conclusión sobre la selección los datos requeridos para la construcción de modelos de minería de datos.

El trabajo efectuado en la selección y depuración de los datos para este proyecto requirió un esfuerzo importante dado que los formatos de los archivos entregados contenían una alta variabilidad en sus formatos, esto requirió aplicar un proceso de limpieza, transformación y carga de los datos a una base de datos relacional facilitando el consumo futuro de la información de forma estructurada.

Dicha tarea facilito la generación de distintos sets de datos con información requerida para los de modelos de segmentación y series de tiempo de este proyecto, los cuales se ejecutaron de forma exitosa además se logró asegurar la calidad de los datos obteniendo así resultados precisos.

6.2 Conclusión sobre aplicar el modelo de minería de datos de series de tiempo en la evaluación del uso y abastecimiento del recurso hídrico.

Los modelos generados brindaron una proyección aproximada con los datos de entrenamiento y prueba, luego de generar las predicciones se pudo poner a prueba la predicción con los datos de los primeros seis meses del año 2017, lo cual muestra que los modelos brindan un resultado bastante aproximado a la realidad siempre y cuando no existan cambios abruptos en el consumo, lo cual puede afectar la efectividad de los modelos y esto se comprueba con la figura



Ilustración 64 Resultado Modelos Predictivos Belén

Fuente: Elaboración Propia

Ambos modelos mantienen un patrón similar al consumo real, en el caso del modelo Arima falló en 5% en promedio manteniéndose por debajo del consumo real.

El Modelo Holt Winters FiT tuvo un 11% de error sobre el consumo real, el modelo genera proyecciones que superan el consumo real.

En este caso es recomendable utilizar el promedio de proyección de ambos modelos lo cual nos brinda un error de un 3%

Tabla 18 Evaluación de los modelos

Mes	Modelo Arima	Consumo Real	Modelos Combinados	Modelo HwFIT
Enero	204593	219469	230652,5	256712
Febrero	201192	221368	219766,5	238341
Marzo	201286	202438	215184,5	229083
Abril	209762	200280	229961,5	250161
Mayo	198180	228845	218669,5	239159
Junio	196590	206689	205874,5	215159
	5%		3%	10%

Se pudo comprobar que el uso de los datos para la predicción y correcta clasificación de clientes brindará beneficios a la comunidad de Belén, asimismo el uso de los datos históricos revela que existe una relación importante en los patrones de comportamiento, por lo que al tomar ventaja de estos análisis los profesionales del departamento de informática, así como los expertos del acueducto, podrán trabajar en conjunto en el futuro para continuar con la estimación de predicciones cada vez más cercanas a la realidad.

6.3 Conclusiones del aplicar modelos de minería k-means en el agrupamiento de las comunidades

La clasificación de consumidores que se realizó permite observar grupos de consumidores que mantienen comportamientos estables.

Los grandes consumidores mantienen su consumo durante todo el año los mismos en todos los afluentes.

