VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA AMBIENTAL

METODOLOGÍA PARA LA VALIDACIÓN DE DATOS DE CALIDAD DEL AIRE GENERADOS POR UNA RED DE MONITOREO AUTOMÁTICO

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN INGENIERÍA
INGENIERÍA AMBIENTAL - AIRE

P R E S E N T A :

Q. MARCOS HIDALGO NAVARRO

TUTOR:

M. EN C. VICENTE FUENTES GEA

2011

Jurado Asignado:

Presidente: Dr. Aguilar Márquez Armando

Secretario: Dr. Sosa Echeverría Rodolfo

Vocal: M.C. Fuentes Gea Vicente

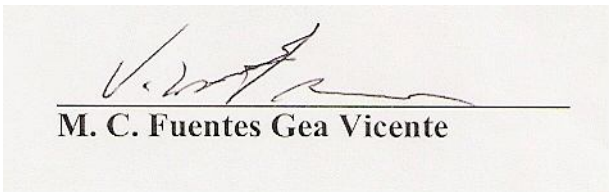
1er Suplente: M.I. Wellens Purnal Ann

2do Suplente: M.C. Audry Sánchez Javier

Lugar donde se realizó la tesis:

Posgrado de Ingeniería de la UNAM, México D.F.

Tutor de tesis:



Para Dorita, Adolfo y Diego, son lo que más amo.

Agradecimientos

A mis padres y hermanos por estar siempre y sin condiciones.
A mi comité tutoral, Dr. Armando Aguilar Márquez y Dr. Rodolfo Sosa Echeverría, por sus acertadas sugerencias y el seguimiento que dieron a mi trabajo.
A mis sinodales, M. en I. Ann Godelieve Wellens Purnal y Dr. Javier Audry Sánchez, por sus oportunas y valiosas observaciones.
A la Universidad Nacional Autónoma de México y al Consejo Nacional de Ciencia y Tecnología, por darme todas las herramientas necesarias durante mis estudios.
A todas las personas del Sistema de Monitoreo Atmosférico de la Ciudad de México, por sus opiniones y la información facilitada.
Finalmente y de manera muy especial a mi director de tesis, M. en C. Vicente Fuentes Gea, por su inquebrantable paciencia y por supuesto, por sus valiosas enseñanzas.

Contenido

Resumen

A través del monitoreo atmosférico automático se generan volúmenes considerables de información sobre la calidad del aire de zonas urbanas. En México, el Sistema Nacional de Información de la Calidad del Aire (SINAICA), reúne y difunde los datos que son generados por las principales redes automáticas de monitoreo atmosférico de la República Mexicana. Estos datos son revisados, validados y emitidos de acuerdo con los criterios propios de cada una de las redes en donde se generan; sin embargo, para tener la posibilidad de que al integrar las diferentes bases de datos estos cuenten con un grado de confianza equivalente y comparable, es necesario validarlos antes de su emisión y posterior conjunción siguiendo una misma metodología básica de validación. En el presente trabajo, se propone una metodología basada en la revisión de las características fundamentales del comportamiento del ozono en la Ciudad de México. Esta metodología incluye la aplicación de técnicas estadísticas, tanto gráficas como numéricas.

Al analizar los datos de ozono generados en la ciudad de México se logro documentar su comportamiento, de manera que fue posible describir sus características fundamentales, tales como el horario de ascenso y descenso de concentraciones diarias, evolución de la curva típica de ozono, disminución paulatina en las concentraciones diarias a partir de la década de los 90´s, entre otros . Partiendo de lo anterior y con la aplicación de la metodología propuesta, se consiguió verificar la existencia de comportamientos típicos, así como la existencia de características de los datos no documentadas referentes a los valores máximos de concentración y a la dispersión de los datos, estos dos últimos con la consideración de que corresponden a información específica generada en una estación de monitoreo que impide generalizar los resultados hasta no realizar la validación de datos de otras estaciones de monitoreo.

Glosario de términos

Antropogénica: Se refiere a los efectos, procesos o materiales que son el resultado de actividades humanas a diferencia de los que tienen causas naturales sin influencia humana.
Bandera: Es un código alfa-numérico que sirve para identificar eventos extraordinarios ajenos a la medición y datos confiables que pueden ser utilizados para análisis posteriores.
Calibración: Conjunto de operaciones que establecen, en condiciones especificadas, la relación entre los valores de las magnitudes indicadas por un instrumento de medición o un sistema de medición, o los valores representados por una medida materializada o un material de referencia, y los valores correspondientes de la magnitud realizada por los patrones.
Calidad del aire: Estado de la concentración de los diferentes contaminantes atmosféricos en un periodo de tiempo y lugar determinados, cuyos niveles máximos de concentración se establecen en las normas oficiales mexicanas y que son catalogados por un índice estadístico atendiendo sus efectos en la salud humana.
Contaminantes criterio: (O_3, CO, SO_2, NO_2, Pb, PST, PM10 y PM2.5). Aquellos contaminantes normados a los que se les han establecido un límite máximo de concentración en el aire ambiente, con la finalidad de proteger la salud humana y asegurar el bienestar de la población. Estos son: el ozono, el monóxido de carbono, el bióxido de azufre, el bióxido de nitrógeno, el plomo, las partículas suspendidas totales, y las partículas suspendidas menores a diez y a 2.5 micrómetros.
Criterio de suficiencia de datos: Cantidad mínima de datos para realizar un análisis estadístico representativo.
Datos crudos: Datos que se generan en las redes de monitoreo de la calidad del aire y muestreo de contaminantes atmosféricos, que no han pasado por ningún proceso de validación o revisión.
Equipo: Dispositivo o conjunto de dispositivos, que son utilizados en la operación del sistema de monitoreo de la calidad del aire.
Estación de monitoreo: Uno o más instrumentos diseñados para medir, de forma continua, la concentración de contaminantes en aire ambiente, con el fin de evaluar la calidad del aire en un área determinada. Una estación de monitoreo es utilizada para indicar en tiempo real cuál es la calidad del aire de la zona en donde está localizada la estación.
FAC: Función de autocorrelación.
INE: Instituto Nacional de Ecología.
Instrumento de medición: Los medios técnicos con los cuales se efectúan las mediciones y que comprende las medidas materializadas y los aparatos medidores.
Monitoreo atmosférico: Conjunto de metodologías diseñadas para muestrear, analizar y procesar en forma continua y sistemática las concentraciones de

sustancias o de contaminantes presentes en el aire.
RAMA: Red Automática de Monitoreo Atmosférico del SIMAT.
Red de muestreo y/o monitoreo: Las redes de medición se conforman por más de una estación de muestreo y/o monitoreo. Representan el conjunto de estaciones que miden la calidad del aire en una región determinada.
SMA: Secretaría del Medio Ambiente de la Ciudad de México.
SIMAT: Sistema de Monitoreo Atmosférico de la Ciudad de México.
SINAICA: Sistema Nacional de Información de la Calidad del Aire.
Sistemas de monitoreo de la calidad del aire: Un sistema de monitoreo consiste en un conjunto organizado de recursos humanos, técnicos y administrativos empleados para operar una o un conjunto de estaciones de monitoreo y/o muestreo que miden la calidad del aire en una zona o región.
SMA: Secretaría del Medio Ambiente de la Ciudad de México.
Validación. Conjunto de actividades dirigidas a la determinación de la validez y confiabilidad de un conjunto de datos.

Capítulo 1

Introducción

En zonas geográficas en las que se presentan actividades generadoras de emisiones contaminantes a la atmósfera y que además están densamente pobladas, es de interés primordial conocer la calidad del aire. La calidad del aire es "*el estado de la concentración de los diferentes contaminantes atmosféricos en un periodo de tiempo y lugar determinados, cuyos niveles máximos de concentración se establecen en las normas oficiales mexicanas y que son catalogados por un índice estadístico atendiendo sus efectos en la salud humana"*(PROY-NOM-156-SEMARNAT-2008**)**.Para conocer la calidad del aire, se requiere saber en que concentración se encuentran determinados compuestos químicos que dependen del tipo de actividades antropogénicas que se desarrollan en la zona. Las concentraciones de tales compuestos o contaminantes pueden ser medidas a través del monitoreo atmosférico continuo; este tipo de monitoreo se utiliza para calcular las concentraciones presentes en intervalos definidos de tiempo para contaminantes como ozono, óxidos de nitrógeno, monóxido de carbono, dióxido de azufre y partículas. Al medir de manera continua y automática, se generan grandes cantidades de datos o bases de datos de concentraciones de los contaminantes atmosféricos monitoreados.

En México existen diversas redes de monitoreo que operan de manera independiente unas de otras; la mayoría de estas redes se encuentran en estados de la República Mexicana como: Distrito Federal, Toluca, Puebla, Monterrey, Chihuahua, Guadalajara, Morelos y Baja California Norte. Cada una de estas redes depende de los gobiernos estatales o locales, y operan de manera autónoma. El gobierno federal a través del SINAICA reúne y difunde a través de la página WEB del INE los datos generados por las principales redes automáticas de monitoreo atmosférico de la República Mexicana con el objeto de dar a conocer la situación actual e histórica de la calidad del aire de diferentes ciudades del país.

El SINAICA integra paulatinamente en un solo portal de internet las bases de datos generadas por las siguientes redes de monitoreo: Ciudad de México, Monterrey, Guadalajara, Toluca, Puebla, Salamanca, León, Celaya, Irapuato, Silao, Ciudad Juárez, Tijuana-Rosarito-Tecate, Mexicali, Cuernavaca, Durango, Gómez Palacio, Torreón, San Luis Potosí y Región Tula-Tepeji. En este intento por crear una gran base de datos a nivel nacional, se han presentado situaciones a tomar en cuenta inherentes al funcionamiento y operación independientes de cada una de las redes, situaciones tales como: marcas de equipo, técnicas analíticas de los equipos, edad de los equipos, mantenimiento y en general todo el diseño técnico y operativo de las redes, así como los criterios y herramientas que se utilizan para la revisión o validación de los datos que se generan. Cada una de las redes de monitoreo integradas al SINAICA cuentan con sus propios procedimientos de validación o revisión de las bases de datos, procedimientos que aplican criterios de revisión diferentes y que por lo tanto son diferentes para cada red.

Es de fundamental importancia conocer el grado de confianza que tienen los datos generados por un sistema de monitoreo atmosférico por varias razones. Los datos obtenidos del monitoreo del aire pueden ser utilizados en la toma de decisiones que pueden afectar la salud humana y la calidad del ambiente, influir en las políticas ambientales y tener impacto económico directo e indirecto en la población. Por otra parte, la obtención de datos de calidad del aire a través del monitoreo atmosférico automático integra una serie de procesos que requieren de precisión y exactitud, además de que económicamente son muy costosos, dos razones más por las que es necesario generar datos confiables.

Para tener la posibilidad de integrar diferentes bases de datos con un grado de confianza y validez equivalente y comparable, es indispensable validarlas antes de su emisión y posterior conjunción siguiendo una misma metodología básica de validación. El proceso de validación de los datos es un procedimiento mediante el cual se determina y denota la calidad o validez de un conjunto de datos. El objetivo principal de este proceso es tener datos de calidad conocida mediante la evaluación documental, técnica y estadística para determinar la confiabilidad de la información generada por el sistema de monitoreo. Es deseable, que al integrar las bases de datos de diversas redes de monitoreo, éstas sean revisadas y validadas mediante un mismo procedimiento de carácter básico y general.

En el capitulo 2 se define en que consiste el proceso de validación de datos y se hace una breve descripción de las partes que integran al proceso de validación completo. La atención primordial se centra en la parte del proceso de validación en que se utiliza un conjunto de herramientas estadísticas para la validación de datos.

Este documento está dirigido a los operadores de los sistemas de información de las diferentes redes de monitoreo atmosférico automático que operan en el país, con lo cual se pretende iniciar un proceso de homologación de los criterios de validación de datos en todas estas.

Objetivo general

- Establecer una metodología para la validación de datos de calidad del aire que homologue el manejo de la información emitida por las diferentes redes de monitoreo atmosférico en el país.

Objetivos específicos

- Establecer una metodología basada en herramientas gráficas y estadísticas para validar datos generados a través del monitoreo atmosférico automático.
- Aplicar la metodología establecida a datos generados por la Red Automática de Monitoreo Atmosférico de la Secretaría del Medio Ambiente de la Ciudad de México.

Alcances

- Se seleccionará una estación de monitoreo de la RAMA para aplicar la metodología a uno de los contaminantes fundamentales de la ciudad de México en un periodo de 10 años.
- Se determinarán las características fundamentales del comportamiento del conjunto de datos de la estación de monitoreo seleccionada en el periodo de 10 años.
- Con la aplicación de la metodología se asignará un nivel de validación a los datos seleccionados.
- Se emitirá una base de datos válida y confiable del contaminante y estación seleccionados, de manera que sirva como ejemplo de cómo se verifica que una base de datos cumple con los criterios de validación a establecer.

Capítulo 2

Marco teórico

La validación de datos, es un proceso por medio del cual se determina y denota la confiabilidad de un conjunto de datos que tienen en común el método de generación. Durante el proceso de validación se evalúa la consistencia interna, temporal y espacial de un conjunto de datos, se identifican datos ausentes, no lógicos, tendencias debidas al mal funcionamiento de instrumentos o a la operación de los mismos, y características no esperadas. El objetivo principal de este proceso es tener datos de calidad conocida mediante la evaluación documental, técnica y estadística para determinar si los datos obtenidos de las mediciones son confiables.

2.1 Las características fundamentales de los datos

Parte del proceso de validación de un conjunto de datos de calidad del aire, se basa en la revisión de características fundamentales del comportamiento de los datos a través de diferentes métodos de revisión. Las características fundamentales de los datos son aquellos comportamientos que se presentan de manera regular o que siguen un patrón determinado, que al hacerlos evidentes, se pueden buscar de manera sistemática en otros conjuntos de datos del mismo tipo, y para este caso, en periodos de tiempo distintos.

El aire limpio está compuesto principalmente por nitrógeno y oxígeno, además, en pequeñas proporciones se puede encontrar vapor de agua y dióxido de carbono. *El aire contaminado contiene gases, polvos, olores y humos en grandes cantidades que dañan la salud de las personas, animales y plantas; estos contaminantes provienen de la adición de sustancias emitidas a la atmósfera que causan un desequilibrio en la composición original* (RAMA, SMA; 2009).
Los contaminantes criterio son *aquellos que están regulados por una norma que*

define los niveles de concentración en el aire recomendables para la protección de la salud humana. Estos son: ozono (O_3), partículas suspendidas totales (PST), partículas menores a 10 micrómetros (PM_{10}), partículas menores a 2.5 micrómetros ($PM_{2.5}$), monóxido de carbono (CO), dióxido de azufre (SO_2), dióxido de nitrógeno (NO_2) y plomo (Pb). El ozono se considera como uno de los contaminantes de mayor preocupación en la actualidad, ya que es altamente oxidante y afecta a los tejidos vivos, se asocia con diversos padecimientos en la salud humana. Los individuos que viven en zonas donde se registran regularmente concentraciones altas de ozono, presentan diversos síntomas, como: irritación ocular, de nariz y garganta, tos, dificultad y dolor durante la respiración profunda, opresión en el pecho, malestar general, debilidad, náusea y dolor de cabeza (RAMA, SMA; 2010).

En el presente documento se utilizan datos de ozono urbano, generados por la Red Automática de Monitoreo Atmosférico de la Secretaría del Medio Ambiente de la Ciudad de México.

De acuerdo con diversas publicaciones (RAMA, SMA; 2004) en donde se habla del comportamiento del ozono en diferentes escenarios tanto temporales como espaciales, se puede observar de manera general que el ozono en la Ciudad de México presenta diversos comportamientos característicos, como los que a continuación se describen:

- Las concentraciones diarias de ozono presentan un comportamiento típico, en donde cada día de manera general se presenta un aumento alrededor de las 8:00 h, un punto máximo alrededor de las 13 hrs y consiguientemente el descenso.

- El análisis del perfil diario de las concentraciones registradas en diversas estaciones, se mostró que de 1990 a 1996 hubo una disminución gradual de los niveles de ozono en todas las horas del día

- El comportamiento diario del ozono muestra una disminución paulatina a lo largo de la década de los 90´s, en los promedios horarios máximos.

- Desde 1990 disminuye la magnitud de las concentraciones máximas de ozono registradas en cada día de la semana.

- En la época seca caliente se presentan las concentraciones habituales mayores y en la época de lluvias las menores, aun cuando valores atípicos pueden ocurrir en cualquier época.

- Las máximas concentraciones diarias se registran en el mes de mayo para la época seca caliente, en noviembre y febrero para la época seca fría, y en agosto para la época de lluvias.

- El comportamiento de las concentraciones máximas diarias muestra diferencias entre épocas del año, la época seca caliente registra la menor variación mientras que la época seca fría y la de lluvias presentan variación similar.

Estas características que determinan el comportamiento de los datos de concentraciones de ozono en la Ciudad de México, se deben buscar de manera intencionada y sistemática en conjuntos de datos del mismo tipo. Para realizar la búsqueda de las características anteriores, de manera que ésta cumpla con un orden, se pueden agrupar tales características como se indica a continuación: **a)** comportamiento general de los datos a través del tiempo, **b)** comportamiento específico de los datos a través del tiempo, **c)** comportamiento y particularidades de los valores extremos, y **d)** comportamiento y particularidades de la dispersión de los datos.

2.2 El proceso de validación de datos

El proceso de validación de datos se distribuye típicamente en cuatro niveles, cada uno de ellos se identifican por medio de números o etiquetas (Knoderer et al., 2003). Un nivel de validación es una asignación numérica que indica el grado de confianza que se tiene en los datos, con la idea de que cada uno de los sitios o estaciones de monitoreo de una red califiquen la calidad de sus datos de manera equivalente y comparable (U.S. EPA, SLAMS, 1998). Cada nivel de validación representa la profundidad con que han sido revisados los datos y comúnmente se identifican con los números del 0 al 3; en cada nivel se utilizan diferentes técnicas o herramientas para la revisión de los datos.

A continuación se expone en que consiste la revisión de datos para cada uno de los niveles de validación, para posteriormente, y ya que el interés principal de esta tesis es el desarrollo de una metodología que permita validar datos de nivel 1 a nivel 2, se realiza una revisión y descripción de herramientas por medio de las cuales se puede calificar un conjunto de datos con el nivel 2 de validación, para finalizar con la exposición de dos pruebas de bondad de ajuste como una de las actividades iniciales propias del nivel 3.

Nivel 0

El nivel 0 se asigna esencialmente, a datos brutos o crudos obtenidos directamente de los sistemas de adquisición de datos. En estos datos puede haber cambiado el formato original de almacenaje o captura (el utilizado en la adquisición de los datos), pero no se editan y no están revisados (Knoderer et al., 2003). Estos datos no han recibido ningún ajuste por tendencias conocidas o

problemas que puedan haber sido identificados durante eventos como:

- mantenimientos preventivos
- revisiones rutinarias
- auditorias
- fallas en el funcionamiento de los equipos de monitoreo
- fallas en la calibración
- fallas en la transducción de datos
- cortes de energía

por mencionar algunos.

El uso común de estos datos es supervisar que la operación de los instrumentos ocurre de manera continua.

El nivel 0, indica un conjunto completo de datos (mas no suficiente) y de calidad no específica, en otras palabras, de esta manera se designa a los datos crudos. Este conjunto contiene todos los datos disponibles y también puede contener algún tipo de nomenclatura específica indicando datos perdidos o inválidos imputables al proceso técnico de generación, fallas de poder, perdida de comunicación y datos incompletos.

Una característica más de los datos de nivel 0, es que están expresados en unidades de medición bien definidas y estandarizadas.

Nivel 1

El nivel 1 indica un conjunto de datos sujetos a la revisión de los procedimientos de aseguramiento y control de calidad. Esto se da con la revisión de los procedimientos de operación estándar, bitácoras de operación, toda la documentación referente a la colección de datos y reportes relevantes del funcionamiento de los equipos (Knoderer et al., 2003). Dichos documentos de aseguramiento y control de calidad proporcionan información valiosa acerca de problemas potenciales o anomalías en el conjunto de los datos.

Algunos reportes específicos que deben revisarse son (PROY-NOM-156-SEMARNAT-2008):

a) Informes y documentos de colección, manejo, comunicación y reducción de datos, así como los procedimientos establecidos para tales efectos y verificación de su utilización.
b) Informes de control de calidad de estaciones de monitoreo, que documentan la operación de los sistemas de medición, y que incluyen los datos de muestras de control y cualquier otra medición de control de calidad interno.

c) Bitácoras de los sistemas de medición, como son las de: operación, mantenimiento preventivo, mantenimiento correctivo, calibración y desempeño, entre otras.
d) Informes de fallas e interrupción de los sistemas de medición debidas a cortes de energía, mantenimiento, calibración y verificación de desempeño.

Las anomalías aparentes en el registro de los datos, valores perdidos, desviación aparente de comportamientos ya conocidos y uso de procedimientos de operación no estandarizados de colección de datos, proveen pistas para la revisión preliminar de los datos en el nivel 2 de validación (U.S. EPA QA/G9R, 2006). Por otro lado, esta actividad familiariza al analista con los rasgos principales de la metodología con que son generados los datos, el diseño de muestreo y la manera en que cada una de las muestras son tomadas o medidas.

Con la revisión de los reportes y resultados de auditorías a los sitio o estaciones de monitoreo, se puede saber si la operación de las estaciones es comparable y si los datos que se generan entre ellas son equivalentes.

Al otorgar una etiqueta de nivel 1, se indica que los datos se están generando de forma homogénea es decir, que tienen consistencia interna en cuanto a la generación de estos y por consiguiente, este nivel de validación de datos emplea como técnica la observación directa de los registros del proceso completo de generación de datos.

Nivel 2

En este nivel, se realiza una revisión de los datos aplicando técnicas de estadística descriptiva así como la generación de gráficos; todo con el fin de conocer la estructura de los datos e identificar tendencias, relaciones y datos inciertos . En esta revisión se conoce el conjunto de datos e identifican anomalías que pueden influir en el análisis. Una manera de saber qué buscar, es basándose en la revisión anticipada de los procedimientos y documentación de la generación de datos mencionados en el nivel 1 (Knoderer et al., 2003).

Los datos de nivel 2 son examinados para verificar la consistencia interna, temporal y espacial cuando así corresponda, y siempre debe existir un registro de cambios que debe retenerse permanentemente.

De acuerdo a lo anterior el método que se emplea para asignar este nivel consiste en la aplicación sistemática de la estadística descriptiva.

Nivel 3

En esta etapa, se pretende explicar el comportamiento de los datos en función de otras variables relevantes que intervienen en el comportamiento de estos (Knoderer et al., 2003), por ejemplo, variables meteorológicas tales como: insolación, humedad, temperatura y velocidad y dirección del viento.

En el nivel 3 de validación, los datos se revisan a través del análisis e interpretación, para lo cual se utilizan datos que cuentan con la etiqueta de nivel 2. Al realizar el análisis a datos de nivel 2, es posible que se descubran inconsistencias, las cuales deben ser verificadas con una nueva revisión del proceso de validación a partir del nivel que se considere necesario.

Técnicas que con frecuencia se utilizan son regresión y regresión múltiple, con las cuales se pueden desarrollar, por ejemplo, modelos de predicción (U.S. EPA QA/G9, 2000). El usuario de los datos recomendará el nivel 3 de validación al generador de estos, en base a los resultados de la evaluación que el mismo usuario realice.

2.3 El nivel 2 de validación

Cuando se obtiene un conjunto de datos a menudo son tan numerosos que pueden parecer carentes de sentido, por lo que estos se deben ordenar, condensar o resumir de forma comprensible (U.S. EPA QA/G8, 2002). Pueden ser clasificados en forma sistemática y presentados en una tabla; para transmitir su significado más sencillamente o destacadamente, los datos pueden ser presentados por medio de gráficos, y se pueden calcular medidas descriptivas, tales como proporciones, promedios y dispersiones (estadísticos). Estos son dos elementos muy importantes en la revisión de los datos: las ***representaciones gráficas*** y ***medidas de la estadística descriptiva***, que básicamente con estos se toma la decisión de otorgar una etiqueta de nivel 2. Las representaciones gráficas se utilizan para identificar patrones y relaciones dentro del conjunto de datos, confirmar o refutar hipótesis e identificar problemas potenciales. Las medidas de la estadística descriptiva son funciones de los datos que describen con valores numéricos al conjunto de ellos; estas pueden ser usadas para proporcionar un cuadro mental de todos ellos y ser útiles para hacer inferencias acerca de la población de la que son extraídos (U.S. EPA QA/G9, 2000). Existe una gran número tanto de representaciones gráficas, como de medidas de estadística descriptiva; las más comúnmente utilizadas por su relativa sencillez de cálculo y que proporcionan mayor y/o mejor información se revisan a continuación.

2.3.1 Representaciones gráficas

El primer paso en el análisis de datos es frecuentemente un estudio gráfico de las características del conjunto de estos. El objeto de las representaciones gráficas es identificar patrones y tendencias en los datos que podrían pasar inadvertidos utilizando métodos completamente numéricos. Pueden usarse gráficos para identificar estos patrones y tendencias, confirmar o refutar hipótesis, descubrir nuevos fenómenos e identificar problemas en el comportamiento. Los descriptores gráficos de datos ilustran el carácter de un conjunto de datos, pero no los resumen.
Las representaciones gráficas incluyen despliegues de puntos de datos individuales, medidas estadísticas, datos temporales, datos espaciales y dos o más variables. Una sola representación gráfica no puede proporcionar un cuadro completo de los datos, por lo que se pueden escoger técnicas gráficas diferentes para remarcar rasgos diferentes de los datos.
A continuación, se describen la utilidad y construcción de cuatro tipos de gráficos que aportan diferente información.

a) Histograma

El histograma es un gráfico que indica la frecuencia con la cual ocurren los eventos (ya sean discretos o continuos) en intervalos discretos (Montgomery et al., 2003). Esto se representa a través de rectángulos de base uniforme y de áreas proporcionales a frecuencias de los datos. Una distribución de frecuencias es a menudo representada en la forma de un histograma, que consiste en un conjunto de rectángulos que tienen sus bases en el eje de las “x”, con centros en lo que se designa marcas de clase y ancho uniforme que se llama intervalo de clase o ancho de clase; el alto de cada rectángulo corresponde a la frecuencia o número de datos en cada clase.

