Identificación de incidentes de tránsito a partir de registros GPS del transporte público

Patricio Álvarez-Mendoza*, Álvaro Olivares
Universidad del Bío Bío, Chile

Tomás Echaveguren
Universidad de Concepción, Chile

* Autor para correspondencia:
palvarez@ubiobio.cl

RESUMEN

Los operadores de transporte público en las rutas licitadas del Gran Concepción están utilizando dispositivos de posicionamiento global (GPS) y cámaras de video para controlar de manera más eficiente la operación y recaudación de su flota. Dichos datos son almacenados en bases de datos administradas por empresas informáticas locales. Dado que los dispositivos GPS ya han sido instalados en los vehículos de transporte público, el costo adicional de almacenar y administrar dichos datos es significativamente bajo comparado con el costo de instalar dichos dispositivos en la flota, y por lo tanto en la medida que la recolección y almacenamiento de los datos capturados por los dispositivos GPS se masifique, y que además se implementen los protocolos para acceder a ellos, resulta natural pensar en explorar el potencial uso de dicha información. En este estudio, se propone una metodología que permite caracterizar la operación del sistema de transporte público por medio de la identificación de patrones y anomalías a partir de los datos censados por los dispositivos GPS. Los resultados de la aplicación de la metodología a un caso de estudio dan cuenta del potencial de información contenida en los registros históricos de la operación del sistema de transporte público. En particular la metodología permite identificar la frecuencia, duración y severidad de incidentes de tránsito, y en consecuencia su explotación es de interés para la implementación de planes de manejo de incidentes de tránsito en corredores de transporte público.

Palabras clave: incidentes de tránsito, GPS

ABSTRACT

The public transportation system in the metropolitan area of Concepción, Chile has already implemented a global positioning system (GPS) and video cameras to improve the system performance, safety and revenue control. Such data is stored in databases in a number of different local technological operators. Considering that the hardware is already in place, the cost of storing and processing such data is relatively low as compared with the cost of implementing the data acquisition hardware. Therefore, as the volume of collected data increases and the protocols to access the data are developed, it sounds natural to explore other potential applications beyond fleet management. In this study a methodology to characterize the system operation is proposed. The methodology is based on the automatic recognition of patterns and anomalies in the trajectory data registered by the GPS system. The results show that the methodology can extract relevant information from the historic GPS records that is not available from other sources. Particularly, the methodology allows recognizing traffic incident frequency, duration and severity. Thus mining the GPS database using the proposed methodology can help the development of traffic incident management plans in public transportation corridors.

Keywords: traffic incidents, GPS

1. INTRODUCCIÓN

En los últimos años el aumento de la demanda de transporte y del volumen de tránsito han causado aumentos significativos en la congestión, y todo indica que este problema seguirá agravándose. Su principal manifestación es la progresiva reducción de las velocidades de circulación, que se traduce en incrementos de tiempos de viaje, tiempos de viaje menos confiables, aumento del consumo de combustible, polución atmosférica y aumento de otros costos de operación. Además, la lentitud de desplazamiento aumenta la frustración de los conductores y fomenta el comportamiento agresivo de ellos.

Dependiendo del origen que pueda tener la congestión de tránsito, esta se puede clasificar en dos tipos: congestión recurrente y congestión no recurrente (Pardillo y Sánchez 2015). La congestión recurrente, se refiere principalmente al hecho de que la demanda por usar la infraestructura de transporte excede la capacidad de la misma y por ende a menudo es considerada un problema de dimensionamiento que es lógicamente combatido aumentando la capacidad del sistema ya sea por medio del incremento de la infraestructura física o por medio de la gestión de la existente. En general este tipo de congestión tiende a concentrarse en períodos cortos de tiempo, típicamente conocidos como "horas punta". Por lo tanto este tipo de congestión se refiere a un fenómeno de carácter repetitivo y predecible. Comúnmente, la congestión recurrente es analizada por medio de procesos de planificación con metodologías probadas y bien conocidas en el ámbito nacional.

Por otro lado, la congestión no recurrente se refiere a un tipo de congestión que ocurre de manera irregular, generalmente asociada a eventos que reducen la capacidad del sistema de transporte y que suceden de forma independiente del aumento de la demanda en las horas punta. La congestión no recurrente es el resultado de accidentes de tránsito, vehículos en "panne", malas prácticas de conducción (estacionar en doble fila, etc.), presencia de basura o elementos extraños en el sistema de transporte, actividades de mantención del sistema de transporte (bacheos, mantención de semáforos, etc.), faenas de construcción, actividad policial, clima adverso, y en general cualquier otra actividad no rutinaria en el sistema de transporte.

Las causas mencionadas anteriormente se denominan en general incidentes de tránsito y se pueden agrupar en tres categorías principales (McGroarty, 2010): incidentes de tránsito (50%), faenas constructivas (15% - 25%) y clima adverso (10%).

Estudios previos han indicado que los incidentes son una de las principales causas de pérdida de tiempo y aumentos de costos en las redes de transporte, por ejemplo, en los Estados Unidos, se determinó que en el 2003, más del 60% de la congestión en las autopistas urbanas fue causada por incidentes, y ese indicador se estima que fue de más del 70% en 2005 (Lomax *et.al.*, 2003).

A diferencia de la congestión recurrente, las características de la congestión no recurrente, esto es severidad (cantidad de pistas afectadas), frecuencia (periodicidad de ocurrencia de incidentes) y duración (tiempo en que se ve afectada la capacidad de la vía), no han sido abordadas en el medio local y por ende no existen antecedentes suficientes que permitan estimar el costo de estas externalidades en la operación del sistema de transporte.

Dado lo anterior, es posible razonar que existirán ciertos umbrales a partir de los cuales es posible proponer planes de manejo de incidentes tales que el costo de su implementación sea inferior a los ahorros obtenidos producto de la operación de dichos planes, y por esta vía mitigar los efectos de la congestión no recurrente. Por tal motivo entonces, resulta relevante cuantificar la magnitud de la congestión no recurrente para evaluar la conveniencia de la implementación de dichos planes.

