FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

APLICACIÓN WEB PARA VISUALIZACIÓN DE INFORMACIÓN METEOROLÓGICA DE QUITO

AUTOR

Rodrigo Andrés Naranjo Ayala

AÑO

2019

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

APLICACIÓN WEB PARA VISUALIZACIÓN DE INFORMACIÓN METEOROLÓGICA DE QUITO

Trabajo de Titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en Sistemas de Computación de Informática

Profesor Guía

PhD. Mario Salvador González Rodríguez

Autor

Rodrigo Andrés Naranjo Ayala

Año

2019

DECLARACIÓN DEL PROFESOR GUÍA

Declaro haber dirigido el trabajo, Aplicación web para la visualización de visualización información meteorológica de Quito, a través de reuniones periódicas con el estudiante Rodrigo Andrés Naranjo Ayala, en el semestre 201910, orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación.

_____

Mario Salvador González Rodríguez

Doctor en Ingeniería Informática y de Telecomunicación

CC:0958376345

DECLARACIÓN DEL PROFESOR CORRECTOR

Declaro haber revisado este trabajo, Aplicación web para la visualización de visualización información meteorológica de Quito, del estudiante Rodrigo Andrés Naranjo Ayala, en el semestre 201910, dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación.

_____

Adonis Ricardo Rosales García

Máster en Informática Aplicada

CC:1756883144

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE

Declaro que este trabajo es original, de mi autoría, que se han citado las fuentes correspondientes y que en su ejecución se respetaron las disposiciones legales que protegen los derechos de autor vigentes.

_____

Rodrigo Andrés Naranjo Ayala

CC:1003976410

RESUMEN

En el presente documento se expone el proceso de diseño, desarrollo e implementación de un sistema web basado en la librería R Shiny para la visualización de información meteorológica de la ciudad de Quito. Se realizan procesos de limpieza de datos sobre los datos históricos de las diferentes estaciones de monitoreo que forman parte de la Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) en un periodo de 2004 a 2018. Se establecen y generan gráficos automatizados basados en las necesidades de investigadores cuyos campos se asocian o están influenciados por los diferentes datos meteorológicos de la ciudad de Quito.

ABSTRACT

The purpose of this project is to design, develop and implement a data visualization system for the meteorological data recovered in the city of Quito based on the package R shiny. The project will also entail the realization of data cleaning processes in the data recovered by the stations that form part of the Quito Metropolitan Network of Atmospheric Monitoring (REMMAQ) in the period of 2004 to 2018. The graphs that the system will be able to generate will be based on the necessity of researchers whose fields of work are related or influenced by the meteorological data of the city of Quito.

ÍNDICE

1. INTRODUCCIÓN

La calidad del aire es una creciente preocupación en todo el mundo debido a los efectos resultante de ser expuestos a contaminantes tanto a corto como largo plazo. Se ha identificado que la exposición a contaminantes puede llevar al desarrollo de enfermedades cardio pulmonares e incluso a muerte prematura. Los efectos de la calidad del aire son sentidos principalmente por individuos de tercera edad, niños y personas con enfermedades crónicas. De acuerdo con Sinharay et al.

> "Exposición a largo plazo a la contaminación puede llevar a un aumento en la tasa de disminución de la función pulmonar, especialmente en personas mayores y en personas con enfermedad pulmonar obstructiva crónica (EPOC), mientras que a corto plazo la exposición a niveles más altos de contaminación se ha implicado en causar un exceso de muertes por cardiopatía isquémica y exacerbaciones de la EPOC" (2018, pp 339-349).

Mientras que la calidad de aire en países del primer mundo ha mejorado en los últimos años principalmente por esfuerzos en ámbitos académicos y legislativos, Latinoamérica aún presenta un alto riego relacionado a problemas causados por baja calidad de aire.

Se han realizado estudios para establecer el nivel de concientización de la gente de países latinoamericanos como Colombia en relación con la calidad del aire. Según Ramirez, Mura & Franco sobre los resultados del estudio "Existe una preocupación generalizada por la calidad del aire dentro de los participantes." (2017, pp 1).

Otros estudios han analizado la información de contaminantes en la ciudad de Quito durante la última década. En relación con legislación implementada para reducir la contaminación del aire por la Ciudad como 'pico y placa', estos encontraron que "los efectos de cada regulación 'se desgastan' después de 2 – 3 años debido a un constante incremento en el número de vehículos." (Zalakeviciute, Rybarczyk, López-Villada & Suares. 2017, pp. 66-75).

Finalmente existen factores ambientales que podrían contribuir al problema de contaminación en ciudades localizadas en altas alturas relativas al nivel del mar

como Quito. De acuerdo con Zalakeviciute, López-Villada & Rybarczyk refiriéndose a la concentración de contaminantes en áreas urbanas de gran altura “Hay una correlación positiva entre la concentración urbana promedio diario de PM2.5 y la humedad relativa en las áreas centrales ocupadas por el tráfico, y una correlación negativa en las afueras de la ciudad en zonas más industriales.” (2018, pp 1-21).

Tomando en consideración toda la información antes presentada sobre los efectos de la calidad del aire sobre la población, la creciente preocupación de individuos sobre los efectos de contaminantes sobre su salud y los diferentes factores que impactan la calidad de aire en ciudad de Quito. Se propone la implementación de una herramienta para el monitoreo y visualización de niveles de contaminación en la ciudad de Quito.

El sistema denominado como AirQ2 (*Quito Air Quality*), es un aplicativo web mediante el cual los usuarios son capaces de generar gráficos de contaminantes y datos atmosféricos, haciendo uso de los datos recopilados por la Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) tras la realización de procesos de limpieza de datos para el aseguramiento de la calidad de los gráficos resultantes. El sistema permite el análisis de los datos desde múltiples perspectivas mediante la generación de: mapas de calor, diagramas de tendencia, diagramas de dispersión, perfiles de tiempo, rosas de viento, diagramas polares, diagramas de densidad, entre otros.

El sistema permite que el público general se informe sobre: el comportamiento de los contaminantes en las diferentes áreas del distrito metropolitano de Quito, las tendencias de los datos de contaminantes y la correlación de contaminantes y variables atmosféricas. A su vez los diferentes grupos de investigadores de campos con relación a la calidad del área y variables atmosféricas son capaces de usar los gráficos generados y los datos limpios disponibles como ayuda en sus procesos de investigación. Finalmente, legisladores pueden utilizar el sistema como base para la toma de decisiones y proposición de leyes, de forma informada y con coherencia al contexto presente en la ciudad de Quito.

1.1. Problema Por Resolver

En la actualidad los procesos de analítica de datos sobre información meteorológica de las diferentes estaciones de Quito requieren de un proceso de limpieza de datos individual por parte de los diferentes grupos de investigadores para la aplicación de métodos matemáticos, estadísticos, aplicación de técnicas descriptivas y modelos predictivos. Estos procesos de limpieza, así como la aplicación de los diferentes modelos buscan el descubrimiento, interpretación y la comunicación de patrones significativos en la información para su uso en la toma de decisiones. Sin embargo, estos procesos de análisis de datos son tareas complejas, exigentes y repetitivas, generando desperdicio en los esfuerzos de los multiples grupos de investigadores en los diferentes campos de interés.

Debido a que no existe una herramienta de visualización de datos enfocada a datos atmosféricos, tanto los procesos de limpieza como la generación de gráficos para el análisis y documentación de hallazgos deben ser realizados en su totalidad desde cero. Otros aspectos por considerar es el conocimiento técnico de los diferentes investigadores en herramientas estadísticas y de visualización ya que, debido a la inexistencia de una herramienta enfocada al análisis de datos meteorológicos, cualquier investigación requiere de un miembro enfocada a la generación de gráficos por medios tradicionales lo que aumenta el esfuerzo del equipo de desarrollo o la inclusión de miembros con conocimientos de sistemas informáticos ampliando el alcance del proyecto de investigación sustancialmente para poder generar gráficos de calidad.

1.2. Objetivo General

Implementar un sistema web para la visualización de datos meteorológicos de la ciudad de quito para facilitar los procesos de análisis multidisciplinario de la información procesada.

1.3. Objetivos Específicos

- Desarrollar un grupo de controles visuales en base a librerías para visualización y modelado de datos incluidas en R, capaces de generar

gráficos de forma dinámica en base a selecciones de datos de meteorológicos de la ciudad de Quito.

- Desarrollar un sistema web intuitivo en base a R-*Shiny* que provea a los usuarios con un grupo de controles para la selección de datos de forma dinámica en base al tipo de grafico que se busque generar.
- Realizar un proceso de limpieza de datos sobre los archivos de variables meteorológicas de las diferentes estaciones de la ciudad de Quito provistos por la secretaria del ambiente, como base para la generación de gráficos.

2. MARCO TEÓRICO

2.1. Sitio Web

Una página web es documento electrónico accesible por medio de internet usado para la compartición de contenido multimedia. De acuerdo con Annis “Las páginas *web* son organizadas en colecciones llamadas sitios web” (Annis, 2014, p 5). Los sitios web tienen la característica de tener un solo autor común para las distintas páginas que lo componen, el cual puede ser un individuo, una organización, etc. Un sitio *web* puede contener una aplicación web que es un sistema el cual esta implementado en varias páginas web.

2.2. Lenguaje R

El lenguaje R es “un lenguaje y ambiente para la computación estadística y la generación de gráficos” (R Project, 2018). Basado en lenguaje S pero con un enfoque de código abierto, R provee a sus usuarios de una gran cantidad de herramientas y técnicas para el análisis estadístico y generación de gráficos. De acuerdo con el proyecto R encargado de la gestión y publicación del lenguaje, la ventaja de R es “la facilidad con la que se pueden producir gráficos con calidad de publicación” (R Project, 2018). R ha ganado un gran número de usuarios en los últimos años en gran número de áreas por su diversidad de paquetes, potencial para la implementación de formulaciones complejas en pocos caracteres y flexibilidad en su uso

2.3. Shiny

Shiny según Resnizky es “un paquete que contiene un grupo de funciones para construir una aplicación web dentro de R.” (Resnizky,2015, p 58). Desde una visión más amplia se puede decir que Shiny es un paquete enfocado a la resolución de peticiones en código R a través de un canal web. Esto lo logra mediante un grupo de funciones que se encargan de reemplazar código R por HTML/JavaScript que permite visualizar el sitio web dentro de un navegador. Shiny es una de las soluciones más populares para la implementación de sistemas web basados en R principalmente por su flexibilidad, capacidad de reflejar cambios dinámicamente, compatibilidad con diferentes paquetes de uso común y facilidad de uso.

2.4. Plotly

Plotly es una librería para la implementación de gráficos con controles interactivos para su uso en ambiente web. *Plotly* implementa sus gráficos en archivos JSON, esto permite su uso en diferentes lenguajes de programación e incluso proveer funcionalidades adicionales a gráficos de paquetes externos como es el caso de ggplot2. Según Baumer, Kaplan & Horton “Funciones como *brushing* (donde los puntos seleccionados son marcados) y anotaciones *mouse-over* (donde los puntos muestran información adicional cuando el ratón se coloca sobre ellos) son automáticas” (Baumer, Kaplan & Horton, 2017, p 244). Al implementar este paquete también se gana la capacidad de exportar los gráficos en archivos vectoriales cuyo uso es común en la generación de artículos y publicaciones.

2.5. Lenguaje Python

Python es un lenguaje de programación para uso general. En la actualidad *Python* es usado por millones de desarrolladores alrededor del mundo por sus capacidades de realizar operación con máximo rendimiento en procesamiento de datos. Las principales características del lenguaje son la fácil interpretación de su código que de acuerdo con Lutz “está diseñado para ser leíble, y por ende mantenible – mucho más que lenguajes tradicionales para *scripts*” (2013) al igual que su Portabilidad en donde dice que “La mayoría de programa en *Python* corren sin cambios en todas

las principales plataformas de computación" (Lutz, 2013). Estos factores aportan a la acogida del lenguaje en varios campos desde finanzas hasta investigación.

2.6. Verificación de Calidad de los datos

La calidad de los datos es una prioridad en la implementación de técnicas de visualización y análisis de información. El uso de datos de baja calidad resulta en procesos de análisis cuyos resultados son erróneos. Como establecen Jugulum & Gray "Los análisis solo puede ser efectivos cuando los datos que son analizados son de alta calidad. Decisiones basadas en conclusiones sacadas de datos de baja calidad pueden concluir en resultados de calidad igualmente baja" (Jugulum & Gray, 2014, p xiii). Siempre que se realicen procesos de análisis es necesario verificar la calidad de los datos, según IBM "Los datos raramente son perfectos. De hecho, la mayoría de los datos contienen errores de programación, valores perdidos, u otros tipos de inconsistencias que hacen que el análisis en momentos sea complejo" (IBM,2018). Entre los tipos comunes de inconsistencias están: valores faltantes, datos erróneos, medidas erróneas, inconsistencias en el código y malos metadatos.

2.7. Visualización de información

De acuerdo con Torres (2010, p. 5) "El término Visualización de Información, como hoy se aplica, se relaciona [con] una disciplina que estudia el diseño de representaciones visuales interactivas que optimicen la carga cognitiva en la representación la misma.", dicho de otra manera, esta disciplina busca generar representaciones visuales de información de tal manera que el conocimiento obtenido de las mismas sea más profundo y entendido con un menor grado de dificultad.

> "En las visualizaciones de información intervienen distintos procesos. Hay una transformación de datos brutos en abstracciones analíticas que, a continuación, se transforman en un modelo espacial-visual abstracto, para que finalmente, mediante procesos de diseño visual, el modelo visual se presente al usuario de forma gráfica y visible. (Olmeda-Gomez, 2014, p 213)

Tomando la visión de Olmeda-Gomez hay que recalcar que los procesos de visualización de información no están relacionados a una sola área de conocimiento y su uso puede beneficiar a múltiples campos en base un solo conjunto de datos si aplicamos diferentes diseños visuales a las abstracciones analíticas obtenidas durante este proceso.

