Development of Open Source Tool using General Transit Feed Specification (GTFS) for Transit Spatiotemporal Connectivity Analysis

S. Kiavash Fayyaz S., E.I.T.
Ph.D. Student, Graduate Research Assistant
Secretary of University of Utah ITE student chapter
Tel: (385) 234-0212, Email: kiavash.fayyaz@utah.edu

1) Department of Civil and Environmental Engineering, University of Utah
110 Central Campus Drive, Suite 2000B, Salt Lake City, UT 84112

Abstract

The social functions of urbanized areas are highly dependent on and supported by the convenient access of public transportation systems, particularly for the less privileged populations who do not have auto ownership. To evaluate the connectivity of the existing transit service and identify access gap of those disadvantaged population groups, it is critical to accurately estimate the travel times between transit stations which is changing throughout the day due to transit schedule variations. In recent years, General Transit Feed Specification (GTFS) data has been gaining popularity for between-station travel time estimation due to its capability to be incorporated into spatiotemporal analytics. Many software packages, such as ArcGIS, have developed toolbox to enable the travel time estimation with GTFS. They perform reasonably well in calculating Travel Time between All Stations (TTAS) for a specific time-of-day (e.g. 8:00 AM), yet become computational inefficient and unpractical with the increase of data dimension (e.g. all times of day travel time estimation). For example, a quad-core machine using ARCGIS toolbox will take approximately 60 days to process the calculation of TTAS for each minute of the day on a relatively small network of 1400 stations and 100 transit routes. To address this problem, we are developing an open-source tool, written in C++, to calculate TTAS for all times of day with enhanced computational efficiency. This stand-alone tool, utilizing object oriented programming and hash-map search efficiency, enables the calculation of TTAS with GTFS for any desirable times of day. The preliminary version of the tool have been applied to Saint George transit network in order to calculate Weighted Average Travel Time (WATT) as an accessibility measure. The results can be further used for public transit planning, visualization, and connectivity analysis.

KEYWORDS: GTFS; Public Transit; Travel Time; Accessibility; Planning
Contents
INTRODUCTION .................................................................................................................. 1
DATA ..................................................................................................................................... 3
METHODOLOGY .................................................................................................................. 5
  Weighted Average Travel Time (WATT) .............................................................................. 5
  Code Complexity .............................................................................................................. 5
RESULTS .............................................................................................................................. 6
RESULTS INTERPRETATION ............................................................................................... 8
CONCLUSION AND DISCUSSIONS .................................................................................... 9
ACKNOWLEDGMENTS ........................................................................................................ 10
REFERENCES ...................................................................................................................... 10
INTRODUCTION
Equitable access to public transport services is a cornerstone of a just society (Golub and Martens, 2014, Martens, 2009, and Martens et al., 2012). To achieve social equity, an efficient access to places must be provided by public transport services, particularly for the less privileged populations who do not have auto ownership (Farber et al., 2016). Transport planning traditionally has focused on improving mobility throughout the network. The increased mobility improved the auto-oriented transport modes, but neglected the other modes and their needs, particularly equal access by modes (Benenson et al., 2011, Golub and Martens, 2014, Kaplan et al., 2014, and Martens et al., 2012). As the results, the needs of transit-dependent population were overlooked in transit service planning avoiding them to participate in desirable activities (Martens et al., 2012 and Welch and Mishra, 2013, Delbosc and Currie, 2011 and Welch, 2013, Church et al., 2000, Hine, 2003, Lucas, 2012, Páez et al., 2009, Preston and Rajé, 2007, and Raje, 2004). This motivates researchers to develop public transit accessibility measures to identify inefficiencies in transit network (e.g. where supply and demand don’t match). Consequently, public transit accessibility measures guides decision making related to transportation investment and land use development (Coffel, 2012).

Accessibility is defined as the ease of travel for an individual to reach a desired destination. By this definition accessibility consists of two elements: an activity element and a transportation element (Burns, 1980, and Koenig, 1980). The activity element reflects the potential opportunities available at destination and usually measured by population density, job density, and/or facilities available at destination. In public transport context, the transportation element of accessibility reflects the ease of travel and is affected by spatial coverage, temporal coverage, cost of travel (e.g. travel time), and comfort of service. The transit accessibility measure can be divided into two categories: travel time discretionary measures and travel time dependent measures. Travel time discretionary measures such as local index of accessibility (Rood, 1998), transit capacity and quality of service (Kittelison, 2003) method, transit level-of-service (Ryus et al., 2000, and Tumlin et al., 2005), and time of day (Polzin et al., 2002) method consider service coverage (spatially or temporally), service frequency, vehicle capacity, and comfort of service. The main shortcoming of travel time discretionary measures are, as their name indicates, neglecting the travel time (or any other cost function) between origin and destination. Travel time is one of the major factors impacting the ease and viability of transit use. Thus, overlooking travel time tend to overestimate the proportion of population with transit access (Polzin et al. 2002).