El ancho de clase seleccionado no debe ser muy grande, de lo contrario la forma verdadera de la distribución se puede enmascarar o deformar; si el ancho de clase seleccionado es muy pequeño, entonces pueden aparecen muchos huecos en el histograma resultante. No existe una regla única para agrupar los datos en clases, pero de manera general se recomienda construir no menos de 7 ni más de 20 clases. El ancho de cada clase, ac, debe ser uniforme y se puede determinar calculando el rango “R” de la distribución (ec. 2.1) (Montgomery et al., 2003) y dividiéndolo por el número de clases ”nc” que se deseen (ec. 2.2).

$$R = VMax - VMin \qquad (2.1)$$

en donde VMax es el valor máximo del conjunto de datos y VMin el valor mínimo

$$ac = \frac{R}{nc} \qquad (2.2)$$

en donde nc es el número de clases.

La primer clase se construye a partir de un valor apenas por debajo del Vmin a la suma de este más el ancho de clase calculado. La segunda clase comienza a partir del límite superior de la primera clase y termina sumándole el ancho de clase calculado. De la misma manera se continúa hasta abarcar el rango completo de datos. Ya construidas las clases se determina cual es la frecuencia para cada una de ellas, esto es, cuantos elementos de la distribución inciden en cada una de las clases. Con las clases en el eje de las abcisas y las correspondientes frecuencias en el eje de las ordenadas se construyen los rectángulos que integran al histograma.

La construcción del histograma se ejemplifica en el anexo A1.

b) Datos ordenados

Un gráfico de datos ordenados muestra los datos del más pequeño al más grande a intervalos uniformemente espaciados (U.S. EPA QA/G9R, 2006). Este gráfico es una representación útil y fácil de construir, fácil de interpretar y no hace ninguna suposición sobre un modelo para los datos. Además, el gráfico de datos ordenados muestra todos los datos, por consiguiente, es una representación gráfica de los datos en lugar de un resumen.
El eje de las ordenadas representa las unidades originales de medición de la variable; el eje de las abcisas representa el orden de los datos por magnitud o tamaño.

La construcción del gráfico de datos ordenados se ejemplifica en el anexo A2.

c) Gráfico de dispersión

Para conjuntos de datos que consisten en observaciones apareadas donde dos o más variables continuas son medidas, un gráfico de dispersión es una de las herramientas más útiles para analizar la relación entre dos o más variables. Los gráficos de dispersión son fáciles de construir para dos variables y muchos paquetes gráficos de computadora pueden construir gráficos tridimensionales. Un gráfico de dispersión muestra claramente la relación entre dos variables, y pueden identificarse datos extraños de una o las dos variables apareadas (U.S. EPA QA/G9R, 2006). Este gráfico también muestra si existe correlación entre ambas variables; los

patrones de no linealidad pueden ser obvios delante de uno de estos gráficos. Otro rasgo importante que puede descubrirse usando este gráfico es cualquier efecto de agrupamiento entre los datos.
En el gráfico de dispersión, el eje de las abcisas representa a la variable independiente y el eje de las ordenadas representa a la variable dependiente.

La construcción de un gráfico de dispersión se ejemplifica en el anexo A3.

d) Datos temporales

A los datos colectados a intervalos regulares de tiempo se les llama datos temporales, pues son un conjunto de magnitudes pertenecientes a diferentes periodos de tiempo de cierta variable o conjunto de variables (U.S. EPA QA/G9, 2000). Por ejemplo, puede coleccionarse datos de monitoreo del aire de un contaminante minuto a minuto o una vez al día. Al examinar datos temporales se puede estar interesado en las tendencias a través del tiempo, correlación entre periodos de tiempo, tendencias cíclicas, tendencias estacionales, tendencias direccionales y patrones estacionarios.

Las tendencias cíclicas son de escala pequeña, como una tendencia diaria donde los datos muestran el mismo modelo durante cada día. Las tendencias estacionales son modelos en que los datos se repiten con el tiempo, es decir, los datos suben y caen regularmente durante más de un periodo de tiempo. Las tendencias estacionales son mayores que las cíclicas, como una tendencia anual donde los datos muestran, por ejemplo, el mismo modelo de subida y bajada a través de todo el año. Los patrones estacionarios se identifican cuando en periodos de tiempo grandes (años) se observa el mismo comportamiento, tal vez estacional, pero sin ninguna tendencia a subir o bajar de manera global.

Algunas representaciones gráficas específicas para los datos temporales son a) las series de tiempo y b) el correlograma o función de autocorrelación (FAC) (U.S. EPA QA/G9, 2000).

d.1) Series de tiempo

Uno de los gráficos más simples de generar que proporciona una gran cantidad de información es un gráfico de tiempo. Este gráfico hace fácil identificar tendencias a gran y a pequeña escala sobre el tiempo (U.S. EPA QA/G9, 2000). Las tendencias en pequeña escala se presentan en un gráfico de tiempo como fluctuaciones en periodos cortos; por ejemplo, el ozono típicamente en el curso de un día muestra un aumento hasta la tarde, entonces disminuye, y este proceso se repite todos los días. Las tendencias de escala más grandes, como fluctuaciones estacionales,

aparecen como levantamientos regulares en el gráfico; por ejemplo, dependiendo de la ubicación de la zona urbana estudiada, los niveles de ozono tienden a ser más altos durante el verano que durante el invierno; los datos de ozono muestran una tendencia diaria y una tendencia estacional. También pueden identificarse fácilmente posibles datos dudosos usando un gráfico de tiempo.
El tiempo se traza en el eje horizontal y la observación correspondiente se traza en el eje vertical. Los puntos trazados en un gráfico de tiempo pueden ser unidos por líneas; sin embargo, se recomienda que los puntos trazados no se conecten para evitar crear un sentido falso de continuidad, a menos que se conozca de antemano un patrón bien definido.

La construcción del gráfico de una serie de tiempo se ejemplifica en el anexo A4.

d.2) Correlograma o función de autocorrelación (FAC)

Cuando se mide una variable X_t a través del tiempo, con frecuencia ésta se relaciona consigo misma, estos es, que el comportamiento de la variable puede depender de los valores pasados o anteriores al momento en que se mide. Un análisis de autocorrelación de una variable con su pasado, estudia el comportamiento histórico de las observaciones, en otras palabras, mide la tendencia de la serie a comportarse como lo hizo antes (U.S. EPA QA/G9, 2000).

El método de autocorrelación simple ó Función de Autocorrelación (FAC), evalúa los componentes de una serie, componentes tales como (Jiménez et al., 2004):

a) Tendencia: es el componente de largo plazo que representa el crecimiento o disminución en la serie sobre un periodo amplio.
b) Cíclico: es la fluctuación en forma de onda alrededor del componente de tendencia, el componente cíclico tiende a repetirse periódicamente.
c) Estacional: es un patrón de cambio que se repite a él mismo cada cierto periodo de la serie.
d) Aleatorio: indica si la serie de datos presenta o no, carácter aleatorio.

y estos se muestran a través de la construcción de un correlograma.

El correlograma es un gráfico que se utiliza para apreciar los componentes de una serie de datos cuando estos se coleccionan a intervalos regulares de tiempo y se construye a través del cálculo de los coeficientes de correlación $r_1, r_2, \ldots\ldots, r_k$. Los coeficientes de correlación corresponden a la relación entre las covarianzas de una variable en un desfase k y la varianza, en donde un desfase k representa el número de unidades de tiempo que separan a los datos involucrados en el r_k con el del inicio (U.S. EPA QA/G9, 2000).

Si X_1, X_2, …, X_n representan los datos ordenados por tiempo para puntos

espaciados regularmente, esto es, X_1 es colectado al tiempo 1, X_2 es colectado al tiempo 2 y así sucesivamente. Para construir un correlograma, primero se calculan los coeficientes de autocorrelación. Así para k = 0, 1, …,n; se calculan los r_k

$$r_k = \frac{g_k}{g_0} \qquad (2.3)$$

en donde g_k es la covarianza para el desfase k y g_0 es la varianza de los datos,

$$g_k = \frac{\sum_{t=k+1}^{n} \left(X_t - \overline{X}\right)\left(X_{t-k} - \overline{X}\right)}{n} \qquad (2.4)$$

$\overline{X}$ es la media aritmética de los datos y "n" el número de datos.

El gráfico se construye con las parejas de datos (k, r_k) para k=0, 1, …, n.

Una autocorrelación será significativa o no según el criterio de Quenouille (Jiménez et al., 2004), que ha demostrado que los coeficientes de autocorrelación de datos aleatorios tiene una distribución que se puede aproximar a una curva normal con media cero y una desviación estándar de $1/\sqrt{n}$.

Si una serie de datos en efecto corresponde a una serie de datos aleatorios, la mayoría de los coeficientes de autocorrelación debe ubicarse dentro de los limites

$$-Z/\sqrt{n} < 0 < Z/\sqrt{n} \qquad (2.5)$$

en donde:

-z/√n = LI, límite inferior;
y z/√n= LS, límite superior;
IZI=1.96, es el valor normal estándar para un nivel de confianza del 95%;
n, es el número de observaciones en la serie de datos.

Como ya se ha mencionado, la construcción de un correlograma tiene como finalidad el apreciar los componentes de una serie de datos, a través de las siguientes observaciones (Jiménez et al., 2004):

- Una serie de datos presenta carácter estacionario si el coeficiente de correlación "r_1" es cercano a 1 y los sucesivos r_2,.....,r_k coeficientes decaen rápidamente a cero, y no existe alguna tendencia ascendente o descendente a lo largo de su comportamiento promedio (comportamiento a largo plazo).

- Presenta carácter estacional, si se observan coeficientes máximos a intervalos regulares de k desfases (comportamiento a mediano plazo).

- Será aleatoria si los coeficientes $r_1, r_2, ..., r_k$ son cercanos estadísticamente a cero, o sea, que estén dentro de los límites establecidos en la ecuación (2.5).
- Tendrá correlación significativa si existen máximos r_k que sobrepasan los límites establecidos en la ecuación (2.5).

La construcción de tres gráficos de la FAC se ejemplifica en el anexo A4.

A4.1) Correlograma sin ausencias de datos.
A4.2) Correlograma con ausencias de datos.
A4.3) Correlograma de datos aleatorios.

2.3.2 Medidas de estadística descriptiva

Las medidas de estadística descriptiva resumen algunas características cuantitativas de los datos usando medidas estadísticas comunes. Algunas medidas útiles son: número de observaciones; medidas de posición relativa como percentiles; medidas de tendencia central como media, mediana y moda; medidas de dispersión como rango, variancia, desviación estándar, coeficiente de variación y rango intercuartil; medidas de simetría de la distribución o forma y medidas de asociación entre dos o más variables como correlación. Estas medidas pueden usarse entonces para describir y comunicar características del comportamiento de la población de la que los datos son extraídos.

A continuación, se describen la utilidad y cálculo de las medidas de estadística descriptiva más relevantes y que aportan diferente información.

a) Medidas de posición relativa

La posición relativa de una o de varias observaciones respecto al total de estas, se puede conocer con los percentiles. Un percentil es el valor de un dato que es mayor que o igual a un porcentaje dado de datos. En general, dado un conjunto de n mediciones, el p-ésimo percentil o percentil p, es el valor de *X* tal que por lo menos p% de las mediciones son menores o iguales al valor *X* y a lo mucho (100-*p*)% son mayores o iguales al valor de *X*.

Si se considera que $x_{.p}$ es el *p*-ésimo percentil y $X_1, X_2, \ldots, X_n$, representan n datos, para calcular el percentil $x_{.p}$ se ordenan los datos del menor al mayor y se

etiquetan como $X_{(1)}, X_{(2)}, \ldots, X_{(i)}, \ldots, X_{(n)}$; así (*U.S. EPA QA/G9, 2000*)

$$t = \frac{p}{100} \qquad (2.6)$$

en donde *"t"* representa el p-ésimo percentil en forma centesimal y "p" es el percentil en porcentaje

$$nt = e.d \qquad (2.7)$$

en donde, "n" el número de datos, "e" la parte entera del producto "nt", "d" la parte decimal del producto "nt".
Entonces el cálculo de del p-ésimo percentil *"$x_{.p}$"*, se calcula de la siguiente manera:

si $d = 0$ entonces $x_{.p} = \frac{\left(X_{(e)} + X_{(e+1)}\right)}{2}$ (2.8)

si $d \neq 0$ entonces $x_{.p} = x_{(e+1)}$ (2.9)

El cálculo de percentiles se ejemplifica en el anexo B1.

b) Medidas de tendencia central

Las medidas de tendencia central caracterizan el valor central de una muestra de datos; son valores típicos en el sentido de que se emplean a veces para representar todos los valores individuales de una serie o de una variable. Las tres medidas más comunes son el promedio, mediana y moda.

La medida usada más comúnmente del valor central de una muestra es el promedio, denotada por $\overline{X}$. Esta estimación del valor central de una muestra puede ser pensada como el "centro de gravedad" de esta y es un promedio aritmético para muestras simples o muestras con el mismo "peso" que está influenciada por los valores extremos (grandes o pequeños). Se expresa en las mismas unidades que la variable.

El promedio es la suma de todos los datos individuales de la muestra dividida por el número total de observaciones (Mc. Clave et al., 1997):

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (2.10)$$

en donde

$\overline{X}$ es la media aritmético o promedio

n el número total de observaciones

La mediana es la segunda medida más comúnmente usada como valor central de los datos, se denota por $\tilde{X}$. Esta medida coincide exactamente con el valor central de los datos cuando estos se ordenan del más pequeño al más grande, esto significa que la mitad de los datos son más pequeños y la otra mitad son más grandes que la mediana. La mediana es otro nombre que se da para el percentil 50 ($x_{.50}$). La mediana no es influenciada por los valores extremos de una distribución, lo cual la hace una medida muy conveniente de localización para distribuciones asimétricas. Al igual que el promedio, se expresa en las mismas unidades que la variable.

El cálculo de la mediana se realiza de la siguiente manera:

Si el número de datos es impar

$$\tilde{X} = X_{\frac{n+1}{2}} \qquad (2.11)$$

Si el número de datos es par, entonces

$$\tilde{X} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \qquad (2.12)$$

La tercera medida de tendencia central es la moda, se denota por Xm. La moda de una muestra es el valor que ocurre con mayor frecuencia (Mc. Clave et al., 1997). Este valor no existe siempre, o puede no ser único; es la medida de tendencia central menos utilizada. Sin embargo, la moda es útil para los datos cualitativos y al igual que el promedio y la mediana, se expresa en las mismas unidades que la variable.
Aunque la moda es un concepto sencillo y útil, su aplicación presenta algunos aspectos menos convenientes:

- Una distribución puede revelar que dos o más valores se repiten un número

igual de veces, y en tal situación no hay forma lógica de determinar que valor debe ser escogido como la moda, por lo cual se dice que es bimodal.

- Puede ocurrir que no se encuentre repetición de valor alguno.

- La moda es un valor muy inestable, puede cambiar drásticamente con el redondeo de los datos.

El cálculo de las medidas de tendencia central se ejemplifica en el anexo B2.

c) Medidas de dispersión

Las medidas de tendencia central son más significativas si se complementan con información de cómo los datos se distribuyen alrededor del centro de una muestra, lo cual se observa con las medidas de dispersión. Las medidas de dispersión en un conjunto de datos incluyen el rango, varianza, desviación estándar, coeficiente de variación y rango intercuartil.

La medida de dispersión más fácil de calcular es el rango de la muestra. Para las muestras pequeñas, el rango es fácil de interpretar y puede representar la dispersión de los datos adecuadamente. Para las muestras grandes, el rango no es muy informativo porque considera solo (y por consiguiente es altamente influenciado) los valores extremos. El rango no es en modo alguno una medida de dispersión de los datos intermedios con relación a la media. Se expresa en las mismas unidades de la variable y como ya se ha visto, se calcula con la ecuación (2.1).

La varianza y la desviación estándar miden la dispersión de los datos alrededor de la media de una muestra y son las medidas de dispersión más frecuentemente empleadas. Una varianza (o una desviación estándar) grande implica que existe una dispersión grande de los datos respecto a la media. Una varianza (o una desviación estándar) pequeña implica que hay poca dispersión de los datos alrededor de la media. La varianza de la muestra es afectada por los valores extremos, se expresa en las unidades de la variable al cuadrado y se calcula de la siguiente forma (Mc. Clave et al., 1997):

$$S^2 = \frac{\left(\sum_{i=1}^{n} X_i - \overline{X} \right)^2}{n-1} \qquad (2.13)$$

en donde

S^2 es la varianza de la muestra

$\overline{X}$ es la media aritmética o promedio

n el número total de observaciones

La desviación estándar es la raíz cuadrada de la varianza de la muestra, por lo tanto:

$$S = \sqrt{S^2} \qquad (2.14)$$

y tiene la misma unidad de medida que los datos, es por esto que es ampliamente utilizada.

Cuando se comparan dos o más conjuntos de datos cuyas unidades de medición son idénticas y sus medias aritméticas son aproximadamente iguales, se puede decir que una muestra tiene un menor grado de dispersión que otra si la primera tiene una menor varianza o desviación estándar. Cuando faltan estas condiciones se puede usar una medida relativa de dispersión, el coeficiente de variación. El coeficiente de variación es una medida adimensional que permite la comparación de la dispersión entre varios conjuntos de datos, aun cuando las unidades de estos no sean iguales. El coeficiente de variación se denota por CV , su cálculo se realiza dividiendo la desviación estándar entre la media y la cantidad obtenida, obviamente, es adimensional (Anderson et al., 1998).

$$CV = \frac{S}{\overline{X}} \qquad (2.15)$$

Ya se vió que el rango está sujeto a la probabilidad de datos extremos erráticos y no toma en cuenta la dispersión dentro de todo el rango. Para superar estas limitaciones se pueden utilizar los rangos interpercentiles. Estas medidas estadísticas no son dependientes de valores extremos.
El rango intercuartil, uno de los más utilizados, solo incluye el 50% de en medio de un conjunto de datos; es decir, un cuarto de las observaciones en el extremo inferior y otro cuarto de las observaciones en el extremo superior de la distribución son excluidos .

El rango intercuartil es la diferencia entre el percentil 25 (cuartil inferior primero) y el percentil 75 (cuartil superior o tercero) y se expresa en las mismas unidades de la variable (U.S. EPA QA/G9, 2000).

$$\text{rango intercuartil} = x_{.75} - x_{.25} \qquad (2.16)$$

El cálculo de las medidas de dispersión se ejemplifica en el anexo B3.

d) Medidas de asociación

Los datos incluyen a menudo más de una variable y puede haber interés en saber cuál es el nivel de asociación entre dos o más de estas variables. Una de las medidas más comunes de asociación es el coeficiente de correlación "r", éste mide la relación entre dos variables, sin embargo, el coeficiente de correlación no implica causa y efecto (Anderson et al., 1998). El coeficiente de correlación puede indicar que la correlación entre dos variables es alta y la relación es fuerte, pero no puede indicar que una variable es causante de que la otra variable aumente o disminuya sin más evidencia.

Una asociación lineal implica que cuando una variable crece la otra crece o decrece linealmente, y cuando una variable decrece la otra crece o decrece linealmente. Un valor del coeficiente de correlación cercano a +1, implica que cuando una variable crece así lo hace la otra, y de manera inversa para los valores cerca de -1. Un valor de +1 implica una correlación lineal positiva perfecta, es decir, todos los pares de datos describen una línea recta con una pendiente positiva. Un valor de -1 implica una correlación lineal negativa perfecta, describiendo así una recta con pendiente negativa. Los valores cercanos a 0 implican correlación débil entre las variables.

Tabla 2.1. Guía de interpretación del tamaño de r.

Irl	Interpretación
0	Correlación nula
0-0.5	Correlación baja
0.5-0.8	Correlación media
0.8-1	Correlación fuerte
1	Correlación perfecta

(Tomado de An Introduction to the Statistical Analysis of data; Anderson & Sclove; Houghton Mifflin Co.)

Si $X_1, X_2, \ldots, X_n$ representan los datos de una variable y $Y_1, Y_2, \ldots, Y_n$ representan los datos de otra variable, el coeficiente de correlación r entre X y Y se calcula con la ecuación (2.17) como sigue (Anderson et al., 1998):

$$r = \frac{s_{xy}}{s_x s_y} \tag{2.17}$$

en donde

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}} \tag{2.18}$$

que es la desviación estándar para la muestra de la variable x,

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}{n-1}} \tag{2.19}$$

es la desviación estándar para la muestra de la variable y,

$$s_{xy} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n-1} \tag{2.20}$$

que es la covariancia para x e y.

El cálculo del coeficiente de correlación se ejemplifica en el anexo B4.

2.4 Pruebas de bondad de ajuste.

Al tener un conjunto de datos que se ha revisado por medio de representaciones gráficas y medidas estadísticas, se puede entonces realizar el análisis e interpretación, independientemente de quién y con qué fin los utilice.

Para iniciar con alguna actividad de interpretación o análisis es necesario conocer el tipo de distribución que mejor describe al conjunto de datos, o en otro caso, saber que los datos no siguen algún tipo de distribución específica. Después de conocer cómo se distribuyen los datos, se pueden hacer inferencias acerca de estos escogiendo el tipo de métodos estadísticos adecuados, esto es, se puede decidir si para el análisis de los datos se debe utilizar estadística paramétrica o estadística no paramétrica. Muchas pruebas estadísticas y modelos son apropiados solo para datos que siguen una distribución particular. Dos de las distribuciones más importantes para pruebas que involucran datos ambientales son la distribución normal y la distribución lognormal; ambas distribuciones remiten al uso de estadística paramétrica. Para determinar si un conjunto de datos sigue una de estas distribuciones específicas, se pueden utilizar las pruebas de bondad de ajuste.

En el análisis de datos se pueden tener conjuntos que se supone siguen una distribución normal, binomial, de Poisson, etc. Para cada caso específico, las frecuencias de las distribuciones teóricas deben contrastar con las frecuencias observadas, a fin de saber si la distribución se adecua al modelo teórico.

2.4.1 Prueba de bondad de ajuste de Kolmogorov-Smirnov

La prueba Kolmogorov-Smirnov permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica específica; el objetivo es señalar si los datos de interés provienen de una población que se comporte como la distribución teórica específica. Mediante la prueba se compara la distribución acumulada de las frecuencias teóricas con la distribución acumulada de las frecuencias observadas, se encuentra el punto de divergencia máxima y se determina qué probabilidad existe de que una diferencia de esa magnitud se deba al azar.

Para aplicar la prueba de bondad de ajuste de Kolmogorov-Smirnov, se recomienda un tamaño de muestra mínimo de n=50. Esta prueba es más apropiada que la prueba chi cuadrada para muestras grandes (n > 50), pues en la construcción de grupos o intervalos de datos, la chi cuadrada requiere que cada grupo contenga al menos 5 elementos.

El procedimiento para realizar la prueba de bondad de ajuste de Kolmogorov-Smirnov es el siguiente (Mc. Bean et al., 1998):

a) Con el fin de obtener los estadísticos necesarios para la construcción de la prueba se construye un histograma de los datos.

b) Planteamiento de la hipótesis

Hipótesis nula (Ho). Las diferencias entre los valores observados y los valores de la distribución teórica se deben al azar.

Hipótesis alterna (Ha). Los valores observados de las frecuencias para cada clase son diferentes de las frecuencias teóricas.

c) Definición del nivel de significancia α

Para todo valor de probabilidad igual o menor que α, se rechaza Ho y se acepta Ha.

d) Definición de la región de rechazo

Para todo valor D' mayor que el valor crítico D, se rechaza la hipótesis nula.

e) Cálculo de las frecuencias observadas acumuladas $(Fobs)$.

f) Cálculo de las frecuencias esperadas acumuladas de la distribución teórica a considerar (Ft).

g) Cálculo del estadístico D', máxima discrepancia entre las frecuencias observadas F_{obs} y las frecuencias teóricas F_t, así

$$D' = \frac{Ft - Fobs}{n} \qquad (2.21)$$

en donde n es el número total de observaciones de la muestra.

h) Comparación del estadístico D' con el valor crítico del estadístico de Kolmogorov-Smirnov (D).

en donde D se calcula con la ecuación (2.22) para α=0.05 (tabla C5)

$$D = \frac{1.358}{\sqrt{n}} \qquad (2.22)$$

en esta, n es el número total de observaciones de la muestra.

i) Criterios para la aceptación o rechazo de hipótesis

Si $D' < D$ entonces no se rechaza Ho (2.23)

Si $D' > D$ entonces

se rechaza Ho y se acepta Ha (2.24)

El procedimiento para aplicar la prueba de bondad de ajuste de Kolmogorov-Smirnov es muy simple, quizá la parte que requiere mayor cuidado corresponde al cálculo de las frecuencias esperadas de cada tipo de distribución teórica.

La prueba de bondad de ajuste de Kolmogorov-Smirnov se ejemplifica en el anexo C1, probando para una distribución teórica normal.

2.4.2 Prueba de bondad de ajuste de Shapiro-Wilk

La prueba "W" de Shapiro-Wilk es otra prueba de bondad de ajuste que es adecuada para muestras de menor tamaño (n<50) y prueba la hipótesis nula de que los datos corresponden a una muestra aleatoria de una distribución normal seguida de una hipótesis alterna de que los datos no pertenecen a una distribución normal. Esta prueba se considera una de las mejores pruebas numéricas de normalidad.
La prueba puede ser aplicada en cualquier muestra de tamaño menor a 50. Conforme se acerca el tamaño de la muestra a 50 también aumenta la potencia de la prueba.

Para aplicar la prueba de bondad de ajuste de Shapiro-Wilk, se lleva a cabo el siguiente procedimiento (Mc. Bean et al., 1998):

a) Con el fin de obtener los estadísticos necesarios para la construcción de la prueba se construye un histograma de los datos.

b) Planteamiento de la hipótesis

   Hipótesis nula (Ho). La muestra de datos proviene de una distribución normal.

   Hipótesis alterna (Ha). La muestra de datos no proviene de una distribución normal.

c) Definición del nivel de significancia α

   Para todo valor de probabilidad igual o menor que α, se acepta Ha y se rechaza Ho.

d) Definición de la región de rechazo

   Para todo valor del estadístico W' menor que el valor crítico W de Shapiro-

Wilk se rechaza la hipótesis nula.

e) Cálculo del estadístico de prueba W'.

El estadístico de prueba W' se calcula con la ecuación (2.25)

$$W' = \left(\frac{b}{S\sqrt{n-1}} \right)^2 \qquad (2.25)$$

en donde para calcular b, es necesario construir la tabla 2.2.

Tabla 2.2. Coeficientes para la prueba de normalidad de Shapiro-Wilk.

I Numero de muestra	II Valor de la muestra	III X(i)	IV X(n-i+1)	V X(n-i+1) - X(i)	VI a(n-i+1)	VII b(i)
1 2 . n						
						b = Σ b(i)

en donde

X(i) son los datos ordenados del menor al mayor

a(n-i+1) son los coeficientes para la prueba de normalidad de Shapiro-Wilk obtenidos de tablas de coeficientes para el estadístico de normalidad (anexo D)

b(i) es el producto de las columna V y VI

f. Comparación del estadístico W' con el valor crítico W de Shapiro-Wilk (tabla C7).

g. Criterios para la aceptación o rechazo de hipótesis.