2. REVISIÓN BIBLIOGRÁFICA

2.1 Incidentes de tránsito y su detección

Los incidentes de tránsito pueden ser definidos como cualquier evento que interrumpe el normal funcionamiento de la infraestructura de transporte, degradando la seguridad y reduciendo la capacidad. Estos eventos incluyen: vehículos averiados, accidentes de tránsito, actividades de mantenimiento de vías, condiciones climáticas adversas, protestas, escombros en la carretera, etc. La congestión de tránsito generada por los incidentes (incluyendo los impactos secundarios) tiene efectos perjudiciales en la seguridad pública, la economía local y el medio ambiente. Por lo tanto la cuantificación y caracterización de estos eventos son esenciales a la hora de la implementación de planes de manejo de estos incidentes y en consecuencia la disminución de sus efectos. El manejo de incidentes conlleva importantes beneficios, tales como la reducción de los retrasos en los tiempos de viaje de los vehículos debido a incidentes, a través de la reducción de la frecuencia de incidentes y la mejora de la respuesta y el tiempo de despacho de unidades de asistencia al incidente.

Los incidentes de tráfico fundamentalmente reducen la capacidad disponible de una carretera o degradan su rendimiento, lo que se expresa en velocidades de operación más bajas y en una mayor congestión. También pueden aumentar la probabilidad de incidentes secundarios y una degradación del rendimiento en calles que ni siquiera están directamente influenciados por el incidente, a través de circunstancias como el conocido fenómeno rubbernecking (Cambridge Sys Inc. *et al.*, 1998). El término se refiere al acto de curiosear por medio de girar o estirar el cuello con el fin de obtener una mejor visión. En operaciones de tráfico dicho acto se da en presencia de incidentes y anormalidades en la vía que típicamente además conlleva una reducción de velocidad con la consecuente pérdida de capacidad de la vía.

2.2 Estudios previos

En los últimos años se han implementado diversos sistemas de control de tránsito que nacieron con el objeto de supervisar, controlar, administrar y mejorar la gestión de tránsito de un sector urbano o vial. Estos sistemas se basan en datos obtenidos mediante diversas técnicas, ya sea mediante el uso de vehículos sonda, equipados con GPS ("probe cars"), mediante equipos de conteo automático basados en espiras, circuitos CCTV y software de visión artificial y más incipientemente en Chile el uso de protocolos de comunicación de corta distancia tales como bluetooth.

La experiencia internacional en relación al uso de la información recolectada con este tipo de dispositivos indica que los detalles adicionales que proporcionan los datos GPS, tanto en términos espaciales como temporales, permiten un mejor entendimiento y representación de la operación y de las condicionantes que determinan el comportamiento de la red de transporte. En particular, los datos GPS que cubren periodos extensos de tiempo entregan la oportunidad de clasificar los días y eventos del año en diferentes patrones de comportamiento y así por ejemplo considerar que en ciertos corredores pudiese resultar relevante distinguir entre diferentes temporadas del año o reconocer de forma especial aquellos días en donde se realicen eventos especiales (Álvarez *et al.*, 2010). El uso de los datos GPS permite entre otras aplicaciones determinar los efectos en la operación de la red producto de variaciones de la demanda horaria, estacional, o por eventos especiales, accidentes de tránsito, trabajos en la vía, y condiciones climáticas entre otras (Cambridge Systematics, 2010; U.S. Department of Transportation, 2004; Courage y Lee, 2008).

De igual manera, a partir de información capturada por dispositivos GPS diversos estudios han utilizado diferentes técnicas de "datamining" o minería de datos, para evaluar los niveles de congestión de tráfico, logrando clasificar segmentos de carretera en varios niveles según la velocidad promedio de desplazamiento de los vehículos. (Diker, 2012; Yong-Chuan, 2011).

Skabardonis *et al.* (2003) desarrollaron en California una metodología preliminar para cuantificar la congestión recurrente y no recurrente usando datos del sistema de medición del desempeño de autopistas de California (PeMS) en conjunto con informes de incidentes de la patrulla de carreteras de California (CHP). Usando estos datos, el estudio fue capaz de caracterizar ambos tipos de congestión, recurrente y no recurrente, en segmentos de carretera seleccionados, e identificar el origen de la congestión no recurrente. El estudio también arrojó que la congestión no recurrente es función de las características del segmento y el grado de congestión recurrente.

En el caso chileno, Cortés *et al.* (2011) presenta una metodología basada en información dinámica proveniente de los buses del Transantiago para la estimación en tiempo real de velocidades de operación media. Para ello se utilizan datos GPS los cuales permiten construir trayectorias rectificadas que pueden ser utilizadas para estimar velocidades de operación.

3. METODOLOGÍA

El objetivo del estudio dice relación con la caracterización de la operación del sistema de transporte público por medio de la identificación de patrones y anomalías a partir de los datos censados por los dispositivos GPS. Para lograr este objetivo se desarrollaron dos funciones que se aplican a los datos contenidos en los archivos históricos de operación. La primera función consiste en la extracción automatizada de la información según el segmento o corredor de la red vial que se desee estudiar. Estos datos constituyen el insumo de la segunda función, en la cual se categorizan (clasifican) las diferentes trayectorias que efectúan los vehículos de transporte público, con el fin de identificar la variabilidad de las trayectorias en función del tiempo y en particular aislar aquellas trayectorias anómalas para un estudio más profundo de las posibles causas que la originan. En las secciones sub siguientes se detallan las funciones antes descritas y se muestra un caso de estudio que ilustra el potencial de la metodología.

La experiencia internacional en relación al uso de la información recolectada con este tipo de dispositivos indica que los detalles adicionales que proporcionan los datos GPS tanto en términos espaciales como temporales permiten un mejor entendimiento y representación de la operación y de las condicionantes que determinan el comportamiento de la red de transporte. En particular, los datos GPS que cubren periodos extensos de tiempo entregan la oportunidad de clasificar los días y eventos del año en diferentes patrones de comportamiento y así por ejemplo considerar que en ciertos corredores pudiese resultar relevante distinguir entre diferentes temporadas del año o reconocer de forma especial aquellos días en donde se realicen eventos especiales. El uso de los datos GPS permite entre otras aplicaciones determinar los efectos en la operación de la red producto de variaciones de la demanda horaria, estacional, o por eventos especiales, accidentes de tránsito, trabajos en la vía, y condiciones climáticas entre otras.

3.1 Función pre-procesamiento de datos y extracción de trayectorias

En este proyecto se desarrollaron herramientas de pre-procesamiento, imputación y extracción de datos que fueron posteriormente implementadas en MATLAB. Estas herramientas se aplican a los datos contenidos en el repositorio ELESIS y producen la información base para alimentar las funciones de clustering o clasificación. ELESIS es el proveedor tecnológico para el sistema de adquisición y almacenamiento de datos GPS de alguna línea de transporte público del Gran Concepción. En este sentido, el pre-procesamiento de datos consiste en leer los datos temporales y de posición desde archivos Excel que contienen los datos crudos de ELESIS, y luego representar la distancia acumulada en función del tiempo. Lo anterior teniendo presente un área de estudio consistente en un corredor o tramo de la red vial la cual queda definida por un punto de inicio y un punto final en la trayectoria del vehículo.

