

# Social-Transportation Analytic Toolbox (STAT) for Transit Networks

Xiaoyue Cathy Liu, Ph.D.

Ran Wei, Ph.D.

Aaron Golub, Ph.D.

Liming Wang, Ph.D.



# **Social-Transportation Analytic Toolbox (STAT) for Transit Networks**

**Final Report**

**NITC-RR-1080**

by

Xiaoyue Cathy Liu (PI)  
Qian Zuo  
Shenruoyang Na  
University of Utah

Ran Wei  
University of California, Riverside

Aaron Golub  
Liming Wang  
Jake Davis  
Portland State University

for

National Institute for Transportation and Communities (NITC)  
P.O. Box 751  
Portland, OR 97207



**July 2019**

## Technical Report Documentation Page

|  |  |  |  |   |           |
|--|--|--|--|---|-----------|
| 1. Report No.<br>NITC-RR-1080  |  | 2. Government Accession No.                                    |  | 3. Recipient's Catalog No.                            |           |
| 4. Title and Subtitle<br>Social-Transportation Analytic Toolbox (STAT) for Transit Networks  |  |  |  | 5. Report Date<br>July 2019                           |           |
|  |  |  |  | 6. Performing Organization Code                       |           |
| 7. Author(s)<br>Xiaoyue Cathy Liu (PI), Ran Wei, Liming Wang, Aaron Golub  |  |  |  | 8. Performing Organization Report No.                 |           |
| 9. Performing Organization Name and Address<br>Department of Civil & Environmental Engineering<br>University of Utah<br>110 Central Camps Drive, Suite 2000<br>Salt Lake City, UT 84112  |  |  |  | 10. Work Unit No. (TRAIS)                             |           |
|  |  |  |  | 11. Contract or Grant No.????                         |           |
| 12. Sponsoring Agency Name and Address<br><br>National Institute for Transportation and Communities (NITC)<br>P.O. Box 751<br>Portland, OR 97207   |  |  |  | 13. Type of Report and Period Covered<br>Final report |           |
|  |  |  |  | 14. Sponsoring Agency Code                            |           |
| 15. Supplementary Notes  |  |  |  |   |           |
| 16. Abstract<br><p>This project builds an open-source, socio-transportation analytic (STAT) toolbox for public transit system planning in an effort to integrate social media and general transit feed specification (GTFS) data for transit agencies in evaluating and enhancing the performance of public transit systems. This toolbox is novel and essential to transit agencies in two aspects. First, it enables the integration, analysis and visualization of two major, new open transportation data, social media and GTFS data, to support transit decision-making. Second, it allows transit agencies to evaluate service network efficiency and access equity of transit systems in a cohesive manner, and identify areas for improvement to better achieve these multidimensional objectives.</p> <p>The toolbox employs a combination of data mining, geographical information systems and transportation network modeling. The STAT is an open-source toolbox and is publicly accessible. The project engages two transit agencies, the Utah Transit Authority (UTA) and TriMet, to test the usability of the toolbox, where Salt Lake City and Portland are used as case studies in the platform for querying, navigating and exploring the interactions between transit users and services. STAT can assist agencies in evaluating overall system performance and identifying existing public transit connectivity gaps, particularly for disadvantaged populations, in reaching essential services. It can also act as a decision support tool for recommending improvements (e.g., prioritize the stations and routes, identify the necessity for introducing a new line within existing infrastructure, etc.) The project ties to the NITC theme of improving mobility of people and creating vibrant communities. We expect that it can be adapted over time to cover different geographies and incorporate new data sources. In addition to serving transit agency staff, the tool can be used in university curriculum and by advocacy organizations engaged in transportation decision-making. Finally, the project lays the foundation for NITC developing other open-source tools using big data.</p> |  |  |  |   |           |
| 17. Key Words<br>Transit accessibility; twitter; performance assessment; GTFS; social media  |  |  | 18. Distribution Statement<br>No restrictions. Copies available from NITC:<br><a href="http://www.nitc-utc.net">www.nitc-utc.net</a> |   |           |
| 19. Security Classification (of this report)<br><br>Unclassified   |  | 20. Security Classification (of this page)<br><br>Unclassified |  | 21. No. of Pages<br><br>30                            | 22. Price |

## **ACKNOWLEDGEMENTS**

This project was funded by the National Institute for Transportation and Communities (NITC; grant number 1080) a U.S. DOT University Transportation Center. The project also benefitted from matches from the University of Utah, Portland State University and the University of California at Riverside. Furthermore, we acknowledge and thank the anonymous peer reviewers who provided helpful insights and corrections to the report, which is published in the Journal of Public Transport (Haghighi et al., 2018), PLoS One (Fayyaz et al., 2017), Transportation Research Part B Methodological (Zhuang et al., 2019), and Computers, Environment, and Urban Systems (Chen et al., 2019)

## **DISCLAIMER**

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the material and information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation University Transportation Centers Program in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the U.S. Government. This report does not constitute a standard, specification, or regulation.

## **RECOMMENDED CITATION**

Liu, Xiaoyue Cathy, Wei, Ran, Aaron Golub and Liming Wang. *Social-Transportation Analytic Toolbox (STAT) for Transit Networks*. NITC-RR-1080. Portland, OR: Transportation Research and Education Center (TREC), 2019.

## TABLE OF CONTENTS

|  |    |
|--|----|
| EXECUTIVE SUMMARY  | 1  |
| 1.0 INTRODUCTION   | 2  |
| 2.0 BACKGROUND   | 5  |
| 2.1 TRANSIT ACCESSIBILITY MEASURES   | 5  |
| 2.2 TWITTER DATA FOR URBAN TRANSPORTATION ANALYTICS                                | 6  |
| 3.0 METHODOLOGY  | 8  |
| 3.1 TRANSIT ACCESSIBILITY COMPUTATION  | 8  |
| 3.2 TWITTER DATA ACQUISITION AND PROCESSING  | 13 |
| 3.2.1 Classification Model to Filter Relevant Tweets                               | 13 |
| 3.2.2 Information Extraction Method  | 17 |
| 4.0 STAT VISUALIZATION AND ANALYTICS   | 19 |
| 4.1 STAT PLATFORM INTERFACE  | 19 |
| 4.2 WATT ACCESSIBILITY INFERENCE   | 22 |
| 4.3 INTERCONNECTED KNOWLEDGE DISCOVERY TO UNDERSTAND SUPPLY-<br>DEMAND INTERACTION | 25 |
| 5.0 CONCLUSIONS  | 26 |
| 6.0 REFERENCES   | 27 |

## LIST OF FIGURES

|   |    |
|---|----|
| Figure 3.1: Methodological framework  | 9  |
| Figure 3.2: Pseudocode for finding all-pairs shortest path and station WATT through a<br>day              | 11 |
| Figure 3.3: WATT calculation code runs in terminal and at GUI   | 12 |
| Figure 3.4: A tweet with complaint semantic   | 13 |
| Figure 3.5: Architecture of the neural network model  | 15 |
| Figure 3.6: Parameter matrix of the input layer of the model, with each row representing<br>a word vector | 16 |
| Figure 3.7: A CNN model for tweets classification   | 17 |
| Figure 3.8: A pos tagging and chunking example  | 17 |
| Figure 4.1: Use Chart.js to visualize WATT pattern for a transit stop                                     | 20 |
| Figure 4.2: Twitter comments visualization on Google Map API in STAT                                      | 21 |
| Figure 4.3: ArcGIS online visualizing the accessibility of each TAZ in Salt Lake City area                | 22 |
| Figure 4.4: WATT distribution across Portland metropolitan area   | 23 |
| Figure 4.5: AMWR distribution across Portland metropolitan area   | 23 |
| Figure 4.6: WATT temporal distribution of a sample transit stop   | 24 |
| Figure 4.7: Transit stop clustering result based on HAC   | 26 |

## EXECUTIVE SUMMARY

This project builds an open-source, socio-transportation analytic (STAT) toolbox for public transit system planning in an effort to integrate social media and general transit feed specification (GTFS) data for transit agencies in evaluating and enhancing the performance of public transit systems. This toolbox is novel and essential to transit agencies in two aspects. First, it enables the integration, analysis and visualization of two major, new open transportation data, social media and GTFS data, to support transit decision-making. Second, it allows transit agencies to evaluate service network efficiency and access equity of transit systems in a cohesive manner, and identify areas for improvement to better achieve these multidimensional objectives. Specifically, leveraging machine learning and natural language processing techniques, we retrieved Twitter data that are related to public transit systems and extracted sentence structures to geomap those tweets to their corresponding transit lines/stations. Combining with transit accessibility measures computed using GTFS, STAT is able to identify the mismatch between the services that the agency is providing versus what the transit users are experiencing.

The toolbox employs a combination of data mining, geographical information systems and transportation network modeling. The STAT is an open-source toolbox and is publicly accessible. The project engages two transit agencies, the Utah Transit Authority (UTA) and TriMet, to test the usability of the toolbox, where Salt Lake City and Portland are used as case studies in the platform to demonstrate how to use STAT for querying, navigating and exploring the interactions between transit users and services. STAT can assist agencies in evaluating overall system performance and identifying existing public transit connectivity gaps, particularly for disadvantaged populations, in reaching essential services. It can also act as a decision support tool for recommending improvements (e.g., prioritize the stations and routes, identify the necessity for introducing a new line within existing infrastructure, etc.) The project ties to the NITC theme of improving the mobility of people and creating vibrant communities. We expect that it can be adapted over time to cover different geographies and incorporate new data sources. In addition to serving transit agency staff, the tool can be used in university curriculum and by advocacy organizations engaged in transportation decision-making. Finally, the project lays the foundation for NITC developing other open-source tools using big data.

