UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

INGENIERÍA DE SISTEMAS – INVESTIGACIÓN DE OPERACIONES

"DEMANDA DE USUARIOS EN TRENES DE PASAJE URBANO, UN CASO DE ESTUDIO"

TESIS QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN INGENIERÍA

PRESENTA:

Actuario: ALEXEI GÓMEZ-EGUIARTE MARTÍNEZ

TUTOR: Maestro en Ingeniería: FRANCISCO JOSÉ ÁLVAREZ Y CASO

FACULTAD DE INGENIERÍA

MÉXICO, D. F; 2013

JURADO ASIGNADO:

Presidente:	DR. JESÚS ACOSTA FLORES
Secretario:	DR. GABRIEL DE LA NIEVES SÁNCHEZ GUERRERO
Vocal:	M.I. FRANCISCO JOSÉ ALVÁREZ Y CASO
1 er. Suplente:	DR. JUAN MANUEL ESTRADA MEDINA
2 do. Suplente:	M.I. JOSÉ ANTONIO RIVERA COLMENEROS

Ciudad Universitaria; MÉXICO D.F.

TUTOR DE TESIS:

M.I. FRANCISCO JOSÉ ÁLVAREZ Y CASO

_____

FIRMA

AGRADECIMIENTOS

Reconozco en mis padres:

José Manuel Gómez-Eguiarte González (R.I.P.) y Martha Martínez Maguey

Su esmerado amor y cariñosos cuidados.

En mis hermanos, el apoyo que siempre me acompaña.

En mi esposa e hijos, la inspiración y alegría.

En mis cuñadas y sobrinos, que engrandecen el núcleo de mi felicidad.

Me considero orgullosamente un hijo de la

UNIVERSIDAD NACIONAL AUTONÓMA DE MÉXICO

Alma máter, institución a quién debo mi formación y el honor de haberme matriculado.

Agradezco a mis profesores en esta noble institución y particularmente a mi tutor:

M.I. Francisco José Álvarez y Caso, quién tuvo la percepción y talento para guiar mis afanes y concretarlos.

Agradezco a los miembros del jurado, su paciencia y bonhomía.

Soy deudor de muchas personas que contribuyeron a lograr éste trabajo, a todos ellos, amigos míos: ofrezco mi gratitud por aportar sus conocimientos, consejos, y opiniones.

Contenido

Nombre de la tesis:

"Demanda de usuarios en trenes de pasaje urbano, un caso de estudio"

Objetivo General:

Se proyectó ofrecer un análisis de la demanda o flujo de pasajeros en una línea de trenes urbanos y modelar matemáticamente el comportamiento aleatorio de éste requerimiento. Se espera que la información generada pueda ser considerada por la administración para optimizar sus recursos y principalmente mejorar el servicio que otorga a los ciudadanos.

Objetivos específicos

1. Determinar la demanda diaria de usuarios en una línea del transporte metropolitano de trenes mediante una muestra representativa.
2. Con base en el resultado anterior, obtener un modelo matemático que permita desarrollar estrategias de investigación de operaciones para optimizar el servicio de trenes.

Hipótesis:

1. El componente principal a conocer en un sistema de transporte es la demanda de usuarios.
2. La demanda de pasajeros de una línea metropolitana de trenes puede modelarse matemáticamente utilizando técnicas del análisis de sistemas apoyándose en herramientas de cómputo.
3. Se asume que la demanda o flujo de pasajeros tiene un comportamiento aleatorio que puede modelarse rigurosamente.

Síntesis

El sistema de trenes metropolitanos referido en el presente trabajo es el más eficiente y popular transporte de pasajeros en el Valle de México.

En cualquier sistema de transporte, la afluencia de pasajeros define condiciones para la planeación de servicios al usuario. Siendo el flujo (demanda) de pasajeros el dato más relevante, se analizó la demanda diaria para obtener un modelo descriptivo de una línea particular del sistema. Tal representación coincide con la función densidad de probabilidad (FDP) de una variable aleatoria (VA); el conocimiento de la FDP de la VA otorga una descripción completa de la misma. Extraer la FDP mediante información proporcionada por una muestra de pasajeros en flujo es el problema que se pretendió resolver.

Se eligió una línea de trenes sin estaciones de transferencia para contabilizar la demanda de manera eficaz. Él flujo de pasajeros se midió mediante conteos automatizados realizados por los dispositivos de acceso al sistema.

Para alcanzar el objetivo se obtuvieron y analizaron conteos correspondientes a la demanda de 2010 y se introdujeron en una base de datos compatible con software estadístico. Mediante análisis preliminar y particionando la muestra en clases mutuamente excluyentes se obtuvo la función de distribución empírica (FDE) para cada clase en ésta línea. Utilizando software especializado, la demanda caracterizada por la FDE, es ajustada por Máxima Verosimilitud a modelos FDP teóricos validados mediante pruebas de bondad de ajuste.

Los modelos obtenidos son de tipo continuo, con soporte en los reales positivos y son funciones de densidad sesgadas que pertenecen a la familia de distribuciones de valor extremo y a la familia de distribuciones logarítmicas, entre otros.

Las FDP obtenidas sirven para calcular la probabilidad de eventos de interés y generar números aleatorios que siguiendo la distribución propuesta permiten obtener una visión muy próxima al comportamiento real del flujo de pasajeros.

Abstract

The metropolitan rail system referred to in this manuscript is the most efficient and popular passenger transport in Mexico's Valley. In any system of transportation, ridership flow planning, defines conditions for user services. Being the flow of passengers the most relevant interest, we will analyze the daily demand to get for a descriptive model of the system for a particular line. This representation coincides with the probability density function (PDF) of a random variable (RV); knowledge of the PDF on the RV provides a complete description of it. Extract the FDP by information provided by a sample of passengers flow is the problem to be solved. We choose a rail line without transfer stations for to account effectively passengers demand. Passenger traffic is measured by automated counting devices made by system access. To achieve the goal of collecting and analyzing demand counts for 2010, we get outputs (data) and it were entered into a database compatible with statistical software. By preliminary analysis and partitioning the sample into mutually exclusive classes, we get the empirical distribution function (E-CDF) for each class in this line. Using specialized software, demand characterized by E-CFD is adjusted for maximum likelihood. The achieved FDP theoretical models are validated by goodness of fit tests. Those FDP models are of continuous type, with support on the positive real line. They are density functions, biased, that belongs to the family distributions of extreme value, logarithmic family, and others.

The FDP obtained will be used to calculate the probability of events of interest and generate random-numbers. When they follow the proposed distribution, this will allow obtaining a very close insight to the actual behavior of passenger flow in this metropolitan rail-line.

Capítulo 1

Introducción

El presente trabajo se ha realizado con la intención de hacerlo accesible a un público amplio y no necesariamente especializado en el tema, esta condición asociada al uso intensivo de metodologías gráficas pudiera explicar la extensión de éste documento. En los anexos se presentan elementos teóricos y funcionales que pudieran ser útiles al lector.

El sistema de trenes metropolitanos al que se refiere el presente trabajo corresponde a un organismo del Gobierno del Distrito Federal; el *sistema* (STC, Sistema de Transporte Colectivo Metro), es desde hace más de 40 años la columna vertebral del transporte masivo y popular en la Ciudad de México y los municipios conurbados del Estado de México.

En cualquier sistema de transporte, público o particular, la afluencia de pasajeros a los medios de transferencia define la manera en que los gobiernos realizan la planeación del servicio que ofrecen a los usuarios. Habitualmente se implementa una planeación de los servicios de transporte en función de una demanda pronosticada, dicha planeación depende de algunas variables ya identificadas como son: las alternativas al transporte público, la infra-estructura de carreteras y caminos, la implementación de políticas de tenencia de la tierra y el análisis de impacto ambiental junto con estudios socio-demográficos. [8] Sin embargo, aún la planeación más cuidadosa de un sistema de transporte colectivo, puede verse rebasada cuándo existen fuertes fenómenos de migración y ocupación irregular del suelo.

El crecimiento poblacional se explica en la Zona Metropolitana del Valle de México, (ZMVM) debido a que existe en ésta un amplio mercado de trabajo y gran cantidad de oferta educativa. Ambas atraen significativamente el desplazamiento de personas del Estado de México hacia la capital. [9] Dicho desplazamiento de personas dispara la demanda de todo tipo de servicios, incluido el transporte. En tal caso, los servicios se vuelven insuficientes y se saturan, en el transporte público esta saturación se refleja en aglomeraciones y pérdida de capacidad para movilizar a los pasajeros, especialmente dentro del mayor medio de transporte masivo en México: el sistema. Éste ha ocupado hasta la tercera posición mundial con respecto al número de usuarios transportados por día; [7] ante semejante relevancia estratégica, se necesita obtener un modelo que describa la demanda de pasajeros bajo las condiciones de operación que correspondan a las características que actúan sobre él.

El presente trabajo se propuso analizar el comportamiento de la demanda de usuarios, en trenes "sub-urbanos" utilizando datos de conteo de afluencia diaria.

1.0 Antecedentes: Historia breve del metro de la Ciudad de México

El metro de la Ciudad de México es un sistema de transporte público que sirve a extensas áreas del Distrito Federal y parte del Estado de México. Su operación y explotación está a cargo del organismo público descentralizado denominado Sistema de Transporte Colectivo, (STC, el *Sistema*) y su construcción está a cargo del Proyecto Metro del Distrito Federal, un organismo desconcentrado perteneciente a la Secretaría de Obras y Servicios del Distrito Federal y se constituyó:

"*Por Decreto publicado en el Diario Oficial de la Federación el 29 de abril de 1967, se instituye el organismo público descentralizado con personalidad jurídica y patrimonio propios, denominado Sistema de Transporte Colectivo…" y se estableció como "…la construcción, operación y explotación de un tren rápido con recorrido subterráneo y superficial para el transporte colectivo de pasajeros en la ZMVM y del Estado de México…*" [38]

La construcción inició el 19 de Junio de 1967, en Línea 1, con un tramo planeado de 12.660 km; y 16 estaciones. En el año 1969 se delineó para la Ciudad de México una red de metro como eje articulador del sistema de transporte de la ciudad. Dicho plan tenía tres líneas que se construirían de acuerdo a la evolución de la demanda. El 4 de septiembre de 1969 comenzó a funcionar el primer tramo Zaragoza-Chapultepec, con 16 estaciones inauguradas correspondientes a la línea 1. A partir de Abril de 1970, la línea 1 se extendió hasta la estación Juanacatlán y más adelante se inauguraron los tramos Juanacatlán-Tacubaya y Tacubaya–Observatorio, el 10 de Junio de 1972. Le siguieron en orden cronológico las líneas dos y tres inauguradas en Agosto y Noviembre de 1970. Posteriormente en Agosto de 1991, el gobierno anunció la inauguración de la Línea Oriente que corre desde las delegaciones Iztapalapa e Iztacalco, al Oriente de la Ciudad de México hacia los municipios de Los reyes y La Paz en el Estado de México, esto junto a un nuevo plan de inversiones para solucionar los problemas de transporte urbano en la ciudad.

Esta inversión, consistió en la reorganización del plan de transporte público a través de un sistema integrado que incluye los Centros de Transferencia Modal *(CETRAM)*, conocidos coloquialmente como "paraderos", estos son espacios en donde confluyen diversos vehículos

y rutas de transporte de pasajeros. Su objetivo es facilitar el movimiento de personas entre los sistemas de transporte que allí convergen. En la mayoría de los CETRAM las líneas de autobuses provienen de la ZMVM.

De los 45 CETRAM existentes en la Ciudad de México, 37 están ubicados en estaciones terminales y de mayor afluencia en el Metro de la Ciudad de México.

Como articulador del entonces nuevo sistema de transporte, llamado "*sistema troncal*" el metro tuvo y conserva un rol primordial, por lo que se consideraron inversiones importantes para mejorar y extender la red. Las inversiones en la red del metro, en el corto plazo, consistieron en extender las líneas 5 y 9 a fin de conectar al Norte y centro de la capital mediante la Línea Oriente y ofrecer el servicio a los municipios de Netzahualcóyotl, Ixtapaluca, Chalco y Chimalhuacán -en el antiguo vaso de Texcoco- y proporcionar mano de obra a la zona industrial de Azcapotzalco, Vallejo e incluso Tlalnepantla en el Norte-Poniente de la Ciudad. Las extensiones de las Líneas 1 y 5 se encuentran en operación desde Diciembre de 1981 y Agosto de 1987 respectivamente. [35]

A nivel económico, el aumento del ingreso personal crea aspiraciones para obtener una calidad superior de transporte urbano, en términos de velocidad, seguridad y efectividad, esto generó un desafortunado incremento en el parque vehicular, cuya consecuencia fue el aumento del tráfico automotriz y la contaminación por gases de efecto invernadero, éstas dos situaciones, que ahora parecen insostenibles hicieron que se inaugurara en Agosto de 1991 la Línea Oriente del Sistema como alternativa de transporte urbano.

La línea de pasajeros a la que se avocó el presente trabajo es una de transporte férreo que se conduce por vía externa; y es la Línea Oriente, arriba mencionada, la novena línea del sistema en ser inaugurada. La Línea Oriente, está integrada por 10 estaciones y su trazo se localiza al Sur- Oriente de la Ciudad de México con dirección predominante Oriente-Poniente.

Esta línea corre desde el Oriente de la Ciudad y conecta entidades de alta densidad poblacional como son los municipios de Netzahualcóyotl, Chalco, Ixtapaluca y otros con la parte Oriental del Distrito federal, las delegaciones de Iztapalapa e Iztacalco principalmente. [9][H]. Dicha línea tiene una longitud de vía de 17.192 kilómetros, de los cuales 14.893 kilómetros son utilizados para el servicio de pasajeros y el restante se emplea para maniobras.

Se distingue del resto de las líneas por tener trenes de 6 carros de rodadura férrea alimentados por catenaria En ferrocarriles se denomina "catenaria" la línea aérea de alimentación eléctrica que transmite energía a una locomotora o cualquier elemento de tracción. Por ese motivo también se le conoce como *metro férreo* o *metro ligero* [42]

La razón para utilizar esta línea como sujeto de estudio corresponde a que las mediciones de afluencia se obtienen de manera directa a partir de los controles en los dispositivos de entrada y debido a que en esta línea no existen estaciones de transferencia, es decir no hay conexión con otras líneas salvo en una de las terminales.

Así, el flujo de pasajeros no se ve distorsionado por la probabilidad de que "k" pasajeros provenientes de "q" posibles líneas, transborden en la línea de interés, digámosle *Línea O*.

Los recursos que sustentan el funcionamiento del metro provienen de las partidas presupuestales otorgadas por la federación al gobierno del D.F y de la propia administración del Sistema y son estos fondos los que permiten subsidiar el costo del pasaje a la población de usuarios de éste medio de transporte. Debido a que el número de pobladores de la ZMVM, es creciente y debido al surgimiento de nuevas zonas urbanas, es que la metrópoli Mexicana continúa en expansión, gran parte de esta población demanda servicios de transporte efectivos que puedan satisfacer sus necesidades de movilización.

Los sistemas de trenes urbanos son caros y frecuentemente constituyen el mayor gasto de recursos públicos de un gobierno, aunque la experiencia internacional ha demostrado que no siempre están localizados en los mejores corredores de transporte, de ésta manera en el caso que nos ocupa, puede decirse que las políticas de planeación sobre uso del suelo y diversificación del transporte urbano han traído como consecuencia que el servicio que proporciona el Sistema se vea cada vez más demandado y que enfrente problemas de operación, a pesar de las recomendaciones que hacen los organismos internacionales de financiamiento. [2]

Aun así, este tipo de transporte tiene una alta aceptación y demanda de usuarios, no solo porque es barato sino porque ha sido el más seguro y eficiente medio de transferencia en la Ciudad de México. En efecto el presente trabajo se enfoca exclusivamente a una línea de trenes que presta servicio al Oriente de la ZMVM, y no a la red general del Sistema.

1.1.- La demanda de usuarios

Con el Sistema funcionando, se han manifestado problemas en el servicio de los trenes y son de diversa índole; aglomeraciones en su infraestructura, tiempos de traslado crecientes e inseguridad. Sin embargo, éstas y otras fallas que van desde la iluminación en las estaciones hasta descarrilamientos no siempre han podido ser anticipadas tal vez por falta de presupuesto, falta de estudios en la materia, etc.

Un problema inmediato por resolver, en el caso de las aglomeraciones y tiempos de traslado, es determinar el comportamiento de la demanda de usuarios en la Línea O. Existen diversas formas para analizar el comportamiento de la demanda por ejemplo, son comunes los modelos de tasas por viaje, (TRM) por sus siglas en inglés; donde la tasa de generación de viajes se define como un índice estadístico calculado a partir de los recuentos realizados en los sitios de estudio. También se dispone de modelos de demanda directa llamados "modelos agregados simultáneos" (ASM) cuya contraparte es el modelo de elección desagregado (MCM). [13] [27]

Estas son aproximaciones desde el punto de vista de la teoría de elección y comportamiento ante opciones de consumo, en tal caso el enfoque se encuentra dentro del área econométrica. Un enfoque alternativo consiste en levantar encuestas donde se le pregunta a las personas las veces que utilizan un determinado tipo de transporte, lo cual es llamado "*intenciones establecidas*" (SI). Sin embargo éste último tratamiento es arriesgado ya que una encuesta que no esté finamente ajustada puede llevar a subestimar ó sobre-estimar los datos de uso en transporte.

Ambas técnicas (elección e intenciones) consisten en análisis de paneles, que son una forma popularmente creciente de análisis de datos entre los investigadores de ciencias del comportamiento. Un "panel" es una sección cruzada o grupo de personas a quienes se les aplican encuestas periódicamente sobre un intervalo de tiempo dado; además existen alternativas que se proponen en el campo de la ingeniería electrónica y de sistemas de cómputo, entre otras.

El Sistema proporciona un medio de transporte que se hace indispensable para el funcionamiento de la ZMVM y que a falta de este servicio público la ciudad y sus habitantes se

verían envueltos en el caos que significa la falta de movilidad, por la cual millones de usuarios faltarían o retrasarían su acceso hacia sus ocupaciones cotidianas.

A pesar de su importancia, hasta la fecha no se conoce una aproximación del comportamiento en la demanda diaria de usuarios, tampoco del costo en "horas hombre" que representan los retrasos en el transporte público. Teniendo en cuenta que actualmente el sistema cuenta con un subsidio importante y que el costo del boleto ($3.00 M.N.) representa sólo una fracción –aproximadamente un tercio- del costo real por pasaje; sería en el campo de la administración pública donde pueda obtenerse una función costo. Debido precisamente al carácter público del Sistema cualquier elemento que ayude a mejorar el servicio se traduce en bienestar para la población. Sin embargo, un mejor servicio representa cumplir uno de los objetivos principales de la administración pública, el cual es servir con eficiencia al usuario mediante la aplicación de los recursos públicos.

Por tal motivo, un paso importante para desarrollar prácticas de eficiencia es conocer el comportamiento de la demanda de usuarios. Esta investigación sobre los flujos de usuarios en la red de transporte del Sistema puede utilizarse para mejorar el funcionamiento y el servicio, *tomando en cuenta que actualmente no se encuentran disponibles los estudios que permitan conocer el flujo de pasajeros.*

Tal pesquisa deberá ser válida para que el Sistema mejore su operación y servicio en función de afirmarse como usuario y generador de tecnologías de ingeniería, información e investigación de operaciones.

Esta investigación se genera como respuesta a una problemática observada y como parte de una necesidad de hacer eficiente el servicio de transporte público que se ofrece a los habitantes de la ZMVM.

1.2.- Métodos para estimar la demanda de usuarios

El estudio del flujo de pasajeros en transporte juega un rol esencial en las sociedades desarrolladas, el transporte es responsable de la movilidad personal y provee acceso a servicios, actividades económicas y de ocio. Debido a la enorme influencia del transporte en la sociedad moderna, los problemas relacionados a éste exigen la utilización de muy diversas disciplinas, que se relacionan entre sí de formas intrincadas. Por ejemplo la complejidad, diversidad y naturaleza aleatoria del problema requieren el uso de diferentes herramientas,

en cuyo caso, la adecuación de éstas en la parte de investigación y análisis final dependen en gran parte del tipo de problema por resolver y de la selección adecuada de las técnicas empleadas en función de los recursos disponibles.

Se proponen diversas soluciones desde diferentes puntos de enfoque, por ejemplo los modelos "probit", son modelos descriptivos de elección binaria y de elección discreta. En estos el conjunto de elección se reduce a solo dos alternativas posibles mutuamente excluyentes y se registran mediante cuestionarios que posteriormente reciben un tratamiento de análisis estadístico.

También existen acercamientos de la ingeniería eléctrica y electrónica que incluyen dispositivos de conteo por medio de sensores con rayos infra-rojos y de otro tipo que permiten evaluar el comportamiento de "bajadas" y "subidas" de usuarios a un medio de transporte.

En este caso, el detalle de las mediciones puede llegar a ser tan preciso como contabilizar el flujo de pasajeros inclusive puerta por puerta. Además, estos dispositivos –contadores- permiten discriminar el sentido en se mueven los usuarios y de ésa manera el conteo de personas que suben o bajan del transporte se lleva a cabo "*in situ*" lo que permite modelar de manera instantánea el comportamiento del flujo de pasajeros.

En otras latitudes se utilizan las tarjetas inteligentes, que son dispositivos electrónicos de complejidad diversa. Estas registran diversos datos del usuario, y dependiendo de la complejidad de la tarjeta, es que se permite llevar solamente un control de entrada-salida hasta producir una descripción completa de las rutas, horarios y frecuencia de los movimientos de un individuo por una red de sistemas de transporte. En ésta situación, la demanda de pasajeros se mide en forma diaria mediante criterios diversos que se encuentran dependientes del objetivo de investigación. De cualquier manera, las diferentes metodologías de solución están acotadas por la disponibilidad de recursos financieros para aplicar la tecnología y procedimientos más adecuados en cada circunstancia.

1.3.- Estudios sobre la estimación de la demanda en transporte

La demanda por servicios de transporte posee características que la diferencian claramente de la demanda de otros bienes y servicios. El primer elemento a destacar es su carácter

derivado, pues, no se demanda viajar “per se”, sino que se hace con el objetivo de realizar una actividad localizada en el espacio y en el tiempo.

En el contexto de los viajes interurbanos también es posible detectar componentes estacionales en el comportamiento de la demanda. Los modelos de demanda desagregados constituyen una herramienta de análisis adecuada para abordar el problema de modelar la demanda de transporte, tales modelos se basan en el análisis del comportamiento del consumidor individual dentro del marco de la microeconomía de las elecciones discretas y de la teoría de la utilidad aleatoria, donde se aborda el problema de modelar la demanda en un contexto de alternativas, para ello es preciso tener en cuenta los aspectos relevantes que rodean el proceso de toma de decisiones de los individuos.

Por esta razón, su aplicación no sólo se extiende dentro del ámbito de la economía del transporte sino en cualquier contexto relacionado con la economía de las elecciones discretas [S].

Entre las ventajas que estos modelos presentan se destaca la posibilidad de realizar un análisis desagregado de elementos tan importantes como las elasticidades de la demanda y el valor subjetivo del tiempo de los viajeros. Las fuentes de información comunes para los modelos desagregados son las *preferencias reveladas* y las *preferencias declaradas*. Las primeras se basan en las elecciones efectivamente realizadas por los individuos; y aportan información acerca de la importancia relativa de las distintas variables que influyen en su decisión. Las preferencias declaradas también capturan esta misma idea pero a diferencia de las preferencias reveladas, se basan en la construcción de escenarios hipotéticos que son presentados al consumidor para que indique su elección. Sin embargo, cuentan con el inconveniente de que no siempre los individuos hacen lo que declaran que van a hacer, amén del oneroso gasto que implica realizar encuestas micro económicas regulares.

En la ciudad de Seúl, estudios llevados a cabo por la Universidad Pohang de Ciencias y Tecnología, en Corea del Sur, se dirigen a la estructura de redes del metro. El enfoque está en las propiedades estadísticas y consecuencias topológicas del sistema metro. Se obtienen varias medidas que incluyen la longitud de las rutas, el radio (coeficiente de agrupación poblacional) y la eficiencia de la red. La longitud de la ruta, el radio, así como la eficiencia son calculadas en términos de la distancia física entre estaciones. El flujo de pasajeros en el sistema se analiza construyendo el árbol de máxima expansión de los flujos.

Esto se logra porque en el metro de Seúl operan tarjetas inteligentes que mantienen un registro de la información de la ruta de viaje de cada pasajero. Las tarjetas inteligentes (tarjetas de circuitos integrados) son usadas en todo el sistema coreano de transporte público y registran el lugar de partida, estaciones de llegada así como fecha y hora. La base de datos de operaciones de tránsito, contiene hasta 10, 000, 000.0 de transacciones por día y contiene también información de tiempo/posición del viaje de cada pasajero, de manera que es posible rastrear sus movimientos individuales, debido a que cada tarjeta tiene su propia identificación (ID).
Se analizan los flujos de viaje por día basándose en datos de las transacciones efectuadas en un día representativo; en el caso mexicano es imposible realizar esto porque el sistema no cuenta con una forma de identificar cual es la ruta que sigue un usuario durante su traslado por la red del metro.

Para Corea por ejemplo, en el día 24 de Junio de 2005, el flujo total de pasajeros en el metro de Seúl fue de 4, 909,316 viajes; y se encontró que la distribución de peso despliega una ley de potencia conductual, donde la fuerza de la distribución sigue una distribución Log- Normal Uno con pico en desviación estándar de aproximadamente 4×10^4 pasajeros en dicho día. [C]

También se observó que en el metro de Seúl, el peso de una liga que conecta dos estaciones, representa el flujo de pasajeros entre ellas y que la fortaleza de una estación corresponde al número de pasajeros que llegan y parten desde ésa estación. En la metrópolis coreana, la mayoría de los servicios están localizados cerca de las estaciones así que cada estación es naturalmente abundante en pasajeros, el número de ellos refleja el grado de servicios localizados alrededor de la estación. El hecho de que la mayoría de las estaciones sean usadas por un número similar de pasajeros, corresponde a los picos en la distribución de fortaleza, lo cual indica que con servicios comerciales y residenciales tomados en cuenta los lugares más cercanos a las estaciones son rápidamente desarrollados para acomodar densas y compactas comunidades.

El pico de la distribución de fortaleza da una medida acerca de la capacidad característica de los servicios en un Seúl totalmente urbanizado. También se revela el comportamiento de la ley de potencias conductual en el grado de distribución del árbol de expansión. Una ley de potencias es una relación matemática entre dos cantidades; en estadística, si estas dos cantidades son la variable aleatoria y su frecuencia, entonces en una distribución de ley de potencias las frecuencias decrecen según un exponente cuando la variable aleatoria aumenta.

En teoría de grafos, un árbol de expansión (T) de un grafo conexo, no dirigido (G) es un árbol compuesto por todos los vértices y algunas (quizá todas) de las aristas de G. El árbol de expansión de G es una selección de aristas de G que forman un árbol que cubre todos los vértices. Esto es, cada vértice está en el árbol, pero no hay ciclos. Además todos los puentes de G deben estar contenidos en T. Un árbol de expansión o árbol "recubridor" de un grafo conexo G puede ser también definido como el mayor conjunto de aristas de G que no contiene ciclos, o como el mínimo conjunto de aristas que conecta todos los vértices.

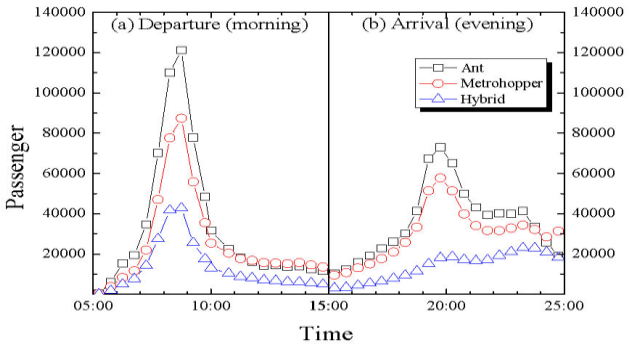
Para clarificar sus observaciones, los científicos coreanos caracterizaron tres tipos de pasajeros con base a la fábula de Esopo, "La hormiga y la cigarra", tales caracteres fueron, hormigas, cigarras y un híbrido entre ambos, los metro-cigarras. Las hormigas representan al grupo de pasajeros madrugadores y que temprano se retiran a casa y las cigarras son quienes disfrutan de la noche en Seúl y dedican tiempo a la diversión, existiendo seres híbridos entre ambas "especies" que merodean en horarios diversos.

Se investigaron las distribuciones de tiempo inter-partida e inter–llegada, así como las distribuciones de horarios de partida de los primeros viajes y los horarios de llegada de los últimos viajes para las tres especies. Todas ellas muestran un comportamiento similar en distribuciones de tiempo de partida, lo cual indica que salen de casa en horario similar por la mañana.

De cualquier forma, la distribución de llegada para los metro-cigarras muestra dos picos, mientras que para las hormigas solamente muestra un pico y un hombro. El segundo pico para los híbridos, mientras regresan a casa en metro es más notable que para los metro-cigarras, quienes toman transportes diferentes al metro.

En la *gráfica 1* siguiente se muestran diferencias en la distribución de partidas por las mañanas; se observa que en la tarde un gran número de personas (los metro cigarras e híbridos) visitan las áreas de diversión en diferentes distritos, lo cual despliega distribuciones de llegada más densas.

Gráfica 1



Se realizaron trabajos posteriores cuya meta fue hallar una "ecuación maestra" que describiera la evolución de los pasajeros en el metro. Se obtuvo una "tasa de transición", construida a partir de consideraciones geográficas, en analogía con redes neuronales artificiales (RNA); lo cual explica un sistema de interconexión de neuronas ó "centros" en una red que colabora para producir un estímulo de salida ó resultado. Una RNA se compone de un conjunto masivamente paralelo de unidades de proceso muy simples y es en las conexiones entre estas unidades donde reside la "inteligencia" de la red.

La ecuación de evolución para la distribución de pasajeros en el metro de Seúl se encuentra entre distribuciones sesgadas que incluyen la Weibull y Log-Normal. Estas aproximaciones se aplicaron al sistema metro de Seúl. El análisis de los datos de viaje de los pasajeros, en la mayoría de los caso se ajusta bien a una distribución log-normal pero sin mencionar los parámetros de escala y localización.

En el caso del transporte público de la Ciudad de México se tiene pensado insertar la tarjeta inteligente paulatinamente a medida que se soluciona el problema laboral de eliminar ó reubicar al personal que expende boletos en taquilla; a pesar de que se tiene planeado utilizar la tarjeta inteligente a partir de 2012 todavía no se conocen las características de la tarjeta inteligente mexicana para su uso en el metro y el metro bus.

Mientras inicia la era de la tarjeta inteligente, en el metro de la Ciudad de México únicamente se cuenta con los datos de entrada que se obtienen de forma directa en cada estación mediante los dispositivos de entrada colocados en el andén de ingreso de las estaciones. Partiendo de que el ingreso de los pasajeros a las estaciones, puede representarse mediante el análisis de los datos recolectados en estas, es que el presente trabajo estudió como describir el comportamiento de los usuarios por día.

Otras alternativas sugieren el uso de detectores de personas instalados en las puertas de acceso, de tal forma que mediante tecnología de rayo infra-rojo, rayo láser o rayos X, se puede contabilizar el número de pasajeros que aborda o desciende en cada estación con una confianza del 98%. El sistema está conectado vía fibra óptica con bases de datos que registran diversos datos de operación del sistema de transporte, las mediciones de flujo de pasajeros se registran y almacenan en tiempo real y la cantidad de usuarios es monitoreada a cada instante y con un detalle que puede especificar inclusive la puerta donde se realiza la transacción.

Esta tecnología se ofrece como un paquete que incluye el dispositivo de conteo a colocarse en el material rodante, su instalación y puesta a punto, así como el software estadístico para modelar el flujo de pasajeros. El software se ofrece con tecnología digital en protocolo de internet (IP) y es tecnología Estadounidense asociada con capital Británico. [U]

Desafortunadamente, este sistema es costoso y tal vez no sea apropiado para el metro de Ciudad de México, debido a que si bien existe la infraestructura de fibra óptica en el Sistema, ésta todavía no está completamente instalada y si lo está, no necesariamente está en servicio.

Es de observar que en todos los casos, excepto el contador con rayo infra-rojo, los flujos de pasajeros se describen por día.

1.4.- Técnicas propuestas para estimar la demanda de usuarios

Si bien existe la disponibilidad de acceso a las diferentes estaciones del sistema, la medición del flujo de usuarios en algunas de ellas representa una labor ardua de muestreo limitada por diversos factores. Por ejemplo, existe el factor de temporalidad, este factor se debe a que las fluctuaciones de personas que ingresan por las estaciones ocurren de manera discontinua y dichos flujos difieren por horario y locación en la ciudad; además, establecerse en las estaciones del Sistema simultáneamente para realizar el muestreo de flujo constituye un gasto oneroso de recursos que no es posible devengar.

La ausencia de conocimiento e información sub-determina la efectividad de la planeación práctica en el sistema. Cualquier proyecto de planeación necesita del conocimiento de la demanda de usuarios; esto es fundamental para pronosticar los efectos y para llevar a cabo los controles con la diligencia adecuada.

Indiscutiblemente, pese a que no se tienen recursos suficientes para realizar el análisis mediante un muestreo, existen métodos alternativos para obtener datos confiables que permitan realizar un estudio de flujo de pasajeros. Por tal motivo se seleccionó una línea en particular para su análisis, tal línea constituye un segmento del Sistema pero es representativa en el aspecto de que recoge una gran cantidad de usuarios y posee propiedades que facilitan su investigación. Las herramientas empleadas para abordar el problema fueron la estadística e investigación de operaciones.

El trabajo presente comprende cubre técnicas diversas de análisis, se llevó a cabo un estudio general y después utilizar las técnicas más adecuadas en función de los equipos utilizados y de los resultados hallados por la investigación inicial. Así se modeló el comportamiento de la cantidad de personas que ingresan a una Línea, *por día* tal como se hace en otras latitudes, mediante el uso de los datos obtenidos por los dispositivos de acceso y que son conocidos en México con el nombre de *torniquetes.*

En particular, se esperó obtener el desarrollo y selección de hipótesis útiles de investigación y utilización de metodologías alternativas señalando sus bondades e inconvenientes. Estas metodologías se utilizan en virtud del tipo de información obtenida y porque son las herramientas con las que fue posible realizar un análisis. En muchas formas, la carencia de recursos implica allegarse de medios alternativos, que ya estén disponibles.

Para realizar el trabajo se analizaron los registros que posee el Sistema sobre el número de usuarios que ingresa a Línea O del Sistema, el registro se obtiene automáticamente cuando el usuario ingresa a la estación donde necesariamente debe atravesar los torniquetes. Cada vez que el torniquete gira se obtiene un conteo que indica el número de personas que han ingresado por ése dispositivo, los torniquetes tienen un solo sentido de conteo (el de ingreso) y personal del Sistema levanta día con día los registros para enviarlos a un departamento donde se les da un tratamiento informático. Puesto que la mayoría de usuarios ingresan por los torniquetes y éstos no "regresan", los conteos son una muy buena aproximación del número de personas (con pasaje cubierto) que ingresa a Línea O.

Se tiene información acerca de estudios que lleva a cabo el sistema, éste realiza análisis mediante series de tiempo con las lecturas de los torniquetes, mediante el uso del paquete estadístico SPSS (mr.) pero por diversas razones, no se conoce el resultado de ésta información.

Tampoco se conocen estudios base precedentes que aproximen la afluencia de usuarios en trenes urbanos que se apoyen en la simple medición de lecturas de los torniquetes. Puesto que no se cuenta con dispositivos de medición hechos ex profeso, como en otros países (Japón o Corea), los datos útiles serán los que están disponibles, así que el trabajo carece de antecedentes.

Los métodos utilizados en la búsqueda del conocimiento generalmente son similares en diversas áreas del saber, estos involucran el reconocimiento y formulación del problema, la recolección empírica o experimental de datos relevantes y frecuentemente el uso de análisis matemático o estadístico para explorar relaciones entre los datos o para verificar hipótesis acerca de las observaciones. La complejidad y ambigüedad del comportamiento humano crean la necesidad de emplear herramientas de la matemática y análisis complejos para investigar tales pautas.

Los sistemas de transporte no consisten únicamente de elementos físicos y organizacionales que interactúan unos con otros para producir oportunidades de transportación, sino también de una "demanda" que se aprovecha de tales oportunidades para viajar de un lugar hacia otro. Esta demanda de viajes, es el resultado de interacciones entre las varias actividades económicas y sociales localizadas en una cierta área. Los modelos de sistemas de transporte representan, para un sistema real o hipotético, el funcionamiento de los elementos físicos y de organización, las interacciones entre ellos y sus efectos sobre el mundo externo.

Dichos modelos se aplican a sistemas reales a gran escala y son las herramientas fundamentales para evaluar y/o diseñar acciones que afecten a los elementos físicos (por ejemplo una línea de trenes) y a los componentes de organización (por ejemplo, una tabla de horarios de llegada) en los sistemas de transporte.