Several travel time dependent accessibility measures have been developed up to date including competition measures (Van Wee et al., 2001, Joseph et al. 1982, and Scheurer et al. 2007), constraints-based measures (Geurs et al., 2004, Bhat et al., 2000, and Scheurer et al. 2007), and composite measures (Miller, 1999). Still the cumulative and gravity-based measures are the most widely used methods to measure accessibility. Cumulative measures are based on number of potential opportunities can be reached within cost (e.g. travel time) threshold (Vickerman, 1974, Wachs et al., 1973, Geurs et al., 2001, and Bhat et al., 2000).
\[ A_i = \sum_{j=1}^{J} B_j \ast a_j \] (1)

Where \( A_i \) is cumulative accessibility measure at point \( i \), \( a_j \) is the potential opportunities at point \( j \), and \( B_j \) is a binary value equal to 1 if point \( j \) is within the predetermined travel time threshold and 0 otherwise. This measure doesn’t account for the impedance (e.g. travel time) of reaching the destination by public transit. In other words, if the travel time to desired destination is slightly longer than predetermined threshold, then this destination is not considered in the accessibility measure calculation.

Gravity-based measures overcomes the shortcoming of cumulative accessibility by weighting the number of potential opportunities that can be reached based on cost function (e.g. travel time) (Hansen, 1959, Geurs et al., 2001, Bhat et al., 2000, and Bhat et al., 2006).

\[ A_i = \sum_{j=1}^{J} O_j \ast f(C_{ij}) \] (2)

Where \( A_i \) is the gravity-based accessibility measure at point \( i \), \( O_j \) is the potential opportunities at point \( j \), and \( f(C_{ij}) \) is the impedance or cost function (e.g. travel time) for travelling between \( i \) and \( j \) by public transport. Major disadvantage of this method is the need of developing an impedance factor between all origin-destinations pairs and estimating the number of potential opportunities at each point (El-Geneidy et al., 2006). This method accounts for spatial coverage, service frequency, attractiveness of destination, and travel time between origin and destinations. Measuring the accessibility for all times of day will enable the method to also account for temporal coverage and fluctuation. Thus, gravity-based accessibility measure for all time of day is the most comprehensive method for calculating public transit accessibility.

Previous studies in public transit accessibility have widely used cumulative and gravity-based accessibility measures to evaluate the public transit network and services (Farber et al., 2014, Foth et al., 2013, Lei and Church, 2010, and O’Sullivan et al., 2000). The major drawback of these studies, as Farber (2016) mentioned, is ignoring the temporal fluctuation in travel time throughout the day due to transit schedule variations. The temporal fluctuation in travel time leads to temporal fluctuation in accessibility throughout the day. Ignoring the fluctuation in accessibility throughout the may lead to over/under-estimation in transit service evaluation. Farber (2016) tried to address this problem by calculating TTAS for each minute of the day. They reported the calculation for Salt Lake City network with 1400 stations, and 100 transit routes in ARCGIS will take about 60 days on a quad-core machine. Hence it is challenging if possible to calculate the gravity-based accessibility measures and its temporal fluctuation. In addition, visualizing and analyzing an accessibility measure with temporal fluctuation is still in preliminary stages, resulting for the need for new analysis methods that can calculate and quantitatively compare the temporal accessibility measure. To fill this gap, this paper introduces an efficient and innovative tool to calculate and evaluate gravity-based accessibility measure for all time of day for transit stations. The contributions of this paper are twofold. First, I developed open-source tool, written in C++, to calculate TTAS for all times of day with enhanced computational efficiency. The tool will take advantage of open-source publicly available datasets (i.e. GTFS and census data) to calculate the...
WATT (a gravity-based accessibility measure described in methodology section). Second, I try to analyze and visualize the accessibility and its temporal fluctuation.

Next I provide a description of data from SUNTRAN’s network. The accessibility calculation method is then described, followed by the analysis and visualization of the results. Finally, the implication are discussed.

DATA
As mentioned before, this study uses publicly available dataset including census data and GTFS for city of St. George, Utah. St. George is small city in southern Utah with 76,817 population. Census block data for state of Utah for 2015 have been collected from Utah Automated Geographic Reference Center (AGRC) website. The census block is the smallest geographic unit used by United States Census Bureau for tabulation of data collected from all houses. The high resolution of census block data increase the accuracy of the accessibility measure.