## 1.0 INTRODUCTION

Urbanization, combined with rapid population growth, strains existing infrastructures to their limits. This is particularly true for urban transportation systems, whose functionality is essential to the quality of life for city residents. Traffic congestion leads to travel delays, potentially resulting in significant economic losses (Schrank et al., 2015). It also impacts air quality and the environment. Given the magnitude of challenges associated with personal vehicles, public transit plays a pivotal role in economic vitality, resource conservation, emission reduction, and traffic congestion relief. For example, there has been over \$7 billion in economic investment along the rail lines in the Salt Lake City metropolitan area since their construction, and major employers have indicated that convenient access to transit is a key factor in their business location decisions (UDOT, 2015). Good transit service encourages active transportation such as walking and biking. Consequently, increased public transit ridership may reduce emissions from personal vehicles and slow the decline in air quality. On the other hand, the social functions of urbanized areas are highly dependent on and supported by convenient access to public transportation systems. Poorly designed transit system can cause social exclusion, particularly for the less-privileged populations with limited auto ownership who rely heavily on public transit to access essential services (i.e., jobs, schools, healthcare and grocery) (SEU, 2003).

To make public transit “appealing,” one of the critical elements is to identify the connectivity gap for both existing and planned public transit systems, which has long been an important yet challenging issue for transportation researchers and authorities. This requires a thorough understanding of both the demand side (transit riders) and the supply side (transit service). Specifically, how the multimodal demand and supply are interacting with each other and how people respond and adapt their behavior to the existing systems. With the profusion of the Internet of Things (IoT), a myriad of new, open transportation data sources, including social media, general transit feed specification (GTFS), census transportation planning products (CTPP), and real-time transit location trackers, have become available and offer an unparalleled opportunity for transit agencies to uncover the dynamic and complex interactions between users and transit services. While the potential value of these new open data is vastly acknowledged, it remains elusive on how to effectively and efficiently integrate these data into transit decision-making to improve overall performance and strengthen the tie between transit riders and agencies.

Take transit accessibility evaluation as an example: Transit accessibility is defined as the ease of travel for an individual to reach a desired destination via public transit. Previous studies have identified several transit accessibility measures which can be categorized on the basis of whether travel time is taken into account. The *travel time-discretionary* measures, which do not consider travel time, focus on service coverage, frequency and comfort of service. However, overlooking the impact of travel time, which is a major contributing factor for evaluating the ease and feasibility of transit use, tends to overestimate the accessibility (Polzin et al., 2002). Consequently, the *travel time-dependent* measures, accounting for the travel time between origin and destination on top of other miscellaneous factors (e.g., service coverage), have been gaining popularity in recent years (Lei and Church, 2010). Most of the relevant studies using the travel time-dependent measures focused on transit accessibility for a specific time of day (e.g., peak hour), yet ignored the temporal fluctuation in travel time throughout the day due to transit schedule variation (Farber et al., 2014). The drawback for such analysis is that it might lead to an over-optimistic evaluation in transit accessibility due to frequent service during peak hours. The availability of GTFS and CTPP, as well as their open-source nature, enable the tracking of such measures in both spatiotemporal dimensions.

On the other hand, social media platforms such as Facebook, Instagram and Twitter have revolutionized the process that information is generated, shared and stored. With the profusion of IoT devices, huge amount of social media data are created. For instance, Twitter reported that 500 million tweets are sent each day (Maghrebi et al., 2015). The social media data have drastically altered the way information is disseminated and exchanged (Kaplan and Haenlein, 2010). With rich semantic and multimedia content, users of these location-based social media services can be seen as “semantic sensors” with the ability to report and describe events by sending messages with geographic footprints (Goodchild, 2007). Such datasets also present unprecedented opportunities for creating a cohesive and seamless integration of urban transportation and technology. It has the potential to provide context to transportation performance monitoring and evaluation. Forward-thinking transportation analytics has started to realize the advantages of using such an explosion of data to manage mobility. For example, the city of Los Angeles partnered with Google Waze to extract information from people using this navigation app and learn where congestion hot spots are (Goldsmith, 2016). The city also partnered with Esri and developed a geospatial data visualization platform. One of the projects, “High Injury Network,” maps the city’s pedestrian and cyclist fatalities related to traffic incidents to identify risk factors and prevention strategies (Vision Zero, 2016). Such developments, integrating the physical transportation assets with virtual structure, allow agencies to improve traffic management and operations and the general public to better understand their local environment. More importantly, it will inform evidence-based and data-driven decision-making in transportation policy and investment choices.

Public transit is in direct competition with automobiles. Transit agencies always aim to achieve the highest ridership possible with the least operational costs, as ridership is generally considered as a surrogate measure for revenues (Wei et al., 2017). A myriad



of factors can affect transit ridership, including service quality (reliability, comfort and convenience), service coverage, station accessibility, and user experience (Fayyaz et al., 2017; Farber et al., 2016). The current practice for transit agencies to evaluate user experience is to conduct customer satisfaction surveys of bus riders. Through these surveys, passengers express their opinions about various attributes describing quality of transit service in terms of a pre-defined scope of evaluation (Kittelsohn & Associates, 2003). The high cost, limited sample size and low resolution have been the major obstacles to make the full use of survey results to inform investment decisions. Moreover, travelers' actual experiences might tell entirely different stories in comparison with these surveys. One alternative to gauge transit riders' experience is through the mining of social media data to augment the data collected via traditional approaches. This method is much less costly and time consuming, and allows transit agencies to leverage synergistic benefits for effective transit planning and management.

To this end, in this project we build the STAT system, an open-source *social-transportation analytic toolbox*, to facilitate the use and integration of new, open transportation data for transit agencies in discovering fundamental patterns of interactions between users and transit services, improving the overall performance of the public transit network, and strengthening the tie between transit riders and agencies. Our goal is to leverage the STAT system to provide a rich, public transit analytical platform to enable transit agencies to effectively explore insights from the integrated transportation data. Specifically, leveraging machine learning and natural language processing techniques, we retrieved Twitter data that are related to public transit systems and extracted sentence structures to geomap those tweets to their corresponding transit lines/stations. Combined with transit accessibility measures computed using GTFS, we will be able to identify the mismatch between the services the agency is providing versus what the transit users are experiencing.

## 2.0 BACKGROUND

### 2.1 TRANSIT ACCESSIBILITY MEASURES

Transit accessibility consists of two core elements, activity element and transportation element (Burns, 1980; Koenig, 1980). The activity element reflects the potential opportunities available at a destination and is usually measured by population density, job density, and/or service/facility availability at the destination. The transportation element reflects the ease of travel and is affected by the spatial and temporal coverage of transit, cost of travel (e.g., travel time), and comfort of service.

Several *travel time-dependent* accessibility measures have been developed to date such as competition measures (Joseph and Bantock, 1982; Scheurer and Curtis, 2007; van Wee et al., 2001); constraints-based measures (Bhat et al., 2000; Geurs and van Wee, 2004; Scheurer and Curtis, 2007); composite measures (Miller, 1999); and cumulative and gravity-based measures (Farber et al., 2014; Foth et al., 2013; Lei and Church, 2010; O'Sullivan et al., 2000). The latter two are the most widely used ones (Anderson et al., 2013; Benenson et al., 2010; Dill et al., 2013). Cumulative measures are based on the number of potential opportunities to be reached within a certain cost (e.g., travel time) threshold (Bhat et al., 2000; Geurs and Ritsema van Eck, 2001), and can be expressed as:

$$A_{ic} = \sum_{j=1}^J B_{ij} * a_j \quad (2-1)$$

where  $A_{ic}$  is the cumulative accessibility measure at a location  $i$ ,  $a_j$  represents the potential opportunities at location  $j$ , and  $B_{ij}$  is a binary value, with 1 indicating that location  $j$  can be reached within a predetermined threshold (e.g., within a one-hour travel time window) and 0 otherwise. This measure assumes that a destination is reachable if and only if the impedance of reaching it is lower than the threshold. As a result, two destinations with the same potential opportunities would have the same impact on the measure as long as the impedance of reaching them are both within the threshold. Additionally, if the travel time to a desired destination is slightly outside the predetermined threshold, then this destination is deemed as inaccessible.

Gravity-based measures attempt to address this single-threshold deficiency by weighting the potential opportunities that can be reached based on a cost function (e.g., travel time) (Bhat et al., 2000; Geurs and Ritsema van Eck, 2001). The general form is:

$$A_{ig} = \sum_{j=1}^J O_j * f(C_{ij}) \quad (2-2)$$

where  $A_{ig}$  is the gravity-based accessibility measure at location  $i$ ,  $O_j$  is the potential opportunities at location  $j$ , and  $f(C_{ij})$  is the impedance or cost function (e.g., travel time) for travelling between  $i$  and  $j$  via public transit. The main challenge of this method is the need to develop an impedance function between all OD pairs, other than estimating the number of potential opportunities at each location (El-Geneidy and Levinson, 2006).

The gravity-based measure is able to account for spatial coverage, service frequency, destination attractiveness and travel time between origins and destinations. By adding the temporal dimension to the gravity-based measure, it provides the most comprehensive picture of transit accessibility.

Weighted average travel time (WATT) is in nature a gravity-based accessibility measure that weights travel times based on the attractiveness (potential opportunities) of destinations. According to Cao et al. (2013), the WATT between stations can be described as:

$$WATT_i = \frac{\sum_{j=1}^J M_j * tt_{ij}}{\sum_{j=1}^J M_j} \quad (2-3)$$

where  $WATT_i$  is the weighted average travel time of station  $i$ , also referred to as location indicator.  $M_j$  is the potential opportunities (e.g., population density) of station  $j$ ,  $tt_{ij}$  is the travel time (including egress, ingress, waiting and transfer time) via public transit from station  $i$  to station  $j$ , and  $J$  is the total number of stations within a transit network. WATT is based on a gravity-like interaction pattern between locations (Geertman and Ritsema van Eck, 1995) - an increase in potential opportunity (gravity) and a decrease in travel time (distance) will increase the accessibility (gravity force) between two stations (masses). WATT is intuitive to interpret. For example,  $WATT_1 = 60$  minutes indicates the average travel time from station 1 to all the other stations is 60 minutes. Calculating WATT for all times of day will provide a comprehensive transit accessibility measure that captures the temporal variation in services.

