CLOGGED ARTERIES: AN EMPIRICAL APPROACH FOR IDENTIFYING AND ADDRESSING LOCALIZED HIGHWAY CONGESTION BOTTLENECKS

Vivek Sakhrani, PhD, CPCS Transcom Inc. (USA)
Tufayel Chowdhury, CPCS Transcom Limited (Canada)

Introduction

Congestion bottlenecks are severe traffic chokepoints where demand far exceeds available highway capacity. The Federal Highway Administration (FHWA) of the US Department of Transportation states that recurring bottlenecks account for the largest share of road delay in the United States (40%), far exceeding traffic incidents (25%), inclement weather (15%), construction (10%) or other causes (Federal Highway Administration, 2012). Recurring bottlenecks in urban areas are the focus of this research.

Relieving congestion by removing bottlenecks requires addressing insufficient capacity. New construction can play an important long-term role, but this solution is capital intensive. Limited resources and immediate demands often require solutions centered on maximizing the efficiency of existing infrastructure. Optimizing the use of existing assets through a variety of approaches can improve performance, i.e. avoid excess fuel consumption, save lost time, and reduce greenhouse gases and other emissions. An important first step in addressing bottlenecks is to identify their precise locations and estimate their impacts along these performance dimensions. This research investigates and demonstrates how to precisely identify urban bottlenecks using empirical vehicle speed data and highway characteristics.

In this study, we develop a screening and prioritization framework for empirically identifying localized highway congestion bottlenecks using observed speeds from vehicle GPS probe data. We apply the framework to a large vehicle speed dataset comprising over 350,000 highway segments across the United States. The five-step procedure includes (i) time-based data sampling and the development of an idealized speed profile for highway segments, (ii) network conflation, (iii) pinpointing congestion bottlenecks, (iv) adjacency analysis for capturing the full length of the congested zone and associated costs, and (v) estimating the potential benefits of alleviating bottlenecks. Spatial visualization techniques succinctly depict the results. We estimate both the multi-dimensional costs of these bottlenecks, as well as the potential benefits of alleviating them.

Definition, Drivers and Impacts of Congestion

Conceptually, congestion is a mismatch between highway capacity and demand. Drivers are forced to reduce speed to accommodate a larger number of vehicles when demand exceeds capacity. Highway design features such as merging lanes, ramps, and reduced visibility around curves also contribute to congestion as they cause drivers to quickly decrease speed. Weather, visual distractions, accidents, construction and maintenance, and special events may further affect the smooth flow of vehicles. In most cases, these factors do not operate in isolation; a number of them interact to exacerbate congestion (Federal Highway Administration, 2015).

There is no single, universal definition of highway traffic congestion. The OECD has proposed both an operational definition of congestion: "*the impedance vehicles impose on each other, due to the speed-flow relationship, in conditions where the use of a transport system approaches capacity*," and from a user's perspective: "*a relative phenomenon that is linked to the difference between the roadway system performance that users expect and how the system actually performs.*" The former definition is about the efficiency of highway capacity use, and is based on objective and empirical measurements of the

relationship between traffic speed and flow. The latter is related to individual experience and expectations and is more subjective in nature. Both the objective and user experience viewpoints are relevant for policy in developing solutions and managing expectations.

Congestion causes "delays", i.e. increases the time it takes to get from point A to B, and results in lost productivity. FHWA estimates that hours of delay per traveler have more than doubled in cities of all sizes since 1982. Other calculated that congestion caused drivers in the US to spend an extra 6.9 billion hours on travel in 2014 Texas A&M Transportation Institute and INRIX, 2015). The lost time impacts both quality of life for individuals and the overall economy. Drivers give up productive work hours, and goods movement becomes more costly. Along with delays, congestion can potentially increase fuel consumption depending on speed conditions and also increase greenhouse gas emissions (Barth & Boriboosomsin, 2009).

Given that vehicle-miles traveled (VMT) are expected to grow at 1% annually over the next 20 years for light vehicles, and 2% per year for trucks over the same period, policymakers are concerned about identifying and addressing congestion hotspots (FHWA, 2015). Reducing congestion could (among other things) reduce capital costs, freeing up public resources for other uses. Users care about congestion (among other reasons) because it potentially limits their access to employment, goods and services or consumes some of their time that could be spent in more productive or enjoyable activities.