Si $W' > W$ entonces no se rechaza Ho (2.26)

Si $W' < W$ entonces
se rechaza Ho y se acepta Ha (2.27)

La prueba de bondad de ajuste de Shapiro-Wilk se ejemplifica en el anexo C2.

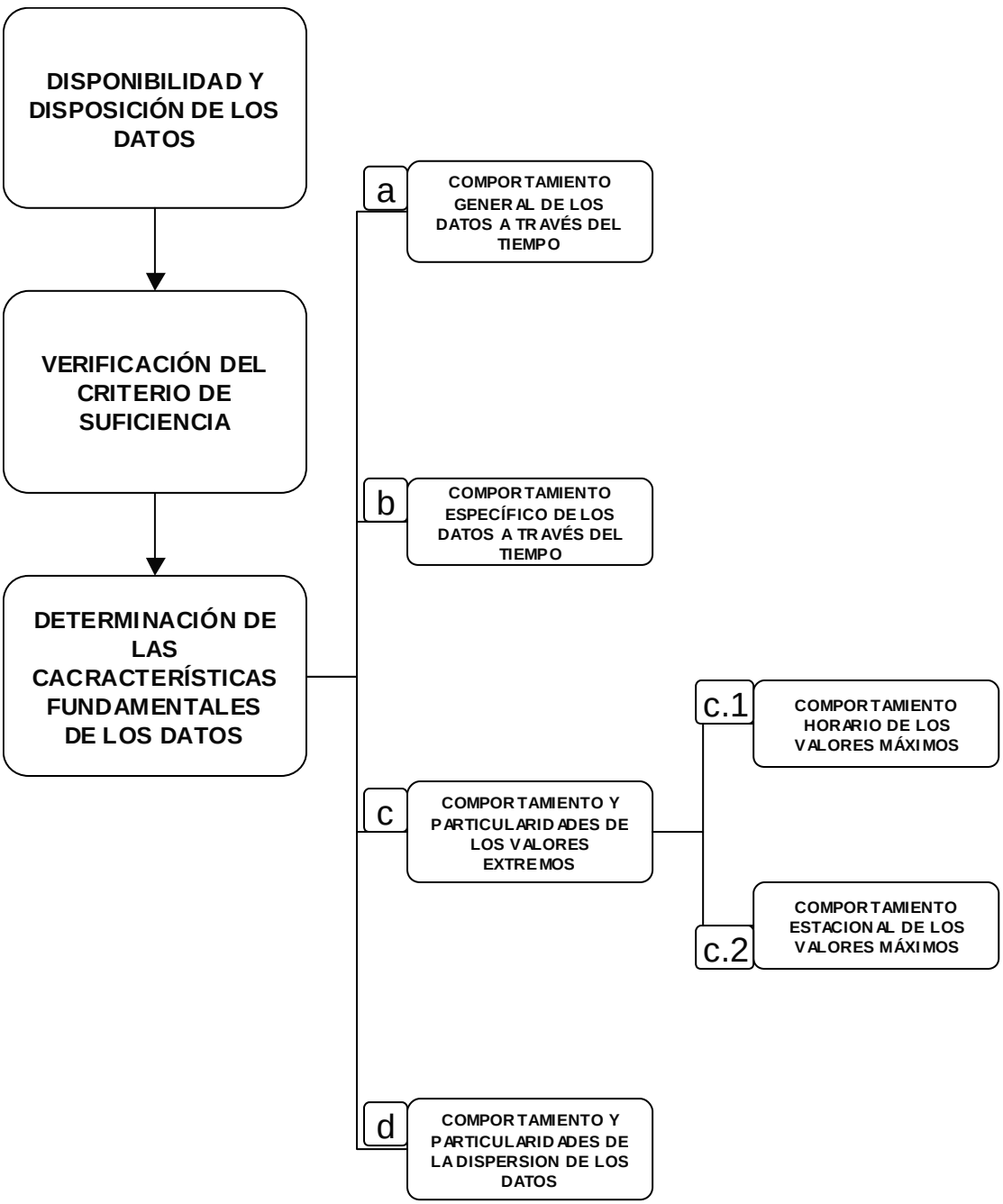
Capítulo 3

Metodología

Como ya se ha explicado en el capítulo 2, la validación de nivel 2 en un conjunto de datos de calidad del aire, se basa en la revisión de las características fundamentales del comportamiento de los datos a través de métodos gráficos y medidas estadísticas. Si se determinan las características fundamentales de un tipo específico de datos, es posible hacer un seguimiento metódico del comportamiento de otros conjuntos de datos del mismo tipo.

Existen diversas herramientas que pueden utilizarse para validar un conjunto de datos, de igual manera, existe una infinidad de formas para hacerlo. La siguiente metodología considera aspectos de los datos tales como disponibilidad, disposición, cantidad, tipo de valores existentes, suficiencia, comportamiento y distribución temporales, valores extremos, características de los valores extremos y dispersión, utilizando las herramientas estadísticas descritas en el capítulo previo; pretendiendo validar los datos de nivel 1 a nivel 2 por medio de la revisión de estas características generales y particulares de su comportamiento.

Figura 3.1 Metodología para la validación de datos de calidad del aire generados por una red de monitoreo automático



3.1 Consideraciones respecto a la disponibilidad y disposición de los datos

De acuerdo con la figura 3.1, en el que se ilustra la metodología a seguir en el proceso de validación de los datos de nivel 1 a nivel 2, se debe iniciar considerando algunos aspectos y características de los datos que son de fundamental importancia, tales como: tipo de datos, disponibilidad, disposición, forma de acceso y cantidad; lo anterior debe ser homogéneo en toda la base de datos, tanto a través del tiempo como entre estaciones de monitoreo.

3.2 Verificación del criterio de suficiencia de datos

De acuerdo con Washington State Department of Ecology, 2004, la suficiencia o criterio de suficiencia de datos es la medición de la cantidad válida de datos con respecto a la cantidad esperada o teórica de estos en condiciones normales, esta debe ser al menos del 75%.
Para verificar la suficiencia de datos se debe calcular el número teórico y el número real de datos arrojados o generados por los sistemas de medición, y que sean válidos para el nivel 1 de validación.

3.2.1 Cálculo de la cantidad teórica de datos

El tamaño de la base de datos o cantidad teórica de datos, se debe determinar considerando que los datos disponibles en las bases públicas son promedios horarios y teóricamente se tienen 24 datos diarios de ozono, aunque técnicamente no sea así todo el tiempo. Por otra parte, debe tomarse en cuenta que no todos los meses ni todos los años, contienen el mismo número de días; lo cual conducirá a un cálculo correcto del tamaño teórico de la base de datos.

3.2.2 Cálculo de la cantidad real de datos

Para calcular la cantidad real de datos, se deben identificar los diferentes tipos de valores existentes en la base de datos y verificar que solamente existan tales tipos de valores. De manera general, solo deben existir valores de texto y valores numéricos, los cuales se deben identificar y localizar en los archivos que conforman la base de datos, verificando que su distribución sea homogénea.

En la figura 3.2 se muestra el procedimiento a seguir para realizar la verificación del criterio de suficiencia.

FIGURA 3.2 Verificación del criterio de suficiencia

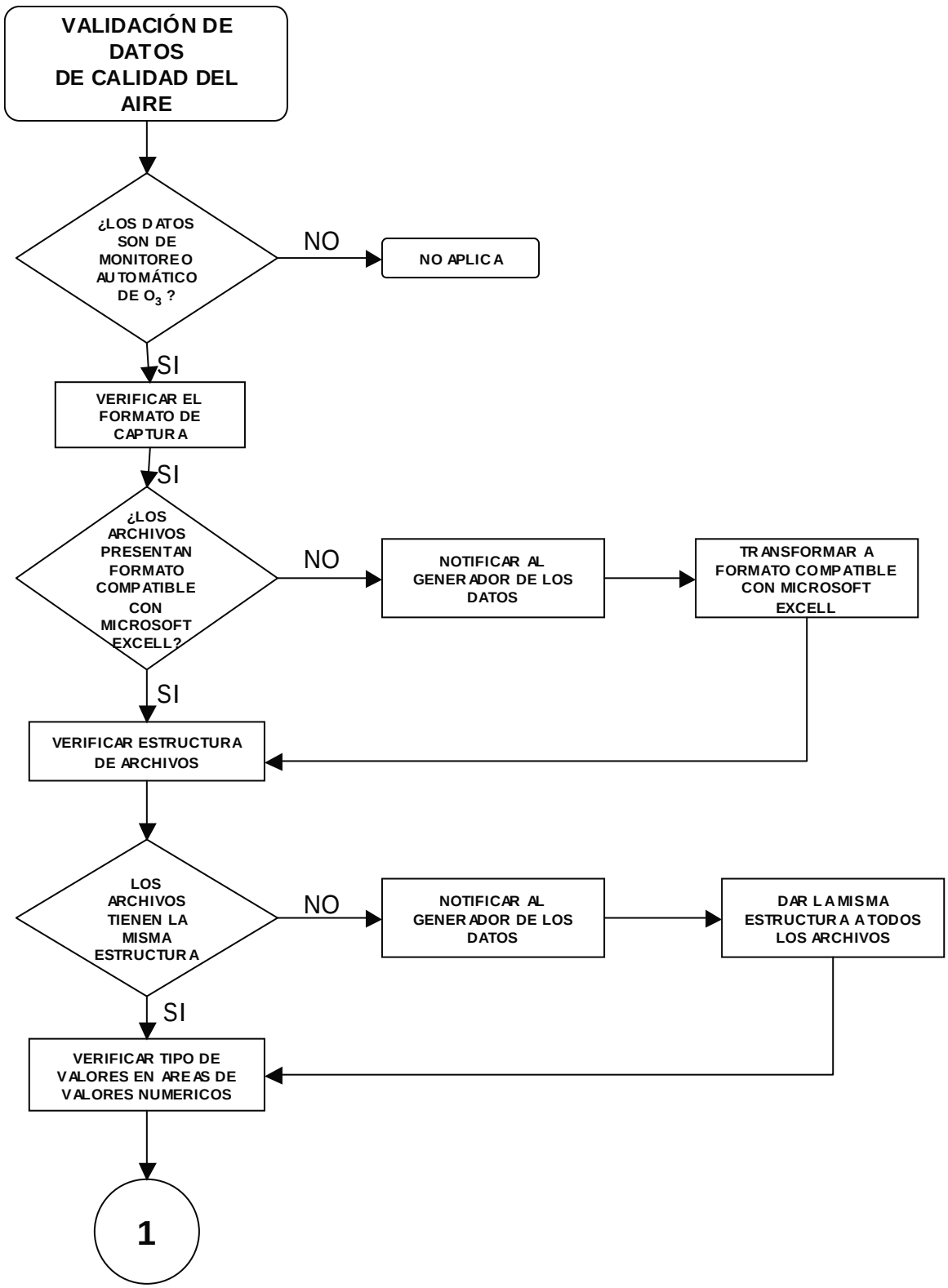
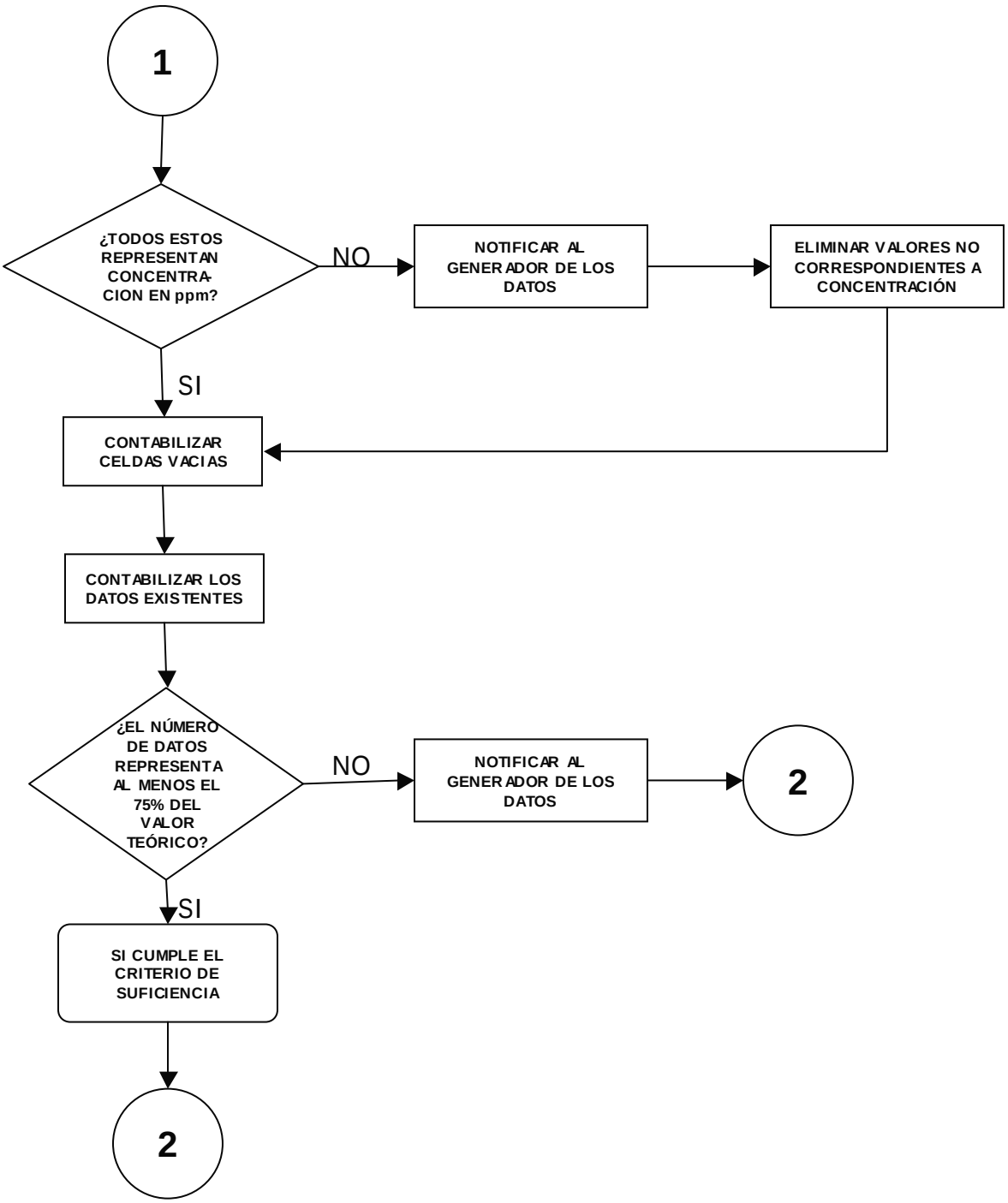


FIGURA 3.2 Verificación del criterio de suficiencia (Continúa)



3.3 Determinación de las características fundamentales de los datos

Según se vio en el capítulo 2, las características fundamentales de los datos son aquellos comportamientos que al hacerlos evidentes, se pueden buscar de manera sistemática en otros conjuntos de datos del mismo tipo. En el mismo capítulo, sección 2.1, se menciona como es el comportamiento característico del ozono en la ciudad de México; el cual se buscará y revisará de manera sistemática como a continuación se indica (figura 3.1):

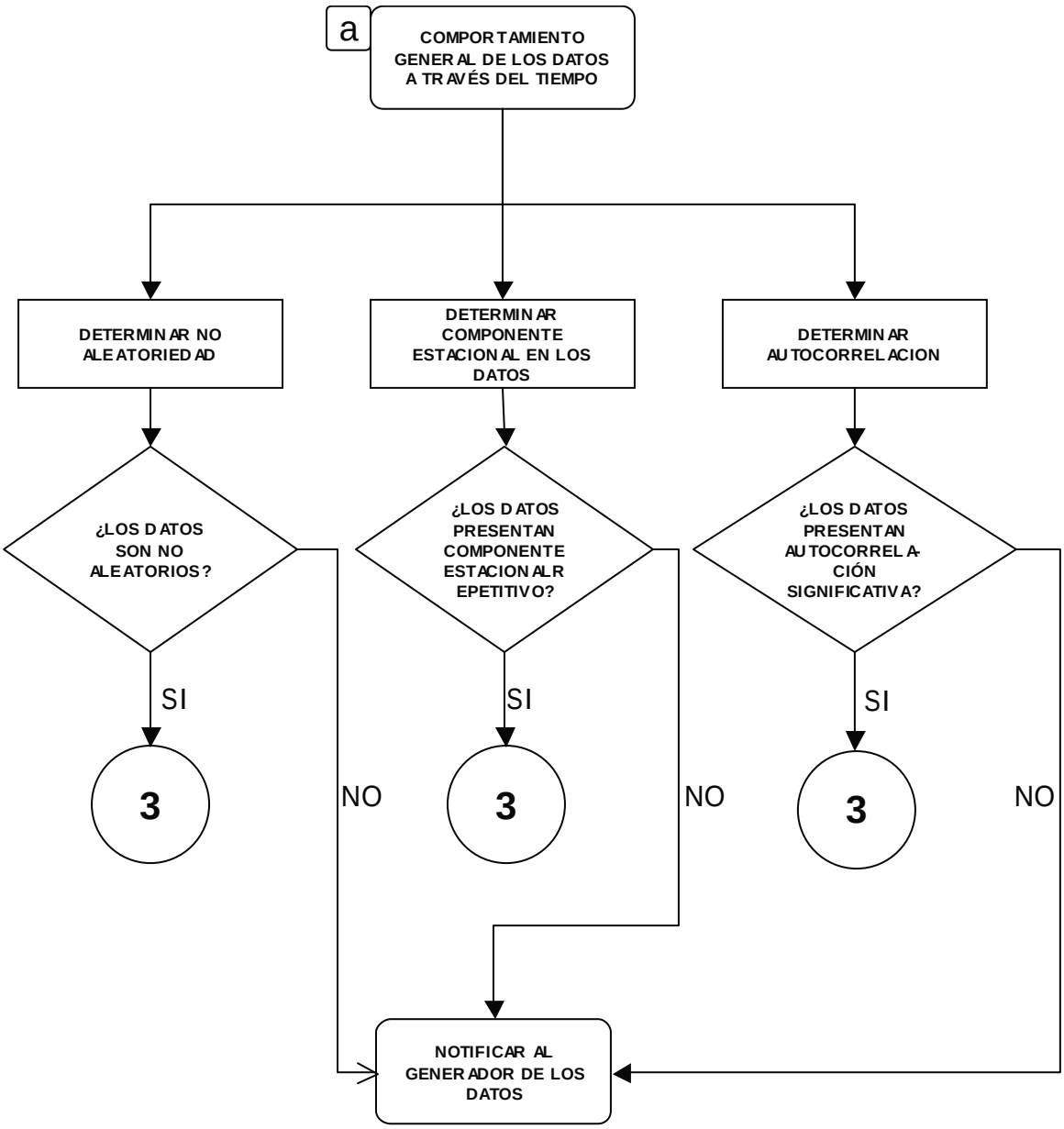
- Comportamiento general de los datos a través del tiempo
- Comportamiento específico de los datos a través del tiempo
- Comportamiento y particularidades de los valores extremos
- Comportamiento y particularidades de la dispersión de los datos

3.3.1 Comportamiento general de los datos a través del tiempo

Desde el enfoque de la Función de Autocorrelación (correlograma) se pueden obtener características del comportamiento de los datos en intervalos amplios de tiempo y se puede saber si un conjunto de ellos presenta autocorrelación, es decir, si el comportamiento de los datos más recientes depende de los datos pasados, o de otra manera, si presentan un comportamiento azaroso. Con este tipo de información, como se menciona en el capítulo anterior, se evalúa la existencia de componentes en la serie de datos tales como tendencia, ciclicidad y estacionalidad; además, se investiga si la variable en estudio es aleatoria o no aleatoria.

Los periodos de datos a utilizar serán aproximadamente anuales, es decir, cercanos a doce meses y con valores del promedio horario mensual de concentraciones. Es importante mencionar que al utilizar valores promedio, las curvas se suavizan pero siguen representando el comportamiento esencial de los datos.
En la figura 3.3 se indica la secuencia a seguir para la determinación del comportamiento general de los datos a través del tiempo.

Figura 3.3 Comportamiento general de los datos a través del tiempo



3.3.2 Comportamiento específico de los datos a través del tiempo

Como se ha mencionado, el ozono presenta un comportamiento característico diario en la Ciudad de México, el cual puede ser observado a través de gráficos de series de tiempo para la concentración horaria. También se ha mencionado que los datos disponibles en las bases de datos públicas son de tipo horario y teóricamente se tienen 24 datos diarios de ozono; esto ofrece una gran cantidad de ellos, por lo que es necesario tomar diferentes periodos de estos para lograr mayor detalle y comparación durante el proceso de revisión, ya que el objetivo principal de un gráfico es evidenciar el comportamiento de todos o de una porción de los datos.

De acuerdo con lo anterior, para determinar el comportamiento básico de los datos a través del tiempo, se construirán series de tiempo con datos de las concentraciones horarias en porciones de 24 h, para pasar a semanas completas y contrastar con semanas completas de diferentes meses del periodo de diez años (figura 3.4).

3.3.3 Comportamiento y particularidades de los valores extremos

Como se apuntó en la sección 2.1, existen valores extremos en el comportamiento diario de los datos, los cuales corresponden a máximos de concentración que se presentan aproximadamente cada 24 hrs.
Con el fin de evaluar el comportamiento de los valores máximos de manera particular, y como se indica en el figura 3.4, se revisarán (c.1) el comportamiento horario de los valores máximos y (c.2) el comportamiento estacional de los valores máximos.

3.3.3.1 Comportamiento horario de los valores máximos

Para realizar la evaluación del comportamiento de los valores máximos diarios, los gráficos de series de tiempo e histogramas son adecuados para realizar observaciones tales como:

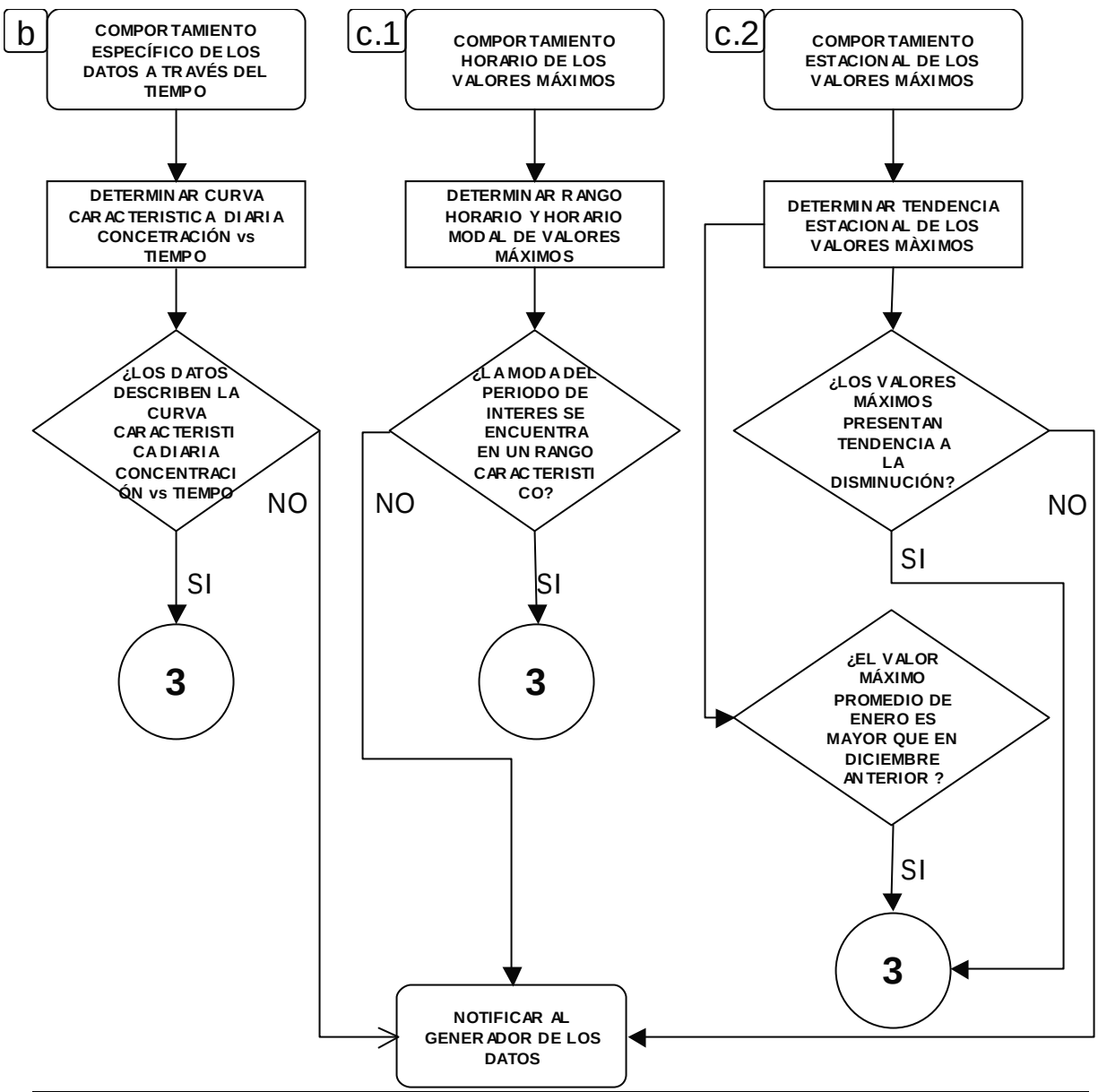
- Rango horario en que se presentan los máximos de concentración
- Distribución de los valores máximos respecto a la hora del día
- Horario modal de máximos
- Eventos poco frecuentes

que son algunas de las observaciones y resultados que se citan en el capítulo 2.

3.3.3.2 Comportamiento estacional de los valores máximos

Según se vio en 2.1, las concentraciones de ozono en la Ciudad de México han presentado tendencia al descenso desde la década de los 90´s, esto referido a los promedios horarios máximos (RAMA, SMA; 2004). Así, la construcción de un correlograma de los promedios horarios máximos, proporcionará información acerca del comportamiento estacional, de tendencia y la existencia de autocorrelación de los datos; información que puede complementarse con la construcción de series de tiempo que particularizan dichos comportamientos, (figura 3.4).