The major drawback of past studies, as Farber et al. (2016) mentioned, is the missing piece of tracking the temporal fluctuation in travel time throughout the day due to computational intensity. This directly results in an over/underestimation in transit accessibility (Owen and Levinson, 2015; Farber et al., 2016). Farber et al. (2016) reported that the calculation of travel time between all stations for all times of day for Salt Lake City's network with 1,400 stations, and 100 transit routes would take approximately 60 days on a quad-core machine in ArcGIS.

## 2.2 TWITTER DATA FOR URBAN TRANSPORTATION ANALYTICS

A myriad of studies have attempted the use of social media for transportation research. These studies can be classified into four major categories including travel demand estimation (Tasse and Hong, 2014; Golder and Macy, 2014; Yin et al., 2015); mobility behavior assessment (Cho et al., 2011; Hasan et al., 2013); traffic condition monitoring (Tian et al., 2016; Steur, 2015); and incidents and natural disasters modeling (Sakaki et al., 2010; Ukkusuri et al., 2014; Mai and Hranac, 2013). Only several studies to date have used social media information for public transit analysis, mostly focusing on sentiment analysis to evaluate transit system performance from transit riders' perspective (Schweitzer, 2014; Collins et al., 2013; Luong and Houston, 2015).

Schweitzer (2014) used tweets to evaluate users' opinions about public transit. She found that Twitter users express more negative sentiments about public transit than

other public services (e.g., police departments). Moreover, transit agencies that respond directly to the questions and criticisms of their users demonstrate more positive sentiments. Collins et al. (2013) analyzed Twitter data to assess transit riders' satisfaction using a sentiment strength detection algorithm. They collected tweets containing keywords of train names in the city of Chicago. Their results revealed that transit riders tend to express negative sentiments to a situation (e.g., power outages) than positive sentiments. Luong and Houston (2015) conducted sentiment analysis to examine Twitter users' attitudes towards light rail services in Los Angeles. Data were collected using the Search Twitter API around Los Angeles using the names of seven light rail lines. Steiger et al. used various social media data including Twitter, Foursquare, Instagram and Flickr to analyze public transit flow and detect major transit hubs in London. They used a latent Dirichlet allocation model to extract train-related tweets and then applied density-based spatial clustering to find clusters with points closely packed together. They found that detected clusters are spatially located along the track segments of London. The results were validated using an overlay of the major rail and public transit network from OpenStreetMap.

These aforementioned studies provided valuable insight on the applications of social media data in public transit analysis. While extracting relevant tweets has significant impacts on the accuracy of results, most of the previous studies only used a simple keyword search to filter transit-related tweets. Yet, based on our preliminary analysis, most of these tweets might not really reflect users' feedback on quality of transit service. And, more importantly, no research attempted to associate these tweets with location-specific features (e.g., infer location information from the tweet itself), and rather just analyzed the overall semantic patterns.

## 3.0 METHODOLOGY

The STAT enables transit agencies to integrate, analyze and visualize major new, open transportation data, including social media, GTFS and CTPP. We use the Salt Lake City and Portland metropolitan areas as two case studies to demonstrate how the retrieval, analytics and visualization of the open data can be used by transit agencies for performance assessment and public opinion collection. This methodology section entails the detailed methods we employed for accessibility computation and Twitter data retrieval and geomapping. The STAT platform is currently hosted at <http://xiaoyueliu.net:8002/>

### 3.1 TRANSIT ACCESSIBILITY COMPUTATION

GTFS was created in 2005 by Google and TriMet for transit agencies to describe their schedules, trips, routes and stops data in an open-source format that can be used for the Google Transit Web-based trip planner. GTFS has evolved ever since based on the feedback from agencies and developers. To date, the majority of transit agencies have made their GTFS data available to the general public (323 transit agencies nationwide) (Google Transit Data Feed, 2016). A GTFS dataset consists of several plain text files which have been formatted as comma-separated values . In public transport networks, stops represent transit stations where vehicles pick up and drop off passengers. Routes are sequences of two or more stops whose schedule is followed by a transit vehicle. Multiple trips can occur following the same route throughout a day. Therefore, a trip is a sequence of two or more stops that occurs at a specific time.

Here we present our algorithm design for computing the accessibility measure (i.e., WATT) using GTFS data. The core component is the capability of finding the shortest path and updating the travel time between stations in both spatiotemporal dimensions. We further decipher the travel time matrix to explore the impact of network connectivity on the accessibility measure. Figure 3.1 presents the overall methodological framework including the core components (e.g., datasets, techniques and formula) and their relations.

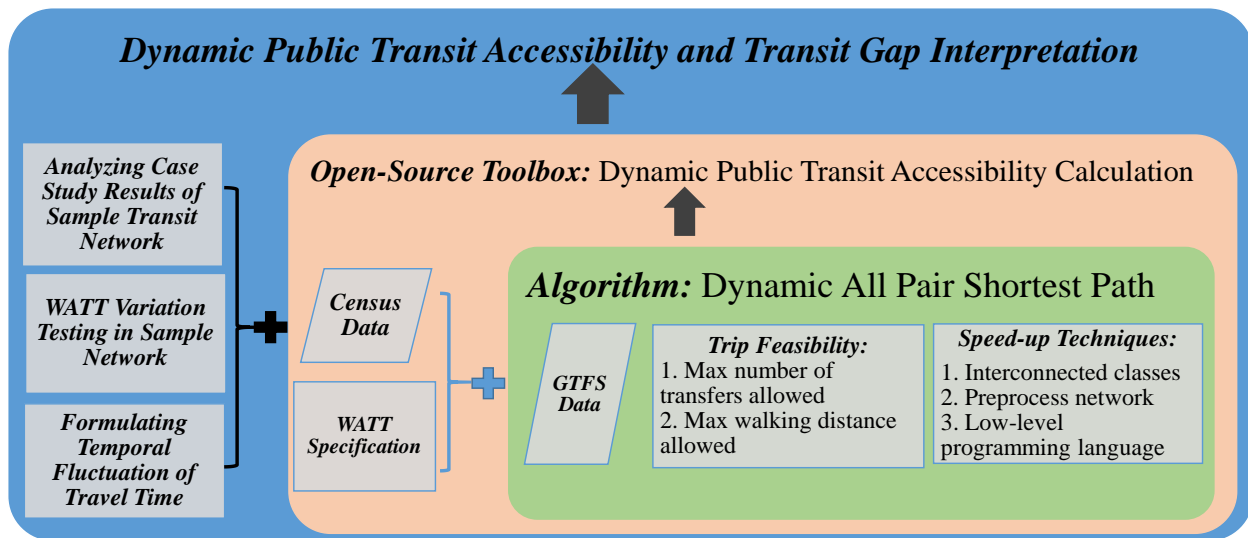


Figure 3.1: Methodological framework

The proposed algorithm starts at each station by finding the next available trip passing through this station and the immediately connected stations at a specific time of day. These trips are further traced and travel times for stations met on these trips are updated. If the met stations are transfer stations (connected to a new route) then the next available trip on the new route is traced as well. This process continues until either all the stations in the network are met or the trips appear unpractical from users' perspective. All the calculations of travel times are based on the time-table reading from GTFS files, so there is no need to build the network graph for each departure time. We assume that any user is willing to take up to four transfers and walk up to 700 meters for transfers to reach a destination (Kittelson et al., 2003). With this assumption, the algorithm is described as follows:

**Step 1: Input data**

The GTFS data is read into three classes including stops, routes and trips. Stop class contains route and trip members that store the passing routes and trip IDs. The route class includes stop member storing station IDs that are visited by the routes. The trips are stored in hash tables in order to improve the process of finding the next available trips.

**Step 2: Find connected routes to each station and update travel time by walking**

In this step, the distances between all stations are calculated and converted to travel time assuming a constant walking speed of 2.98 miles per hour (O'Sullivan and Morrall, 1996). The values are stored in a stop class member vector called travel time ( $tt_i$ ). In addition, if a stop is in close vicinity of another stop within 700 meter radius that serves different routes, those routes will be added to the route member and both stops will be added as connected stops members of the stop class. The time complexity of this step is  $O(V^2)$ .

It is important to mention that when a destination was not reachable within four transfers, the walking time between origin and destination stations was selected as the travel time. This prevents the WATT value from becoming extremely small or large. Specifically, the impact of travel time to reachable destinations will be undermined if a

large travel time value is selected for non-accessible destinations. The walking time is used as travel time between origin and destination stations only in cases where transit travel time is longer than walking time and walking distance is less than 700 meters.