La demanda de viajes se deriva de la necesidad de acceder a servicios urbanos en diferentes lugares y está determinada por la ubicación de las viviendas y las actividades económicas que se desarrollan en cierta área. Los habitantes de aquellas realizan elecciones de "movilidad" y "elecciones de viaje", estas elecciones resultan en flujos de pasajeros que demandan viajar. Los viajes se realizan por diferentes propósitos y en diferentes períodos de tiempo, en diversos medios de transporte. Para llevar a cabo el trabajo se analizaron los registros sobre el número de usuarios que ingresa a Línea O del Sistema. Este examen de los registros fue la base para obtener un modelo de la demanda

Se analizaron observaciones que han sido generadas de acuerdo a un experimento aleatorio ó ley probabilística de cuyos parámetros no se tiene idea; tal como el número de pasajeros que abordará cierto transporte en un día cualquiera. Analizando las observaciones se logró conocer un comportamiento estadístico que permite realizar inferencia sobre las propiedades de la demanda de usuarios y determinar su verosimilitud

Nota: un comportamiento estadístico es una sucesión de eventos aleatorios en un intervalo de tiempo.

La totalidad de elementos que se desea estudiar y sobre los cuales se desea obtener información es llamada *población objetivo* o simplemente población.

Un problema que se presenta cuando uno quiere saber las características de un grupo de objetos o individuos, es que tal vez éstos sean demasiados. Así, uno de los principales problemas que la estadística busca resolver es lograr un conocimiento acerca de la totalidad de la población mediante el análisis de un pequeño grupo de ésta. El problema que inmediatamente surge es: ¿cómo seleccionar una parte de la población objetivo para investigarla?

Una *muestra estadística* (también llamada *muestra aleatoria* o simplemente *muestra*) es un subconjunto de casos o individuos de una población estadística. Donde se tiene un conjunto de variables aleatorias digamos:

$X_1, X_2,\ldots, X_n$ que comparten una función de densidad conjunta $f_{X1, X2,\ldots, Xn}(x_1, x_2,\ldots, x_n)$ *tales que cumplen* $f_{X1, X2\ldots Xn}(x_1, x_2,\ldots, x_n) = f(x_1), f(x_2),\ldots, f(x_n)$. *Para* $i = 1,2,\ldots, n$ *(principio de independencia)*

Donde f (*) es la función común de densidad de cada X_i. Entonces $X_1, X_2,\ldots, X_n$ está definida como una muestra aleatoria de tamaño "n" de una población con densidad f (*).

Podemos hacer afirmaciones basados en probabilidad, si la muestra es seleccionada con cierta dinámica. Si tomamos una muestra de tamaño "n" y capturamos sus valores numéricos observados, tendremos los valores: $x_1, x_2, \ldots, x_n$ conocidos y quisiéramos poder realizar alguna operación con estos valores. [45]

En el caso de una muestra aleatoria simple: $X_1, \ldots, X_n$, (variables aleatorias) vale suponer que cada elemento en nuestra población tiene algún valor numérico asociado y la distribución de estos valores numéricos está dada por una función de densidad. [4]

Un caso distinto o caso límite en este tipo de análisis son los censos. En ese caso nos interesa contar el número total de individuos dentro de una población objetivo.

Una de las herramientas que se usan en el conteo de individuos que cumplen con cierta característica pero que a diferencia de los censos, sólo estudia un subconjunto de la población objetivo, es el muestreo. Hacer un muestreo equivale a aproximar un número usando sólo una pequeña porción de la información disponible. Existen diversas técnicas y enfoques en el muestreo, que hacen muy confiables los resultados y se logra aproximar el resultado a un censo, aunque ésta técnica resulta costosa económicamente hablando.
En nuestro caso, una muestra se obtuvo mediante los dispositivos mecánicos que captan el ingreso al Sistema (los torniquetes) y que sirven para verificar el acceso con boleto o tarjeta electrónica, además de contabilizar el número de pasajeros que ingresan a las estaciones por día. Los torniquetes miden -mediante dispositivos electromecánicos- la cantidad de usuarios que ingresan a la estación, cada giro implica el ingreso de un usuario; estos dispositivos son del tipo "sentido positivo" es decir, cuando el aparato da vuelta en sentido contrario al de "entrada", la contabilidad del número de usuarios no decrece. La cantidad observada de usuarios corresponde sólo a una Línea del Sistema, a un período determinado de tiempo y a un conjunto limitado de estaciones.

Una de las razones principales para seleccionar Línea O se basa en que se evita establecer el número de combinaciones que hay en un conjunto de k-individuos que transbordan de una línea hacia Línea O tomados de "q en q" $\binom{k}{p}$. Esto implica evadir el cálculo de probabilidades que significa obtener la distribución de pasajeros que hipotéticamente transbordarían desde una línea del Sistema hacia Línea O.

Para obtener la muestra de interés, se recolectaron las mediciones realizadas de manera ininterrumpida durante el año de 2010, a partir del primero de Enero y hasta el 31 de Diciembre, que corresponden a la Línea O del Sistema. Dicha Línea carece de intersecciones que la conecten a otras líneas del Sistema, no posee estaciones de transferencia o transbordo entre sus terminales Pantitlán y La Paz.

Cabe destacar que el registro de los datos en el Sistema es diario, pero que el proceso de captura y depuración de estos lleva un desfasamiento de 3 meses o más con respecto a su registro inicial (en las estaciones); por tal motivo, y debido a la falta de disponibilidad de los datos, el análisis parecerá muy retrasado.

Los datos que nos incumben consisten de una serie de mediciones realizadas sobre un número de objetos, personas o entidades de interés. Tales datos pueden ser representados de una manera general por una matriz dada por:

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

En este caso el elemento típico $x_{i,j}$ representa el valor de la j-ésima entidad de interés o variable para el i-ésimo individuo. El número de individuos bajo observación es representado por "n" y el número de mediciones tomadas de cada sujeto es "p". En diversos casos, las variables observadas serán de diferentes tipos y cada una de ellas corresponderá a alguna de las siguientes escalas de medición:

Nominal: Corresponde a variables que solo muestran categorías, por ejemplo:

SEXO (M, F), Verdadero, Falso.

Ordinal: Donde existe un orden entre las variables pero éste no implica la distancia entre diferentes puntos de la escala. Por ejemplo: coeficiente intelectual, I.Q

Intervalo: En cuyo caso hay diferencias iguales entre puntos sucesivos en la escala pero donde el cero es un punto arbitrario. Las escalas de temperatura son ejemplo de variables tipo intervalo, el "cero" es arbitrario según la escala Fahrenheit y Celsius, aunque ambas refieren la misma medida: temperatura

Razón o racional: Es el máximo nivel de medición, donde pueden compararse diferencias en puntuaciones tanto como en la relativa magnitud de los puntajes. Así, las mediciones que corresponden al número de pasajeros que ingresan en las estaciones de la Línea O, corresponden a una muestra aleatoria de mediciones de tipo discreto en escala racional.

1.5.- Desarrollo de la investigación

Muchos de los problemas relacionados al transporte de pasajeros involucran procesos estocásticos, los cuales son influenciados por factores observados y no observados de forma desconocida. La naturaleza estocástica de los problemas de transportación es, en su mayor parte el resultado del rol que juegan las personas durante la toma de decisiones que utilizan para su transportación.

Esta complejidad estocástica requirió que en el análisis realizado sobre la demanda de transporte se utilizaran diversos instrumentos o herramientas analíticas, en diferentes etapas.

La primera parte de la investigación se ocupó del pre-procesamiento de la información, transformación de los datos y obtención de estadísticos sumarios simples, gráficas y análisis exploratorio de datos.

La segunda parte correspondió a los fundamentos de la inferencia estadística, enfocándose en modelos dependientes de variables aleatorias continuas y análisis paramétrico clásico, en cuyo caso se revisará el impacto que causa la no verificación de los supuestos del análisis paramétrico y opciones de solución.

En la tercera parte se revisó la pertinencia de examinar las unidades de muestreo desde un enfoque de ajuste de distribuciones. Más aún, se presentará el análisis de modelos de variables independientes e idénticamente distribuidas llevados a cabo sobre una línea de tiempo a fin de obtener la distribución de probabilidad que mejor se ajuste al comportamiento estocástico de las observaciones.

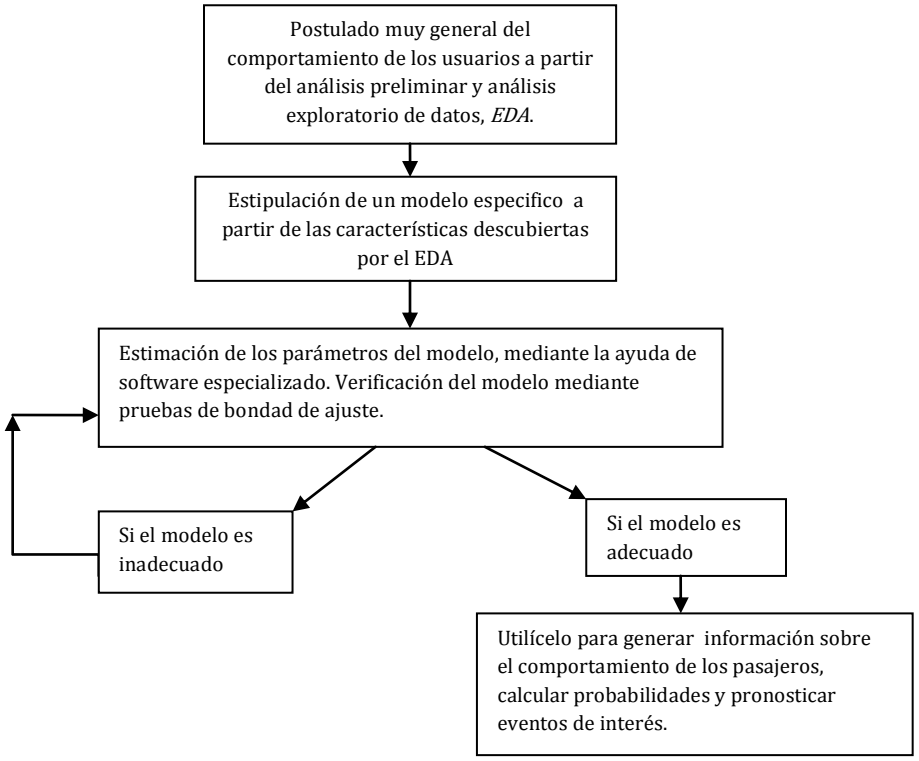
Posteriormente, en la cuarta parte se presentan los resultados obtenidos y recomendaciones pertinentes a la administración. Para concluir, en la quinta parte se incluyen los anexos y complementos que sirvieron para soportar la investigación realizada.

Capítulo 2. Estimación de la demanda de Usuarios

Introducción

El objetivo de las teorías aquí descritas es proporcionar en perspectiva técnicas que se utilizan en el análisis estadístico para obtener una representación simplificada de la estructura subyacente al comportamiento de observaciones obtenidas y alcanzar tal representación mediante lo que justamente es llamado *modelo.* Éste debe servir para proveer un perfil del comportamiento realizado por el flujo de individuos que ingresan a un medio de transporte. Un modelo puede gestarse desde una descripción verbal sencillamente imprecisa hasta una representación geométrica o una ecuación matemática. La intención de construir tal modelo es proveer una descripción que sea consistente con los datos. En ese contexto, el modelo se considera una simplificación y soporte de la estructura de los datos y no necesariamente implica ligas causales o mecanismos ya que estos pueden ser multifactoriales y no es necesario obtener un perfil de tales características. El proceso a seguir para la construcción del modelo se presenta en la figura 2.1. Inicialmente un conocimiento de la manera en que los datos han sido coleccionados y los resultados del análisis exploratorio de datos (Exploratory Data Analysis, EDA) permitirán postular una clase muy general de modelos que pudieran ser adecuados y que posteriormente sean refinados hasta obtener el que sea útil.

Figura 2.1



2.1 Supuestos subyacentes de la teoría de la probabilidad

2.1.1 Propiedades del modelo

La estadística del transporte combina ambas disciplinas para indagar acerca de leyes y principios generales que no son observables y que rigen un comportamiento bajo investigación. Puesto que estas leyes y principios no son directamente notorias entonces se formulan en términos de una hipótesis de investigación; en modelación matemática tal hipótesis sobre la estructura y comportamiento del proceso de interés se establece en términos de familias de distribuciones paramétricas de probabilidad: Modelos. El objetivo del modelaje es deducir la forma del proceso subyacente y verificar la viabilidad de tales modelos. Una vez colectados los datos y después de obtener información de ellos, la meta será utilizar diversos procesos que permitan a los datos revelar su "forma interna" para especificar un modelo.

Antes de trabajar con modelos complejos para analizar un conjunto de datos, es muy recomendable obtener un patrón simple de éstos en términos de gráficos y estadísticos que revelen un resumen de las principales propiedades de los datos.

Un estadístico es una función medible (T) que dada una muestra de valores observados ($x_1, x_2,..., x_n$), les asigna un número, $T(x_1, x_2,..., x_n)$, que sirve para estimar determinado parámetro de la distribución de la que procede la muestra. [4]

Una medida sobre un conjunto es una forma sistemática de asignar un número adecuado a cada subconjunto de ése conjunto. La función de medida otorga números reales no-negativos a ciertos subconjuntos de un conjunto "X". Por principio, ésta debe asignar cero al conjunto vacío y ser contablemente aditiva, es decir, la medida de un "gran" subconjunto -que pueda descomponerse en intervalos finitos (contables)- es la suma de las medidas de los subconjuntos más pequeños. Para definir la medida, ésta se asigna inicialmente a una sub-colección de todos los subconjuntos, los así llamados subconjuntos medibles que se requieren para formar una sigma-álgebra $(\sigma - álgebra)$. Ser parte de una $\sigma - álgebra$ implica que las uniones e intersecciones de subconjuntos medibles, junto con sus complementos son por tanto medibles. [26]

En la siguiente figura (**2.2**), se muestra el mapeo existente entre diversos subconjuntos y la recta real, asignando una medida consistente y monótona, es decir: si A es un subconjunto de B, entonces la medida del subconjunto A es menor o igual que la medida del subconjunto B.

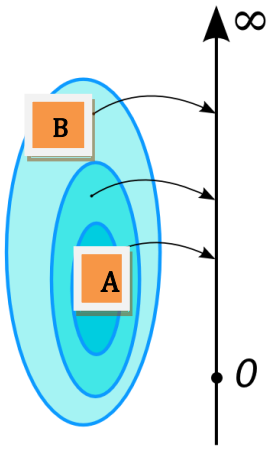


Figura 2.2 Función de medida

En lenguaje simple un parámetro es el término utilizado para identificar una característica o factor conmensurable que ayude a identificar un objeto en particular. En estadística, un parámetro es un valor numérico que describe alguna de las características de una población o distribución de probabilidad. Por ejemplo la distribución Binomial con parámetros (**n, p**) está completamente determinada si el número de ensayos (**n**) y la probabilidad de éxito (**p**) son conocidos. Por su parte una distribución de probabilidad se representa mediante una tabla o ecuación que liga cada resultado de un experimento estadístico con su probabilidad de ocurrencia. [45]

Un experimento estadístico consiste en realizar una acción que selecciona un punto de un conjunto omega, (**Ω**), llamado espacio muestral. El punto seleccionado es llamado resultado del experimento, en probabilidad y estadística el enfoque se aplica a experimentos aleatorios o estocásticos donde el resultado determinado no puede predecirse por adelantado. Subconjuntos del espacio muestral son llamados eventos y son miembros de una colección de subconjuntos de (**Ω**), en cuyo caso omega es llamada una sigma-algebra de eventos. Cualquier experimento estadístico posee tres características principales: el experimento puede tener más de un resultado, el conjunto de resultados puede especificarse por adelantado pero el resultado del experimento depende del azar.

Una variable es un símbolo (A, B, x, y,…) que puede tomar uno entre cualquier conjunto de valores. Cuándo el valor de la variable es el resultado de un experimento estadístico, la

variable se torna en una variable aleatoria (VA). Una VA es una función que asigna un valor numérico único a todos los posibles resultados de un experimento llevado a cabo bajo condiciones experimentales fijas. Matemáticamente la VA se define como una función medible que mapea un espacio de probabilidad o espacio muestral ($\boldsymbol{\Omega}$) a números, éste espacio medible es el espacio de todos los posibles resultados ($\boldsymbol{\omega}$) del experimento estadístico, donde ($\boldsymbol{\omega} \in \boldsymbol{\Omega}$) y los valores de la VA pertenecen a un mapeo correspondiente a los números reales. Al conjunto de resultados obtenidos mediante uniones, intersecciones y complemento de resultados simples ($\boldsymbol{\omega}$), se les denomina álgebra de eventos ($\mathscr{F}$) y constituyen una sigma álgebra ($\sigma - \acute{a}lgebra$) [26]

Generalmente se utilizan letras mayúsculas para representar variables aleatorias (VA's) y letras minúsculas para representar los valores que toma la VA, así como los valores observados durante el experimento. Digamos que "X" representa la VA, entonces la probabilidad de que ocurra el evento representado por X será P (X). P (X = x) se refiere a la probabilidad de que la VA sea igual a un cierto valor "x". Una VA puede ser discreta si el rango de la VA es contable, es decir, si existe un conjunto finito o numerable de números reales: $\{x_1, x_2, \ldots, x_n\}$ tales que "X" únicamente puede tomar valores en ése conjunto. Una VA continua, por el contrario tiene un rango definido sobre un intervalo (a, b), donde: $P(X = x) = 0$; como se verá más adelante.

La exploración estadística actual sin el concepto de distribución de probabilidad de una VA sería muy difícil de concebir, la distribución de probabilidad es un modelo matemático que describe el comportamiento probabilístico de la VA. Cualquier utilización posterior de la distribución de la VA: cálculo de probabilidades, inferencia estadística, etc. utilizan y dependen de la distribución de probabilidad que se presupone para la VA. La representación matemática más tangible de la distribución de una variable aleatoria se corresponde con las denominadas funciones de distribución acumulativa (CDF) y densidad de probabilidad (PDF) de la VA, íntimamente relacionadas entre sí.

La función de distribución acumulativa (CDF) de una VA "X", denotada por $F_X(x)$ se define como una función con dominio en la línea real y contradominio en el intervalo [0,1], la cual satisface:

$$F_X(x) = P[X \leq x) = P\{\omega : X(\omega) \leq x\} \_\_\_\_\_ 1)$$

Para todo número real "x"; donde "ω" es un elemento del espacio muestral $\boldsymbol{\Omega}$, es decir uno de todos los posibles resultados del experimento. Una VA X se dice continua si existe una función $f_X\,(x)$ tal que:

$$f_X(x) = \int_{-\infty}^{x} f_X(u)du \text{ Para todo número real "x"} \_\_\_\_\_ 2)$$

Donde "u" es una variable auxiliar dependiente de "x". La función de distribución $F_X(x)$ de una VA continua "X" es llamada absolutamente continua. La función $f_X\,(x)$ en la ecuación **2** es llamada la *función de densidad* de la VA X.

Obsérvese que la palabra "continua" implica que:

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\left(\int_{-\infty}^{x} f_X(u)du\right) = f_X(x) \_\_\_\_\_ 3)$$

Por lo tanto, la probabilidad de que la VA X= x, es cero por ser la función de densidad $f_X\,(x)$ continua. Es decir, la función $f_X\,(x)$ puede escribirse como la derivada de $F_X\,(x)$. [4]

Si se conoce la función de densidad de una VA entonces se tiene una descripción completa de la misma. Es por tanto un problema fundamental de la estadística estimar la función de densidad de una VA a partir de la información proporcionada por una muestra. El problema de estimación consiste en suponer que algunas características de los elementos de una población pueden representarse por una VA X cuya densidad es de la forma:

$$f_X(*;\ \theta) = f(*;\ \theta)$$

Donde la función de densidad es dependiente de algún(os) parámetro(s) $\boldsymbol{\theta}$

Un primer enfoque es considerar que la función de densidad que se desea estimar pertenece a una determinada clase de funciones paramétricas, digamos: Normal, Exponencial, Weibull, etc. Donde la "forma" de la densidad se asume conocida excepto por algún(os) parámetro(s) desconocidos ($\boldsymbol{\theta}$). Posteriormente se asume que los valores $\{x_1, x_2,\ldots, x_n\}$ de una muestra aleatoria de las VA's $X_1, X_2,\ldots, X_n$ de la densidad f_X pueden ser observados. En base a los valores muestrales $\{x_1, x_2,\ldots, x_n\}$ se desea estimar el valor de los parámetros desconocidos ($\boldsymbol{\theta}$) o el valor de alguna función (estadístico) digamos T ($\boldsymbol{\theta}$), de los parámetros desconocidos.

Tal suposición comúnmente se basa en informaciones sobre la variable que son externas a la muestra, pero cuya validez puede ser comprobada posteriormente mediante pruebas de bondad de ajuste. Bajo esta suposición la estimación se reduce a determinar el valor de los parámetros del modelo a partir de la muestra. Esta estimación es la que se denomina estimación paramétrica de la densidad.

Ahora bien, cualquier estadístico cuyos valores sean utilizados para estimar T ($\boldsymbol{\theta}$) donde T (*) es alguna función del parámetro $\boldsymbol{\theta}$; es llamado *estimador* de T ($\boldsymbol{\theta}$). Un estimador es una función que permite obtener el valor de un parámetro poblacional basándose en una muestra aleatoria proveniente de la población. Al mismo tiempo, el estimador es una VA porque su valor depende de una muestra particular, la cual es aleatoria. El método utilizado para obtener los parámetros de la función de densidad será el método de máxima verosimilitud. [4]

Por lo tanto, un estimador es tanto una VA como una función. Por instancia y suponiendo que se tiene: $X_1, X_2,\ldots, X_n$ una muestra aleatoria de una densidad $f(*, \boldsymbol{\theta})$ y se desea estimar el parámetro ($\boldsymbol{\theta}$) de tal densidad, entonces T (*) debe ser una función de $\boldsymbol{\theta}$.

Así entonces $T = t(X_1, X_2,\ldots, X_n)$ es un estimador de ($\boldsymbol{\theta}$), resaltando el hecho de que ($\boldsymbol{\theta}$) puede ser un valor simple o un conjunto de valores, vector. De manera que el estimador:

$T = t\ (X_1, X_2, X_3\ldots,)$ puede pensarse de dos formas relacionadas, la primera como una VA, digamos T (*), donde $T= t\ (X_1, X_2,\ldots, X_n)$ y la segunda como la función $t(X_1, X_2,\ldots, X_n)$.

Generalmente y por convención se llama al estadístico (VA) que es utilizado como estimador *el estimador* y a los valores que ése estadístico toma se les llama *estimados*. De manera que la palabra *estimador* sirve para nombrar la función y la palabra *estimado* se utiliza para los valores de ésa función.

En cuanto al problema de estimación de la función de densidad y sus parámetros, la posibilidad alternativa es no predeterminar a priori ningún modelo para la densidad de probabilidad de la VA y dejar que la función densidad pueda adoptar cualquier forma, sin más límites que los impuestos por las propiedades que se exigen a estas funciones para ser consideradas como tales. La técnica tiene sus orígenes en los trabajos de Fix y Hodges en 1951, que buscaban una alternativa a las técnicas clásicas de análisis discriminante para liberarse de las rígidas restricciones sobre la distribución de las variables implicadas. En cierta manera el enfoque no paramétrico permite que los datos determinen de forma

totalmente independiente, sin prohibiciones, la forma de la densidad que los ha de representar.

Para construir el modelo una hipótesis común es aquella que permite suponer que "los datos observados son realizaciones de variables aleatorias, independientes e idénticamente distribuidas, (IID)". La ventaja de asumir tal hipótesis es que el experimento del flujo de pasajeros por las estaciones de Línea O no puede ser reproducido a plenitud, es decir, estamos frente a un proceso aleatorio que ocurre de manera única y por tal motivo es útil presumir que el resultado es la realización de una variable aleatoria. La teoría de la probabilidad provee herramientas como la Ley débil de los Grandes Números, el teorema Central del Límite y el teorema de Glivenko-Cantelli que permiten extraer de los datos un comportamiento general y que podrá por tanto ser el fundamento para obtener un modelo.

Anticipadamente a la obtención de un modelo, será necesario realizar procedimientos de análisis exploratorio de datos (EDA) y obtener información sobre el comportamiento del flujo de pasajeros, en base a estos resultados preliminares, se podrá proponer alguna distribución de probabilidad que se ajuste a los datos observados. Con el modelo especifico y conocidos sus parámetros, se está en posición de evaluar el modelo mediante pruebas de bondad de ajuste, las cuales consisten en verificar que tan bien éste modelo se ajusta a los datos observados. La bondad del ajuste se coteja comparando la distribución teórica (o sugerida) contra la distribución que siguen los datos en forma *empírica.* Como ya se mencionó una función de distribución acumulativa da la probabilidad de que una variable aleatoria X, sea menor o igual a un cierto valor digamos "t".

$$F_X(t) = F\ (t) = P\ \{X\ \leq t\}$$

Una distribución empírica es muy similar, la diferencia es que se trabaja con datos observados en lugar de funciones teóricas.

Sean (X_1, X_2,…, X_n), VA's independientes e idénticamente distribuidas (IID) cuya función de distribución acumulativa común (CDF) para ello, sea la función F_{Xi} (t), para $i = 1, 2, \ldots, n$

La función de distribución empírica se define como:

$$\widehat{F_n}(t) = \frac{número\ de\ elementos\ en\ la\ muestra\ \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I_A}\{x_i\ \leq t\}$$

Donde I_A es la función indicador del evento "A" la cual es igual a uno cuando el evento A ocurre y cero en cualquier otro caso.

Teniendo en cuenta que la función de distribución es una función no decreciente, y continua por la derecha con valores en el intervalo [0,1] donde "n" es el número de elementos en la muestra, se tiene que:

$$\lim_{t\to-\infty} F_t = 0 \ \textit{Y que:} \ \lim_{t\to\infty} F_t = 1$$

Con esto, se desea establecer que, F_X (t) es una cantidad teórica, que puede estimarse en términos de los datos utilizando la función de distribución empírica:

$$\widehat{F_n} = \frac{1}{n}\sum_{i=1}^{n} I_A\{x_i \leq t\} \_\_\_\_\_ 4)$$

De manera que mientras la función de distribución suministra como función de "t" *la probabilidad de que cada una de las VA's* X'_i *sea menor igual a un número "t"*; la función de distribución empírica **–**calculada mediante los datos**-** proporciona la frecuencia relativa con la que *los valores observados son menores o iguales a cierta cantidad* ***"t"***. *Debe notarse que* $\widehat{F_n}(t)$ *es un estimador insesgado para F (t).* Anexo 5.1.2

2.2 Análisis Exploratorio de Datos

El análisis exploratorio de datos (Exploratory Data Analysis, por sus siglas en inglés) es un enfoque de aproximación al análisis que emplea una variedad de técnicas no paramétricas -la mayoría gráficas- para:

1.-Maximizar la visión del conjunto de datos

2.- Descubrir la estructura subyacente

3.- Extraer variables importantes, detectar observaciones aberrantes y anomalías

El EDA consiste en realizar un trabajo cuantitativo sobre los datos e implica que éstos deben ser explorados primeramente sin ningún prejuicio acerca de modelos probabilísticos, distribución de errores, número de grupos o relaciones entre las variables. La meta será explorar los datos para revelar patrones y características que serán útiles para entender y analizar el comportamiento que describen las observaciones.

El EDA permite revelar información acerca de los datos y contiene metodologías que permiten visualizar las características de éstos. Diversas características descriptivas de la densidad, tales como multimodalidad, asimetrías, comportamiento en las colas, etc., enfocadas desde un punto de vista no paramétrico, y por tanto más flexible, pueden ser más reveladoras y no quedar enmascaradas por suposiciones rígidas. Las técnicas del EDA persiguen algunas metas que se aplican comúnmente a diversos tipos de datos:

A) Inicialmente es necesario obtener estadísticos que reduzcan la información a un conjunto de medidas descriptivas, al tiempo que se hace necesario hallar observaciones aberrantes (*outliers*) o que carezcan de sentido, verificar si hay datos faltantes y realizar un análisis gráfico que permita ver el comportamiento de los datos.
B) A partir de una muestra, ésta se utilizará para desarrollar un modelo; el modelo puede ser incluido en la simulación de un proceso físico y el EDA puede servir para determinar las técnicas que ayuden a determinar la distribución de los datos y qué modelo puede ser de mayor utilidad.
C) Si se tienen observaciones que relacionan dos o más características de una variable y se hace necesario describir como es tal relación, entonces, un gráfico inicial entre ellas puede definir el tipo de relación, ya sea lineal o de algún otro tipo.

Así, existen dos objetivos principales en el EDA, 1) localizar aquellas observaciones que parecieran fuera de contexto y analizarlas 2) determinar un modelo razonable para describir el proceso que genera los datos y mediante técnicas sencillas, empleadas por el EDA, realizar las siguientes tareas:

1.- Graficar los datos "crudos" (en grafías como diagramas de "caja y bigotes", histogramas, diagramas cuantil-cuantil, etc.).

2.- Obtener estadísticos simples como, la media, desviación estándar y moda, etc.

En cierta manera el enfoque no paramétrico permite que los datos determinen de forma totalmente libre, sin restricciones, la forma de la densidad que los ha de representar. El análisis exploratorio de datos (EDA) constituye una tradición estadística bien establecida que provee herramientas conceptuales y de cómputo muy útiles para descubrir patrones que permiten fomentar el desarrollo y refinamiento de la hipótesis de investigación. Estas herramientas y actitudes complementan el uso de las pruebas significancia e hipótesis utilizadas en el análisis confirmatorio de datos y ajuste de distribuciones.

Capítulo 3. Construcción del modelo

Introducción

La importancia de la dinámica en el comportamiento del transporte ha sido reconocida desde hace tiempo. Esta dinámica incluye la capacidad presente de los hábitos, la incertidumbre e información imperfecta que sobreviene al registrar alternativas, costos de transacción y valores esperados concernientes al ingreso y ciclo de vida de los pasajeros que utilizan el transporte. Se supone que alrededor de la motivación que dispone el comportamiento humano, las decisiones son hechas mediante ensayo y error, un proceso que ocurre naturalmente en el tiempo e implica que las decisiones óptimas no sean alcanzadas instantáneamente.

En general, la modelación del transporte continúa basándose en muestras de comportamiento del viajero en algún punto del tiempo. Tales datos se refieren a diferencias entre las circunstancias y el comportamiento de los individuos entre los viajes, digamos "escenarios de elección" dados en un lapso; sin constatar que posiblemente los cambios en las circunstancias implican cambios en el comportamiento de viaje para un individuo determinado. De manera que hasta solo recientemente los modelos estadísticos constituyen el fundamento de la mayoría de los modelos de transporte y una razón para esto tiene que ver con el tipo de datos utilizados para analizar modelos de flujo de pasajeros.

La representación matemática del flujo de pasajeros coincide con la denominada función de densidad de probabilidad de una variable aleatoria (VA), el conocimiento de la función de densidad de la VA otorga una descripción completa de la misma. De manera que la estimación de la función de densidad de una VA a partir de la información proporcionada por una muestra de pasajeros en flujo hacia las estaciones del Sistema es el problema que se solventó.

3.1 El análisis no paramétrico

Dejando de lado la estimación paramétrica, la opción alternativa es no predeterminar a priori ningún modelo para la densidad de probabilidad de la VA y dejar que la función de densidad pueda adoptar cualquier forma sugerida por los datos, manteniendo los supuestos requeridos para que una función sea considerada una función de densidad. La estadística no paramétrica estudia las pruebas y modelos estadísticos cuya distribución subyacente no se ajusta a los llamados criterios paramétricos.

Así que la función de probabilidad no se definió por adelantado, pues fueron los datos observados los que la determinaron. La utilización de estos métodos se hizo recomendable porque no fue posible asumir que los datos se ajustaran a una distribución conocida cuando el nivel de medición empleado no fuera, mínimo de intervalo. En tal caso el nivel de medición corresponde al máximo nivel de medición, nivel racional o razón.

El motivo para preferir el enfoque no paramétrico consiste en pensar que las circunstancias del viajero están cambiando continuamente y los ajustes a estos cambios toman un tiempo determinado, así que una encuesta sobre hábitos y comportamientos en los modos de viaje resulta generalmente inadecuada. La inconsistencia entre los supuestos de modelos teóricos y los datos subyacentes lleva a estimados sesgados del comportamiento a largo plazo. Por tal motivo, una manera de eliminar el sesgo entre comportamiento observado y esperado consistió en verificar lo más aproximadamente posible el comportamiento observado mediante un análisis de datos que verificó la manera en que se comportan los viajeros, lo cual ocurre cuando hay formas de cotejar el comportamiento subyacente a la conducta del viajero mediante una función de probabilidad.

Tal comportamiento está intrínsecamente asociado al modo en que los usuarios acceden al Sistema mediante las lecturas en torniquetes, de ahí la pertinencia de estudiar las lecturas tomadas durante un año, el 2010 y analizarlas desde el punto de vista de la estimación no-paramétrica de la densidad y sus alternativas.

Los datos fueron proporcionados por personal del Sistema perteneciente al departamento de peaje y se otorgaron “crudos”, es decir, corresponden a los conteos capturados en dicho departamento y el sistema no les aplicó ningún tipo de análisis o transformación posterior a su captura.

El análisis se realizó con el auxilio de algunos sistemas de estadística: SPSS, R, Mathw y Micro Soft Excel. SPSS es un software estadístico de uso generalizado en ciencias sociales. R es un software libre que se ofrece sin ningún tipo de garantía y que puede redistribuirse bajo ciertas circunstancias. R [5] tiene una línea de comandos interactivos que ofrece considerables ventajas sobre los sistemas que se basan en menús, en cuestiones de eficiencia y velocidad de procesamiento, además de ofrecer herramientas muy poderosas de graficación y algoritmos seleccionados de análisis. Microsoft Excel es la herramienta habitual de análisis estadístico y hoja de cálculo. Además se utilizó la ayuda del sistema de análisis Mathw. Este producto permite ajustar distribuciones de probabilidad a datos empíricos mediante el método de máxima verosimilitud, tomando como fundamento teórico el hecho de que a una colección de

datos empíricos puede ajustársele una distribución teórica, según el teorema de Glivenko-Cantelli. Si existe tal distribución, entonces ésta puede servir para calcular la probabilidad de eventos de interés y para generar números aleatorios, que siguiendo la distribución, permitan obtener una visión muy aproximada al comportamiento real del flujo de pasajeros en Línea O. Tal software se ofrece de forma libre durante tiempo limitado y es compatible con varios paquetes estadísticos, después del tiempo de gracia hay que adquirirlo bajo licencia, estas herramientas permitieron llevar a cabo el análisis y construcción del modelo.

3.2 Contexto sobre las bases de datos que serán usadas

Es común que en la investigación del comportamiento de pasajeros el interés resida en hallar tendencias de movilización con respecto a diversos factores tales como son: el medio de transporte, la ruta a seguir, el horario etc.

En muchas otras ocasiones lo que se desea comprobar es la probable utilización de un tipo específico de transporte y entonces describir la forma en que los consumidores de éste van "moviéndose" hacia el tipo de transporte bajo observación a fin de hallar el comportamiento que sigue tal concentración de usuarios. Este tipo de experimento se conoce como flujo de pasajeros o demanda de pasajeros a un sistema de transporte.

En este caso, se posee un registro del número de pasajeros que entran al Sistema con porte pagado, contenido en archivos electrónicos, los conteos se registraron diariamente y fueron recabados por el personal del Sistema para un análisis posterior.

Los datos fueron entregados en archivos "planos", con extensión del tipo .dbf o archivos del tipo D-Base, archivos carentes de una estructura que permitiera llevar a cabo algún tipo de búsqueda relacional o "*query*" para obtener información sobre diversos tópicos de interés. Los registros obtenidos correspondientes al tipo antes mencionado, contenían lecturas u observaciones de los 365 días del año 2010 en cada una de las estaciones que conforman la Línea O del Sistema. Los datos de afluencia del archivo original (aflto110.dbf) se hallaban capturados de la siguiente forma, Tabla 3.1

La simbología de la tabla indica el año de acopio de datos (2010) a continuación se tiene el mes y la estación correspondiente de Línea O, la columna ACCES indica la posición donde se encuentra la batería de torniquetes, el acceso indica el tipo de torniquete (U =único), la FASE es un control temporal interno, NTORN indica en que torniquete se está efectuando la

"lectura" y a continuación las columnas AFTO1…AFTO31 indican los totales obtenidos desde el día uno al día 31 del mes.

Tabla 3.1

ANO	MES	ESTAC	ACCES	TIPOT	FASE	NTORN	AFT01	AFT02	AFT03		AFT31
10	1	PAN	NTE	U	1	1	0	0	0	…	0
10	1	PAN	NTE	U	1	2	0	12	3	…	0
10	1	PAN	NTE	U	1	3	13	19	24	…	10
10	1	PAN	NTE	U	1	4	89	116	111	…	210
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	12	LPA	U	U	2	134	234	321	112	…	124
10	12	LPA	U	U	2	135	213	53	321	…	345
10	12	LPA	U	U	3	136	0	125	38	…	237
10	12	TIP	000	0	000	0	0	0	0	…	0

Un examen inicial revela que la manera de agrupar los datos no es útil para realizar algún tipo de análisis, es decir las variables (o estaciones) no están en las columnas y los renglones contienen variables. Debe notarse que las mediciones de afluencia se encuentran en una escala de orden, también llamada escala de razón o racional y que corresponde al máximo nivel de medición, donde pueden compararse diferencias en puntuaciones tanto como en la relativa magnitud de los puntajes y donde el cero representa la ausencia de la característica o propiedad. De manera que las mediciones obtenidas corresponden al número de pasajeros que ingresan en las estaciones de la Línea O y pertenecen a una muestra aleatoria de mediciones de tipo finito contable en escala racional, por tal motivo fue necesario transformar los archivos de lecturas.