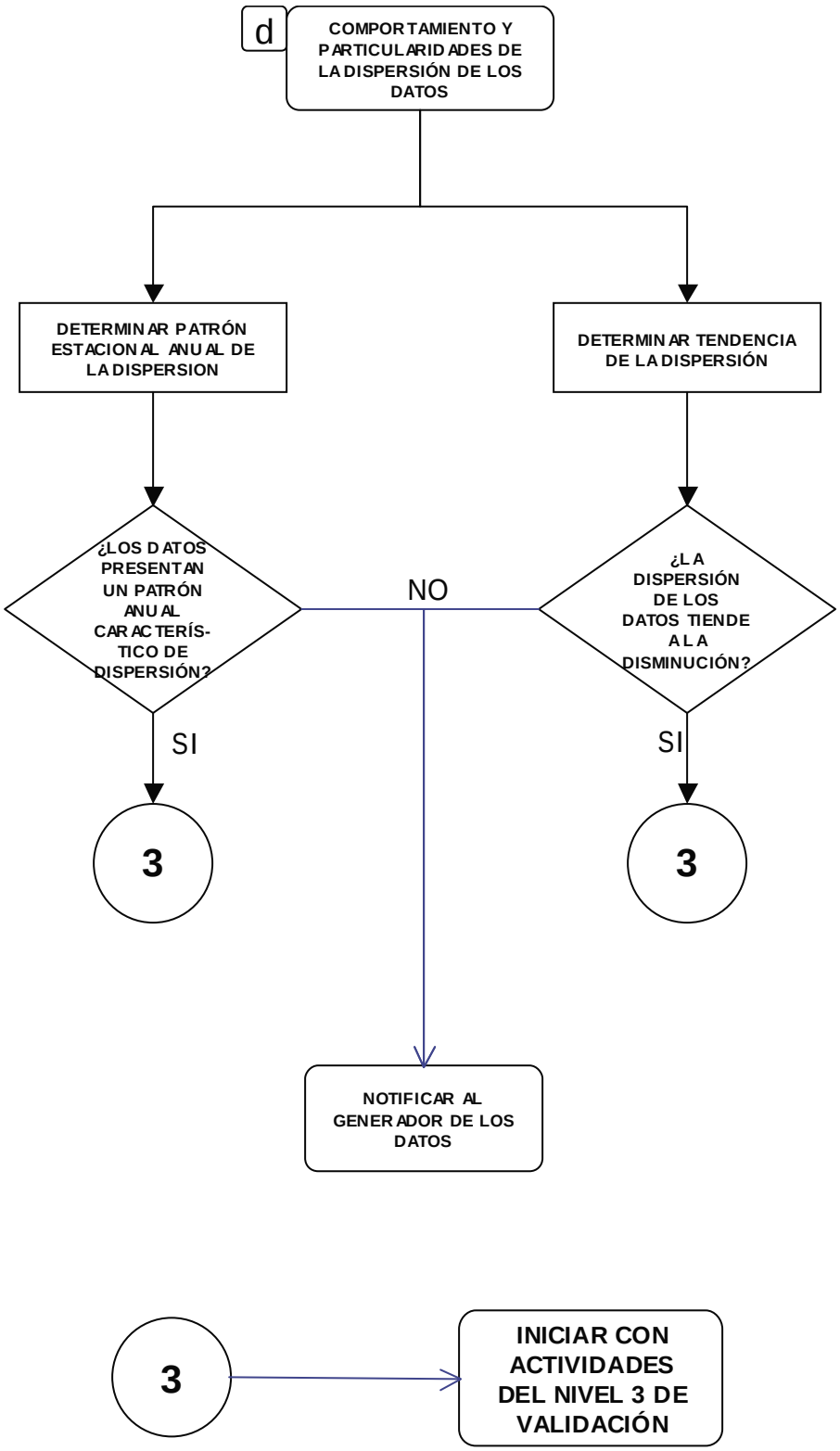
Figura 3.4 Comportamiento específico de los datos a través del tiempo y Comportamiento de los valores máximos



3.3.4 Comportamiento y particularidades de la dispersión de los datos

Derivado de la tendencia al descenso en las concentraciones de ozono desde la década de los 90´s (RAMA, SMA; 2004), se puede pensar como consecuencia en la disminución del rango de concentraciones y disminución en la dispersión de los datos a través del tiempo.
Con la construcción de un gráfico del rango mensual de concentraciones para el periodo de 10 años, se deberá evidenciar dicha tendencia al descenso y debido a que las concentraciones de ozono dependen en gran medida de la radiación solar y la temperatura ambiente (entre otros parámetros), deberá observarse un cierto patrón en el comportamiento de la dispersión de los datos respecto al cambio de las estaciones climáticas. Esto último se puede analizar a través del coeficiente de variación de los datos obtenidos por medio de la representación gráfica de éste a través del tiempo (figura 3.5).

Figura 3.5. Comportamiento y particularidades de la dispersión de los datos



3.4 Aplicación de los datos validados

Teniendo en cuenta los conceptos descritos en el capítulo 2, las pruebas de bondad de ajuste pueden ser una herramienta muy útil antes de iniciar con alguna actividad de análisis e interpretación de los datos. Dependiendo de los objetivos que se persigan, puede iniciarse averiguando el tipo de distribución estadística que mejor describe al conjunto de datos, o de otra forma, saber que los datos no siguen alguna distribución específica.

Los datos de monitoreo automático generados por la RAMA son datos de los que se espera no provengan de una sola población, en realidad se espera que provengan de 24 poblaciones distintas, pues cada una de estas tiene características inherentes al horario de muestreo, principalmente:

- Radiación solar
- Temperatura ambiental
- Humedad ambiental
- Actividades antropogénicas

Tales características prácticamente son únicas para cada hora del día, por lo que cada horario de monitoreo se puede tomar como una variable independiente, lo que de alguna manera da como resultado 24 variables independientes a tratar.

Para ejemplificar lo anterior, es adecuado aplicar la prueba de bondad de ajuste de Kolmogorov-Smirnov (prueba de normalidad) a datos de horarios consecutivos y a datos de horarios específicos del periodo de 10 años.

Capítulo 4

Aplicación de la metodología

En el presente capítulo se muestran el desarrollo y los resultados obtenidos de la aplicación de la metodología propuesta, utilizando las herramientas ya descritas y puntualizando las observaciones o resultados obtenidos de todas ellas.

4.1 Consideraciones respecto a la disponibilidad y disposición de los datos

Según se menciona en la sección 3.1, se deben considerar aspectos que caracterizan a un conjunto de datos como: su tipo, disponibilidad, disposición, cómo se accede a ellos y la cantidad.

Como ejemplo de la metodología propuestas se utilizaron datos de ozono generados por RAMA en el periodo enero de 1996 a diciembre de 2005, los cuales están disponibles en las bases públicas localizadas en la página electrónica de la misma Secretaría: www.sma.df.gob.mx. La RAMA realiza el monitoreo automático continuo de O_3 durante las 24 h del día los 365 días del año, cuenta con 16 estaciones de monitoreo en el área metropolitana de la Ciudad de México en las que se generan o generaron datos de ozono; en la tabla 4.1 se muestra cuáles son estas 16 estaciones de monitoreo y la clave de identificación de cada una.

Cabe señalar, que se utilizaron datos de las 16 estaciones de monitoreo citadas en la tabla 4.1, hasta completar la verificación del criterio de suficiencia; a partir de ese punto se utilizaron datos de la estación de monitoreo automático Merced, una de las estaciones que mejor cumplen con el criterio de suficiencia y además, es una de las estaciones con mejor funcionamiento histórico de acuerdo con la RAMA.

Tabla 4.1. Estaciones de monitoreo que generan o generaron datos de ozono en la Ciudad de México en el periodo 1986 - 2005

Clave	Estación	Inicio	Clave	Estación	Inicio
AZC	Azcapotzalco	1986	BJU	Benito Juárez	1992
CES	Cerro de la Estrella	1986	TAX	Taxqueña	1992
HAN	Hangares	1986	CUA	Cuajimalpa	1993
MER	Merced	1986	TAC	Tacuba	1993
PED	Pedregal	1986	TAH	Tláhuac	1993
PLA	Plateros	1986	TPN	Tlalpan	1993
UIZ	UAM-I	1987	SUR	Santa Úrsula	2000
LAG	Lagunilla	1990	COY	Coyoacán	2005

Las bases de datos se consultaron tal y como las publicó la SMA durante el año 2005 y 2006, encontrándose organizadas en carpetas anuales, como archivos comprimidos compatibles con ambiente “Windows”. En cada carpeta existen 12 subcarpetas también comprimidas y cada una correspondiente a uno de los meses del año. Dentro de las subcarpetas se encuentran archivos en lenguaje d-base de cada uno de los contaminantes atmosféricos que se monitorean, los cuales se visualizaron y manipularon directamente con el software Microsoft Excel. Los archivos DBF visualizados por medio de Microsoft Excel presentan la configuración mostrada en la tabla 4.2, en donde aparece solo un fragmento del año 2000 para ozono de la estación Merced.

Tabla 4.2.Visualización de archivos DBF en MS Excel

	A	B	C	D	E	F	-----	-----	AM	AN
1	FECHA	HORA	MO3LAG	LAG	MO3TAC	TAC	-----	-----	MO3TAH	TAH
2	01/01/2000	1	0.004		0.008		-----	-----	0.009	
3	01/01/2000	2	0.004		0.007		-----	-----	0.013	
4	01/01/2000	3	0.005		0.006		-----	-----	0.010	
5	01/01/2000	4	0.005		0.005		-----	-----	0.013	
6	01/01/2000	5	0.007		0.006		-----	-----	0.009	
7	01/01/2000	6	0.007		0.005		-----	-----	0.005	
8	01/01/2000	7	0.005		0.004		-----	-----	0.004	
9	01/01/2000	8	0.008		0.005		-----	-----	0.012	
10	-----	-----	0.009		0.006		-----	-----	0.008	
11	-----	-----	-----		0.013		-----	-----	0.021	

En la tabla 4.2, la primera columna indica la fecha de monitoreo desde el primero hasta el último día de mes. La segunda columna indica la hora diaria de monitoreo y en las siguientes columnas la concentración del contaminante en ppm para cada una de las estaciones o sitios de monitoreo. Tal configuración se cumple para todos y cada uno de los archivos incluidos en la base de datos, lo cual se verificó

durante el cálculo de las cantidades teórica y real de datos como se muestra en la siguiente sección.

4.2 Verificación del criterio de suficiencia de datos

Tomando en cuenta la información descrita en la sección 3.2, para verificar el criterio de suficiencia se procedió a calcular el número teórico y el número real de datos disponibles debiendo representar este último, al menos el 75% del número teórico.

4.2.1 Cálculo de la cantidad teórica de datos

Considerando que la cantidad teórica de datos que se generan al día es de 24, en la tabla 4.3 se indica la cantidad de datos horarios de ozono que se deben generar por mes y año en cada una de las estaciones de monitoreo automático de la SMA; lo cual resultó de multiplicar el número de días por 24 datos diarios.

Tabla 4.3. Cantidad teórica de datos de ozono por mes y año que debe generar una estación de monitoreo

Días por mes	No. Datos Horarios
28	672
29	696
30	720
31	744

Días por año	No. Datos Horarios
365	8760
366	8784

4.2.2 Cálculo de la cantidad real de datos

De acuerdo con la configuración de los archivos mostrada en la tabla 4.2, se identificaron tres tipos de valores en las celdas: numéricos, de texto y fechas. Una vez identificados los tipos de valores y su distribución en los archivos, se procedió a revisar cada archivo verificando la correcta distribución y contenido como se describe a continuación.

4.2.2.1 Valores de texto

Como se observa en la tabla 4.2, los valores de texto se localizan en la primer fila de cada archivo y corresponde a la identificación de las estaciones de monitoreo. Ha sido muy importante verificar que la estructura de todos los archivos de la base

de datos es homogénea, esto es, para cada uno de los meses de cada año los archivos cuentan exactamente con la misma construcción mostrada en la tabla 4.2. En el caso de haber encontrado archivos con una construcción diferente (lo cual no ocurrió), estos no se habrían tomado en cuenta en el resto del proceso de validación, siendo necesario notificarlo al responsable de la generación de los datos, para verificar y corregir el error.

Con la lectura de los encabezados de columnas de cada uno de los archivos, se verificó la homogeneidad en la estructura de estos; lo anterior se realizó construyendo un archivo de lectura de los encabezados de cada archivo.
En la tabla 4.4 se muestra un fragmento del archivo de lectura con el cual se verificó que todos los archivos de ozono que constituyen la base de datos en el periodo 1996 a 2005, efectivamente tienen la misma construcción y las celdas correspondientes a valores de texto contienen solamente texto.

Tabla 4.4. Verificación de encabezados de columnas (fragmento)

	A	B	C	D	E	F	G	H	I
1	Ene-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
2	Feb-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
3	Mar-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
4	Abr-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
5	May-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
6	Jun-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
7	Jul-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
8	Ago-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
9	Sep-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
10	Oct-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
11	Nov-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
12	Dic-96	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
13	Ene-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
14	Feb-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
15	Mar-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
16	Abr-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
17	May-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG
18	Jun-97	MO3LAG	LAG	MO3TAC	TAC	MO3EAC	EAC	MO3SAG	SAG

4.2.2.2 Valores numéricos

Las celdas correspondientes a valores numéricos muestran tres tipos de valores:

- Valores de concentración en ppm.
- El valor –99.999.
- El valor 0.000

Un valor –99.999 es la manera en que la RAMA identifica la no disponibilidad de un dato, o dato faltante. El 0.000 es un valor no disponible que no fue identificado, pues en el periodo de tiempo en que se registraron estos datos el limite de detección de los instrumentos analíticos no permitía medir cero de concentración. A estos dos tipos de valores se les llamó datos falsos, los cuales se eliminaron de la base de datos por la siguiente razón: por que al no corresponder a datos de concentración, se producirían resultados erróneos o no lógicos al aplicar las técnicas gráficas y estadísticas para su análisis.

Con la eliminación de los datos falsos, se obtuvo el tamaño real de la base de datos para cada uno de los meses del periodo completo de 10 años y para las 16 estaciones. Estos resultados no se muestran por tratarse de 10 archivos muy extensos, sin embargo, los resultados se utilizaron para verificar el criterio de suficiencia, los cuales se muestran en la tabla 4.5.

En la tabla 4.5 se observa que solo la estación TPN no cumple con el criterio de suficiencia de datos en el periodo 1996-2005, pues 3 de 10 años no lo cumplen. Estos resultados son relativos, pues el criterio de suficiencia toma sentido según el periodo de tiempo que se selecciona; esto es, si se escoge el periodo 1997-2003 entonces la estación TPN si cumple el criterio de suficiencia al 100%. También se observa, que LAG y MER si cumplen con el criterio de suficiencia en el periodo de 10 años y para cada uno de esos 10 años, pero de manera mensual, cada estación tiene un mes en el que no cumple, LAG en enero de 2004 y MER en diciembre del mismo año. Es necesario recordar que el criterio de suficiencia es relativo, pues solo estableciendo el periodo de tiempo a validar se puede decidir si se cumple o no.

Tabla 4.5. Numero de meses y años en que no se cumple el criterio de suficiencia en el periodo 1996-2005

		LAG	TAC	AZC	MER	PED	CES	PLA	HAN	UIZ	BJU	TAX	CUA	TPN	TAH
1996	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	5	6	3	10	10	10
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	58	50	75	17	17	17
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	NO	NO	NO	NO	NO	NO
1997	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
1998	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
1999	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
2000	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
2001	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
2002	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
2003	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
2004	MESES CON INSUFICIENCIA	1	0	0	1	0	0	0	0	1	0	1	0	12	0
	P0RCENTAJE ANUAL	92	100	100	92	100	100	100	100	92	100	92	100	0	100
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	NO	SI
2005	MESES CON INSUFICIENCIA	0	0	0	0	0	0	0	2	1	12	0	0	7	1
	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	83	92	0	100	100	42	92
	CUMPLE ANUALMENTE	SI	SI	SI	SI	SI	SI	SI	SI	SI	NO	SI	SI	NO	SI
1996	AÑOS CON INSUFICIENCIA	0	0	0	0	0	0	0	0	1	2	1	1	3	1
A	P0RCENTAJE ANUAL	100	100	100	100	100	100	100	100	90	80	90	90	70	90
2005	CUMPLE	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	NO	SI

4.3 Determinación de las características fundamentales de los datos

Según se vio en el capítulo 2, las características fundamentales de los datos son aquellos comportamientos que al hacerlos evidentes, se pueden buscar de manera sistemática en otros conjuntos de datos del mismo tipo. También se menciona como es el comportamiento característico del ozono en la ciudad de México; este comportamiento se buscó y revisó de manera sistemática tal como se indica en la sección 3.3.

4.3.1 Comportamiento general de los datos a través del tiempo

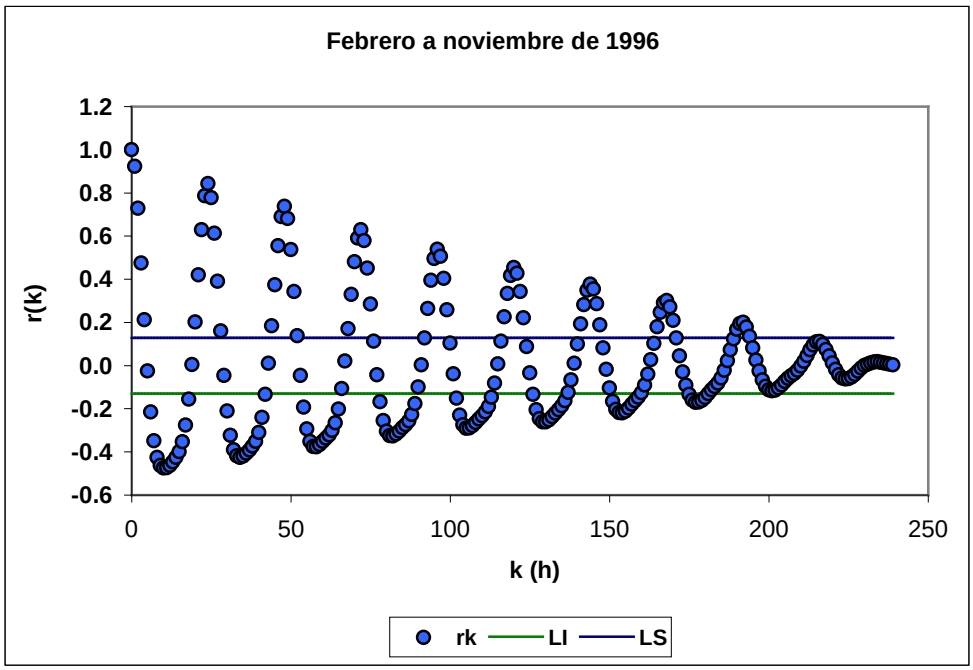
Como se menciona en 3.3, a través de la función de autocorrelación o correlograma, se pueden obtener características generales del comportamiento de los datos en el tiempo y se puede saber si un conjunto de datos presenta correlación con ellos mismos; así, para la construcción del correlograma, se utilizaron los valores del promedio horario mensual de concentraciones mostrados en la tabla 4.6.

Para un periodo inicial de 10 meses, se construyó el correlograma de acuerdo con las ecuaciones 2.3, 2.4 y 2.5.

Tabla 4.6. Promedio horario mensual, estación Merced

h	Ene-96	Feb-96	Mar-96	----------	Oct-05	Nov-05	Dic-05
1	0.007	0.007	0.007	----------	0.007	0.004	0.012
2	0.007	0.007	0.007	----------	0.007	0.004	0.012
3	0.007	0.007	0.007	----------	0.008	0.004	0.012
4	0.007	0.007	0.008	----------	0.007	0.004	0.013
5	0.006	0.006	0.007	----------	0.005	0.003	0.013
6	0.006	0.006	0.006	----------	0.004	0.003	0.013
7	0.011	0.012	0.012	----------	0.004	0.004	0.012
8	0.013	0.013	0.012	----------	0.004	0.004	0.012
9	0.014	0.015	0.013	----------	0.007	0.007	0.013
.....	....	....	....	....	....	....	....
14	0.119	0.106	0.097	----------	0.066	0.070	0.078
15	0.104	0.108	0.087	----------	0.059	0.072	0.087
16	0.100	0.092	0.075	----------	0.048	0.068	0.081
.....	....	....	....	....	....	....	....
20	0.016	0.019	0.019	----------	0.008	0.009	0.018
21	0.012	0.014	0.015	----------	0.007	0.005	0.015
22	0.009	0.009	0.009	----------	0.007	0.005	0.013
23	0.008	0.008	0.008	----------	0.006	0.005	0.013
24	0.007	0.007	0.008	----------	0.006	0.004	0.012

Figura 4.1. Correlograma del promedio horario mensual 1996, estación Merced

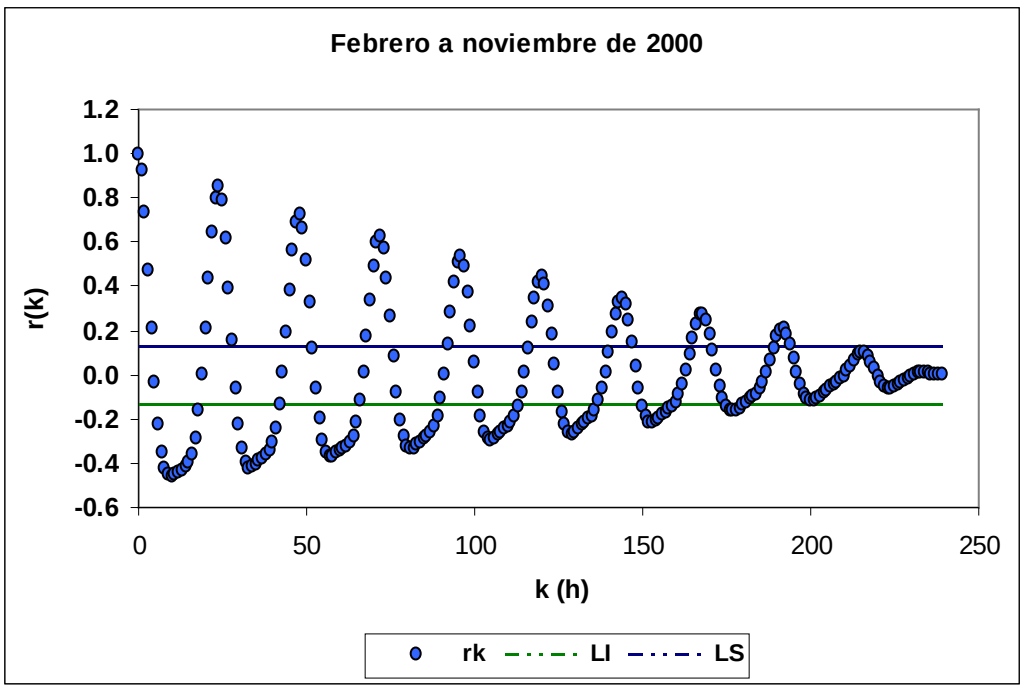


El gráfico 4.1 muestra una variable no aleatoria, debido a la existencia de coeficientes $r_1,r_2,.....,r_k$ que estadísticamente son diferentes a cero. Es razonable haber obtenido información de no aleatoriedad, pues el mismo diseño del monitoreo no lo es, es un monitoreo programado que genera datos en intervalos de tiempo uniformes. Los datos presentaron carácter cíclico, pues se aprecian r(k) máximos a intervalos regulares de 24 desfases a lo largo de toda la serie.
Al presentarse r(k) máximos que sobrepasan los límites LI y LS (Ec. 2.5) a lo largo de toda la serie, se indica que existe autocorrelación o que el comportamiento de datos recientes tiene relación con los datos pasados.

Un segundo correlograma para otro periodo de 10 meses (Figura 4.2), muestra que los datos del mismo periodo pero del año 2000, presentan exactamente el mismo comportamiento que el año anterior: son no aleatorios, tienen carácter cíclico y presentan autocorrelación significativa.

El mismo periodo de meses para el resto de los 10 años, presenta el mismo comportamiento, por lo cual se omite el resto de los correlogramas.

Figura 4.2. Correlograma del promedio horario mensual 2000, estación Merced



De la función de autocorrelación o correlograma, se obtuvieron tres características del comportamiento general de los datos a través del tiempo:

- Datos no aleatorios, pues los coeficientes $r_1, r_2, \ldots, r_k$ son estadísticamente diferentes a cero. El diseño del muestreo para la medición de contaminantes atmosféricos, como el ozono, se realiza de manera sistemática y programada en función del tiempo, lo cual convierte a la determinación o medición en una variable con carácter no aleatorio.

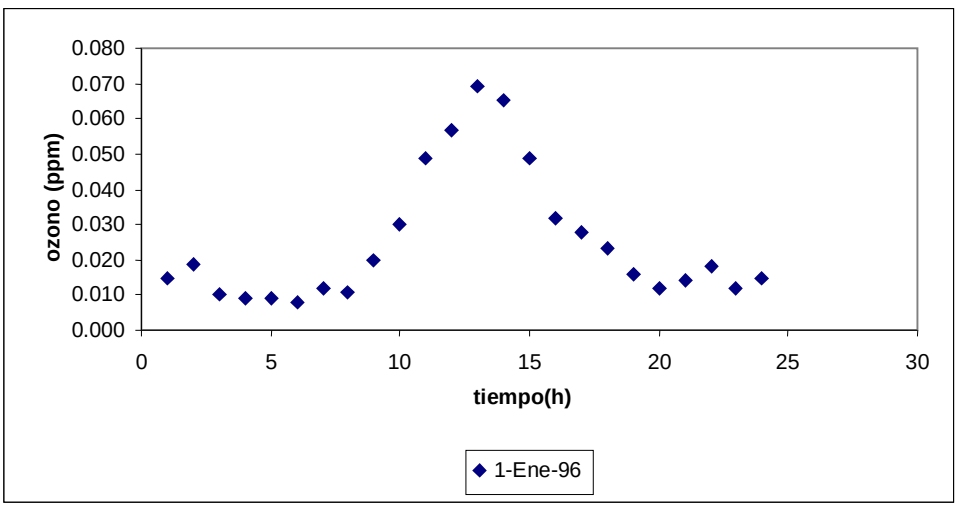
- En las figuras 4.1, 4.2 y otras que se omiten, se observa que cada 24 desfases o puntos en el figura, se presenta un valor máximo perfectamente definido y de manera repetitiva a lo largo de toda la serie, además, la fluctuación de los coeficientes en forma de onda; de esta manera se determinó que la serie de datos tiene un componente cíclico en su comportamiento a través del tiempo.

- Todas las series o periodos de datos presentaron autocorrelación significativa, pues existen máximos r_k que sobrepasan los límites establecidos en la ecuación (2.5) a lo largo de toda la serie, en otras palabras, el comportamiento de datos recientes tiene relación con el comportamiento de los datos pasados.

4.3.2 Comportamiento específico de los datos a través del tiempo

De acuerdo con lo descrito en la sección 3.3.2, se construyó una primera serie de tiempo de 24 h. Así, para las primeras 24 h del año 1996 se obtuvo la serie de tiempo concentración de ozono contra horas (figura 4.3), en donde cabe señalar que los datos utilizados para el análisis en este apartado, son datos consecutivos.