**Step 3: Find all-pairs travel time and WATT for each station for all times of day**

The pseudocode for calculating all-pairs travel time for all times of day is shown in Figure 3-2. In the pseudocode  $k$  represents the number of transfers allowed; *shortest path* function finds and updates the travel time from stop  $i$  (origin) to other stops that are connected to stop  $i$  without transfer; *shortest path T* function finds and updates the travel time from stop  $i$  (origin) to other stops that are connected to stop  $i$  with 1, 2, 3, and 4 transfers, respectively, in each  $k$ ; *while loop*,  $t_o$  represents the earliest time to arrive at stop  $o$  from stop  $i$  and it is directly read from trips class,  $t$  represents the departure time from stop  $i$ , and  $t_o - t$  is the shortest path (travel time) from stop  $i$  to stop  $o$ .

```

For each time  $t \in T$ :                                     'T represents all departure times of day
{
    For each station  $i \in J$ :                               'J is total number of stations in transit network
    {
         $TT = tt_i$ ;                                       'tti is the travel time vector from  $i$  to all other stations
         $TS = \text{Shortest Path}(i, t)$ ;                   'Shortest Path is shortest path function without transfer
        Set  $k = 0$ ;
        While  $k < 4$ :
        {
             $TS = \text{Shortest Path}_T(i, t, TS)$ ;         'with transfer
             $k = k + 1$ ;
        }
        Calculate and store  $WATT_{it}$ ;
    }
}
Shortest Path( $i, t$ ):
{
    Set vector <pair (stop, time)>  $TS = \text{empty}$ ;
    For each route  $m \in CR_i$ :                               'CRi is connected routes of station  $i$ 
    {
        Find closest trip  $n$  on route  $m$ ;
        Find arriving time  $t_o$  at stops  $o \in n$ ;
        For each  $o \in n$ :
            If  $t_o - t < TT_o$ :
                 $TT_o = t_o - t$ ;
            If  $o$  is a transfer stop:
                Store pair of  $o$  and  $t_o$  in  $TS$ ;
    }
    Return  $TS$ ;
}
Shortest Path_T( $i, t, TS$ ):
{
     $TS1 = TS$ ;
    Set vector <pair (stop, time)>  $TS = \text{empty}$ ;
    For each pair of  $o$  and  $t_o \in TS1$ :
    {
        For each route  $m \in CR_x$ :                           'CRx is connected routes of transfer stops in  $TS1$  vector
        {
            Find closest trip  $n$  on route  $m$ ;
            Find arriving time  $t_o$  at stops  $o \in n$ ;
            For each  $o \in n$ :
                If  $t_o - t < TT_o$ :
                     $TT_o = t_o - t$ ;
            If  $o$  is a transfer stop:
                Store pair of  $o$  and  $t_o$  in  $TS$ ;
        }
    }
    Return  $TS$ ;
}

```

Figure 3.2: Pseudocode for finding all-pairs shortest path and station WATT through a day



The WATT computation across all bus stops in STAT is implemented using the programming language C++. The reason for using C++ as the primary language is that C++ is powerful and efficient. It is fast in dealing with tens of thousands of data and returning the result within an acceptable time frame. Visual Studio 2017 is used to build the environment and implement the algorithms. Figure 3-3 shows the interface at our back-end terminal and at the front end. Several Standard Template Library in C++ are used to achieve the goal of fast access to the city's stop, trip, route and population information, which speeds up the calculation significantly. Taking the hashmap as an example, the time complexity for search, insert and delete operations is only  $O(1)$  time, which is exactly the reason why it is widely used in the project and helps produce an impressive library of algorithms and data types that greatly improve productivity.

In this project, Dijkstra's algorithm is implemented for finding the shortest path between two bus stops. In the Dijkstra's algorithm, each bus stop is considered as a node, and the rest of the nodes are marked as unvisited. The algorithm first sets the tentative distance to zero for the initial node and infinity for all other nodes. Then, if there is a bus line, the algorithm keeps updating the tentative distance once a shorter distance is found, and marks the visited nodes as "visited." When all the nodes are marked as "visited" or if the bus line transfer is more than four times, the algorithm is completed. If a route is found, the tentative distance is the shortest distance; otherwise, there is no valid route from the initial bus stop to the destination bus stop. That is how Dijkstra's algorithm works to find the shortest route between one bus stop and all the other bus stops in a city. The time complexity for this algorithm is just  $O(|V|^2)$  (where  $V$  is the number of nodes). However, other algorithms such as the one based on a min-priority queue implemented by a Fibonacci heap takes  $O(|E| + |V| \log |V|)$  time (where  $|E|$  is the number of edges between bus stops).

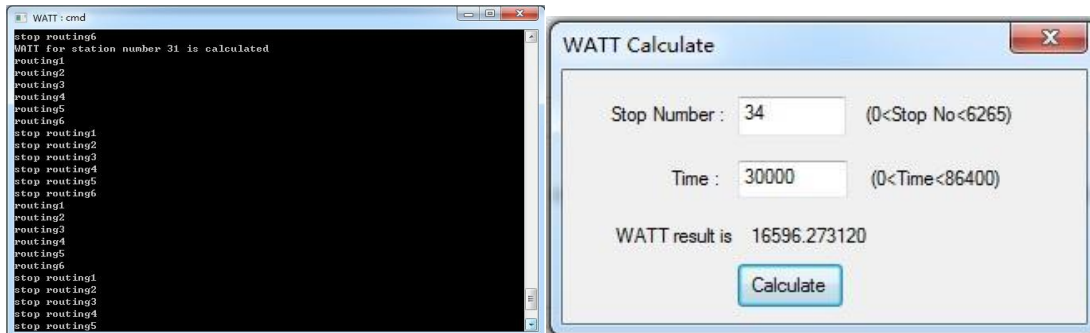


Figure 3.3: WATT calculation code runs in terminal and at GUI

As mentioned in Section 2.1, WATT weighs travel time based on opportunities; in our case, it weighs travel time based on job density retrieved from CTPP. Our algorithm described in Figure 3.2 fully captures the accessibility in temporal dimension (over a day) for each specific transit stop. We further use average to median WATT ratio (AMWR) as a unified ratio to capture the spatiotemporal variation of transit service provisions. Based on our finding (Fayyaz et al., 2017), when  $AMWR < 1$  (i.e., WATT distribution negatively skewed), temporal fluctuation in service was large (compared against the WATT range). A majority of the WATTs during the day were closer to the

maximum. On the other hand, when  $AMWR > 1$  (i.e., WATT distribution positively skewed), the temporal fluctuation in service was small (compared against the WATT range). In the latter case, the transit service appeared to be frequent and consistent.

## 3.2 TWITTER DATA ACQUISITION AND PROCESSING

To determine whether a tweet is relevant to public transit, first it is discriminated by whether it has the semantic of “complaint on the public transportation that is not on time.”. Take the tweet in the following figure as an example.

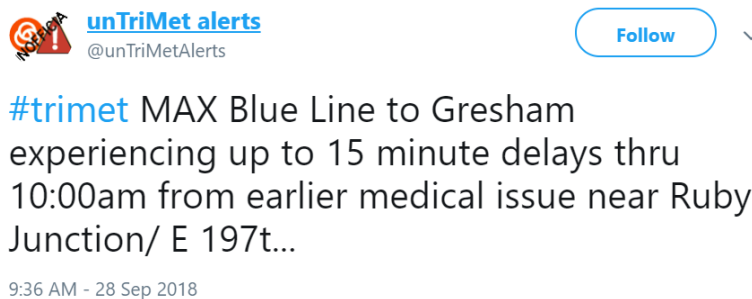


Figure 3.4: A tweet with complaint semantic

The tweet apparently has the semantic that the MAX Blue Line didn't arrive on time at Ruby Junction/E 197t station. Therefore, such a tweet will be classified into the “transit complaint” category, and tweets without such semantic will be classified into the other category.

Once all “transit complaint” tweets are filtered out, an information extraction method is applied to extract the geo information of such tweets. Take the previous tweet as an example, “Ruby Junction/ E 197t” should be extracted and treated as the geo label of the tweet.

Tweet processing involves two major components. One is a **classification model** to filter the tweets with semantics that are relevant to public transit complaints. The other one is an **information extraction method** that would extract the geo information of the complaint tweets.

### 3.2.1 Classification Model to Filter Relevant Tweets

Twitter's standard search API allows simple queries against the indices of recent or popular tweets and behaves similarly to, but not exactly like, the Search UI feature available in Twitter mobile or web clients. The Twitter Search API searches against a sampling of recent tweets published in the past seven days (<https://developer.twitter.com/en/docs/tweets/search/overview>). The standard search API is focused on relevance and not completeness. This means that some tweets and users may be missing from search results. To retrieve more eligible data, we employed

the method described in <https://github.com/Mottl/GetOldTweets3>. The library mimics how Twitter Search through a browser works. Basically, when a user enters the Twitter page, a scroll loader starts, one would get more and more tweets as he/she scrolls down the page, all through calls to a JSON provider. We are able to search through some of the older tweets via the library.

The following attributes describe a specific tweet.

- id: a unique number describes the tweet
- permalink: the URL link of the tweet
- username: the user who publishes the tweet
- text: the content of the tweet
- date: date the user tweets
- retweets: number of retweets
- favorites: number of favorites
- mentions: mentioned user
- hashtags: hashtags of the tweet
- geo: geo info of the tweet

With the library, we can retrieve the tweets by user-defined criteria. The following are the search parameters to be used to retrieve the tweets with specific constraints.

- setUsername (): An optional specific username(s) from a twitter account (with or without "@").
- setSince ("yyyy-mm-dd"): A lower bound date (UTC) to restrict search.
- setUntil ("yyyy-mm-dd"): An upper bound date (not included) to restrict search.
- setQuerySearch (str): A query text to be matched.
- setTopTweets (bool): If True only the Top Tweets will be retrieved.
- setNear(str): A reference location area from where tweets were generated.
- setWithin (str): A distance radius from "near" location (e.g. 15mi).
- setMaxTweets (int): The maximum number of tweets to be retrieved. If this number is unsetted or lower than 1 all possible tweets will be retrieved.

In this project we obtained 15,000 tweets from the search query which set "setQuerySearch" parameter as "@rideuta" and "trimet." Tweets retrieved from "@rideuta" are used as the training data. First the data are naively and automatically labeled as complaints by whether a tweet has words like "delay," "early," and "on time." Then the tweets are manually filtered to make sure the labels are correct.