3.1.1 Determinación de límites del área de estudio

Uno de los propósitos del pre-procesamiento de datos es obtener la distancia y tiempo acumulados producto del recorrido hecho por cada bus, desde el momento en que este ingresa hasta que sale de un corredor o tramo específico. Para esto, en primera instancia se fijan los límites geográficos (en UTM) del corredor en estudio, de manera de descartar los puntos de la trayectoria del vehículo que se encuentran fuera de dicha sección. Para esto fueron implementados dos tipos de límites. El primer límite corresponde a las ecuaciones de las rectas paralelas a la sección del recorrido de interés (línea segmentada). El segundo límite corresponde simplemente a las latitudes o longitudes que definen el inicio y final del tramo de interés (línea punteada). La definición de estos límites permite verificar si el punto reportado por el GPS está dentro o fuera del área de estudio y en consecuencia decidir si dicho punto debe ser considerado como parte de la trayectoria acumulada en función del tiempo. En este sentido, y a modo de ejemplificar este proceso, en la Figura 1 se muestran los límites trazados para trayectorias realizadas en una trayectoria de prueba, donde se marca el inicio y fin de las trayectorias a analizar, además de descartar puntos pertenecientes a tramos paralelos a la sección en estudio.

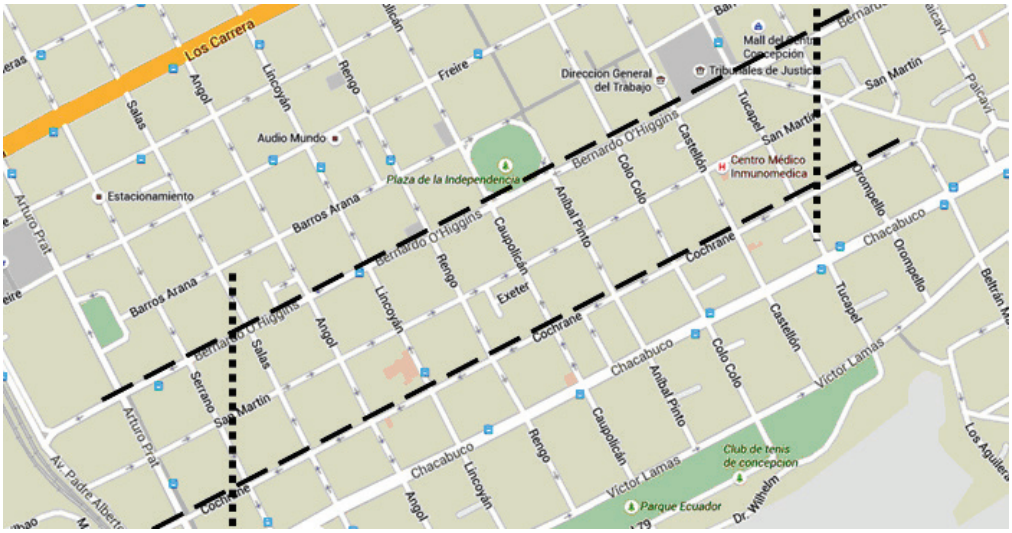
3.1.2 Imputación de datos

Un problema frecuente en la base de datos ELESIS está dado por la ausencia de datos o repetición de estos en algunos períodos de tiempo. Como se mencionó anteriormente el criterio para manejar este problema fue remover dichos registros, generando archivos compactos sin vacíos de información.

3.1.3 Conversión de coordenadas

Los archivos con los datos de ELESIS en su formato original no permiten el cálculo de distancias en un sentido cartesiano. Es por esto que como parte de la función de pre- procesamiento se incluyó una rutina que permite la conversión del sistema coordenadas geográficas proporcionado por ELESIS, a coordenadas UTM (Universal Transverse Mercator). Esta conversión permite transformar grados sexagesimales a metros. Este proceso constó de dos etapas; en la primera se modificó el formato de la posición de texto, por ejemplo 36495082 a formato grados sexagesimales (36,82447°), para luego realizar la conversión a coordenadas UTM (666972,9450 m).

Figura 1: Visualización de límites trazados



3.1.4 Cálculo de distancia y tiempo acumulados

Una vez conocidas las coordenadas X,Y se obtiene el tiempo de viaje entre dos puntos consecutivos simplemente restando el horario (time stamp) entre dichos puntos. En el caso de trayectorias que crucen los límites del área de estudio, dicho tiempo se obtiene interpolando proporcionalmente a la posición del límite. Análogamente se obtienen las distancias, utilizando para esto la fórmula clásica para la distancia euclidiana, esta vez con la latitud y longitud de dos puntos consecutivos, e interpolando en caso de que las trayectorias crucen los límites del área de estudio. Finalmente, las relaciones distancias acumuladas versus tiempo de viaje se obtienen sumando exclusivamente los tiempos y distancias incluidas dentro del corredor en estudio.

La Tabla 1 muestra el resultado de dicho proceso aplicado a una trayectoria de prueba, donde se observa una distancia recorrida de 2034,5 metros en 98,5 segundos. Además, se observa en los puntos 865 y 874 de la columna "Δtiempo", que el intervalo de tiempo es menor al resto, producto de la interpolación aplicada según lo explicado con anterioridad. Por otro lado, en los puntos 864 y 876, la distancia y tiempo acumulados es 0, debido a que dichos puntos se realizaron fuera de los límites del tramo en estudio.

Tabla 1: Visualización de distancias (m) y tiempos acumulados (seg.)

Punto	latitud (m)	longitud(m)	Δtiempo (s)	tiempo acumulado (s)	Δdistancia (m)	distancia acumulada (m)
...	...	...	...	...	...	...
864	672284,4	5922252,5	10	0	157,901	0
865	672141,7	5922185,4	6,82	6,820	107,482	107,482
866	671912,8	5922080,7	15	21,820	251,702	359,184
867	671737,5	5921999,2	10	31,820	193,372	552,556
868	671549,7	5921912,5	10	41,820	206,812	759,368
869	671352,1	5921823,1	10	51,820	216,934	976,302
870	671146,3	5921732,9	10	61,820	224,727	1201,029
871	670944,3	5921640,2	10	71,820	222,258	1423,287
872	670733,3	5921543,0	10	81,820	232,299	1655,585
873	670521,2	5921445,5	10	91,820	233,394	1888,979
874	670338,3	5921326,4	6,67	98,486	145,519	2034,498
876	670257,0	5921194,6	10	0,000	154,810	0
...	...	...	...	...	...	...