Figura 4.3. Serie de tiempo, 24 h., Estación Merced

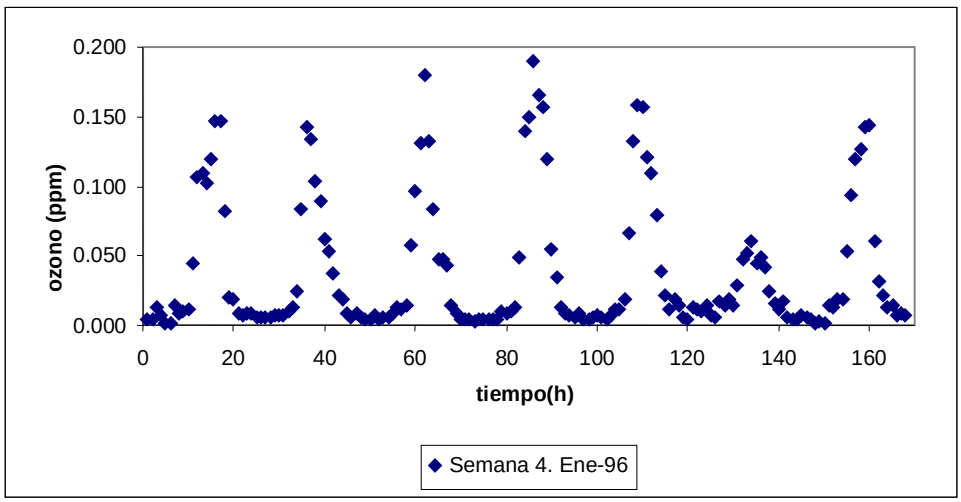


De la figura 4.3 se obtuvieron dos observaciones inmediatas:

- La existencia un máximo en la concentración de ozono que se presentó a las 13:00 h.
- El aumento bien definido de las concentraciones a partir de las 8:00 h. y la disminución a partir de las 13:00 h.

Estas dos observaciones hasta ese momento, solo indicaron que en ese día hubo un máximo de concentración, ascenso y descenso bien definidos; sin embargo, con los datos de un solo día no se puede sacar conclusiones sobre un comportamiento general.
Posteriormente se revisó un periodo mayor de tiempo, 7 días en horas consecutivas, correspondiente a la cuarta semana de enero.

Figura 4.4. Serie de tiempo, 4ª semana de enero estación Merced



En la figura 4.4 se observó lo siguiente:

- Se presenta una serie de 7 máximos, los cuales ocurren aproximadamente cada 24 h.
- Cada uno de los siete máximos corresponde a uno de los siete días de la semana. Los horarios en que se presentaron los máximos se presentan en la tabla 4.6:

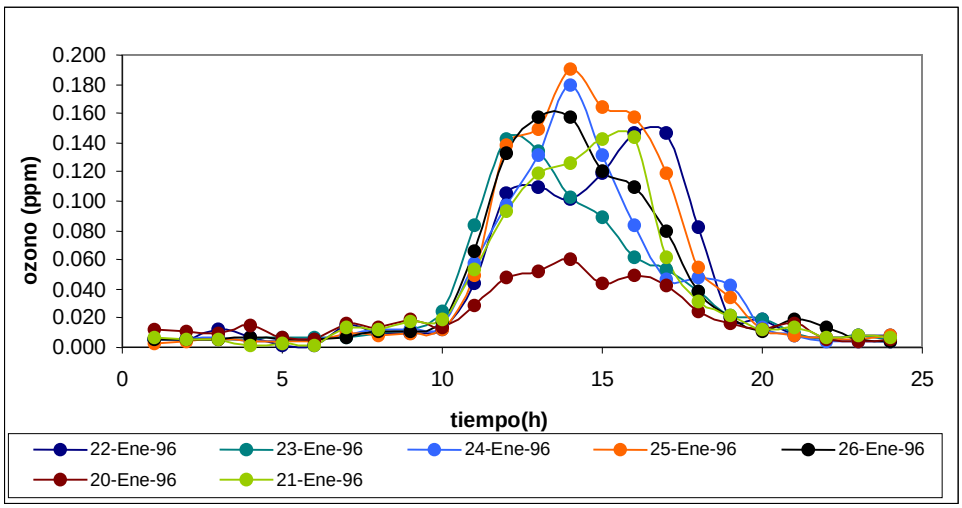
Tabla 4.6. Horario de los máximos de concentración de ozono en la cuarta semana de enero de 1996 en la estación Merced

Fecha	Hora
22-Ene-96	16.00
23-Ene-96	12.00
24-Ene-96	14.00
25-Ene-96	14.00
26-Ene-96	13.00
27-Ene-96	14.00
28-Ene-96	16.00

- Existen ascensos y descensos bien definidos, presentándose aproximadamente cada 24 h.

El periodo anterior de 7 días, se graficó en una escala de tiempo de solo 24 h:

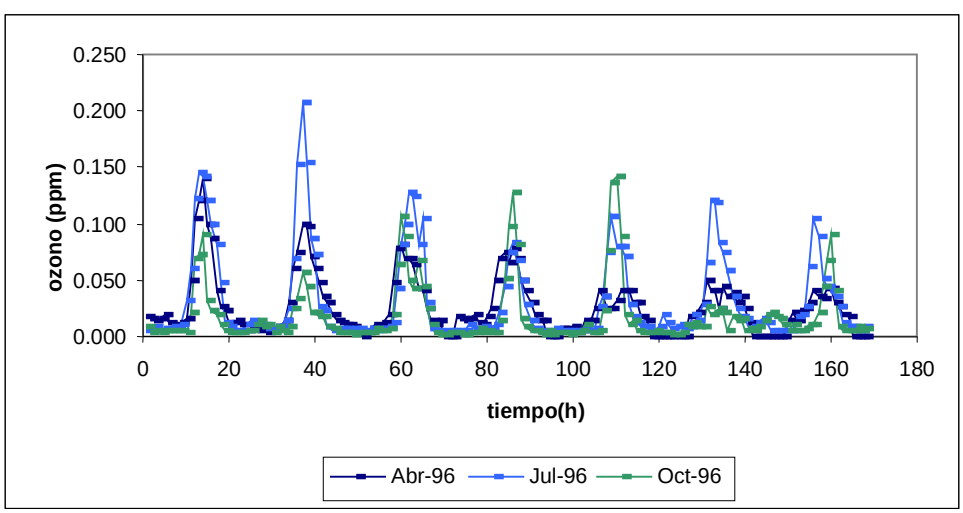
Figura 4.5. Serie de tiempo, 4ª semana de enero, Estación Merced



Con la superposición de los siete días en la figura 4.5, se observa mas claramente la gran similitud en el comportamiento de las concentraciones de ozono diarias y como los horarios de ascenso, descenso y máximo son bastante regulares.

Las observaciones de las figuras 4.3 a 4.5, coinciden con que el comportamiento de la concentración de ozono es un fenómeno que cumple ciclos de 24 h, en donde cada día el valor máximo se encuentra después de las 12:00 h.
Para averiguar el comportamiento el resto del año, se escogieron 3 periodos de tiempo similares; una semana pero de los meses: abril, julio y octubre de 1996.

Figura 4.6. Serie de tiempo, 1ª semana de mes 1996, Estación Merced



En la figura 4.6, se verifica para los tres meses el mismo comportamiento cíclico de 24 h, lo cual sigue coincidiendo con que la concentración de ozono describen un fenómeno que cumple ciclos de 24 h. De la misma manera, se escogieron al azar periodos iguales de tiempo para otros años y el comportamiento fué el mismo, como se observa en los siguientes dos gráficos (marzo, junio y septiembre de 1999; mayo, agosto y noviembre de 2002).

Figura 4.7. Serie de tiempo, 2ª semana de mes 1999, estación Merced

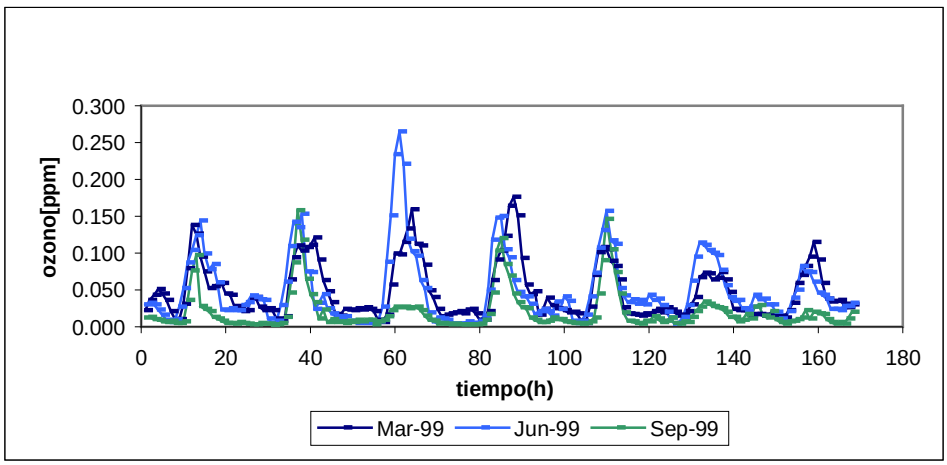
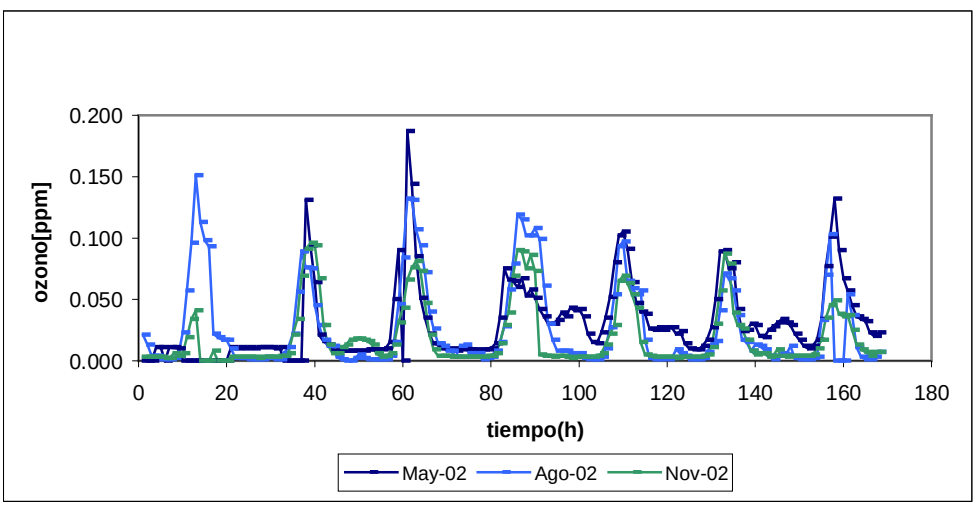


Figura 4.8. Serie de tiempo, 3ª semana de mes 2002, estación Merced



Como se observa en las figuras 4.5 a 4.8, los datos presentan el mismo comportamiento cíclico sin importar semana, mes o año; esto es, cada 24 horas se cumple un ciclo característico y perfectamente bien definido, en donde se registra un valor máximo de concentración de ozono con los correspondientes ascensos y descensos.

Con dos tipos de gráficos se observan dos resultados coincidentes, por una parte la función de autocorrelación indica un componente cíclico en el comportamiento de las concentraciones de ozono, y por otra, los figuras de serie de tiempo detallan este comportamiento cíclico, tal como se concluye en el párrafo anterior.

Es importante recordar que el periodo de tiempo a graficar debe ser tal que permita visualizar claramente la manera en que se comportan los datos, es decir, un periodo muy corto no da información suficiente y un periodo muy grande provoca pérdida de detalles en el figura.

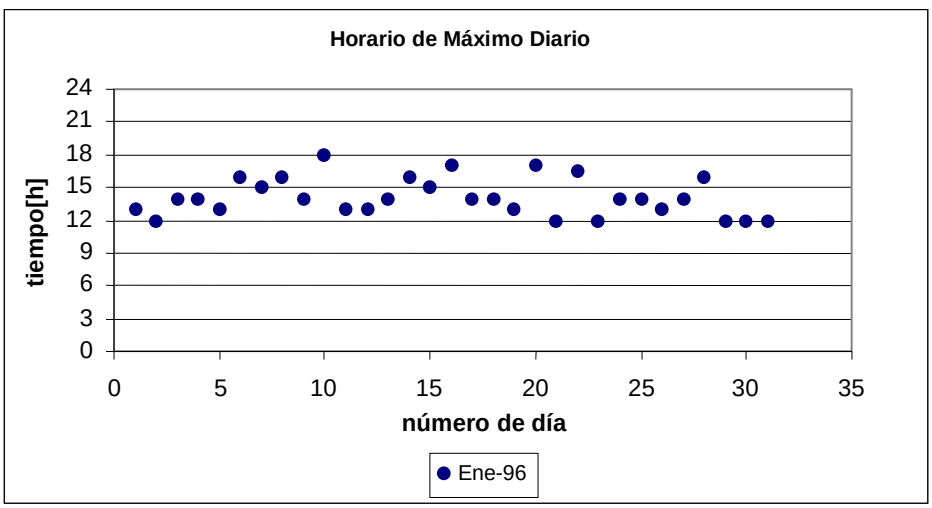
4.3.3 Comportamiento y particularidades de los valores extremos

En la sección 4.3.2, se observó que existen valores extremos de concentraciones de ozono en el comportamiento diario de los datos, estos valores extremos corresponden a los máximos de concentración que se presentan aproximadamente cada 24 h.
Por lo anterior se revisó el comportamiento particular de los valores máximos en cuanto al horario en que se presentan y al comportamiento a mediano plazo.

4.3.3.1 Comportamiento horario de los valores máximos

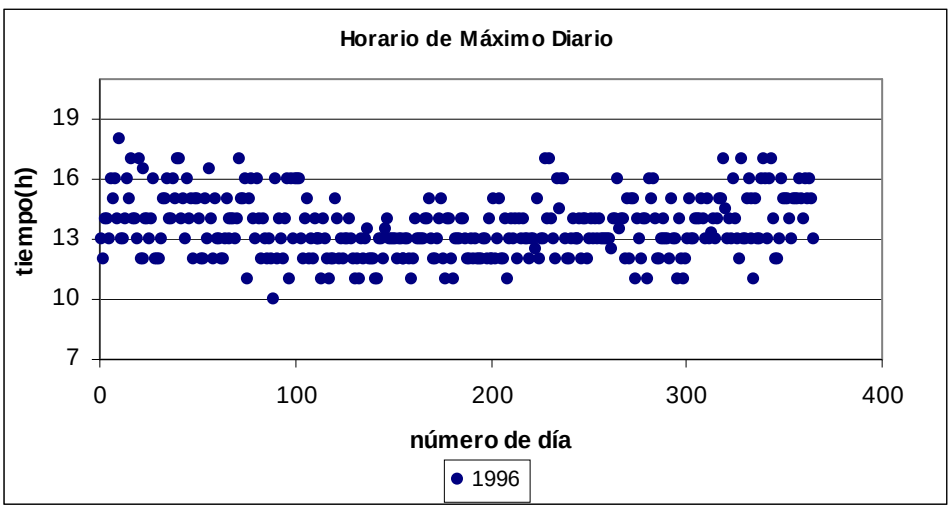
Para verificar el comportamiento horario de los valores máximos, se inició la revisión con una serie de tiempo de valores máximos diarios para el mes de enero de 1996.

Grafico 4.9. Horario de máximo diario, estación Merced.



En la figura 4.9 se observa que los valores máximos se presentan en un horario entre las 12 y las 18 h., tal como se observó en la tabla 4.6. Con las características observadas en las secciones 4.3.1 y 4.3.2, se estableció que el comportamiento general de los datos a través del tiempo cumple ciclos de 24 h, con lo que se pudo inferir que el horario de los máximos de concentración diaria se presenta dentro de un rango horario más o menos definido. Así, se realizó una revisión de todo el año 1996 con la construcción de la serie de tiempo representada en la figura 4.10, en donde se visualizan tres características más:

Figura 4.10. Horario de máximo diario de 1996, estación Merced.



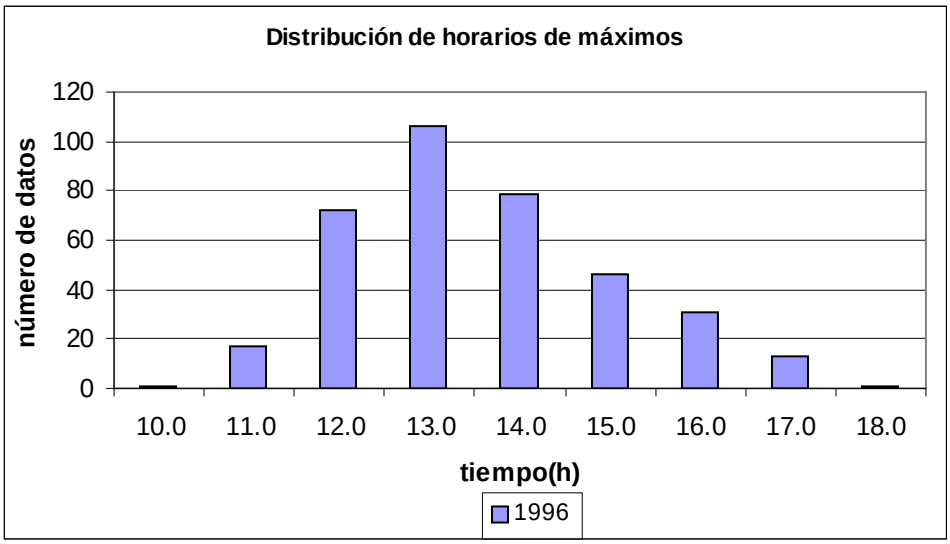
- El rango de horas en que se presentan los máximos de concentración es de 10 a 18 h
- La mayor cantidad de datos se encuentra entre las 12 y 15 h
- Como eventos poco comunes, se localizan datos en las 10, 17 y 18 h

Las tres características anteriores se visualizan claramente utilizando un histograma.

Tabla 4.7. Frecuencias de horario máximo diario de 1996, estación Merced.

		Clase	*# de datos*	*%*
Máximo	18.0	10.0	1	0.3
Mínimo	10.0	11.0	17	4.6
Rango	8.0	12.0	72	19.7
Clases	8.0	13.0	106	29.0
Ancho Clase	1.0	14.0	79	21.6
		15.0	46	12.6
		16.0	31	8.5
		17.0	13	3.6
		18.0	1	0.3

Figura 4.11. Horario de máximo diario de 1996, estación Merced.



Con la construcción del gráfico anterior (figura 4.11), se ven de manera clara las características del comportamiento de los datos y son las siguientes:

- El rango horario de los valores máximos es de 10 a 18 h
- La mayor cantidad de estos se encuentra entre las 12 y 15 h (82.8 %); (303 de 366 datos), siendo las 13 h. el horario modal con 106 datos
- Los eventos poco frecuentes son los siguientes (hr.,frecuencia): (10,1); (11,17); (17,13) y (18,1); que representan el 8.7 %

Para los siguientes 9 años el comportamiento de los valores máximos es el mostrado en la tabla 4.8.

Tabla 4.8. Frecuencias de horario máximo diario, 1996 a 2005, estación Merced.

Clase	# de datos									
	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
10	1	0	0	0	0	0	0	1	0	1
11	17	14	12	3	3	4	8	6	3	4
12	72	60	50	57	39	29	33	37	41	36
13	106	109	119	102	106	107	91	81	86	81
14	79	94	83	85	92	103	107	89	109	94
15	46	47	61	55	71	60	61	74	68	69
16	31	26	23	34	37	35	38	41	36	53
17	13	13	14	20	9	20	15	17	14	19
18	1	0	0	1	4	0	8	4	4	4

Frecuencia Máxima.

En la tabla 4.8 se observan dos comportamientos; el primero, de 1996 a 2001 en donde el horario modal de máximos es las 13:00 h; el segundo, de 2002 a 2005 en donde el horario modal es las 14:00 h. Sería interesante estudiar y analizar esta información, lo cual es una actividad propia para el nivel 3 de validación.

Con la construcción de los siguientes dos histogramas y de acuerdo con la información que se presenta en la tabla 4.8, se detalla la distribución de horarios máximos para los dos siguientes periodos: 1997 a 2001 y 2002 a 2005, que junto con los datos graficados en la figura 4.11 se complementa el periodo de 10 años.

Figura 4.12. Horario de máximo diario de 1997 a 2001, Estación Merced.

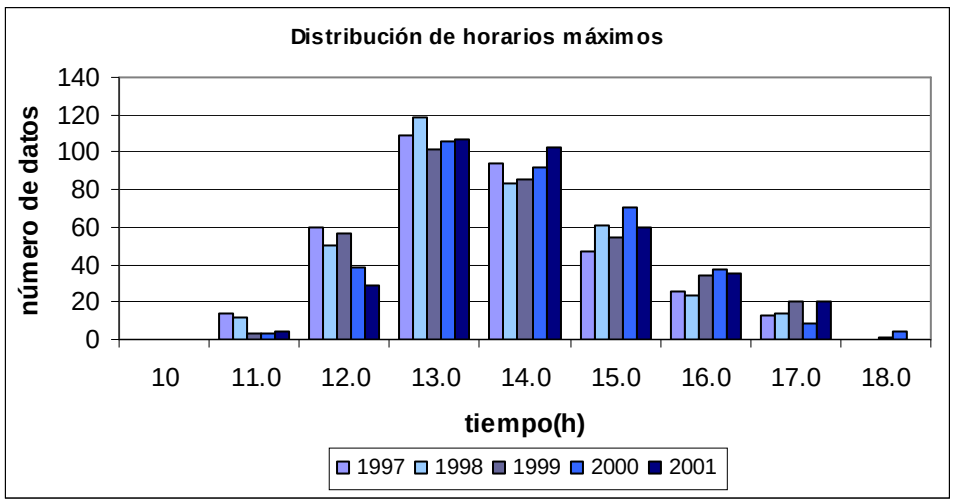
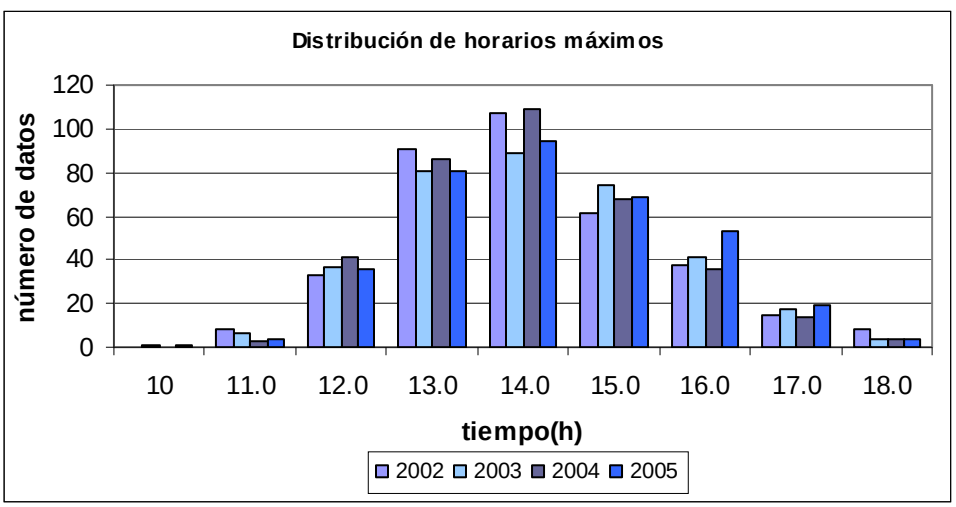


Figura 4.13. Horario de máximo diario de 2002 a 2005, Estación Merced.



Es importante observar de las figuras 4.11, 4.12 y 4.13, lo siguiente:

- De 1996 a 2001 la mayor cantidad de datos se encuentra entre las 12 y 15 h, siendo las 13 h el horario modal.
- Los eventos poco frecuentes son los siguientes: 10, 11, 17 y 18 h.
- De 2002 a 2005 la mayor cantidad de datos se encuentra entre las 12 y 15 h, pero es a las 14 h el horario modal.
- Los eventos poco frecuentes son los mismos que para el periodo de 1996 a 2001: 10, 11, 17 y 18 h.
- Es evidente que el comportamiento de las frecuencias a través de los años es prácticamente el mismo, la única variante es que antes de 2002 la clase modal es las 13 h, y después de 2002 incluido este, la clase modal es las 14 h; lo cual sería muy interesante analizar de manera independiente a este documento.

De este modo se ha establecido el componente cíclico de la concentración de ozono, con la ocurrencia de ciclos de 24 h; dentro de cada uno de estos se registra un valor máximo de concentración que incide en horarios muy específicos. En el periodo analizado de diez años, se tienen dos comportamientos dados por el horario modal en el que se presentan los máximos de concentración, el primero para el periodo 1996 a 2001 en donde las 13 h es el horario modal, y el segundo para el periodo 2002 a 2005, en donde las 14 h es el horario modal. En ambos casos la mayor cantidad de datos se encuentra entre las 12 y 15 h.

Por otra parte, para los dos periodos de tiempo considerados en las figuras 4.11 a 4.13, en base al horario modal de máximos de concentración existe un mismo rango horario en el que se presentan todos los valores máximos, este es de 10 a 18 h, teniendo como eventos poco frecuentes las 10, 11, 17 y 18 h.

Hasta ahora se ha descrito y verificado el comportamiento anual de los valores máximos de concentración, lo cual implica investigar el comportamiento a través de los meses, o dicho de otra forma, es necesario conocer el comportamiento estacional o de mediano plazo para los valores máximos de concentración.

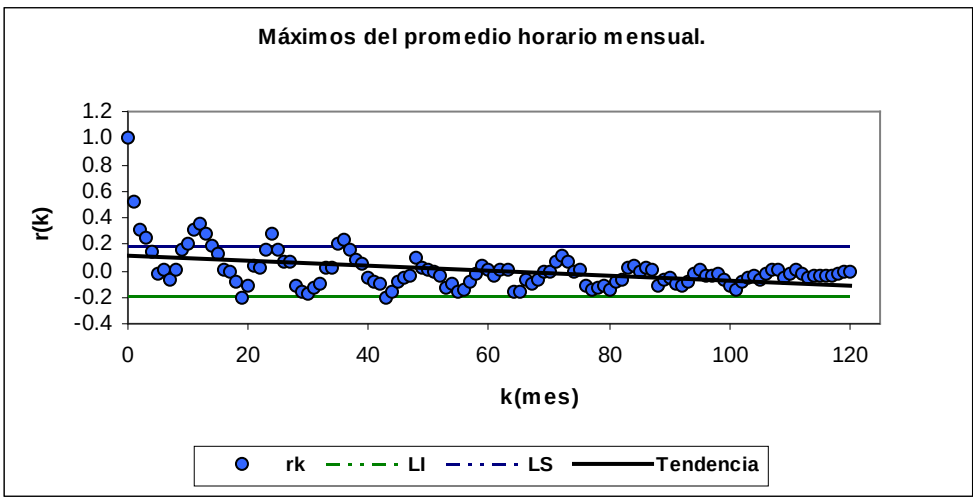
4.3.3.2 Comportamiento estacional de los valores máximos

Para determinar y verificar el comportamiento de mediano plazo de los máximos de concentración, se construyó un correlograma con los valores máximos de los promedios horarios de cada mes a lo largo del periodo de 10 años, en la tabla 4.9 se muestra un fragmento de estos datos, lo cual muestra un panorama general y representativo del conjunto completo de valores máximos.