GTFS was created in 2005 by Google and TriMet for transit agencies to represent their schedules, trips, routes, and stops data in an open-source format that is usable for Google Transit Web-based trip planner. GTFS has evolved since 2005 regarding agencies and developers feedback. Currently many cities and agencies participate in sharing their GTFS data (“google transit data feed,” 2016). A GTFS dataset consists of several plain text files which been formatted as Comma-separated Values (CSV) and are contained within a single zip file, which is hosted on transit agency’s website and accessible for public. The main files in GTFS for this study are Stops, Routes, Trips, Stop-Times, and Calendar. Figure 1 shows three of these files for St. George transit network in their original format. These files contain a very detail information about transit schedule for each minute of each day of a week.

FIGURE 1 Stop-Times, stops, and trips files of GTFS for St. George, UT
The research identified examples of service evaluation metrics that can be determined using only GTFS data (Wong, 2013, Hillsman et al., 2011, and Catala, 2011) include service area calculation, service coverage, time and distance service calculation, stop location and spacing optimization, service frequency, and span of service. The strength of GTFS data are better understood when combined with other datasets. GTFS combination with Automatic Passenger Count (APC) data can provide several important transit performance measures including ridership by hour, trip, and stop, trip activity ranking, stop activity ranking, and activity by period. GTFS combination with census data can be used to provide key information for planners. Several researchers have used this combination for analyzing spatial accessibility metrics such as number of job accessible by transit in a time limit (e.g. 30 minutes) and population accessible by transit in a time limit.

The GTFS data for St. George transit network operated by SUNTRAN were collected from GTFS data exchange website (“GTFS Data Exchange,” 2016). SUNTRAN’s transit network consists of six bus routes and 134 bus stops. Four of these routes operate at 40 minutes fixed headway and the other two operate at 80 minutes fixed headway. Figure 2 shows SUNTRAN’s transit network map.

FIGURE 2 SUNTRAN’s transit network map; 1: “Taucahn” Station, 2: “Sunset Corner” Station
METHODOLOGY

The methodology part is separated in two sections: WATT and code complexity. WATT section will define the accessibility measure used in this study. The code complexity section is describing the algorithm developed and its impact on enhancing the computation efficiency.

Weighted Average Travel Time (WATT)

According to (Cao et al. 2013), the WATT between stations can be described as:

\[
WATT_i = \frac{\sum_{j=1}^{J} M_j \times tt_{ij}}{\sum_{j=1}^{J} M_j} \quad j = 1,2,...,J
\]  

where \(WATT_i\) is the weighted average travel time of station \(i\), also referred to as location indicator (Gutiérrez et al., 1996 and Gutiérrez, 2001). \(M_j\) is the population in the 700 meter radius of the station \(j\), \(tt_{ij}\) is the travel time (including egress, ingress, and transfer time) using public transit from station \(i\) to station \(j\), and \(J\) is the total number of stations in transit network. The WATT is in nature a gravity-based accessibility method, meaning it is based on gravity-like interaction pattern assumption between locations (Geertman et al., 1995). In other words, increase in population (gravity) and decrease in travel time (distance) will increase the accessibility (gravity force) between two stations (masses).

WATT accounts for spatial coverage, temporal coverage, travel time between stations, and attractiveness of different destinations. In addition, WATT unlike other accessibility measures, is a measure of time which makes it understandable and tangible. For example, \(WATT_i = 60\) minutes, means the average travel time from station 1 to all other stations regarding their attractiveness is 60 minutes.

Code Complexity

Shortest path (SP) is the most commonly used approach in calculating the travel time between different nodes of the network. The first step in SP is to build the shortest path graph, then travel time between stations can be calculated from the SP graph. However, in transit network SP graph is unique for each time day which makes the computation time intensive. The complexity of SP algorithm is \(O(V^2)\), where \(V\) is the number of vertices (in this study vertices are the stations) in SP graph (Dijkstra, 1959). Thus the complexity of algorithm that calculates travel time for each minute of a day (24*60) will be 1440 * \(O(V^2)\). The complexity (computation time) will increase by the second power of number of stations, meaning for large networks the algorithm will have large computation time. To overcome this problem, this study developed a new algorithm that uses public transit networks constraints to decrease the complexity (computation time).

The developed algorithm will not build the SP graph, instead it will start from each station and follow the next trips passing through this stations and all transfers station. Then update the travel time read from stop-times file in GTFS for the passed stations. Assuming that each transit user will not take more than four different routes to reach the desired destination, the algorithm will break after four transfer and repeat the process for next station (figure 3). The complexity of this algorithm is \(O(V * TS^3)\), where TS is the average number of transfer station connected to a station directly or with one transfer. In typical public transport network the TS value is usually smaller than 4. Thus, as network size grows the computation time for this algorithm is less than SP...
algorithm. In addition, the developed algorithm code is written in C++ which will compute the same task much faster than software-implemented tools (ARGIS GTFS tool), because C++ is a lower-level programming language.