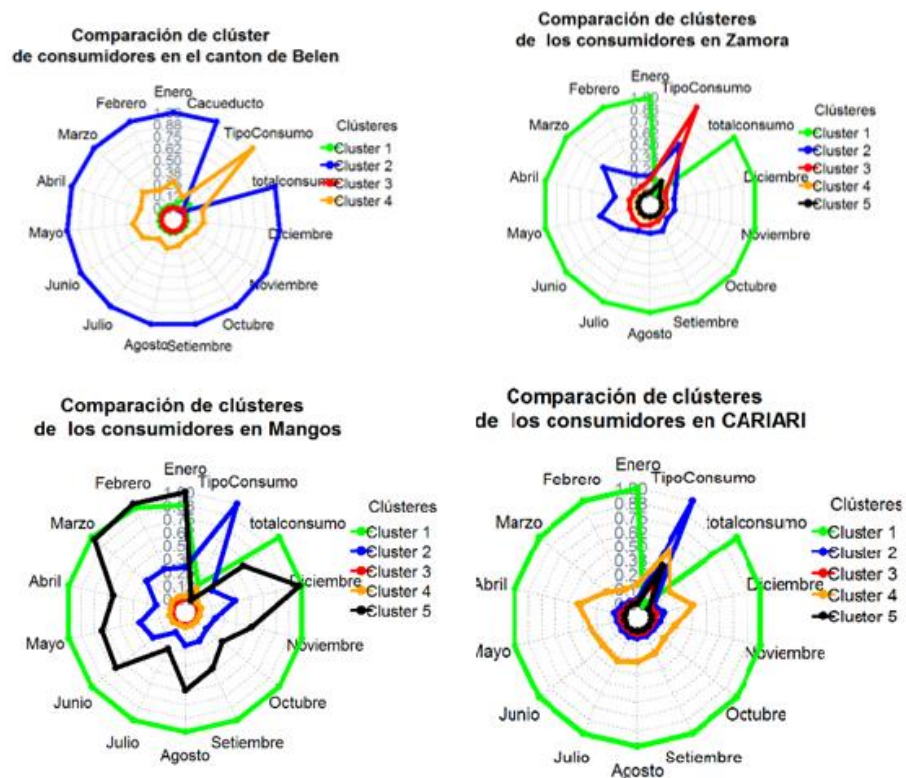


Ilustración 65 Comparación entre las segmentaciones

Fuente: Elaboración Propia

Se observa que entre los resultados de los mayores consumidores de los acueductos dominan los Condominios habitacionales y centros corporativos.

Por esta razón, se considera relevante evaluar la cantidad de casas de habitación que se desarrollan en cada uno de los condominios, esto con el fin de determinar si existe una proporción adecuada entre consumo y cantidad de hogares.

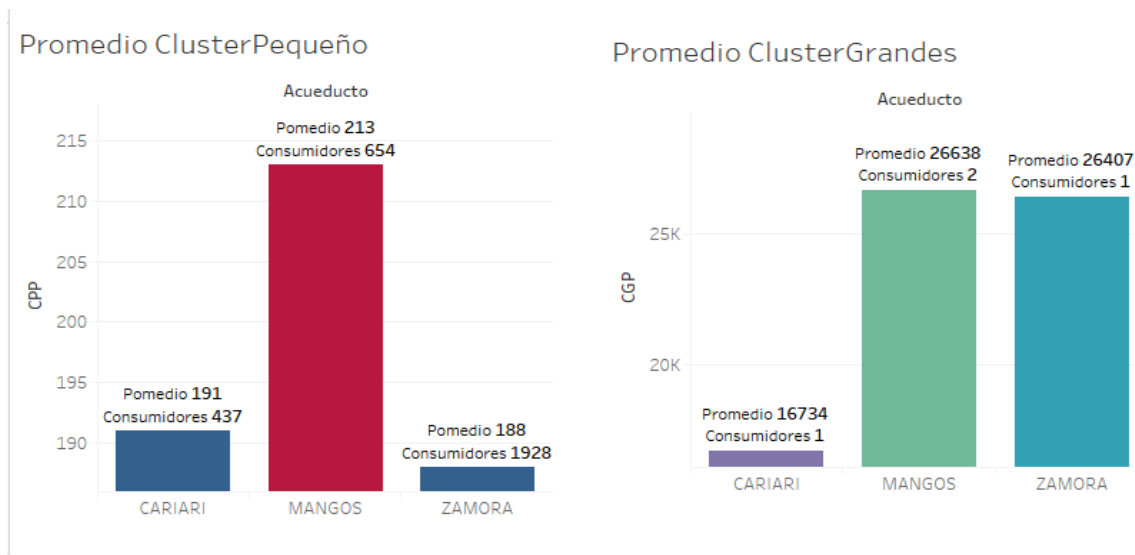


Ilustración 66 Rangos consumo

Fuente: Elaboración Propia

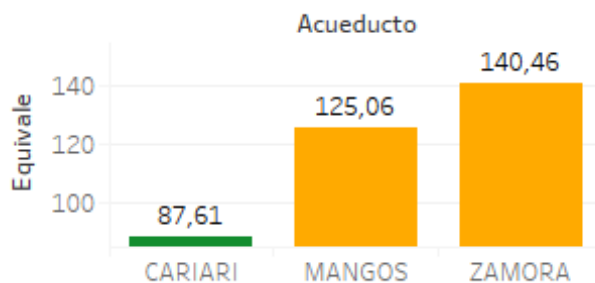
En el caso de los centros corporativos se debe evaluar la implementación de planes de disminución en el consumo de agua, la estrategia debe incluir el uso de servicios sanitarios eficientes, así como evitar la irrigación constante de las áreas verdes.