Tabla 4.9. Promedio horario mensual y su máximo, estación Merced

h	Ene-96	Feb-96	Mar-96	----------	Oct-05	Nov-05	Dic-05
1	0.007	0.007	0.007	----------	0.007	0.004	0.012
2	0.007	0.007	0.007	----------	0.007	0.004	0.012
3	0.007	0.007	0.007	----------	0.008	0.004	0.012
4	0.007	0.007	0.008	----------	0.007	0.004	0.013
5	0.006	0.006	0.007	----------	0.005	0.003	0.013
6	0.006	0.006	0.006	----------	0.004	0.003	0.013
7	0.011	0.012	0.012	----------	0.004	0.004	0.012
8	0.013	0.013	0.012	----------	0.004	0.004	0.012
9	0.014	0.015	0.013	----------	0.007	0.007	0.013
10	0.018	0.019	0.034	----------	0.015	0.014	0.019
11	0.053	0.056	0.069	----------	0.029	0.029	0.033
12	0.101	0.090	0.083	----------	0.045	0.047	0.053
13	0.120	0.100	0.095	----------	0.059	0.059	0.067
14	0.119	0.106	0.097	----------	0.066	0.070	0.078
15	0.104	0.108	0.087	----------	0.059	0.072	0.087
16	0.100	0.092	0.075	----------	0.048	0.068	0.081
17	0.076	0.078	0.058	----------	0.034	0.057	0.068
18	0.051	0.056	0.043	----------	0.019	0.031	0.050
19	0.024	0.030	0.028	----------	0.011	0.013	0.023
20	0.016	0.019	0.019	----------	0.008	0.009	0.018
21	0.012	0.014	0.015	----------	0.007	0.005	0.015
22	0.009	0.009	0.009	----------	0.007	0.005	0.013
23	0.008	0.008	0.008	----------	0.006	0.005	0.013
24	0.007	0.007	0.008	----------	0.006	0.004	0.012
Máximos del promedio horario mensual							
	0.120	0.108	0.097	----------	0.066	0.072	0.087

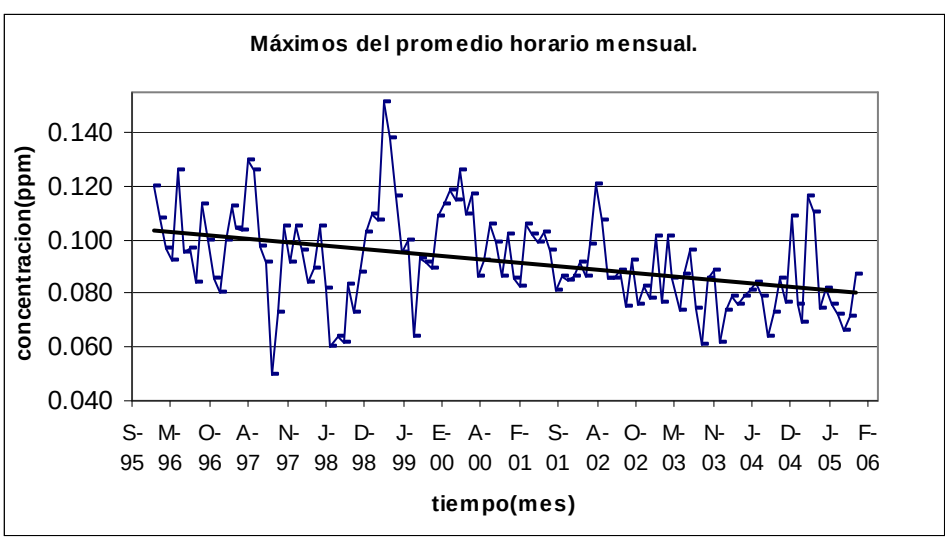
Figura 4.14. Correlograma de los máximos del promedio horario mensual para el periodo 1996 a 2005, estación Merced



En la figura 4.14, se observa un componente cíclico cada 12 desfases, que para este caso corresponde a cada 12 meses, lo cual indica que se trata de un componente estacional o de mediano plazo; por otra parte, se observa una tendencia a la disminución en la concentración de los valores máximos graficados del periodo de 10 años, o dicho de otra manera, el comportamiento a largo plazo no es estacionario. Así mismo, no se observa autocorrelación significativa, lo cual indica que un valor máximo no depende de los valores máximos anteriores.

Para particularizar el patrón cíclico indicado por el correlograma anterior, se construyó una serie de tiempo de los mismos datos (figura 4.15).

Figura 4.15. Máximos del promedio horario mensual, estación Merced



En la figura 4.15, se observa un rango de concentraciones para estos valores máximos muy variable entre años consecutivos, sin embargo, se evidencia una clara tendencia a la disminución de la concentración de valores máximos a través del tiempo. Adicionalmente, se observa un aumento bien definido en el valor máximo de la concentración promedio de enero de cada año respecto a la de diciembre del año anterior.
Por otra parte, cada año el valor máximo de las concentraciones promedio corresponde al mes de abril o mayo para el periodo completo de 10 años, con excepción del año 2004 en donde este valor máximo se presenta en noviembre y para 1998 en enero, como se muestra en la tabla 4.10.

Tabla 4.10. Mes en que se presenta el valor máximo para el promedio horario mensual, estación Merced

Año	Mes
1996	mayo
1997	mayo
1998	enero
1999	abril
2000	mayo
2001	abril
2002	abril
2003	abril
2004	noviembre
2005	abril

Particularizando el comportamiento de los valores máximos de concentración de ozono, se observó que se cumplen ciclos anuales con tendencia a disminuir a través de los años, en donde el valor máximo del promedio horario de cada mes de enero es superior al de diciembre del año inmediato anterior; además, de manera mensual el valor máximo de las concentraciones promedio para cada año corresponde generalmente a los meses de mayo o abril.

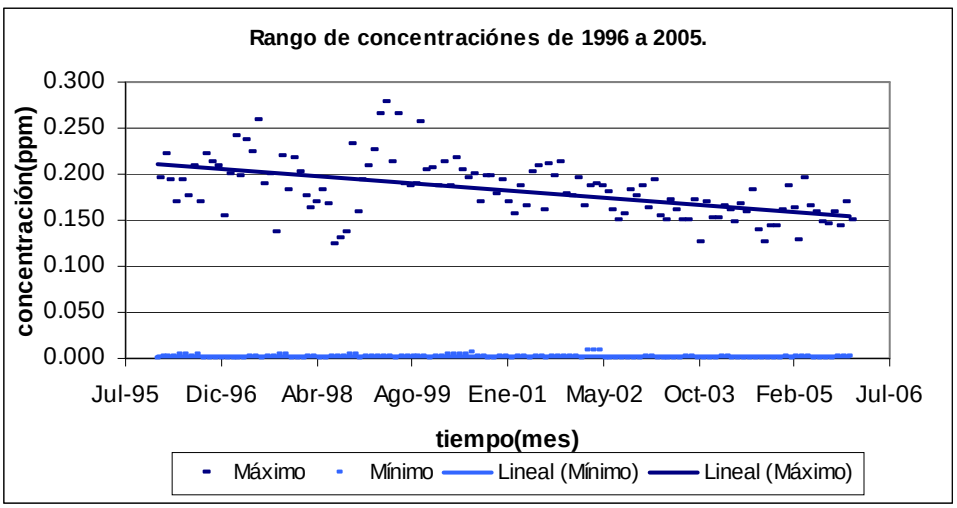
4.3.4 Comportamiento y particularidades de la dispersión de los datos

Con la revisión del comportamiento de los valores máximos de concentración, quedó evidente una gran variabilidad en el rango de los datos, sin embargo, se observó la posible existencia de un patrón de comportamiento de esta misma variabilidad (figura 4.15). Para evidenciar la existencia de un patrón en el comportamiento de la dispersión de los datos, se construyó un gráfico del rango mensual de los datos con ayuda de la tabla 4.11, tal como se muestra a continuación:

Tabla 4.11. Rango mensual de concentraciones, estación Merced

	Ene-96	Feb-96	Jul-00	Ago-00	Dic-01	Ene-02	-------	-------	Nov-02	-------	Nov-05	Dic-05
Máximo	0.195	0.221	0.199	0.169	0.176	0.195	-------	-------	0.176	-------	0.169	0.149
Mínimo	0.001	0.002	0.006	0.002	0.002	0.002	-------	-------	0.001	-------	0.002	0.002
Rango	0.194	0.219	0.193	0.167	0.174	0.193	-------	-------	0.175	-------	0.167	0.147

Figura 4.16. Rango mensual de concentraciones, estación Merced



En la figura 4.16, se observa que aunque el rango mensual de concentraciones es muy variable, existe una tendencia a la disminución de este rango para el periodo de 10 años, lo cual fue sugerido por el gráfico 4.15, con la diferencia de que para este gráfico no se utilizan valores promedio.

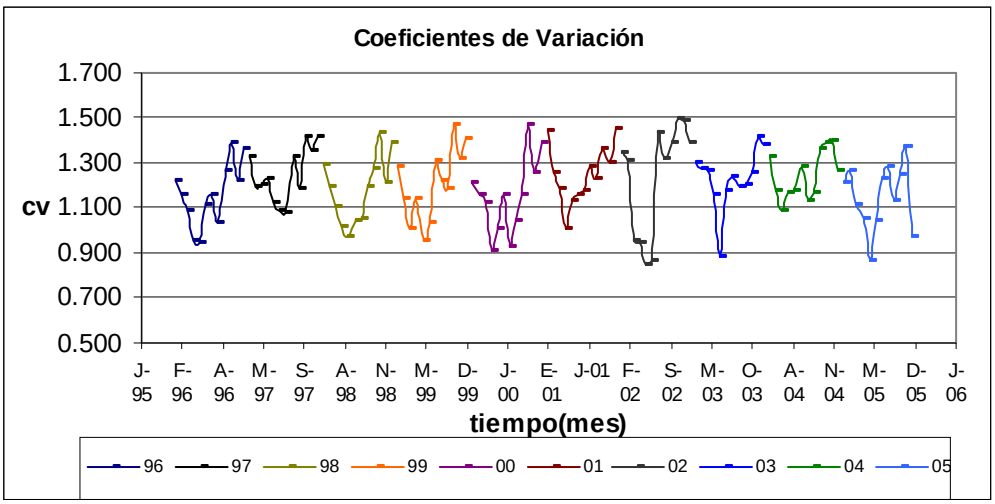
Para comparar la dispersión mensual de estos datos, se utilizó el coeficiente de variación, ya que aunque las unidades de medición son las mismas entre los rangos de concentración, el valor de la dispersión es muy diferente de un mes a otro.

Tabla 4.12.Coeficientes de variación, estación Merced.

	Ene-96	Feb-96	Mar-98	Abr-98	Oct-00	Nov-00	---------- ----------	Jun-03	Jul-03	Nov-05	Dic-05	min
S	0.046	0.0425	0.0324	0.034	0.0405	0.041	---------- ----------	0.027	0.03	0.03	0.03	0.0192
X	0.0376	0.0367	0.0293	0.034	0.0275	0.033	---------- ----------	0.0229	0.024	0.022	0.031	0.0162
CV	1.224	1.1558	1.1041	1.014	1.4703	1.253	---------- ----------	1.1784	1.238	1.375	0.97	0.8428

En la tabla 4.12 (fragmento de la tabla completa) se registran ascensos y descensos aparentemente periódicos del CV, lo cual es mas adecuado visualizar a través de una gráfica o serie de tiempo, como la que se muestra en la figura 4.17.

Figura 4.17.Coeficientes de variación, estación Merced



De este modo, en la figura 4.17, se observa que cada año la dispersión de los datos realiza un ciclo que a continuación se detalla:

- Para cada mes de enero la dispersión fue menor que en diciembre del año anterior.
- Es el mes de octubre el que con mayor frecuencia presenta el CV máximo (6 de 10 años); solo dos meses más lo presentan en el periodo de 10 años, noviembre (3 veces) y diciembre (una vez) (Tabla 4.13).
- De manera análoga, es el mes de mayo el que con mayor frecuencia presenta el CV mínimo (6 de 10 años); tres meses más presentan el CV mínimo en el periodo de 10 años (Tabla 4.13).

Tabla 4.13.CV máximo y mínimo de cada año, Estación Merced.

CV mínimo	May-96	Jul-97	May-98	May-99	Abr-00	Abr-01	May-02	May-03	Mar-04	May-05
CV máximo	Oct-96	Oct-97	Oct-98	Oct-99	Oct-00	Dic-01	Oct-02	Nov-03	Nov-04	Nov-05

De esta manera, se tiene que la dispersión mensual de los datos sigue prácticamente el mismo patrón de comportamiento año con año, en donde, aunque efectivamente el rango mensual de concentraciones es muy variable, se ha evidenciado la tendencia a disminuir de los valores máximos a través del periodo de 10 años, como se muestra en la figura 4.16. Adicionalmente, se establecieron tres particularidades de la dispersión de los datos: 1) un patrón cíclico anual de dispersión en donde se observa que para cada mes de enero ésta es menor que en diciembre del año anterior; 2) en los meses de octubre a diciembre de cada año se presenta mayor dispersión, principalmente en octubre; y 3) en los meses de marzo a mayo se presentan los valores de dispersión más bajos, principalmente en el mes de mayo.

De acuerdo con todo lo anterior y a manera de resumen, después de revisar y buscar las características fundamentales en el comportamiento del ozono, aplicando la metodología propuesta a los datos seleccionados, se obtuvo lo siguiente:

- Todas las bases de datos de las 16 estaciones de monitoreo automático de la RAMA en el Distrito Federal, que se encuentran disponibles en la página electrónica de la SMA, cumplen con un mismo formato de construcción y edición.

- El criterio de suficiencia se cumple para el periodo de 10 años, de 1995 a 1996, para 15 de las 16 estaciones de monitoreo. Cabe recordar que el criterio de suficiencia es relativo a los datos incluidos de manera teórica en el periodo de tiempo elegido.

- Los datos correspondientes a la estación de monitoreo “Merced” muestran carácter no aleatorio, presentan un componente cíclico y autocorrelación significativa. Lo cual, como ya se ha mencionado, son resultados esperados y lógicos, pues el diseño del monitoreo es automático y programado por tiempo.

- Se muestra un componente cíclico descrito por ciclos de 24 horas en donde se registra un valor máximo y los correspondientes ascenso y descenso; hallazgos coincidentes con el comportamiento característico del ozono descrito en el capítulo 2.

- Los valores máximos de concentración presentan comportamientos característicos, como la incidencia de estos en un rango horario específico y la incidencia en horarios modales, como se apunta en el marco teórico; adicionalmente se especificaron incidencias de valores máximos en horarios poco comunes.

- Los valores máximos de los datos seleccionados, efectivamente presentan una marcada tendencia a disminuir a través de los años. Por otra parte, el valor máximo para el promedio horario de cada mes de enero, es superior al de diciembre del año inmediato anterior y el valor máximo anual para el promedio horario mensual de las concentraciones, corresponde al mes de mayo o abril de manera predominante.

- La dispersión de los valores máximos de concentración, presenta un patrón cíclico anual para el periodo completo de 10 años en donde, en cada mes de enero la dispersión es menor que en diciembre del año previo. Así mismo, de octubre a diciembre se presenta la mayor dispersión anual, principalmente en octubre; y por último, los menores valores de dispersión se presenta principalmente en los meses de abril y mayo. Estos últimos

resultados también coinciden con el comportamiento característico que presenta el ozono en la Ciudad de México.

Finalmente, tomando en cuenta esta serie de resultados que como ya se mencionó, se obtuvieron de aplicar la "Metodología para la Validación de Datos de Calidad del Aire Generados por una Red de Monitoreo Automático", se considera pertinente otorgar la etiqueta de Nivel 2 de validación a los datos que se generaron en la "Estación Merced" de la Red Automática de Monitoreo Atmosférico de la Secretaría del Medio Ambiente de la Ciudad de México, durante el periodo enero de 1996 a diciembre de 2005.

4.4 Aplicación de los datos validados

Partiendo de los resultados anteriores, en donde se concluye que los datos seleccionados para la aplicación de la metodología propuesta han sido etiquetados como datos que cumplen con el nivel 2 de validación, y para ejemplificar una de las actividades propias del nivel 3 de validación que se pueden realizar de manera inicial con éste tipo de datos, se aplicó la prueba de bondad de ajuste de Kolmogorov-Smirnov (prueba de normalidad) a datos de horario consecutivo y a datos de horario específico para el periodo de 10 años.

4.4.1 Prueba de bondad de ajuste a datos en horario consecutivo

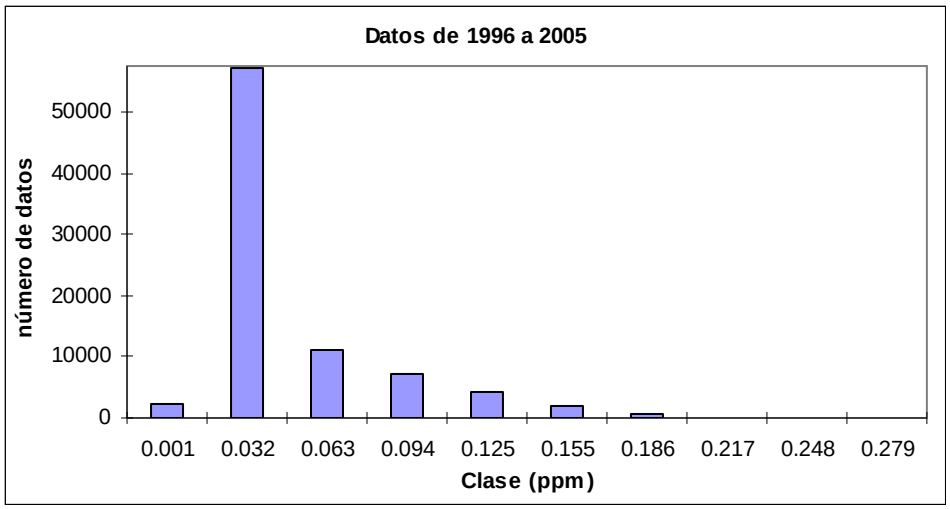
Como un primer ejemplo de aplicación, se utilizó la prueba de normalidad de Kolmogorov-Smirnov con datos de horario consecutivo, es decir, considerando las 24 h del día y la totalidad de la base de datos de O_3, correspondiente a la estación de monitoreo Merced para el periodo enero de 1996 a diciembre de 2005.

- Construcción del histograma

Tabla 4.14. Frecuencias 1996 a 2005 , estación Merced

CLASE		# de datos
<	0.0010	2184
0.0010	0.0319	57252
0.0319	0.0628	11166
0.0628	0.0937	7129
0.0937	0.1246	4113
0.1246	0.1554	1911
0.1554	0.1863	593
0.1863	0.2172	126
0.2172	0.2481	28
0.2481	0.2790	7
0.2790	>	0

Figura 4.18. Histograma 1996 a 2005, estación Merced



- Planteamiento de la hipótesis

  (Ho). Las diferencias entre los valores observados y los valores de la distribución normal se deben al azar

  (Ha). Los valores observados de las frecuencias para cada clase son diferentes de las frecuencias de la distribución normal

- Nivel de significancia

  Para todo valor de probabilidad igual o menor que 0.05, se acepta Ha y se rechaza Ho

- Zona de rechazo

  Para todo valor D' menor o igual que el valor crítico D, se acepta la hipótesis nula

- Cálculo de las frecuencias observadas acumuladas (*Fobs*)

Tabla 4.15. Frecuencias Observadas Acumuladas.

CLASE		Observado	Fobs
<	0.0010	2184	2184
0.0010	0.0319	57252	59436
0.0319	0.0628	11166	70602
0.0628	0.0937	7129	77731
0.0937	0.1246	4113	81844
0.1246	0.1554	1911	83755
0.1554	0.1863	593	84348
0.1863	0.2172	126	84474
0.2172	0.2481	28	84502
0.2481	0.2790	7	84509
0.2790	>	0	84509

- Cálculo de las frecuencias normales acumuladas(*Ft*)

Tabla 4.16. Frecuencias normales acumuladas.

CLASE		Z_i	Z_s	$F(Z_i)$	$F(Z_s)$	p	Esperado	Ft
<	0.0010		-0.7909		0.2145	0.2145	18128	18128
0.0010	0.0319	-0.7909	0.0646	0.2145	0.5258	0.3113	26303	44431
0.0319	0.0628	0.0646	0.9201	0.5258	0.8212	0.2955	24971	69403
0.0628	0.0937	0.9201	1.7756	0.8212	0.9621	0.1409	11904	81306
0.0937	0.1246	1.7756	2.6311	0.9621	0.9957	0.0336	2843	84149
0.1246	0.1554	2.6311	3.4866	0.9957	0.9998	0.0040	339	84488
0.1554	0.1863	3.4866	4.3421	0.9998	1.0000	0.0002	20	84508
0.1863	0.2172	4.3421	5.1976	1.0000	1.0000	0.0000	1	84509
0.2172	0.2481	5.1976	6.0531	1.0000	1.0000	0.0000	0	84509
0.2481	0.2790	6.0531	6.9086	1.0000	1.0000	0.0000	0	84509
0.2790	>	6.9086		1.0000		0.0000	0	84509

- Calculo de D’ con la ecuación (2.21): $D' = \frac{Ft - Fobs}{n}$

Tabla 4.17. Diferencia Máxima D'

CLASE		Fobs	Ft	(Ft-Fobs)/n
<	0.0010	2184	18128	0.189
0.0010	0.0319	59436	44431	-0.178
0.0319	0.0628	70602	69403	-0.014
0.0628	0.0937	77731	81306	0.042
0.0937	0.1246	81844	84149	0.027
0.1246	0.1554	83755	84488	0.009
0.1554	0.1863	84348	84508	0.002
0.1863	0.2172	84474	84509	0.000
0.2172	0.2481	84502	84509	0.000
0.2481	0.2790	84509	84509	0.000
0.2790	>	84509	84509	0.000

en donde $D' = 0.189$

- Cálculo del valor crítico del estadístico de Kolmogorov-Smirnov (ec. 2.22) y comparación con (2.23) y (2.24)

$$D = \frac{1.358}{\sqrt{n}} = 0.00467 \qquad \text{por lo tanto} \quad D' > D$$

De acuerdo con lo anterior, el estadístico (*D'*) de Kolmogorov-Smirnov obtenido es mayor que el valor crítico (*D*), por lo tanto, se rechaza Ho y se acepta Ha. En otras palabras, las frecuencias observadas en la distribución de los datos seleccionados y la distribución normal calculadas, difieren significativamente, por lo tanto, las observaciones del conjunto de datos considerado no describen una distribución normal.

Del mismo modo, cualquier otro periodo de años o meses que incluya el horario consecutivo o completo (24h), no se comporta de acuerdo a una distribución normal. Por otra parte, se aplicó la prueba de bondad de ajuste de Kolmogorov-Smirnov con la distribución lognormal, obteniendo que en ningún caso los datos se comportan de acuerdo a la distribución teórica.

4.4.2 Prueba de bondad de ajuste a datos de horario específico

Como segundo ejemplo, se aplicó la prueba de normalidad de Kolmogorov-Smirnov a datos en horario específico en periodos de tiempo también específicos, considerando la base de datos de O_3 correspondiente a la estación de monitoreo "Merced" para los siguientes dos casos: I) datos de las 13:00 h para el periodo 1996 a 2001 y II) datos de las 14:00 h. para el periodo 2002 a 2005.

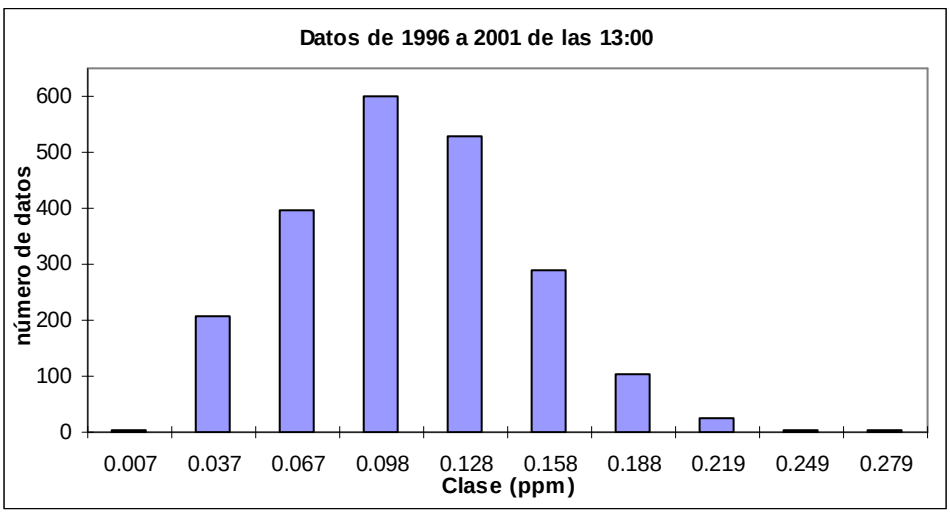
Caso I. Aplicación de los datos de las 13:00 h para el periodo 1996 a 2001

- Planteamiento de la hipótesis

  (Ho). Las diferencias entre los valores observados y los valores de la distribución normal se deben al azar

  (Ha). Los valores observados de las frecuencias para cada clase son diferentes de las frecuencias de la distribución normal

Gráfico 4.19. Histograma 1996 a 2001 , estación Merced.



D' = 0.01817

$$D = \frac{1.358}{\sqrt{n}} = 0.02919$$

$$D' < D$$

De acuerdo con el estadístico (*D'*) de Kolmogorov-Smirnov obtenido, el cual resultó menor que el valor crítico (*D*) , se acepta Ho y se rechaza Ha. En otras palabras, las frecuencias observadas del conjunto de datos considerado y las frecuencias calculadas de la distribución normal, no difieren significativamente, por lo tanto, los datos de ozono de las 13:00 h para el periodo 1996 a 2001 siguen una distribución normal al cinco por ciento de nivel de significancia.