En algunos corredores existe doble sentido de tránsito, por lo cual se hace necesario además identificar aquellos recorridos que se hicieron en el sentido de interés. Para ello, se extraen los datos de distancia acumulada y tiempo acumulado correspondientes a las trayectorias realizadas por el tramo en estudio, y se verifica en cada una de ellas si el sentido del desplazamiento es el que se desea. Esto se logra comprobando si se cumple un incremento constante de latitud o longitud en el sentido necesario. Cabe señalar que el error aleatorio de los sensores GPS puede tornarse en un problema en la estimación de la distancia. Sin embargo, el objetivo de la metodología propuesta no es monitorear el comportamiento de los servicios de transporte público sino monitorear el estado de secciones críticas de la infraestructura. En ese sentido dicho problema se mitiga por la vía de elegir secciones de control rectas.

3.1.5 Agregación y normalización de datos

Los datos contenidos en ELESIS son registrados en intervalos del orden de segundos, sin embargo a partir de dicho intervalo es posible construir una descripción de la distancia acumulada versus tiempo para cualquier intervalo de tiempo. Lo anterior debido a que el potencial o la efectividad de la función de clasificación dependerá del nivel de agregación de los datos. De esta forma y pese a disponer de intervalos pequeños de tiempo, el análisis se realizará con períodos que eventualmente y dependiendo de la extensión del área de estudio pudiesen ser de mayor magnitud.

Del resultado del pre-procesamiento se extraen los datos de posición y tiempo. Estos datos resultantes se muestran en la Tabla 2, que contiene información para individualizar el bus que originó los datos, la fecha en donde se leyeron los datos, la hora a la cual se cruza el límite donde comienza el diagrama distancia acumulada versus tiempo y finalmente el registro detallado de la posición en metros para sucesivos intervalos regulares de tiempo hasta cruzar el límite de salida del área de estudio.

Tabla 2: Matriz de datos agregados

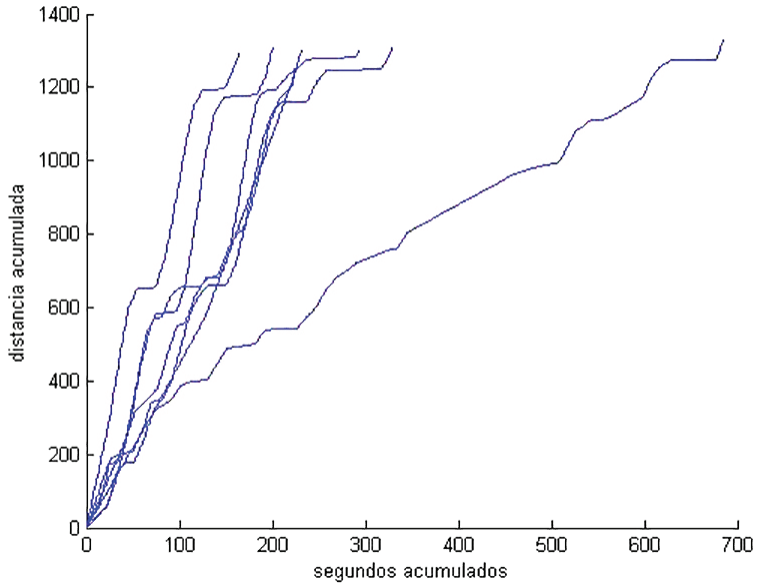
Abril-01-2013 Maq9.xlsx						
04-01-2013	04-01-2013	04-01-2013	04-01-2013	04-01-2013	04-01-2013	04-01-2013
8:22:36	10:26:31	12:28:20	14:45:33	17:03:11	19:20:58	21:38:34
0	0	0	0	0	0	0
27,76	57,53	119,30	26,82	48,30	74,83	43,80
55,52	114,16	238,60	52,87	125,36	149,65	87,60
127,21	169,10	377,87	114,49	193,25	176,37	131,40
170,45	228,01	528,57	212,28	203,20	231,40	175,20
177,82	303,50	626,63	329,33	209,97	340,76	219,00
229,65	336,09	653,03	459,15	264,48	473,20	262,79
329,96	411,16	702,45	573,40	347,84	584,41	352,96
...	...	...	...	...	...	...

Una vez finalizado el pre-procesamiento, es posible la representación gráfica de los datos de la matriz de datos agregados, como se muestra en la Figura 2, donde es posible apreciar las diferencias entre las trayectorias efectuadas en distintos momentos del día. En particular es posible observar las diferencias en los tiempos totales de viaje para la misma sección, cambios locales de velocidad y tendencias globales de velocidad de operación, expresadas por la pendiente de las curvas.

3.1.6 Síntesis de base de datos

La automatización de los procesos anteriores permite procesar todos los archivos Excel disponibles para el análisis, para finalmente consolidar en una planilla única todas las trayectorias de todos los buses que en algún momento transitaron el área de estudio. Para esto, se unen todas las matrices de datos agregados, correspondientes a cada archivo Excel procesado. Esta nueva matriz resumen, constituye el insumo básico para la función de clasificación, donde cada columna contiene la información necesaria correspondiente a cada trayectoria realizada dentro de un tramo que se desee estudiar.

Figura 2: Trayectorias efectuadas por un bus, en un día en un corredor determinado



3.2 Función de identificación y clasificación de patrones en trayectorias

Como ya se mencionó, el objetivo final de esta metodología tiene que ver con caracterizar la operación del transporte público por medio de la identificación de patrones y anomalías en el sistema. Para lograr este objetivo se desarrolló una función para poder categorizar (clasificar) las diferentes trayectorias que efectúan vehículos de transporte público a lo largo del día dentro de un corredor en particular. El objetivo de esta clasificación busca separar las trayectorias que se efectúan en condiciones normales de operación, es decir condiciones que se observan sistemáticamente todos los días en los distintos momentos del día, de aquellas que por una u otra razón se realizan en condiciones diferenciables de las anteriores. La lógica de la separación propuesta tiene que ver con la ocurrencia de incidentes o eventos no típicos asociables a eventos de congestión no recurrente. Aún más, se propone que por medio del análisis de la relación distancia acumulada tiempo acumulado de diversos vehículos afectados por el mismo evento se puede inferir información del mismo, por ejemplo: la duración e intensidad de dicho evento. Usando la misma lógica, y por medio del análisis de muchos eventos, es posible determinar la frecuencia de estos eventos no recurrentes.