The model of skip-gram, which is commonly employed on large corpus, can convert the words to vectors. In this model we are given a corpus of words  $w$  and their contexts  $c$ . Generally speaking, for a sentence of  $n$  words  $w_1, \dots, w_n$  contexts of a word  $w_i$  comes from a window of size  $k$  around the word:  $C(w) = w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+k}$ , where  $k$  is a parameter. We consider the conditional probabilities  $p(w)$  and, given a corpus Text, the goal is to set the parameters  $\theta$  of  $p(w; \theta)$ , so as to maximize the corpus probability:

$$\arg \max_{\theta} \prod_{w \in Text} \left[ \prod_{c \in C(w)} p(c|w; \theta) \right] \tag{3-1}$$

One approach for parameterizing the skip-gram model follows the neural-network language models literature, and models the conditional probability  $p(w; \theta)$ , using softmax:

$$p(c|w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}} \tag{3-2}$$

where  $v_c$  and  $v_w \in R_d$  are vector representations for  $c$  and  $w$ , respectively, and  $C$  is the set of all available contexts. The parameters  $\theta$  are  $v_{c_i}, v_{w_i}$  for  $w \in V, c \in C, i \in 1, \dots, d$  (a total of  $|C| * |V| * d$ ). We would like to set the parameters such that the corpus probability is maximized, which is the same as minimizing the negative log of the corpus probability. An example can be seen at <http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/>. From the perspective of neural networks, the architecture of the network is shown in Figure 3.5 and the word vector representation shown in Figure 3.6.

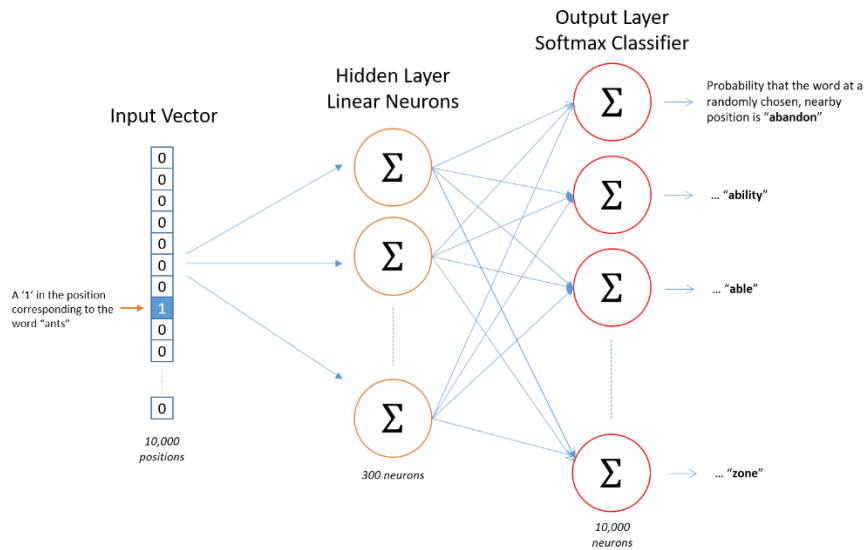


Figure 3.5: Architecture of the neural network model

The “word vector” in this model is represented as the parameter matrix from the input layer to the hidden layer, which is  $v_w$  from the equation. The matrix is depicted in Figure 3.6.

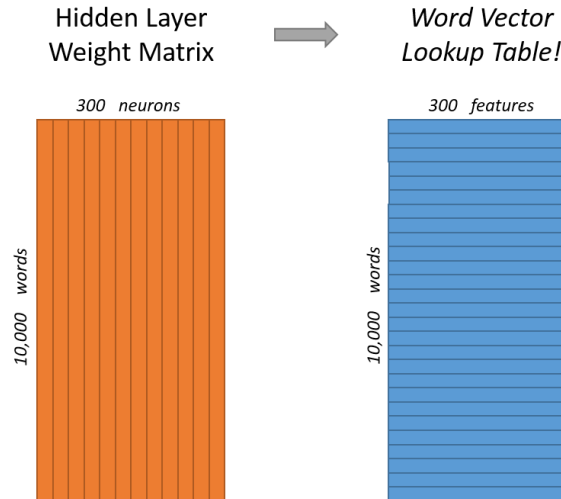


Figure 3.6: Parameter matrix of the input layer of the model, with each row representing a word vector

The example in Figure 3.6 shows that the parameters space is huge, which is two matrices with  $300 \times 10000$  parameters each. To minimize the negative log of the corpus probability with the gradient descent method. There is a huge computation cost in updating the parameters. To avoid such cost, methods such as subsampling frequent words and negative sampling is employed. The detail of the process can be found in Mikolov (2013).

The convolutional neural network model is used for tweets classification to distinguish comments and non-comments among all tweets. The model architecture is shown in Figure 3.6. Assume a tweet is constructed by  $n$  words (padded if less than  $n$ ) and each word is represented as a  $k$ -dimensional vector. The  $i$ -th word vector of the tweet is represented as  $v_i \in R^k$ . Let  $v_{i:i+j}$  be part of a tweet, which is the concatenation of words  $v_i, v_{i+1}, \dots, v_{i+j}$ . A feature is obtained by applying a convolution filter  $K \in R^{hk}$  to every  $h$  words of a sentence. With the convolution filter  $c_i = f(w \cdot v_{i:i+h-1} + b)$ ,  $c_i$  is the generated feature from the words  $v_{i:i+h-1}$ . By applying such a filter with each possible window of words in the sentence  $\{w_{1:h}, w_{2:h+1}, \dots, w_{n-h+1:n}\}$ , the feature map  $c = [c_1, c_2, \dots, c_{n-h+1}]$  is obtained. A max pooling operation  $c_{max} = \max\{c\}$  which takes the maximum value from the feature map is applied, then the captured max value is treated as the feature of this filter. Thus, the most important feature is captured from the max pooling operation. Multiple filters with different window size is used to obtain different features. These features are passed to the layer which is a softmax output. The final output is a probability distribution over labels.

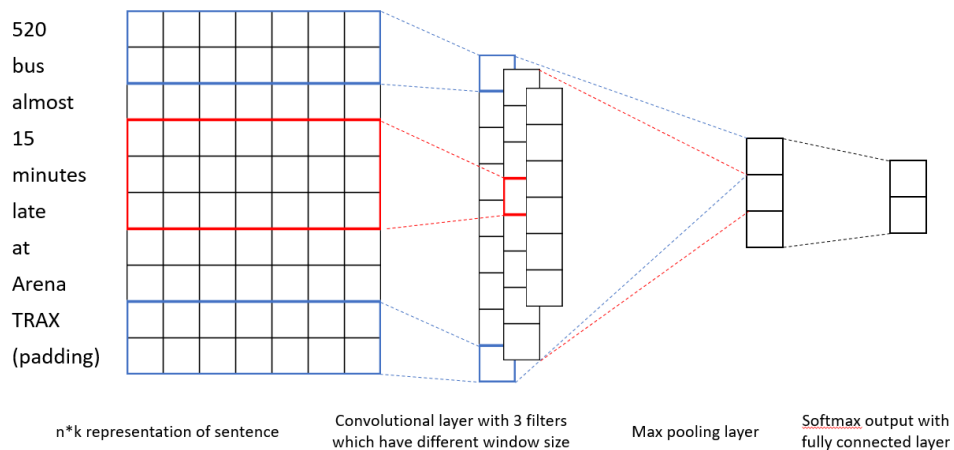


Figure 3.7: A CNN model for tweets classification

### 3.2.2 Information Extraction Method

Detecting geolocation information of tweets is also referred to as entity detection in natural language processing. Most of the named entity recognition problem can be solved by applying machine learning method directly. Due to the lack of a training dataset, we apply noun phrase (NP) chunking before named entity recognition. The result from NP chunking is used first to tag all the noun phrases in the tweet. Since the named entities in this problem are all noun phrases (e.g., stop name, bus name, time), the noun phrases then can be recognized as entities by looking up the gazetteer.

We'll take an example from (<https://www.nltk.org/book/ch07.html>) to show how chunking works.

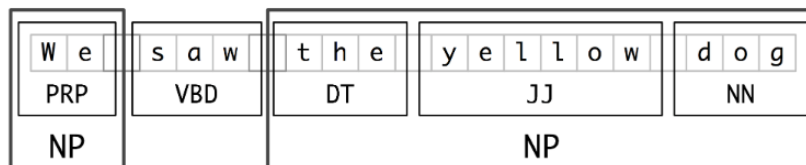


Figure 3.8: A pos tagging and chunking example

Chunking which segments and labels multitoken sequences is as illustrated in Figure 3.8. The word-level tokenization and part-of-speech tagging is shown in the smaller boxes, while the large boxes show phrase-level chunking.

Part-of-speech tagging will give a tag to each word of the sentence. Mainly the tag is used to represent the lexical category of the word. For example, the PRP here means

personal pronoun. For the rest of the tagging, Marcus et al. (1993) can be a good reference.

Usually, a subset of the tokens is chosen as a chunk. The larger boxes are called a chunk. Also like a tagger, the produced chunks do not overlap in the source text. Here NP means noun phrase. We designed a simple method to detect the geo information based on the NP chunking model. Basically, after detecting the NPs of the tweets, a measurement based on word embedding similarity is employed to evaluate whether such a NP is in the Public Transportation Stop List. For each stop in the stop list, if the similarity with the NP is larger than the threshold, the largest similarity is chosen as the detected geo information. Otherwise, the NP is not a geo information entity. Such a threshold is empirically set.