Measuring Congestion

Congestion measures should align with the objectives for addressing congestion. Following the definitions above, "pure" operational measures such as peak hour travel speeds (relative to a baseline speed) and peak hour flows maybe be useful for quantifying the severity of traffic congestion. This raises the issues of temporal and spatial scales. Are we concerned with only peak hours? Are we concerned with the duration of peak hours? Are link level observations, i.e. highway segments, for the purpose identifying bottlenecks most important? Are observations aggregated to larger spatial scales more important? How do we handle the aggregation? From a user perspective, if the opportunity cost of time is the primary reason why users are concerned, perhaps total hours of delay is the best measure. This in turn raises a host of other practical problems, such as determining the appropriate baseline for delay calculations and (if calculating the cost of congestion) monetizing the delay. We briefly discuss the most common congestion metrics with these issues in mind.

Level of Service (LOS): A traditional approach to measuring congestion is the level-of-service (LOS) as defined in the Transportation Research Board's Highway Capacity Manual (HCM). The LOS provides "scores" to a road ranging from A (best LOS) to F (worst LOS). Typically, the estimation of LOS is dictated by volume-capacity ratio and modelled speed which can be projected by traffic forecasting. However, LOS is not scalable to all roads in a metropolitan area because it is defined for certain highway classes. In addition, non-recurring congestion or travel time reliability cannot be measured based on volume and capacity without observed speed information (Transportation Research Board, 2010).

Travel Time Index (tti): This congestion metric is based on speed. In the annual *Urban Mobility Scorecard,* the Texas Transportation Institute uses *tti*, defined as the ratio of observed travel time to a baseline travel time. This metric does not isolate recurring and non-recurring congestion. Further, the use of free-flow conditions as a baseline, while practical and familiar, has been questioned by practitioners who are interested in the level of "excess" congestion (Texas A&M Transportation Institute and INRIX, 2015).

Total Delay: Another simple measure of congestion is total delay. Typically, an average daily delay per vehicle is estimated and then multiplied by both Annual Average Daily Traffic and an annual recurrence factor (number of weekdays, for instance). The delay is calculated as the difference between observed

travel time and a baseline travel time. Congestion reports usually identify and rank bottlenecks based on total annual delays. The navigation company TomTom also provides a similar ranking, but at the city level. It aggregates the average delay per link across highways, arterials and local roads and the cities are ranked based on the percent of congested roads.

Buffer Index (BI) and Planning Time Index (PTI): While TTI and delay estimates provide the average level of congestion, from a user's perspective, it is also important to measure congestion variability or reliability. Typical reliability measures are Buffer Index (BI) and Planning Time Index (PTI). The BI represents how much additional time a traveller should add to their average travel time in order to reach a destination on time. The PTI measures the total travel time a traveller plan for on-time arrival. PTI is the sum of average travel time and the buffer time. Studies often use the 95th percentile travel time for BI which means that a traveller would be late one weekday per month (Lyman & Bertini, 2008; Pu, 2011).

A challenge with congestion measurement is factoring in behavioural responses to hypothetical investments or policies designed to reduce delay. Several studies such as Urban Mobility Scorecard cite yearly hours of delay per commuter as one measure of congestion. Leaving aside the choice of the baseline for travel speeds (free-flow or "optimal"), we are still left without a clear understanding of what the user response to investments or policies that are designed to reduce or eliminate that delay would be. For example, if road or transit capacity were expanded to increase peak hour travel speeds (thereby reducing total delay), some users would change their travel behaviour through residential location, employment and other choices. If so, a new equilibrium with higher total traffic volumes but the same level of average delay may be reached. One might argue that even if many users ultimately "trade in" their time savings for other things (such as larger homes, better jobs, etc.), measuring delay and delay costs is still meaningful because users would only trade in those time savings for goods and services that are of greater value. However, if in practice the alternatives are expected to achieve other outcomes (such as improving access), we might be as interested in direct measurements of those outcomes? Emerging research has sought to develop "accessibility measures" as alternatives to traditional congestion measures. For instance, these measures take advantage of GIS tools to develop isochronic-based indices, which indicate the number of opportunities that are reachable within a given travel time. These measures attempt to directly address the reasons why we care about congestion. However, they are also more data intensive and complex. That said, if complexity is a barrier to implementation, the concept of access remains useful to explain for practitioners and the general public to put traditional congestion measures into context.