La función de clasificación se basa en la comparación de un índice de similaridad que puede ser calculado a partir del diagrama posición versus tiempo para las diferentes trayectorias registradas por diferentes vehículos en diferentes horas del día y en diferentes días de un período de análisis dado. En particular, para este problema se consideró la utilización del algoritmo de clasificación k-means (Alpaydin, 2004). Este algoritmo utiliza un procedimiento iterativo de particionamiento, en donde se minimiza la suma de las distancias de los puntos que pertenecen a una misma trayectoria respecto del centroide de su clase; al mismo tiempo que las distancias entre centroides se maximiza.

3.2.1 Algoritmo de clasificación k-means

Para avanzar en la identificación de patrones, se desarrolló una función para poder categorizar (clasificar) las diferentes trayectorias que efectúan vehículos de transporte público a lo largo del día. Esta clasificación se basa en una comparación de la similaridad que presentan las diferentes trayectorias registradas por diferentes vehículos en diferentes horas del día y en diferentes días del período de evaluación (Alpaydin, 2004). En particular, para este problema se consideró la utilización del algoritmo de clasificación k-means. Este algoritmo es un procedimiento iterativo de particionamiento, en que se divide un conjunto de n elementos, en k clusters o clases, donde los elementos pertenecientes a una misma clase se asemejan entre ellos, y a su vez se diferencian de los elementos de las clases restantes. El algoritmo comienza seleccionando k elementos aleatoriamente, los cuales representarán el centroide o media de cada clase. Luego, cada elemento restante es asignado a la clase que más se le asemeje, basándose en una medida de la distancia entre el elemento y el centroide del cluster. A

continuación se recalcula el centroide de cada clase y se vuelve a asignar cada elemento a la clase de mayor similitud. El algoritmo itera hasta que los centroides no se modifiquen, y por ende, los elementos no cambien de clase.

Para este proyecto en particular, los elementos a clasificar corresponden a las trayectorias realizadas por los buses, descritas mediante curvas de posición versus tiempo. La distancia entre los puntos que pertenecen a una trayectoria y su centroide de clase es medida por medio de la distancia euclidiana definida mediante las ecuaciones 1 y 2.

$$dist\ (v_j, c_k) = \sum_i (v_j(t_i) - c_k(t_i))^2, \forall j \in k \quad (1)$$

$$c_k(t_i) = \frac{1}{n_k} \sum_j v_j(t_i)^2, \forall j \in k \quad (2)$$

Donde

$v_j\ (t_i)$ = posición j en el intervalo de tiempo i

$v_k\ (t_i)$ = centroide de la clase k en el intervalo de tiempo i, y

n_k = total de trayectorias en la clase k

La implementación de esta función permite especificar el número de clases que se espera como resultado del proceso, posibilitándose encontrar una clase en que se aprecien trayectorias realizadas de manera irregular o anómala con respecto al resto de las clases.

3.2.2 Especificación de número de clases

La implementación de esta función permite al analista especificar el número de clases o clusters (k) que espera como resultado del proceso, es decir en cuántas clases se desea dividir las trayectorias incluidas en la matriz de datos. Por supuesto, a mayor número de clases, estas más homogéneas son, sin embargo un número elevado de clases no es de interés dado que el objetivo es ayudar a identificar un número limitado de patrones en el sistema estudiado. Posteriormente, el algoritmo k-means, procede a realizar la clasificación de datos según el número de clusters definido por el usuario entregando en pantalla una descripción gráfica de los patrones encontrados. Esto último permite posteriormente identificar la fecha y hora en que realizaron las trayectorias que pertenecen a las clases de interés.

El proceso de la identificación de clases o clusters cuyas trayectorias se diferencian notoriamente del resto, es claramente un proceso iterativo que depende de la habilidad del analista para manejar la rutina de clasificación. En consecuencia, si dentro de los clusters resultantes no es posible distinguir alguno que agrupe trayectorias claramente diferenciables de la tendencia, se hace necesario repetir el proceso de clasificación desde el comienzo, aumentando o disminuyendo el número de clases o clusters según lo defina el analista.

3.2.3 Análisis de posibles incidentes detectados

Una vez identificado un clúster cuyas trayectorias presenten características de posibles incidentes de tránsito, se procede a individualizar y analizar los días y períodos de tiempo asociados a las trayectorias en dicho cluster. Para ello basta revisar la marca de clase que la rutina de clasificación registra con los datos consolidados originales. Con esto es posible analizar en detalle las trayectorias que son candidatas a incidente, y revertir el análisis a partir de la fecha y hora de duración para determinar otras trayectorias que eventualmente pudiesen contener información del incidente y que pueden haber quedado relevadas a un cluster de mayor duración o menor duración. Esto se debe principalmente a que posiblemente estos buses se vieron afectados durante menor cantidad de tiempo por el incidente de ese período, pero de igual manera presentan retrasos con respecto a las condiciones normales.

Además, si el tramo en estudio tuviese doble sentido de tránsito, entonces de manera complementaria es posible hacer un análisis del comportamiento del sentido contrario en el mismo horario del posible incidente, de manera de verificar si hay

efectos producto de la operación debido a la reducción de velocidad de los usuarios para poder observar lo que sucede en el sentido contrario de circulación.

3.3 Estimación de las características de la congestión no recurrente

La severidad corresponde al porcentaje de la capacidad afectada o a la cantidad de pistas que pudiesen bloquearse producto de la ocurrencia de un incidente de tránsito. La frecuencia corresponde a la periodicidad con que se produce determinado tipo de evento, y se puede determinar estudiando la proporción entre la cantidad de trayectorias con características de incidente, por sobre el total de trayectorias que se generen. Por último, la duración corresponde al tiempo que dura un incidente, desde que este se produce hasta que la capacidad del corredor vuelve a la normalidad.