Here is an example how we do NP chunking: B-NP stands for the beginning of a noun phrase and I-NP means the word is inside of a noun phrase. The rest can be represented as O, which means outside of a noun phrase. NP chunking is a typical sequence labeling problem, here we apply the bi-LSTM-CRF model to solve this problem which is commonly applied in a sequence labeling problem. We choose the training dataset from the CoNLL2000 task (Sang et al., 2000).

The bi-LSTM-CRF model combines a bidirectional LSTM model and a CRF model. Here is a brief description how the bi-LSTM and CRF model combined. Let input sequence be  $x = (x_1, \dots, x_m)$ , which can be seen as the words of a sentence in our application. Let the sequence of output states be  $s = (s_1, \dots, s_m)$ , e.g. np chunking tags in our output result. In conditional random fields (CRF),  $p(s_1, \dots, s_m | x_1, \dots, x_m)$  is modeled as the conditional probability of the output sequence given a input sequence. To model such probability, a feature map  $\Phi(x_1, \dots, x_m, s_1, \dots, s_m) \in R^d$  is defined to map the input sequence with the output sequence to some d-dimensional feature vector. The probability now can be represented by the log-linear model:

$$p(x; w) = \frac{\exp(w \cdot \Phi(x, s))}{\sum_{s'} \exp(w \cdot \Phi(x, s'))} \quad (3-3)$$

where  $s'$  ranges over all possible output sequences.  $w \cdot \Phi(x, s) = score_{crf}(x, s)$  is a score function (sometimes also called potential function) which can be viewed as scoring how well the state sequence fits the given input sequence. Such a log-linear model is a typical CRF model. Now replacing the linear scoring function by a non-linear neural network is how the models are combined. We define the score:

$$score_{LSTM-CRF}(x, s) = \sum_{i=0}^N W_{s_{i-1}, s_i} \cdot LSTM(x)_i + b_{s_{i-1}, s_i} \quad (3-4)$$

where  $W_{s_{i-1}, s_i}$  and  $b$  are the weight and the bias corresponding to the transition from  $s_{i-1}$  to  $s_i$ , respectively. Here the LSTM function is the neural network model; we use a bi-LSTM in our model (see Huang et al., 2015 for more details).

After getting the NP result, in order to figure out what the named entity of the NP is algorithms like dynamic programming can be applied to calculate the minimum edit distance (also called Levenshtein distance) between the NP and the entities in the gazetteer.

## **4.0 STAT VISUALIZATION AND ANALYTICS**

### **4.1 STAT PLATFORM INTERFACE**

STAT is designed as a unifying interactive visualization platform to support querying, navigating and exploring the interactions between transit users and services. The platform consists of three components: temporal distribution of transit stops' accessibility of chosen, transit-stop positioning in Google Maps with geomapped tweets around that stop, and overall accessibility visualization at the traffic analysis zone (TAZ) level shown in ArcGIS. The platform integrated data from two metropolitan areas: Salt Lake City and Portland. With multiple visualization methods, Utah Transit Authority



(UTA) and TriMet can assess the usability of the toolbox and have an overall picture from their respective service networks.

For the visualization of temporal distribution of a transit stop's accessibility, Chart.js is used in STAT to dynamically display the WATT of the selected bus stop. Chart.js is an open-source tool and is easy to include animated, interactive graphics. The interactive feature allows for the observation of an accessibility pattern across different times of day (as shown in Figure 4.1).

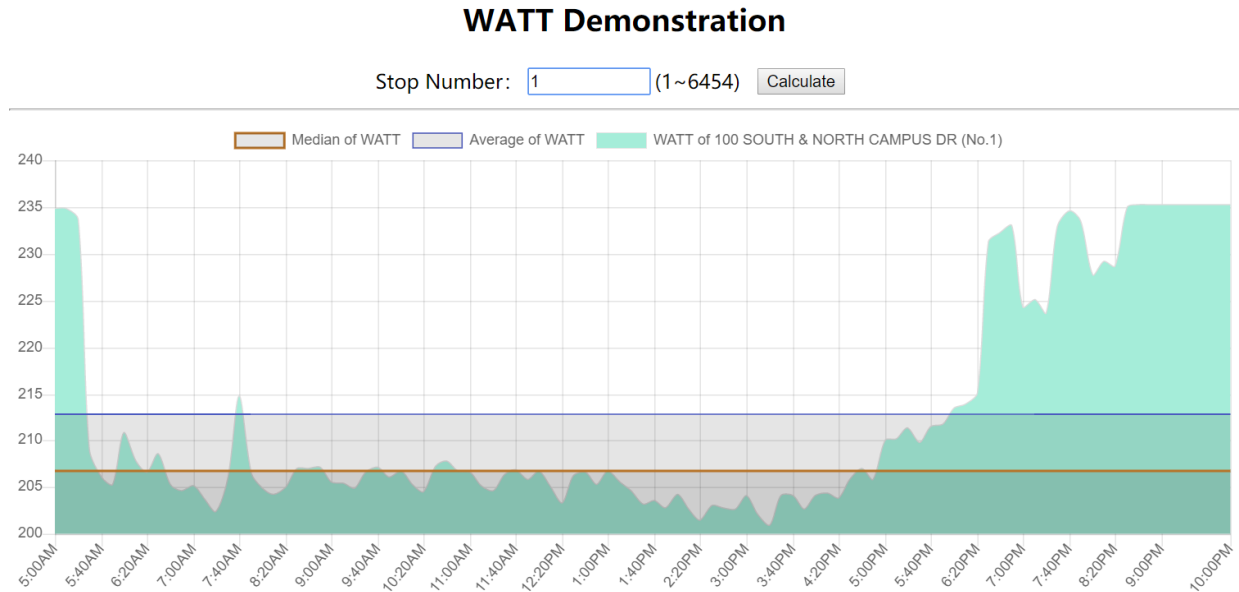


Figure 4.1: Use Chart.js to visualize WATT pattern for a transit stop

For transit stop positioning, Google Map API is used to dynamically display the location of bus stops and STAT calls jQuery.js to retrieve the current location's tweets from the captured database. Google Maps is favorable for its geolocation services worldwide, which include a database of geographical features, small businesses and street images across the globe. STAT is able to scale based on the number of tweets on the map. When the map is zoomed out, STAT will show the number of tweets around a location. And when the map is zoomed in, it will display the contents of the tweets, as shown in Figure 4.2. The jQuery is a JavaScript function library. It is lightweight and has a large number of open-source JS frameworks and extensions on the web, which makes it an ideal tool to pinpoint and edit the data crawled from Twitter.

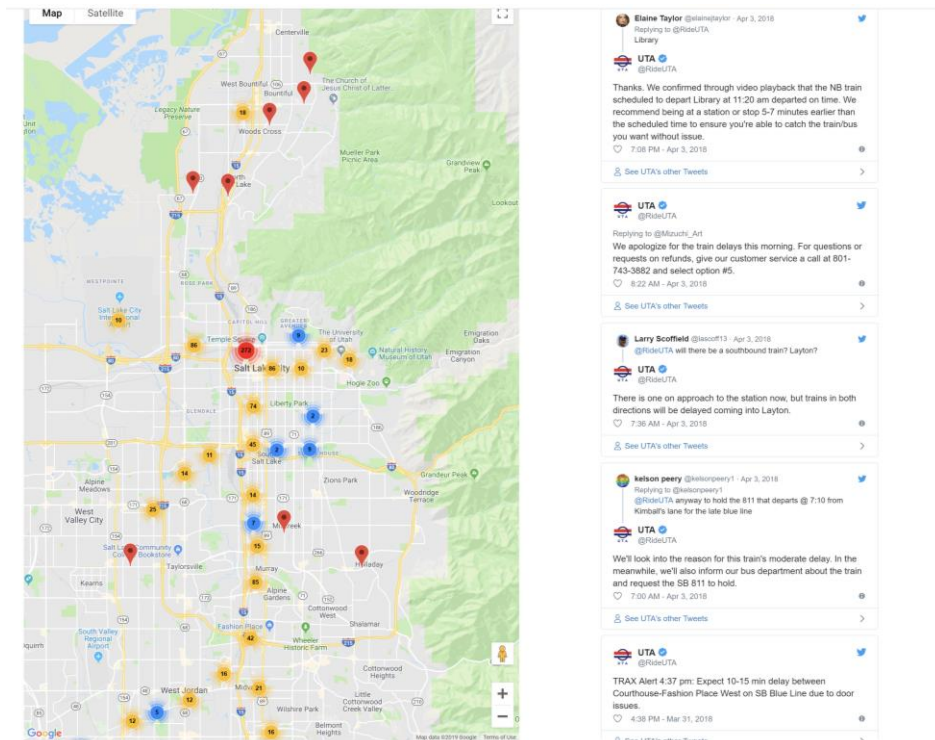


Figure 4.2: Twitter comments visualization on Google Map API in STAT

The last component of WATT is the visualization of overall WATT within a region, at the TAZ level. ArcGIS online is used to realize such a function. Users can not only refer to specific TAZ transportation accessibility information, but also overlay their own shapefiles, geodatabases or imagery.