Data Needs for Congestion Estimation

Congestion indices are generally estimated using three types of data: volume and capacity; speed from loop detectors; and, speed from GPS probes. Volume data are collected by local, provincial and state governments through traffic counting stations. The volume is typically expressed in Annual Average Daily Traffic (AADT) which is estimated by applying some seasonal adjustment factors to the count data. AADT is allocated throughout the road network by each link based on the locations of the counters and the characteristics of road links between the counters. In some databases, certain lengths of adjacent links are assumed to carry similar AADT, thus are allocated the same volume. In the US, the State Departments of Transportation (DOTs) are mandated to report AADT, along with some other information to the Federal Highway Administration (FHWA) through a program called the Highway Performance Monitoring System (HPMS). AADT information is publicly available in HPMS GIS Shapefiles.

The same databases also contain information related to road capacity, such as number of lanes, with which one can estimate link capacity using procedures described in the HCM. Link volume and capacity are the most commonly available and traditional inputs in estimating traffic congestion measures. While volume and capacity data are quite reliable, they cannot be used to measure more robust, speed-based indices of congestion.

Loop detectors can identify a vehicle and its speed at a particular location where the loop system is installed in a road. The loop technology identifies a vehicle and its speed, based on the type and duration of magnetic disruption the metal surface of a vehicle causes to the loop's electric field. The data is transferred to a central database which is programmed to produce outputs such as traffic count and speed. While loop detectors are largely reliable source of speed data, especially the dual loops, Washington DOT reports that around 8% of the freeway loops provide erroneous data at any given time (Bremer, Cotton, Cotey, Prestrud & Westby, 2004). This requires careful quality control measures. A drawback of loop installations is that they provide counts only for their specific locations. Thus, the traffic patterns and congestion between the locations where loops are not installed need to be estimated based on assumptions.

GPS probe data address this limitation. In its raw form, a GPS dataset contains an anonymous vehicle ID, a time stamp of a GPS ping (i.e. recorded signal), the geographic coordinates (latitude and longitude) of the ping and an instantaneous speed called the "spot speed". In North America, there are several providers of such GPS data, notably HERE, TomTom, INRIX, ATRI and Shaw Tracking. Most datasets contain spot speeds for each ping which are allocated to a road network by GIS map matching whereby an average spot speed of all pings during a time interval is allocated to the nearest road link. The data provided by Shaw Tracking does not have spot speed, thus, speed needs to be imputed based on distance and time between two pings.

Research Approach

Many authors have published model-based analyses of highway congestion in urban areas. These studies often rely on statistical relationships for estimating congestion, which may not suit specific local conditions. Many studies also develop congestion estimates for an urban area as a whole. The area-wide modeling approach does not allow decision makers to target precise locations for appropriate interventions. The objective of our research was to identify and rank the top bottlenecks in the US freeway system to enable the development of targeted interventions. The study measured, for each bottleneck, the annual hours of delays, cost associated to delays and other congestion-related externalities, such as, fuel loss, CO2 emissions and accidents. The following procedures document our analysis:

(i) *Time-based Sampling and Idealized Speed Profiles*: The study used preprocessed GPS speed data allocated to the highway network in the US. The data was collected by HERE and obtained through National Performance Management Research Dataset (NPMRDS). The GPS probe data contains an instantaneous speed, called the "spot speed" for each ping. During the preprocessing the spot speeds are allocated to the nearest road segment and the average speed is taken for a 5 minute time period for each day. For this study, an average hourly speed was estimated from the 5 minute bin file for 24 hours of a typical weekday. The typical weekday hourly speeds were estimated from 40 representative weekdays throughout the year of 2014.