![Algorithm Diagram]

**FIGURE 3 Travel time calculation algorithm**

**RESULTS**
The tool was implemented to SUNTRAN’s network, and WATT were calculated for all station for each 5 minute interval throughout the day (figure 4). The computation time was about 3.5 minutes. In order to better visualize the results WATT were plotted regarding the population in 700 meter radius of each station (figure 5). The error bars in figure 4 and 5 show the temporal fluctuation in WATT throughout the day.

In order to better show the temporal fluctuation in WATT (figure 6 and 7), two stations were selected including “Tuacahn” station and “Sunset Corner” station with highest and lowest WATT value respectively. “Tuacahn” station is located in recreational area with the population of 20 people living in 700 meter radius of it. Bus route 5 is only route serving the station that operates on 80 minutes headway. On the other hand, “Sunset Corner” station is located close to shopping centers and residential areas with the population of 1600 people living in the 700 meter radius of it. Bus routes 3, 4, 5, and 6 are serving this station.
The sequence of station IDs are usually based on the routes. Thus, most of the stations with close station ID value are on the same route.

FIGURE 5 WATT plotted regarding the station population living in 700 meter radius
RESULTS INTERPRETATION

The computation time for St. George network was about 3:30 second. Considering the complexity of the new algorithm computation time for a network with 1400 station will be around two hours. The developed algorithm will decrease the computation time for network with 1400 stations by 1,438 hours.

Figure 4 shows the general accessibility (WATT) for each station. The results shows station with WATT value above 300 minutes have the worst accessibilities in the network. In addition figure
4 shows that many stations with high or low accessibility have close station ID, this is due to the fact that station IDs are assigned sequentially to stations on the same route in St. George network.

A station that have low accessibility value and low population living in its vicinity is not concerning. The problem arises at stations with high accessibility and low population or at stations with low accessibility and high population. In the first case, the supply is more than demand leading to network inefficiencies and waste of resources. In the second case, the demand is more than supply leading to customer dissatisfaction and violate equitable access. This two cases are visualized bet by figure 5. The stations that have excess demand are highlighted with the red circle in figure 5. These stations are the first priority in improving the transit network. The stations with excess supply are highlighted with the green circle. These stations must be studied closely and their existence must be justified. If they do not exist to serve justifiable purpose (e.g. as transfer station), they can removed from the network to prevent the waste of resources. In order to better capture the equity issue, the socioeconomic characteristics of the population around the stations can be considered to better filter the results. Unfortunately, in this study area the differences between socioeconomic characteristic of the population around different stations including age, gender, and average salary were small and negligible.

The “Tuacahn” station WATT graph shows the temporal fluctuation of accessibility throughout the day. As figure 1 shows the station is placed in a remote area, and there is only one route serving it. Thus the accessibility of the station maximize when the bus is at the station and right after it will reach its minimum point. Then it will gradually increase as approaching arrival time of next bus. The time interval between the two consecutive maximum accessibility moments are equal to 80 minutes (operating headway of serving route). As the bus is in a remote area the temporal fluctuation in accessibility is relatively small. On the other hand, “Sunset Corner” station is placed closely to city center and served by four transit routes. Figure 7 shows that the time interval between the two consecutive maximum accessibility moments for this station varies from 15 to 30 minutes. This differences are dependent to the coordination between the four routes serving the station. Figure 7 shows that the coordination between the four routes at “Sunset Corner” station is at worst condition around 9:00 AM. This is due to the unavailability of service from route 5 at that hour for the station.

CONCLUSION AND DISCUSSIONS
Rather than using a accessibility measure for specific time of day, this study has developed a new tool to analyze accessibility and its temporal fluctuation. The method will not only consider the spatial coverage, attractiveness of stations, and travel time between stations but also analyze the temporal coverage and service inefficiencies. To overcome the computational inefficiencies of this method, this study introduced a new tool that decrease the computational time from days to hours. The study introduce a new visualizing method for accessibility of the stations. The comparative visualization (that include all stations on one plot) is powerful method to identify the stations with excessive supply or demand. From this perspective, the method will provide a performance measure for macro-level analysis. The station visualization can be used to analyze temporal fluctuation of accessibility throughout the day. In addition, the station visualization can show inefficiencies in bus coordination. In this setting, the method can be used to analyze the network
in micro-level. Thus, the WATT calculated for all time of day can provide both a macro-level and micro-level performance measure for public transit network.

Future expansion of this study will focus on incorporating socioeconomic characteristic of suited population and implementing the method to larger network. For example, incorporating the average income of the population living in the vicinity of the station can be used as the filtering measure for prioritizing stations need for development. Implementing the method to larger network with more routes with fixed and variable headways and more stations will provide better insights regarding the network coordination.

ACKNOWLEDGEMENT
The author would like to thank Dr. Cathy Liu for her support and feedback on this study.

REFERENCES