3.2.1 Metodología Aplicada a los datos iniciales

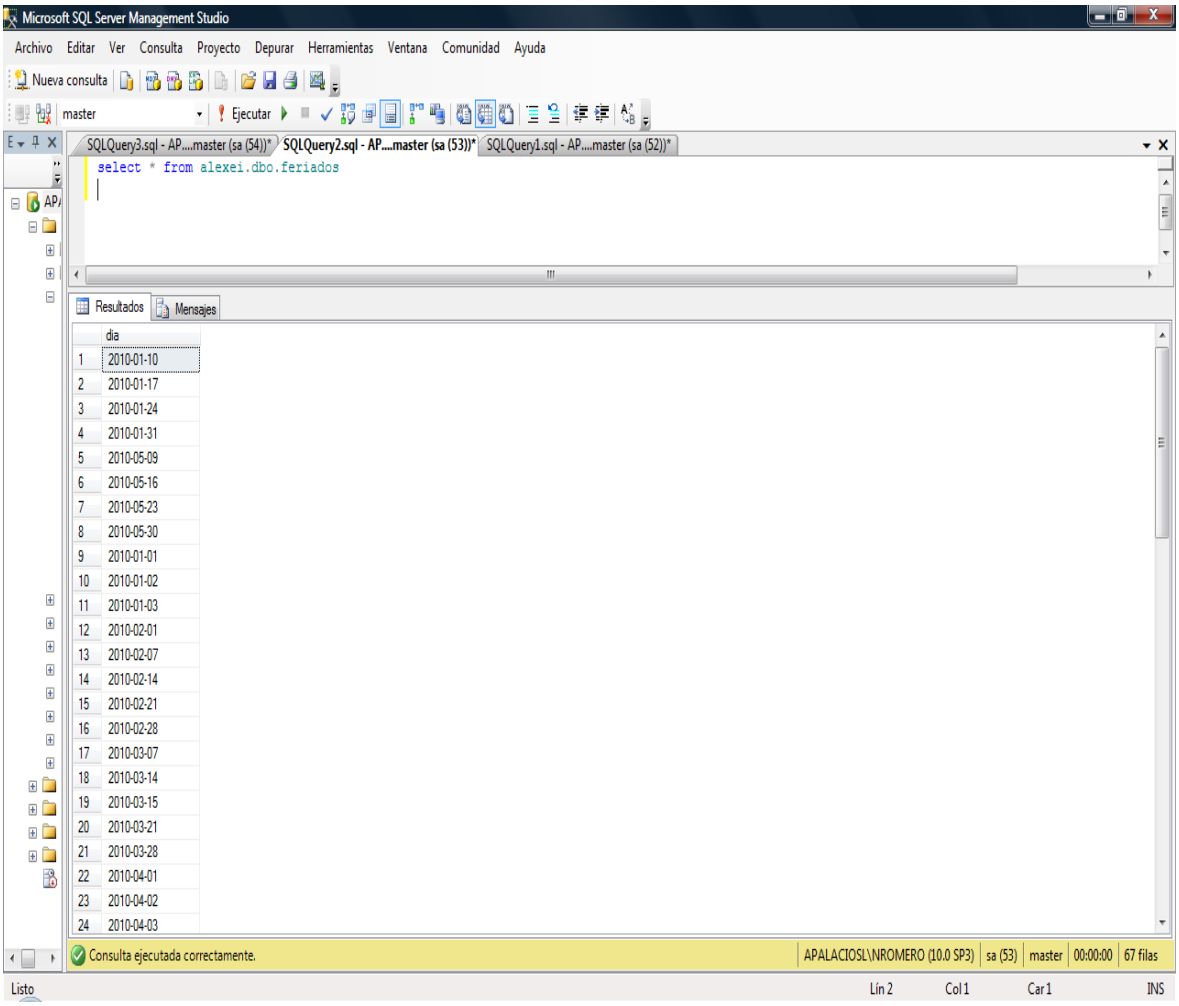
La primera parte del proceso de análisis consistió en realizar transformaciones del archivo plano inicial (**aflto110.dbf**) con la finalidad de obtener una serie de archivos compatibles con diversas herramientas de software estadístico. Este archivo propuesto, "**db_llena**" no posee las propiedades particulares de una base de datos, con relaciones entre sus variables y que permita la búsqueda de caracterizaciones a partir de elementos comunes del archivo.

Su construcción se llevó a cabo mediante operaciones en un sistema manejador de bases de datos muy popular, el así llamado SQL [54]; éste es un sistema de administración muy poderoso que permite llevar a cabo múltiples ejecuciones de búsqueda de datos o "*querys*"

mediante elementos de programación y librerías especializadas para la administración de bases de datos. De manera que para transformar el archivo inicial, la construcción de la tabla "**alexei.dbo.usuarios**" tuvo que realizarse mediante la ejecución de múltiples líneas de código de programación. Para una descripción completa del código se remite al lector al anexo.

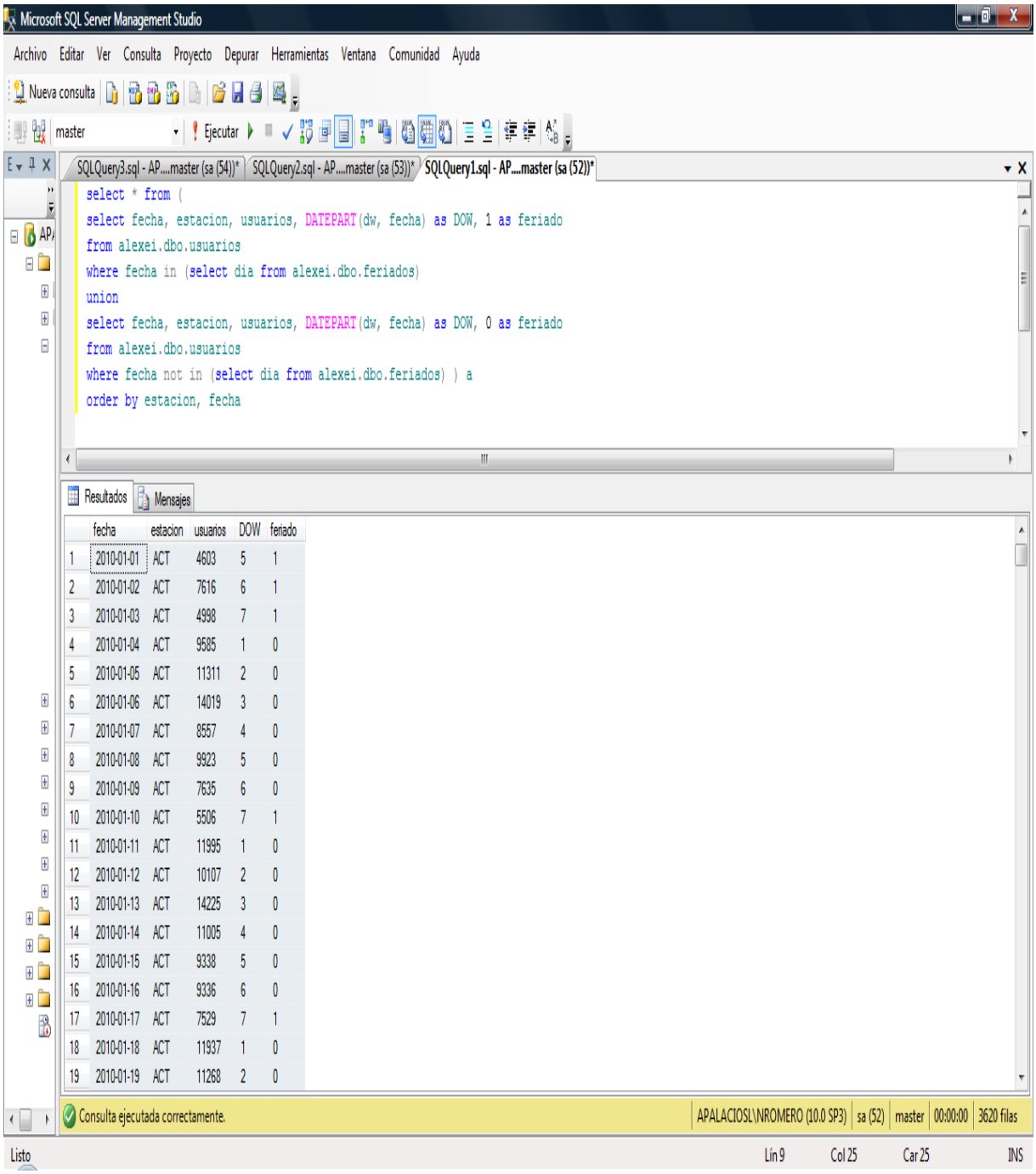
```
	select * from (select fecha, estacion, usuarios, DATEPART(dw,
fecha) as DOW, 1 as
	feriado
	from alexei.dbo.usuarios
	where fecha in (select dia from alexei.dbo.feriados)
	union
	select fecha, estacion, usuarios, DATEPART(dw, fecha) as DOW, 0 as
feriado
	from alexei.dbo.usuarios
	where fecha not in (select dia from alexei.dbo.feriados) ) a
	order by estacion, fecha
```

El código de búsqueda ó *query* anterior hace uso de otra tabla que contiene los días feriados, ésta tabla se muestra en la figura siguiente. **Figura 3.2.1.**



Las dos graficas muestran como se modifican los registros, de simplemente números asignados a una lista por fechas hasta obtener una relación entre *fechas, estaciones, número de usuarios por días (Laborales, Sábados y festivos ó Domingos).* Finalmente se obtienen los resultados que se copian y pegan en el archivo **Acumulados.xls** y se muestran a continuación

Figura 3.2.2



De ésta manera se obtiene un archivo con la forma general de una base de datos compatible con las herramientas de Microsoft EXCEL, del lenguaje R y de los paquetes del sistema Mathw. Finalmente, la conjunción de diversos tipos de datos y archivos da como resultado el archivo **Acumulados.XLS** que representa el "súmmum" de los datos que fueron requeridos para utilizarse en EXCEL y que permitieron obtener los datos más relevantes en la investigación. El código completo se encuentra en el anexo.

3.3 Aplicación del Análisis exploratorio de los datos, EDA

La segunda parte del proceso consistió en obtener de forma detallada varias medidas aplicables a los datos y utilizar elementos gráficos para auxiliarse del EDA con lo cual se obtuvieron conclusiones preliminares sobre el comportamiento de los usuarios con respecto a la afluencia en las estaciones de Línea O. Los datos corresponden al flujo de pasajeros que ingresan en una determinada estación, que a partir de ahora designamos como "variable" del tipo aleatorio, la caracterización de cada estación es mediante la letra $\mathbf{X}$ con subíndice $i = 1$, 2…,10 y la variable Y (T) representa el comportamiento de toda la Línea O.

Esta labor tuvo como objetivo indagar las características principales de los datos contenidos en la base de datos y la tendencia que presentaban las observaciones para cada una de las estaciones, en su conjunto y por separado.

Para ello la información de las siguientes secciones se organizó de la siguiente manera, primero se presentan formas descriptivas mediante las representación en diagramas de caja de las variables de estudio, en la segunda (3.3.2) se hizo un análisis de las estadísticas descriptivas, histogramas y gráficas cuantil-cuantil (*QQ-plots*) que verificaron si la distribución general de los datos sigue una distribución continua por ejemplo la distribución Normal o no, empleando el paquete SPSS. A continuación en la sección 3.3.3 se revisaron los resultados preliminares obtenidos por los análisis anteriores y se propusieron funciones de densidad adecuadas al tipo de los resultados obtenidos. En la 3.3.4 se planteó el análisis mediante funciones de distribución de probabilidad, metodología y alcances del modelo. A partir de la sección 3.4 se obtuvieron diversos modelos para las variables aleatorias en forma Total, Laboral, Sábado y Domingos en la parte 3.7; estos modelos se compararon y validaron mediante pruebas de bondad de ajuste que determinaron la compatibilidad de los modelos con respecto al flujo de pasajeros.

Los resultados de esta ejercicio pretendieron obtener una relación entre cantidades numéricas y aunque ésta sea por sí misma una explicación de la realidad, deberá tenerse en cuenta que tales relaciones son descripciones simples que a su vez necesitan de una explicación, tomando en cuenta que la conducta del individuo, responde muchas veces a factores impredecibles.

Luego, los datos provenientes de una muestra de 365 observaciones fueron procesados utilizando el programa SPSS versión 19.0. Se pretendió cumplir con un objetivo general: determinar el modelo estadístico de distribución que se ajuste a las variables del estudio y alcanzar dos objetivos específicos: primero, analizar la dispersión (mediante diagramas diversos) y tablas estadísticas de cada una de las variables de estudio para verificar la tendencia de las observaciones en análisis no paramétrico y segundo, mostrar la representación descriptiva de cada una de las variables del estudio para saber si se comportan como una distribución Normal, que es un requisito de las pruebas paramétricas y no paramétricas para verificar si las densidades continuas son modelos útiles. Se construyeron gráficos cuantil-cuantil para determinar si el modelo Normal, correspondió al comportamiento de los usuarios.

El estudio realizado consideró la base de datos está compuesta por las siguientes variables de análisis: X_1 Pantitlán; X_2, "Agrícola Oriental"; X_3, "Canal de San Juan"; X_4, "Tepalcates"; X_5, "Guelatao"; X_6, "Peñón Viejo"; X_7, "Acatitla"; X_8, "Santa Martha"; X_9, "Los Reyes", X_{10}, "La paz"; Y (T) Línea O ver figura 3.3.1

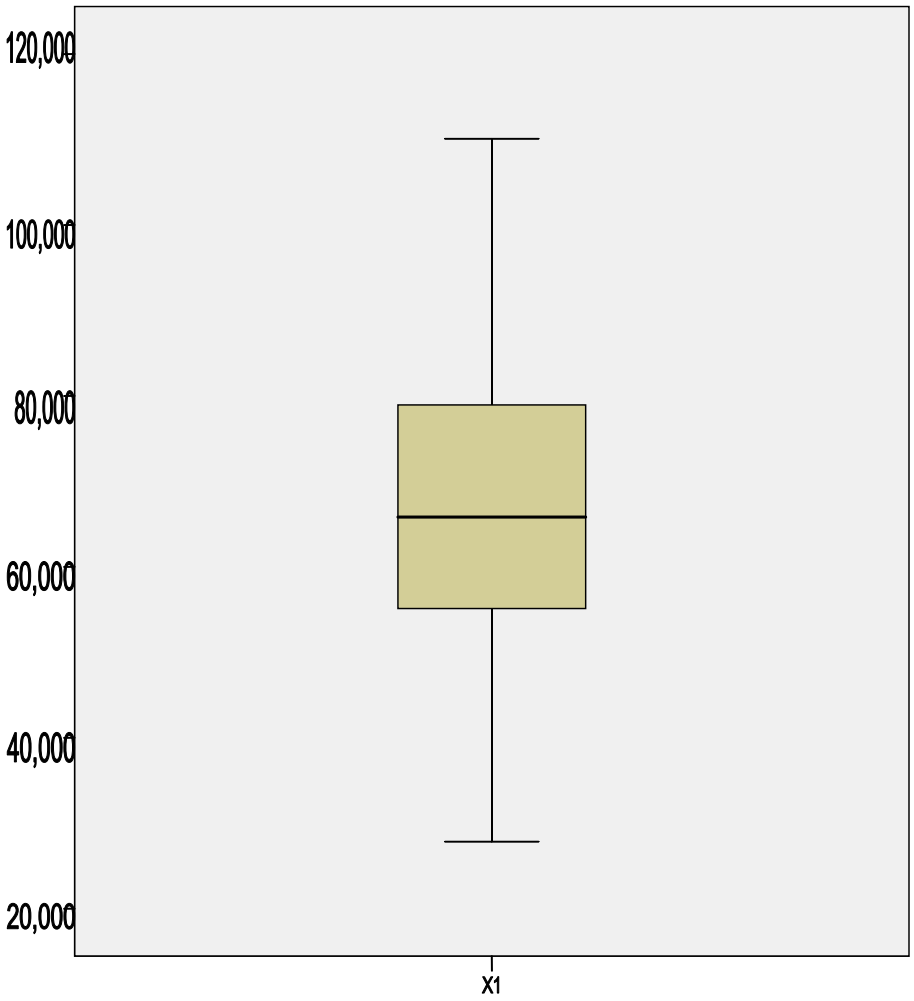
Figura **3.3.1**

Pantitlán	Agrícola O	Canal de San Juan	Tepalcates	Guelatao	Peñón Viejo	Acatitla	Sta. Martha	Los Reyes	La Paz
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

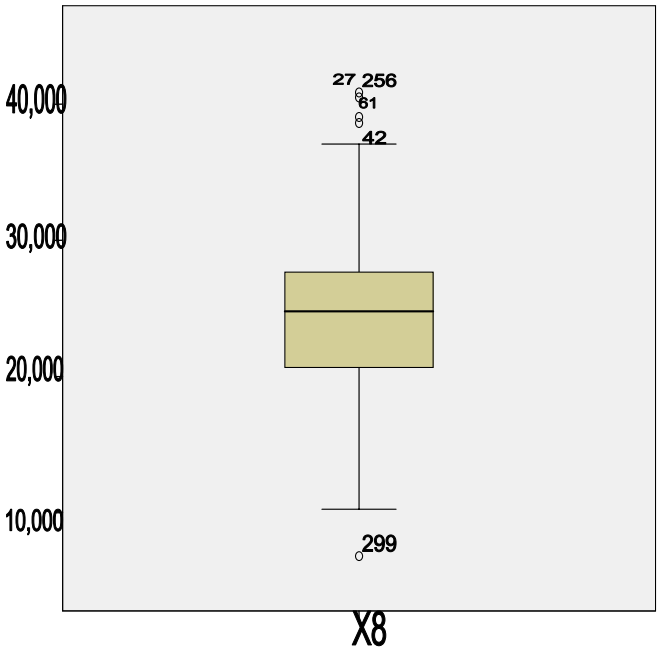
3.3.1 Análisis preliminar

A continuación se muestra la aplicación del EDA, los diagramas de caja muestran diversos estadísticos obtenidos de las variables de estudio y se presentan características poblacionales y los porcentajes de la escala de respuesta a un total de 365 observaciones sobre los

elementos Xi de Línea O. Los diagramas y gráficas aquí mostradas corresponden a las variables que se consideraron más importantes, en función del volumen en flujo de pasajeros como son: X_1, X_8, X_{10} y el total de Línea O, Y (T). En total se analizaron las 10 estaciones de Línea O. En seguida se muestran los análisis mediante diagramas de "caja y bigotes". Se observó que la mayoría de las variables bajo estudio mostraron la mediana ligeramente alejada del valor de la media, es decir presentaron cierta variabilidad en las medidas centrales, lo cual implicó un sesgo en la distribución del flujo de pasajeros. El análisis gráfico es mostrado a continuación, en el **Diagrama de "caja" de la VA X_1**

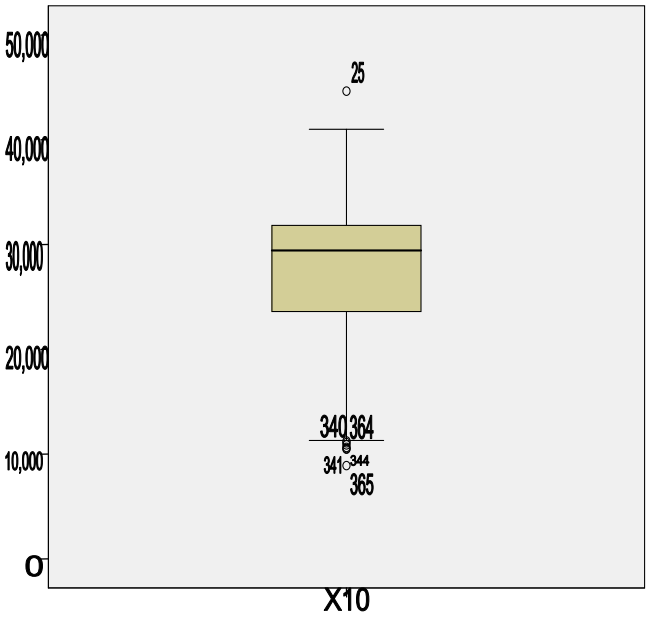


Para X_1 la distribución de los datos se encuentra sesgada a la derecha y sin datos fuera de rango, lo que implica una regularidad en la cantidad de pasajeros que ingresan a la estación X_1, Pantitlán. Se observa un flujo máximo de 110, 000 usuarios aproximadamente y un mínimo de 30 mil. En el siguiente **Diagrama de caja de la VA X_8**

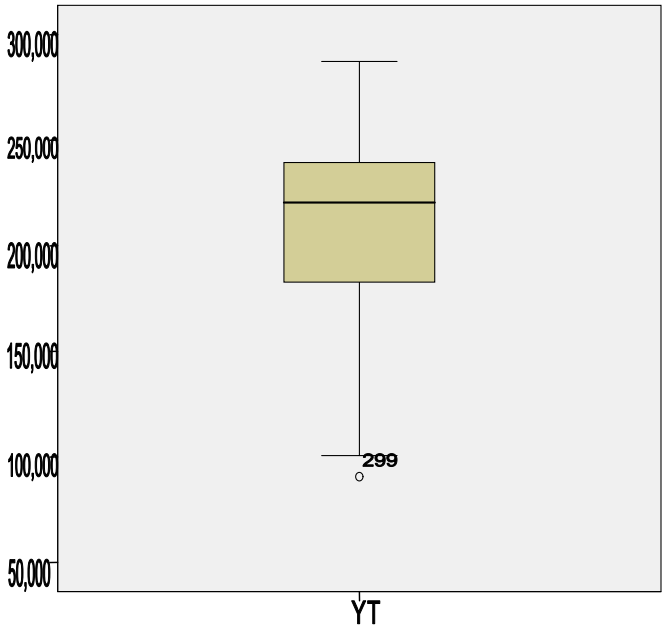


En la variable X_8 llamó la atención una variedad de valores atípicos, que estuvieron por encima del máximo (40 mil) y que ocurrieron en las observaciones 27, 42, 61 y 256, esto condujo a indagar la fuente de tal dispersión. La investigación posterior mostró que estos valores corresponden a días laboralmente "pesados", lunes o miércoles. Caso similar ocurre en la observación del día 299, por debajo del flujo mínimo, día martes, sin motivo verificable.

Diagrama de caja de la VA X_{10}



En la variable X_{10} se observa un sesgo negativo marcado en la distribución de los datos, la mediana se encuentra a corta distancia del tercer cuartil, lo que implica un corto rango intercuartil ó una aglomeración de los datos alrededor de la mediana. También se observan valores atípicos superiores al máximo de 44 mil personas por día y esto ocurre en la observación 25, lunes; es decir un día laboral de alta demanda. Existen valores muy por abajo del mínimo, 9 mil personas por día, y estos valores se registran en las observaciones 340, 341, 344, 364 y 365, esto indica que existe un período de baja demanda especialmente al final del año. La distribución de los datos aparece sin valores atípicos por encima del máximo, pero existe un valor por abajo del mínimo que implica una fluctuación irregular en la observación 299, que corresponde a un día martes (26 Octubre) día laboral pero sin mayor esclarecimiento. **Diagrama de caja de Y (T)**



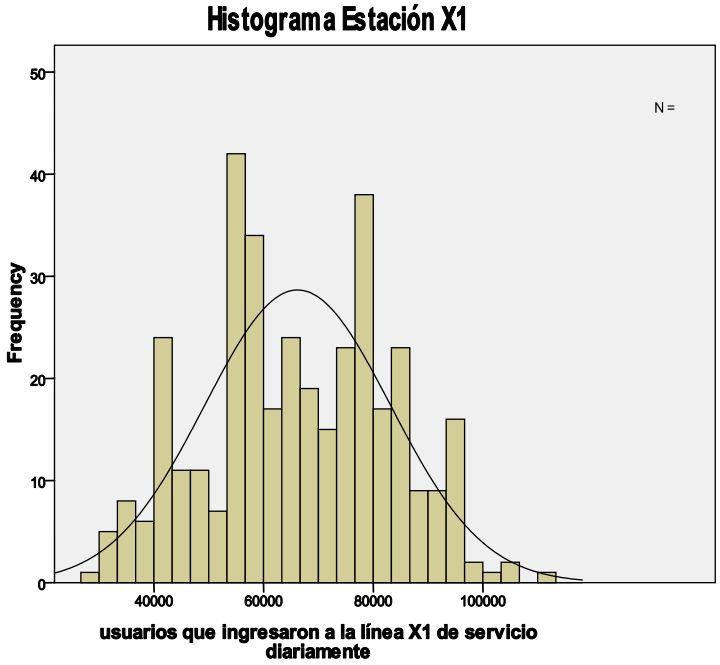
3.3.2 Estadística de las variables, histograma de frecuencias con ajuste Normal y gráficos cuantil-cuantil.

De manera auxiliar a los gráficos de caja se obtuvieron distintas mediciones estadísticas que permitan ahondar en la forma en que se distribuyen los usuarios de la Línea O por las estaciones del sistema; estos datos nos permitirán discriminar entre los diversos tipos de distribuciones de probabilidad que puedan describir el flujo de pasajeros en Línea O.

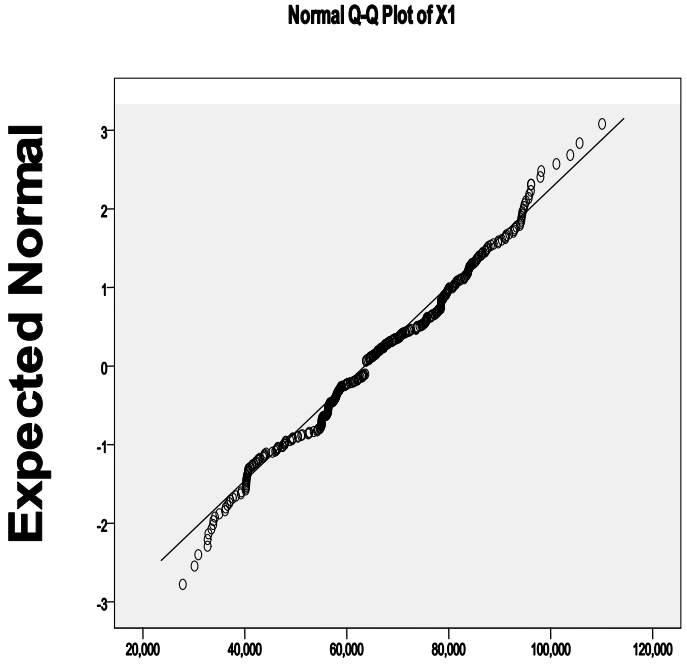
Datos estadísticos de la VA X_1

			Statistic	Std. Error
X1	Mean		66209.93	886.443
	95% Confidence Interval for Mean	Lower Bound	64466.74	
		Upper Bound	67953.13	
	5% Trimmed Mean		66231.45	
	Median		65825.00	
	Variance		2.868E8	
	Std. Deviation		16935.463	
	Minimum		27855	
	Maximum		110099	
	Range		82244	
	Interquartile Range		23848	
	Skewness		-.025	.128
	Kurtosis		-.702	.255

Los datos de la tabla muestran que el valor de la mediana está por encima del valor esperado para la media, esto denota un sesgo negativo en la distribución de los datos. El histograma de la variable X_1 muestra la distribución de los datos en un histograma comparándolo con una distribución Normal pero puede verse que es multimodal y que una Normal no necesariamente ajusta adecuadamente.



Puede verse en el histograma que los datos tienen una cierta asimetría y que presentan múltiples valores multimodales, la gráfica de comportamiento con respecto a una distribución Normal (*QQ-plot*) muestra que los datos de la variable X_1, tienen un comportamiento poco asintótico con una distribución Normal para valores esperados teóricos y flujo empírico de pasajeros entre [95, 000, 120, 000] individuos.

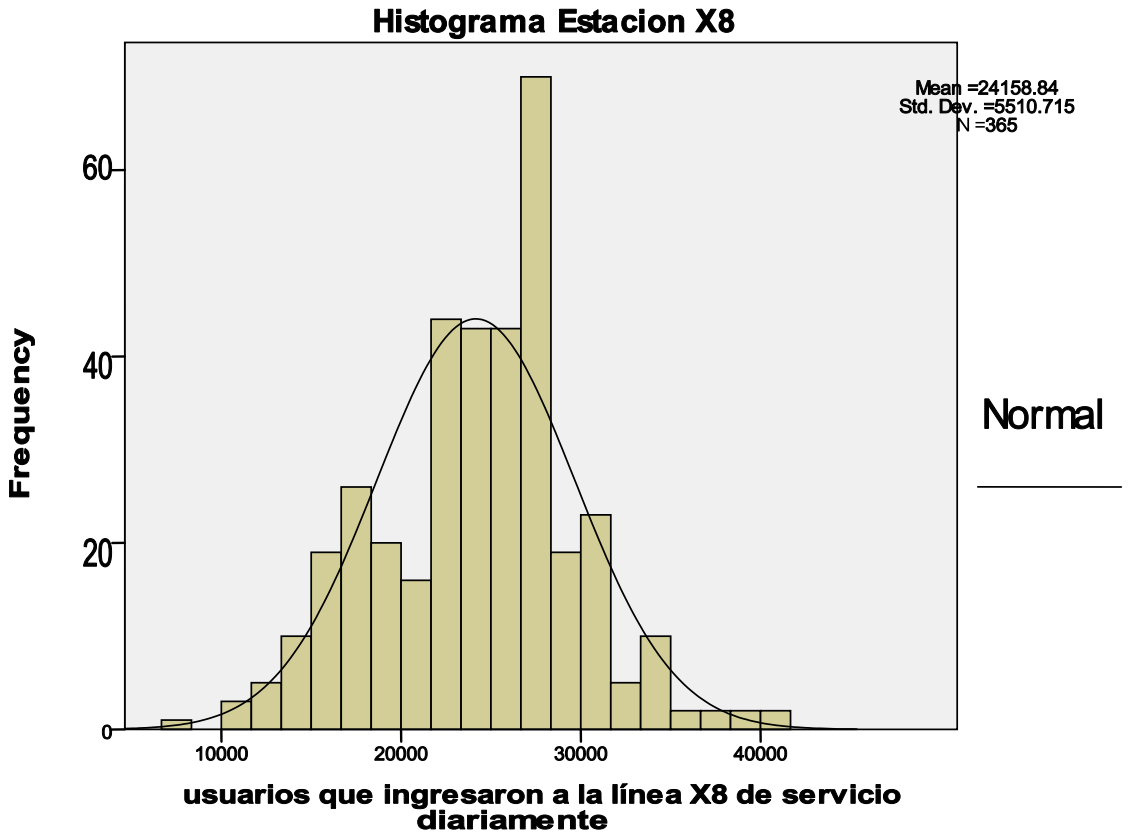


La gráfica cuantil-cuantil (QQ-plot) muestra que los datos observados en la VA X_1=Pantitlán siguen un comportamiento que puede ser descrito mediante un modelo Normal (66209.3; 2.868 E8), pero el histograma nos indica que existe un sesgo hacia la derecha, amén de que existen varios "picos" en la distribución empírica de los datos, lo cual sugiere algún otro modelo que acumule la multimodalidad.

Datos estadísticos de la VA X_8

			Statistic	Std. Error
X8	Mean		24158.84	288.444
	95% Confidence Interval for Mean	Lower Bound	23591.62	
		Upper Bound	24726.07	
	5% Trimmed Mean		24160.87	
	Median		24779.00	
	Variance		3.037E7	
	Std. Deviation		5510.715	
	Minimum		6809	
	Maximum		40856	
	Range		34047	
	Interquartile Range		7020	
	Skewness		-.143	.128
	Kurtosis		.145	.255

A continuación se muestra el histograma de la VA X_8 en la cual se notan elementos atípicos que sugieren días en que la afluencia es mayor con respecto al flujo representativo normal.



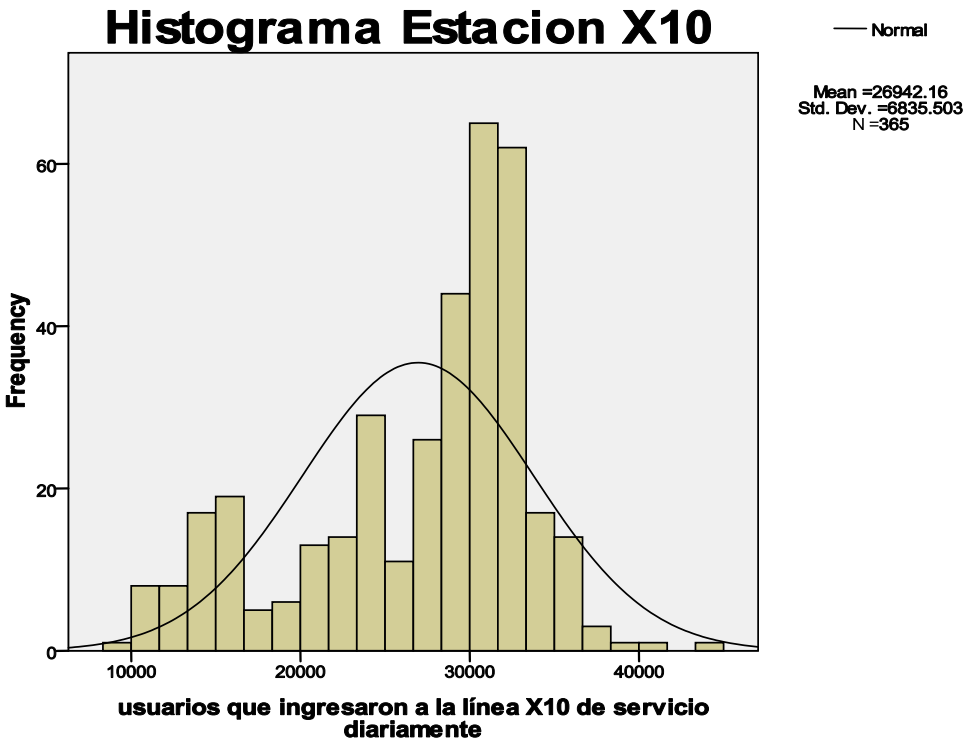
A continuación en el diagrama cuantil-cuantil se observa que la distribución Normal está lejos de proveer un ajuste adecuado al flujo de pasajeros en la estación X_8 Santa Martha.

La gráfica cuantil-cuantil de X_8 muestra una dispersión de datos empíricos que se aleja del modelo Normal en valores extremos al flujo de personas, entre los valores (10,000; 40,000). La Tabla de datos estadísticos de la VA X_{10} ofrece los valores característicos del flujo en la estación La Paz.

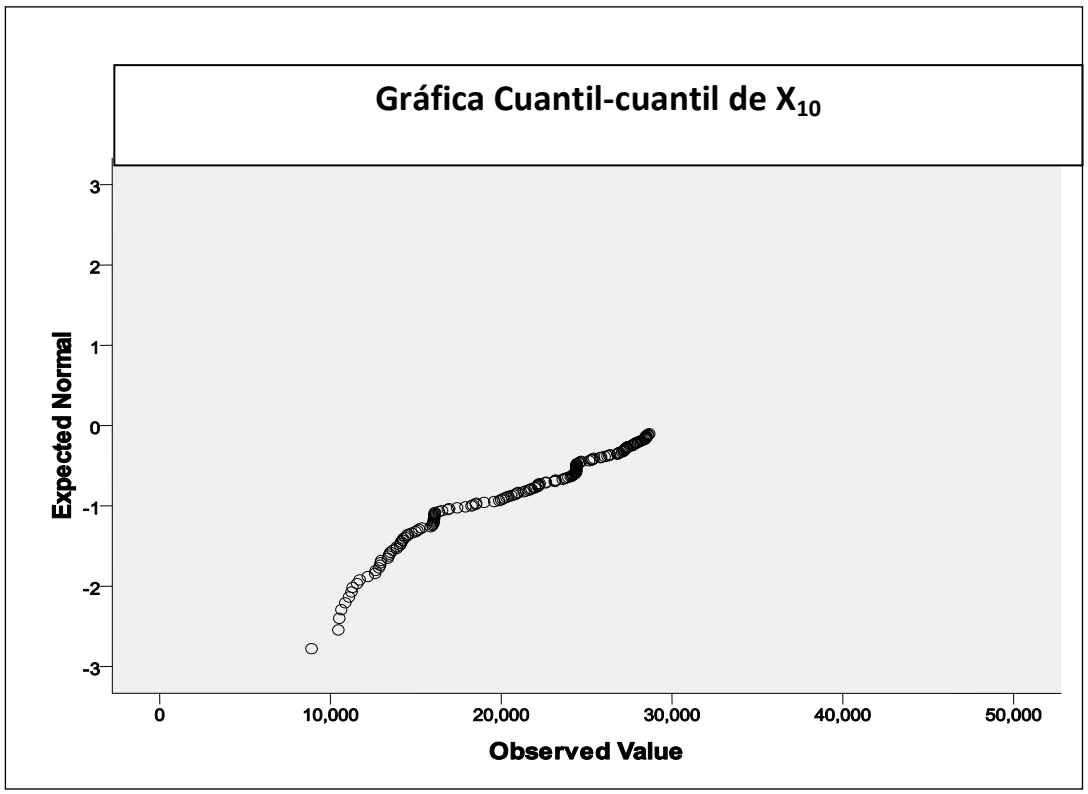
Datos estadísticos de la VA X_8

			Statistic	Std. Error
X10	Mean		26942.16	357.787
	95% Confidence Interval for Mean	Lower Bound	26238.57	
		Upper Bound	27645.74	
	5% Trimmed Mean		27214.43	
	Median		29428.00	
	Variance		4.672E7	
	Std. Deviation		6835.503	
	Minimum		8896	
	Maximum		44636	
	Range		35740	
	Interquartile Range		8429	
	Skewness		-.793	.128
	Kurtosis		-.180	.255

El histograma de X_{10} muestra una distribución irregular, con sesgo negativo que no se ajusta a una distribución Normal, lo cual puede constatarse en el diagrama cuantil-cuantil.