2.8. Sistemas de análisis de datos

"El análisis de datos es el proceso en el cual los datos puros son organizados y ordenados, para su uso en métodos que ayuden a explica el pasado y predecir el futuro. El análisis de datos no se trata de los números, se trata de hacer/formular preguntas, desarrollar explicaciones y probar hipótesis." (Cuestas, 2013, p 7)

Partiendo de la visión de Cuestas se pueden hacer paralelos con el análisis de datos y los sistemas computacionales. Ambos tienen un crecimiento debido a esta relación, "Las herramientas computacionales crean las herramientas para el análisis de datos. La basta cantidad de datos generados ha hecho critico al análisis computacional (…)". (Cuestas, 2013, p 7). Estos procesos de análisis alimentan métodos para la proyección de datos como lo son Big Data, Smart Data, Inteligencia del negocio y el aprendizaje automático. Múltiples áreas usan este tipo de análisis según Iafrate "La mayoría de los negocios usa la información (muchas veces generada por sus propios sistemas de información, vía sus soluciones transaccionales cuyo objetivo es mejorar la productividad en procesos operacionales) que tiene para monitorear y optimizar sus actividades "(Iafrate, 2015, p xvi).

2.9. Scrum

De acuerdo con Schwaber & Sutherland "*Scrum* no es un proceso, una técnica o método definitivo. En lugar de eso, es un marco de trabajo dentro del cual se pueden emplear varios procesos y técnicas. "(Schwaber & Sutherland, 2017, p 1). *Scrum* se utiliza para la implementación de proyecto de múltiples áreas de forma eficaz y eficiente. *Scrum* es iterativo, en el cual cada ciclo se denomina *Sprint*, cada *Sprint* debe resultar en la implementación de un entregable que agregue valor al proyecto final. Al ser un marco de trabajo *Scrum* provee al usuario de un grupo de actividades

para ser desempeñadas, pero no establece necesariamente que se debe hacer en cada una.

3. ANÁLISIS DEL PROBLEMA

3.1. Problemática

En 2002 se puso en marcha la Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) que de acuerdo a la Secretaria del Ambiente " tiene como finalidad producir datos confiables sobre la concentración de contaminantes atmosféricos en el territorio del Distrito Metropolitano de Quito que sirvan como insumo para la planificación, formulación, ejecución y evaluación de políticas y acciones orientadas al mejoramiento de la calidad del aire y difundir esta información en condiciones comprensibles para el público en general" (2018).

La REMMAQ tiene puesto en funcionamiento nueve estaciones a lo largo del distrito metropolitano de Quito. Cada una de estas estaciones tiene equipado un numero de sensores usados para la recolección de datos meteorológicos y concentración de contaminantes. Los datos recopilados por estas estaciones pueden ser visto en un grupo de gráficos accesibles en el sitio web de la secretaria del ambiente http://www.quitoambiente.gob.ec/ambiente, en donde también es posible descargar datos históricos.

Los datos históricos accesibles mediante la secretaria del ambiente están seccionados en base a los diferentes contaminantes y variables meteorológicas, usando una categoría para identificar de que estación provienen. A su vez estos archivos están separados en base a diferentes periodos de tiempo y no existe control de inconsistencias comunes para estos datos. Debido al volumen y diversidad de tipos de archivos, la agrupación de datos y realización de procesos de limpieza requieren conocimientos sobre el uso de sistemas computaciones, por ejemplo: R, S y *Python*.

Los datos proporcionados por la secretaria del ambiente tienen un impacto en una gran cantidad de áreas desde campos ambientales hasta jurídicos donde no es posible asegurar que posean conocimientos técnicos sobre: el correcto desempeño

de métodos de limpieza de datos, aseguramiento de su calidad y generación de representaciones mediante sistemas computacionales. De igual manera la realización de procesos de análisis de datos requiere conocimientos sobre características individuales de los diferentes contaminantes y variables atmosféricas que está relacionado con el estudio del ambiente.

3.2. Historias de Usuario

La recolección de requerimientos para el sistema fue realizada en base a historias de usuario que fueron recopiladas mediante interacciones con expertos en calidad de aire, visualización y análisis de datos. Estas historias fueron organizadas y dadas un formato según los métodos para limpieza de pendientes. El proceso de limpieza de pendientes de acuerdo con Vanderjack es “el proceso de convertir Historias de Usuario en un formato que el equipo de trabajo esté dispuesto a aceptar en una iteración ágil de *Scrum*.” (2015, p 15).

Las historias recopiladas fueron redactadas en base al formato común propuesto por *Scrum* y se presentan a continuación:

- Como investigador quiero generar mapas de calor en base a contaminantes y su distribución relativa a la posición de las diferentes estaciones, para analizar el movimiento, orígenes y concentración de los contaminantes por la ciudad de Quito.
- Como investigador quiero poder generar gráficos de densidad de datos basados en promedios de diferentes periodos de tiempo, para analizar tendencias de los datos.
- Como investigador necesito poder agregar variables categóricas a gráficos de densidad basados en valores provisto por la organización mundial de la salud, para facilitar procesos de análisis de concentración de contaminantes en niveles peligrosos.
- Como investigador quiero poder generar gráficos de trayectoria de los datos en diferentes escalas de tiempo (horarios, mensuales, anuales, etc), para analizar la variación de valores y como ayuda en proceso de predicción.

- Como investigador quiero generar gráficos de perfiles horarios de variables atmosféricas, para establecer correlaciones entre las variables, comprobar expectativas en diferentes estaciones climáticas y comparar valores durante diferentes eventos climáticos como el niño/niña.
- Como investigar quiero seleccionar las variables disponibles en los perfiles horarios (máximos, mínimos, promedios, acumulados, valores provistos por la organización mundial de la salud), para personalizar los gráficos para mejor representar el comportamiento de las variables.
- Como investigador quiero generar gráficos de rosas de viento y de dirección de viento en base a la librería OpenAir, para validar hipótesis de la dirección del viento por temporada (temporada seca contra temporada lluvia) mediante la comparación de sus valores.
- Como investigar quiero poder personalizar aspectos de: periodos de tiempo, colores para representación, métodos estadísticos, entre otros, para poder mejorar la representación de la información contenida en el gráfico.
- Como investigador quiero generar gráficos de la correlación de contaminantes y variables atmosféricas, en base a periodos de 24 horas, días de la semana y anuales, para analizar su variación en el tiempo y relación entre sus valores.
- Como investigador quiero poder descargar datos del sistema seccionados, para realizar procesos de investigación y comprobar la generación de gráficos del sistema.
- Como investigador quiero poder generar gráficos de la calidad de los datos en diferentes estaciones y periodos anuales, para comprobar cantidades de datos validas e inválidos (negativos y no existentes).

4. IMPLEMENTACIÓN DE LA SOLUCIÓN

4.1. Propuesta de la solución

En base a la problemática y las historias de usuarios recopiladas se propone la implementación de un sitio web en lenguaje R mediante el uso del paquete *Shiny*, en el cual se harán disponibles los conjuntos de datos provistos por la secretaria del

ambiente tras realizar la limpieza y estandarización de estos datos para su acceso por cualquier individuo. Además, se propone la implementación de un grupo de gráficos basados en los diferentes paquetes de R para el análisis de datos meteorológicos.

Para la realización de los procesos de limpieza de datos se hará uso de lenguaje *Python* y de sus diferentes librerías para facilitar la generación de archivos csv estandarizados para su uso en la generación de gráficos en la aplicación web en R. El uso de archivos csv facilita los procesos de mantenimiento de la aplicación, permite la gestión del sistema por personal sin conocimiento técnico de sistemas de información, agiliza la modificación de los datos usados para el desarrollo de una aplicación iterativa y permite la implementación de funcionalidades para la carga de datos personalizados en un futuro desarrollo.

La aplicación web utilizara controles dinámicos comunes para la generación de gráficos personalizables por los usuarios del sistema, esto se implementarán en base a las funciones provistas por *Shiny*. El uso de controles comunes facilitaría el análisis de una estación desde varias perspectivas y disminuiría la carga de procesamiento generada por él envió de múltiples peticiones por un usuario para generar un solo gráfico. El uso de controles dinámicos limita la carga del servidor al solo mostrar los controles necesarios, además de permitir limitar opciones en base a estaciones, variables y tipos gráficos acorde con las especificaciones de investigadores.

Las funciones para la generación de gráficos se implementarán en base a los paquetes ggplot2 y *Plotly*. El paquete *Plotly* se utilizará para agregar funcionalidades adicionales web a los gráficos generados mediante el paquete ggplot2. Ciertos gráficos implementados en el sistema serán directamente generados en base al paquete Open Air, el cual es desarrollado por el consejo de investigación del ambiente natural para el análisis de datos de contaminantes. Entre los gráficos generados en base a este paquete están: flores de viento, flores de contaminantes y diagramas de dispersión.

4.2. Metodología de Desarrollo

El desarrollo del proyecto se realizó en base a *Scrum*, en donde se estableció un ciclo de iteración de una semana. La pila de producto del proyecto se generó en base las historias de usuarios previamente recopiladas. Las historias de usuario fueron priorizadas por el dueño del producto, asignadas un valor de esfuerzo relativo por el equipo de desarrollo y ordenadas para su implementación.

Al ser un proceso iterativo, la pila del producto fue modificada durante el proceso de desarrollo con el objetivo de descomponer funciones muy complejas para ser desarrolladas durante un ciclo de iteración o en base a nuevas prioridades provistas por el dueño del producto. La Pila del producto puede ser vista en Tabla 1, en esta pueden ser vistas las historias de usuario implementadas en el sistema, el orden de ingreso a la pila de producto y el ciclo de iteración donde fueron implementadas.

Tabla 1.

Pila del Producto

ID (Orden ingreso)	Alias	Estado	Esfuerzo (Puntos de Historia)	Iteración
1	Comprensión de fuentes de datos	Hecho	8	1
2	Limpieza Datos	Hecho	24	2
3	Estructura Sitio	Hecho	16	3
4	Implementación Servidor	Hecho	8	4
5	Diagrama Tendencia	Hecho	8	6
6	Diagrama Dispersión	Hecho	8	7
9	Diagramas Open Air	Hecho	12	8
10	Diagrama Calidad Estación	Hecho	8	5
8	Mapas Calor	Hecho	14	9
12	Tabla de datos	Hecho	5	10
13	Descarga de Datos	Hecho	5	10
11	Resumen de datos	Hecho	5	10
14	Diagrama Calidad Barras	Hecho	12	11
15	Diagrama Dispersión - Open Air	Hecho	10	12
16	Diagrama Perfiles Tiempo	Hecho	14	12
7	Diagrama Distribución	Hecho	18	13

4.3. Desarrollo del Sistema

En el sistema se implementaron múltiples funciones para la visualización de datos en un ambiente web en base al paquete R *Shiny*. Estas funciones hacen uso de un grupo de controles de usuario dinámicamente generados para la personalización de los resultados, mediante la selección de parámetros como: estación para el análisis, variables atmosféricas analizadas, colores usados para la representación de valores, etc. Los controles implementados son compartidos en múltiples gráficos considerando aspectos de compatibilidad para facilitar la navegación del sitio web cuando se analiza cualquiera de las diferentes estaciones. El uso de controles comunes también permite el análisis de la misma información desde múltiples perspectivas haciendo uso de diferentes funciones, pero los mismos parámetros.

Las actividades realizadas por el sistema y el usuario, durante el proceso de selección y parametrización de gráficos se pueden ver representadas mediante un árbol de tareas concurrentes en Fig.1, donde se puede identificar el flujo normal en la generación de gráficos por la aplicación. En este el usuario selecciona el tipo de grafico usando el menú de navegación preestablecido, el sistema carga los controles asociados y validaciones específicas de los controles en base al tipo de gráfico. Tras este proceso el sistema realiza la carga de datos basados en el tipo de gráfico. Finalmente, el sistema genera el grafico con el cual el usuario puede interactuar de manera paralela haciendo uso de los controles previamente cargados.

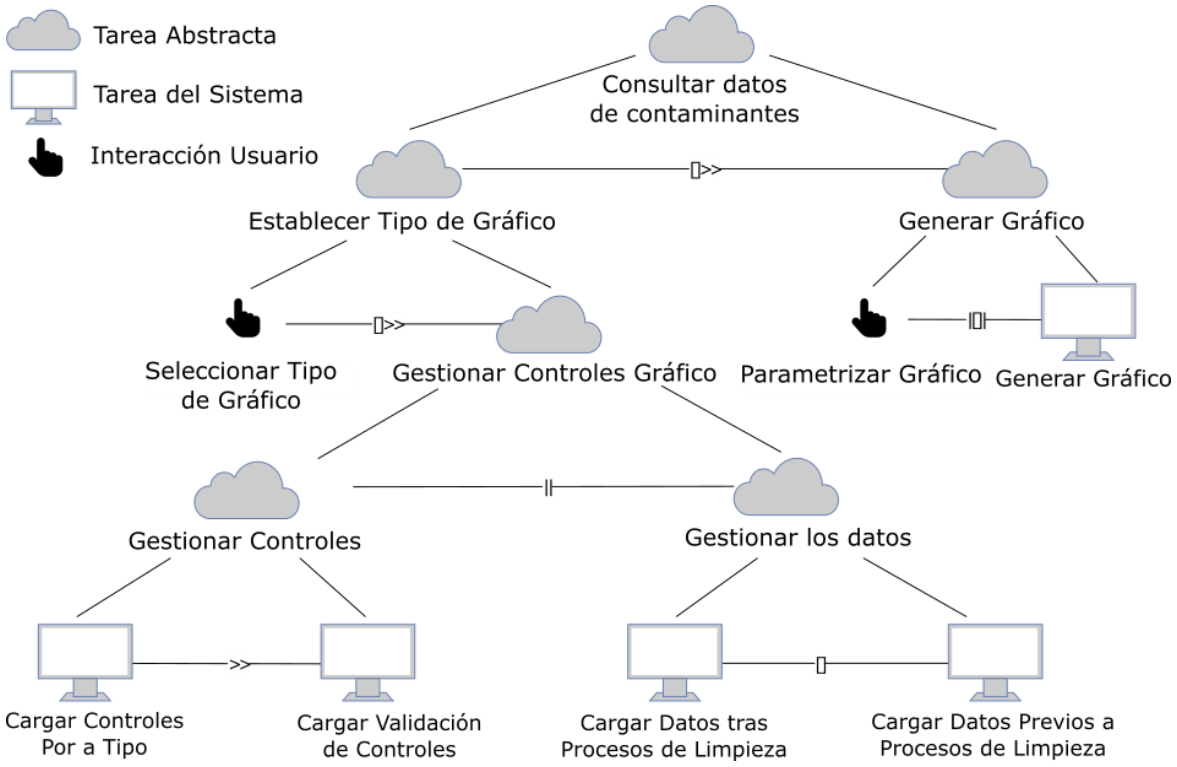

Figura 1. Árbol de Actividades Concurrentes de Aplicación AirQ2.