(ii) *Network Conflation*: Network conflation is a process of transferring data from a network to another. Generally, two networks, albeit similar, don't share exactly the same spatial alignment. The GPS speed from HERE was allocated to a network called Traffic Messaging Channel (TMC) during preprocessing. However, the TMC network does not have traffic volume data which is necessary to estimate total delays across all vehicles (average GPS speed × traffic volume). The volume information is available in a different network, called HPMS (Highway Performance Management System). A network conflation between TMC and HPMS was necessary to either transfer the speed or volume from one to the other. It seemed logical to transfer GPS speed from TMC to HPMS because the former is a directional network with parallel lines and the latter is non-directional with one line. Doing a HPMS-to-TMC conflation would require some assumptions on how the volume from HPMS would split by direction in TMC. A

safer, less erroneous approach was opted instead, involving TMC-to-HPMS conflation and the average of directional speeds were brought into non-directional HPMS. The process is illustrated in Figure 1.

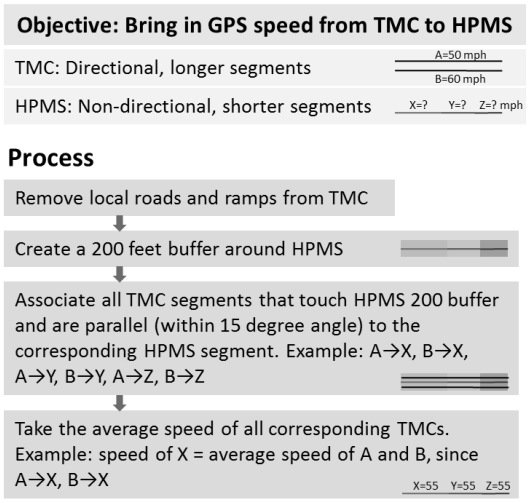

Figure 1. Network Conflation Process

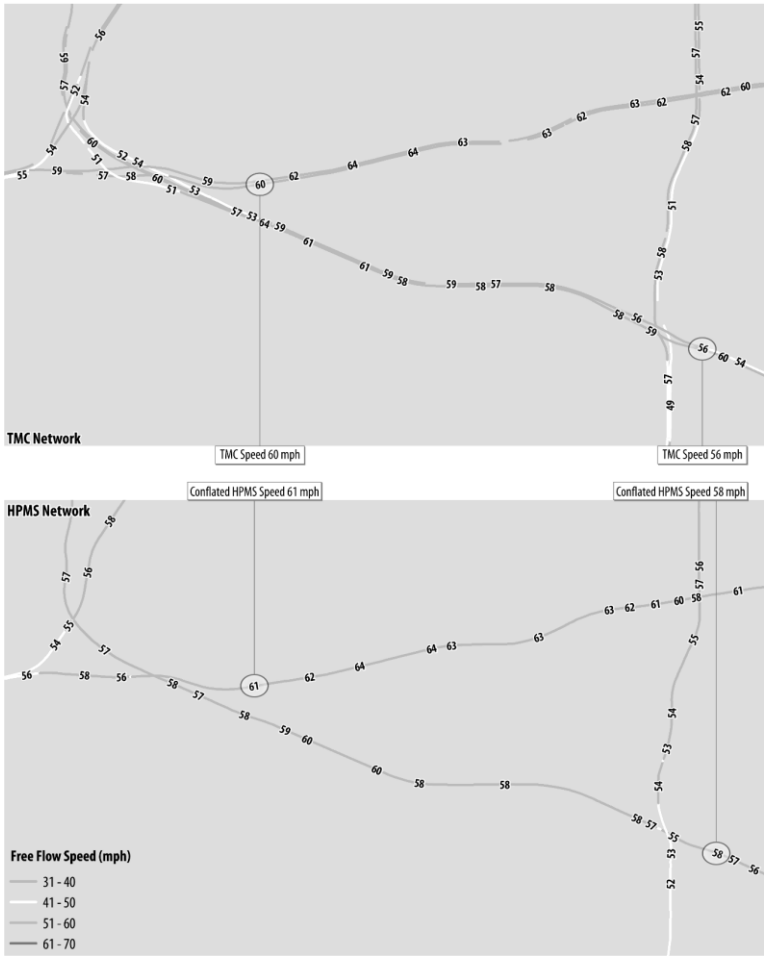

Figure 2. Network Conflation Result: before conflation (upper) and after conflation (lower)

Visualization in GIS was an integral part during the conflation process, especially deciding on the buffer distance and defining what road segments were parallel, thus relevant for conflation. Also, the before-after map validated the accuracy of the process.