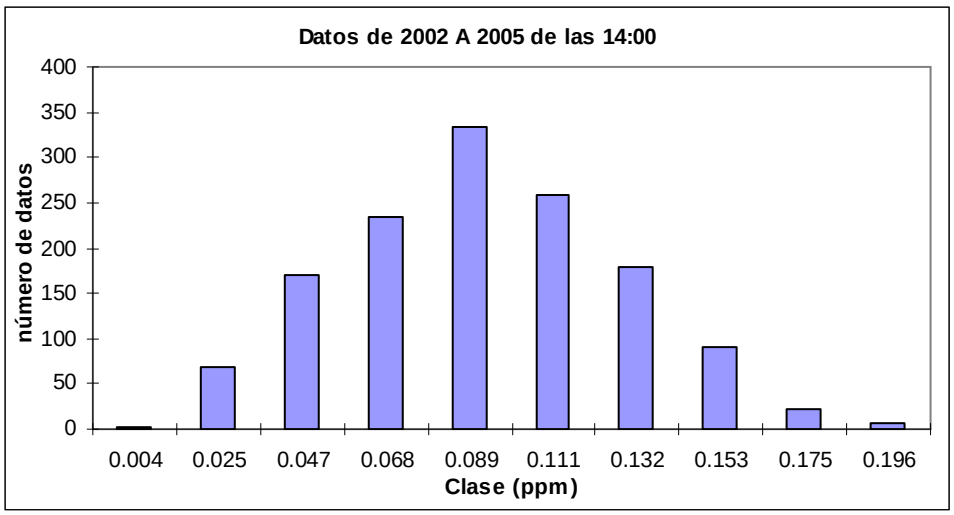
Caso II. Datos de las 14:00 h. para el periodo 2002 a 2005. Aplicación de los datos de las 14:00 h. para el periodo 2002 a 2005.

- Planteamiento de la hipótesis.

  (Ho). Las diferencias entre los valores observados y los valores de la distribución normal se deben al azar.

  (Ha). Los valores observados de las frecuencias para cada clase son diferentes de las frecuencias de la distribución normal.

Gráfico 4.20. Histograma 2002 a 2005, estación Merced.



$$D' = 0.01440$$

$$D = \frac{1.358}{\sqrt{n}} = 0.03676$$

$$D' < D$$

De acuerdo con el estadístico (*D'*) de Kolmogorov-Smirnov obtenido, el cual resultó menor que el valor crítico (*D*) , se acepta Ho y se rechaza Ha. En otras palabras, las frecuencias observadas del el conjunto de datos considerado y las frecuencias calculadas de la distribución normal, no difieren significativamente, por lo tanto, los datos de ozono de las 14:00 h para el periodo 2002 a 2005 siguen una distribución normal al cinco por ciento de nivel de significancia.

Por el contrario, el obtener comportamiento de distribución normal en los datos de los dos horarios específicos, 13:0 h y 14:00 h, no implica que las 22 h restantes deban comportarse de igual manera, de hecho, la misma prueba de bondad de ajuste arroja resultados de no normalidad para estos casos. Una de las

explicaciones para este comportamiento es que, en los horarios 13:00 h y 14:00 h, condiciones meteorológicas como radiación solar y temperatura son más estables, principalmente la radiación solar que en tales horarios es máxima.

De manera análoga, para los datos de horarios específicos diferentes a las 13:00 h y 14:00 h, y para los datos en horarios consecutivos, se ha realizado la misma prueba de bondad de ajuste tanto con la distribución normal como con la aproximación lognormal, obteniendo de manera generalizada que estos datos no describen un comportamiento semejante a estas dos distribuciones teóricas. De aquí que, para un análisis posterior de estos últimos datos, es adecuado utilizar métodos de estadística no paramétrica para su análisis.

Capítulo 5

Conclusiones

Derivado de la aplicación de la metodología propuesta, se ha obtenido una serie de resultados, que si bien no todos están referenciados en el marco teórico, si son resultados lógicos, como se apunta en los párrafos posteriores.

A través de la revisión bibliográfica se ha logrado definir cada uno de los niveles de validación, lo cual ha sido de fundamental importancia para lograr establecer el tipo de herramientas aplicables, para la realización de la propuesta metodologica citada en el título de esta tesis.

Durante la revisión hecha sobre las disponibilidad y disposición de los datos, se encuentra que todos los archivos incluidos en la base de datos correspondiente a las 16 estaciones de monitoreo automático en el periodo enero de 1996 a diciembre de 2005, cuentan con la configuración descrita en 4.1. Además, todos estos archivos se encontraron disponibles en la página electrónica de la SMA con las mismas características de formato y visualización también descritas en la sección 4.1. Lo anterior, ha permitido extraer y manipular los datos, a fin de realizar las revisiones y verificaciones subsecuentes.

Referente a la verificación del criterio de suficiencia, se ha demostrado, como se indica en la sección 4.2, cuales son las estaciones de monitoreo que no cumplen con la cantidad mínima aceptable de datos para poder continuar con el proceso de validación. De ésta manera, ha sido posible identificar y seleccionar las estaciones de monitoreo susceptibles a la aplicación de la metodología propuesta para la validación de los datos a un nivel 2. Así mismo, se ha podido puntualizar, que el cumplimiento del criterio de suficiencia es relativo al periodo de tiempo que se considera.

Al continuar con la determinación de las características fundamentales de los datos, se obtuvo una serie de resultados los cuales han sido coincidentes con las características referenciadas en el capítulo 2 sobre el comportamiento del ozono

en la Ciudad de México, como la descripción de una curva característica en el comportamiento del ozono para la ciudad de México. En la sección 4.3, se mencionan específicamente los resultados coincidentes de la estación de monitoreo Merced, algunos son:

- El ascenso de concentraciones alrededor de las 8:00 h y la disminución a partir de las 13:00 o 14:00 h.
- En el periodo de 1996 a 2005 se presenta una clara tendencia a la disminución en las concentraciones máximas de ozono.
- El intervalo de horas en que se presentan los máximos de concentración es de 10 a 18 h.
- La mayor cantidad de datos se encuentra entre las 12 y 14 h (70.2%).
- Como eventos poco comunes, incidencias en las 10, 17 y 18 h (4.1%).
- En la tabla 4.11 se observan dos comportamientos; el primero, de 1996 a 2001 en donde el horario modal de máximos es las 13:00 h.; el segundo, de 2002 a 2005 en donde el horario modal es las 14:00 h.

Se mencionó el hallazgo de comportamientos no documentados para la Ciudad de México, los cuales, aunque resultan lógicos o esperados como ya se ha dicho, no es adecuado generalizarlos hasta no confirmarlos o descartarlos por medio del análisis de datos del mismo tipo para otras estaciones de monitoreo de la misma red; éstos hasta ahora, son hallazgos particulares de la estación de monitoreo considerada (Merced), hallazgos como los siguientes:

- Del comportamiento de los valores máximos de concentración de ozono, se observó que se cumplen ciclos anuales con tendencia a disminuir a través de los años; en donde el valor máximo del promedio horario de cada mes de enero, es superior al de diciembre del año inmediato anterior; además, de manera mensual el valor máximo de las concentraciones promedio para cada año, corresponde generalmente a los meses de mayo o abril.
- La dispersión de los valores máximos de concentración, presenta un patrón cíclico anual para el periodo completo de 10 años en donde, en cada mes de enero la dispersión es menor que en diciembre del año previo. Así mismo, de octubre a diciembre se presenta la mayor dispersión anual (temporada fría-seca), principalmente en octubre; y por último, los menores valores de dispersión se presenta principalmente en los meses de abril y mayo que corresponde a la temporada más cálida del año. Estos últimos resultados también coinciden con el comportamiento característico que presenta el ozono en la Ciudad de México.

Por lo anterior, se debe considerar que los datos presentan consistencia temporal y espacial concluyendo que el conjunto de herramientas estadísticas seleccionadas, ha sido el adecuado y suficiente para verificar la existencia de los comportamientos característicos del ozono en datos generados por la RAMA en su estación Merced y así, se recomienda otorgar a estos la etiqueta de nivel 2 de validación.

Recomendaciones.

Es de suma importancia considerar, que para iniciar le revisión de datos correspondiente al nivel 2 de validación, se deben realizar de manera exhaustiva las revisiones que involucran los niveles previos de validación (nivel 0 y nivel 1). De manera general, en estos dos niveles se revisan las evidencias referentes a aseguramiento y control de calidad tanto de la parte documental como de la parte técnica de la generación de los datos, lo cual se ha planteado en el marco teórico y se han mencionado ejemplos de los documentos que se revisan en estos dos niveles de validación.

Por otra parte, es recomendable que al emitir la información de las bases de datos, se considere un formato que integre periodos más amplios de tiempo, pues al manipularlos se requiere conjuntar gran cantidad de archivos (uno por cada mes de cada año) aun tratándose del mismo contaminante.

Es necesario que el generador de los datos que han sido validados con la presente metodología, realice una revisión de las bases de datos emitidas, ya que se han encontrado valores en estas que no corresponden a concentraciones en ppm; específicamente datos como: -99.999, 0.000 y celdas vacías, los cuales fueron eliminados.

De acuerdo con la información que el SINAICA publica en la página WEB del INE, cada una de las redes de monitoreo integradas al SINAICA cuentan con procedimientos propios de validación y revisión de datos, que como se menciona en la introducción, son procedimientos que aplican criterios de revisión diferentes y que por lo tanto son diferentes para cada red. Lo recomendable es que se unifiquen y estandaricen tales procedimientos a través del establecimiento de una metodología que integre los mismos criterios básicos de revisión de datos, lo cual arrojara bases de datos con criterios de validación homólogos.

Referencias

Charley A. Knoderer, Timothy S. Dye, Sonoma Technology, Inc., Petaluma, CA (2003). AMS Short Course on the Fundamentals of Boundary Layer Wind and Temperature Profiling Using Radar and Acoustic Techniques Long Beach, CA.

Chow Ya Lun; 1994. Análisis Estadístico; 2ª edición; Mc. Graw Hill.

Ginevan Michael E., Splistone Douglas E; 2003. Statistical Tools for Environmental Quality Measurement; Chapman and Hall – CRC.

Grant Eugene L., Leavenworth Richard S; 1988. Statistical Quality Control; six edition; Mc. Gaw Hill.

Guzmán Gómez D., López Ramírez G., Rodríguez Baracaldo R.; 2004. Determinación de la Capacidad Requerida Para la Prestación del Servicio de Mantenimiento en Plantas de Generación de Energía Hidroeléctrica. Revista de la Facultad de Ingeniería de la Universidad de Antioquia; Medellín, Colombia.

Gilbert O. R.; 1997. Statistical Methods for Environmental Pollution Monitoring; Van Nostrand, New York.

Jaimes Palomera, M., Muñoz Cruz, R.; 2004. Análisis del Comportamiento del Ozono por Época del Año en la Ciudad de México. Secretaría del Medio Ambiente del Distrito Federal; México.

Jiménez P y col.; 2004. Estudio de Variaciones Climáticas e Hidrológicas Ocurridas Durante el Último Siglo A partir del Análisis Correlatorio y Espectral de Series Temporales de Datos Registrados en el Sur de la Península Ibérica. Universidad de Málaga, Grupo de Hidrogeología; Instituto Geológico y Minero de España, Dirección de Hidrogeología y Aguas Subterráneas.

Mc. Clave James T., Dietrich Frank A., Sincich Terry; 1997. Statistics; vol. II; seven edition; Prentice Hall.

Mc. Bean Edward, Rovers Frank; 1998. Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment; vol 3; Prentice Hall PTR.

Molinero Luis M.; 2004. Análisis de series temporales; Sociedad Española de Hipertensión.

Montgomery Douglas, Runger George; 2003. Applied Statistics and Probability for Engineers; third edición; John Wiley and Sons.

T. W. Anderson, Stanley I. Sclove, 1998. An introduction to the Statistical analysis of Data; Houghton Mifflin.

Trueba V. Leopoldo; 1998. Elementos Estadísticos en el Análisis de las Series de Tiempo; Universidad Autónoma de Zacatecas.

Muñoz Cruz R., Ortuño Mojica C.; 2004. Evaluación Comparativa de las Mediciones de Ozono en Estaciones de Monitoreo del SIMAT. Secretaría del Medio Ambiente del Distrito Federal; México.

Ortuño Mojica C, Jaimes Palomera, M.; 2004. Análisis del Comportamiento Semanal de Ozono en la Ciudad de México, 1990-2003. Secretaría del Medio Ambiente del Distrito Federal; México.

Programa Para Mejorar la Calidad del Aire ZMVM 2002-2010, Capitulo 3: Los Indicadores de la Calidad del Aire. Secretaría del Medio Ambiente del Distrito Federal; México.

Programa Para Mejorar la Calidad del Aire ZMVM 2002-2010, Capitulo 6: Comportamiento Diario Típico de los Contaminantes Atmosféricos en la ZMVM. Secretaría del Medio Ambiente del Distrito Federal; México.

Proyecto de Norma Oficial Mexicana PROY-NOM-156-SEMARNAT-2008, establecimiento y operación de sistemas de monitoreo de la calidad del aire.

U.S. Environmental Protection Agency; 2000. Guidance For Data Quality Assessment; Practical Methods for Data Analysis; EPA QA/G9.

U.S. Environmental Protection Agency; 1996. The Data Quality Evaluation Statistical Toolbox; EPA QA/G9D.

U.S. Environmental Protection Agency; 2002. Guidance on Environmental Data Verification and Data Validation; EPA QA/G8.

U.S. Environmental Protection Agency; 2000. Quality Assurance Handbook for Air Pollution Measurement Systems. vol II; Ambient Air Quality Monitoring Program Quality System Development.

U.S. Environmental Protection Agency; 1998. Quality Assurance Guidance Document Model Quality Assurance Project Plan for the PM Ambient Air 2.5 Monitoring Program at State and Local Air Monitoring Stations (SLAMS).

U.S. National Institute of Standards and Technology; Engineering Statistics Handbook. Kolmogorov-Smirnov Goodness of Fit Test.

Washington State Department of Ecology; 2004. Air Quality Program; Automated Method Data Documentation and Validation Procedure.

Watson J.G., DuBois D.W., DeMandel R., Kaduwela A., Magliano K., McDade C., Mueller P.K., Ranzieri A., Roth P.M., and Tanrikulu S. (1998) Aerometric monitoring program plan for the California Regional PM10/PM2.5 Air Quality Study. Prepared for California Regional PM10/PM2.5 Air Quality Study Technical Committee, California Air Resources Board, Sacramento, CA by Desert Research Institute, Reno, NV, DRI Document No. 9801.1D5, December.

Anexo A

Representaciones Gráficas

En este anexo se ejemplifica la aplicación de las representaciones gráficas descritas en el capítulo 2. La mayoría de estas representaciones gráficas se construyen de manera muy sencilla, aun así, no se omiten detalles que se consideran importantes como señalar las ecuaciones utilizadas. Por otra parte, todos los cálculos y gráficos se realizaron por medio de Microsoft Excel, y solo se muestran los resultados y despliegue de gráficos.

A1

Histograma

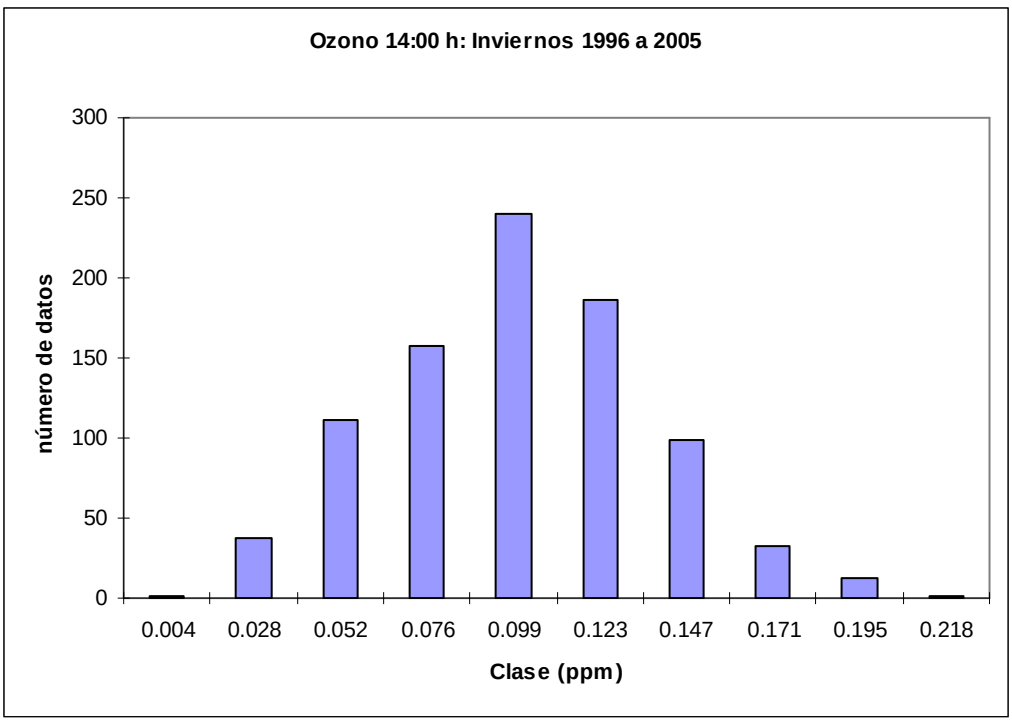
Con datos de monitoreo de O_3 en ppm de las 14:00 h., para los meses enero, febrero y diciembre de 1996 a 2005 de la estación de monitoreo Merced, se construye un histograma.

El rango de los datos se calcula con la ecuación (2.1) y el ancho de clases se calcula con la ecuación (2.2).

Tabla A1. Construcción de clases

Máximo	0.242	
Mínimo	0.004	
Rango	0.238	
No. de clases	10.000	
Ancho de clases	0.0238	
CLASE		*Frecuencia*
<	0.0042	1
0.0042	0.0280	38
0.0280	0.0518	111
0.0518	0.0756	157
0.0756	0.0994	240
0.0994	0.1232	186
0.1232	0.1470	99
0.1470	0.1708	33
0.1708	0.1946	12
0.1946	0.2184	1
0.2184	>	1

Figura A1. Histograma



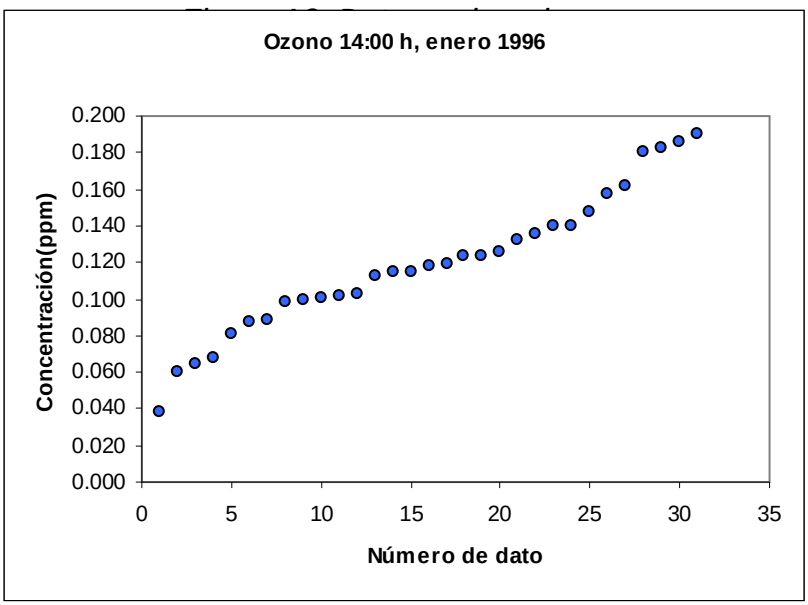
A2

Datos ordenados

Con datos de monitoreo de O_3 en ppm de las 14:00 h., para el mes de enero de 1996 de la estación de monitoreo Merced, se construye un gráfico de datos ordenados por magnitud. En la primera columna se tienen los datos en el orden en que fueron tomados, en la segunda columna los datos ordenados por magnitud o tamaño.

Tabla A2. Datos de O_3 de las 14:00 h, enero 1996, Merced

Ordenados	
Por día	Por magnitud
ppm	ppm
0.065	0.038
0.038	0.060
0.115	0.065
0.132	0.068
0.124	0.081
0.140	0.087
0.162	0.089
0.098	0.098
0.186	0.100
0.115	0.101
0.101	0.102
0.089	0.103
0.183	0.113
0.119	0.115
0.118	0.115
0.087	0.118
0.148	0.119
0.140	0.123
0.136	0.124
0.113	0.126
0.123	0.132
0.102	0.136
0.103	0.140
0.180	0.140
0.190	0.148
0.157	0.157
0.060	0.162
0.126	0.180
0.081	0.183
0.068	0.186
0.100	0.190



A3

Gráfico de dispersión

Con los datos de la tabla A3 se construye un gráfico de dispersión para dos variables.

Tabla A3. Velocidad del viento y corriente de salida de un molino de viento.

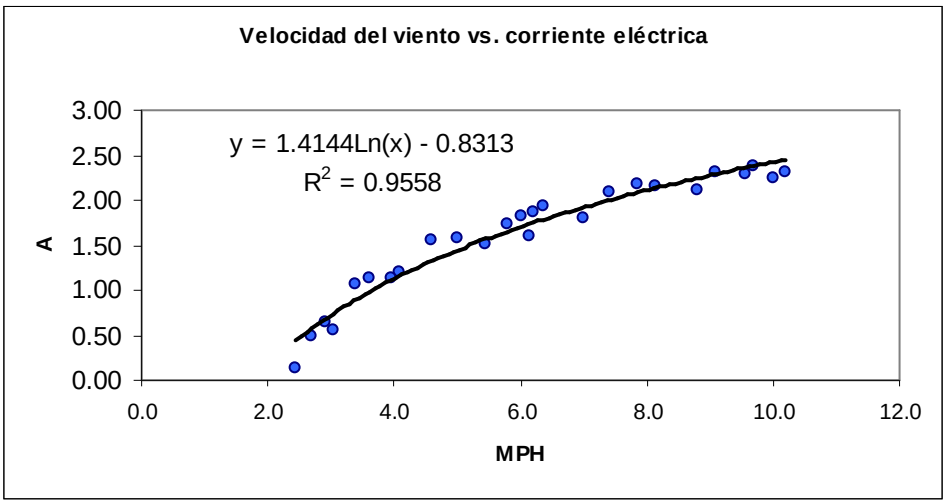
n	Velocidad del viento (mph)	Corriente de salida (A)
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

(Tomado y modificado de Applied Statistics and Probability for Engineers; Montgomery & Runger; John Wiley & Sons)

X → Velocidad del viento (mph)

Y → Corriente eléctrica (A)

Figura A3. Dispersión.



A4

Datos temporales

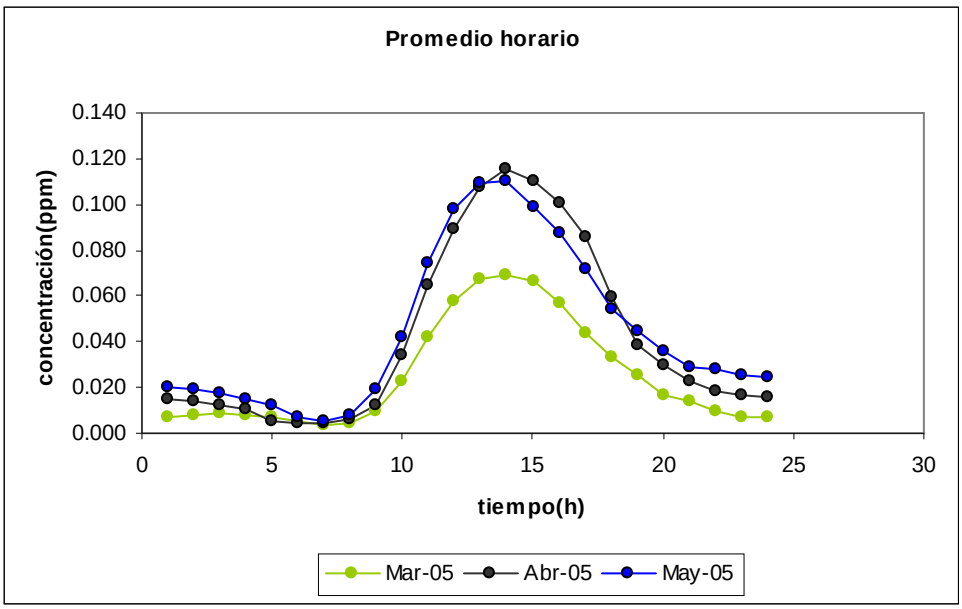
Series de tiempo

Con datos de monitoreo de O_3 en ppm para los meses marzo, abril y mayo de 2005 de la estación Merced, se construyen series de tiempo de los promedios horarios mensuales. El promedio horario mensual, es el promedio aritmético del valor de la concentración de la misma hora de todo el mes, el cual se calcula con la ecuación (2.10).

Tabla A4. Promedio horario mensual MER.

h	Mar-05	Abr-05	May-05
1	0.007	0.015	0.020
2	0.008	0.014	0.019
3	0.009	0.013	0.017
4	0.008	0.011	0.015
5	0.007	0.005	0.012
6	0.005	0.004	0.007
7	0.004	0.004	0.006
8	0.005	0.006	0.008
9	0.010	0.012	0.019
10	0.023	0.034	0.042
11	0.042	0.065	0.074
12	0.057	0.089	0.098
13	0.067	0.107	0.109
14	0.069	0.116	0.110
15	0.066	0.110	0.099
16	0.057	0.100	0.088
17	0.044	0.086	0.072
18	0.033	0.059	0.054
19	0.025	0.039	0.044
20	0.017	0.030	0.036
21	0.014	0.023	0.029
22	0.009	0.018	0.028
23	0.007	0.016	0.026
24	0.007	0.016	0.024

Figura A4. Series de Tiempo.