La aplicación organiza sus gráficos en base a su modo de implementación y el tipo de análisis al que están relacionados. Se presenta las siguientes secciones visibles en Fig.2: Análisis descriptivo, Open Air y Limpieza de datos. La sección Análisis descriptivo presenta gráficos implementados mediante diferentes técnicas usados para el análisis de los datos tras procesos de limpieza, en esta sección se encuentran: diagramas de distribución, diagramas de dispersión, diagramas de tendencia, perfiles de tiempo y mapas de calor. La segunda sección *Open Air*, presenta gráficos implementados en base a funciones de dicho paquete para el análisis de datos tras procesos de limpieza, en esta sección se pueden encontrar: rosas de viento, rosas de contaminantes, gráficos polares y gráficos de dispersión. Finalmente, en la sección Limpieza de datos se presentan gráficos para el análisis de la calidad de las diferentes estaciones, los cuales se generan en base a los datos previos a procesos de limpieza, en esta sección se pueden encontrar: diagrama de calidad de datos (control de nulos) y diagramas de barra.

En Fig.2 se puede visualizar una captura de la aplicación mostrando los controles disponibles para un gráfico (distribución) de la variable (PM2.5) en la estación (Los chillos) para un periodo (diario). En este grafico se pueden visualizar controles (de arriba para abajo) para: seleccionar la estación de estudio, seleccionar la variable meteorológica representada en el gráfico, seleccionar el número de columnas usadas para la representación de los datos, seleccionar periodos para la agrupación de los datos, agregar una variable categórica basada en especificaciones de la organización mundial de la salud y seleccionar rangos de tiempo para el análisis.

Los controles disponibles en la aplicación tienen dos modos de uso. En primera instancia existen controles comunes que pueden ser reutilizados en múltiples gráficos en una misma sección, este es el caso de controles para selección de estaciones, selección de rangos de tiempo y selección de variables meteorológicas. Otro caso son controles que presentan modificaciones basadas en el grafico donde son implementados, estos pueden ser controles específicos para un tipo de gráfico, controles que necesitan validaciones especificas en base a su valor, etc. No existen controles comunes entre múltiples secciones por limitaciones de la tecnología para compartir/actualizar dichos valores de manera dinámica.

En el proceso de implementación de controles dinámicos para el sistema se debió considerar diferentes aspectos relativos a: características de los diferentes gráficos, información disponible en cada estación y expectativas de los usuarios del sistema. En primera instancia, los gráficos individualmente presentan características propias que limitan la cantidad y tipo de controles que es posible generar para la personalización de gráficos. Otro aspecto por considerar es la disponibilidad de información en cada estación, al no poseer un grupo de sensores en común cada estación posee un grupo de variables meteorológicas diferentes lo que establece los valores válidos para la selección en procesos de visualización. Finalmente, con respecto a las expectativas del usuario del sistema. Los usuarios esperan que en casos donde un valor no es posible, este no se muestre como una opción este limitante se puede ver más claramente en casos relativos a periodos de tiempo, no existen valores sugeridos por la organización mundial de la salud para periodos

mensuales. Por ende, no se debería mostrar esta selección al generar una variable categórica en un gráfico de distribución para este periodo.

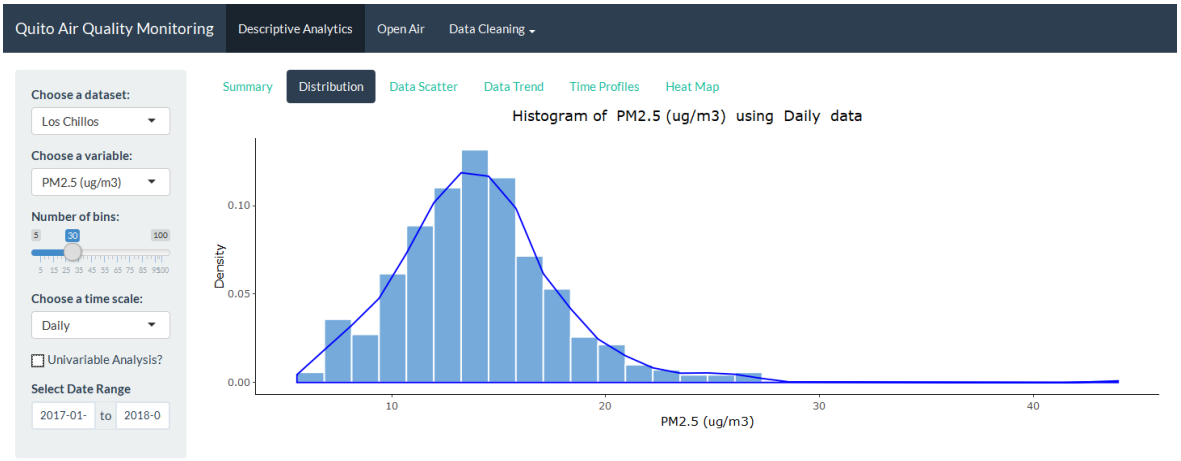

Figura 2. Captura de controles aplicación AirQ2 – Diagrama de Distribución de PM2.5 en un periodo Diario basado en datos de estación Los Chillos.

Otro aspecto que considerar en la implementación del sistema fue el manejo de los datos tras procesos de limpieza y previos a los mismos. En Figura 3 se presentan las relaciones establecidas entre los datos las cuales se aseguraron mediante los procesos de organización previos a la limpieza de datos. Es notable que las variables atmosféricas y contaminantes mantienen una estructura similar que se diferencia por la posible presencia de valores recomendados por la organización mundial de la salud. Finalmente es importante destacar que las funciones para el registro y gestión del tiempo fueron generas usando el paquete lubridate.

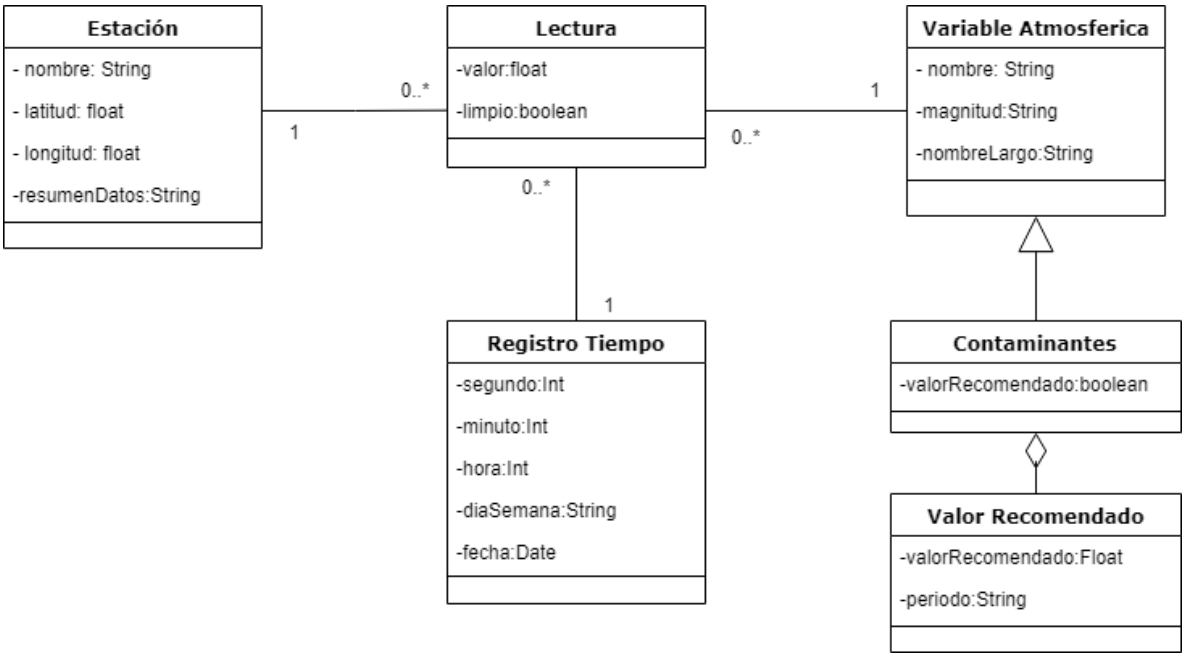

Figura 3. Diagrama UML de Clases

5. RESULTADOS

5.1. Comprensión de datos

“La comprensión de los datos implica acceder a los datos y explorarlos mediante tablas y gráficos” (IBM, 2016, p 11). La compresión de datos es el proceso parte del análisis de información por el cual se recopilan los datos de las diferentes fuentes, se validan sus formatos y se establecen aspectos de calidad para su uso en procesos de análisis. Esto proceso se puede dividir en tres actividades: recolección de datos inicial, descripción de los datos y verificación de la calidad de los datos.

La recolección de datos inicial es el proceso por el cual se recopilan datos de las diferentes fuentes disponibles y que tengan alguna relación válida para uso en el proceso de análisis. En relación con el desarrollo del proyecto, se identificaron dos fuentes potenciales de datos en base a las fuentes hechas disponibles por parte de la secretaria del ambiente, estas son: datos históricos de variables atmosféricas recopiladas por las diferentes estaciones y datos de las diferentes estaciones del distrito metropolitano de quito. Se identifico una tercera fuente potencial en los valores sugeridos por la organización mundial de la salud para la evaluación de la concentración de contaminantes.

En segundo lugar, se llevó a cabo el proceso de descripción de los datos. En este se deben identificar aspectos de volúmenes, formatos y esquemas de representación. En base a la información obtenida de la secretaria del ambiente se identificaron en total siete variables atmosféricas y seis contaminantes, los cuales fueron recopilados desde 2004 hasta la fecha actual en periodos horarios. Los datos se encontraron en archivos csv seccionados por variable climática/contaminante y a su vez dividido en periodos de tiempo. En total se identificaron tres formatos para el almacenamiento de datos: 2004 – 2007 (i), 2008 – 2016 (ii) y 2017 – actual (iii). Cada uno de estos formatos posee variaciones en la representación de estación de origen y formato para almacenamiento de fecha/hora, a continuación, se explican las diferencias encontradas:

- Formato (i): En el formato se identifican 4 columnas: estación, magnitud, fecha y dato. La columna estación identifica el origen de la lectura, magnitud establece la variable atmosférica/contaminantes medido, fecha establece la fecha/hora de la lectura del dato y dato establece el valor de la lectura. Se identifica una fila denominada unidad medida que establece la escala del valor leído. Se puede el formato representado en Tabla 2.
- Formato (ii): Este formato mantiene la estructura de estación y magnitud del formato (i). La principal diferencia viene de la separación de fecha en columnas individuales año, mes y dia. Con el almacenamiento de las lecturas en un grupo de columnas denominadas HORA, existe una columna hora por cada lectura de 24 horas. Una sección del formato puede ser visto en Tabla 3.
- Formato (iii): este formato es el usado actualmente para el registro de datos históricos, en este se presentan columnas individuales para el almacenamiento de lectura de las diferentes estaciones y una columna “date” para el registro de la fecha y hora de la lectura del dato. Tanto el contaminante como la medida usada pueden ser visto debajo del nombre de la columna de cada estación. Este formato puede ser visto en Tabla 4.

Tabla 2.

Formato 2004 - 2007

UNIDAD MEDIDA: ug/m3			
ESTACION	MAGNITUD	FECHA	DATO
COTOCOLLAO	PM2.5	2005-03-17 11:00:00.0	57.63
COTOCOLLAO	PM2.5	2005-03-17 12:00:00.0	27.74
COTOCOLLAO	PM2.5	2005-03-17 13:00:00.0	14.61
COTOCOLLAO	PM2.5	2005-03-17 14:00:00.0	9.6
COTOCOLLAO	PM2.5	2005-03-17 15:00:00.0	7.73
COTOCOLLAO	PM2.5	2005-03-17 16:00:00.0	10.1
COTOCOLLAO	PM2.5	2005-03-17 17:00:00.0	6.96
COTOCOLLAO	PM2.5	2005-03-17 19:00:00.0	9.35
COTOCOLLAO	PM2.5	2005-03-17 20:00:00.0	15.86
COTOCOLLAO	PM2.5	2005-03-17 21:00:00.0	10.4

Tabla 3.

Formato 2008 – 2016

UNIDAD MEDIDA: ug/m3							
ESTACION	MAGNITUD	AÑO	MES	DIA	HORA 1	HORA 2	HORA 3
COTOCOLLAO	PM2.5	2008	1	1	82.94	165.31	59.23
COTOCOLLAO	PM2.5	2008	1	2	1.52	2.76	2.7
COTOCOLLAO	PM2.5	2008	1	3	0.41	8.54	3.57
COTOCOLLAO	PM2.5	2008	1	4	13.31	15.26	19.53
COTOCOLLAO	PM2.5	2008	1	5	17.77	20.65	13.3
COTOCOLLAO	PM2.5	2008	1	6	1.42	5.61	5.57
COTOCOLLAO	PM2.5	2008	1	7	4.01	2.28	1.88
COTOCOLLAO	PM2.5	2008	1	8	13.44	8.92	7.25
COTOCOLLAO	PM2.5	2008	1	9	3.68	3.49	0.56

Tabla 4.