(iii) Pinpointing Congestion Bottlenecks: After conflation, we calculated length-normalized hour-indexed delays (hours per mile) for every urban freeway segment *i* using this relationship

$$delay\ (d_{ji}) = [Vehicles\ per\ hour]_{ji} * (1/[Observed\ Speed]_{ji} - 1/[Baseline\ Speed]_{i})$$

Where,

Vehicles per hour is the hourly volume estimated.

Observed Speed is the weekday profile speed for every hour j in the day, as calculated above;

Baseline Speed is the Maximum Throughput Speed (MTS) for that freeway segment *i*, a counterfactual speed based on ideal travel conditions, developed using relationships published in the Transportation Research Board's (TRB) Highway Capacity Manual .

$$MTS_i = 39 + 0.2 * FFS_i$$

Where, FFS is the free flow speed as estimated by 95th percentile of calculated weekday hourly speeds.

NOTE: The relationship above holds ONLY in weekday hours when the observed speed is lower than the Maximum Throughput Speed (MTS), i.e. drivers experience slow down due to congestion. The following explains the process of congestion build up in relation to the Maximum Throughput Speed. When observed speeds exceeds MTS, delays in those hours -> 0. Although the MTS baseline is lower than the design Free Flow Speed, it represents a better use of available freeway capacity and is therefore a better reference point for estimating delays due to congestion. Assuming a constant volume of vehicles in this range of speeds, using the MTS as a baseline also gives us a more conservative estimate of congestion. In other words, we most likely underestimate hourly delays.

(iv) Adjacency Analysis: After identifying heavily congested segments, the question of "what is a bottleneck?" became a critical question because HPMS segments are highly variable in length, ranging from 50 ft to several miles, with an average length of around 500 ft. A typical bottleneck would consist of several such adjacent segments. Previous studies have either assumed a bottleneck to be 5 miles long and restructured the freeway network to consist of 5 mi long segments or used delay per mile for each freeway corridor (Eisele, Schrank, & Lomax, 2011). A corridor was defined as a group of adjacent freeway segments that experienced at least 4 hours of delay per week. The minimum length of a corridor was set to be at least 3 miles. Our process did not make any ad hoc assumptions on bottleneck length. However, a delay threshold needed to be set, otherwise most urban freeways would have been part of a bottleneck, since they all experience some delays during a weekday. A cut-off of 3000 hours/mile of daily delay was chosen based on the distribution of daily delay across all freeway segments with non-zero delays. The chosen cut-off represents the 99.7th percentile, meaning the top 0.3% of congested freeway segments qualified for national bottleneck ranking. If two bottlenecks were located within 0.5 mile of each other, they, along with the segments in between were considered as being part of one bottleneck.

After the bottlenecks were identified and mapped, some anomalies appeared (see Figure 3). They were deemed "false positives" rather than actual bottlenecks because they were located far away from the urban centers and there were no other bottlenecks nearby. In some cases, those false positives had more delays than the ones near urban areas or some of the known bottlenecks (based on consultation with various local parties, including State DOTs). The visual examination was critical to validate and weed out those anomalies.