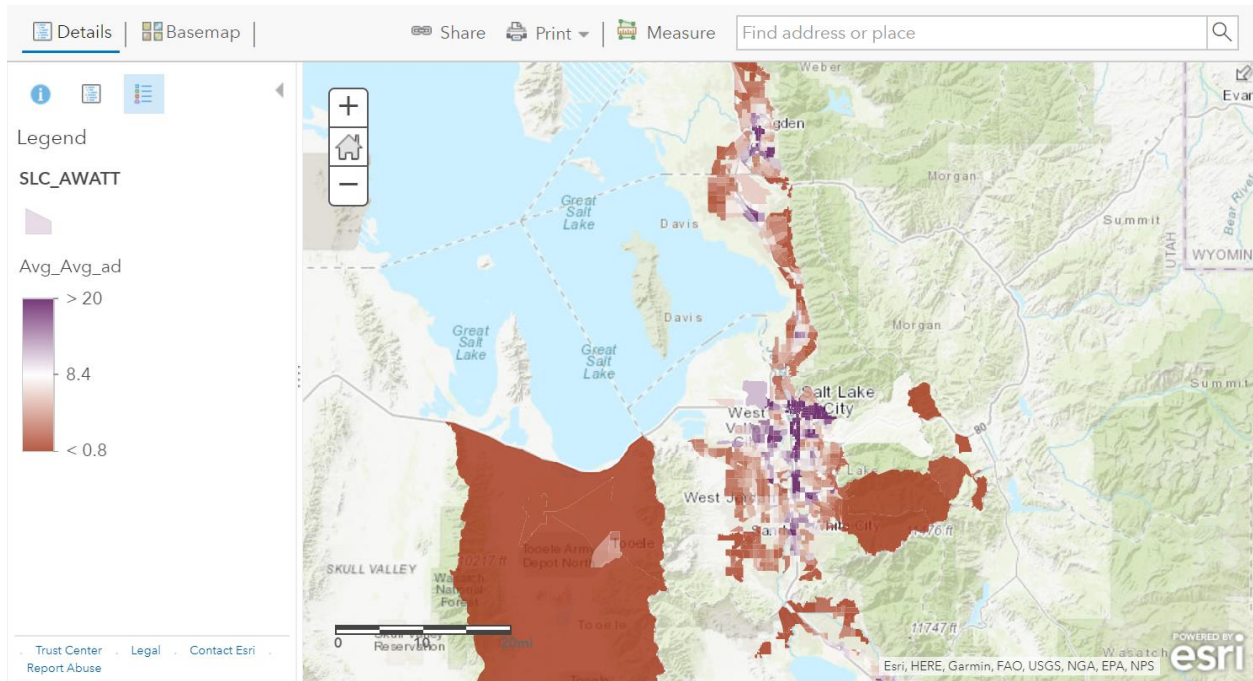


Figure 4.3: ArcGIS online visualizing the accessibility of each TAZ in Salt Lake City area

## 4.2 WATT ACCESSIBILITY INFERENCE

We use the Portland metropolitan area as a demonstration to showcase how WATT (and its derived measure, AMWR) can be used to measure and interpret transit accessibility. As shown in Figure 4.4, stations closer to the city core exhibited lower WATT, implying better access to opportunity, with diminishing returns as the distance to the core increases. This may be exacerbated by mostly radial bus lines surrounding the downtown area, with limited service in between suburbs on the periphery of the service area. Service in Forest Grove, the most westwardly town in the TriMet service area, and Estacada, southeast of Portland, have the least access to opportunity per WATT. Forest Grove has access to one transit line, the 57-TV Highway/Forest Grove, which runs frequently but only to the central core of Forest Grove. The Estacada line, 30-Estacada, runs half-hour service to Clackamas, another city southeast of Portland. This requires riders to transfer in order to reach Portland, where the majority of jobs are located.

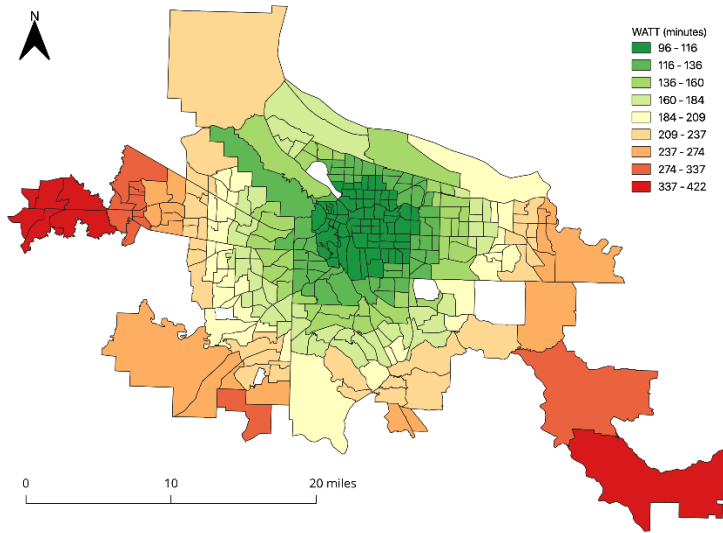


Figure 4.4: WATT distribution across Portland metropolitan area

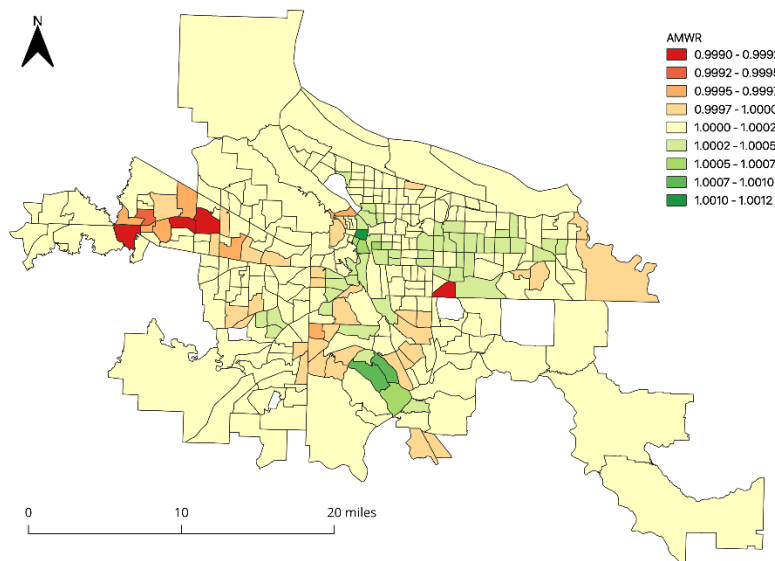


Figure 4.5: AMWR distribution across Portland metropolitan area

Figure 4.5 shows the AMWR range in the TriMet service area where a rather consistent pattern is observed. This may be contributable to the more consistent timing of bus lines across the region; one example WATT temporal profile is shown in Figure 4.6. TAZs with AMWR greater than 1 exhibit more consistent performance throughout the day. As might be expected, this occurs more near downtown Portland and some nearby suburbs, the former being the site near the Portland Transit Mall, which has access to over 20 transit lines, many of which provide daylong frequent service. Conversely, more distant TAZs generally experience less consistent service, often owing to limited bus lines or concentration on rush-hour service, as in the Forest Grove case.

In general, TAZs with transit centers and closer to major downtowns have higher AMWR, including Clackamas, Gresham, Portland, Gateway Transit Center and Barbur Transit Center. This is not steadfast, however, and is particularly untrue west of Portland, where there is high temporal variability through Beaverton, Hillsboro and Forest Grove. This may be due to commuter-oriented transit that serves higher frequencies around the morning and evening commute, but subsides in off hours. Several bus lines serve these areas which provide service only during common commuter hours, which may explain the variability. In general, however, due to the narrow band of AMWR scores, TriMet service is generally consistent throughout the day.

Stop no. 13473, AMWR=1.0000

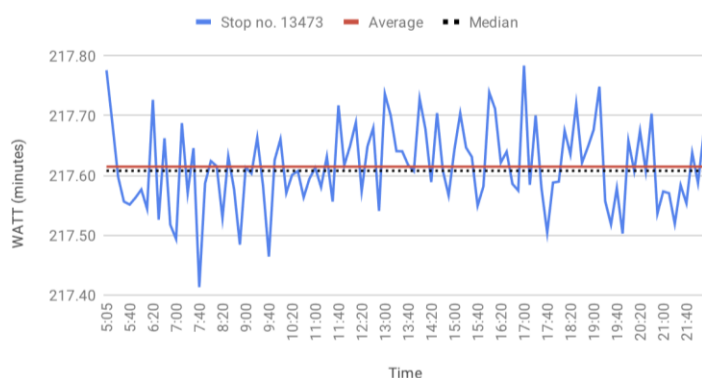


Figure 4.6: WATT temporal distribution of a sample transit stop

Forest Grove and Hillsboro are clear candidates for better transit service. Compared to Estacada, which also has poor transit service, the big difference is in the weighted income of the TAZs. Wages are lower in Forest Grove, necessitating improved transit service per the methodology. Compounding effects of the need for transit access is temporal variability, with service in Forest Grove and parts of Hillsboro varying sometimes at the availability of only one or possibly two bus lines. The lack of redundancy reduces reliability and increases off-hours demand for transit.

To the southwest of Portland are several key geographies, including the cities of Tigard and Tualatin, plus portions of unincorporated Washington County, all of which have significant commuter populations. It is also unserved by light rail, a need for which has been recognized such that the next planned Metropolitan Area Express (MAX) line in the TriMet service area will travel from southwest Portland to Tualatin. In the meantime, riders are heavily dependent on bus service and the distance between many of the lines, which travel predominantly into or out of Portland, but not in between other cities, limits transit access.

The narrow AMWR range means that improvements to transit service depend more on frequency to satisfy demand than temporal variability. The relative consistency of transit lines is a smaller factor than the number of bus lines and how often they arrive.

Much of the Portland region has consistent and frequent transit service with good coverage. There are significant opportunities for improvement, especially in poorer peripheral cities and suburbs, that could greatly increase transit access and equity. Going forward, TriMet would benefit from adding service improvements in the western part of their service area, especially around Hillsboro and Forest Grove, to capture a high desire for better transit service.