Formato 2017 – Actual

	Belisario	**Carapungo**	**Centro**	**Cotocollao**
	PM2.5_ug	PM2.5_ug	PM2.5_ug	PM2.5_ug
Date				
01-Jan-2018 00:00	29.97	162.58	51.68	64.96
01-Jan-2018 01:00	155.09	178	276.67	257.26
01-Jan-2018 02:00	184	38.07	215.73	104.97
01-Jan-2018 03:00	96.74	19.08	115.53	46.85
01-Jan-2018 04:00	30.67	21.13	23.97	6.38

01-Jan-2018 05:00	51.28	10.63	0	5.96
01-Jan-2018 06:00	38.18	8.29	0	3.62
01-Jan-2018 07:00	1.87	2.1	0.12	5.97
01-Jan-2018 08:00	7.88	6.76	0	0.05

Una vez identificadas las fuentes, se realizaron proceso para la verificación de la calidad de los datos obtenidos. Esto se realizó en base al paquete *summarytools*, que fue usado para generar un reporte de la calidad de los datos en las diferentes estaciones en base a las diferentes variables y contaminantes, un ejemplo de los reportes generados puede ser visto en Fig.4 en relación con la estación Los chillos. En el reporte generado se pueden identificar aspectos relacionados a las diferentes variables como: promedios, valores mínimos, valores máximos, valores medios, frecuencia de repetición de valores válidos, cantidad de valores válidos y cantidad de valores no existentes.

Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
CO [numeric]	mean (sd) : 10631.52 (1479724.67) min < med < max : -0.64 < 0.48 < 238377734.19 IQR (CV) : 0.3 (139.18)	266 distinct values	37158 (29.36%)	89386 (70.64%)
DIR [numeric]	mean (sd) : 189.87 (100.74) min < med < max : 0 < 176.2 < 360 IQR (CV) : 171.66 (0.53)	32790 distinct values	123773 (97.81%)	2771 (2.19%)
HUM [numeric]	mean (sd) : 72.59 (21.47) min < med < max : 7.67 < 78.33 < 100 IQR (CV) : 37.5 (0.3)	7797 distinct values	123333 (97.46%)	3211 (2.54%)
LLU [numeric]	mean (sd) : 0.14 (1.09) min < med < max : 0 < 0 < 45.5 IQR (CV) : 0 (7.79)	246 distinct values	119642 (94.55%)	6902 (5.45%)
NO2 [numeric]	mean (sd) : 21.81 (11.35) min < med < max : 1.14 < 20.26 < 106.42 IQR (CV) : 14.87 (0.52)	5364 distinct values	46763 (36.95%)	79781 (63.05%)
O3 [numeric]	mean (sd) : 24.04 (21.68) min < med < max : -12.52 < 18.63 < 152.23 IQR (CV) : 30.86 (0.9)	9424 distinct values	122827 (97.06%)	3717 (2.94%)
PM2.5 [numeric]	mean (sd) : 15.56 (10.27) min < med < max : -21.03 < 14.3 < 404.56 IQR (CV) : 12.61 (0.66)	4494 distinct values	37585 (29.7%)	88959 (70.3%)
PM10 [logical]		All NA's	0 (0%)	126544 (100%)
PRE [numeric]	mean (sd) : 759.24 (1.84) min < med < max : 749.46 < 759.48 < 785.14 IQR (CV) : 2.39 (0)	1341 distinct values	125552 (99.22%)	992 (0.78%)
RS [numeric]	mean (sd) : 200.28 (304.38) min < med < max : 0 < 0.4 < 1279.65 IQR (CV) : 351.71 (1.52)	43913 distinct values	125480 (99.16%)	1064 (0.84%)
SO2 [numeric]	mean (sd) : 8.09 (12.03) min < med < max : -1.3 < 4.5 < 245.15 IQR (CV) : 6.85 (1.49)	4463 distinct values	46903 (37.06%)	79641 (62.94%)
TMP [numeric]	mean (sd) : 15.11 (4.48) min < med < max : -0.45 < 14.08 < 27.87 IQR (CV) : 6.75 (0.3)	2403 distinct values	125866 (99.46%)	678 (0.54%)
VEL [numeric]	mean (sd) : 1.54 (1.08) min < med < max : 0 < 1.19 < 9.24 IQR (CV) : 1.18 (0.7)	704 distinct values	125651 (99.29%)	893 (0.71%)

Figura 4. Reporte de Datos – Estación los chillos.

5.2. Limpieza de datos

El proceso de limpieza de datos consistió en la recopilación de las diferentes fuentes provistas por la secretaria del ambiente de Quito. Estas fuentes fueron obtenidas en base a archivos Excel descargados del sitio web de la secretaria del ambiente. Cada archivo recibió un proceso de estandarización basados en librerías de lenguaje *Python* para obtener un formato común entre las diferentes variables y periodos. El flujo de este proceso se puede ver documentado en Fig.5.

Una vez se consiguió estandarizar el formato de los datos se procedió a realizar un proceso de consolidación en un solo archivo común. En este proceso se juntaron los datos en base a los identificadores de estación, fecha y hora. Los procesos de consolidación de datos fueron realizados mediante el uso de librerías *Numpy* y

Pandas. El resultado final fue un archivo organizado con todos los datos obtenidos de las fuentes, el cual fue seccionado en base a las diferentes estaciones para su uso en el sistema. El proceso de generación de archivos puede ser visto en Fig.5.

Los archivos resultantes mantenían las discrepancias encontradas en los datos iniciales, principalmente la existencia de valores fuera de rangos y valores no existentes. Para la resolución de este problema se hizo uso de paquetes en lenguaje R para el remplazo de los valores antes mencionados con un promedio de los datos validos encontrados en la aplicación, este proceso se realizó en base a las especificaciones de expertos en análisis de datos atmosféricos para mejorar la calidad de los resultados obtenidos en el sistema y evitar conflictos con el uso de variables no existentes en proceso de análisis y visualización. El resultado del proceso fueron los mismos archivos sin presencia de datos no existentes.

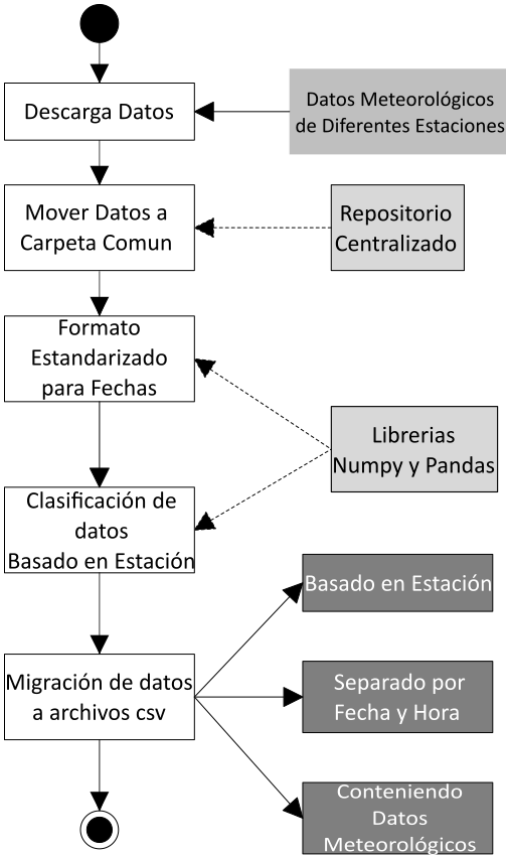

Figura 5. Proceso de limpieza de datos.

5.3. Visualización de datos

El sistema funciona como una herramienta para el análisis de información histórica de contaminantes en el distrito metropolitano de Quito en base a los datos de las estaciones de la REMMAQ recopilados durante el periodo de 2004 hasta junio de 2018.El sistema implementa múltiples funciones para la generación de gráficos en base a diferentes métodos para la visualización de datos usados en el análisis de información. En total en se implementaron 6 métodos de análisis con controles dinámicos usados para la personalización de los resultados. Los métodos implementados son: exploración de la calidad de los datos, perfiles de tiempo (anuales, mensuales, diarios y horarios), rosas de contaminantes, mapas de calor, análisis multivariables y de tendencia.

5.3.1. **Exploración de la calidad de los datos**

El análisis de exploración de la calidad de los datos se centra en identificar y entender el material fuente con el cual se va a realizar el proceso de análisis y visualización de los datos. En estos gráficos se utilizan los datos previos a procesos de limpieza de valores negativos y vacíos. Se pueden entender estos datos como los archivos agrupados y con un formato común generados en los procesos de limpieza de datos. En este tipo de análisis se provee también la capacidad de descargar los datos antes mencionados mediante un botón de descarga presente en todas las interfaces de usuario de gráficos de este tipo, véase Fig.6, Fig.7 y Fig.8.

Existen dos gráficos asociados al análisis de exploración de la calidad de los datos, estos son: diagramas de calidad de datos y diagramas de barra. En primer lugar, los diagramas de calidad de datos realizan un análisis de cada una de las lecturas obtenidas por una estación, considerando si el valor recolectado es vacío o no. Mediante una trasformación matemática, este valor es convertido a una escala de unos y ceros que se representa como un mapa de calor mediante el paquete *plotly*, se puede ver un ejemplo basado en la estación (los chillos) en Fig.6.

El segundo diagrama que conforma este método, diagrama de barras. Se genera en base a la agrupación de los valores recolectados por un periodo anual o en base a las variables contenidas en dicha estación. En el caso de una variable anual, la

agrupación se realiza por funciones integradas en los paquetes base de R con ayuda de la librería *lubridate* la cual contiene “Funciones para trabajar con fecha-horas y periodos de tiempo” (Spinu et al., 2018), Los datos se agrupan en periodos anuales donde cíclicamente se determinan total de datos válidos, datos con valores negativos y datos vacíos. Estos se agregan en un gráfico de barras en posiciones determinadas por el año de origen de los datos y representados mediante el paquete *ggplot2* con funcionalidades web provistas por *plotly*. Se pueden ver ejemplos de gráficos de barra por variable y por año, respectivamente en Fig.7 y Fig.8.

Existe una tercera funcionalidad relacionada al análisis de calidad de datos implementada en esta sección del sitio denominada como *data tables* o tablas de datos, véase Fig.9. Esta funcionalidad permite la descarga de datos limpios de las diferentes estaciones seccionados por periodos de tiempo o variables para su uso en procesos de análisis y visualización externos a la aplicación.