A5

Datos temporales

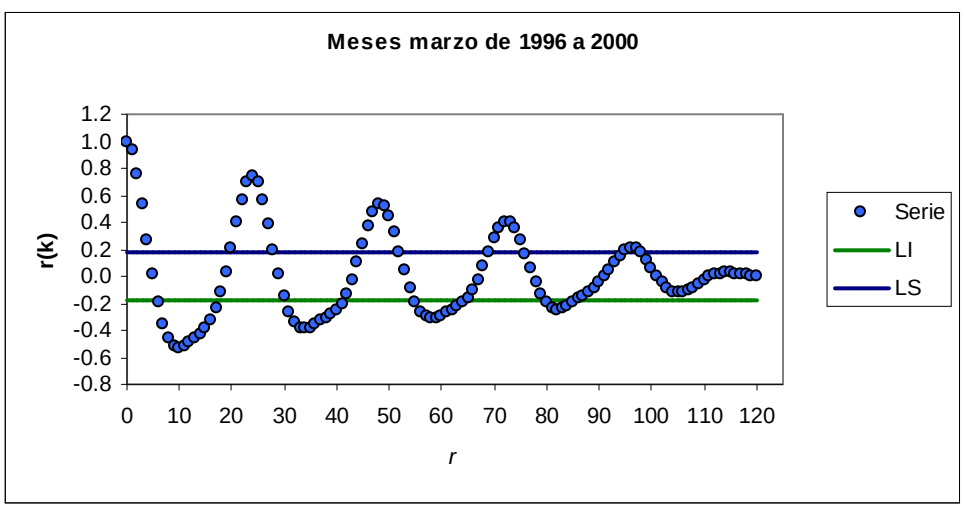
Correlograma o función de autocorrelación

En las siguientes páginas se dan tres ejemplos de la función de autocorrelación, todos los correlogramas se construyeron con el procedimiento ya descrito en el capítulo 2, sección 2.3.1. Como ya se ha mencionado al inicio del anexo, el cálculo de los coeficientes de autocorrelación se realizó con Microsoft Excel y por tratarse de hojas de cálculo muy grandes (y sin sentido aparente) no se muestran los resultados numéricos, solamente los gráficos generados.

A5.1. Correlograma sin ausencia de datos

La construcción de un correlograma de un conjunto de datos completo (sin ausencias), permite obtener información real del comportamiento de los datos a través del tiempo. Con datos de los promedios horarios para los meses marzo de 1996 a 2000, se construye el correlograma de la serie de tiempo.

Figura A5. Correlograma.



De la figura A5 se sabe que los datos son no aleatorios, pues existen coeficientes $r_1, r_2, \ldots, r_k$ que no son cercanos estadísticamente a cero a lo largo de toda la serie; son estacionales porque cada 24 desfases se repite el mismo comportamiento de

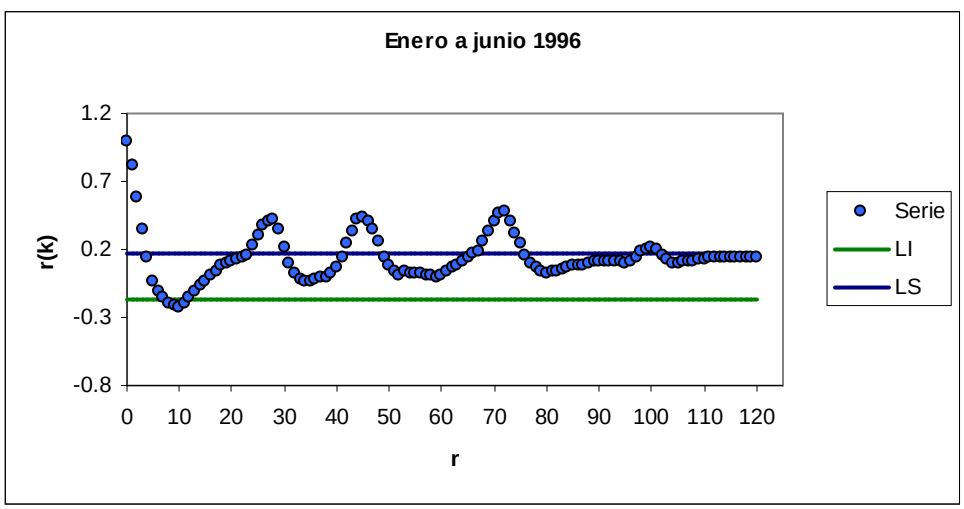
máximos y estacionarios porque r_1 es cercano a 1 y los sucesivos r_2,.....,r_k tienden a cero rápidamente; además, presentan una fuerte autocorrelación indicada por los r(k) máximos que rebasan los límites $-1.96/\sqrt{n} < 0 < 1.96/\sqrt{n}$, como indica la ec. 2.5.

A5.2 Correlograma con ausencias de datos

Un caso particular en la construcción de un correlograma es cuando existe ausencia de datos; los datos previos y los posteriores a las ausencias mostrarán distorsión del carácter estacionario (si es que existe); los datos no aleatorios pueden parecer aleatorios; y sería muy aventurado decidir si existe o no estacionalidad y por consecuencia si existe o no autocorrelación. Para solventar este problema existente en casi todos los meses de casi todos los años, se calculan promedios horarios y de esta manera se minimizan los efectos adversos de la ausencia de datos, como en el gráfico A5. Se debe recordar que en cualquier caso las ausencias no deben exceder el 25% de los datos del periodo de tiempo en estudio.

El siguiente gráfico muestra una serie de tiempo en donde la ausencia de datos distorsiona un correlograma. En este caso se utilizan datos del promedio horario mensual para los meses de enero a junio de 2006, de donde se eliminaron al azar 20 de 144 datos (14%).

Figura A6. Correlograma del promedio horario mensual con ausencia de datos.

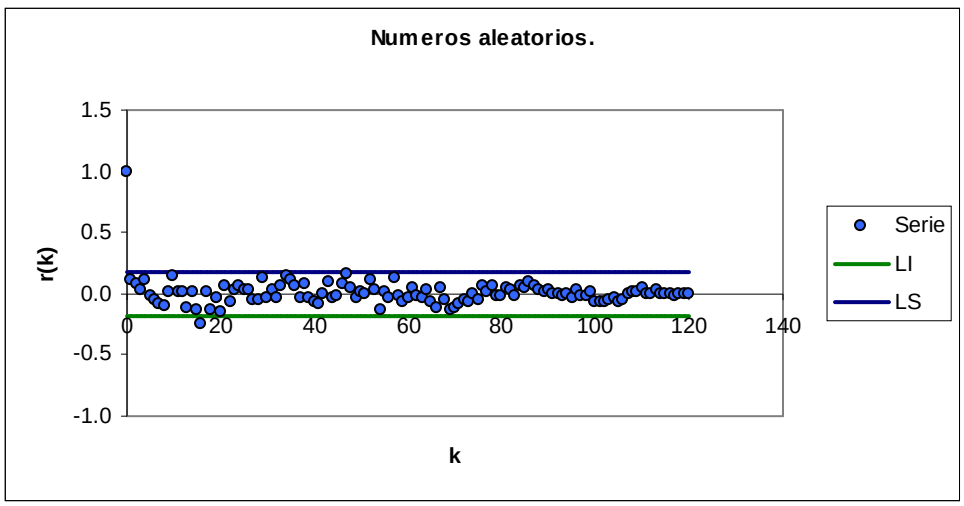


En la figura A6 se esperarían las mismas características que en la figura A5, pero al existir ausencias de datos no es posible realizar observaciones determinantes pues el gráfico se deforma y pierde las características que permiten determinar si los datos son aleatorios y si tienen carácter estacional y/o estacionario.

A5.3 Correlograma de datos aleatorios

Suponiendo un muestreo aleatorio con repetición, de 120 números en el rango [0 a 1000] se construye el siguiente correlograma (los datos se generaron por medio de microsoft excel).

Figura A7. Correlograma de datos aleatorios.



En la figura A7 se muestra un conjunto de datos aleatorios, pues los coeficientes r_2,.....,r_k son estadísticamente cercanos a cero; no se presenta carácter estacional ni estacionario y tampoco autocorrelación.

Anexo B

Medidas de estadística descriptiva

En este anexo se ejemplifica la aplicación de las cantidades estadísticas descritas en el capítulo 2 indicando cuales son las ecuaciones utilizadas. Aunque todos los cálculos se realizaron por medio de Microsoft Excel, se verificó que el software utiliza las mismas ecuaciones que se describen en capítulo 2.

B1

Medidas de posición relativa

Se calculan el percentil 90 ($X_{.90}$) y 95 ($X_{.95}$) para los datos de la tabla B1.

Tabla B1. Concentración química de datos de monitoreo.

n	1	2	3	4	5	6	7	8	9	10
ppb	4	4	4	5	5	6	7	7	8	10

(Tomado de Guidance for Data Quality Assessment, Practical Methods for Data Analysis; EPA QA/G-9)

para $X_{.90}$: con la ecuación (2.6) se calcula t y con la ecuación (2.7) se calcula nt

$$t = \frac{90}{100} = 0.90 \qquad \text{y} \qquad nt = (10)(0.90) = 9.0\,;$$

en donde, $e = 9$ y $d = 0$, por lo que se utiliza la ecuación (2.8) para calcular $X_{.90}$

$$x_{.90} = \frac{(8+10)}{2} = 9\,ppb\,.$$

para $X_{.95}$: $t = \frac{95}{100} = 0.95$ y $nt = (10)(0.95) = 9.5$

en donde, $e = 9$ y $d = 5$, por lo que se utiliza la ecuación (2.9) para calcular $X_{.95}$

entonces $x_{.95} = x_{(10)} = 10\,ppb\,.$

B2

Medidas de tendencia central

Se calculan la media, mediana y moda para los datos de la tabla B2.

Tabla B2. Concentración química de datos de monitoreo.

n	*Cloro (mg/L)*	*Ordenados*
1	*25.25*	*11.74*
2	*13.32*	*13.32*
3	*15.78*	*14.09*
4	*22.63*	*14.90*
5	*20.36*	*15.64*
6	*21.43*	*15.78*
7	*11.74*	*16.09*
8	*22.27*	*16.16*
9	*18.11*	*16.39*
10	*27.50*	*16.66*
11	*18.12*	*18.11*
12	*16.16*	*18.12*
13	*16.39*	*18.71*
14	*16.66*	*20.36*
15	*18.71*	*20.47*
16	*14.90*	*21.43*
17	*14.09*	*22.27*
18	*15.64*	*22.63*
19	*16.09*	*25.25*
20	*20.47*	*27.50*

(Tomado de Statistical Procedures for Análisis of Environmental Monitoring Data & Risk Assessment; McBean & Rovers; PTR-PH)

con la ecuación (2.10) se calcula el promedio: $\overline{X} = \frac{1}{20}(365.89) = 18.29 mg/L$

como el número de datos es par, con la ecuación (2.12) se calcula la mediana:

$$\tilde{X} = \frac{16.66 + 18.11}{2} = 17.39 mg/l$$

y finalmente se calcula la moda, que es el valor mas frecuente. En este caso ningún valor se repite: $Xm = \phi$

B3

Medidas de dispersión

Para los datos de la tabla B2, calcular: rango, variancia, desviación estándar, coeficiente de variación y rango intercuartil.

Con la ecuación (2.11) se calcula el rango

$$\mathrm{R} = 27.50 - 11.74 = 15.76 mg/l$$

Con la ecuación (2.13) se calcula la variancia

$$S^2 = \frac{311.82}{20-1} = 16.41 mg^2/l^2$$

Con la ecuación (2.14) se calcula la desviación estándar

$$S = 4.05 mg/l$$

Con la ecuación (2.15) se calcula el coeficiente de variación

$$CV = \frac{4.05}{18.29} = 0.22$$

con la ecuación (2.16) se calcula el rango intercuartil

rango intercuartil = 20.71 - 15.75= 4.97 mg/l

B4

Medidas de asociación

Se calcula el coeficiente de correlación para los datos de la tabla A3

El coeficiente de correlación se calcula con la ecuación (2.17) y las desviaciones estándar y covarianza con las ecuaciones (2.18) a (2.20) respectivamente.Con el fin de facilitar el cálculo se construye la tabla B4, en la que se obtienen los numeradores de las ecuaciones (2.18) a (2.20)

Tabla B4. Calculo del coeficiente de correlación.

#	Velocidad del viento (MPH)	Corriente de salida (A)	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	5.00	1.582	1.281	0.001	0.031
2	6.00	1.822	0.017	0.045	-0.028
3	3.40	1.057	7.464	0.305	1.510
4	2.70	0.500	11.779	1.231	3.808
5	10.00	2.236	14.961	0.392	2.423
6	9.70	2.386	12.731	0.603	2.770
7	9.55	2.294	11.683	0.468	2.339
8	3.05	0.558	9.499	1.106	3.241
9	8.15	2.166	4.072	0.310	1.123
10	6.20	1.866	0.005	0.066	0.017
11	2.90	0.653	10.446	0.915	3.092
12	6.35	1.930	0.048	0.103	0.070
13	4.60	1.562	2.347	0.002	0.073
14	5.80	1.737	0.110	0.016	-0.042
15	7.40	2.088	1.608	0.229	0.607
16	3.60	1.137	6.411	0.223	1.197
17	7.85	2.179	2.952	0.324	0.978
18	8.80	2.112	7.118	0.252	1.340
19	7.00	1.800	0.753	0.036	0.165
20	5.45	1.501	0.465	0.012	0.074
21	9.10	2.303	8.809	0.481	2.058
22	10.20	2.310	16.549	0.491	2.849
23	4.10	1.194	4.129	0.173	0.844
24	3.95	1.144	4.761	0.217	1.016
25	2.45	0.123	13.557	2.210	5.474
n	$\bar{x}$	$\bar{y}$	$\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
25	6.13	1.610	153.554	10.211	37.029

(Tomado y modificado de Applied Statistics and Probability for Engineers; Montgomery & Runger; John Wiley & Sons)

$$s_x = \sqrt{\frac{153.554}{24}} = 2.529 MPH \qquad s_y = \sqrt{\frac{10.211}{24}} = 0.652 A$$

$$s_{xy} = \frac{37.029}{24} = 1.543 MPHA \qquad \boxed{r = 0.93514}$$

De acuerdo con la tabla 2.1, el valor de r indica que los datos presentan una correlación positiva fuerte.

Anexo C

Pruebas de bondad de ajuste

En este anexo se ejemplifica la aplicación de las pruebas de bondad de ajuste detalladas en el capítulo 2. El procedimiento está integrado por varios pasos que se detallan en los ejemplos, en donde se indican las ecuaciones utilizadas. Como en los anexos anteriores, se utilizó Microsoft Excel para realizar los diversos cálculos como determinación de frecuencias, frecuencias acumuladas, valores de probabilidad normal, despliegue de gráficas, etc.

C1

Prueba de bondad de ajuste de Kolmogorov-Smirnov

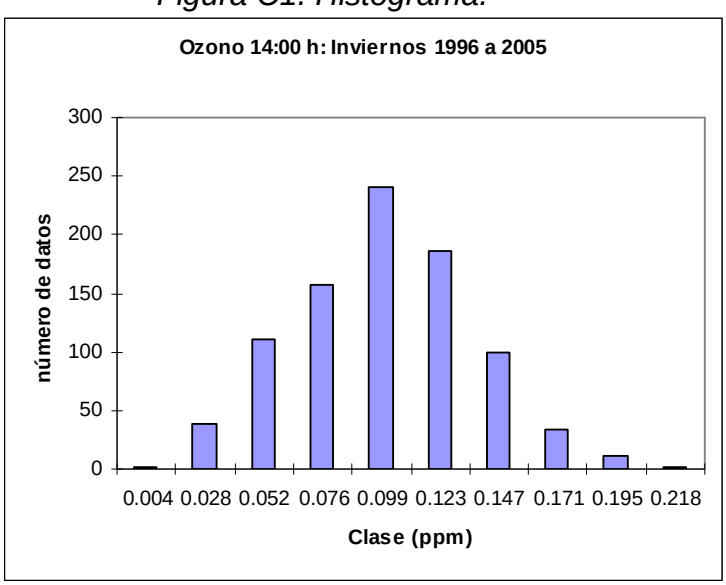
Para datos de monitoreo de O_3 en ppm de las 14:00 h, para los meses enero, febrero y diciembre de 1996 a 2005 de la estación de monitoreo Merced, aplicar la prueba de bondad de ajuste de Kolmogorov-Smirnov y determinar si los datos se ajustan a una distribución normal.

a) Construcción de un histograma de los datos

Tabla C1. Frecuencias.

CLASE		Observado
<	0.0042	1
0.0042	0.0280	38
0.0280	0.0518	111
0.0518	0.0756	157
0.0756	0.0994	240
0.0994	0.1232	186
0.1232	0.1470	99
0.1470	0.1708	33
0.1708	0.1946	12
0.1946	0.2184	1
0.2184	>	1

Figura C1. Histograma.



b) Planteamiento de la hipótesis

(Ho). Los datos provienen de una distribución normal.

(Ha). Los datos no provienen de una distribución normal.

c) Nivel de significancia

Para todo valor de probabilidad igual o menor que 0.05, se rechaza Ho y se acepta Ha.

d) Zona de rechazo.

Para todo valor D' mayor que el valor crítico D, se rechaza la hipótesis nula.

e) Cálculo de las frecuencias observadas acumuladas $(Fobs)$.

Tabla C2. Frecuencias observadas acumuladas

CLASE		Observado	Fobs
<	0.0042	1	1
0.0042	0.0280	38	39
0.0280	0.0518	111	150
0.0518	0.0756	157	307
0.0756	0.0994	240	547
0.0994	0.1232	186	733
0.1232	0.1470	99	832
0.1470	0.1708	33	865
0.1708	0.1946	12	877
0.1946	0.2184	1	878
0.2184	>	1	879

f) Cálculo de las frecuencias normales acumuladas (Ft).

Tabla C3. Frecuencias normales acumuladas.

CLASE		Z_i	Z_s	$F(Z_i)$	$F(Z_s)$	p	Esperado	Ft
<	0.0042		-2.321		0.010	0.010	9	9
0.0042	0.0280	-2.321	-1.668	0.010	0.048	0.038	33	42
0.0280	0.0518	-1.668	-1.014	0.048	0.155	0.108	95	136
0.0518	0.0756	-1.014	-0.361	0.155	0.359	0.204	179	316
0.0756	0.0994	-0.361	0.293	0.359	0.615	0.256	225	541
0.0994	0.1232	0.293	0.947	0.615	0.828	0.213	187	728
0.1232	0.1470	0.947	1.600	0.828	0.945	0.117	103	831
0.1470	0.1708	1.600	2.254	0.945	0.988	0.043	38	868
0.1708	0.1946	2.254	2.907	0.988	0.998	0.010	9	877
0.1946	0.2184	2.907	3.561	0.998	1.000	0.002	1	879
0.2184	>	3.561		1.000		0.000	0	879

En donde

$$Z_i = \frac{\overline{X} - L_i}{S} \qquad \text{(C1)} \qquad , \qquad Zs = \frac{\overline{X} - L_s}{S} \qquad \text{(C2)}$$

L_i límite inferior de clase

L_s límite superior de clase

F(Z) área bajo la curva de distribución normal

p probabilidad normal del intervalo

g) Cálculo del estadístico D'.

Con la ecuación (2.21) se calcula D' para cada par de frecuencias, correspondiente a la última columna de la tabla C4.

Tabla C4. Diferencia máxima D'

CLASE		Fobs	Ft	(Ft-Fobs)/n
<	0.0042	1	9	0.009
0.0042	0.0280	39	42	0.003
0.0280	0.0518	150	136	-0.015
0.0518	0.0756	307	316	0.010
0.0756	0.0994	547	541	-0.007
0.0994	0.1232	733	728	-0.006
0.1232	0.1470	832	831	-0.001
0.1470	0.1708	865	868	0.004
0.1708	0.1946	877	877	0.000
0.1946	0.2184	878	879	0.001
0.2184	>	879	879	0.000

en donde la máxima discrepancia es *D'* = 0.010

h) Comparación del estadístico D' con el valor crítico del estadístico de Kolmogorov-Smirnov (D) (tabla C5).

El estadístico D de Kolmogorov-Smirnov se calcula con la ecuación (2.22)

$$D = \frac{1.358}{\sqrt{879}} = 0.0458$$

$$D' < D$$

i) Aceptación o rechazo de la hipótesis.

De acuerdo con la ec. (2.23)

Como $D' < D$ no se rechaza Ho

Debido a que el estadístico *D'* obtenido es menor que el crítico *D* se acepta Ho y se rechaza Ha; la distribución de los datos se aproxima a la distribución normal.

Tabla C5. Valores críticos del estadístico de Kolmogorov-Smirnov

n	α = 0,1	0,05	0,01
1	0,95000	0,97500	0,99500
2	0,77639	0,84189	0,92929
3	0,63604	0,70760	0,82900
4	0,56522	0,62394	0,73424
5	0,50945	0,56328	0,66853
6	0,46799	0,51926	0,61661
7	0,43607	0,48342	0,57581
8	0,40962	0,45427	0,54179
9	0,38746	0,43001	0,51332
10	0,36866	0,40925	0,48893
11	0,35242	0,39122	0,46770
12	0,33815	0,37543	0,44905
13	0,32549	0,36143	0,43247
14	0,31417	0,34890	0,41762
15	0,30397	0,33760	0,40420
16	0,29472	0,32733	0,39201
17	0,28627	0,31796	0,38086
18	0,27851	0,30936	0,37062
19	0,27136	0,30143	0,36117
20	0,26473	0,29408	0,35241
21	0,25858	0,28724	0,34427
22	0,25283	0,28087	0,33666
23	0,24746	0,27490	0,32954
24	0,24242	0,26931	0,32286
25	0,23768	0,26404	0,31657
26	0,23320	0,25907	0,31064
27	0,22898	0,25438	0,30502
28	0,22497	0,24993	0,29971
29	0,22117	0,24571	0,29466
30	0,21756	0,24170	0,28987
40	0,18913	0,21012	0,25205
50	0,16959	0,18841	0,22604
60	0,15511	0,17231	0,20673
70	0,14381	0,15975	0,19167
80	0,13467	0,14960	0,17949
90	0,12709	0,14117	0,16938
100	0,12067	0,13403	0,16081
200	0,80579	0,09518	0,11411
500	0,54440	0,06030	0,07228
Asint	$1{,}244 / \sqrt{n}$	$1{,}358 / \sqrt{n}$	$1{,}628 / \sqrt{n}$

Tomado de Mc. Bean Edward, Rovers Frank; 1998. Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment; vol 3; Prentice Hall PTR.

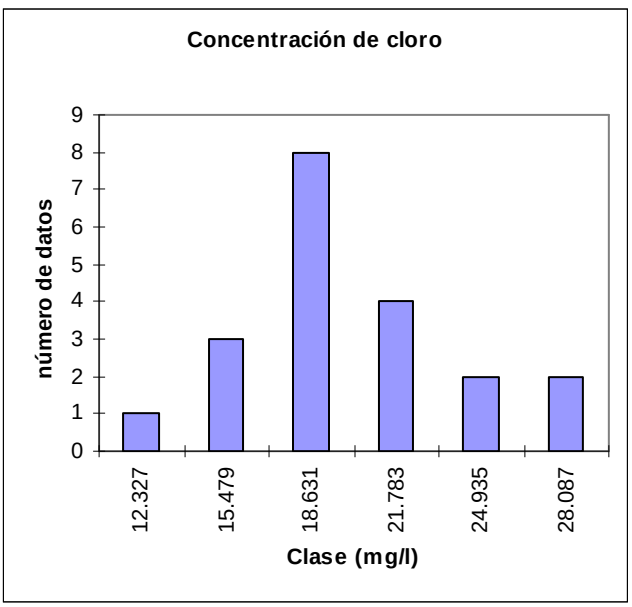
C2

Prueba de bondad de ajuste de Shapiro-Wilk

Para los datos de la tabla B1 del anexo B, se aplica la prueba de bondad de ajuste de Shapiro-Wilk, para determinar si la distribución de los datos se ajusta a una distribución normal.

a) Construcción del histograma de los datos

Figura C2. Histograma.



b) Planteamiento de la hipótesis

(Ho). La muestra de datos proviene de una distribución normal

(Ha). La muestra de datos no sigue una distribución normal

c) Zona de rechazo

A un nivel de significancia del 5%, para todo valor del estadístico W' menor

que el valor critico W se rechaza la hipótesis nula

d) Calculo del estadístico de prueba W'.

Tabla C6. Calculo de b del estadístico W' de Shapiro-Wilk.

I	II	III	IV	V	VI	VII
Numero de muestra	Valor de la muestra	X(i)	X(n-i+1)	X(n-i+1) - X(i)	a(n-i+1)	b(i)
1	25.25	11.74	27.50	15.76	0.4734	7.461
2	13.32	13.32	25.25	11.93	0.3211	3.831
3	15.78	14.09	22.63	8.54	0.2565	2.191
4	22.63	14.90	22.27	7.37	0.2085	1.537
5	20.36	15.64	21.43	5.79	0.1686	0.976
6	21.43	15.78	20.47	4.69	0.1334	0.626
7	11.74	16.09	20.36	4.27	0.1013	0.433
8	22.27	16.16	18.71	2.55	0.0711	0.181
9	18.11	16.39	18.12	1.73	0.0422	0.073
10	27.50	16.66	18.11	1.45	0.0140	0.020
11	18.12	18.11	16.66	-1.45		
12	16.16	18.12	16.39	-1.73		
13	16.39	18.71	16.16	-2.55		
14	16.66	20.36	16.09	-4.27		
15	18.71	20.47	15.78	-4.69		
16	14.90	21.43	15.64	-5.79		
17	14.09	22.27	14.90	-7.37		
18	15.64	22.63	14.09	-8.54		
19	16.09	25.25	13.32	-11.93		b = Σ b(i)
20	20.47	27.50	11.74	-15.76		17.328

$\overline{X}$	18.281
S	4.057

De la ecuación (2.25) $$W' = \left(\frac{17.328}{4.057\sqrt{20-1}} \right)^2 = 0.959$$

e) Comparación del estadístico W' con el valor crítico W de Shapiro-Wilk (tabla C7).

$$W = 0.905$$

$$W' > W$$

f) Aceptación o rechazo de la hipótesis.

De acuerdo con la ec. (2.6)

Como $W' > W$ entonces no se rechaza Ho

Dado que el estadístico *W'* de *Shapiro-Wilk* obtenido es mayor que el crítico W, se acepta Ho y se rechaza Ha. En otras palabras, los datos no muestran evidencia significativa de no normalidad, la distribución de los datos se aproxima a la distribución normal.

Tabla C7. Valores críticos del estadístico de W de Shapiro-Wilk

j \ n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.6134	0.3244	0.3291
3				0.0000	0.0875	0.1401	0.01743	0.1976	0.2141
4						0.0000	0.0561	0.0947	0.1224
5								0.0000	0.0399

j \ n	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.515	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.329	0.3273	0.3253	0.3232	0.3211
3	0.226	0.2347	0.2412	0.246	0.2495	0.2521	0.254	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0927	0.1099	0.124	0.1353	0.1447	0.1109	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.088	0.1005	0.0725	0.1197	0.1271	0.1334
7			0.0000	0.024	0.0433	0.0593	0.359	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0000	0.0496	0.0612	0.0711
9								0.0163	0.0303	0.0422
10									0.0000	0.014

Tomado de Mc. Bean Edward, Rovers Frank; 1998. Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment; vol 3; Prentice Hall PTR.