Adicionalmente el estudio demostró que el área abastecida por el acueducto de los Mangos es una de las más cargadas con condominios de alto consumo, por lo que regular la construcción de estos proyectos es una buena alternativa.

La ilustración 65 muestra la cantidad de hogares que pueden ser abastecido por cada gran consumidor en cada acueducto.



40



41

Ilustración 67 Equivalente grandes consumidores hogares

Fuente: Elaboración Propia

En el mes de mayo del 2017 se inició en desarrollo un nuevo centro corporativo en la zona de los Mangos, además una extensión del centro corporativo el Cafetal con una construcción de 3 edificios, lo cual incrementara el consumo en la zona hasta 40 % debido al comportamiento histórico de consumo.

6.4 Entrega a los encargados del acueducto de Belén los resultados del proyecto, así como las recomendaciones en la optimización de los recursos

Luego de mostrar los resultados obtenidos a los responsables del acueducto de Belén, sobre logró comprobar que los datos relacionados al consumo y comportamientos de los consumidores obtenidos las segmentaciones de la población permitirán abordar de forma confiable a consumidores en campañas de reducción del consumo del agua.

⁴⁰ <http://icons.mysitemyway.com/legacy-icon/078552-blue-jelly-icon-business-home5/>

⁴¹ <http://clipartix.com/building-clipart-image-23046/>

Adicionalmente se comprobó que el crecimiento de la industria y población en Belén podrá ser proyectada mediante el uso de series de tiempo logrando anticipar el consumo con porcentajes de error muy bajos.

Adicionalmente los dashboards generados durante el proyecto permitirán a los responsables explorar y evaluar comportamientos de consumo de una forma sencilla.

Finalmente se mostró especial interés en continuar aplicando este tipo de técnicas en los datos con distinta regularidad con los datos obtenidos de forma mensual.

6.5 Recomendaciones

- Durante el desarrollo de este proyecto se identificó que los encargados del acueducto contaban con la información del consumo hasta el momento que se realizó el cambio de sistema. Posteriormente, la consulta de los datos debe ser solicitada al proveedor, lo cual evita obtener la información de manera expedita, por lo que contar con un módulo de auto consulta facilitará que se obtenga la información de manera ágil.
- Se recomienda contar un único set de datos de los consumidores, el cual debe incluir el código o nombre del acueducto del cual el consumidor se abastece, esto permitirá tener los datos necesarios consolidados. Una manera de realizar esta inclusión es mediante un reporte utilizando una vista SQL, ya que anteriormente la información suministrada estaba almacenada en archivos de Excel lo cual requiere aplicar un proceso de depuración más lento.

- Se identificó que algunos condominios o locales comerciales no tienen definido su tipo de paja, aunque para este estudio no se realizó una subdivisión a mayor detalle, es recomendable depurar dicho identificador para estudios futuros.
- Es importante agregar la cantidad de consumidores internos por medidor, sobre todo en condominios o centros corporativos, estas estadísticas permitirán patrones de consumo aún más detallados.
- El criterio de los encargados del acueducto permitirá validar y ajustar los modelos generados, los datos que se representan visualmente son una herramienta más para optimizar la entrega del agua a todos los habitantes del cantón de Belén.
- Los tanques de abastecimiento deben llegar a restringir el consumo de altos consumidores para poder garantizar el recurso a los habitantes más críticos del cantón.
- Iniciar con un registro automatizado relacionado con los niveles de agua en los tanques de abastecimiento permitirá detectar consumos anormales y posibles fugas en el futuro, este evento puede ser notificados por medio de *triggers*⁴² en el servidor que almacene la información.

Capítulo 7. Reflexiones finales

Como profesional he entendido que la constante adquisición de conocimiento junto a la innovación debe ser parte de nuestro ADN, por esto debemos cuestionar y visualizar patrones

⁴² Procedimiento que se ejecuta automáticamente cuando se produce un evento en el servidor de bases de datos.

en nuestro entorno que pueden ser ajustados y permitirán optimizar el futuro. De la misma manera, los datos pueden ser explorados con las técnicas adecuadas para crear procesos más eficientes e innovar con la entrega de servicios de mayor calidad.