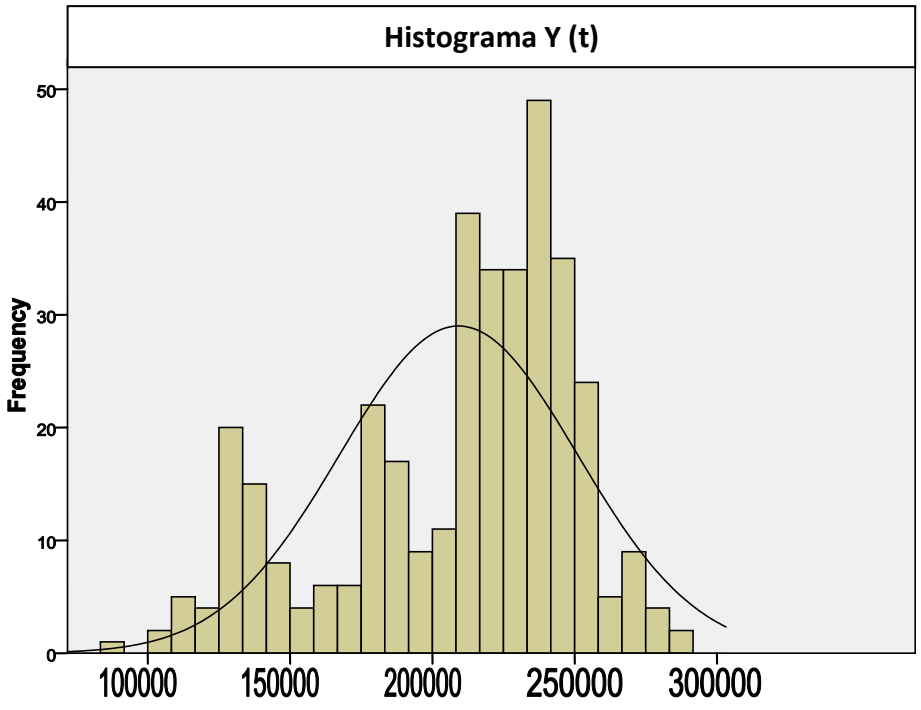
A continuación se muestra la gráfica cuantil-cuantil de X_{10} muestra una dispersión muy alejada de un modelo Normal lo cual sugiere que una distribución sesgada y con valores positivos sea más adecuada.



Las estadísticas siguientes corresponden a la VA X_{10} que representa al flujo total de pasajeros.

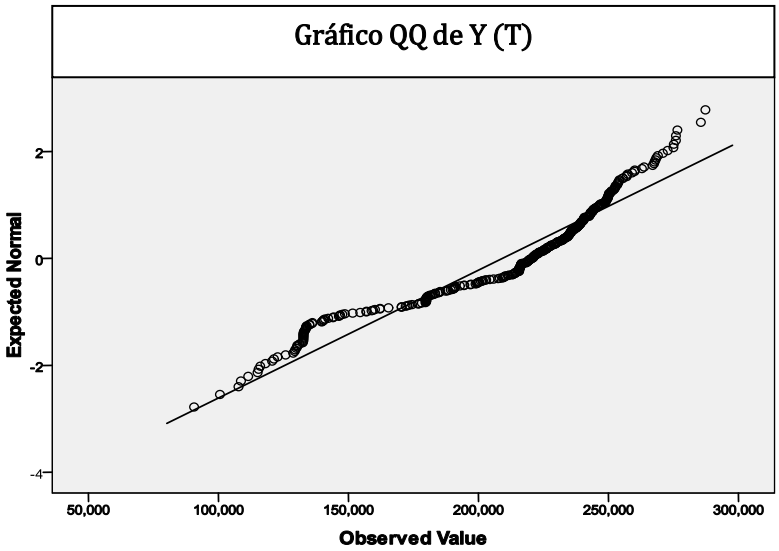
Datos estadísticos de la VA Y (T); Totales

			Statistic	Std. Error
YT	Mean		209231.66	2189.568
	95% Confidence Interval for Mean	Lower Bound	204925.87	
		Upper Bound	213537.45	
	5% Trimmed Mean		210837.83	
	Median		220502.00	
	Variance		1.750E9	
	Std. Deviation		41831.643	
	Minimum		90616	
	Maximum		287299	
	Range		196683	
	Interquartile Range		56905	
	Skewness		-.785	.128
	Kurtosis		-.281	.255



Cantidad de usuarios que ingresaron a Y (t).

El histograma de Y (T) muestra valores que tienen una dispersión pronunciada y un sesgo que no permite ajustar un modelo Normal a los datos, tal como se muestra en el gráfico siguiente.



Obsérvese que la distribución de Línea O se aleja de un comportamiento Normal tanto en los extremos como en valores centrales, esto implica la necesidad de proponer modelos cuya densidad de probabilidad ofrezca un mejor ajuste.

3.3.3 Resultados del EDA

De éste análisis preliminar, se observó que la forma en que fluyen los pasajeros hacia las estaciones de Línea O fue generalmente multimodal, es decir, la distribución presentó varios "picos" o valores extremos y esto dificultó asignar un modelo Normal que describiera adecuadamente el comportamiento de estas "oleadas" de personas confluyendo hacia un servicio de transporte.

Los diagramas de caja y bigotes muestraron que hay diferencias entre la afluencia media hacia las estaciones Xi y su mediana, este es un indicador de que el flujo de personas está sesgado, lo cual implica que existen "colas" pesadas que "inclinan" la distribución hacia uno u otro lado y permite concluir que los datos no son representativos de distribuciones equilibradas, como la Normal. El grado de sesgo es inconstante entre cada estación, por lo que será necesario analizar cada una de ellas por separado o bien agrupándolas en función de volúmenes similares de afluencia.

Del análisis en los gráficos QQ (*cuantil-cuantil*) pudo deducirse que la distribución del pasaje se explicaría adecuadamente mediante alguna distribución continua aunque no precisamente la Normal, por las razones mencionadas anteriormente. No obstante que los datos provienen de un conteo –números discretos- el hecho de que tales conteos se lleven a cabo en una escala de tiempo (continua) implicaría que *la probabilidad de un evento discreto será cero* (por ejemplo: la probabilidad de que accedan exactamente 20,000 personas cierto día, en la estación X_1).

Estas condiciones sugirieron llevar a cabo un ajuste con distribuciones continuas y positivas. Los modelos sugeridos nos permitirán eliminar errores; por ejemplo considerar una observación cuya lectura sea *-12,425.5* carece de sentido (no se puede fraccionar individuos) y por otra parte el signo negativo implicaría que se contabiliza el número de personas "que salen" de la estación $\mathbf{X_i}$.

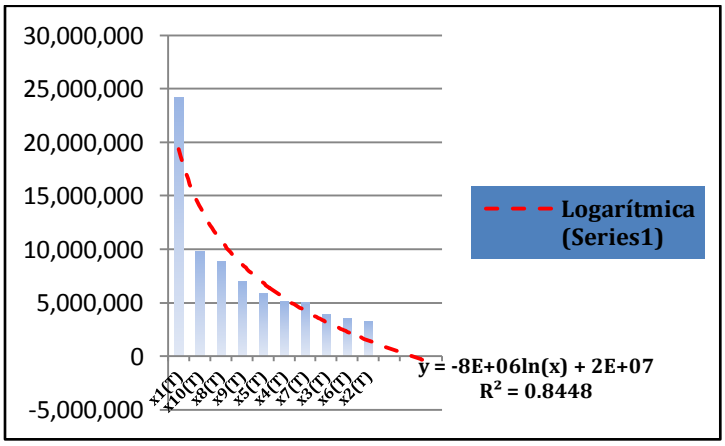
Esto descartó utilizar distribuciones continuas cuyo soporte esté en los números reales, por ejemplo la Normal, en tal caso el estudio debió limitarse a funciones de distribución continuas con soporte en los reales positivos, discretizando los resultados mediante truncamiento. Debido a la forma de los histogramas y a los valores extremos mostrados en los diagramas de "caja" -con diversos valores fuera de rango, outliers- en días específicos, es que se concluye la conveniencia de hacer una "reclasificación" del tipo de datos que se posee. Puesto que se tienen varias observaciones inusitadamente "altas" de flujo de pasajeros, se puede inferir que hay ciertas ocasiones en que "sube" el número de personas que viajan con respecto a otros días. Por tal motivo, se ha decidido clasificar en tres tipos los días para su análisis; ésta clasificación incluye los días Laborales que incluyen los días Lunes a Viernes o semana inglesa, con excepción de los días feriados o "puentes" que se incluyen como festividades del calendario laboral. Puesto que estos "puentes" incluyen un día laboral adyacente al fin de semana, se decide clasificar en un mismo tipo a los días Domingo y feriados, esto implica que sería provechoso definir una categoría diferente, que sea excluyente de las otras dos, puesto que los días Sábado no pertenecen a las otras dos categorías, entonces ellos conforman la tercera categoría en que se clasifican los días de la semana: Laborales, Sábados y Domingos o festivos (LSD); a éstas tres categorías relacionadas al comportamiento de Línea O se les agrega el comportamiento general sin discriminar el tipo de día, en cuyo caso se les designará como Totales. En la siguiente tabla (Afluencia 2010) se muestra la cantidad de personas que circuló en cada una de las estaciones de Línea O durante 2010, el orden concuerda con

volumen o número de usuarios registrado, y de ella puede verificarse cuáles son las estaciones más concurridas y por ende las de mayor importancia para el estudio.

Tabla: Afluencia 2010

RANGO	EST.	Usuarios Año (2010)
1	$X_1(T)$	24,166,626
2	$X_{10}(T)$	9,833,887
3	$X_8(T)$	8,817,978
4	$X_9(T)$	6,965,835
5	$X_5(T)$	5,834,860
6	$X_4(T)$	5,099,594
7	$X_7(T)$	4,995,321
8	$X_3(T)$	3,868,356
9	$X_6(T)$	3,510,557
10	$X_2(T)$	3,276,542
Y(T)	TOTAL	76,369,556

Gráfico de Comportamiento por Volumen de Afluencia Año 2010: En la siguiente gráfica puede verse la diferencia del volumen de pasajeros en las diferentes estaciones, sin incluir a Y (T); una curva logarítmica es ajustada a los datos a fin de señalar el comportamiento de los usuarios en las diferentes estaciones.



3.4 Análisis: Y (T), X_1, X_{10}, X_8 Totales

Cuando se haga referencia a las variables **X_i, Y (T)** con la denominación "totales", se pretende identificarlas como variables que describen el comportamiento de la Línea O durante los 365 días del año. El ajuste de las distribuciones se complicó debido a la influencia que tienen algunos valores extremos en el proceso de adecuación, los extremos corresponden a la conjunción de días Laborales –de gran demanda de usuarios- con las otras clases con menor afluencia de pasajeros. Así que los modelos propuestos no son óptimos, sin embargo son lo mejor que se puede obtener dada la muestra y los recursos utilizados.

Tabla Afluencia 2010

X_1(T)	X_{10}(T)	X_8(T)	X_9(T)	X_5(T)
24,166,626.0	9,833,887.0	8,817,978.0	6,965,835.0	5,834,860.0
X_4 (T)	**X_7 (T)**	**X_3 (T)**	**X_6 (T)**	**X_2 (T)**
5,099, 594.0	4, 995, 321.0	3, 868, 356.0	3, 510, 557.0	3, 276, 542.0

Observe que la variable que almacena el total de los accesos a Línea O es: $Y\ (T) = 76,\ 369{,}556.0$ usuarios durante el año 2010. El ajuste se realizó por volumen de afluencia en el año 2010 en el orden, que indica la Tabla de Afluencia 2010, de mayor a menor cantidad de pasajeros que abordaron Línea O en dichas estaciones: **Y (T), X_1, X_{10}, X_8, X_9, X_5, X_4, X_7, X_3, X_6, X_2,** para una muestra correspondiente a 365 días.

Ajuste para la variable aleatoria Y (T) totales. La siguiente tabla muestra los resultados de un análisis por distribuciones de probabilidad y los estadísticos representativos de tal examen.

Estadística, Totales: Y(T)	Valor	Percentil	Valor
Tamaño de la muestra	365	min	90616
Rango	1.9668E+5	5%	1.3027E+5
Media	2.0923E+5	10%	1.3375E+5
Varianza	1.7499E+9	25% (Q1)	1.8258E+5
Desviación estándar	41832.0	50% (Mediana)	2.2050E+5
Coef. de variación	0.19993	75% (Q3)	2.3949E+5
Error estándar	2189.6	90%	2.5199E+5
Asimetría	-0.78493	95%	2.6010E+5
Curtosis	-0.28146	Max	2.8730E+5

Observando los resultados de la tabla anterior, el rango indica una afluencia concurrida entre los dos puntos extremos en la curva de la distribución (Máximo y mínimo). Se observa la diferencia entre el valor de la media y la mediana, siendo la mediana mayor a la media esto implica una distribución sesgada hacia la izquierda ó con sesgo negativo, que indica el coeficiente de asimetría.

El coeficiente de variación (CV. del 19.93 %) describe la dispersión de los datos a manera que no dependan de las unidades de medición. Un CV mayor indica una mayor dispersión en la variable. El error estándar es la desviación estándar de la muestra de una distribución de muestreo. El error estándar puede utilizarse para proveer un indicio sobre la magnitud de la incertidumbre conjuntamente al coeficiente de variación.

Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles son calculados para ordenar los valores de la variable del menor al mayor y entonces hallar el valor que corresponde al percentil.

La tabla anterior muestra los estadísticos más importantes de la VA Y (T), donde se desprende que la distribución está sesgada negativamente y con una variación aproximada al 20%

alrededor del valor central, se trata de una distribución cuya curtosis negativa la convierte en "platicurtica" (aplanada ó "chata") con valores persistentes y asimétrica ó sesgada .

Distribuciones propuestas

Solamente se propuso un modelo que ajusta "adecuadamente" bajo la prueba bondad de Kolmogorov-Smirnov (KS), la distribución general de valor extremo, DVE. Esto se debe al enorme volumen de personas que concurren en Y (T), éste comportamiento equivale a "ríos de gente" en circulación y es la única distribución que puede tomarse en cuenta, puesto que obtiene el mejor rango de ajuste entre las tres pruebas de bondad. Se observa que bajo el criterio Ji-cuadrado, no hay disponible un valor para el estadístico de prueba.

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
4	General de Valor extremo	Estadística	Rango	Estadística	Rango	Estadística NA	Rango NA
		.07608	1	37.325	11		

Parámetros de la distribución Y (T) totales

#	**Distribución**	**Parámetros**
1	General de valor extremo	k=-0.68699; σ=46113.0 μ=2.0294E+5

Detalles del ajuste, bajo el criterio de Kolmogorov-Smirnov (KS**): Detalles del ajuste:** en la siguiente tabla se muestran los valores de los estadísticos de prueba (D) para KS y (A) para AD. En este caso el estadístico de prueba Anderson-Darling (AD) ubica la distribución propuesta en el lugar 37 y como puede verse en la tabla, rechaza la distribución propuesta y no le asigna un valor de aceptación, por tal motivo, la alternativa es utilizar el estadístico de KS y mostrar porqué la distribución general de valor extremo sí es un modelo adecuado para explicar el flujo de pasajeros en la Línea O.

Detalles del ajuste, distribución General de valor extremo [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	365 0.07608 0.02771 1				
α	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**
Valor crítico	0.05616	0.06401	0.07108	0.07946	0.08527
¿Rechazar?	**Sí**	**Sí**	**Sí**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	365 37.325 11				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

La hipótesis sobre la forma de la distribución es rechazada en el nivel de significación elegido (alfa) si el estadístico de prueba de Kolmogorov-Smirnov (*D)* es mayor que el valor crítico obtenido en una tabla de valores..La prueba Kolmogorov-Smirnov no rechaza la distribución propuesta en los niveles de significación alfa 2% y 1%, lo cual implica no rechazar la hipótesis nula "*los datos provienen de una Distribución general de Valor Extremo, (DVE) cuyos parámetros son*":

$$k = 0.68699;\ \sigma = 46113.0\ \mu = 2.0294E+5$$

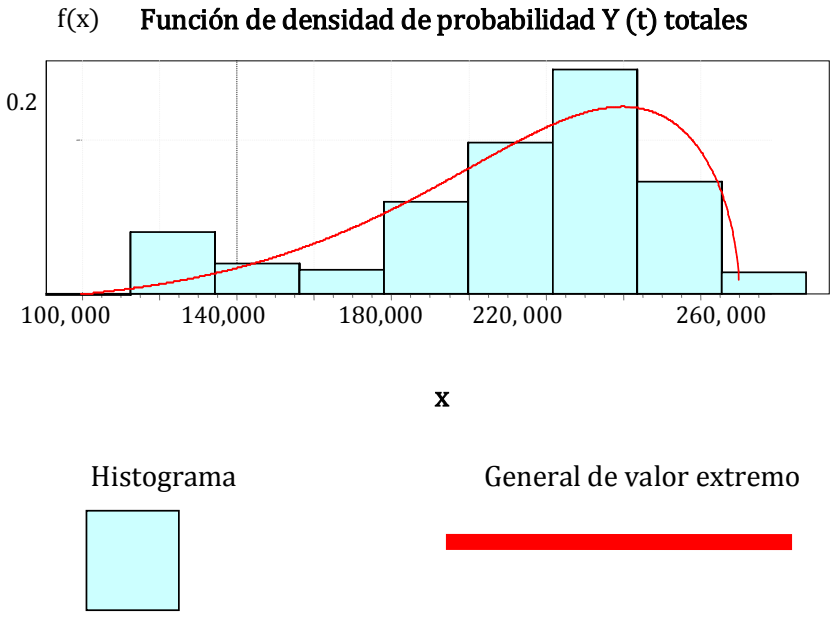
Hay tres tipos de DVE que pueden combinarse en una sola distribución con parametrización común, propuesta por von Mises (1954) y Jenkins (1955), que se conoce como la Distribución

Generalizada de Valores Extremos, DGVE o GEV por sus siglas en inglés. La forma de esta distribución es$G_\varepsilon(X) = \ exp\ \left\{-(1 + \ kx)_{y+}^{-1/k}\right\}$

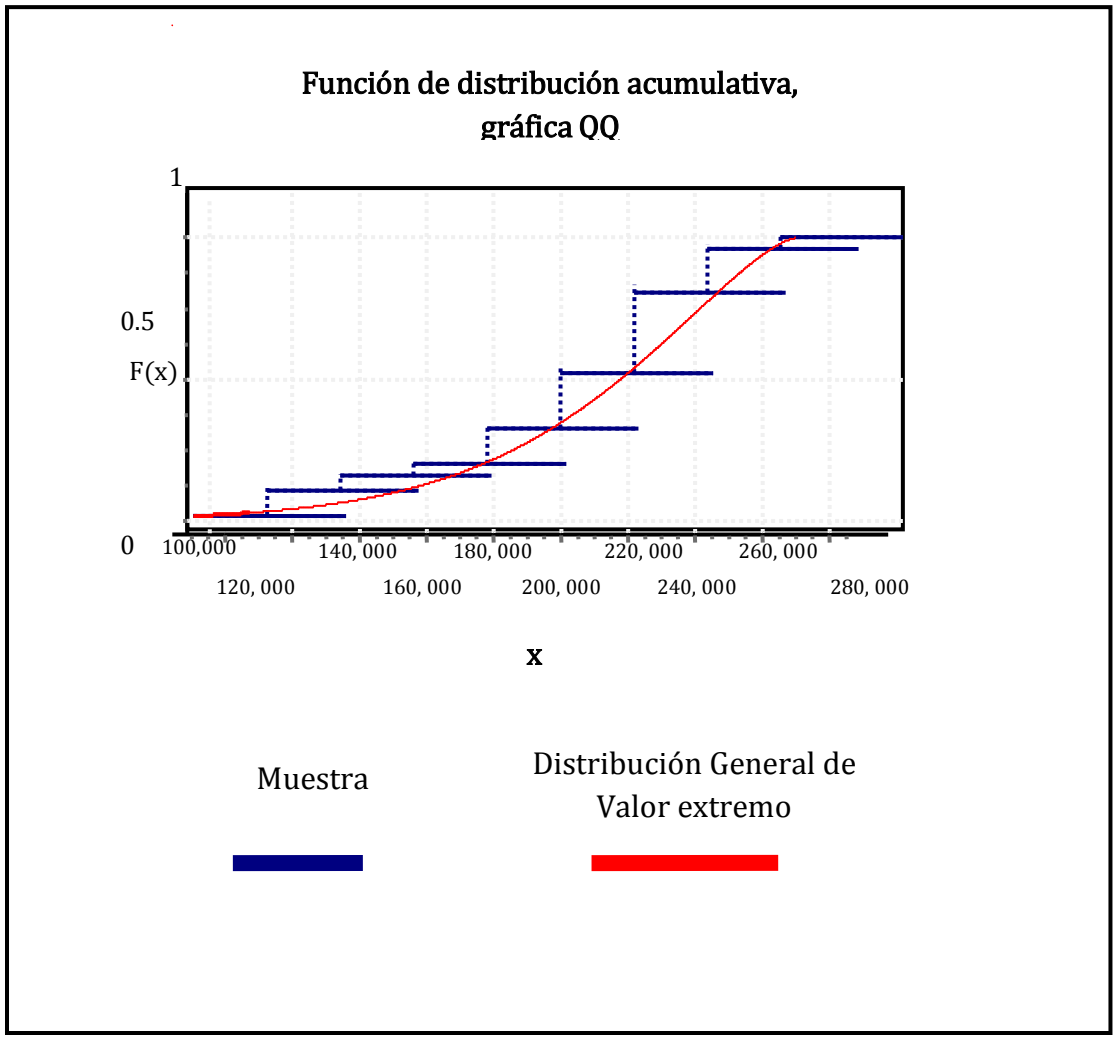
Donde y+ es el máximo entre (y, 0). Para $k > 0$ se tiene la distribución de Frechét con: $\alpha = \frac{1}{k}$

Para $k < 0$ se tiene la distribución Weibull con $\alpha = \frac{-1}{k}$

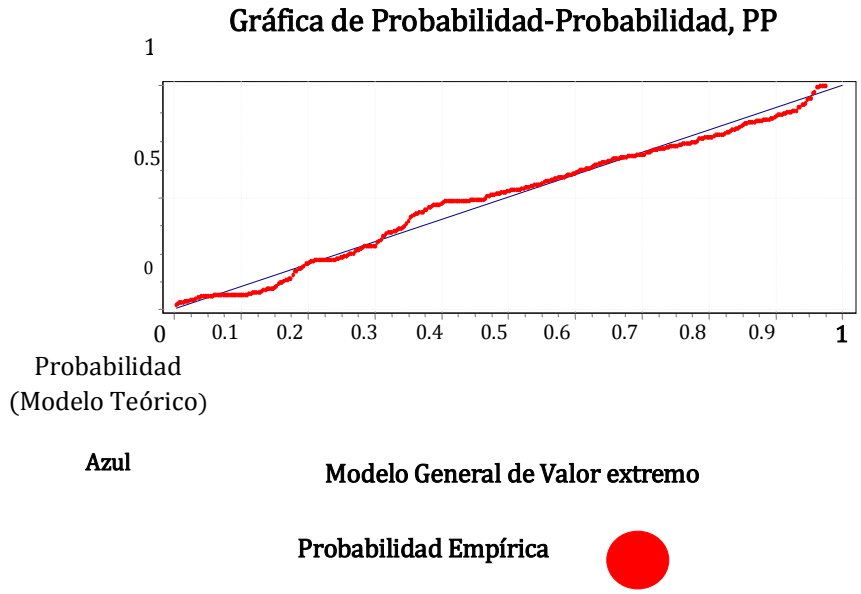
La distribución de Gumbel surge cuando $k \ \rightarrow cero$, el parámetro (k) se conoce como parámetro de forma. Los resultados de las pruebas de bondad de ajuste sugieren que el modelo propuesto es adecuado para describir el flujo de pasajeros que transitan por Y (T).



La gráfica anterior muestra el ajuste de la DVE con respecto al histograma, la distribución empírica (en rojo) está bastante sesgada. La DVE ajusta adecuadamente en las partes centrales del histograma y débilmente en los extremos, sin embargo es el modelo más adecuado para datos tan dispersos.. Ver la siguiente gráfica.



La gráfica muestra la función de distribución acumulativa empírica (en azul) y el ajuste que hace la distribución general del valor extremo (en rojo), puede observarse que la grafica de la DVE está muy próxima a los valores correspondientes al flujo de pasajeros, esto indica que la probabilidad de obtener valores en el intervalo es mayor a la pronosticada por el modelo, lo cual puede confirmar la gráfica siguiente.



La gráfica de probabilidad-probabilidad (arriba) muestra que el modelo no es adecuado para los déciles 4 y 5 y también está poco ajustada a los extremos, no se proponen otros modelos porque ajustan aún más débilmente que la DVE y son rechazados por la prueba AD.

Ajuste para la VA X_1: totales.

Estadísticas X_1	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	365	Mínimo	27, 855
Rango	82, 244	5%	38, 558
Media	66, 210	10%	41, 286
Varianza	2.868E +8	25% (Q1)	55, 111
Desv. Estándar	16, 935	50% Mediana	65, 825
Coef. de Variación	.25578	75% (Q3)	7895
Error estándar	886.44	90%	87, 862
Asimetría	-.02480	90%	87, 862
Curtosis	-.70221	Máximo	1.1010 E +5

De la tabla anterior puede verse que el rango de operación de X_1 es mucho menor que en Y (T) –menos de la mitad- pero que el coeficiente de variación es mucho mayor (un coeficiente de variación CV = 25.5778 %) en la VA X_1 que en el total de Y (T), ésta terminal es por la que fluye el mayor número de pasajeros y tal variabilidad confirma que debe existir una diferenciación entre los diversos días de estudio. Nuevamente la distribución de los datos corresponde a un modelo sesgado a la derecha positivamente -la media es mayor que la mediana- y una distribución "aplanada" (platicurtico, que indica la existencia de valores redundantes.

Distribuciones propuestas: hay concurrencia de criterios en el ajuste, bajo las pruebas AD y KS, el criterio Ji-cuadrado es confirmatorios de los otros dos.

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	General de Valor Extremo	0.05919	1	1.2828	1	52.218	7
10	Log-Pearson 3	0.06394	2	1.6215	2	33.971	1
14	Weibull	0.07269	3	1.8495	3	51.757	6

En la siguiente tabla se muestran los parámetros de las distribuciones propuestas

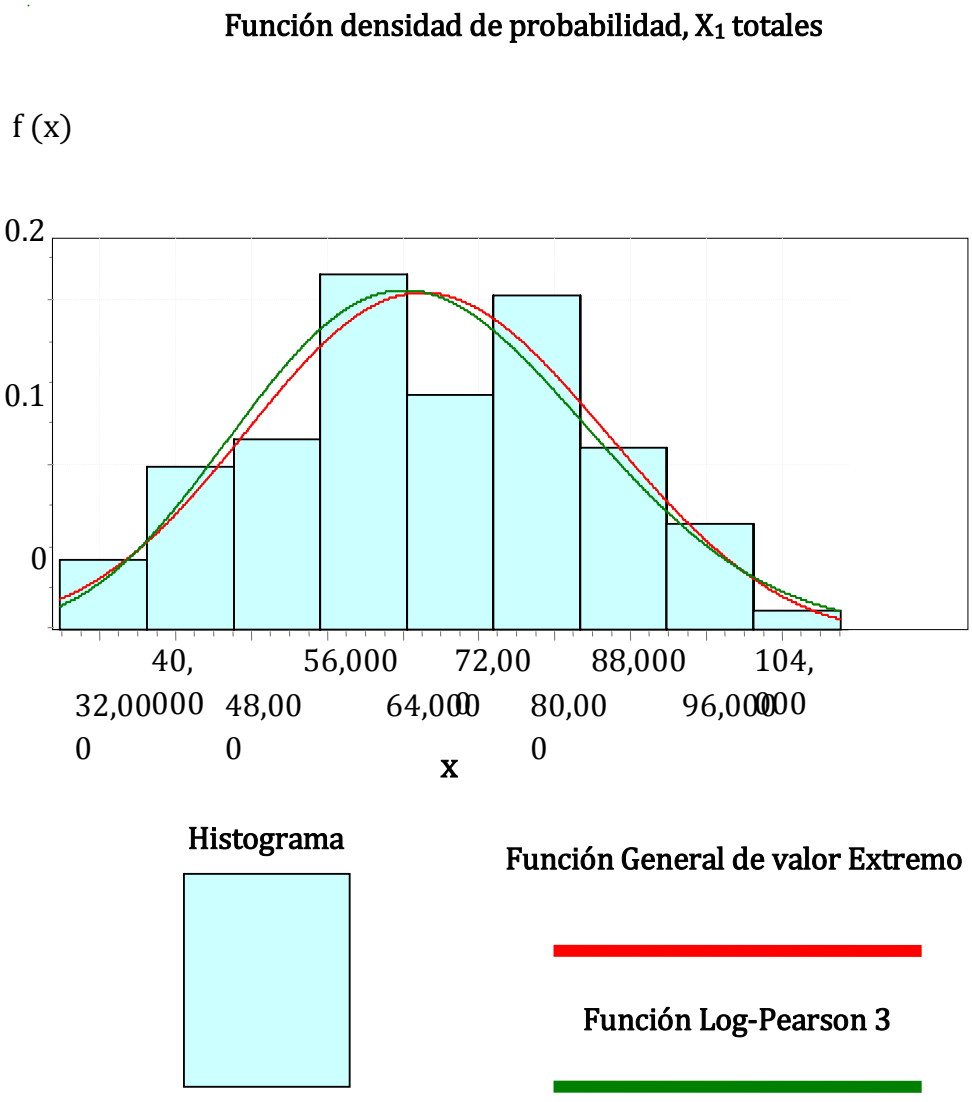
Parámetros de las distribuciones propuestas X_1 totales.

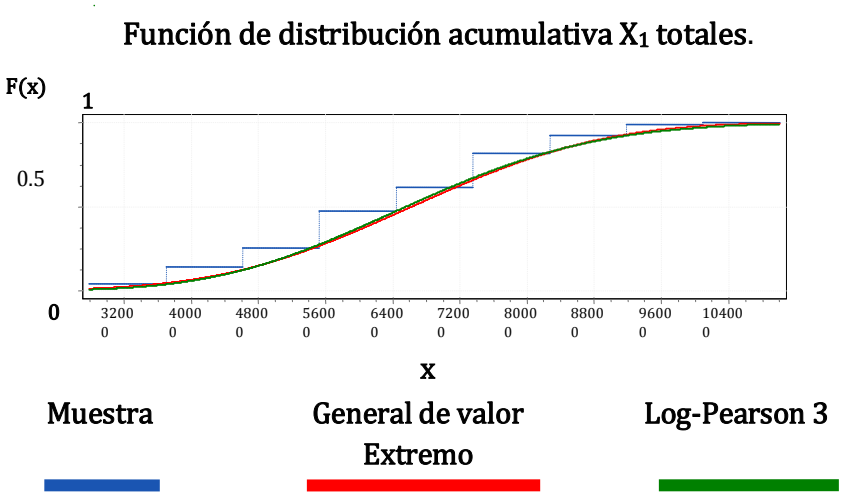
#	**Distribución**	**Parámetros**
1	General Valor extremo	k=-0.29706 σ=17236.0 μ=60287.0
2	Log-Pearson 3	α=12.127 β=-0.07885 γ=12.021
3	Weibull	α =4.5349 β =72360.0

Detalles del ajuste: en la siguiente tabla se muestran los valores de los estadísticos de prueba (D) para KS y (A) para AD, la prueba Ji-cuadrado es confirmatoria de las otras dos.

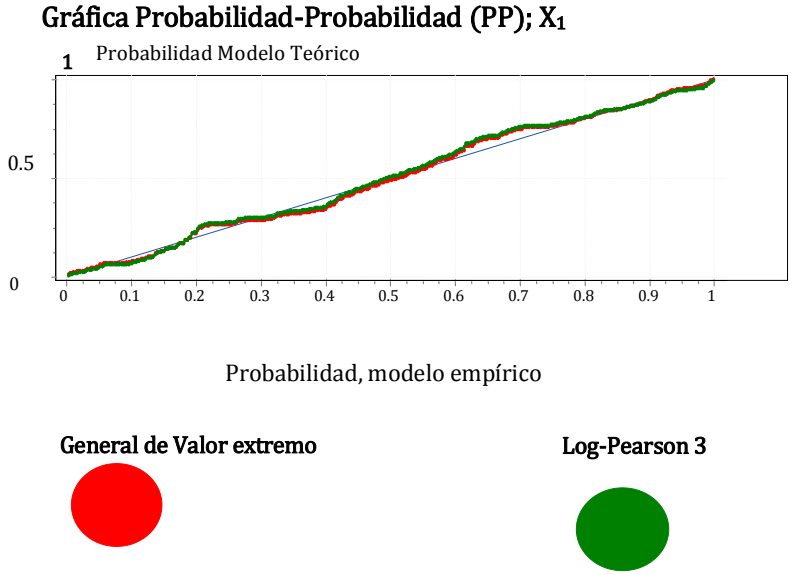
X_1: totales. General de valor extremo [#1]					
Kolmogorov-Smirnov (D)					
Tamaño de la muestra	365				
Estadística	0.05919				
Valor P	0.14892				
Rango	1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.05616	0.06401	0.07108	0.07946	0.08527
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Anderson-Darling (**A**)					
Tamaño de la muestra	365				
Estadística	1.2828				
Rango	1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Ji-cuadrado					
Grados de libertad	8				
Estadística	52.218				
Valor P	1.5274E-8				
Rango	7				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	11.03	13.362	15.507	18.168	20.09
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

Histograma y densidad X_1 totales: la gráfica muestra dos distribuciones propuestas para ajustarse al flujo de X_1 y como se mostró en la tabla anterior los criterios de ajuste coinciden en al menos dos criterios de bondad de ajuste.





La gráfica de la distribución acumulativa muestra dos modelos propuestos, el DVE y Log-Pearson 3; los cuales podrían ser utilizados indistintamente para modelar el flujo de pasajeros en X_1.



La gráfica PP de X_1 muestra la congruencia entre los modelos DVE y Log-Pearson3, que ajustan adecuadamente al flujo en X_1.

Ajuste para la VA X_{10} totales.

Estadística totales X10	Valor	Percentil	Valor
Tamaño de la muestra	365	Min	8896
Rango	35740	5%	13374.0
Media	26942.0	10%	15289.0
Varianza	4.6724E+7	25% (Q1)	23394.0
Desviación estándar	6835.5	50% (Mediana)	29428
Coef. de variación	0.25371	75% (Q3)	31823.0
Error estándar	357.79	90%	33591.0
Asimetría	-0.7934	95%	35468.0
Curtosis	-0.18021	Max	44636

Los estadísticos muestran una gran variabilidad (del 25.37 %) y sesgo negativo en la distribución lo que implica colas-pesadas, por tal motivo se elige el criterio de Anderson-Darling como criterio principal de ajuste al flujo de pasajeros en X_{10}.

Distribuciones propuestas X_{10} totales.

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.13343	4	11.194	1	93.868	1
14	Weibull	.13189	3	11.735	2	183.46	3

Parámetros de las distribuciones

#	**Distribución**	**Parámetros**
1	Burr	k=61.609 α=4.9832 β=67333.0
2	Weibull	α =3.9844 β =29801.0

Detalles del ajuste X_{10} totales , **Burr** [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística **Valor P** Rango	365 0.13343 **3.9486 E-6 = .0000039468** 4				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.05616	0.06401	0.07108	0.07946	0.08527
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**
Anderson-Darling (**A^2**)					
Tamaño de la muestra Estadística Rango	365 11.194 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

Aunque se rechaza la hipótesis nula “los datos provienen de las distribuciones propuestas” el p-valor permite seleccionar un ajuste moderado pero conveniente.