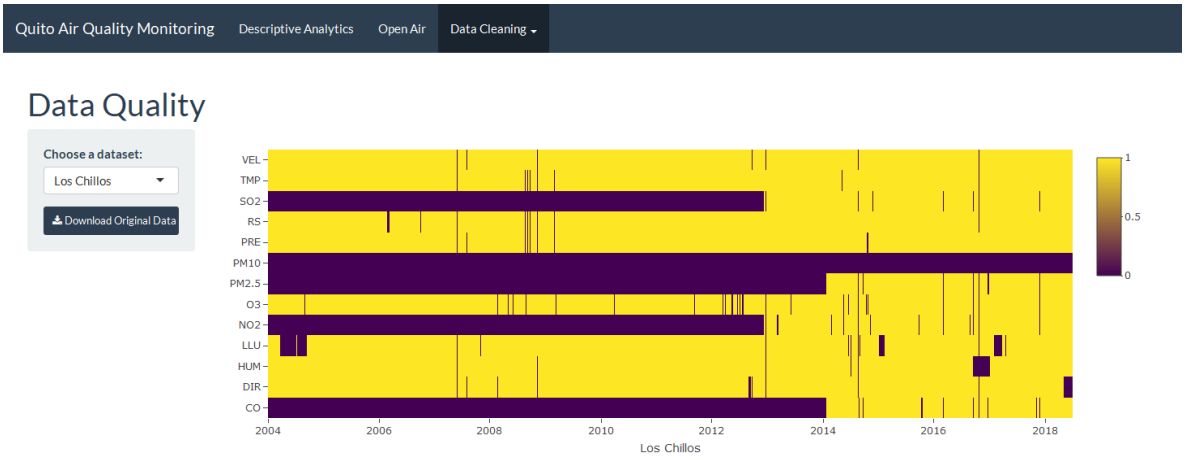

Figura 6. Calidad de datos en estación Los Chillos – Valores Validos

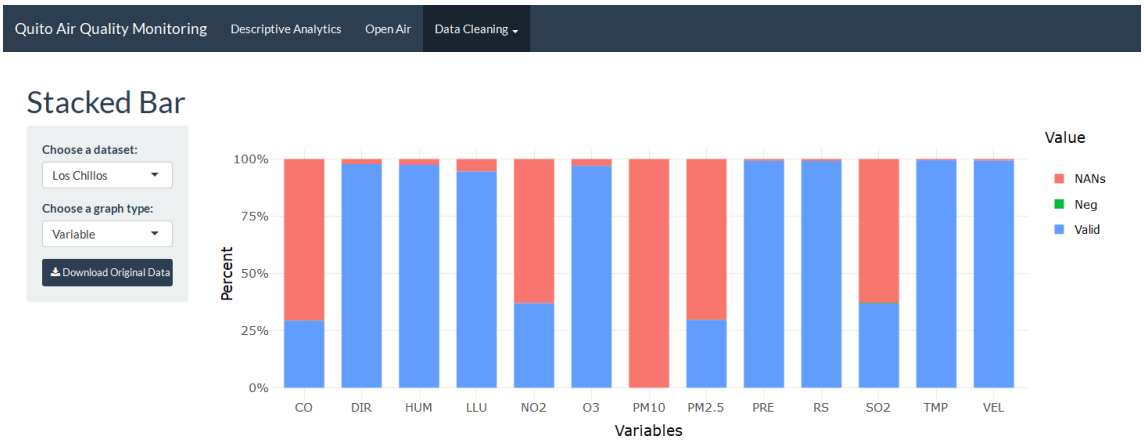

Figura 7. Calidad de datos en estación Los Chillos – Barras por variable

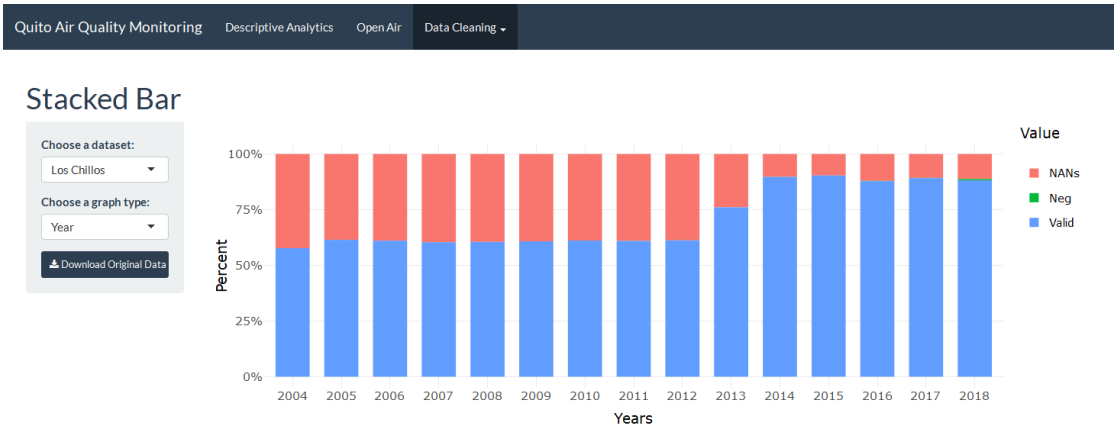

Figura 8. Calidad de datos en estación Los Chillos – Barras por año

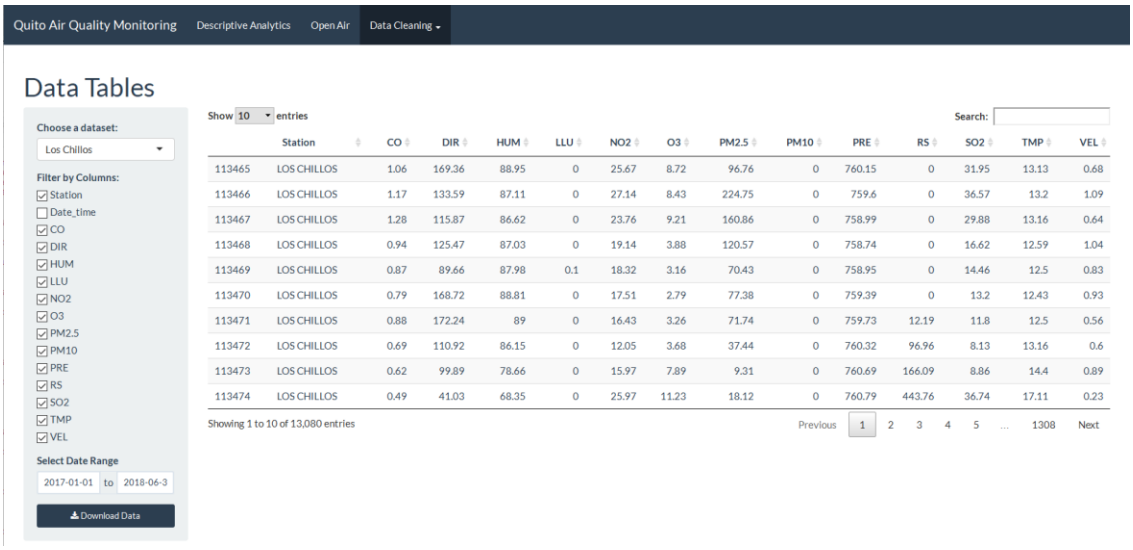

Figura 9. Tablas de datos en estación Los Chillos – Periodo Año, datos enero 2006 – junio 2018.

5.3.2. **Perfiles de tiempo**

Los perfiles de tiempo son un método de análisis de datos relacionados a la agrupación y verificaciones de valores en una escala de tiempo común. Los valores representados pueden ser máximos, mínimos, promedios, errores estándar y acumulaciones. Como se implementaron en la aplicación, los perfiles de tiempo detallan el comportamiento de un contaminante o variable atmosférica a lo largo un periodo de tiempo, representando en esta escala valores resultantes de procesos de agrupación en base a la selección de periodo por parte del usuario.

En Fig.10 se pueden ver los controles comunes implementados en esta función, que incluyen: selección de estación de estudio, selección de variable atmosférica, selección de periodo de tiempo, selección de representación de las diferentes variables (máximo, mínimo, acumulado y promedio) y selector de rango de tiempo. En relación con los controles dinámicos implementados en esta función, las variables atmosféricas disponibles son modificadas según la selección de estación de estudio y el control para agregar los valores propuestos por la organización mundial de la salud solo se visualiza cuando la variable presenta dicho valor para el periodo seleccionado, véase Fig.11.

Los datos representados se obtienen mediante el uso de los paquetes *lubridate* para la gestión de periodos de tiempo, paquetes base de R para la agrupación de valores y cálculo de valores representados (promedios, máximos, mínimos, acumulados, errores estándar) para el periodo. Una vez se tienen los datos, mediante un proceso de comprobación lineal de las variables seleccionadas por el usuario se añaden gráficos representando el comportamiento a un objeto *ggplot2*. El cual finalmente se presenta al usuario mediante la librería *plotly* para proveer funcionalidades adicionales. Algunos ejemplos del uso de esta función se pueden ver en Fig.10 – 14, relacionadas al estudio de la estación (los chillos) en diferentes periodos de tiempo (2017-2018 y 2006-2018) y variables atmosféricas (Ozono y PM 2.5).

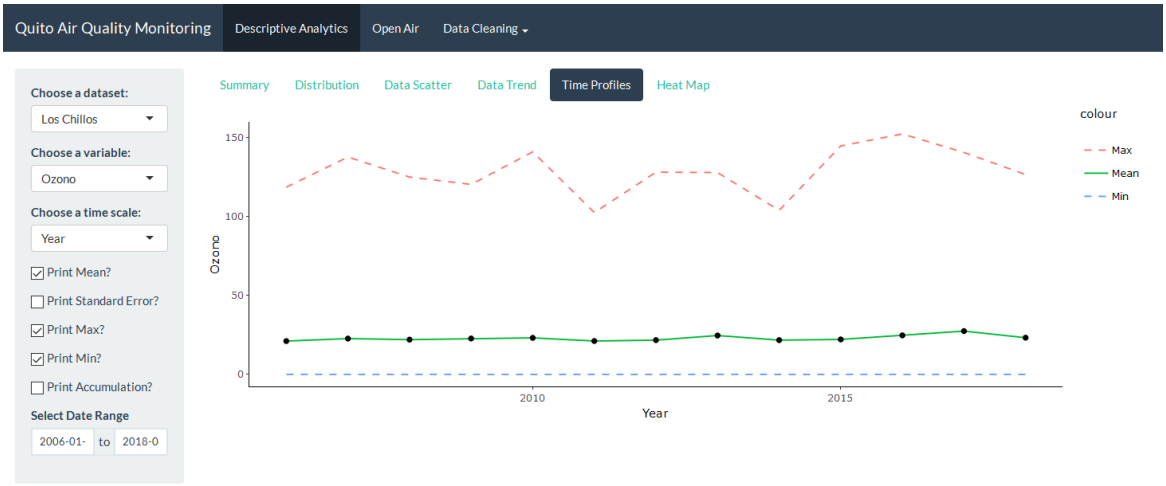

Figura 10. Perfiles de tiempo de variable Ozono en estación Los Chillos – Periodo Año, datos enero 2006 – junio 2018.

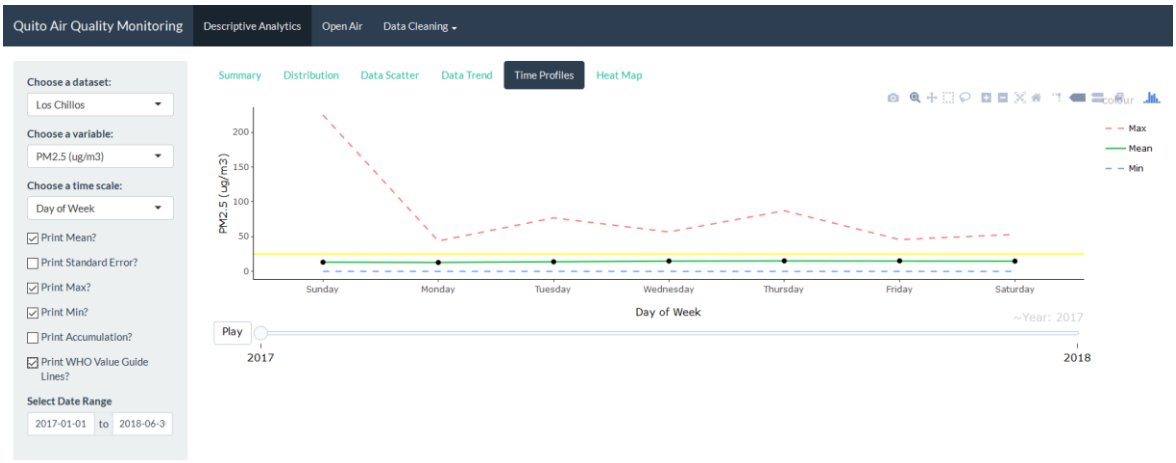

Figura 11. Perfiles de tiempo de variable PM 2.5 en estación Los Chillos – Periodo Año, datos enero 2017 – junio 2018.

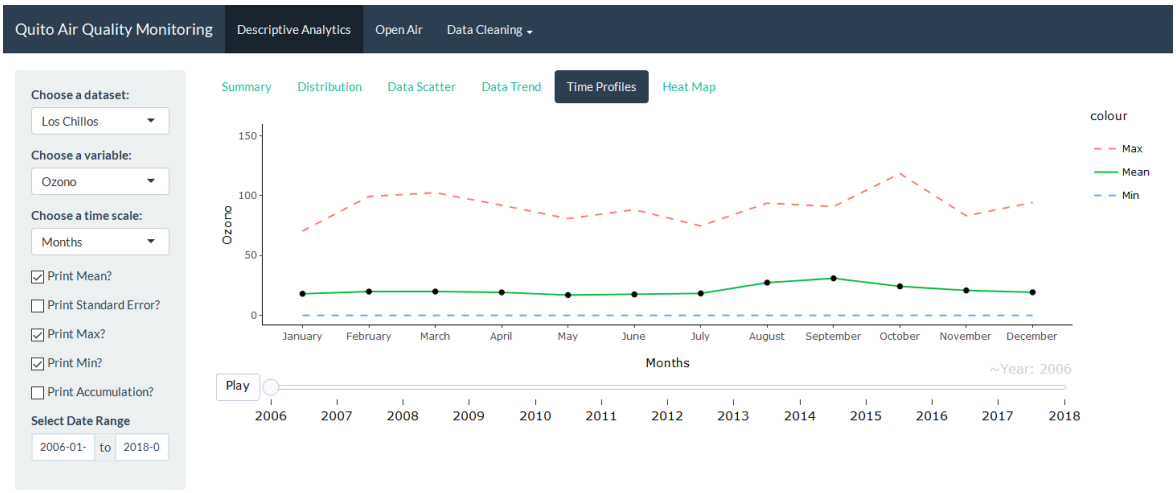

Figura 12. Perfiles de tiempo de variable Ozono en estación Los Chillos – Periodo Mes, datos enero 2006 – junio 2018.

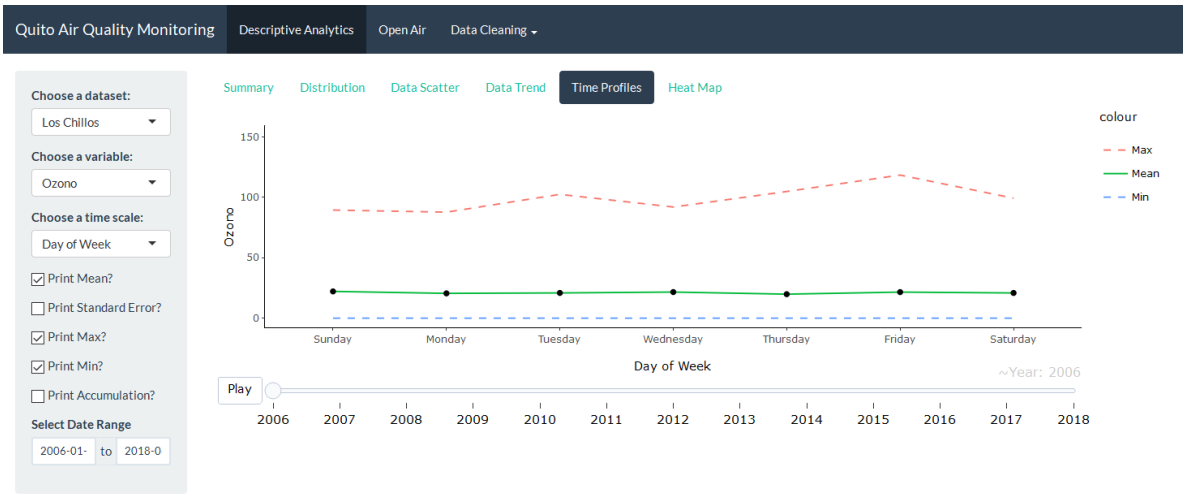

Figura 13. Perfiles de tiempo de variable Ozono en estación Los Chillos – Periodo Día, datos enero 2006 – junio 2018.

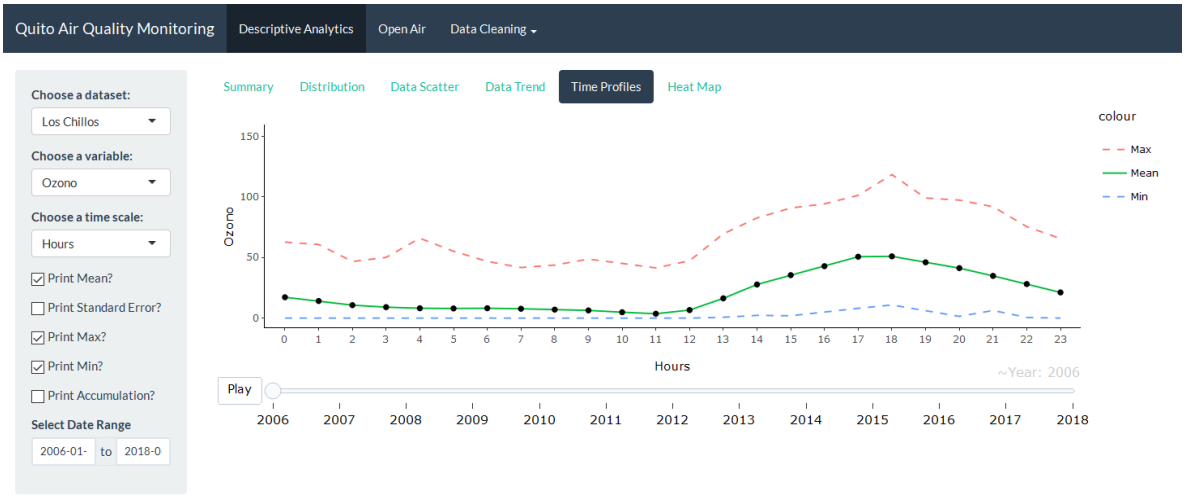

Figura 14. Perfiles de tiempo de variable Ozono en estación Los Chillos – Periodo Hora, datos enero 2006 – junio 2018.