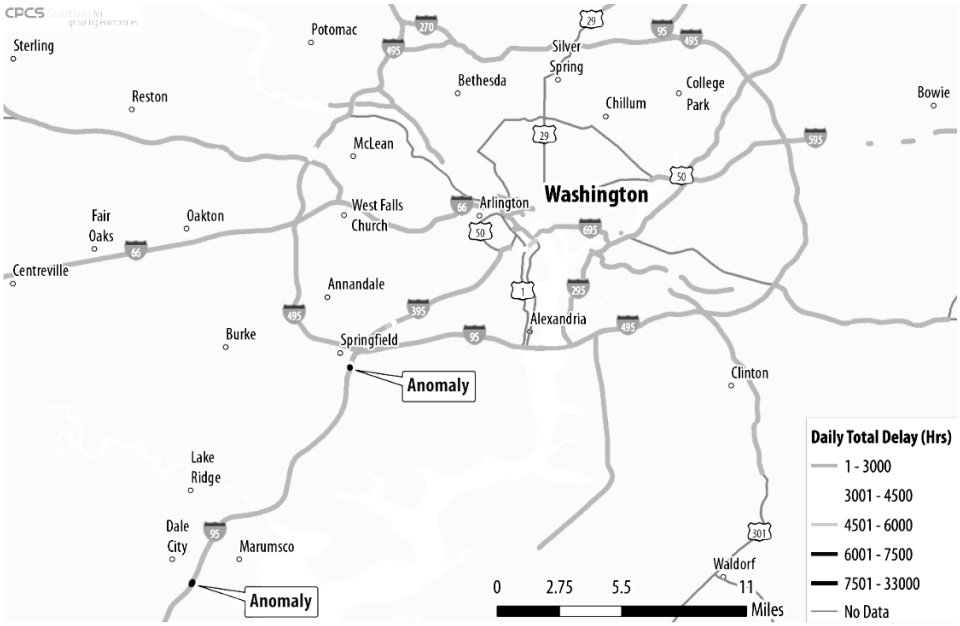

Figure 3. Anomalies in bottleneck identification

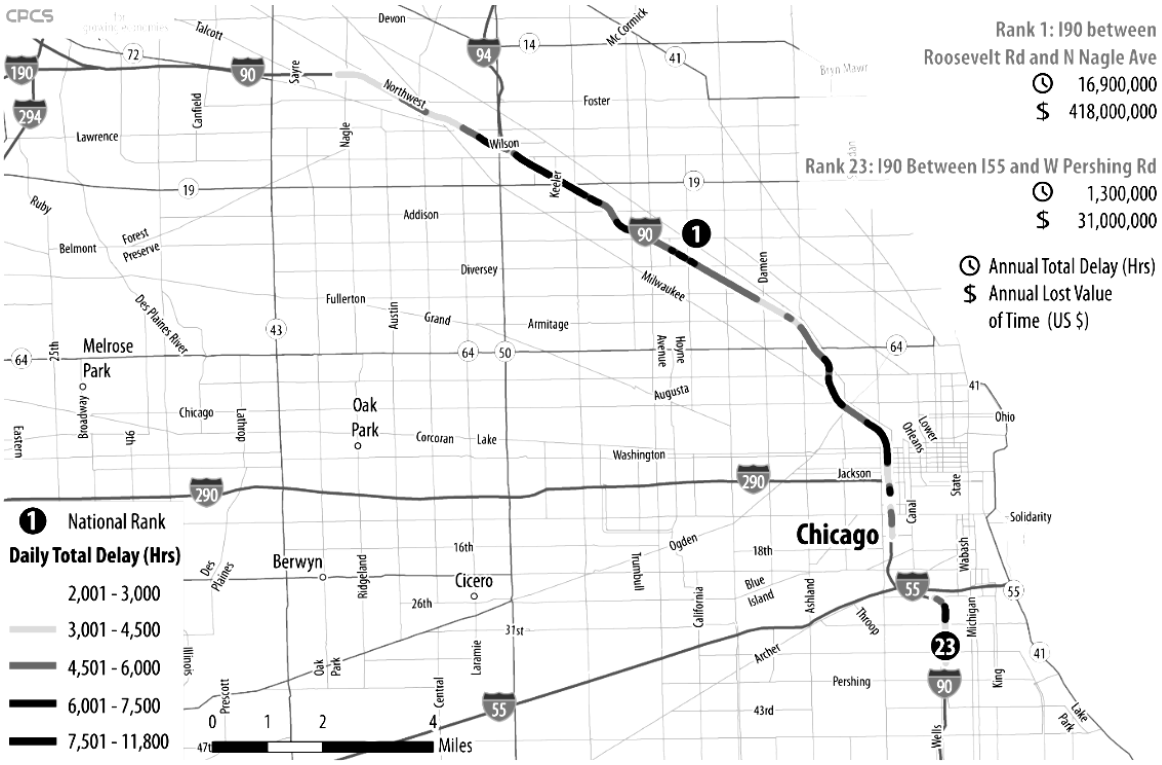

Figure 4. The top-ranked bottleneck in the US in Chicago, based on 2014 data

"A picture is worth a thousand words" was no exaggeration for this study, particularly because the bottleneck location and length were precisely identified and mapped. Figure 4 shows the precise location and length of the top-ranked bottleneck in the US, in terms of total hours of delay annually. Other bottlenecks were also ranked similarly. The analysis involved several rounds of validation, with local