Ante un mayor valor del p-valor, menos se puede creer en que la relación observada entre las variables en la muestra sea un indicador confiable de la relación entre las variables respectivas de la población y viceversa. [44]

La VA X_{10} posee valores extremos que bien pueden ser "outliers"; tal punto corresponde al Día 8 FEB 2010 con un registro de 44636 usuarios, en un día Laboral, veamos qué pasa si eliminamos tal registro y lo cambiamos por alguno cercano a la mediana y dos desviaciones estándar.

Nuevo punto = [29428 + (2*6835)]= 43098.

A pesar de eliminarse el "outlier" y reemplazarse por un dato cercano al comportamiento general, persiste la No obtención de una distribución que no sea rechazada, en tal caso se propone utilizar la mejor distribución bajo el criterio AD y separar los datos en Laborales, Sábados y Domingos para observar el comportamiento de X_{10} por tipo de día.

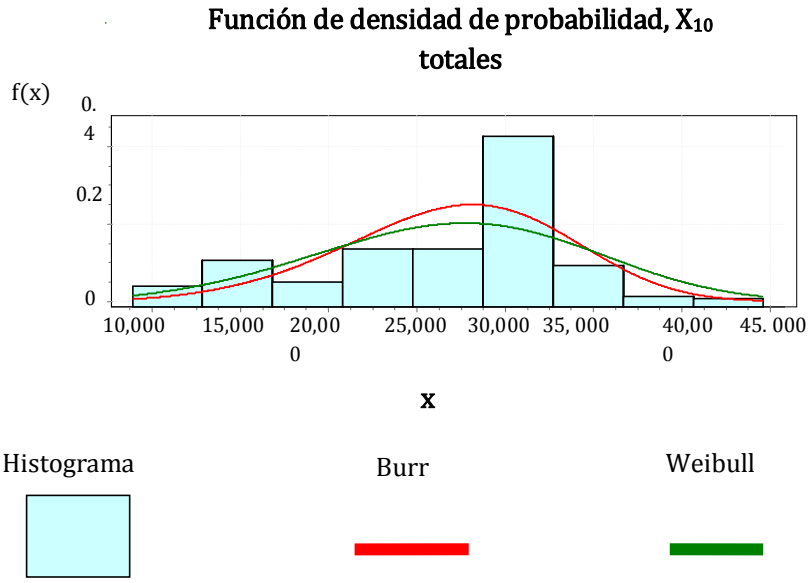
Las hipótesis nula y alternativa son respectivamente:

H_0: Los datos siguen la distribución propuesta

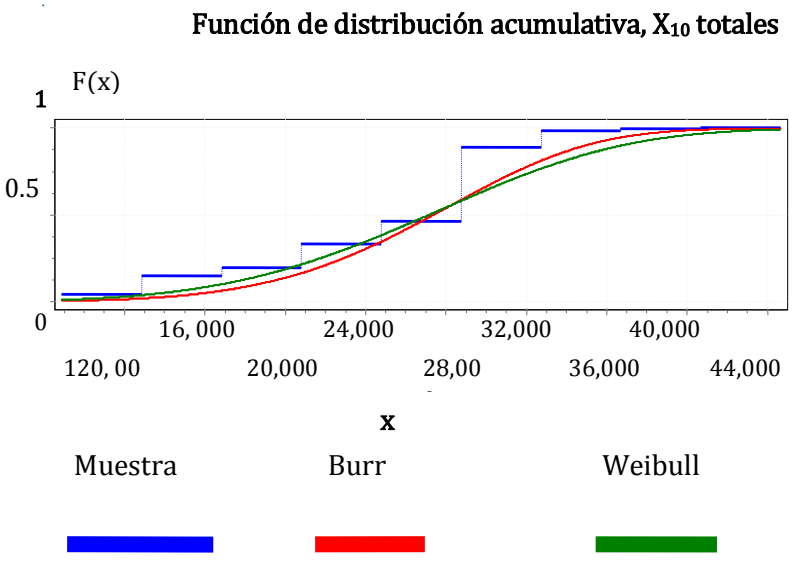
Ha: Los datos no siguen la distribución propuesta

En éste caso no se está postulando que H_0 sea necesariamente falsa y que la hipótesis Ha sea verdadera. Lo que se está exponiendo es que la hipótesis nula es improbable de producir valores más extremos que los valores observados. Cuando no se rechaza ("*se acepta*") la hipótesis nula no se prueba que ésta sea necesariamente verdadera, solo hallamos evidencia de que "no es muy improbable que la hipótesis nula sea verdadera". Especificando un nivel de confianza, indirectamente hallamos valores de la prueba estadística que nos llevan a rechazar. Las cotas entre la región de aceptación y la región de rechazo son los llamados *valores críticos*. El uso de la región de rechazo evita calcular el p-valor, rechazar si el valor observado está en la región de rechazo y aceptar en otro caso. Aunque es preferible hallar y reportar el p-valor en lugar de emitir un veredicto simple de "rechazar" ó "aceptar propuesto es el mejor que se puede obtener.

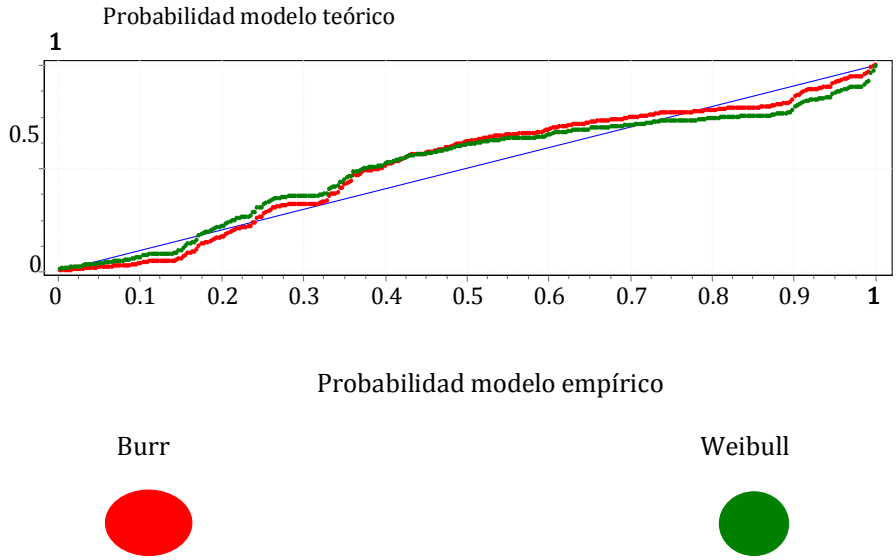
A continuación, en la siguiente gráfica se muestra el histograma y gráfico de densidad.



De la gráfica anterior puede verse que la distribución que sigue el flujo de pasajeros contiene varios "picos" que indican una gran variabilidad en el "tamaño" del flujo de personas en ésa estación. Puesto que el modelo propuesto no es totalmente aceptado se decide particionar los datos y buscar cual es la distribución de X_{10} en días Laborales, Sábados y Domingos.



Probabilidad-Probabilidad X_{10} totales



Las gráficas QQ y PP muestran la comparación entre los modelos Burr y Weibull, que tienen poca diferencia de ajuste, sin embargo ambos se separan de los valores óptimos que indica la recta en color azul, el modelo empírico. La decisión es aceptar momentáneamente los modelos propuestos e indagar si el ajuste progresa particionando los días en clases mutuamente excluyentes

Estadística para la VA X_8 totales

Estadística totales: X8	Valor	Percentil	Valor
Tamaño de la muestra	365	Mín.	6809
Rango	34,047	5%	14, 606
Media	24, 259	10%	16, 573
Varianza	3.0368 E+7	25% (Q1)	20, 656
Desv. Estándar	5510.7	50% Med.	24, 779
Coef. De Variación	.2281	75% (Q3)	27, 675
Error Estándar	288.44	95%	33, 196

Asimetría -0.14294 Con una curtosis de 0.14502 Y un Máximo de 40856. La tabla de estadísticas anterior indica que nuevamente hay una diferencia entre la media y la mediana aunque mínima, el coeficiente de variación es pequeño comparado con las otras variables aleatorias lo

que habla de una regularidad estadística en la VA X8. La distribución es ligeramente sesgada negativamente y es aplanada.

Distribuciones propuestas, bajo el criterio AD X8 totales

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
14	Weibull	.05655	2	1.6922	1	41.108	1
1	Burr	.05604	1	1.8613	2	51.693	2

En este caso no hay discrepancia notable entre los criterios de la prueba AD y KS, de manera que las distribuciones propuestas son las más adecuadas. Ambas distribuciones son casi idénticas y ajustan muy similarmente.

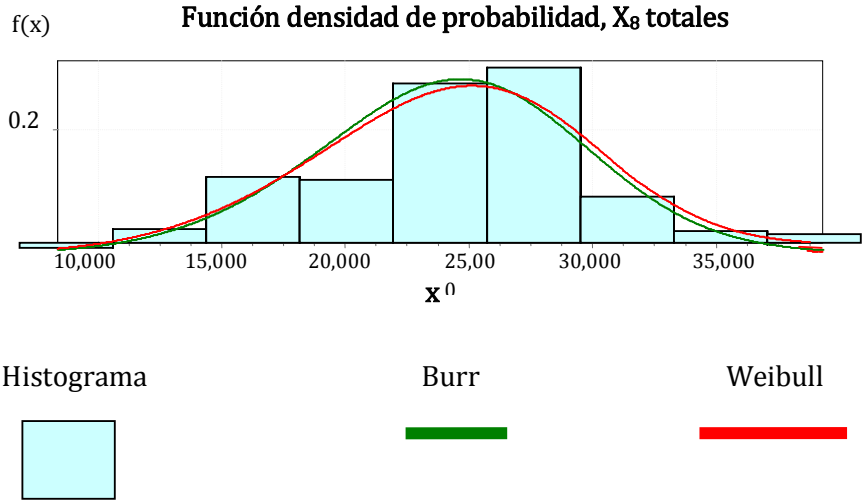
Parámetros de las distribuciones propuestas X8 totales.

#	**Distribución**	**Parámetros**
1	Weibull	α =4.9905 β =26270.0
2	Burr	k=6.0674 α=5.4951 β=35694.0

Detalles del ajuste X8 totales., **Weibull** [#14]	
Kolmogorov-Smirnov	
Tamaño de la muestra	365
Estadística	0.05655
Valor P	0.18629
Rango	2

α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.05616	0.06401	0.07108	0.07946	0.08527
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	365 1.6922 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	8 41.108 1.9899E-6 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	11.03	13.362	15.507	18.168	20.09
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

De los datos que proporciona la tabla anterior puede verse que la prueba AD es la que mejor se ajusta a los datos del flujo de pasajeros y que en ambas pruebas de bondad de ajuste, se rechaza la distribución propuesta en el nivel alfa de .20 A continuación se muestran los diagramas del histograma con la densidad propuesta ajustada a éste.



Los modelos propuestos (Burr y Weibull) ajustan de manera muy similar y pudieran ser utilizados indistintamente, se muestran las distribuciones con buen ajuste al histograma, puede verse que la distribución está sesgada y es casi simétrica, reflejado en asimetría y curtosis, moderados.

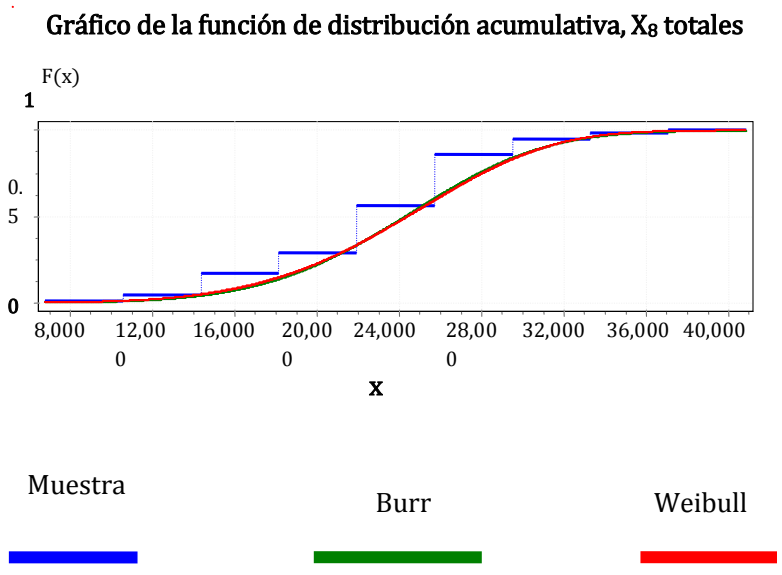
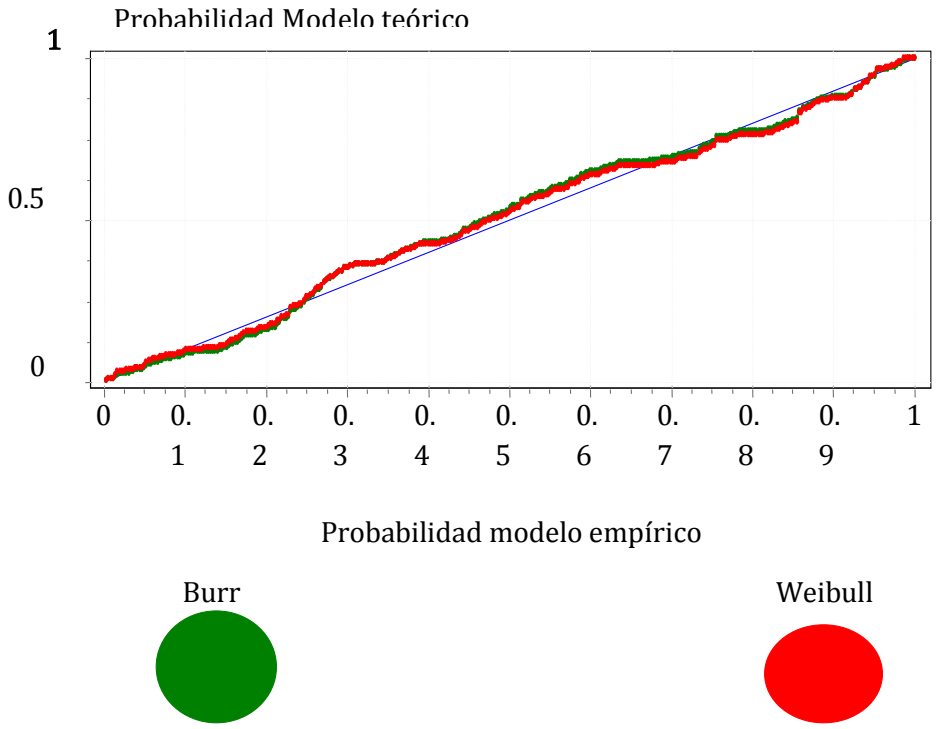


Gráfico Probabilidad-Probabilidad X_8 totales



Las gráficas cuantil-cuantil y probabilidad-probabilidad (QQ y PP) muestran el ajuste de las distribuciones Burr y Weibull al flujo empírico de pasajeros, se observa que ambos modelos ajustan muy adecuadamente y describen apropiadamente la demanda de pasajeros en X_8..

3.5 Análisis: días Laborales, Y (T)

Estadística Laborales Y (T)	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	248	Min	1.4406E+5

Rango	1.4324E+5	5%	1.9097E+5
Media	2.3157E+5	10%	2.0956E+5
Varianza	4.2403E+8	25% (Q1)	2.1874E+5
Desviación estándar	20592.0	50% (Mediana)	2.3287E+5
Coef. de variación	0.08893	75% (Q3)	2.4415E+5
		90%	2.5414E+5
Error estándar	1307.6	95%	2.6765E+5
Asimetría	-0.33445	Max	2.8730E+5
Curtosis	1.2575		

La tabla anterior muestra los valores de los estadísticos principales de la VA Y (T) en días Laborales, es de notar que el coeficiente de variación (CV) es de 8.89%, que implica muy poca dispersión de los datos. Comparando con el CV de Y (T)-Total que fue de 19.93 %; esto indica un acierto al particionar los datos en clases independientes para su análisis. Puesto que la mediana es mayor a la media, esto induce al sesgo negativo de la distribución.

Distribuciones propuestas bajo el criterio AD Laborales Y (T).

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
3	Gamma	0.07689	6	1.261	1	10.371	2
7	Gaussiana Inversa	0.06532	2	1.3043	2	11.531	4

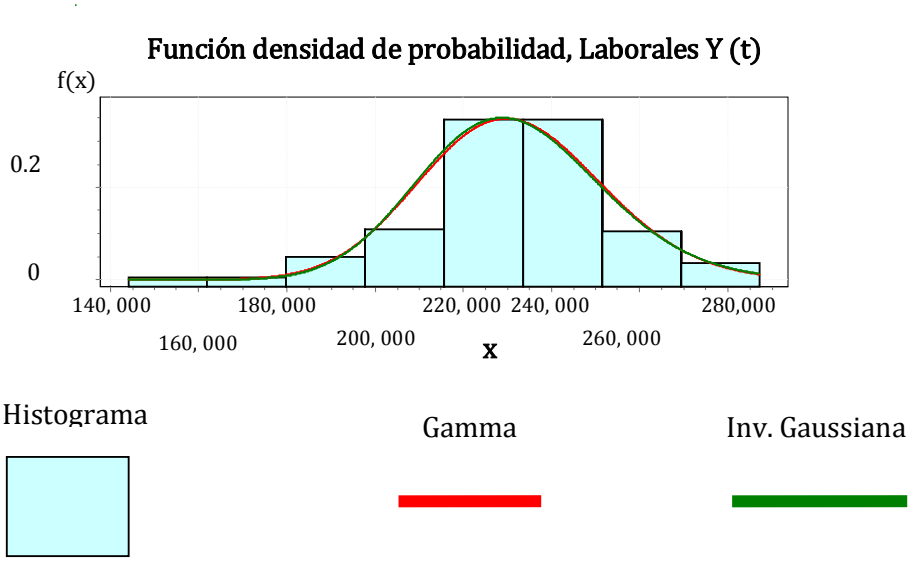
Parámetros de las distribuciones propuestas

	Distribución	**Parámetros**
1	Gamma	α =126.46 β=1831.2, gama = 0.0
2	Inv. Gaussiana	λ=2.9284E+7 μ=2.3157E+5

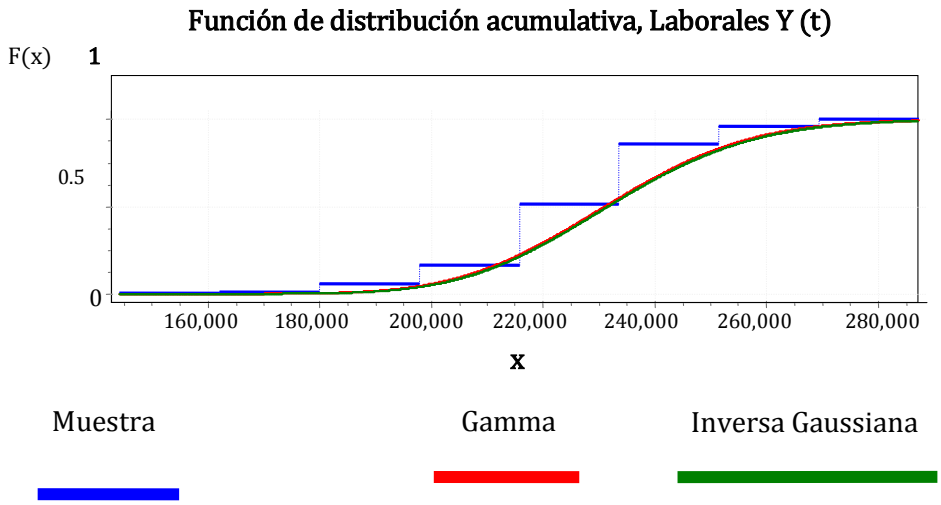
Detalles del ajuste Laborales Y (T), Gamma [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	248 0.07689 **0.10106** 6				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.06814	0.07766	0.08623	0.09639	0.10344
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	248 1.261 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Ji-cuadrado					
Grados de libertad Estadística Valor P Rango	7 10.371 0.16848 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	9.8032	12.017	14.067	16.622	18.475
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**

Los detalles del ajuste nos indican que la distribución teórica propuesta (Gamma) es un modelo confiable a niveles de significancia menores al 2%, donde existe un criterio unificado por las tres pruebas de bondad para No rechazar la hipótesis nula "los datos provienen de una

densidad Gamma", de la cual se muestra el diagrama de densidad ajustado al histograma a continuación.

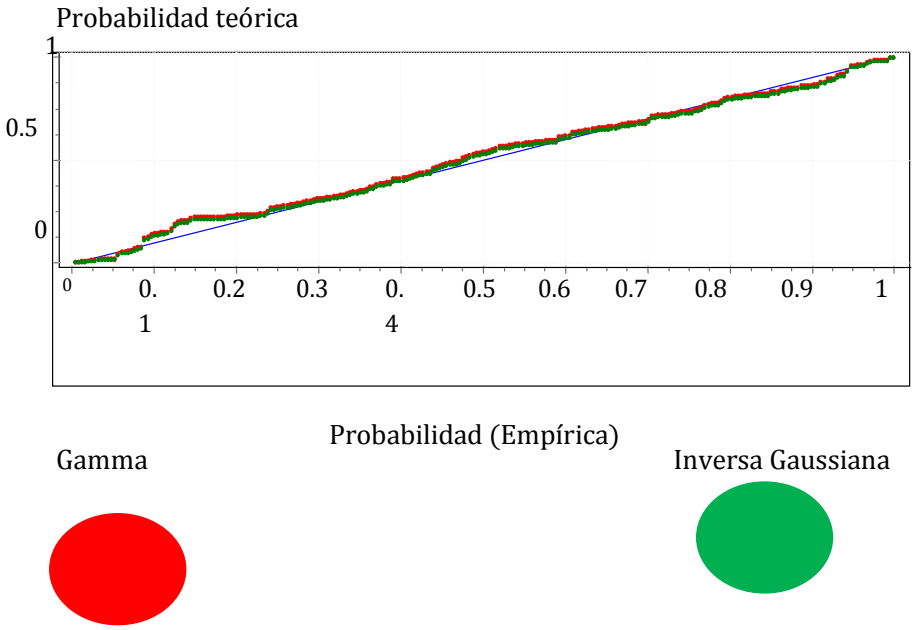


La gráfica del histograma muestra un buen ajuste de los modelos Gamma e Inversa-Gaussiana al flujo de pasajeros en Y (T) Laborales, se observa poca dispersión de los datos tal como lo indica el coeficiente de variación. La distribución se observa con sesgo negativo.



La gráfica de la distribución acumulativa muestra el ajuste muy similar entre ambos modelos.

Gráfica Probabilidad-Probabilidad Laborales Y (t)



El ajuste de ambos los modelos de distribución Gamma e Inversa-Gaussiana es bastante adecuado al flujo de pasajeros en la Línea O en día laborales, lo cual confirma la gráfica PP.

AJUSTE: X_1 Laborales

Estadísticas X1 Laborales	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	248	Mínimo	43,776
Rango	66, 323	5 %	54, 881.0
Media	73,246	10 %	55, 575.0
Desviación Estándar	12,926.0	50 % Mediana	75, 162.0
Coef. De Variación	.17647	75 % Q3	83, 045.0
Error estándar	820.79	90%	9, 180.0

Continuación Estadísticos	**Valor**	**Percentil**	**Valor**
Asimetría	0.03603	95%	94573.0
Curtosis	-0.75116	Max	1.1010E+5

La tabla de estadísticos anterior indica nuevamente una distribución sesgada ligeramente a la derecha y ligeramente aplanada, resalta el nivel del coeficiente de variación, (CV = 17.647 %) alto a comparación de los anteriores. Con un rango anual de pasajeros de 66,323 usuarios y valores mínimos de 43, 776 y Máximo de 110, 100 por año.

Distribuciones propuestas, según el criterio de la prueba AD

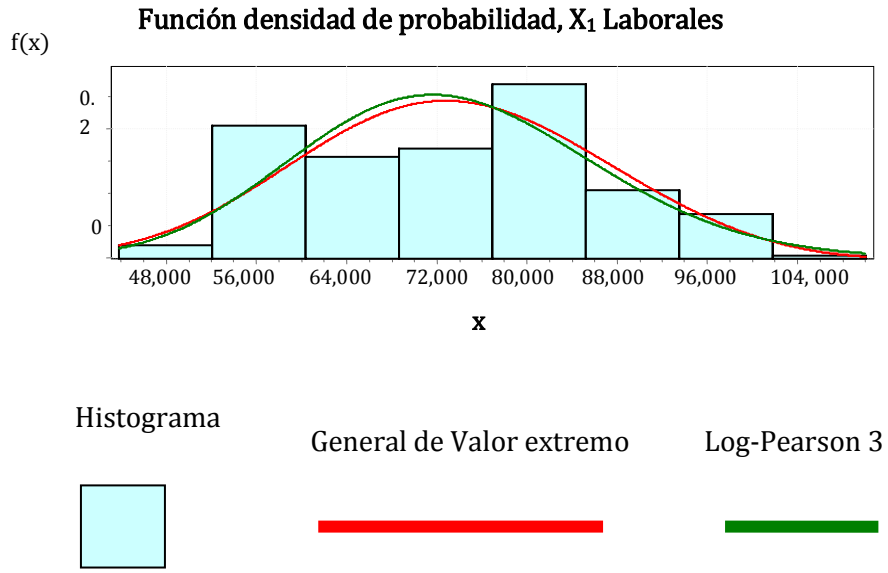
#	Distribución, X1 Laborales	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	General de Valor extremo	0.07017	2	1.5754	**1**	22.792	2
10	Log-Pearson 3	0.07984	4	2.0628	**2**	27.815	3

Parámetros de las distribuciones propuestas X_1 Laborales.

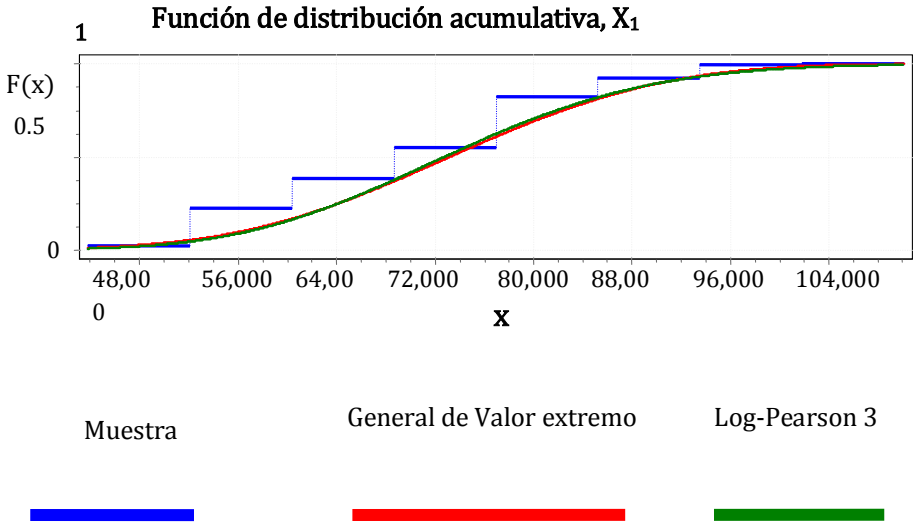
#	**Distribución**	**Parámetros**
1	General de valor extremo	k=-0.28318 σ=13072.0 μ=68638.0
2	Log-Pearson 3	α=46.766 β=-0.02647 μ=12.423

Detalles del ajuste X_1, General de Valor Extremo [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	248 0.07017 0.16585 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.06814	0.07766	0.08623	0.09639	0.10344
¿Rechazar?	Sí	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	248 1.5754 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	Sí	**No**	**No**	**No**	**No**

La tabla muestra los valores críticos de los estimadores (D y A) de las pruebas KS y AD que rechazan la hipótesis de que los datos provienen de una distribución de valor extremo para un nivel de confianza del 20%. En los demás niveles la hipótesis no se rechaza y se acepta el ajuste.

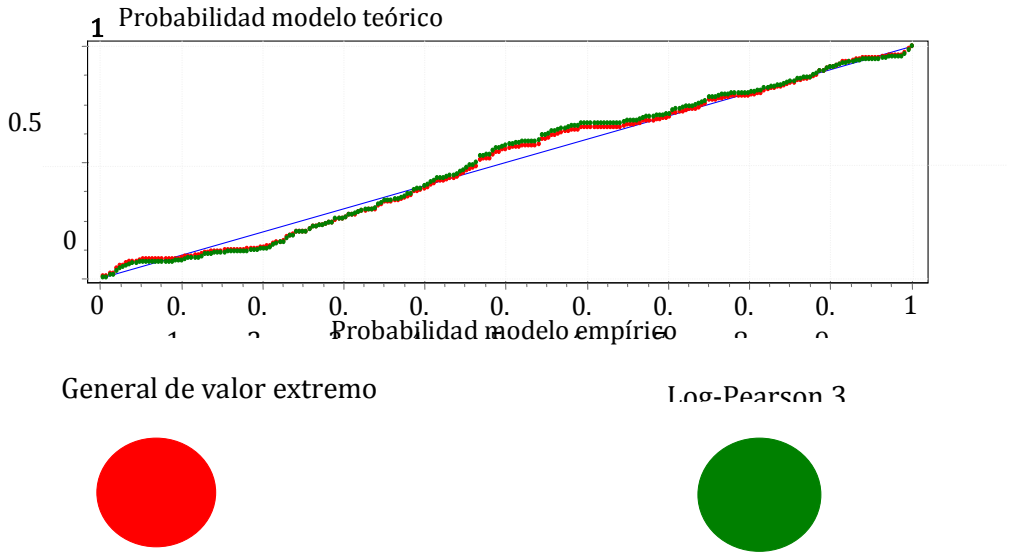


La gráfica del histograma muestra los ajustes que proporcionan los modelos generales de Valor Extremo (DVE) y Log-Pearson



La gráfica de distribución acumulativa muestra que los modelos DVE y Log-Pearson-3 ajustan muy similarmente a los datos empíricos del flujo de pasajeros en X1 durante los días Laborales. Ambos modelos pueden ser representativos del comportamiento de los usuarios.

Gráfico Probabilidad-Probabilidad, X1 Laborales



La gráfica PP de X1 Laborales muestra ajustes muy similares entre los modelos DVE y log-Pearson 3, ambos con valores por debajo (en el extremo izquierdo) y por arriba (en el extremo derecho) a los esperados por la distribución empírica del flujo de pasajeros.

Ajuste X_{10}, Laborales.

Estadísticas X_1	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	248	Mínimo	17, 427
Rango	27209	5%	24, 606
Media	30, 660	10%	27, 071
Varianza	1.1598 E+7	25% (Q1)	2943
Desv. Estándar	3405.6	50% Mediana	31, 057
Coef. de Variación	.11108	75% (Q3)	32, 295
Error estándar	216.26	90%	34, 568
Asimetría	-.32108	90%	36, 255

Con una curtosis de 2.66 y un máximo de 44, 636.0 pasajeros al día. Se observa una distribución ligeramente sesgada a la izquierda con datos que varían poco alrededor del valor central (CV = 11.10%).

Distribuciones propuestas; según el criterio KS X10, Laborales.

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
3	Gamma	0.10973	**1**	4.7753	2	52.345	4
9	Log-logística	0.11141	**2**	5.0294	4	58.254	9

Cuándo uno de los criterios de bondad de ajuste no es suficiente para designar algún modelo de distribución, se toma como criterio de selección seleccionar aquellas pruebas que sí arrojan un modelo, en éste caso, la prueba de bondad de ajuste de Kolmogorov-Smirnov define dos distribuciones que se ajustan a los datos empíricos de la variable aleatoria X_{10} y que tienen un mejor estatus que las proporcionadas por la prueba Anderson-Darling.

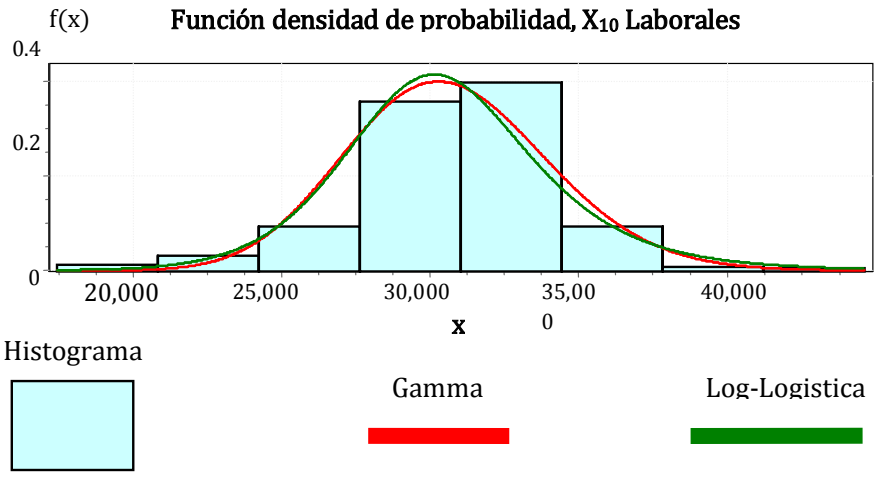
Parámetros de las distribuciones propuestas X10, Laborales.

#	**Distribución**	**Parámetros**
1	Gamma	α=81.05 β=378.28, gama = 0.0
2	Log-Logística	α =14.795 β =30413.0, gama = 0.0

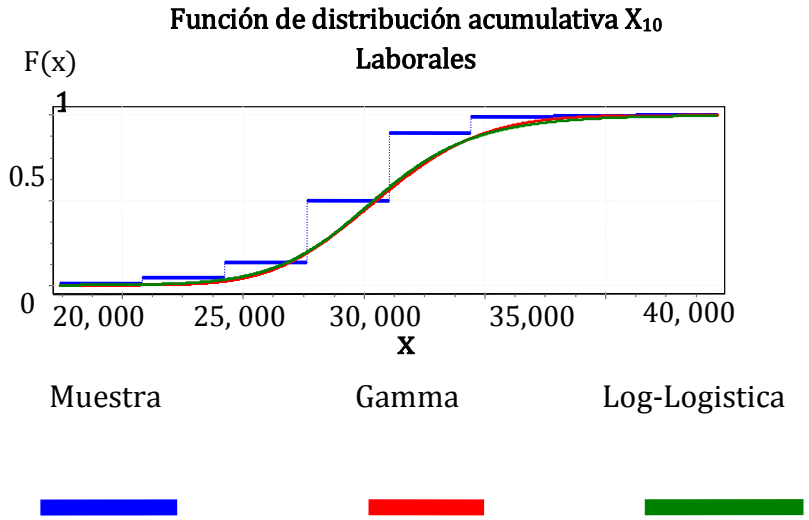
Detalles del ajuste, X_{10} Laborales.					
Gamma [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra	248				
Estadística	0.10973				
Valor P	**0.00468**				
Rango	1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.06814	0.07766	0.08623	0.09639	0.10344
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

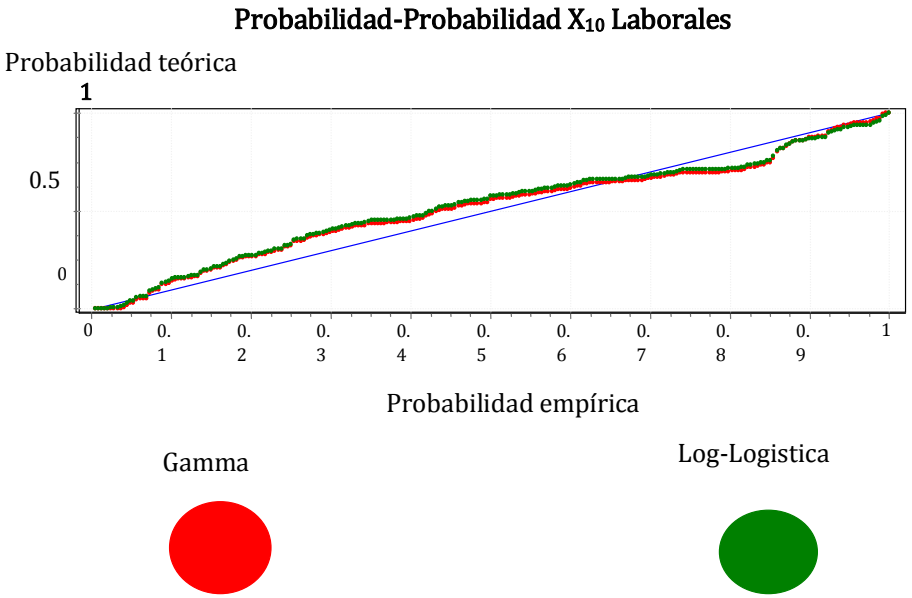
Anderson-Darling [#2]					
Tamaño de la muestra Estadística Rango	248 4.7753 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

A pesar de que se rechaza el modelo Gamma, bajo los distintos valores de alfa, el valor obtenido por el estadístico p-valor permite discriminar entre las dos mejores distribuciones: Gamma y Log-Logística. Aunque ambos modelos son suficientes para describir el flujo en X_{10} los días Laborales, pese a ser rechazados por las pruebas de bondad de ajuste.



La gráfica anterior muestra que los dos modelos propuestos son adecuados; del análisis en la gráfica puede deducirse que el modelo Gamma ajusta ligeramente mejor a los datos empíricos.