Debemos recordar que el uso de metodologías para la extracción de información de los datos serán un proceso recursivo y supervisado el cual con la ayuda de datos limpios brindará resultados más cercanos a la realidad.

Capítulo 8. Trabajos a futuro

Considero que este proyecto es el inicio de una serie de mejoras que podremos entregar a nuestra comunidad, así como al país, existen muchos estudios alrededor del mundo sobre la captación del agua para suministrar a servicios de no consumo humano. Nuestra realidad hace posible que mediante el uso de *hardware* miniaturizado y *software* podamos contar con dispositivos que generan información de altamente precisa.

Por ende, es necesario implementar el uso del internet de las cosas para recolectar y automatizar las lecturas en tiempo real de los niveles de abastecimiento y caudales. Con esto se pretende aumentar el grado de eficiencia y tener un panorama real y con un alto grado de precisión de lo que está ocurriendo en cada fuente de agua.

Las nuevas fuentes de información, así como los datos obtenidos, permitirán controlar y generar alertas automáticas, con el fin de actuar de forma expedita ante cualquier comportamiento inusual en el acueducto y sus consumidores.

Adicionalmente, es necesario generar un perfil de consumo para cada abonado. Con esto se pretende disminuir las pérdidas económicas relacionadas con la fugas no detectadas. La implementación de *triggers* que evalúen el consumo promedio con las nuevas lecturas del abonado, permitirán detectar y notificar a los encargados del acueducto cuando exista un consumo irregular. Con esto se pretende informar a los inspectores, con el fin de normalizar la situación que se esté presentando con el abonado afectado.

Finalmente, incentivar la captación de aguas llovidas en los nuevos desarrollos urbanísticos permitirá disminuir el impacto ocasionado por el cambio climático, por lo que

estimar el beneficio que la comunidad de Belén puede obtener por la implementación de esta buena práctica en grandes consumidores , podrá generar un ahorro considerable del agua para todos.

Glosario

Algoritmo: conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución de un tipo de problemas.

Minería de Datos: es un conjunto de técnicas y tecnologías que permiten explorar grandes conjuntos de datos, de manera semiautomatizada con el objetivo de encontrar patrones que muestran el comportamiento de los datos para un tema determinado.

Almacenamiento del agua: capacidad de guardar agua por un periodo de tiempo en un tanque o depósito destinado a esta función específica.

Alto Consumo: derecho que tiene el cliente para que realicen un estudio sobre el alto consumo marcado en su hidrómetro.

Pozo: perforación vertical en la tierra, profunda y de boca estrecha construida para sacar agua.

Rebombeo: agua que fue bombeada, pero necesita volver a ser bombeada para ser trasladada a otro sitio. Sobrepresión Presiones mayores las presiones de trabajo en las tuberías o accesorios de las redes.

Tratamiento: proceso que se le da al agua para hacerla potable.

CRISP-DM : se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos.

Series de tiempo: es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente.

Análisis de las series de tiempo: se usan métodos que ayudan a interpretarlas y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series.

Kmeans: es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Lenguaje R: es un entorno y lenguaje de programación con un enfoque al análisis estadístico.