planning officials and relevant State DOTs. The maps with precise locations were immensely helpful for communication across various parties. More importantly, the visuals provided a clear picture of exact locations where interventions would be needed.

(v) Potential Benefits of Bottleneck Elimination: We valued each hour of delay using the state-specific estimate of the value of a volunteer hour (US $ / hour). This value is a weighted average of employment wage rates across many labor and skill sectors, and based on data collected by the US Bureau of Labor Statistics (BLS). This approach most likely underestimates the lost value of time. The lost value of time for the US's top ranked bottleneck is $418 million per year, as shown in Figure 4, and this is the estimate of available economic savings from eliminating that particular bottleneck.

We estimated the fuel wasted due to congestion and potential fuel savings (gallons) using relationships between vehicle speed (miles per hour, mph) and fuel economy (miles per gallon, mpg) (Oakridge National Laboratory, 2015). These relationships are based on lab tests as well as observed data from a large fleet of vehicles. Only the excess fuel used when vehicles are traveling at slow speeds during congested conditions are counted. We then calculated the potential emissions avoided (pounds CO2) using standard parameter values (US Environmental Protection Agency, 2014), with 8,887 grams CO2/ gallon of gasoline for cars and 10,180 grams CO2/ gallon of diesel for trucks. If the top ranked bottleneck in Chicago were to be eliminated, it would save 6.4 million gallons of fuel per year and almost 133 million pounds of CO2 annually.

Conclusions

In this research we have made the case for precisely identifying the location of highway congestion bottlenecks so as to enable targeted interventions at those locations. Our five step approach demonstrated the use of GPS probe data with vehicle spot speeds to identify bottlenecks. We also estimate the costs of congestion, and hence the available potential savings of bottleneck elimination along the dimensions of time value of money, fuel consumption and emissions reduction.

References

Barth, M., & Boriboosomsin, K. (2009). Traffic Congestion and Greenhouse Gases. *35*.

Bremer, D., Cotton, K., Cotey, D., Prestrud, C., & Westby, G. (2004). Measuring Congestion Based on Operatinal Data. *Transportation Research Record, 1895*, 188-196.

Eisele, B., Schrank, D., & Lomax, T. (2011). *Congestion Corridors Report.* Texas Transportation Institute.

Federal Highway Administration. (2012). *2012 Urban Congestion Trends*. Retrieved February 2016, from http://www.ops.fhwa.dot.gov/publications/fhwahop13016/index.htm

Federal Highway Administration. (2015). *Describing the Congestion Problem*. Retrieved February 2016, from https://www.fhwa.dot.gov/congestion/describing_problem.htm

FHWA. (2015, June). *Forecast of Vehicle-Miles Traveled*. Retrieved from https://www.fhwa.dot.gov/policyinformation/tables/vmt/vmt_forecast_sum.pdf

Lyman, K., & Bertini, R. (2008). Using travel time reliability measures to improve regional transportation planning and operations. *Transportation Research Record*, 1-10.

Oakridge National Laboratory. (2015). *Transportation Energy Data Book, Chapters 4 and 5.*

Pu, W. (2011). Analytic Relationships between Travel Time Reliability Measures. *Transportation Research Record, 2254*, 122-130.

Texas A&M Transportation Institute and INRIX. (2015). *2015 Urban Mobility Scorecard.*

Transportation Research Board. (2010). *Highway Capacity Manual.* Washington DC: National Academies.

US Environmental Protection Agency. (2014). *Light-Duty Automotive Technology, Carbon Dioxide Emissions, and Fuel Economy Trends: 1975 through 2014. EPA-420-R -14-023a.*