La gráfica PP muestra el ajuste de los modelos Gamma y Log-Logístico, la tabla que detalla el ajuste y el ranking de los modelos propuestos hacen ligeramente preferible al modelo Gamma.

Ajuste X_8, Laborales

Estadísticas X8, Laborales	Valor	Percentil	Valor
Tamaño de la muestra	248	Min	12939
Rango	27917	5%	19380.0
Media	26413.0	10%	21660.0
Varianza	1.6820E+7	25% (Q1)	23886.0
Desviación estándar	4101.2	50% (Mediana)	26757.0
Coef. de variación	0.15527	75% (Q3)	28288.0
Error estándar	260.43	90%	30728.0
Asimetría	0.07665	95%	33647.0
Curtosis	1.5611	Max	40856

Los estadísticos obtenidos muestran un ligero sesgo negativo y una dispersión del 15.52 % de los datos alrededor de la media, el análisis posterior deberá indicar cuáles son los mejores modelos para ajustarse al flujo de pasajeros en X_8 los días Laborales y se proponen las sig:

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.06095	1	.98302	1	21.352	3
3	Gamma	.07672	2	2.2315	2	26.804	5

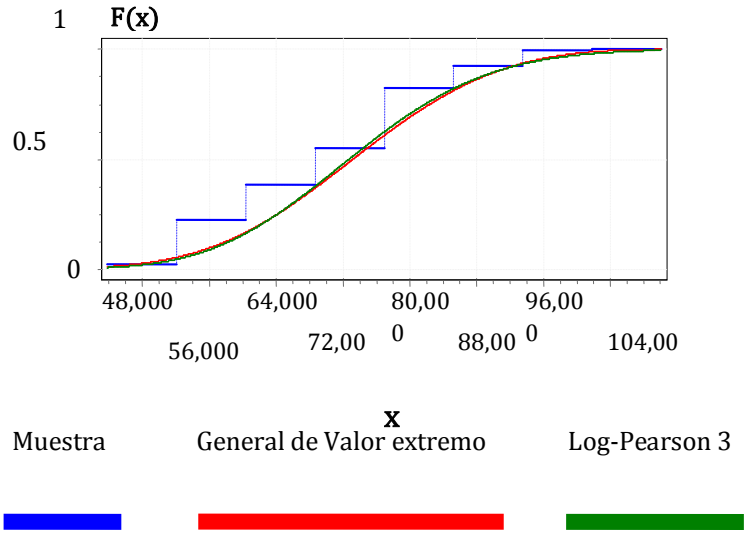
Parámetros de las distribuciones propuestas

#	Distribución	Parámetros
1	Burr	k=1.5782 α=10.358 β=27913.0
2	Gamma	α=41.478 β=636.8

Detalles del ajuste, Burr [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	248 0.06095 0.30315 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.06814	0.07766	0.08623	0.09639	0.10344
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	248 0.98302 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**No**	**No**	**No**	**No**	**No**

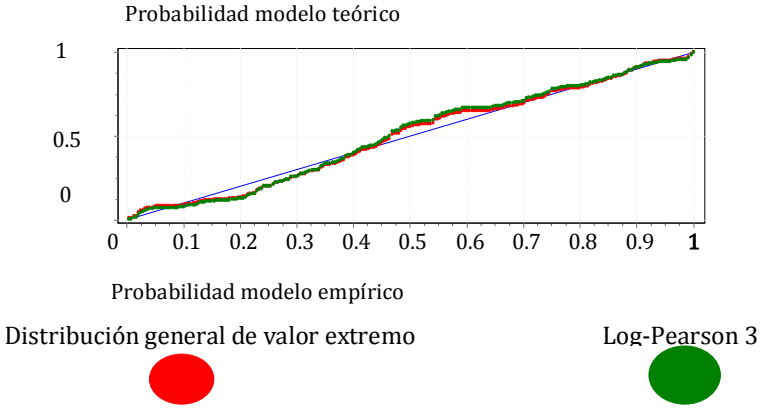
No hay discrepancia entre las diversas pruebas de bondad de ajuste y el modelo Burr es el más adecuado para representar el flujo de pasajeros que circula en días Laborales por X_8.. se prefiere el modelo Burr porque en las tres pruebas de bondad de ajuste

Función de distribución acumulativa, X_8 Laborales



No se rechaza la hipótesis de que los datos provienen de una distribución Burr.

Probabilidad-Probabilidad X_8 Laborales



3.6 Análisis Sábados, Sábados: Y (T)

Estadística Y (T) Sabatino	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	50	Min	1.3121E+5
Rango	1.1456E+5	5%	1.4684E+5
Media	1.8238E+5	10%	1.5896E+5
Varianza	3.5103E+8	25% (Q1)	1.7626E+5
Desviación estándar	18736.0	50% (Mediana)	1.8017E+5
Coef. de variación	0.10273	75% (Q3)	1.9124E+5
Error estándar	2649.6	90%	2.0402E+5
Asimetría	0.22083	95%	2.1307E+5
Curtosis	2.6291	Max	2.4577E+5

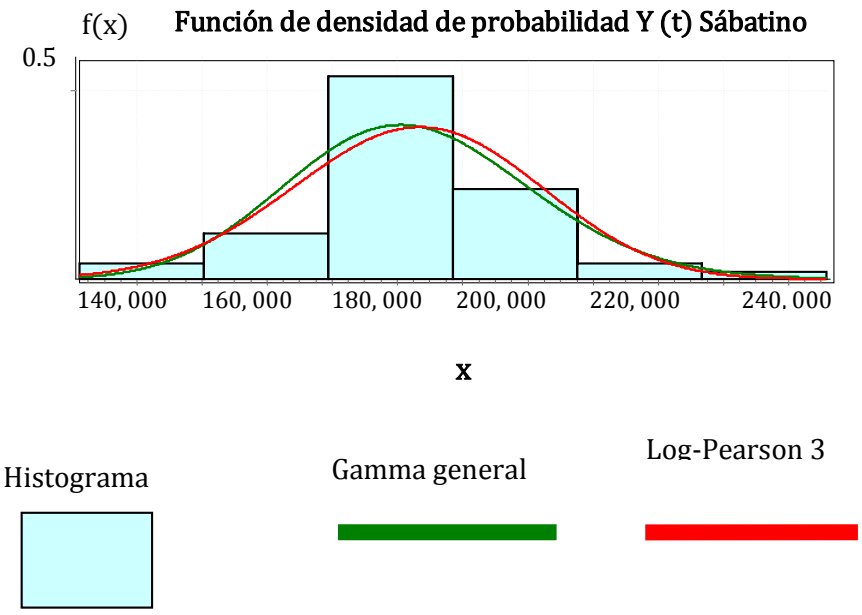
La tabla anterior Y (T) sábados muestra una distribución sesgada positivamente y con una variación del 10.27% lo que implica poca dispersión de los datos a pesar de una muestra aparente pequeña, que se constituye de 50 observaciones.

Distribuciones propuestas y parámetros, sábados: Y (T).

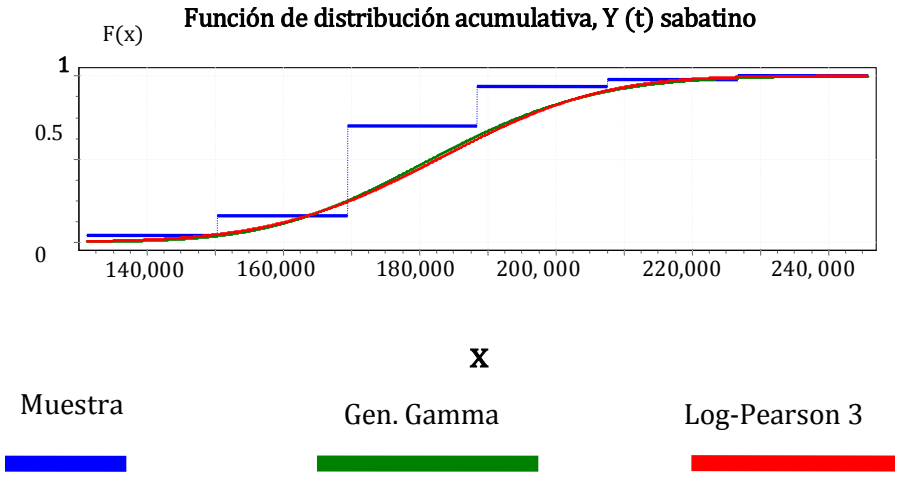
#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
10	Log-Pearson 3	0.15449	2	1.3166	**1**	14.429	9
5	General Gamma	0.17427	5	1.3232	**2**	2.8391	6

#	**Distribución** sábados: Y (T).	**Parámetros**
1	Log-Pearson 3	α=23.22 β=-0.02158 γ=12.61
2	General Gamma	k=1.0017 α=95.495 β=1924.7

Detalles del ajuste.: **Log-Pearson 3** [#10]; sábados: Y (T).					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	50 0.15449 0.16535 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.1484	0.16959	0.18841	0.21068	0.22604
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	50 1.3166 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	5 14.429 0.0131 9				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	7.2893	9.2364	11.07	13.388	15.086
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**No**

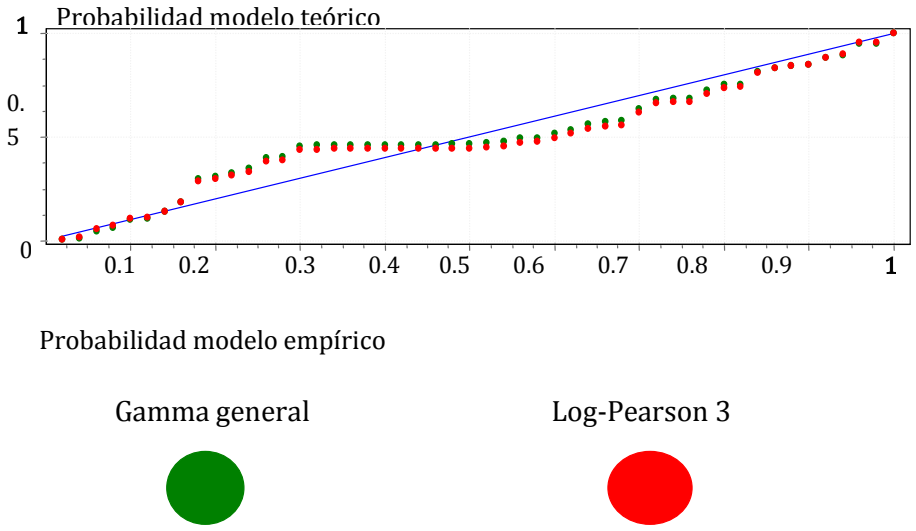


En esta caso puede notarse que ambos modelos propuestos cubren adecuadamente la manera en que se distribuyen las personas en Y (t) sabatino.



Las gráficas distribución acumulativa y probabilidad-probabilidad muestran el ajuste de las distribuciones Gamma general y log-Pearson al flujo empírico de pasajeros, se observa que ambos modelos describen apropiadamente la demanda de pasajeros en Y (T).

Probabilidad-Probabilidad, Y (t) sabatino



Sábados: X_1 Percentil	Valor
Mínimo	36,991
5%	40310
10%	43792
25 %	54,853
50 % Med.	56, 382
Curtosis	3.8273

Percentil	Valor
Min.	36,991
5%	40310
10%	43792
25 %	54,853
50 % Med.	56, 382

Los estadísticos representativos de la VA X1 Sábado muestran que los pasajeros fluyen en una distribución sesgada positivamente, pero con una variación importante de más del 16% que hace medianamente confiables los modelos, tal como se mostrará en los análisis gráficos.

Distribuciones propuestas, AD; sábados: X1

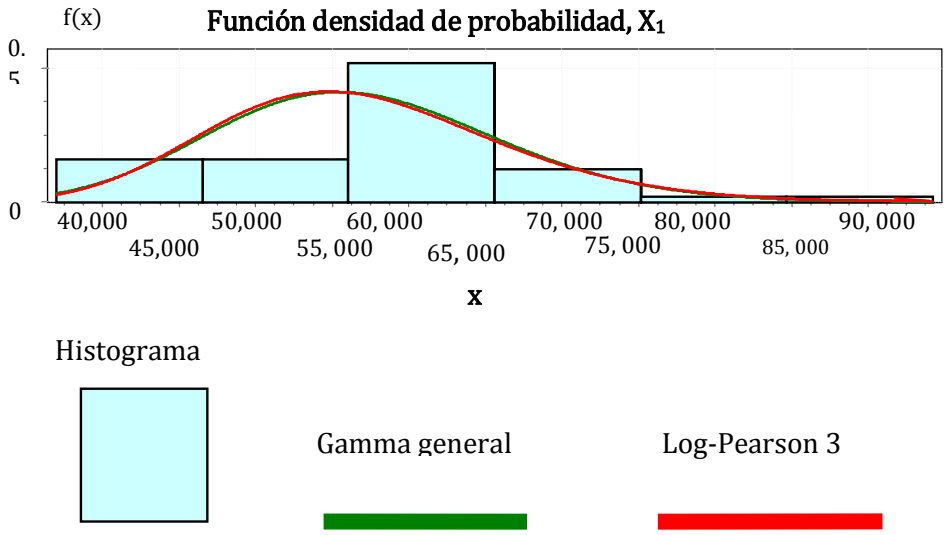
#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
5	General Gamma	0.20906	4	1.6772	**1**	2.9652	2
3	Log-Pearson	0.21857	5	1.7408	**2**	20.749	5

Parámetros de las distribuciones sábados: X1

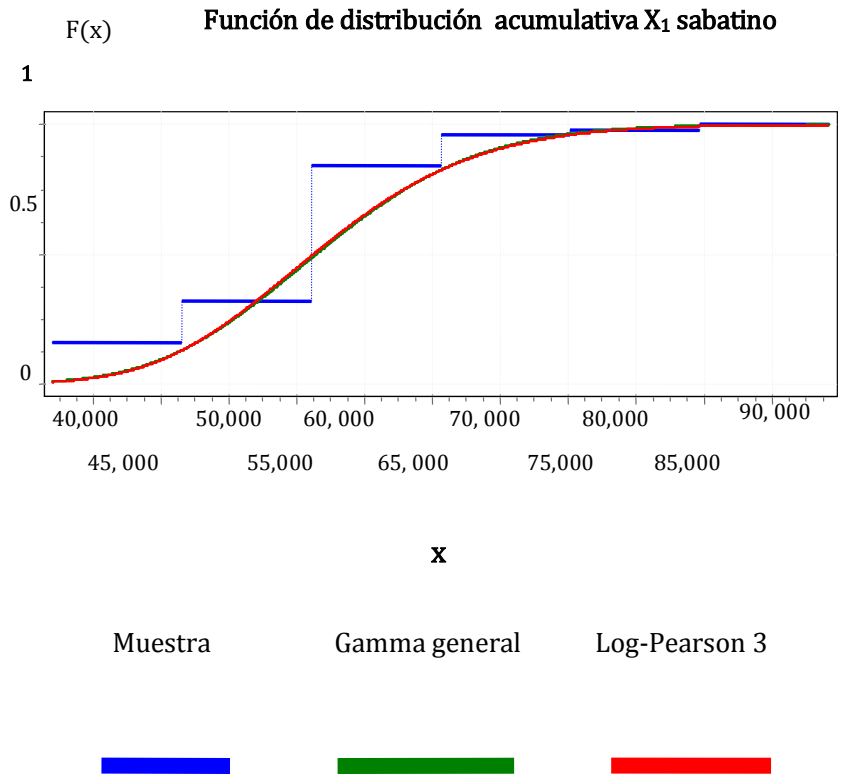
#	**Distribución**	**Parámetros**
1	General. Gamma	k=1.0098 α=36.366 β=1622.6
2	Log-Pearson 3	α=8087.5 β=-0.00184 γ=25.853

Detalles del ajuste, sábados: X1, **Gamma General** [#1], sábados: X1					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	50 0.20906 0.02145 4				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.1484	0.16959	0.18841	0.21068	0.22604
¿Rechazar?	**Sí**	**Sí**	**Sí**	**No**	**No**
Anderson-Darling					
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	3 2.9652 0.39702 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	4.6416	6.2514	7.8147	9.8374	11.345
¿Rechazar?	**No**	**No**	**No**	**No**	**No**

Los modelos propuestos están validados por las pruebas de bondad de ajuste,, el resultado puede confirmarse en la gráfica siguiente.

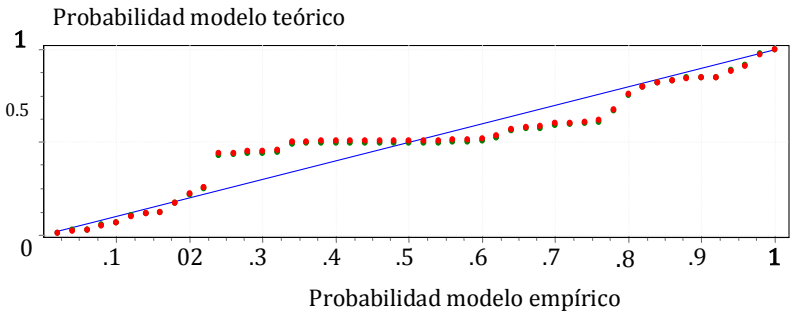


Las gráficas densidad e histograma junto a la distribución acumulativa muestran el ajuste de las distribuciones Gamma general y Log-Pearson 3 al flujo empírico de pasajeros, se observa que ambos modelos ajustan adecuadamente y representan apropiadamente la demanda de pasajeros en X1 sábados. Tal como confirma la gráfica probabilidad-probabilidad, en la página siguiente.



Las gráficas densidad e histograma junto a la distribución acumulativa muestran el ajuste de las distribuciones Gamma general y Log-Pearson 3 al flujo empírico de pasajeros en X_1 sabatino, se observa que ambos modelos ajustan adecuadamente y representan apropiadamente la demanda de pasajeros en X1 sábados. Tal como confirma la gráfica probabilidad-probabilidad, siguiente.

Probabilidad-Probabilidad, X_1 sabatino



Gamma general

Log-Pearson 3

Sábados: X_{10}

Estadística Sábados: X10	Valor	Percentil	Valor
Tamaño de la muestra	50	Min	14381
Rango	14703	5%	16722.0
Media	22655.0	10%	19062.0
Varianza	7.0856E+6	25% (Q1)	20917
Desviación estándar	2661.9	50% (Mediana)	23657.0
Coef. de variación	0.1175	75% (Q3)	24394.0
Error estándar	376.45	90%	24533.0
Asimetría	-0.87287	95%	25828.0
Curtosis	1.446	Max	29084

El análisis estadístico realizado muestra que la distribución del flujo de pasajeros está negativamente sesgada y que tiene una variación de casi el 12% que es considerable dados los valores Máximo y mínimo.. Parámetros de Log-normal: sigma = .12533; mu = 10.021;

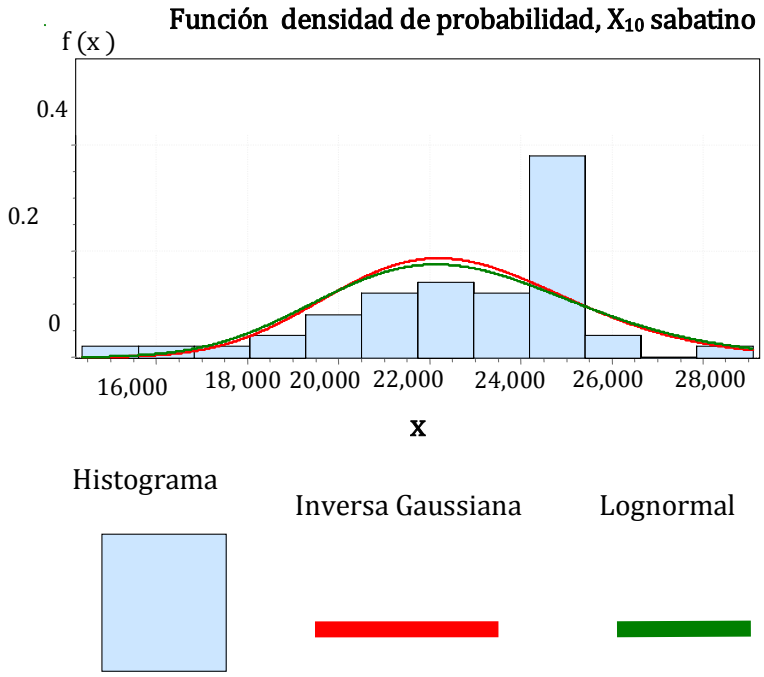
Detalles del ajuste: Log normal [#1], sábados: X10					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	50 0.17239 0.09063 3				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.1484	0.16959	0.18841	0.21068	0.22604
¿Rechazar?	**Sí**	**Sí**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	50 2.3436 5				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
Rechazar?	**Sí**	**Sí**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	3 1.2271 0.74651 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	4.6416	6.2514	7.8147	9.8374	11.345
¿Rechazar?	**No**	**No**	**No**	**No**	**No**

La tabla anterior muestra que el modelo log-normal es el más adecuado ya que se valida mediante dos de las tres pruebas de bondad de ajuste. En segundo lugar estaría la Inversa-Gaussiana, con parámetros: Lambda = 1.6410 Exp. (+6); mu = 22655.0; gamma = cero y a continuación se dan detalles del ajuste.

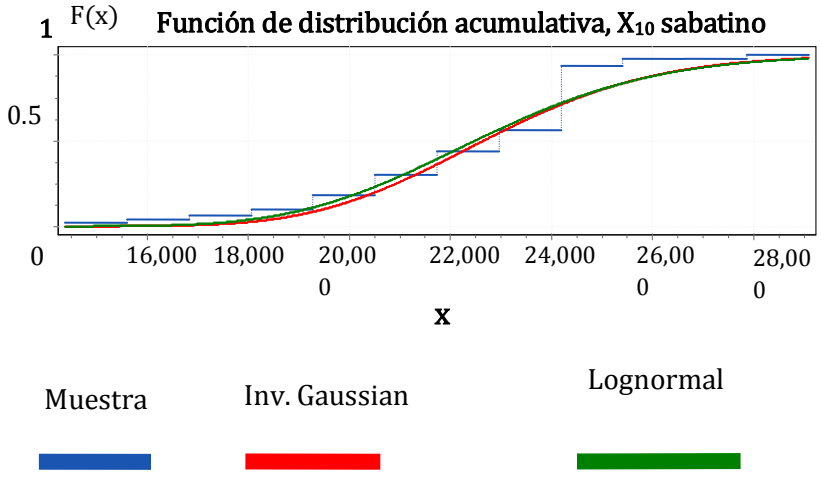
Inversa Gaussiana [#2], **sábados: X10**					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	50 0.16757 0.10725 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.1484	0.16959	0.18841	0.21068	0.22604
¿Rechazar?	**Sí**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	50 2.1498 2				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
Rechazar?	**Sí**	**Sí**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	4 5.4173 0.2471 8				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	5.9886	7.7794	9.4877	11.668	13.277
¿Rechazar?	**No**	**No**	**No**	**No**	**No**

De la tabla anterior, puede verse que los modelos Log-Normal e Inversa Gaussiana son los adecuados a un nivel de confianza del 5% en adelante, es decir hasta 2 % y 1 % lo que probablemente sea una consecuencia del tamaño de la muestra, que en este caso es de únicamente 50 observaciones sabatinas. Las pruebas más confiables en éste caso son las de Anderson-Darling y Kolmogorov-Smirnov. La prueba Chi-cuadrada rechaza ambos modelos porque es una prueba que necesita una mayor cantidad de datos

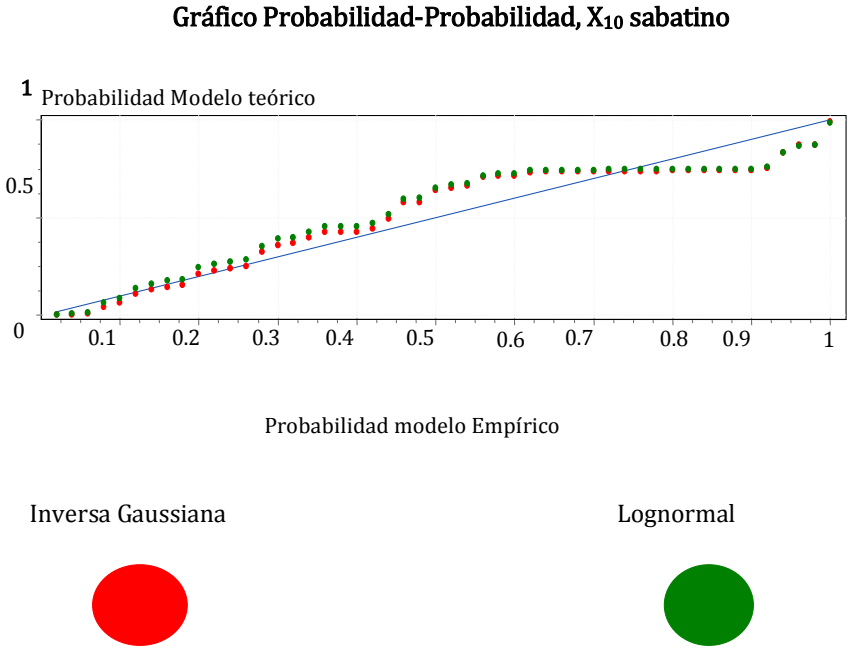
La gráfica del histograma y densidad siguientes muestra que los modelos Log-Normal e Inversa-Gaussiana ajustan adecuadamente a los datos de X_{10} sabatino.



Las gráficas siguientes muestran el ajuste de las distribuciones Inversa-Gaussiana y Log-normal al flujo empírico de pasajeros, ambos modelos ajustan convenientemente y describen la demanda de pasajeros en X_{10} sabatino.



Ambos modelos ajustan y describen la demanda de pasajeros en X_{10} sabatino.



La gráfica anterior del tipo PP confirma el ajuste de las distribuciones Inversa-Gaussiana y Log-normal al movimiento de pasajeros en Sábados X_{10}.

Estadística X_8 sabatino	**Valor**	**Percentil**	**Valor**
Tamaño de la muestra	50	Min	10260
Rango	30230	5%	15424.0
Media	21726.0	10%	17813.0
Varianza	1.6557E+7	25% (Q1)	19326.0
Desviación estándar	4069.1	50% (Mediana)	22370.0
Coef. de variación	0.18729	75% (Q3)	23220.0
Error estándar	575.46	90%	25105.0
Asimetría	1.3809	95%	25756.0
Curtosis	9.2579	Max	40490

Los estadísticos muestran una variación considerable de casi 19% lo cual hace poco confiable el ajuste de los modelos propuestos. Una muestra de mayor tamaño evitaría tal dispersión.

Distribuciones propuestas sábados: X8.

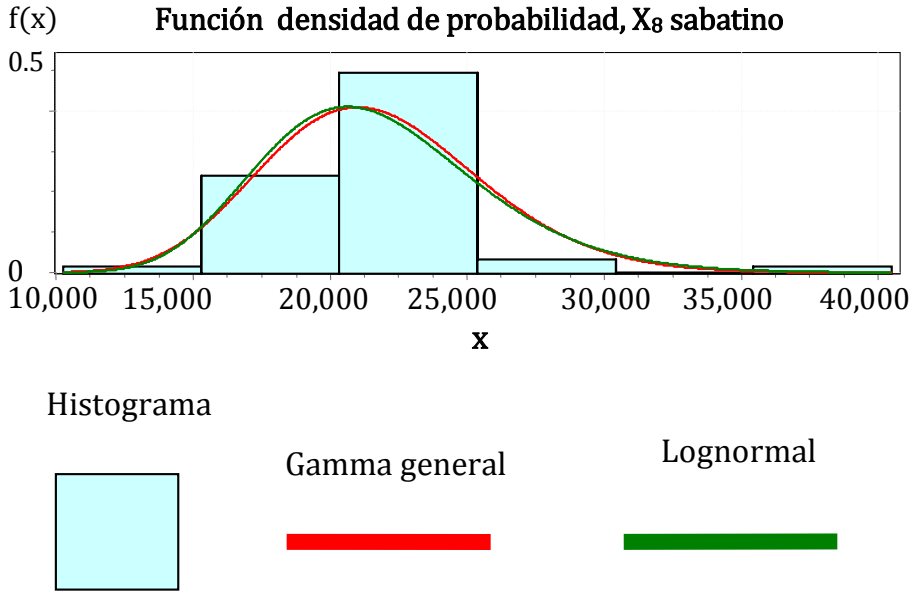
#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Chi-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	General Gamma	0.21241	7	2.3109	**1**	5.9284	6
10	Log-Normal	0.22307	8	2.4552	**2**	4.6088	4

Parámetros de las distribuciones sábados: X_8.

#	**Distribución**	**Parámetros**
1	General Gamma	k=1.0084 α=29.329 β=762.1
2	Log normal	σ=0.18631 μ=9.9694

Detalles del ajuste sábados: X_8. Gamma General [#4].					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	50 0.21241 0.01854 7				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.1484	0.16959	0.18841	0.21068	0.22604
¿Rechazar?	Sí	Sí	Sí	Sí	No
Anderson-Darling					
Tamaño de la muestra Estadística Rango	50 2.3109 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	Sí	Sí	No	No	No
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	3 5.9284 0.11515 6				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	4.6416	6.2514	7.8147	9.8374	11.345
¿Rechazar?	Sí	No	No	No	No

La tabla anterior confirma la sospecha de la poca viabilidad para ajustar un modelo que fuera No rechazado por los tres criterios de bondad de ajuste.



Se muestra el ajuste de las distribuciones Gamma general y Log-normal al flujo empírico de pasajeros en X_8, sábado.

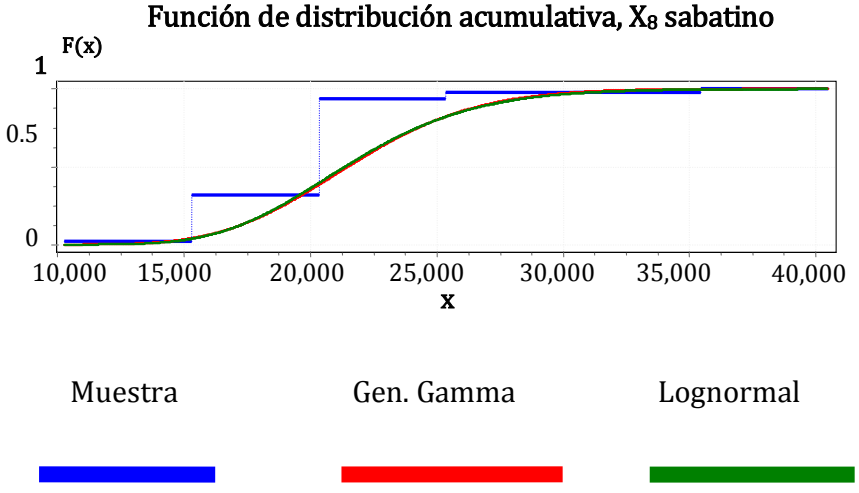
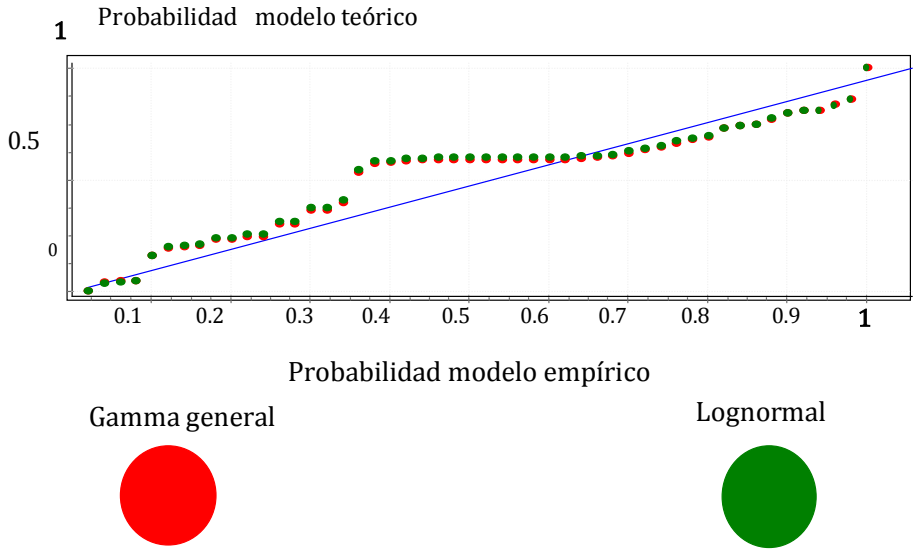


Gráfico de Probabilidad-Probabilidad, X_8 sabatino



Las tres gráficas, densidad e histograma, cuantil-cuantil y probabilidad-probabilidad (QQ y PP) muestran el ajuste de las distribuciones Gamma general y Log-normal al flujo empírico de pasajeros, los dos modelos ajustan adecuadamente y la demanda de pasajeros es descrita apropiadamente en X_8 Sábados.

3.7 Análisis: Domingos

Estadística DOM Y (T)	Valor	Percentil	Valor
Tamaño de la muestra	67	Min	90616
Rango	1.7748E+5	5%	1.0810E+5
Media	1.4660E+5	10%	1.1538E+5
Varianza	1.4003E+9	25% (Q1)	1.2996E+5
Desviación estándar	37421.0	50% (Mediana)	1.3358E+5

Coef. de variación	0**.25526**	75% (Q3)	1.4765E+5
Error estándar	4571.7	90%	2.1214E+5
Asimetría	1.8626	95%	2.4667E+5
Curtosis	3.1667	Max	2.6810E+5

La tabla anterior indica una variación inconveniente (25.52%) que hace poco probable ajustar un modelo adecuado al flujo de pasajeros. Por tal motivo se ensaya ajustar 3 modelos.

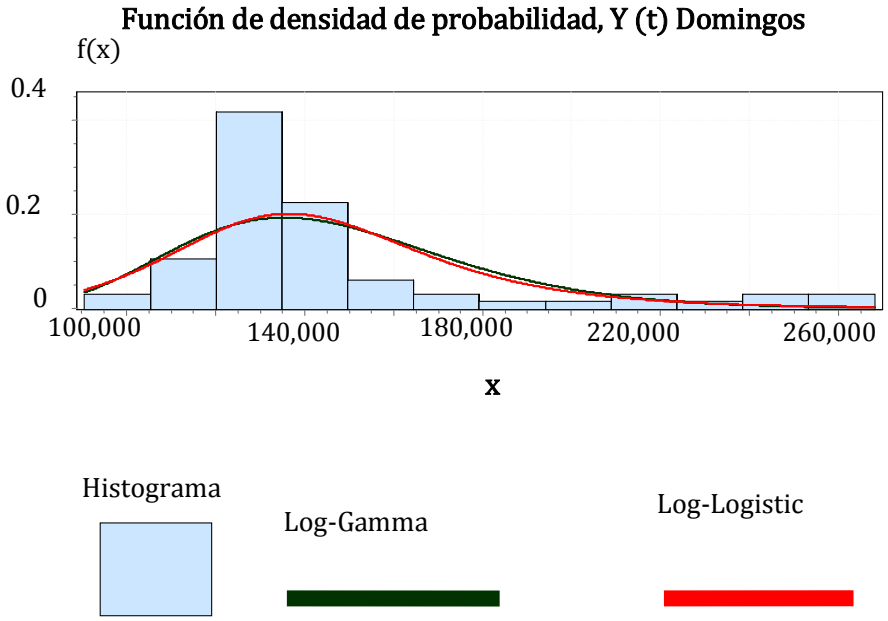
Distribuciones propuestas DOM Y (T).

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Chi-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
2	Log-logística	0.18403	1	3.5573	1	49.512	3
1	Log- Gamma	0.2018	2	4.0421	2	48.087	1
3	Log-normal	0.20361	3	4.1398	3	49.175	2

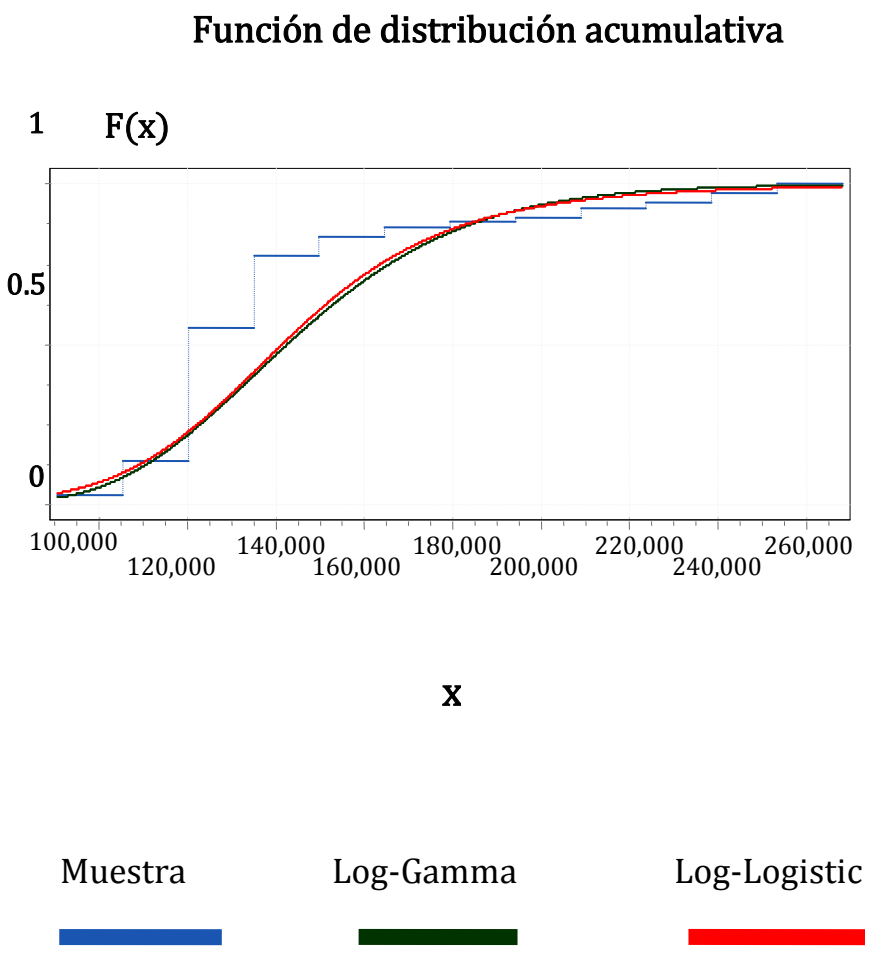
Los parámetros de las distribuciones se presentan a continuación, en orden descendente.