Con el fin de estimar la duración del incidente, se propone utilizar las ecuaciones (3) y (4) las cuales provienen del análisis clásico de teoría de colas y que entregan la duración media de tiempo en cola y la duración máxima de un vehículo en cola (May, 1990).

$$\overline{d}_R = \frac{30 t_R \left(\lambda - \mu_R\right)}{\lambda} \qquad (3) \qquad , \qquad d_M = \frac{60 t_R \left(\lambda - \mu_R\right)}{\lambda} \qquad (4)$$

Donde

t_R = Duración del incidente (horas)
λ = Demanda (Veq/h)
μ_R = Capacidad reducida debido a la presencia de un incidente (Veq/h)
$\overline{d}_R$ = Duración promedio de cada vehículo en cola (minutos)
d_M = Duración máxima de un vehículo en cola (minutos)

Para analizar la duración de los incidentes de tránsito, es posible extraer desde las trayectorias la duración máxima (d_M) y duración promedio ($\overline{d}_R$) de los vehículos en cola producto del incidente, siendo ambos datos necesarios para el uso de las ecuaciones 3 y 4.

Por otro lado, la severidad de estos incidentes puede ser analizada mediante la velocidad de desplazamiento que tengan los buses, representada por la pendiente de las curvas. Si la velocidad tiende a 0 durante la ocurrencia de un incidente de tránsito, se podría concluir que se está ante un bloqueo total de pistas, mientras que si la velocidad se viera reducida, respecto de los valores usuales observados en el corredor en estudio, se estaría ante un bloqueo parcial de pistas.

4. CASO DE ESTUDIO

De manera de ejemplificar la metodología propuesta, se analizó una muestra aleatoria de la base de datos para el puente Llacolén en Concepción, en sentido desde San Pedro de la Paz hacia Concepción, cuyo procedimiento y resultados se presentan a continuación.

4.1 Base de datos

Para este proyecto fueron puestos a disposición una muestra aleatoria de datos correspondientes a recorridos efectuados por la línea de buses San Remo S.A. Esta muestra consta de 245 archivos Excel acompañados por sus respectivos archivos kml (Google Earth). Cabe señalar que cada archivo Excel contiene el recorrido efectuado por una máquina dentro de un día entero de operación.

Estas 245 trayectorias diarias, efectuadas por distintas máquinas, están distribuidas aleatoriamente en 44 días entre los meses de enero a mayo del 2013, además del mes de junio del 2012, como se indica en la Tabla 2.

Tabla 2: Distribución diaria de datos

Junio 2012						
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Marzo 2013						
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Enero 2013						
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Abril 2013						
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

Febrero 2013						
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28			

Mayo 2013						
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

4.2 Clasificación de trayectorias y detección de incidentes

Como se ha indicado, la metodología comienza con el pre-procesamiento de la base de datos disponible. En este caso, como el corredor es relativamente corto, y los tiempos de viaje también lo son, se ha optado por visualizar los datos cada 5 segundos (por simple conveniencia dado que la sección de control tiene sólo 2 km de largo), la mitad del intervalo promedio de captación de datos. Para ello se han interpolado datos aceptando las consecuencias de dicho proceso.

Producto del pre-procesamiento de datos, se obtuvieron 1204 trayectorias de buses en el puente Llacolén en el sentido antes señalado. Luego, el analista puede especificar el número de clases que espera como resultado del proceso de clustering. La Figura 3 muestra los resultados de aplicar esta función a los recorridos realizados por la línea de buses San Remo, usando diferente número de clases para la categorización. El set de datos inicial contiene registros de volumen en días laborales, fin de semana, días con incidentes de tránsito, diferentes estados del clima, eventos especiales y errores en el registro de datos. Por supuesto, a mayor número de clases, estas más homogéneas son, sin embargo un número elevado de clases no es de interés dado que el objetivo es identificar un número limitado de patrones del sistema estudiado.

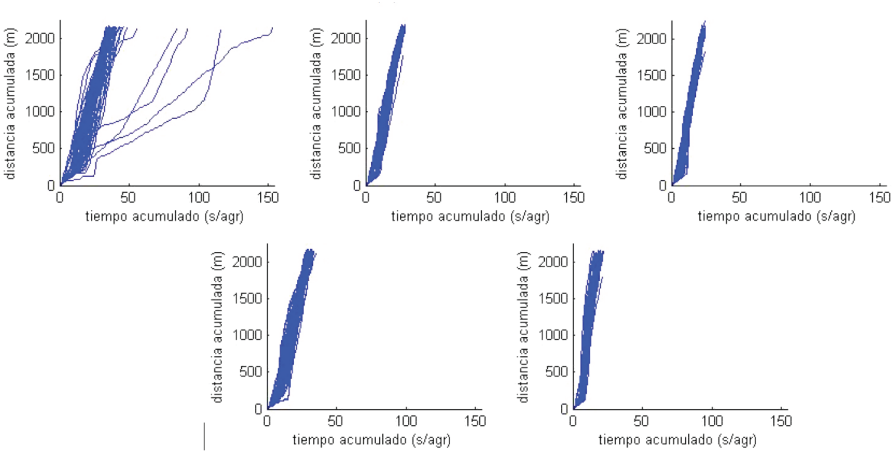
Como se ve en la Figura 3-a, la especificación de dos patrones no es suficiente, dado que si bien se aprecia un clúster con trayectorias de mayor duración, esta incluye eventos de congestión debido a las horas punta (congestión recurrente) e incidentes de tránsito, ya sean accidentes, protestas, obras de mantención vial, etc. (congestión no recurrente). Lo mismo sucede en la primera clase de la Figura 3-b. Finalmente la Figura 3-c muestra los resultados del análisis cuando se especifican 10 clases en donde como resultado del criterio elegido es posible observar cuatro posibles incidentes de tránsito.

Como se puede analizar de la figura anterior, las aplicaciones desarrolladas básicamente dividen los viajes según la velocidad en que estos fueron realizados, y por ende, los tiempos promedio de viajes de cada cluster o clase son similares. Es por esto, que los cluster que posean viajes de notoria mayor duración con respecto a los demás, como se aprecia en la clase 10 de la Figura 3-c, eventualmente pudiesen corresponder a incidentes de tránsito. Para la corroboración de estos probables incidentes, se crea una planilla Excel en que a cada trayectoria se asocia el número de cluster al que este pertenece, determinándose que tres de las trayectorias se produjeron el día 22 de junio del año 2012, entre las 13:00 y las 14:00 horas, todas ellas graficadas en la Figura 5.