Referencias

- Arce, R. d. (n.d.). *Técnicas de Previsión de variables financieras*. Retrieved from Universidad Autónoma de Madrid:
https://www.uam.es/personal_pdi/economicas/anadelsur/pdf/Box-Jenkins.PDF
- Autoridad Reguladora de los servicios Públicos. (2017). *Aresep*. Retrieved from
<https://aresep.go.cr/usuarios/noticias/85-normativa/805-ley-no-1634>
- Belen, M. d. (n.d.). *Acerca de la Municipalidad*. Retrieved from
<https://www.belen.go.cr/web/guest/acerca-de-la-municipalidad?inheritRedirect=true>
- Belen, M. d. (s.f.). *Filosofía Municipal*. Obtenido de Municipalidad de Belen:
<https://www.belen.go.cr/web/guest/inicio?inheritRedirect=true>
- Belen, M. (n.d.). *Historia Canton Belen*. Retrieved from Sitio Web Municipalidad de Belen:
<https://www.belen.go.cr/web/guest/historia>
- Bhattacharyya, S. (2017). *Intelligent Multidimensional Data Clustering and Analysis*. IGI Global.
- Caricari, A. (1 de Febrero de 2014). Obtenido de SlideShare:
<https://www.slideshare.net/caricarisoft/busqueda-avanzada-informacion-30704878>
- Gobierno de la República de Costa Rica. (20 de Febrero de 2017). *AyA confirma merma en producción de agua potable*. Obtenido de
<http://presidencia.go.cr/comunicados/2017/02/aya-confirma-merma-en-produccion-de-agua-potable/>
- González, M. P. (2010, Noviembre). *Instituto Nacional de Estadística Español*. Retrieved from Error cuadrático medio de predicción para modelos estructurales de series temporales.:

<http://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3B+filename%3D1259924936446&blobkey=id&blobtable=MungoBlobs&blobwhere=1259936727951&ssbinary=true>

Kweku-Muata. (2015). *Knowledge Discovery Process and Methods to Enhance Organizational Performance*. Auerbach Publications.

Laboratorio de Análisis Ambiental. (2017). *Calidad de las aguas superficiales*. Belen: Universidad Nacional de Costa Rica.

NCR, P. C. (2000). *Step-by-step data mining guide*. SPSS Inc.

Óscar Marbán, G. M. (2009, January). *A Data Mining & Knowledge*. Retrieved from World's largest Science, Technology & Medicine Open Access book publisher.: http://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf

Prodromou, T. (2017). *Data Visualization and Statistical Literacy for Open and Big Data*. IGI Global.

Shmueli, G. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner, Third Edition*. Massachusetts: John Wiley & Sons .

Trubetskoy, G. (2016, January 29). <https://grisha.org>. Retrieved from Holt-Winters Forecasting for or Developers: <https://grisha.org/blog/2016/01/29/triple-exponential-smoothing-forecasting/>

Wise Owl Business Solutions Ltd. (2016). *Downloading our SQL Server movies or movie database*. Retrieved from <http://www.wiseowl.co.uk/sundry/movies-database/>
<http://www.wiseowl.co.uk/sundry/movies-database/>

Wu, J. (2012). *Advances in K-means Clustering a Data Mining Thinking*. New York: Springer
Verlag.

Anexo A

Se incluye el documento generado y su código fuente generado en el lenguaje R para generar la segmentación en el cantón de Belén.

El archivo KMeansBelen incluye los pasos necesarios de forma detallada para poder guiar la carga y exploración de los datos, adicionalmente este documento contiene los códigos necesarios para realizar las segmentaciones y sus respectivos gráficos para este y futuros proyectos relacionados.



KMeansBelen.Rmd



KMeansBelen.html

Anexo B

Se incluye el documento y código fuente del lenguaje R para generar las series de tiempo en el cantón de Belén.

El documento SeriesTiempo funciona como un manual detallado donde destaca los puntos más relevantes para la confección de las series de tiempo, con los cuales podrán tomar como base para futuros estudios.



SeriesTiempo.Rmd



SeriesTiempo.html

Anexo C

Se incluyen los dashboard confeccionados con el software TABLEAU utilizados para explorar los años 2014, 2015 y 2016.

Estos documentos permiten observar el comportamiento de los consumidores en cada periodo de forma gráfica.

Para facilitar la exploración de los datos en por cada periodo evaluado se crearon dos historias las cuales amplían la comprensión de patrones de consumo.



ConsultasDB2015.twbx



ConsultasDB2014.twbx



ConsultasDB2016.twbx

Anexo D

Se incluyen el reporte creado con el software TABLEAU para analizar el comportamiento de los grandes consumidores en Belén para los años 2014,2015 y 2016.

En el documento se muestra de forma ágil la tendencia de cada consumidor , además de aplicar un parámetro de comparación como la cantidad de piscinas olímpicas de agua potable que consume cada uno de estos grandes abonados , una piscina olímpica representa 2.500.000 de litros de agua potable.



TopConsumidores.t
wb