#	**Distribución** DOM Y (T).	**Parámetros**
1	**Log-Logística**	α=7.5605 β=1.4145E+5, gamma = 0.0
2	**Log-Gamma**	α=2903.0 β=0.00409, sin parámetro gamma
3	**Log normal**	σ=0.21864 μ=11.869, gamma = 0.0

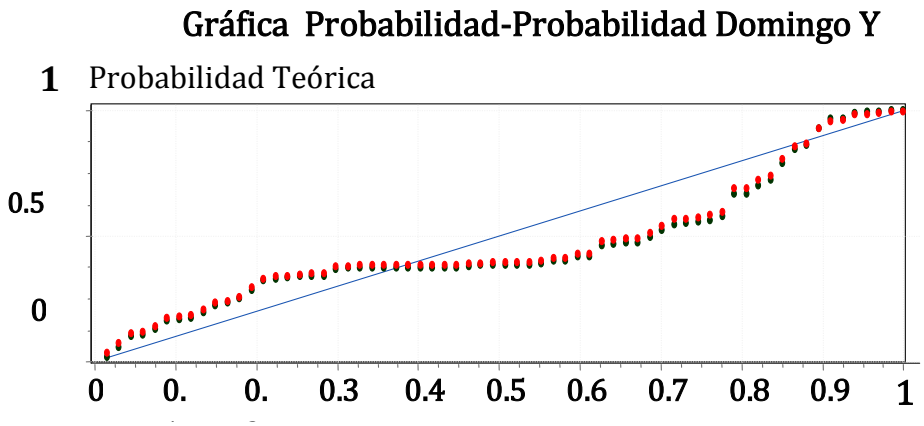
Entre estas distribuciones de ajuste, las que ocupan los tres mejores lugares de acuerdo al criterio ya señalado, no existen diferencias significativas y podría seleccionarse cualquiera de ellas para describir el comportamiento de los datos que corresponden al total de estaciones en la variable Y (T) los días Domingos y festivos del año 2010. En lo siguiente se muestran los diagramas correspondientes al histograma y densidad de la VA Y (T) en los día Domingo.



La gráfica anterior muestra el ajuste de las distribuciones Log-logística y Log-gamma al flujo de pasajeros en Y (T) Domingo. Esto se verá confirmado por las gráficas de distribución acumulativa y probabilidad-probabilidad que ratifican la validez de los modelos propuestos para describir el flujo de personas en esta estación en días Domingo ó festivo.



La gráfica siguiente confirma que la variabilidad de los datos impida obtener un mejor ajuste del modelo teórico al flujo de pasajeros en Y (T) Domingos. Los modelos obtenidos son rechazados por las pruebas de bondad de ajuste, sin embargo son los mejor ubicados en la clasificación que hace el software utilizado



.

En la siguiente página se muestran las estadísticas de la variable aleatoria que corresponde a la estación Pantitlán, (X_1) en días Domingo ó festivos

Estadística DOM X1	Valor
Tamaño de la muestra	67
Rango	77822
Media	47057.0
Varianza	2.6483E+8
Desviación estándar	16274.0
Coef. de variación	0.34583
Error estándar	1988.1
Asimetría	2.1036
Curtosis	4.6141

Percentil	Valor
Min	27855
5%	31616.0
10%	33352.0
25% (Q1)	38267
50% (Mediana)	41455
75% (Q3)	49302
90%	69198.0
95%	90917.0
Max	1.0568E+5

La tabla anterior muestra un coeficiente de variación del 34.58% que posiblemente dificultará ajustar un modelo adecuado. Sin embargo el tamaño de la muestra puede favorecer el ajuste.

Distribuciones propuestas, DOMINGOS: X1.

#	Distribución Dom X_1	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	0.10404	1	0.56674	1	8.363	1
9	Log-logística	0.14746	2	2.0076	2	20.012	3

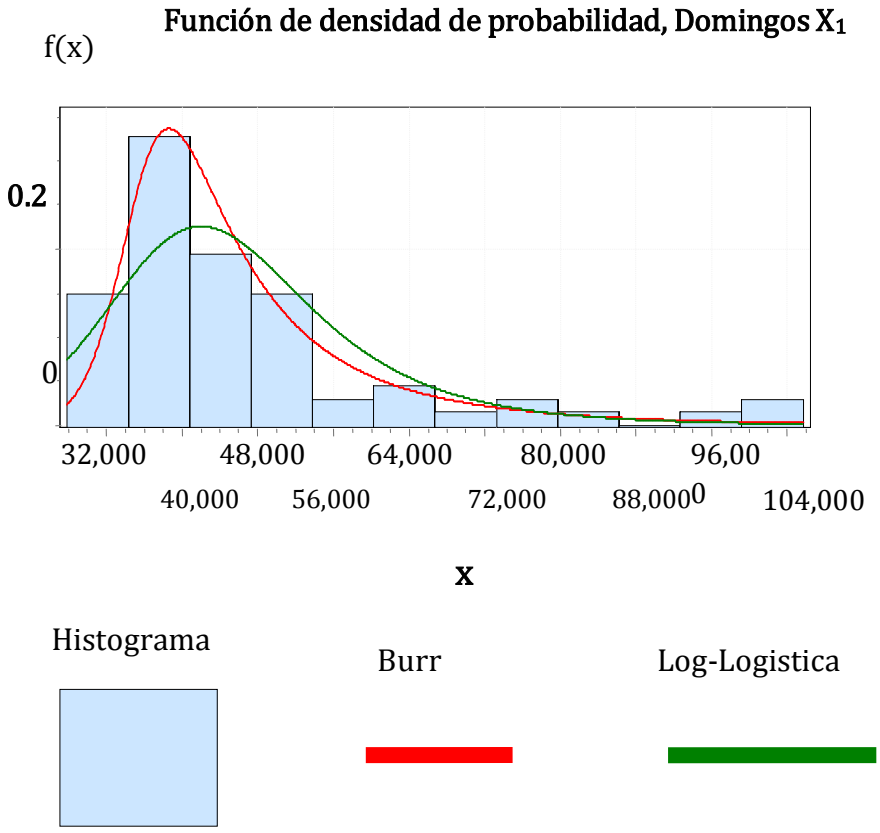
Parámetros de las distribuciones DOMINGOS: X1.

#	Distribución DOM: X1.	Parámetros
1	Burr	k=0.28568 α=14.01 β=36044.0, gamma = 0.0
2	Log-Logística	α=6.0242 β=44413.0

La distribución con mejor ajuste es la Burr con parámetros: k = .28568; alfa = 14.01; beta = 36044.0; gamma = cero.

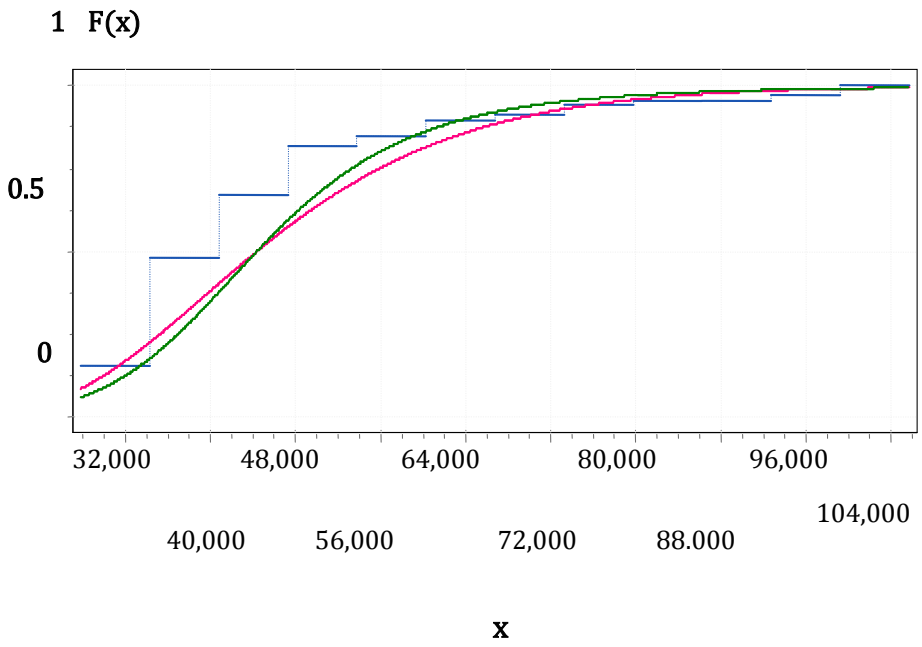
Detalles del ajuste Burr [#1], DOMINGOS: X1.					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	67 0.10404 0.43364 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.12859	0.14693	0.16322	0.18252	0.19584
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Anderson-Darling					
Tamaño de la muestra Estadística Rango	67 0.56674 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
¿Rechazar?	**No**	**No**	**No**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	6 8.363 0.2127 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	8.5581	10.645	12.592	15.033	16.812
¿Rechazar?	**No**	**No**	**No**	**No**	**No**

La tabla muestra que el modelo Burr logra ajustarse al flujo de pasajeros en X1 Domingos; obsérvese que en éste caso el tamaño de la muestra es muy influyente. El segundo modelo sugerido es el Log-logístico con parámetros: alfa = 6.0242; beta = 44413.0; gamma = cero, con un comportamiento de ajuste y rechazo similar.

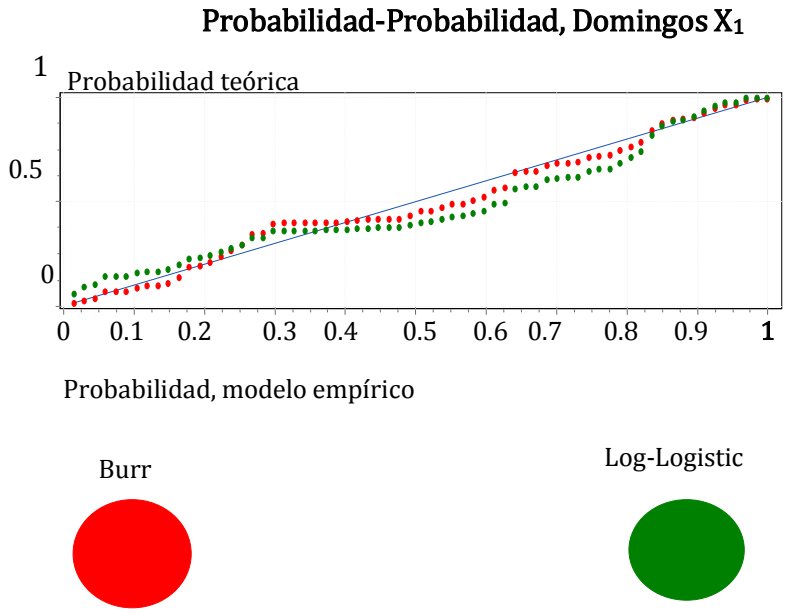


La gráfica anterior del tipo densidad e histograma amén de las gráficas de distribución acumulativa y probabilidad-probabilidad (página siguiente) muestran el ajuste de los modelos Burr y Log-logístico al flujo de los días festivos ó Domingos en X_{10}. (DOM X_1).

Función de distribución acumulativa, Domingos X_1



Se muestran el ajuste de los modelos Burr y Log-logístico en los días festivos ó Domingos en X_{10}. (DOM X_1).



Domingos, X_8.

Estadística DOM X8	Valor	Percentil	Valor
Tamaño de la muestra	67	Min	6809
Rango	27113	5%	11358.0
Media	17630.0	10%	13233.0
Varianza	2.5389E+7	25% (Q1)	15479
Desviación estándar	5038.8	50% (Mediana)	16776
Coef. de variación	0.28581	75% (Q3)	18353
Error estándar	615.58	90%	27127.0
Asimetría	1.5676	95%	30818.0
Curtosis	2.9698	Max	33922

Se muestra un coeficiente de variación del 28.58% en una distribución con sesgo positivo. La VA (estación) es de mediana importancia pero recibe un flujo importante los domingos, debido a su ubicación, cercana a un tianguis lo cual puede explicar su variabilidad.

Distribuciones propuestas, DOM: X8.

#	Distribución Dom X_8	Kolmogorov-Smirnov		Anderson-Darling		Chi-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
3	Log-logística	0.16292	1	2.7562	1	27.07	3
4	Log-normal	0.18024	2	3.1323	2	25.296	2

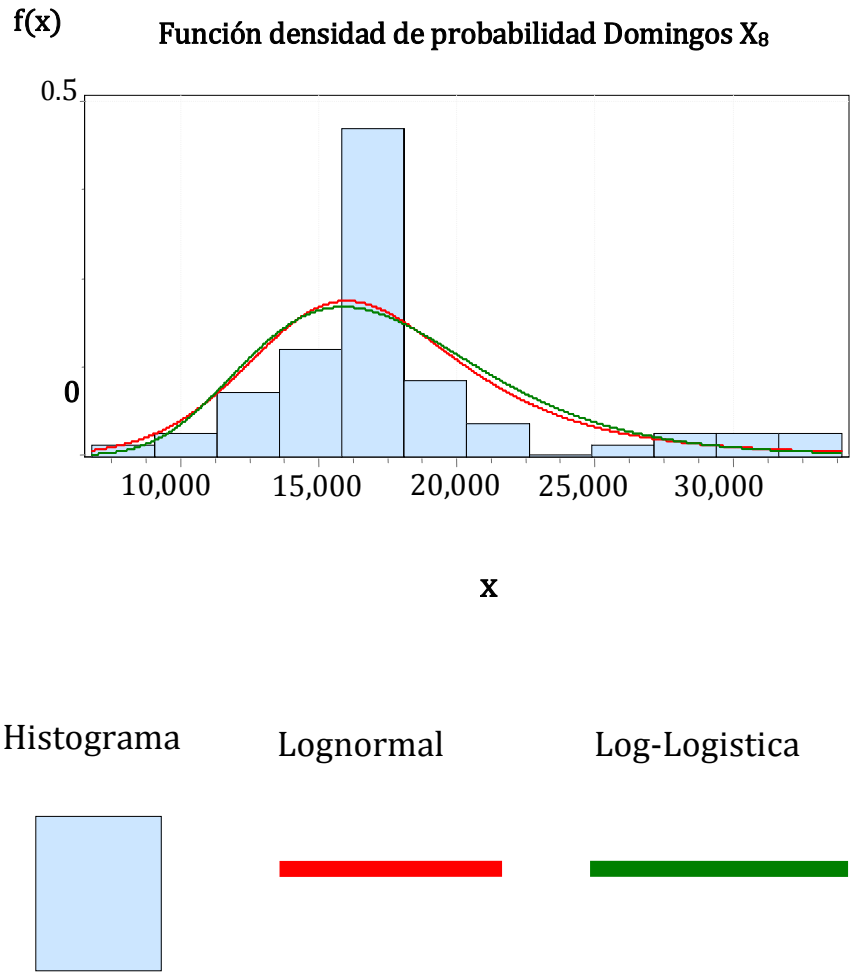
Parámetros de las distribuciones DOM: X8.

#	**Distribución DOM: X8.**	**Parámetros**
1	Log-Logística	α=6.3448 β=16843.0
2	Log normal	σ=0.26109 μ=9.7421

Detalles del ajuste DOM. X_8 distribución Log-Logística [#1]					
Kolmogorov-Smirnov					
Tamaño de la muestra Estadística Valor P Rango	67 0.16292 0.05068 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	0.12859	0.14693	0.16322	0.18252	0.19584
Rechazar?	**Sí**	**Sí**	**No**	**No**	**No**
Anderson-Darling					

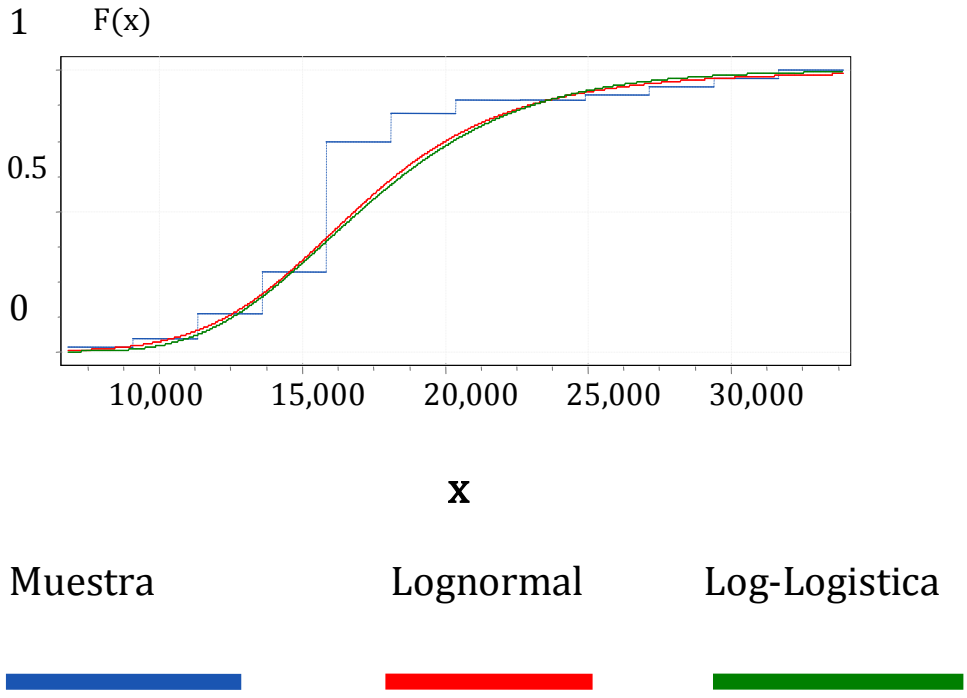
Tamaño de la muestra Estadística Rango	67 2.7562 1				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	1.3749	1.9286	2.5018	3.2892	3.9074
Rechazar?	**Sí**	**Sí**	**Sí**	**No**	**No**
Chi-cuadrado					
Grados de libertad Estadística Valor P Rango	4 27.07 1.9239E-5 3				
α	0.2	0.1	0.05	0.02	0.01
Valor crítico	5.9886	7.7794	9.4877	11.668	13.277
¿Rechazar?	**Sí**	**Sí**	**Sí**	**Sí**	**Sí**

Los modelos propuestos son validados bajo los tres criterios de bondad de ajuste. En la gráfica siguiente se muestra el histograma y las funciones densidad que mejor ajustan a los datos de la estación Santa Martha, denominada X_8 los modelos propuesto son del tipo logarítmico.

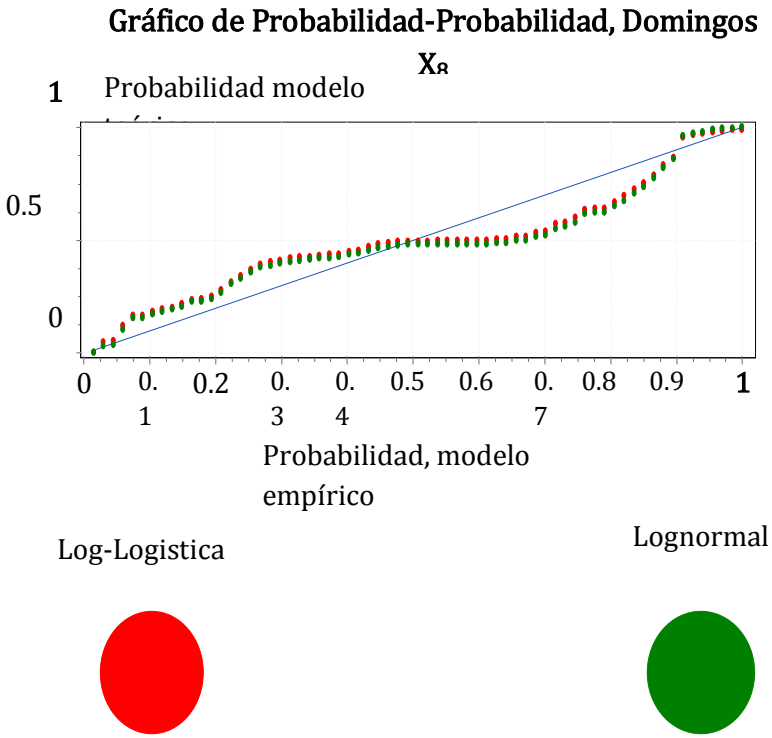


La gráfica muestra al ajuste de los modelos propuestos, que pese a su variabilidad es aceptable bajo los criterios de bondad de ajuste, lo cual confirman las siguientes dos gráficas.

Función de distribución acumulativa X_8 Domingos



Los dos modelos propuestos ajustan casi idénticamente, lo cual los hace intercambiables entre sí.



DOMINGOS X_{10}.

Estadística DOM X10	Valor	Percentil	Valor
Tamaño de la muestra	67	Min	8896
Rango	25477	5%	10568.0
Media	16380.0	10%	11206.0
Varianza	2.9915E+7	25% (Q1)	12974
Desviación estándar	5469.4	50% (Mediana)	15089
Coef. de variación	0.33391	75% (Q3)	16355
Error estándar	668.2	90%	25121.0
Asimetría	1.6575	95%	30500.0
Curtosis	2.4164	Max	34373

Los estadísticos revelan una dispersión que puede dificultar asignar un modelo al flujo de personas que viajan por X_{10} en días festivos, se observa una distribución sesgada positivamente.

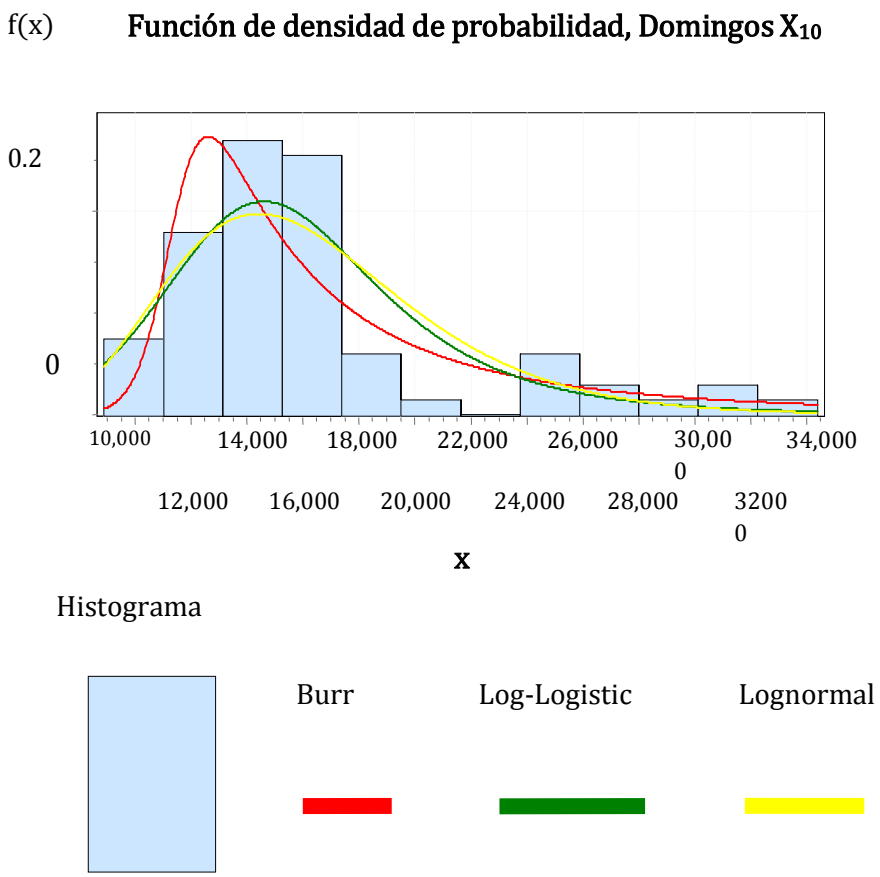
Distribuciones propuestas, no hay discrepancia entre criterios DOM X10.

#	Distribución Dom X_{10}	Kolmogorov-Smirnov		Anderson-Darling		Chi-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	0.17954	1	1.693	1	10.972	1
9	Log-logística	0.18612	2	1.6989	2	20.095	3
4	Log-normal	0.20597	3	2.157	3	18.282	2

Las tres distribuciones propuestas se aceptan bajo los criterios de bondad de ajuste de Anderson-Darling, Kolmogorov-Smirnov y Chi-cuadrado. Los modelos con mejor ajuste se obtienen con las siguientes distribuciones en orden descendente, Burr, Log-logística y Log-Normal.

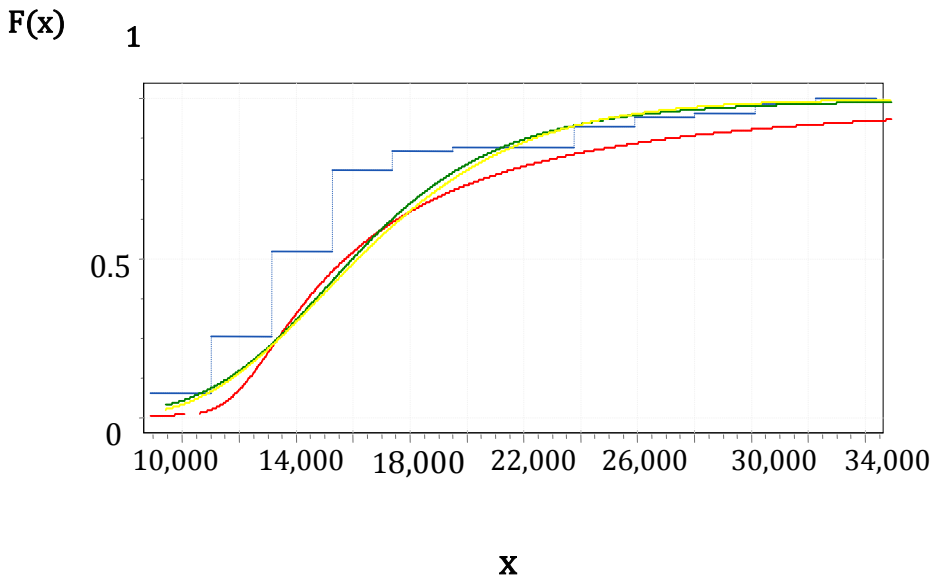
Parámetros de las distribuciones propuestas DOMINGOS X_{10}.

#	Distribución DOM: X_{10}	Parámetros
1	Burr	k=0.14174 α=17.686 β=11551.0, gamma = 0.0
2	Log-logística	α=5.9086 β=15477.0; gamma = 0.0
3	Log-Normal	σ=0.28664 μ=9.659, gamma = 0.0



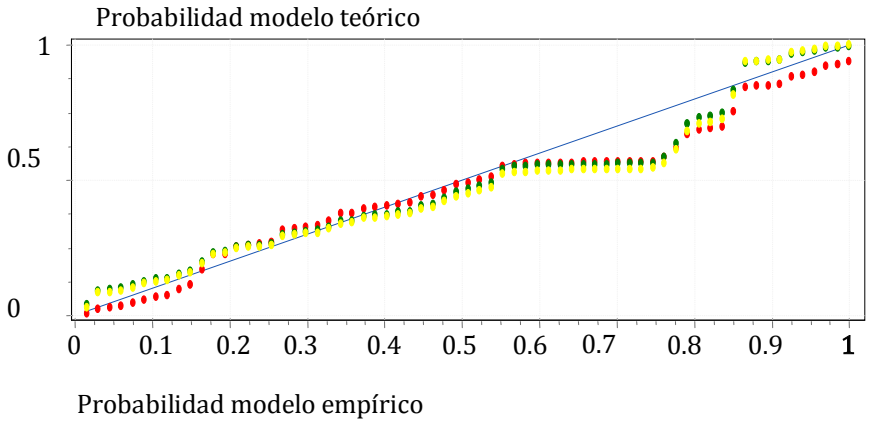
La gráfica de densidad e histograma muestra los tres modelos propuestos donde puede observarse que los modelos Log-logístico y Log-normal ajustan de manera casi idéntica y ambos pueden considerarse como sustitutos del modelo Burr, el cual recoge las variaciones a los extremos de la distribución de pasajeros en tránsito por la VA X_{10}. A continuación se muestran los gráficos y los diagramas probabilidad-probabilidad y cuantil-cuantil para la VA X_{10} (estación La Paz) en días Domingo.

Función de distribución acumulativa, Domingos X_{10}



Una vez más puede constatarse que los modelos propuestos ajustan al modelo empírico a pesar de que existen déciles de la distribución empírica que no son cubiertos por el modelo teórico. Los modelos Log-logístico y Log-normal son bastante similares y pueden ser sustitutos uno del otro y son confiables, en cambio el modelo Burr ajusta por defecto, es decir la probabilidad de que concurran entre 10,000 y 34,000 pasajeros es mucho menor que en los otros dos modelos, con lo cual se tienen argumentos suficientes para desechar el modelo Burr y mantener los dos modelos logísticos.

Gráfica de Probabilidad-Probabilidad Domingos X_{10}



Burr Log-Logistic Lognormal

Las pruebas de bondad de ajuste logran validar los tres modelos propuestos para permitir su posterior utilización con fines de pronóstico. En el siguiente capítulo se muestra la selección de patrones probabilísticos que fueron considerados para utilizarlos como fuente de modelación en cada una de las estaciones de Línea O.

Capítulo 4. Resultados y Conclusión

4.1 Resultado del ajuste de distribuciones

Después de realizar el análisis de la afluencia de pasajeros en las estaciones de Línea O, se obtiene como resultado un conjunto de modelos teóricos de densidad probabilística. Éste tipo de función resultante describe estrechamente el comportamiento del flujo de pasajeros que circula por las estaciones del sistema para trasladarse a sus actividades. En cuyo caso, tal flujo corresponde en aproximación al comportamiento de la variable aleatoria (VA) propuesta (estación de Línea O). El grupo de estaciones se dividió en cuatro grupos y para ello se realizó una partición que divide los datos de afluencia en cuatro clases mutuamente excluyentes: afluencia total, afluencia por días laborales, afluencia en sábados y afluencia en domingos (o días festivos).

En total, se analizaron las once variables aleatorias que corresponden a cada una de las tres categorías de la partición. Debido a que el número total de objetos analizados es muy extenso, en éste reporte se incluye solamente a las cuatro VA's de mayor afluencia y por lo tanto de mayor significancia estadística. Los modelos encontrados corresponden al tipo de patrones sugeridos por el análisis inicial, se trata de modelos continuos cuyo soporte está en los números reales positivos y cuya cota inferior es el cero. Los modelos que mostraron el mejor ajuste a los datos empíricos son aquellos que corresponden a distribuciones de valor extremo general y sus derivados: principalmente la distribución de Weibull con diferentes parámetros. Otros modelos que sugiere el análisis realizado son: Burr, Log-Pearson 3, Gamma general, Log-Gamma y Gamma de dos parámetros, Inversa–Gaussiana, Log-logística y Log-Normal; cuyos resultados se presentan a continuación.

1.- DISTRIBUCIÓN Y (T) total: el modelo que mejor se ajusta a los datos empíricos es el modelo General de valor extremo bajo la prueba de bondad de ajuste de Kolmogorov-Smirnov, el modelo no es recomendable a juicio de las otras dos pruebas.

#	Distribución Y(T)_total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
5	General de Valor extremo	0.07608	1	37.325	11	NA	

Parámetros de la distribución Y (T) total

#	Distribución Y(T)_total	Parámetros
1	General de valor extremo	k=-0.68699 σ=46113.0 μ=2.0294E+5

2.- DISTRIBUCIÓN X_1 total: en éste caso hay concurrencia de criterios entre las pruebas de bondad de ajuste, bajo los criterios de AD y KS; por lo que no se rechaza el modelo General de Valor extremo.

#	Distribución X_1 total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	General Valor Extremo	.05919	1	1.2828	1	52.218	7

Parámetros de la distribución propuesta, X_1 total

#	**Distribución,** X1 total	**Parámetros**
1	General Valor extremo	k=-0.29706 σ=17236.0 μ=60287.0

3.- DISTRIBUCIÓN X_{10} TOTAL

#	Distribución X_1 total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.13343	4	11.194	1	93.868	1

Parámetros de la distribución X_{10} total: el modelo propuesto ajusta bien al flujo de pasajeros en X_{10} bajo el criterio de AD

#	**Distribución, X_{10} total**	**Parámetros**
1	Burr	k=61.609 α=4.9832 β=67333.0

4.- DISTRIBUCIÓN X_8 TOTAL

#	Distribución X_8 total.	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
14	Weibull	.05655	2	1.6922	1	41.108	1

No hay discrepancia notable entre los criterios de la prueba AD y KS, de manera que el ajuste es adecuado. Ambas distribuciones son casi idénticas y ajustan muy similarmente pero se elige la Weibull por tener una mejor clasificación que otras.

Parámetros de la distribución propuesta.

#	**Distribución, X_8 total**	**Parámetros**
1	Weibull	α=4.9905 β=26270.0

5.- DISTRIBUCIÓN LABORALES Y (t): Distribución propuesta según el criterio AD para colas pesadas.

#	Distribución X_8 total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
3	Gamma	.07689	6	1.261	1	10.371	2

Parámetros de la distribución propuesta, Laborales Y (T).

	Distribución, Laborales Y (T)	**Parámetros**
1	Gamma	α=126.46 β=1831.2, gama = 0.0

6.- LABORALES X1: distribución propuesta, según el criterio de la prueba AD

#	Distribución X_1 total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	General Valor Extremo	.07017	2	1.5754	**1**	22.792	2

Parámetros de las distribución propuesta para Laborales X_1

#	**Distribución Laborales X_1**	**Parámetros**
1	General de Valor extremo	k=-0.28318 σ=13072.0 μ=68638.0

7.- Distribución propuesta; según el criterio KS, para X_{10}, Laborales

Cuándo los criterios de bondad de ajuste no concuerdan, para designar algún modelo de distribución se toma como criterio de selección aquella prueba que sí arroja un modelo y una

clasificación. Como en éste caso, donde la prueba de bondad de ajuste de Kolmogorov-Smirnov define una distribución que se ajusta a los datos empíricos de la variable aleatoria X_{10} y que tiene concordancia al criterio Anderson-Darling.

#	Distribución	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrado	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
3	**Gamma**	0.10973	**1**	4.7753	**2**	52.345	4

Parámetros de las distribuciones propuestas, X_{10}, Laborales

#	**Distribución, X_{10} Laborales**	**Parámetros**
1	Gamma	α=81.05 β=378.28, gama = 0.0

8.- Distribución propuesta para la VA X_8 Laborales: En éste caso no hay discrepancia entre dos pruebas de bondad de ajuste y el modelo propuesto no es rechazado, es decir los criterios de bondad de ajuste concuerdan en cuanto al modelo.

#	Distribución X_8 Laborales	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.06095	**1**	.98302	**1**	21.352	3

Parámetros de las distribución propuesta X_8, Laborales

#	**Distribución,** X_8, Laborales	**Parámetros**
1	Burr	k=1.5782 α=10.358 β=27913.0

9.- Distribución propuesta, según AD; para Sábados Y (T)

#	Distribución Y(T) Sábados	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
10	Log-Pearson3	.15449	2	1.3166	**1**	14.429	9

Parámetros de la distribución sábados Y (T)

#	**Distribución Sábados Y (T)**	**Parámetros**
1	Log-Pearson 3	α=23.22 β=-0.02158 γ=12.61

10.- Distribución propuesta, criterio AD para sábados X_1

#	Distribución X_1 Sábados	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
5	Gamma-general	.20906	4	1.6772	1	2.9652	2

Parámetros de la distribución, Sábados X_1

#	**Distribución** Sábados X_1	**Parámetros**
1	General. Gamma	k=1.0098 α=36.366 β=1622.6

11.- Distribución propuesta, Sábados X_8.