5.3.3. **Rosas de contaminantes, rosas de viento y diagramas polares**

Las rosas de contaminantes, rosas de viento y diagramas polares son métodos para el análisis de frecuencias de distribución de valores relativos a la distancia y ángulo de dirección desde el punto de análisis. Estos diagramas permiten analizar aspectos de fuentes de contaminantes o puntos que afecten la lectura de variables de atmosféricas. Los gráficos implementados en la aplicación provienen del paquete open air.

En la implementación de las funciones en la aplicación se hizo énfasis a la generación de controles dinámicos para la personalización de los gráficos resultantes. Estos controles incluyen: selección de estación de estudio, selección de variable atmosférica disponible en la estación, colores para la representación de valores, métodos de conteo de valores, periodos de tiempo y tamaño de representación del gráfico. Estos controles se pueden en uso en Fig.15, donde se genera una rosa de viento de los datos de la estación (Los chillos) mediante el método (conteo proporcional) representando los datos con colores (por defecto).

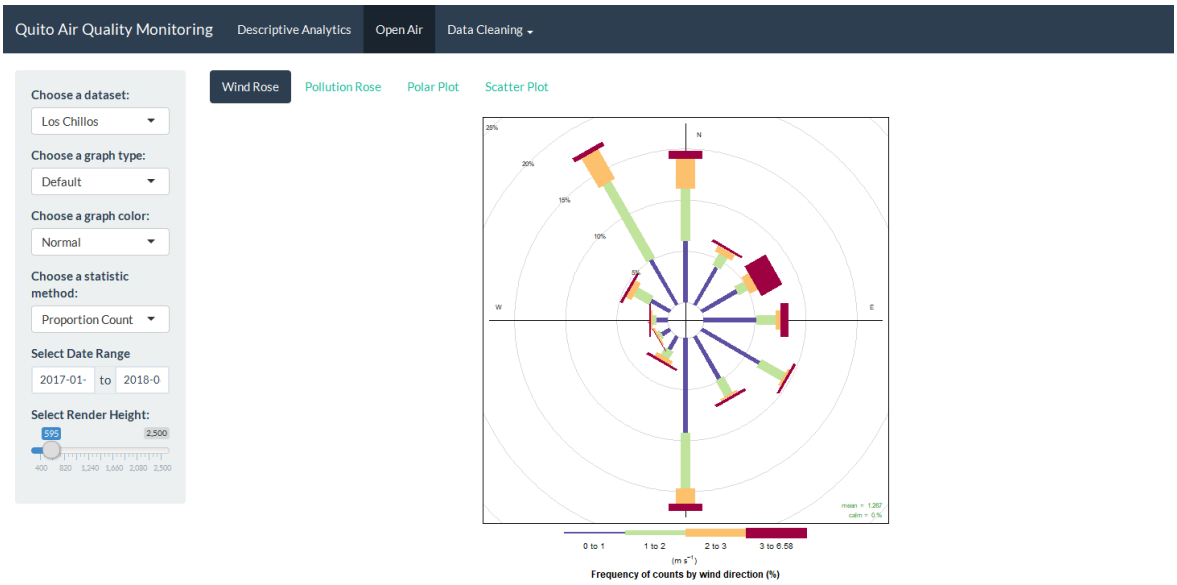

Figura 15. Rosa de Viento en estación Los Chillos, datos enero 2017 – junio 2018.

En relación con los diagramas polares, existen dos capacidades de esta función que la diferencian de las funciones de rosa de viento y rosa de contaminantes. Estas son: cálculo de incertidumbre y nivel de resolución para el cálculo. El cálculo de incertidumbre es la capacidad para en base a los datos provistos generar un cálculo predictivo de valores futuros. El nivel de resolución permite generar un gráfico en dos niveles: normal y alta resolución. Esto aumenta la exactitud de los resultados de los cálculos realizados, necesitando un mayor poder de procesamiento. Ambas opciones no son activadas por defecto dejándose en su lugar como opciones para el usuario, se pueden ver ambas activadas en Fig.16.

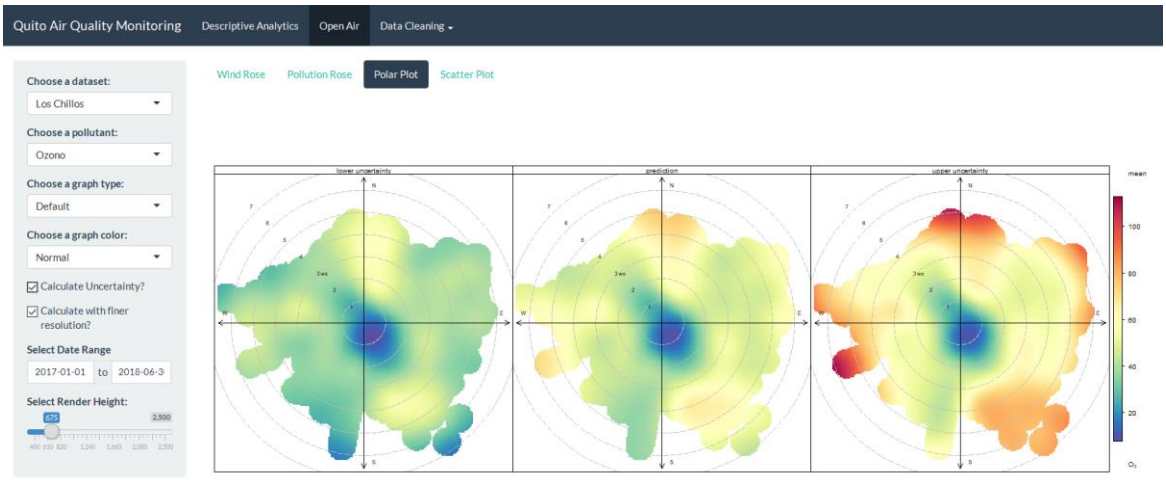

Figura 16. Diagrama polar en estación Los Chillos con incertidumbre y alta resolución, datos enero 2017 – junio 2018.

El paquete open air implementa múltiples maneras de agrupar los datos climáticos recopilados en base a periodos de tiempo como: día de la semana, meses, años, estaciones climáticas, etc. A su vez permite la modificación de los colores usados para la representación de los valores representados en los gráficos, en Fig.17-19 podemos ver múltiples ejemplos de rosa de contaminantes generados en base a los mismos datos (variable ozono en estación los chillos en periodo enero 2017 – junio 2018) agrupados por diferentes variables de tiempo y colores de presentación de datos: día de la semana – *increment*, año – normal y estación climática – *Jet* respectivamente.

Las capacidades de agrupación y representación de colores son compartidas por todos los gráficos de esta librería, incluyendo también cálculos como incertidumbre con ciertas limitaciones. En Fig.20 podemos ver un diagrama polar de variable ozono agrupado por mes y haciendo uso del color *viridis*. En Fig.21 podemos ver un diagrama polar haciendo uso del cálculo de incertidumbre agrupado por día de la semana, podemos ver el cambio en el color de representación sin embargo el diagrama resultante no presenta la estructura de agrupación esperada, en su lugar refleja la estructura predictiva haciendo uso de los datos agrupados por día de la semana como base para el cálculo.

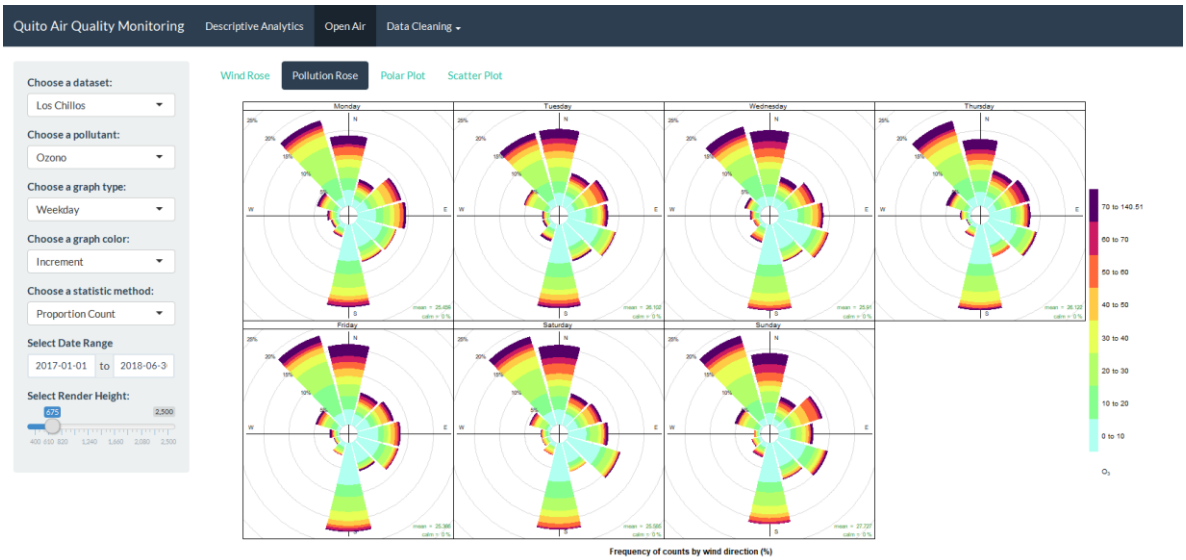

Figura 17. Rosa de Contaminantes de variable Ozono en estación Los Chillos – seccionado por día de la semana usando color (*Increment*), datos enero 2017 – junio 2018.

Figura 18. Rosa de Contaminantes de variable Ozono en estación Los Chillos – seccionado por año usando color (Normal), datos enero 2017 – junio 2018.

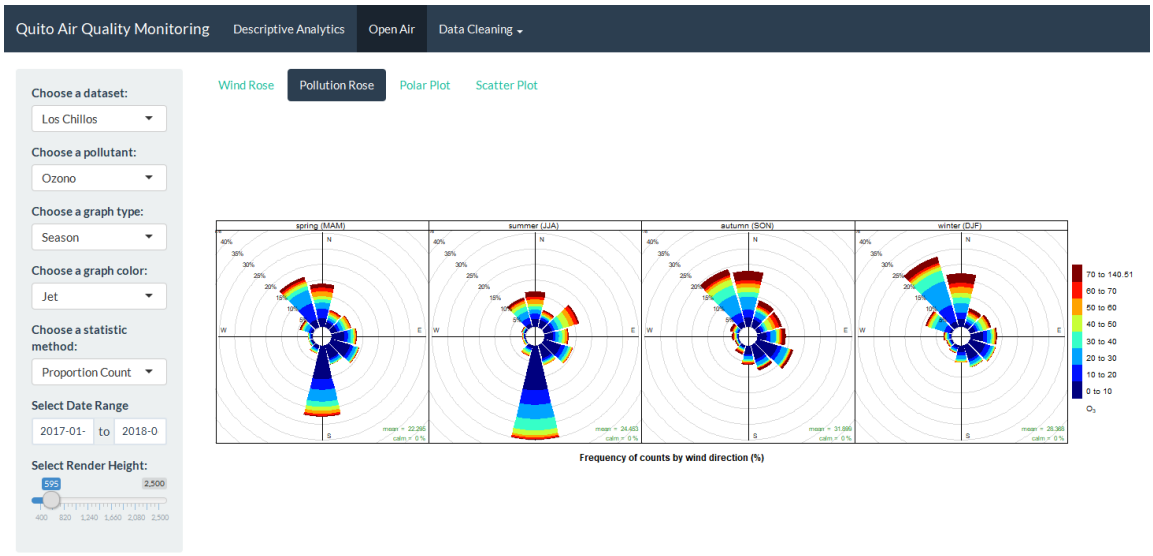

Figura 19. Rosa de Contaminantes de variable Ozono en estación Los Chillos – seccionado por estación climática usando color (JET), datos enero 2017 – junio 2018.

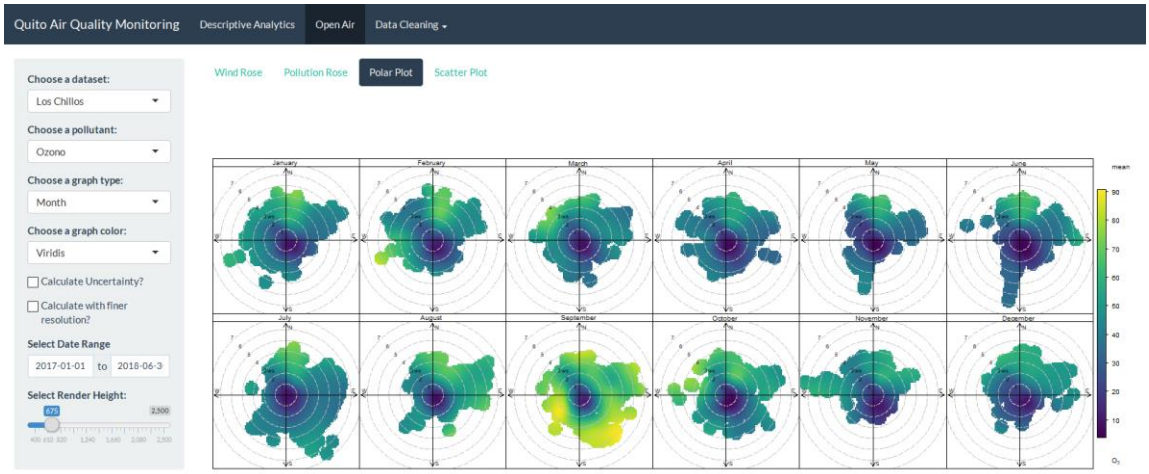

Figura 20. Diagrama polar de variable Ozono en estación Los Chillos – seccionado por mes usando color (*Viridis*), datos enero 2017 – junio 2018.