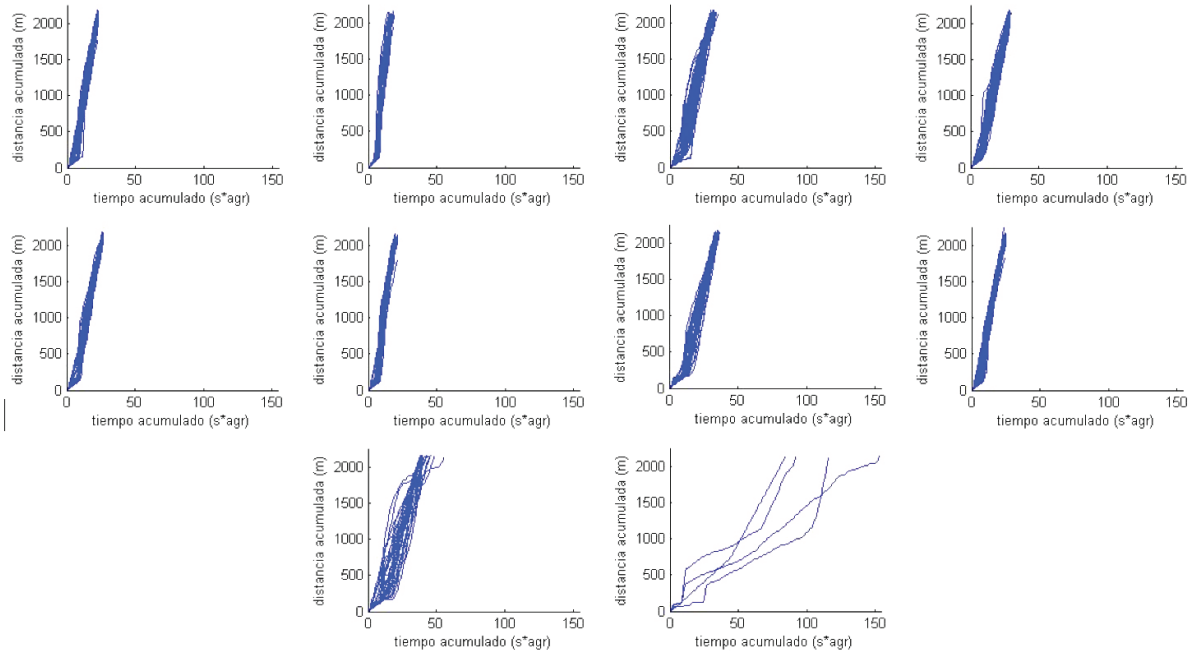
Por lo cual fue descartada la cuarta trayectoria de la clase en análisis, efectuada el día 06 de abril del 2013 a las 21:54 horas, ya que solo se tenía información de un solo bus para ese día. La Figura 4 muestra las trayectorias de tres diferentes buses afectados por el mismo incidente, y que serán utilizadas para la determinación de las características del mismo.

Figura 3: Resultado de la clasificación de datos, considerando 2, 5 y 10 clases

(a) Dos clases



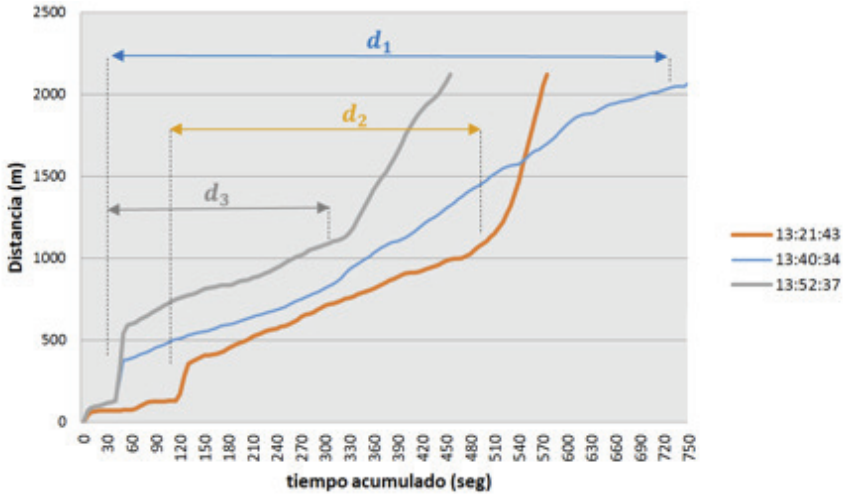
(b) Cinco clases



(c) Diez clases

Como se puede analizar de la figura anterior, las aplicaciones desarrolladas básicamente dividen los viajes según la velocidad en que estos fueron realizados, y por ende, los tiempos promedio de viajes de cada cluster o clase son similares. Es por esto, que los cluster que posean viajes de notoria mayor duración con respecto a los demás, como se aprecia en la clase 10 de la Figura 3-c, eventualmente pudiesen corresponder a incidentes de tránsito. Para la corroboración de estos probables incidentes, se crea una planilla Excel en que a cada trayectoria se asocia el número de cluster al que este pertenece, determinándose que tres de las trayectorias se produjeron el día 22 de junio del año 2012, entre las 13:00 y las 14:00 horas, todas ellas graficadas en la Figura 5. Por lo cual fue descartada la cuarta trayectoria de la clase en análisis, efectuada el día 06 de abril del 2013 a las 21:54 horas, ya que solo se tenía información de un solo bus para ese día. La Figura 4 muestra las trayectorias de tres diferentes buses afectados por el mismo incidente, y que serán utilizadas para la determinación de las características del mismo.

Figura 4: Trayectorias de incidente de tránsito en Puente Llacolén, 22 de junio, 2012



4.3 Caracterización de incidentes

Luego de tener identificados los incidentes, se procede a estimar los factores que caracterizan la congestión, que son la frecuencia, severidad y duración.

4.3.1 Frecuencia de incidentes

Como se comentó anteriormente, en la base de datos disponible se detectaron 1204 trayectorias efectuadas por buses de la línea San Remo S.A. por sobre el Puente Llacolén, de las cuales cuatro correspondieron a incidentes de tránsito. Por lo tanto, un 0,33% de las trayectorias efectuadas por este puente, desde San Pedro de la Paz hacia Concepción, se ven influenciadas por el fenómeno de congestión no recurrente. Sin embargo, este dato es posible de ser enriquecido significativamente si se utiliza una base de datos de tamaño superior.

4.3.2 Severidad de incidentes

El Puente Llacolén, posee dos pistas de circulación vehicular en cada sentido, por lo que hay dos alternativas de bloqueo de pista, bloqueo total, o bloqueo de una pista. Al analizar las trayectorias, es posible detectar que en estas jamás la velocidad es 0, por lo que se estima que la severidad del incidente detectado corresponde al bloqueo de una pista.