#	Distribución Sábados X_8	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
4	Gamma-general	.221241	7	2.3109	**1**	5.9284	6

La prueba Anderson-Darling coloca en primer lugar al modelo Gamma-general para describir los datos de la VA Sábados X_8, las pruebas KS y Ji-cuadrado colocan muy por debajo de la clasificación el modelo propuesto. Se acepta el modelo Gamma porque la prueba AD es más potente que las otras y está definid para distribuciones sesgadas (colas pesadas)

Parámetros de las distribución propuesta, Sábados X_8

#	**Distribución Sábados X_8**	**Parámetros**
1	General Gamma	k=1.0084 α=29.329 β=762.1

12.- Distribución propuesta, criterio Anderson-Darling, Sábados X_{10}

#	Distribución	Parámetros
1	Log-Normal	sigma = .12533; mu = 10.021; gamma = cero

13.- Distribución propuesta Domingos Y (T).

#	Distribución X_1 total	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
2	Log-logística	.18403	1	3.5573	1	45.512	3

Parámetros de la distribución propuesta.

#	**Distribución** DOM Y (T)	**Parámetros**
1	**Log-Logística**	α=7.5605 β=1.4145E+5, gamma = 0.0

14.- Distribución propuesta según el criterio AD, para Domingos X_1

#	Distribución DOM X_1	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.10404	1	.56674	1	8.363	1

Parámetros de las distribuciones Domingos X_1

#	Distribución DOM X1	Parámetros
1	Burr	k=0.28568 α=14.01 β=36044.0, gamma = 0.0

La mejor ajustada es la Burr con parámetros: K = .28568; alfa = 14.01; beta = 36044.0 y gamma = cero.

15.- Distribución propuesta, criterio AD para Domingos X_8. Las pruebas de bondad de ajuste convergen al mismo resultado en las pruebas KS y AD, de manera que el modelo propuesto es adecuado a los datos del flujo de pasajeros en X_8 Domingos.

#	Distribución DOM X_8	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
3	**Log-logística**	**.16292**	**1**	**2.7562**	**1**	**27.07**	**3**

Parámetros de las distribución Domingos X_8

#	**Distribución Domingos X_8**	**Parámetros**
1	**Log-Logística**	**α=6.3448 β=16843.0**

16.- Distribución propuestas para la VA Dom. X_{10}

#	**Distribución DOM X_{10}**	Kolmogorov-Smirnov		Anderson-Darling		Ji-cuadrada	
		Estadística	Rango	Estadística	Rango	Estadística	Rango
1	Burr	.17954	1	1.693	1	10.972	2

Los mejores ajustes se consiguen con las siguientes distribuciones en orden descendente, Burr, Log-logística y Log-Normal que pueden utilizarse de manera indiscriminada para modelar el flujo de pasajeros en la estación X_{10}.

Parámetros de la distribución propuesta: no hay discrepancia entre criterios de bondad de ajuste, el modelo Burr será seleccionado como el más adecuado.

#	Distribución DOM X_{10}	Parámetros
1	Burr	k=0.14174 α=17.686 β=11551.0, gamma = 0.0

Resumen de los parámetros de las distribuciones obtenidas.

Distribución Burr:

"k" parámetro de forma (continuo) k > 0

Alfa, parámetro de forma continuo alfa >0

Beta, parámetro de escala continuo (beta > 0)

Gama, parámetro de localización continuo, gama > 0

Distribución Gamma, parámetros:

α- Alfa, parámetro de forma continuo ($\alpha > 0$)
β- Beta, parámetro de escala continuo ($\beta > 0$)
γ- Gama, parámetro de localización continuo. Si ($\gamma \equiv 0$) entonces se tiene la distribución Gamma de dos parámetros. Dominio: $\gamma \leq x < +\infty$

Distribución General de Valor extremo, parámetros:

k- parámetro de forma continuo
σ- Sigma, parámetro de escala continuo ($\sigma > 0$)
μ- Mu, parámetro de localización continuo

Dominio: $1 + k\frac{(x-\mu)}{\sigma} > 0 \quad \text{for} \quad k \neq 0$; $-\infty < x < +\infty \quad \text{for} \quad k = 0$

Distribución Log-Logística, parámetros

α- Alfa, parámetro de forma continuo ($\alpha > 0$)
β- Beta, parámetro de escala continuo ($\beta > 0$)
γ- .Gama, parámetro de localización continuo, si $\gamma \equiv 0$ se obtiene la distribución Log-logística de dos parámetros. Dominio $\gamma \leq x < +\infty$

Distribución Log-normal, parámetros:

σ-Sigma, parámetro continuo ($\sigma > 0$); μ- Mu, parámetro continuo
γ- Gama, parámetro de localización continuo, si ($\gamma \equiv 0$ se obtiene la distribución Log-normal de dos parámetros. Dominio: $\gamma < x < +\infty$

Distribución de Weibull, parámetros:

α- Alfa, parámetro de forma continuo ($\alpha > 0$)
β- Beta, parámetro de escala continuo ($\beta > 0$)
γ- Gama, parámetro de localización continuo ($\gamma \equiv 0$ si gama es cero se obtiene la distribución Weibull de dos parámetros. Dominio $\gamma \leq x < +\infty$

4.2 Conclusión

Los modelos obtenidos para cada una de las variables aleatorias en cuestión (estaciones) son presentados en la siguiente tabla.

Tabla Final

Variable Aleatoria	Función de Distribución	Parámetros
Y(T) total	General valor extremo	k=-0.68699 σ=46113.0 μ=2.0294E+5
X_1 total	General Valor extremo	k=-0.29706 σ=17236.0 μ=60287.0
X_{10} total	Burr	k=61.609 α=4.9832 β=67333.0
X_8 total	Weibull	α=4.9905 β=26270.0
Laborales Y(T)	Gamma	α=126.46 β=1831.2, gama = 0.0
Laborales X_1	General Valor extremo	k=-0.28318 σ=13072.0 μ=68638.0
Laborales X_{10}	Gamma	α=81.05 β=378.28, gama = 0.0
Laborales X_8	Burr	k=1.5782 α=10.358 β=27913.0
Sábados Y(T)	Log-Pearson 3	α=23.22 β=-0.02158 γ=12.61
Sábados X_1	General. Gamma	k=1.0098 α=36.366 β=1622.6
Sábados X_8	General Gamma	k=1.0084 α=29.329 β=762.1
Sábados X_{10}	Log-Normal	sigma = .12533; mu = 10.021; gamma = cero
DOM Y (T)	Log-Logística	α=7.5605 β=1.4145E+5, gamma = 0.0
DOM X1	Burr	k=0.28568 α=14.01 β=36044.0, gamma = 0.0
DOM X_8	Log-Logística	α=6.3448 β=16843.0
DOM X_{10}	Burr	k=0.14174 α=17.686 β=11551.0, gamma = 0.0

Tras asignar una distribución de probabilidad ajustada a la demanda de usuarios, esta distribución deberá permitir modelar adecuadamente el comportamiento correspondiente a una Línea de transporte y a una estación; en cuyo caso, deberán tomarse las medidas adecuadas para mejorar la asignación de los siguientes recursos como son:

- Número de trenes adecuado al flujo de pasajeros por cada estación (flujo en la variable aleatoria: estación) y conveniente para satisfacer tal demanda.
- Por consiguiente, el flujo de pasajeros implica realizar ajustes en las planillas de conducción de acuerdo a dicha instancia.
- Realizar la asignación de tiempos adecuados de mantenimiento a trenes e instalaciones fijas a fin de obtener resultados favorables en la planeación, abasto y destino de refacciones e insumos que permitan mantener la mayor cantidad de operaciones de servicio al usuario, conforme a la demanda de pasajeros.

De esta manera el beneficiario de obtener éste conocimiento será el usuario habitual de Línea O, quién mejoraría en un "épsilon" su nivel de vida simplemente desplazándose eficientemente por la ciudad a realizar sus actividades.

Conociendo el flujo de la demanda de usuarios pueden obtenerse cantidades óptimas de recursos que permitan al Sistema brindar un mejor servicio a los ciudadanos del Valle de México y sus municipios conurbados. Cabe mencionar que al optimizar sus recursos, el Sistema puede avocarse a mejorar aquellos aspectos de operación que hubiesen quedado detenidos por falta de planeación, financiamiento ó alguna otra causa atribuible a la falta de información.

El ajuste que se obtiene de los modelos o densidades de probabilidad está respaldado por las pruebas de bondad de Kolmogorov-Smirnov y Anderson-Darling principalmente, aunque también se utilizó la distribución Ji-cuadrada en ocasiones; validando hasta con tres criterios el ajuste de la distribución a los datos empíricos. De ésta manera el flujo de pasajeros ha podido ser modelado mediante modelos probabilísticos teóricos.

Se observó que en algunos casos las diferentes pruebas de bondad de ajuste pueden llegar a rechazar la hipótesis de que los datos empíricos provienen de algún modelo particular; sin embargo, debido a la fortaleza del algoritmo de ajuste (máxima verosimilitud) la distribución propuesta puede considerarse un modelo adecuado para los datos empíricos, ya que a pesar

del rechazo que sugiera alguna de las pruebas de bondad, las distribuciones propuestas se consideran el mejor modelo que es posible obtener y son sujetas de tomarse en cuenta.

Mediante éste trabajo ha sido posible obtener una metodología para llevar a cabo el ajuste de modelos de densidad a datos empíricos. Queda aún la posibilidad de obtener más datos, llevar a cabo los procedimientos anteriores y ejecutar las técnicas de modelado para obtener los resultados que mejor describan el flujo de pasajeros. Estas pruebas son perfectibles.

Se confirma que tiene validez dividir el flujo de pasajeros en diferentes clases mutuamente excluyentes para realizar el análisis de la demanda de pasajeros, la veracidad de tal afirmación se sostiene al confirmar que los mejores modelos obtenidos corresponden a las clases: Laborales, Domingos y Sábados respectivamente. El llamado flujo Total que reúne los datos de la Línea O sin discriminar entre diferentes tipos de día resulta con ajustes pobres y que generalmente caen en la región de rechazo de las pruebas de bondad de ajuste, en cuanto a los días "Sábado y Domingos" Totales, un mejor ajuste puede obtenerse aumentando el tamaño de la muestra en diferentes años, la muestra actual puede considerarse insuficiente.

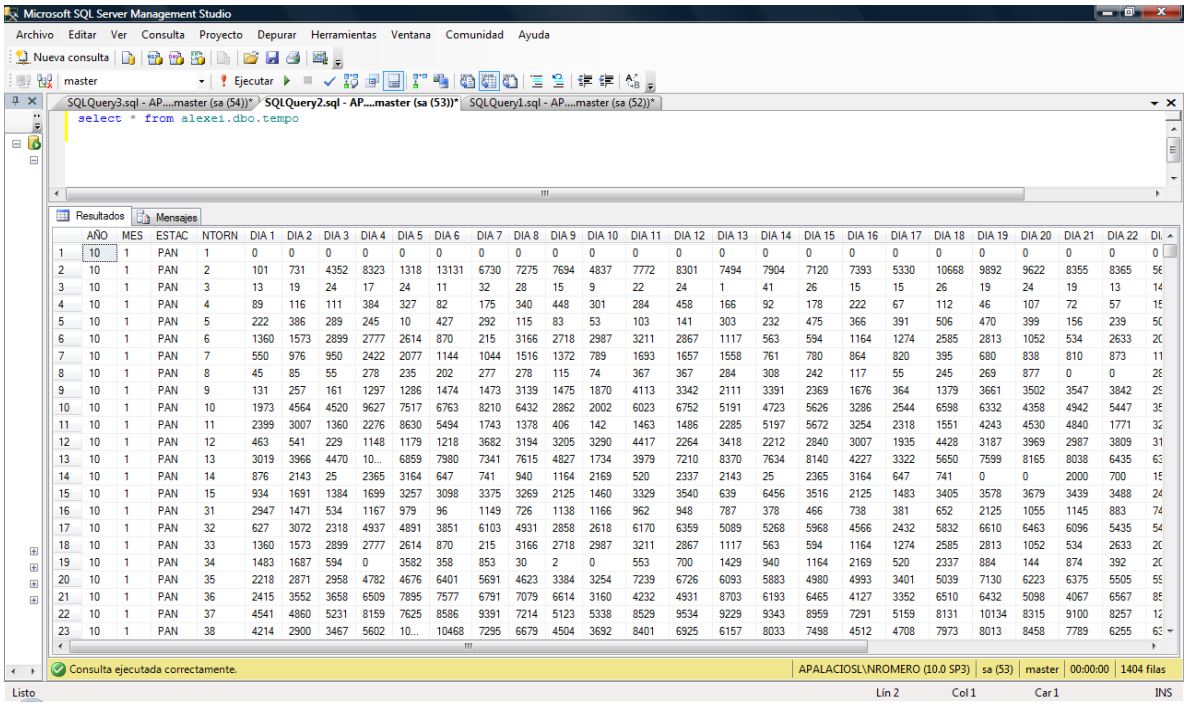
De cualquier forma los modelos aquí obtenidos siempre se verán sujetos a validaciones, a condición de que sean provistos de nuevas "lecturas" que en años posteriores permitan su utilización. Es el factor tiempo la principal razón para terminar el estudio hasta la etapa actual puesto que el presente trabajo está sujeto a un plazo de entrega que ya no puede diferirse.

Los resultados alcanzados confirman la hipótesis de que es posible obtener modelos probabilísticos a partir de los datos empíricos que proporcionan las "lecturas".

El trabajo por llevar a cabo consistirá en utilizar los modelos obtenidos para realizar la simulación del comportamiento del flujo de pasajeros y verificar su utilidad en la predicción de conductas posteriores, esto queda condicionado a diversos factores como son el tiempo y los recursos necesarios para obtener los datos y la licencia del software especializado. Las herramientas de cómputo utilizadas se manejaron en forma estandarizada, es decir con los recursos mínimos ofrecidos por él fabricante; existen versiones que proveen con un mayor número de herramientas y poder de análisis. Por tal motivo se sugiere al Sistema continuar el análisis con datos anteriores y realizar parte ó toda la metodología para obtener los modelos de densidad de probabilidad. Estos estudios posteriores le permitirán mejorar el servicio de transporte que brinda no solo a los usuarios de Línea O sino tal vez a los de otras líneas de trenes.

Anexos

En el Capítulo 3, la base de datos **db_llena.dbf** se importa a SQL Server en la tabla *"alexei.dbo.tempo"*, y se ejecuta la siguiente búsqueda (query) con las siguientes líneas de código:



*select **

into alexei.dbo.usuarios

from (

select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-01', 120) as fecha, sum([dia 1]) as usuarios

from alexei.dbo.tempo

group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-01', 120), estac

union

select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-02', 120) as fecha, sum([dia 2]) as usuarios

from alexei.dbo.tempo

group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-02', 120), estac

union

select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-03', 120) as fecha, sum([dia 3]) as usuarios

from alexei.dbo.tempo

group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-03', 120), estac

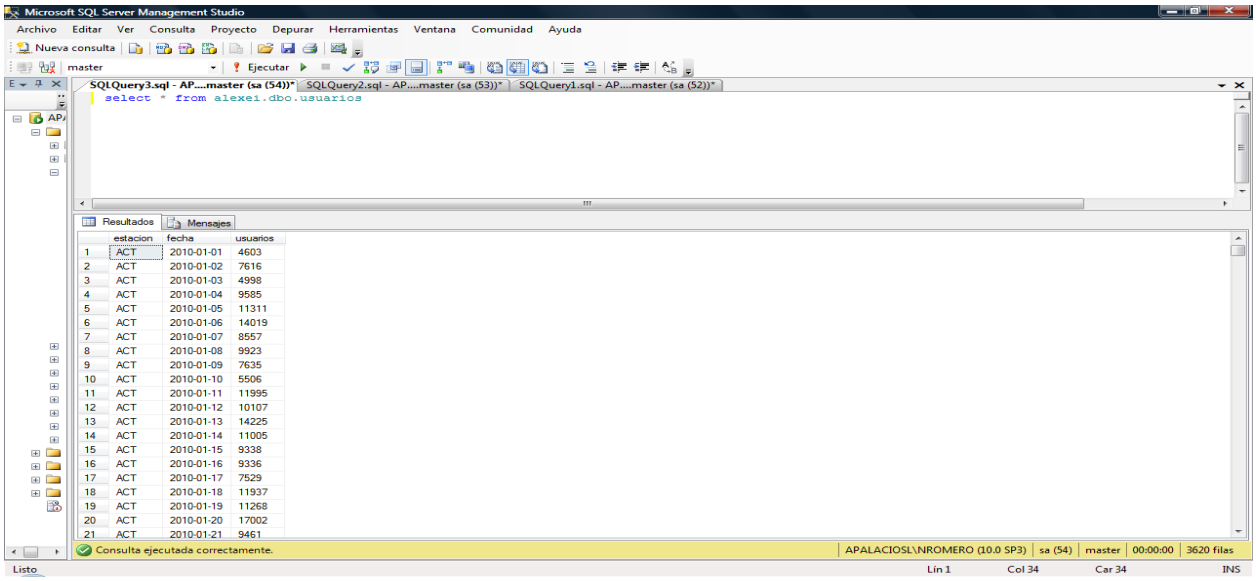
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-04', 120) as fecha, sum([dia 4]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-04', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-05', 120) as fecha, sum([dia 5]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-05', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-06', 120) as fecha, sum([dia 6]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-06', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-07', 120) as fecha, sum([dia 7]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-07', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-08', 120) as fecha, sum([dia 8]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-08', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-09', 120) as fecha, sum([dia 9]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-09', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-10', 120) as fecha, sum([dia 10]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-10', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-11', 120) as fecha, sum([dia 11]) as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-11', 120), estac
union

```
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-12', 120) as fecha, sum([dia 12])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-12', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-13', 120) as fecha, sum([dia 13])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-13', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-14', 120) as fecha, sum([dia 14])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-14', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-15', 120) as fecha, sum([dia 15])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-15', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-16', 120) as fecha, sum([dia 16])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-16', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-17', 120) as fecha, sum([dia 17])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-17', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-18', 120) as fecha, sum([dia 18])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-18', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-19', 120) as fecha, sum([dia 19])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-19', 120), estac
union
```

```
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-20', 120) as fecha, sum([dia 20])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-20', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-21', 120) as fecha, sum([dia 21])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-21', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-22', 120) as fecha, sum([dia 22])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-22', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-23', 120) as fecha, sum([dia 23])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-23', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-24', 120) as fecha, sum([dia 24])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-24', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-25', 120) as fecha, sum([dia 25])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-25', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-26', 120) as fecha, sum([dia 26])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-26', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-27', 120) as fecha, sum([dia 27])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-27', 120), estac
union
```

select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-28', 120) as fecha, sum([dia 28])
as usuarios
from alexei.dbo.tempo
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-28', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-29', 120) as fecha, sum([dia 29])
as usuarios
from alexei.dbo.tempo
where MES not in (2)
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-29', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-30', 120) as fecha, sum([dia 30])
as usuarios
from alexei.dbo.tempo
where MES not in (2)
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-30', 120), estac
union
select ESTAC as estacion, CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-31', 120) as fecha, sum([dia 31])
as usuarios
from alexei.dbo.tempo
where MES not in (2, 4, 6, 9, 11)
group by CONVERT(date, '2010-' + convert(CHAR(2), MES) + '-31', 120), estac
) a
order by estacion, fecha

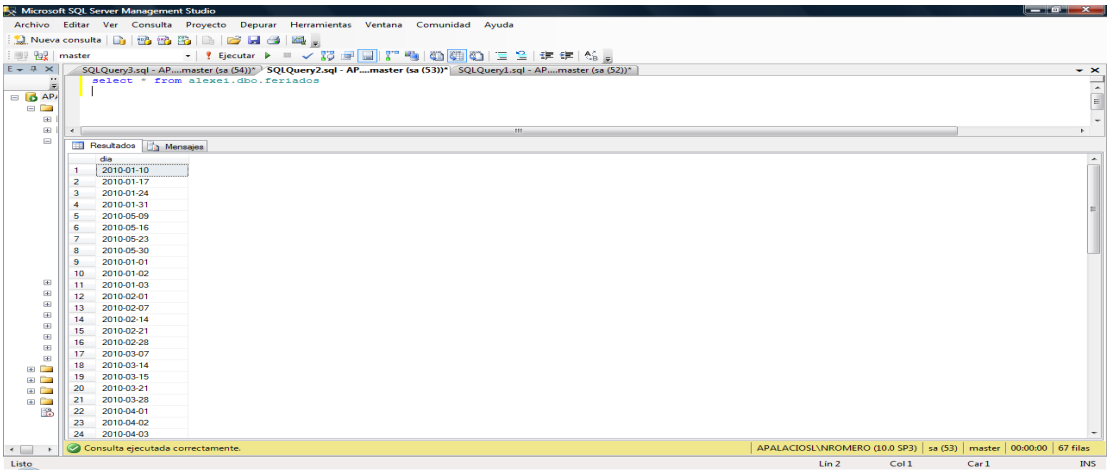
Las líneas anteriores de código, se realizaron para obtener la tabla *alexei.dbo.usuarios.*



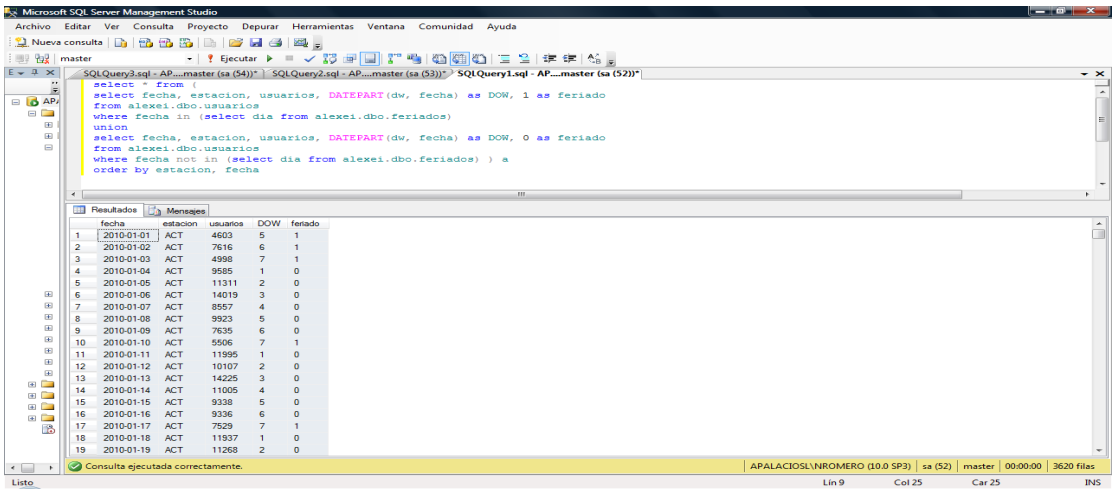
A continuación se ejecutan las siguientes líneas del código para efectuar el siguiente query:

*select * from (*
select fecha, estacion, usuarios, DATEPART(dw, fecha) as DOW, 1 as feriado
from alexei.dbo.usuarios
where fecha in (select dia from alexei.dbo.feriados)
union
select fecha, estacion, usuarios, DATEPART(dw, fecha) as DOW, 0 as feriado
from alexei.dbo.usuarios
where fecha not in (select dia from alexei.dbo.feriados)) a
order by estacion, fecha

El query anterior hace uso de otra tabla que contiene los días feriados:



Para obtener los resultados que se copian y pegan en el archivo *"Acumulados.xls"*



Referencias: libros, publicaciones periódicas y páginas web

Libros:

1. **Anderson,** T W; (1972). ***"The Statistical Analysis of Time Series";*** ed. John Wiley & Sons; ISBN--0-471-02900-9.
2. **Australian Government & AusAID**; (2008). ***"Urban rail Development in China, Issues and Options";*** ed. Institute of Comprehensive Transport (NDRC) & World Bank.
3. **Beat,** Kleiner; **Chambers** John, **Cleveland**, William and **Tukey**, Paul; (1983). ***"Graphical Methods for Data Analysis";*** Wadsworth International Editions.
4. **Boes**, Duane C; **Graybill,** Franklin A; **McFarland Mood**, Alexander; (1974). ***"Introduction to the theory of Statistics";*** ed. McGraw Hill International, Third Edition.
5. **Burt**, Holland, **Simon** P. Washington; (1998). ***"Statistical Analysis and Data Display with R, SAS y S-Plus";*** ed. Cambridge University Press.
6. **Chatfield,** C; (1990). **"*The Analysis of Time Series: An Introduction*";** 3th Edition, Chapman & Hall; ISBN: 0-412-26030-1.
7. **Ciudadanos en red-Metropolitana**; (6 de noviembre de 2006). **"*Ocupa el Metro del DF; el tercer lugar mundial en captación de usuarios*";** ed. Sistema de Transporte Colectivo, Revista Metrópoli 2025.
8. **Conover** W.J; (2009). ***"Practical non parameter statistics";*** Third Edition, John Wiley & Sons, Inc.
9. **Consejo Nacional de Población** (CONAPO), (2010). ***"Escenarios Demográficos y Urbanos de la Zona Metropolitana del Valle de México";*** ed. CONAPO.
10. **Cover**, T.M. & **Thomas** J. A. (1991). **"*Elements of Information Theory*";** ed. John Wiley and Sons, Inc.
11. **Cramer**, H. **(**1946). ***"Mathematical Methods of Statistics";*** ed. Princeton University Press, Princeton, NJ.
12. **Davidson**, R. and **MacKinnon**, J.G. (1993). ***"Estimation and Inference in Econometrics";*** New York: Oxford University Press.
13. **Espitia González,** Carlos Giovanni, (2009). ***"Modelos con variable dependiente discreta y limitada";*** ed. Departamento de Economía Universidad ICESI, España.

14. **Gaceta Oficial del Distrito Federal,** Consejería Jurídica y de Servicios Legales del Distrito Federal, (6 de noviembre de 2007); "***Estatuto Orgánico del Sistema de Transporte Colectivo, Metro***"; Gobierno del Distrito Federal.
15. **Girija,** G; **Raol,** J.R; & Singh, J; (2004). ***"Modeling and Parameter Estimation of Dynamic Systems";*** IEE control Series 65, ISBN: 0-86341-363-3.
16. **González**, Ovidio, **Navarro,** Bernardo; (1989). "***Antecedentes del Transporte en la Ciudad de México***"; Universidad Nacional Autónoma de México, Universidad Autónoma Metropolitana, Instituto de Investigaciones Económicas. ISBN-968- 36-1106-0.
17. **Greenberg**, D; (2003). ***"Longitudinal Data Analysis, personal communication: referring to the research of Nathaniel Beck";*** John Wiley & Sons.
18. **Greene**, W. H; (2003). "***Econometric Analysis***"; Upper Saddle River, 5th ed. Prentice Hall.
19. **Greene**, W. H. (2003). ***"LIMDEP Version 8 Econometric Modeling Guide";*** Vol. 1. Plainview, NY: Econometric Software.
20. **Gujarati**, D. (2003), "***Basic Econometrics***"; 4th Ed. New York: McGraw Hill.
21. **Heiberger,** Richard M; **Holland,** Burt; **Washington,** Simon P; (1999). ***"Statistical & Economical Methods for Transportation Data Analysis";*** ed. Cambridge University Press.
22. **Hoaglin,** David & **Velleman,** Paul (1981). ***"The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis";*** Duxbury Editions.
23. **INEGI** (2010). "***Censo de población y vivienda 2010, resultados preliminares***"; ed. Instituto Nacional de Estadística, Geografía e Informática.
24. **Johnson,** Don; **(**1999). ***"Probability Density Estimation";*** Ed. The Connexions Project and licensed under the Creative Commons Attribution License.
25. **LEE**, In-Keun PhD, (2004). "***Experiences in Seoul Subway Development***"; Planning and Design Office of Subway Construction, ed. Seoul Metropolitan Government.
26. **Lifton**, Joshua H; (31 March 1999). ***"Measure Theory and Lebesgue Integration";*** ed. Swarthmore College Mathematics, Senior Conference.
27. **Louviere,** J. J; **Swait,** J. D; & **Hensher,** D. A; (2000). ***"Stated choice methods: analysis and applications";*** Cambridge: Cambridge University Press.
28. **Martinez**, Angel R. & **Martinez,** Wendy L; (2005). "***Computational Statistics Handbook with MATLAB***"; Ed. Chapman & Hall/CRC, 2002; ISBN: 1-58488-229-8.

29. **Myojo**, Shuichi, (2004). **"*Daily Estimation of Passenger Flow in Large and Complicated Urban Railway Network*";** ed. Railway Technical Institute; Tokyo, Japan.
30. **Nasser,** E Nahi; (1969). ***"Estimation Theory & Applications";*** John Wiley and Sons ISBN: 471-62870-0.
31. **Polanski,** Alan M; (2011). ***"Introduction to Statistical limits Theory";*** CRC-Press, ISBN: 0-978-1-4200-7660-8.
32. **Shiryayev,** A. N; **Lipster,** R.S; (1977). ***"Statistics of Random Processes I: General Theory";*** ED: Springer-Verlag, New York.
33. **Silverman**, B. W. (1986). **"*Density Estimation*";** ed. Chapman and Hall, London.
34. **Sistema de Transporte Colectivo,** (1997). ***"Los constructores: los hombres del metro";*** ed. Departamento del Distrito Federal; ISBN 968-816-138-1.
35. **Sistema de Transporte Colectivo,** (7 Noviembre 2001). **"*Plan maestro de Transporte Colectivo*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
36. **Sistema de Transporte Colectivo,** (2006). **"*Cifras de Operación en 2006*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
37. **Sistema de Transporte Colectivo,** (2007); **"*Cifras de Operación 2007*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
38. **Sistema de Transporte Colectivo**, (2007); **"*Decreto de creación del Sistema de Transporte Colectivo*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
39. **Sistema de Transporte Colectivo,** (2007); **"*Etapas de la Construcción de la red del STC Metro*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
40. **Sistema de Transporte Colectivo,** (2010), **"*Cifras de Operación 2010*";** ed. Sistema de Transporte Colectivo, Ciudad de México, México.
41. **Taylor,** Allen G; (2009). ***"SQL for Dummies";*** 7th Edition; ISBN: 978-0-470-55741-9
42. **TCRP Web Document 5**, (1999). *"Transit Capacity and Quality of Service Manual";* First Edition, ed. TRB, Washington, DC.
43. **United Nations Organization;** (2009). ***"Ranking de las ciudades más pobladas del mundo";*** ed. World Urbanization Prospects.
44. **Verzani,** John; (2009); ***"Using R for Introductory Statistics";*** Chapman & Hall/ CRC; ISBN: 1–58488–450-9

45. **Wilks,** Samuel S. (1962); ***"Mathematical Statistics";*** ed. John Wiley and Sons, New York.
46. **Zegras,** P. Christopher; Diciembre de 2005; ***"II Conferencia Internacional: El Transporte y el uso del Suelo, Teoría y ejemplos";*** ed. Departamento de Usos Urbanos y Planificación (DUSP).

Publicaciones periódicas

A. **Abraham,** Wald; (1949). "A note on the consistency of the maximum likelihood estimate"; Annals of Mathematical Statistics; Volume 20: page: 595-601.
B. **Castañeda Narváez,** Carlos y Noreña Casado, Francisco (1985). "Planeación y Construcción en líneas del metro"; Ingeniería civil (Volumen 231): pp. 9-64.
C. **Choi,** Hannah; **Park,** Jong-Soo; Pohang University Of Science and Technology, **Lee,** Keumsook; Department of Geography, Sungshin Women's University, Seoul 136-742; **Jung,** Woo-Sung, Department of Physics and Basic Science Research Institute; (October 2010). ***"Sleepless in Seoul: The Ant and the Metro hopper"***; Volume 6, arXiv: 1010.1165v1 [physics.soc-ph].
D. **Choi,** M. Y; **Lee,** Keumsook; Park, Jong-Soo; **Jung,** Woo-Sung,; Department of Physics, Pohang University of Science and Technology, Republic of Korea, (2010). "***Statistical analysis of the Metropolitan Seoul Subway System: Network Structure and Passenger Flows***"**;** Volume 6, arXiv: 1010.1165v1 [physics.soc-ph].
E. **Choi,** Hannah; **Park,** Jong-Soo; Pohang University Of Science and Technology; **Lee,** Keumsook; Department of Geography, Sungshin Women's University, Seoul; Jung, Woo-Sung, Department of Physics and Basic Science Research Institute; (23 February 2011). ***"Master equation approach to the intra-urban passenger flow and application to the Metropolitan Seoul Subway system";*** IOP Publishing Journal of Physics: Mathematical and Theorical.
F. **Conover,** W. J; (Sep., 1972). ***"A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions";*** Journal of the American Statistical Association Vol. 67, No. 339, pp. 591-596; American Statistical Association.
G. **Consejería Jurídica y de Servicios Legales del Distrito Federal,** (21 de abril de 2009). "***Decreto por el que se adicionan y derogan diversas disposiciones del reglamento interior de la administración pública del Distrito Federal***"**;** ed. Gaceta Oficial del

Distrito Federal, Ciudad de México, México: Gobierno del Distrito Federal; Volumen 575: pp. 3-5.

H. **Davis,** Diane E; ***"Urban Leviathan: Mexico City in the Twentieth Century";*** Political Science Quarterly, Vol. 111, No. 4 (Winter, 1996-1997), pp. 737-738.

I. **De Bono,** Arthur: Associate Professor, Mr. Coxon, Selby; Dr. Burns, Karen; Dr. Napper, Robbie; (2011). ***"Un examen de los tres enfoques del Metro en Material Rodante: Diseño para mejorar los tiempos de permanencia prolongada debido al crecimiento de pasajeros y el desplazamiento asociado"***; Australasian Transport Research Forum; Proceedings 28-30, September 11; Adelaide Australia.

J. **Farrel;** P. J. and Stewart; K. R; (2006). ***"Comprehensive Study Of Test of Normality And Symmetry: Extending the Spiegel Halter Test";*** Journal of Statistical Computation and Simulation, Vol. 76, N° 9; pp. 803-816.

K. **Gnanadesikan,** R. and **Wilk,** M. B. (1968). ***"Probability Plotting Methods for the Analysis of Data"***; Biometrika, Volume #5, pp. 1-19.

L. **Goodwin,** P.B., (1992), ***"A Review of New Demand Elasticities with Special Reference to Short and Long Run Effects of Price Changes";*** Journal of Transport Economics and Policy, Vol. 26, pp. 155-169.

M. **Jones,** M. C. & **Silverman,** B.W; (1989). ***"E. Fix and J. & L. Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation";*** International Statistical Review, Volume 57(3), pp. 233–247.

N. **Khale,** Thomas; August 21-2006. ***"The Glivenko-Cantelli Theorem and his Generalizations";*** Annals of Probability, Volume 15, pages: 837-870.

O. **Louviere,** J. J. (1988). ***"Conjoint analysis modeling of stated preferences";*** Journal of Transport Economics and Policy, Volume 22: page: 93-119.

P. **Nornadiah,** Modh-Razali & **Yap Bee** Wah, (2011). ***"Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests";*** Journal of Statistical Modeling and Analytics; Vol. 2, N° 1, 21-33; Malaysian Institute of Statistics.

Q. **Ortúzar,** Juan de Dios & **Román,** Concepción, (2003). ***"El problema de modelación de demanda desde una perspectiva desagregada: el caso del transporte";*** Revista Eure (Vol. XXIX, Nº 88), pp. 149-171.

R. **Slud;** Eric, (February 6, 2009). ***"Handout on empirical Distribution Function and descriptive Statistics";*** Statistics # 430.

S. **Tamin,** Ofyar Z; **Sulistyorini,** Rahayu; (2009). ***"Public Transport Demand by Calibrating the Combined Trip Distribution Mode Choice (TDCM) Model from Passenger Counts";*** World Academy of Science, Engineering and Technology, Volume 54.

T. **Vickrey,** William S; (May. 1963). ***"Pricing in Urban and Suburban Transport"***; the American Economic Review, Vol. 53: No. 2, Papers and Proceedings of the Seventy-Fifth Annual Meeting of the American Economic Association, pp. 452-465.

Páginas Web

1. **http://mapserver.inegi.gob.mx/geografia/espanol/datosgeogra/basicos/estados/df_geo.cfm**
2. **Artículo en donde se define la nueva delimitación del Área Metropolitana del Valle de México.**
3. **http://www.conapo.gob.mx/publicaciones/dzm2005/zm_2005.pdf**
4. **http://www.inegi.org.mx/**
5. **http://www.demographia.com/db-worldua2015.pdf**
6. **Intern@te en el metro-El Plan Maestro del Metro**
7. **http://www.metropoli.org.mx/htm/areas/5/tranvia.pdf pág. 11.**
8. **http://gulliver.trb.org/publications/tcrp/tcrp_webdoc_5-a.pdf.**
9. **http://www.consejeria.df.gob.mx//uploads/gacetas/ABRIL_24_09.pdf.**
10. **http://www.consejeria.df.gob.mx//uploads/gacetas/Noviembre07_06_206.pdf.**
11. **http://onlinepubs.trb.org/onlinepubs/nchrp/cd-22/v2chapter5.html**
12. **www.petards.com**
13. **http://statpages.org/javasta2.html#Excel**
14. **http://cnx.org/content/m13418/latest/**
15. **http://en.wikipedia.org/wiki/Measure_theory**
16. **http://www.pindling.org/Math/Statistics/Textbook/Chapter7_inference_mean_proportion/estimator.htm**
17. **http://www.vosesoftware.com/ModelRiskHelp/index.htm#Probability_theory_and_statistics/Probability_theory_and_statistics_introduction.htm**
18. **http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm#e**