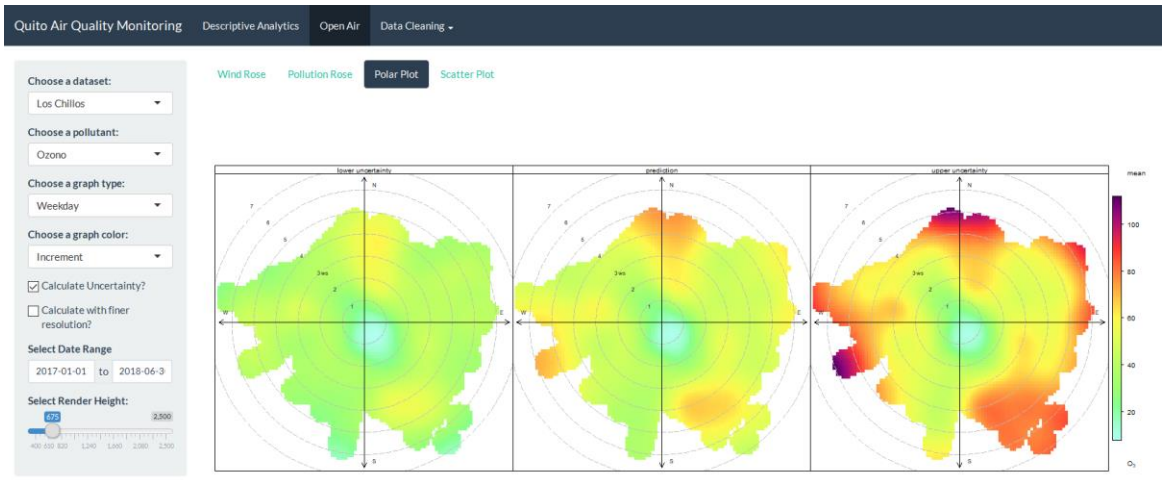

Figura 21. Diagrama polar de variable Ozono en estación Los Chillos – seccionado por día de la semana usando color (*Increment*) con cálculo de incertidumbre, datos enero 2017 – junio 2018.

5.3.4. **Mapas de calor**

El mapa de calor es un gráfico usado para visualización de intensidad de valores haciendo uso de una escala de colores para representar la relación entre cada dato. En el contexto de la aplicación los mapas de calor difieren de las rosas de viento, rosas de contaminantes y diagramas polares en su representación de manera geolocalizada de valores dentro de un mapa de la ciudad de quito, haciendo un punto de referencia centralizado en la posición de la estación.

El usuario se encarga de parametrizar diferentes valores mediante los controles provistos, estos son: estación de análisis, variable atmosférica, tamaño de representación de cada punto, escala de representación de los datos, opacidad mínima de cada punto y opacidad total del diagrama generado. Estos controles se pueden representados en Fig.22, la cual muestra un mapa de calor para la variable (ozono) basado en datos de la estación (los chillos).

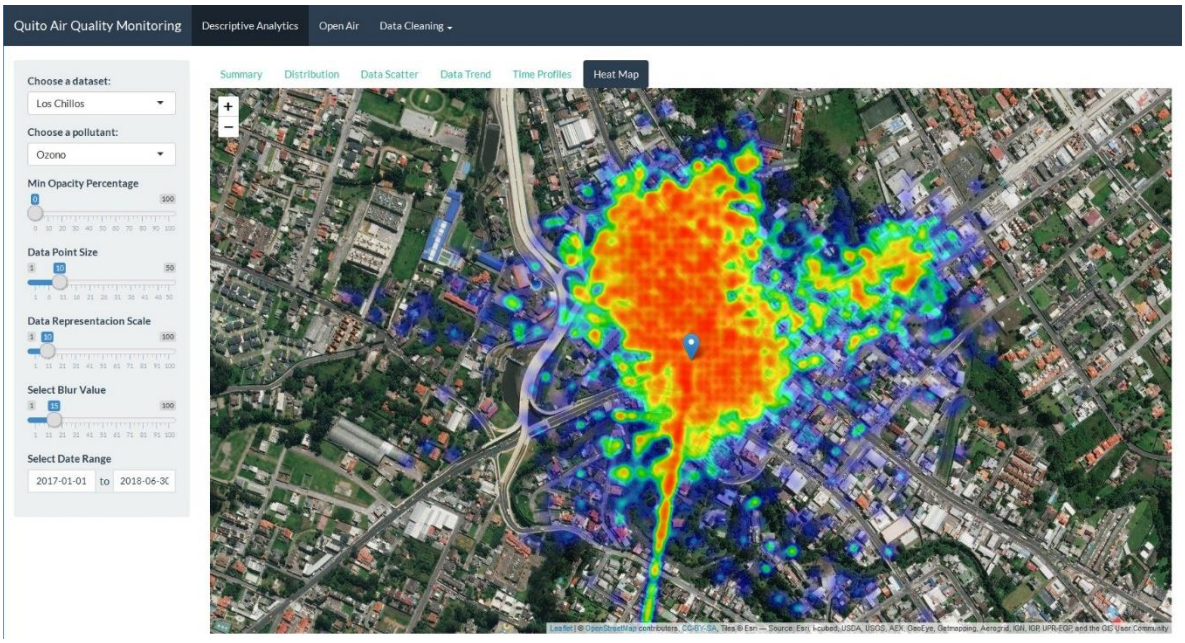

Figura 22. Mapa de calor de variable Ozono en estación Los Chillos, periodo enero 2017 – junio 2018.

Con los parámetros antes mencionados, el sistema hace uso del paquete *geosphere* que, mediante los datos de velocidad de viento, dirección de viento y escala se encarga de generar coordinadas relativas a la ubicación de la estación para su representación en un objeto del paquete leaflet. La generación del mapa de calor se realizó en base al paquete leaflet.extra, el cual es una recopilación de librerías de uso común con leaflet para la representación de escalas en un plano geolocalizado. El mapa de calor generado hace uso de los valores de la variable seleccionada para establecer la intensidad de color y su ubicación se basa en uso de las coordinas antes generadas. El resultado final es una representación geolocalizada basada en los valores de una variable, como se puede ver en Fig.23 para la variable (PM 2.5) en la estación (los chillos).

Aspectos de tamaño cada punto, escala de representación, opacidad mínima relativa y opacidad total del mapa generado son utilizados para mejorar la calidad de representación de la información por parte del usuario. Sin embargo, la aplicación provee valores por defecto que funcionan para la generación de gráficos comunes.

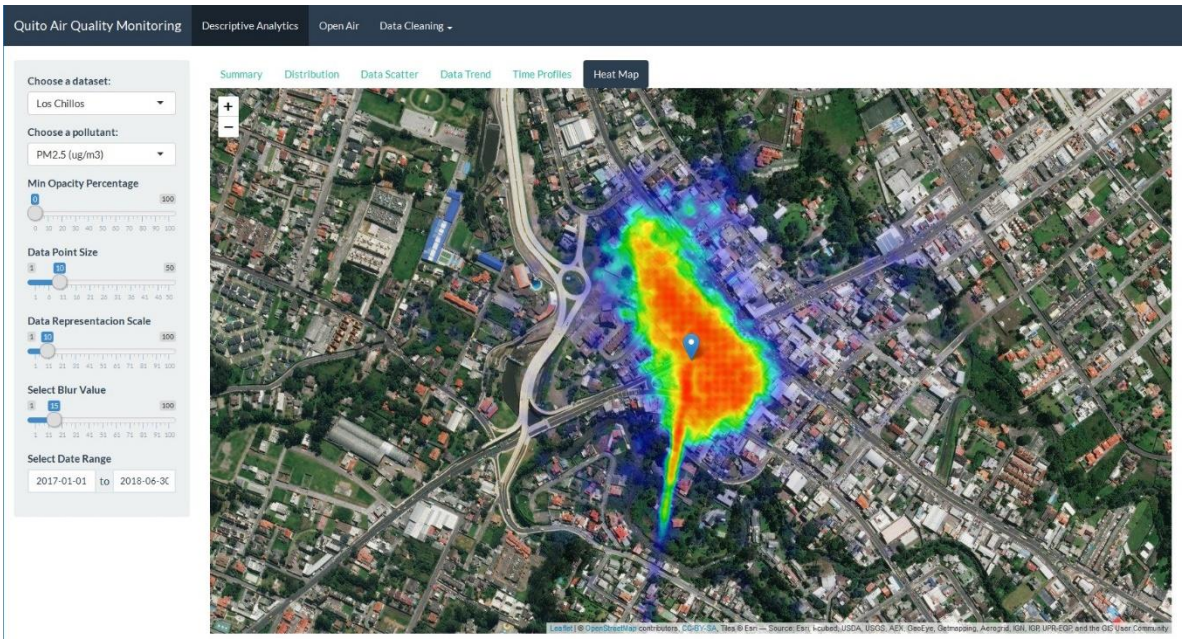

Figura 23. Mapa de calor de variable PM 2.5 en estación Los Chillos, periodo enero 2017 – junio 2018.

5.3.5. **Análisis multivariable y de tendencia**

Las funciones de análisis multivariable y de tendencia forman parte de la sección análisis descriptivo de la aplicación. Existen en total tres funciones que forman parte de esta categoría: diagramas de tendencia, diagramas de densidad y diagramas de dispersión.

En primer lugar, los diagramas de tendencia nos permiten visualizar aspectos de las variables relacionados con comportamiento de datos en una escala de tiempo, en contraste con los perfiles de tiempo, este grafico se genera en base a datos reales en lugar de agrupaciones. El uso del paquete lubridate nos permite generar gráficos de este tipo para múltiples escalas de tiempo de forma sencilla, lo que facilita el análisis de tendencias de datos en periodos de tiempo: muy extensos, relativamente raros o poco explorados.

Para utilizar esta función el usuario parametriza aspectos comunes de: estación de estudio, variable de análisis, rango de fechas y escala de tiempo. El sistema realiza el cálculo de puntos en los datos acordes con el periodo de selección y rango de valores para su representación en el grafico generado mediante *plotly*. Los controles

se pueden ver documentados en Fig.24 generando un gráfico de tendencia de datos de la variable (ozono) de la estación (los chillos) en un periodo de (años) en un rango (enero 2006 a junio 2018), se puede ver la diferencia presente en relación con Fig.25 donde se genera el mismo grafico con un periodo de tiempo (mensual). En ambos se identifica la escala progresiva de valores durante los últimos años, pero con relación a Fig.24, Fig.25 muestra la tendencia de valores más altos presentes a inicio de año y también mejor representa el cambio generado. Ambos análisis son útiles pero su utilidad varía de acuerdo con el estudio realizado.

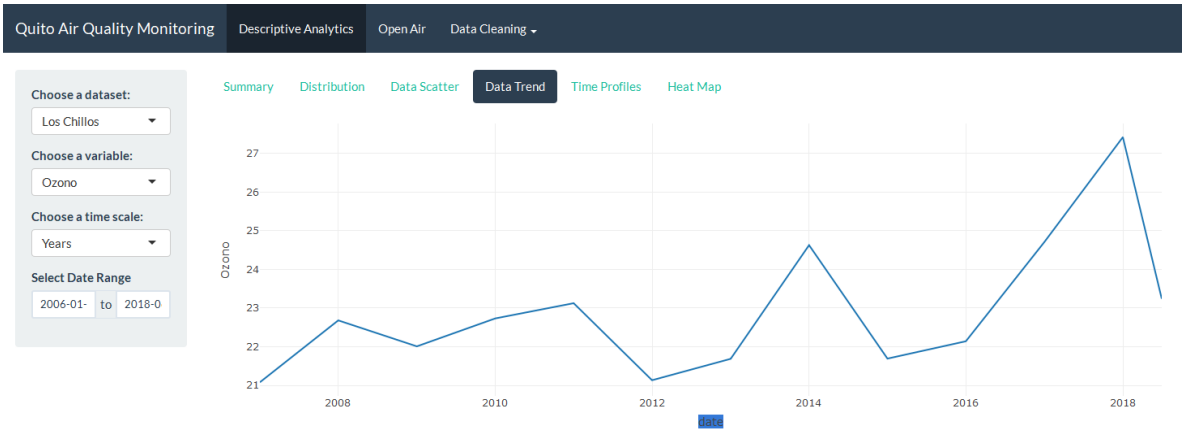

Figura 24. Diagrama de tendencia de variable Ozono en estación Los Chillos por periodo anual, enero 2006 - junio 2018

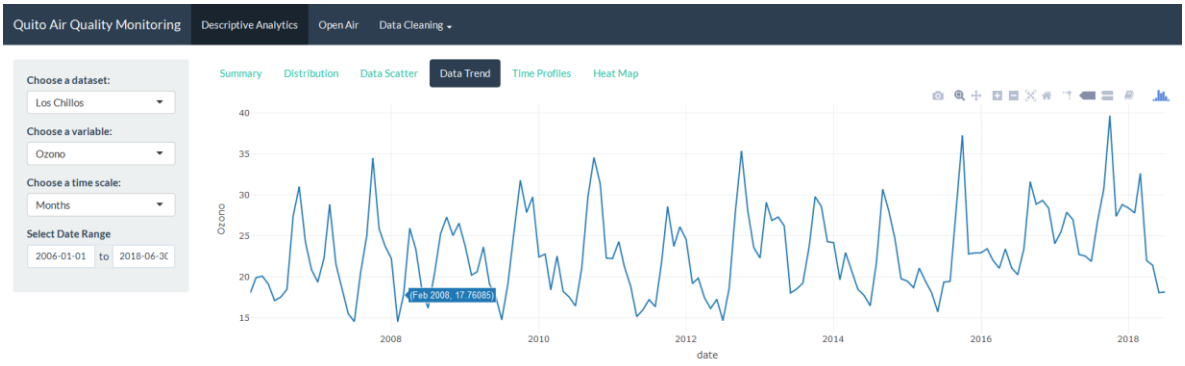

Figura 25. Diagrama de tendencia de variable Ozono en estación Los Chillos por periodo mensual, enero 2006 - junio 2018.

En segundo lugar, los diagramas de densidad nos permiten ver la distribución de valores presentes en un conjunto de datos de una variable en una estación. Esta

función presenta controles para la selección de variables: estación de estudio, variable de análisis, numero de secciones para dividir el conjunto de datos, escala de tiempo para promediar los datos y la capacidad de generar un análisis uni variable del conjunto de datos. Todos los controles son visibles en Fig.26 que representa la generación de un diagrama de densidad de la variable (ozono) en un periodo (diario) basado en datos de la estación (los chillos).