4.3.3 Duración de incidentes

Como se propone en este proyecto, la duración (t_R) será calculada mediante las ecuaciones 3 y 4, donde d_R corresponde al promedio de tiempo en minutos en que los buses se vieron afectados por el incidente. Este valor se calculó desde el gráfico de la Figura 4, como promedio entre las duraciones d_1 d_2 y d_3, resultando ser 7,5 minutos. Por otro lado la capacidad del puente y la demanda en ese horario es 3.100 y 2.245 [veq/h] (Bizama, 2012). Sin embargo es requerida la capacidad del puente producto de la severidad del incidente, la que reduce la capacidad a un 35% de la capacidad en condiciones normales (NAS, 2000). Por lo tanto μ_R corresponde a 1.085 [veq/h]. Finalmente, resolviendo para t_R en la ecuación 3, con los datos presentados en los párrafos anteriores, se obtiene que la duración del incidente o la duración en que la capacidad de la vía se vio afectada por este fue de 29 minutos.

Utilizando la ecuación 4 para d_M, (máxima demora individual) asumiendo que la trayectoria de máxima duración es un buen índice que representa esta condición se considera que 11.75 min se obtiene un valor de 23 minutos para la duración del incidente. Entonces se puede considerar que este incidente tuvo una duración aproximada similar al promedio de los valores antes calculados, es decir, 25 minutos.

5. CONCLUSIONES Y RECOMENDACIONES

Producto del análisis y procesamiento de la base de datos ELESIS, se hace notoria una falta de depuración de los datos, encontrándose variados errores tales como intervalos de tiempo de diversa duración sin registro de datos, períodos sin avance de tiempo e intervalos de tiempo en que se repiten los datos.

Una de las funciones desarrolladas en este proyecto logra la depuración de la base de datos, generando archivos compactos sin errores ni vacíos de información. Además logra la extracción de información referente a trayectorias realizadas por algún tramo vial que se desee estudiar. De este modo la información de la base de datos queda en condiciones para la aplicación de los procesos de clasificación de trayectorias.

El uso del algoritmo k-means es efectivo para separar los registros de trayectorias, facilitando la identificación de los potenciales candidatos a incidentes de tránsito, a través de la aislación de aquellas trayectorias realizadas de manera anómala con respecto al resto.

La aplicación de la metodología desarrollada es efectiva para la caracterización de los parámetros de la congestión no recurrente, esto es, frecuencia, duración y severidad.

La metodología propuesta y su aplicación, dan cuenta del potencial de información contenida en los registros históricos de la operación del sistema de transporte público. El potencial de información a extraer dice relación con información que hoy en día no está disponible, como lo es la frecuencia, duración y severidad de incidentes de tránsito, y en consecuencia su explotación es de interés para la implementación de planes de manejo de incidentes de tránsito en determinados corredores de transporte público del Gran Concepción.

Si bien el proceso de depuración de la base de datos, parte de la clasificación y la identificación de las características de los incidentes detectados han sido realizados manualmente por el analista, el potencial completo del concepto pasa por la automatización de gran parte de estas tareas. En este sentido existe abundante literatura que trata del problema de la depuración de bases de datos. Desde el punto de vista de la identificación de parámetros de los incidentes, el desafío más importante consiste en identificar el período de tiempo sobre el cual aplicar las métricas de interés. La identificación de dicho período se puede lograr por medio de la aplicación de métricas de similitud entre trayectorias. El desafio más importante consiste en automatizar el proceso de clasificación de trayectorias. Si bien el algoritmo KNN clasifica eficientemente el conjunto de datos, el número de grupos a utilizar es un input externo. Este último problema constituye parte de las líneas de investigación futura del proyecto.

Se recomienda que para la obtención de valores más representativos en cuanto a la duración, frecuencia y severidad de incidentes de tránsito, se utilice una base de datos de mayor tamaño, con registros continuos en el tiempo.

REFERENCIAS

Alpaydin, E. (2004) **Introduction to Machine Learning**. The MIT Press, Massachusetts.

Álvarez, P., Hadi, M. y Zhan, C. (2010) Using Intelligent transportation systems data archives for traffic simulation applications, **Transportation Research Record, Journal of the Transportation Research Board**, 2161, 29-39.

Bizama, J. (2012) Modelación y simulación mediante un microsimulador de la zona de influencia del Puente Llacolén. Memoria de Título, Universidad del Bio Bio.

Cambridge Sys Inc., H. Cohen and Science Applications International Corporation (1998) Sketch Methods for Estimating Incident Related Impacts. Report No. DTFH61- 95-00060: 21. Office of Environment and Planning, Federal Highway Administration, Washington, D. C.

Cambridge Systematics, Texas A&M University, Dowling Associates, Street Smarts, H. Levinson and H. Rakha. (2010) Analytical Procedures for Determining the Impacts of Reliability Mitigation Strategies.

Cortés, C.E., Gibson, J., Gschwender, A., Munizaga, M. y Zúñiga, M. (2011) Commercial bus speed diagnosis based on GPS-monitored data. **Transportation Research Part C**, 19(4), 695-707.

Courage, K.G. y Lee, S. (2008) Development of a central data warehouse for statewide ITS and transportation data in Florida: Phase II Proof of Concept. Florida Department of Transportation.

Diker, A.C. (2012) Estimation of traffic congestion level via FN-DBSCAN algorithm by using GPA data. Problems of Cybernetics and Informatics (PCI), 2012 IV International Conference, Baku, Azerbaijan.

National Academy of Sciences (NAS) (2000) **Highway Capacity Manual**.

Lomax T., Schrank, D., Turner, S. y Margiotta, R. (2003) Report for selecting travel reliability measures. Federal Highway Administration, Washington, D. C.

McGroarty, J. (2010) Recurring and non-recurring congestion: causes, impacts, and solutions. Working paper https://www.uc.edu/cdc/niehoff_studio/programs/great_streets/w10/reports/recurring_non-recurring.pdf

Pardillo, J. y Sánchez V. (2015) **Apuntes de Ingeniería de Tránsito**. ETS Ingenieros de Caminos, Canales y Puertos, Madrid, España

Skabardonis A., Varaiya, P. y Petty, K. (2003) Measuring recurrent and non-recurrent traffic congestion. **Transportation Research Record, Journal of the Transportation Research Board**, 1856, 60-68.

U.S. Department of Transportation (2004) Archived data management systems - A cross-cutting study. Publication FHWA-JPO-05-044. FHWA, U.S. Department of Transportation.

Yong-chuan, Z., Xiao-qing, Z., li-ting, Z. y Zhen-ting, C. (2011) Traffic congestion detection based on GPS floating-car data. **Proceedings of Engineering**, 15, 5541-5546.