La función en base a los valores parametrizados se encarga de agrupar los valores en base al periodo de tiempo seleccionado y obtener el promedio de dichos datos. El promedio se utiliza tomando en cuenta que los valores provistos por la organización mundial de la salud mantienen esta escala. En un análisis común los datos se seccionan en base a los rangos generados usando el valor parametrizado por el usuario. Estos se agregan a un gráfico de tipo ggplo2 que se encarga de realizar el cálculo de densidad para mantener una escala orgánica en el gráfico. Véase Fig.26.

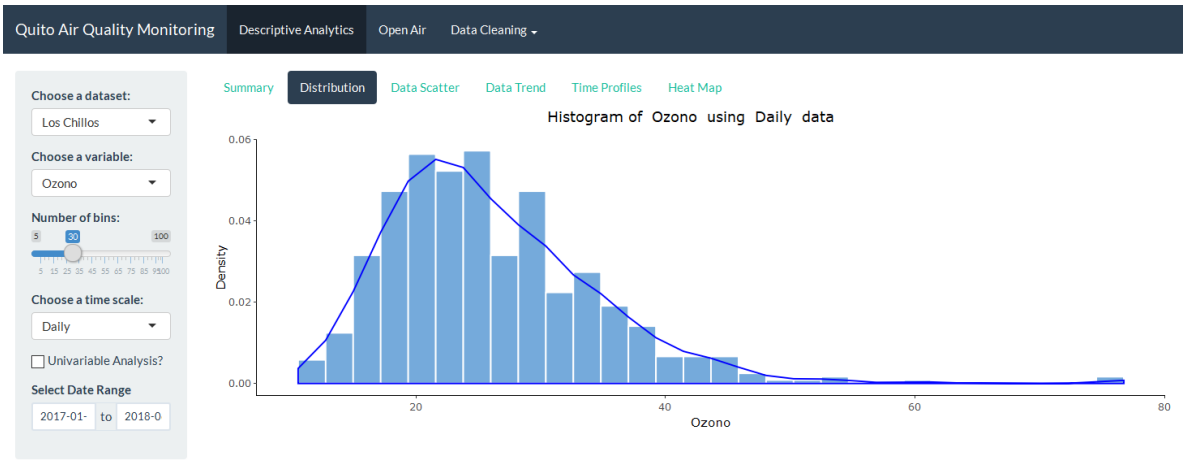

Figura 26. Diagrama de densidad de variable Ozono en estación Los Chillos por periodo diario, enero 2017 – junio 2018.

Al considerar un análisis uni variable, la interfaz de usuario se modifica para agregar la capacidad de seleccionar un contaminante y parametrizar el agregar un gráfico de densidad individual para cada variable. Véase Fig.27. Al generar un análisis uni variable el sistema realiza una búsqueda dinámica del valor recomendado por la organización mundial de la salud en base al contaminante seleccionado y el periodo

de tiempo, en base al valor encontrado se realiza el cálculo de una variable categórica denominada “saludable” la cual establece si el valor está por debajo el recomendado. Esta nueva variable se utiliza para dividir los datos antes mostrados en categorías: cumple (verde) y no cumple (rojo). Véase en Fig.27 un gráfico de densidad de datos de (temperatura) frente al cumplimiento del contaminante (Ozono). En este vemos que la totalidad de datos “no saludable” se encuentran en temperaturas mayores a 22 grados centígrados. Es importante notar que ciertos contaminantes no tienen indicaciones provistas para múltiples periodos de tiempo, el usuario deberá seleccionar solo los periodos de tiempo en los que existen dichos valores para verlos como una opción para su selección.

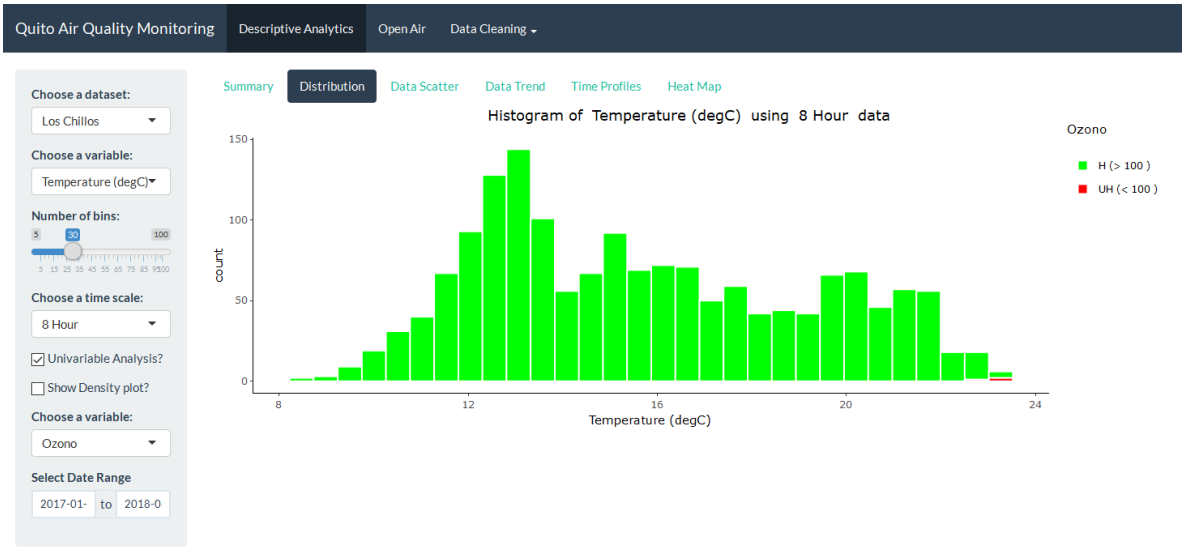

Figura 27. Diagrama de densidad de variable Temperatura con datos categóricos basados en Ozono en estación Los Chillos por periodo ocho horas, enero 2017 – junio 2018.

Finalmente, los diagramas de dispersión son herramientas para el análisis de relaciones de variables y contaminantes. Ciertos contaminantes son productos de temperaturas altas y otros contaminantes, como es el caso de O3. Debido a su importancia existen dos implementaciones diferentes en la aplicación basadas en: paquetes básicos de R y el paquete open air. La principal diferencia entre los resultados es la visualización de los datos, la función de open air provee la

capacidad de modificar colores de representación, modificar métodos para visualizar (puntos y hexágonos) y agrupación de datos. La función generada en base a los paquetes básicos de R se limita a la representación de valores, sin embargo, su simplicidad genera facilidad de uso que es ideal en individuos con capacidad limitadas de uso informático.

En ambos casos el usuario selecciona: estación de análisis, variable de estudio, variable de comparación y rango de tiempo. En el caso de la librería open air, se aumentan aspectos de selección de color, método de representación de datos y tipo de agrupación de datos. En Fig.28 vemos un ejemplo de esta función open air, en donde se genera un diagrama de dispersión en (hexágonos) de (radiación solar) frente (ozono) basado en datos de la estación (los chillos). En el caso de la función usando paquetes básico R, el usuario selecciona una variable para establecer los valores de color de representación y tamaño. En Fig.29 podemos ver un diagrama de dispersión generado en base a los datos de (temperatura) frente (ozono) basado en datos de la estación (los chillos). Es notable Fig.29 al mostrar claramente los valores promedios usados para el remplazo de valores fuera de rango y no existentes, estos se ven en el caso de O3 en la línea de valores visibles alrededor de 20 y en TMP alrededor de 15. Ambos se ven como una línea generado por los valores.

Figura 28. Diagrama de dispersión (Open Air) de variable ozono frente a variable radiación solar en estación Los Chillos método hexbin, datos enero 2017 – junio 2018.

Figura 29. Diagrama de dispersión de variable temperatura frente a variable Ozono en estación Los Chillos por periodo anual, enero 2017 – junio 2018.

5.4. Despliegue de la solución

Actualmente la aplicación se encuentra implementada en un servidor del servicio *Shinyapps* por R Studio dentro de la dirección https://rnaranjo.shinyapps.io/app-clima/. El servicio provee soporte, gestión de aplicaciones y mantenimiento de los paquetes para aplicaciones generadas mediante R *Shiny* a través del programa R Studio. El servicio posee licencias gratuitas que permiten la implementación de

aplicaciones con limites en el tiempo de disponibilidad de la aplicación al mes, con licencias pagadas que proveen mejoras en el procesamiento de la aplicación, tiempos de disponibilidad y funciones de tolerancias a fallos como respaldos activos.

El servicio se espera tenga un servidor físico basado en el sistema operativo Ubuntu Linux, que permita el uso del servicio a los investigadores del distrito metropolitano de Quito sin las limitaciones actualmente provistas por las capacidades del servidor en relación con tiempo de uso, capacidad limitada de procesamiento y tolerancia a fallos.

6. CONCLUSIONES y RECOMENDACIONES

6.1. Conclusiones

El uso de R en la implementación de la aplicación permitió gestionar de manera eficiente y efectiva el procesamiento de altos volúmenes de datos. La integración del paquete *Shiny* facilito la implementación de trece funciones con respuesta dinámica a las selecciones de los usuarios y más treinta controles de usuario para el manejo de parámetros. El paquete facilito el desarrollo de la interfaz del usuario, la parametrización de valores y la clara visualización de cambios en los gráficos generados.

La implementación de controles dinámicos en R complico la implementación de la aplicación sustancialmente en relación al uso de otras tecnologías para el desarrollo de sitios web. Mientras es posible generar controles que cambien en base a interacciones del usuario, su implementación se ve limitada por los altos tiempos de procesamiento vinculados a la generación de controles de este modo. En total se implementaron alrededor de diez controles dinámicos para la parametrización de variables meteorológica en base a la selección de estaciones y otros factores de visualización.

Finalmente, con respecto a la realización de procesos de limpieza de datos sobre los datos recopilados por la REMMAQ sobre las nueve estaciones y trece variables meteorológicas, se identificó que mientras el uso de *Python* permite agilizar y reducir el impacto estos procesos. El uso de un segundo lenguaje de programación complicaba aspectos de carga de archivos a futuro por miembros del equipo. El uso de un sistema multilenguaje web también complico aspectos de la implementación de la aplicación en un servidor.

6.2. Recomendaciones

Se sugiere la implementación de un servidor propio para la aplicación, esto ayudaría a mejorar tiempos de procesamiento, capacidades para la resolución de peticiones de múltiples usuarios, mantenimiento del servicio e implementación de tolerancias a fallos.

REFERENCIAS

Anniss, M. (2014). *What Is a Website and How Do I Use It?* Nueva York, Estados Unidos: Britannica Educational. Recuperado el 27 de Noviembre de 2018, de https://books.google.es/books?isbn= 1622750748

Baumer, B. S., Kaplan, D., & Horton, N. J. (2017). *Modern Data Science with R.* Boca Raton, Estados Unidos: CRS Press. Recuperado el 21 de Noviembre de 2018, de https://books.google.es/books?isbn=1498724493

Cuesta, H. (2013). *Practical data analysis.* Birmingham: Packt Publishing.

Iafrate, F. (2015). *From big data to smart data.* Hoboken, Estados Unidos: John Wiley & Sons.

IBM. (2016). *IBM SPSS Modeler CRISP-DM Guide.* Nueva York, Estados Unidos: IBM.

Jugulum, R., & Gray, D. H. (2014). *Competing with Data Quality.* Hoboken, Estados Unidos: John Wiley & Sons.

Lutz, M. (2013). *Leaning Python: Powerful Objetc-Oriented Programming.* Sebastopol, Canada: O´ Reilly Media. Recuperado el 21 de Noviembre de 2018, de https://books.google.es/books?isbn=9781449355692

Olmeda-Gómez, C. (2014). Visualización de Información. *El profesional de la información, 23*(3), 213-219.

R Project. (2018). *What is R?* Recuperado el 15 de Noviembre de 2018, de https://www.r-project.org/about.html

Ramírez, O., Mura, I., & Franco, J. (2017). How do people understand urban air pollution? Exploring citizens' perception on air quality, its causes and impacts in Colombian cities. *Open Journal of Air Pollution, 6*(1), 1-1.

Resnizky, H. G. (2015). *Learning Shiny.* Birmingham, Reino Unido: Packt Publishing.

Schwaber, K., & Sutherland, J. (2017). *La Guía Definitiva de Scrum: Las Reglas del Juego.* Recuperado el 2018 de Noviembre de 29, de https://www.scrumguides.org/index.html

Scrum, T. A. (2015). *Vanderjack, Brian.* Nueva York, Nueva York: Business Expert Press. Recuperado el 27 de Noviembre de 2018, de https://ebookcentral.proquest.com/lib/udlap/detail.action?docID=2145193

Secretaría de Ambiente del Municipio del Distrito Metropolitano de Quito. (2018). *Generalidades:Red de Monitoreo Atmosférico.* Recuperado el 27 de Noviembre de 2018, de http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades

Sinharay, R., Gong, J., Barratt, B., Ohman-Strickland, P., Ernst, S., Kelly, F. J., . . . Chung, K. F. (2018). Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls. *The Lancet, 391*(10118), 339-349.

Spinu, V., Grolemund, G., Wickham, H., Lyttle, I., Constigan, I., Law, J., . . . Lee, C. H. (11 de Abril de 2018). *Package 'lubridate'.* Recuperado el 17 de Diciembre de 2018, de R Project: https://cran.r-project.org/web/packages/lubridate/lubridate.pdf

Torres Ponjuán, D., & Herrero-Solana, V. (2010). *La visualización de la información en el entorno de la Ciencia de la Información.* La Habana: Universidad de La Habana.

Vanderjack, B. (2015). *The agile edge: managing projects effectively using agile scrum.* Nueva York, Estados Unidos: Business Expert Press.

Zalakeviciute, R., López-Villada, J., & Rybarczyk, Y. (2018). Contrasted effects of relative humidity and precipitation on urban pm 2.5 pollution in high elevation urban areas. *Sustainability, 10*(6), 1-21.

Zalakeviciute, R., Rybarczyk, Y., López-Villada, J., & Suarez, M. V. (2018). Quantifying decade-long effects of fuel and traffic regulations on urban ambient pm2. 5 pollution in a mid-size south american city. *Atmospheric Pollution Research, 9*(1), 66-75.