### 4.3 INTERCONNECTED KNOWLEDGE DISCOVERY TO UNDERSTAND SUPPLY-DEMAND INTERACTION

Essentially, users' experience extracted from social media and transit connectivity analysis complement each other to present a more comprehensive picture to the agencies about how the existing system is performing. To this end, we implemented a hierarchical agglomerative clustering (HAC) algorithm as a knowledge discovery process to better understand the relationship between transit users' experience and transit service quality. HAC is a data mining method that seeks to build a hierarchy of clusters. The basic idea is to take a set of transit stop locations and partition them into subsets, so each subset has similar features. HAC is used in our project to determine whether certain patterns exist for tweets' locations.

Before HAC is implemented, data preprocess is performed. For example, in Salt Lake City, all bus stops' latitude is around  $40.7608^{\circ}N$ , and their longitude is around  $111.8910^{\circ}W$ . The position( $40.7608^{\circ}N, 111.8910^{\circ}W$ ) could be assumed as the center of Salt Lake City (downtown). To check the congregation and distribution of tweets, position information should be the primary attribute as it implies congregation directly. For each tweet location, a new set of latitude and longitude is created by subtracting ( $40.7608^{\circ}N, 111.8910^{\circ}W$ ) from its original coordinates. Correspondingly, the distance between a transit stop and the center of Salt Lake City is calculated by the following:

$$distance = \sqrt{newLat^2 + newLong^2} \quad (4-1)$$

This new distance data is added to the data file and used for HAC. The pseudocode shown below demonstrates how HAC works.

---

**Algorithm 1:** Hierarchical Agglomerative Clustering

---

Each  $x_i \in X$  is a separate cluster  $S_i$ .  
**while** *Two clusters are close enough* **do**  
    Find the closest two clusters  $S_i, S_j$   
    Merge  $S_i$  and  $S_j$  into a single cluster

---

Note:  $x_i$  is a bus stop's location (*latitude, longitude*)  
 $X$  is the whole set of bus stops in Salt Lake City

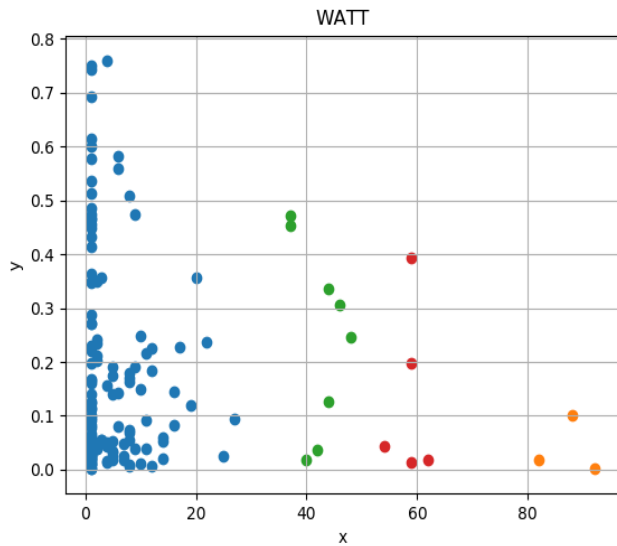


Figure 4.7: Transit stop clustering result based on HAC

Based on the HAC result, Figure 4.7 shows the congregation pattern. X-axis represents the number of tweets for a specific transit stop, and Y-axis represents the relative distance to the center of Salt Lake City as calculated above. There are several key findings from the graph. First, there are more points close to the origin along the Y-axis, which indicates higher frequency of tweets about transit services near downtown Salt Lake City. Second, transit stops close to downtown areas have a wide variation of accessibility (i.e., WATT). Since the tweets we retrieved are focusing on complaints, such a result indicates either more people are not satisfied with the transit service in the downtown areas, or people who travel around downtown tend to resort to Twitter to complain about transit service. Yet, also note that the transit stops with a large number of tweets also correspond to high WATT, which demonstrates consistency regarding the service quality and further verified that WATT, as a measure of accessibility, reflects the actual transit operation condition.

## 5.0 CONCLUSIONS

Transit agencies are oftentimes pressured to provide high-quality service. Performance measures such as percentage of on-time and accessibility are frequently used and considered critical in assessing service quality. In the meantime, agencies usually resort to annual surveys to obtain transit riders' opinions about the service. With the proliferation of open data, it offers an unparalleled opportunity for agencies to leverage this new source to uncover the dynamic and complex interaction between transit users and the services they are offering. STAT is designed as a unifying interactive visualization platform to support querying, navigating and exploring the interactions between transit users and services. The platform consists of three components: temporal distribution of transit stops' accessibility, transit stop positioning in Google

Maps with geomapped tweets around that stop, and overall accessibility visualization at TAZ level shown in ArcGIS. The platform integrated data from two metropolitan areas: Salt Lake City and Portland. With multiple visualization methods, UTA and TriMet can assess the usability of the toolbox and have an overall picture from their respective service networks. The developed platform ties to the NITC theme of improving mobility of people and creating vibrant communities. We expect that it can be adapted over time to cover different geographies and incorporate new data sources. In addition to serving transit agency staff, the tool can be used in university curriculum and by advocacy organizations engaged in transportation decision-making.

## 6.0 REFERENCES

- Anderson PA, Owen A, Levinson DM. The Time Between: Continuously-defined accessibility functions for schedule-based transportation systems. 2012.
- Benenson I, Martens K, Rofé Y. Measuring the gap between car and transit accessibility: estimating access using a high-resolution transit network geographic information system. *Transportation Research Record*. 2010 Jan;2144(1):28-35.
- Burns LD. *Transportation, temporal, and spatial components of accessibility*, 1980.
- Bhat C, Handy S, Kockelman K, Mahmassani H, Chen Q, Weston L. Development of an urban accessibility index: Literature review. University of Texas at Austin. Center for Transportation Research; 2000 May 1.
- Cao J, Liu XC, Wang Y, Li Q. Accessibility impacts of China's high-speed rail network. *Journal of Transport Geography*. 2013 Apr 1;28:12-21.
- Cho E, Myers SA, Leskovec J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining 2011 Aug 21 (pp. 1082-1090)*. ACM.
- Chen Z, Liu XC, Wei R. Agent-based approach to analyzing the effects of dynamic ridesharing in a multimodal network. *Computers, Environment and Urban Systems*. 2019 Mar 1;74:126-35.
- Collins C, Hasan S, Ukkusuri SV. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*. 2013;16(2):2.
- Dai Z, Liu XC, Chen Z, Guo R, Ma X. A predictive headway-based bus-holding strategy with dynamic control point selection: A cooperative game theory approach. *Transportation Research Part B: Methodological*. 2019 Jul 1;125:29-51.
- Dill J, Schlossberg M, Ma L, Meyer C. Predicting transit ridership at the stop level: The role of service and urban form. In *92nd annual meeting of the Transportation Research Board, Washington, DC 2013 Jan*.



- El-Geneidy, A. and Levinson, D., Access to destinations: Development of accessibility measures. 2006. Minnesota Department of Transportation: Minnesota, p.124.
- Farber S, Morang MZ, Widener MJ. Temporal variability in transit-based accessibility to supermarkets. *Applied Geography*. 2014 Sep 1;53:149-59.
- Fayyaz K, Liu XC, Zhang G. An efficient General Transit Feed Specification (GTFS) enabled algorithm for dynamic transit accessibility analysis. *PloS one*. 2017 Oct 5;12(10):e0185333.
- Fayyaz SK, Liu XC, Porter RJ. Dynamic transit accessibility and transit gap causality analysis. *Journal of Transport Geography*. 2017 Feb 1;59:27-39.
- Farber S, Ritter B, Fu L. Space–time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective. *Travel Behaviour and Society*. 2016 May 1;4:40-8.
- Foth N, Manaugh K, El-Geneidy AM. Towards equitable transit: examining transit accessibility and social need in Toronto, Canada, 1996–2006. *Journal of transport geography*. 2013 May 1;29:1-0.
- Geurs KT, Van Wee B. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport geography*. 2004 Jun 1;12(2):127-40.
- Geurs KT, Ritsema van Eck JR. Accessibility measures: review and applications. Evaluation of accessibility impacts of land-use transportation scenarios, and related social and economic impact. RIVM rapport 408505006. 2001 Jun 9.
- Geertman SC, Ritsema Van Eck JR. GIS and models of accessibility potential: an application in planning. *International journal of geographical information systems*. 1995 Jan 1;9(1):67-80.
- Golder SA, Macy MW. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*. 2011 Sep 30;333(6051):1878-81.
- Google transit data feed.  
<https://code.google.com/archive/p/googletransitdatafeed/wikis/PublicFeeds.wiki>.
- Goodchild MF. Citizens as sensors: the world of volunteered geography. *GeoJournal*. 2007 Aug 1;69(4):211-21.
- Goldsmith, S. Let's Testing Ground for Transportation Efficiency, <http://www.governing.com/blogs/bfc/gov-los-angeles-transportation-efficiency-mobility-management.html>. Mar. 2016.
- Hasan S, Zhan X, Ukkusuri SV. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing 2013* Aug 11 (p. 6). ACM.
- Haghighi NN, Liu XC, Wei R, Li W, Shao H. Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport*. 2018 Aug 1;10(2):363-77.

- Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. 2015 Aug 9.
- Joseph AE, Bantock PR. Measuring potential physical accessibility to general practitioners in rural areas: a method and case study. *Social science & medicine*. 1982 Jan 1;16(1):85-90.
- Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*. 2010 Jan 1;53(1):59-68.
- Kittelsohn & Associates, Transit Cooperative Research Program, Transit Development Corporation. Transit capacity and quality of service manual. Transportation Research Board; 2003.
- Koenig JG. Indicators of urban accessibility: theory and application. *Transportation*. 1980 Jun 1;9(2):145-72.
- Lei TL, Church RL. Mapping transit-based access: integrating GIS, routes and schedules. *International Journal of Geographical Information Science*. 2010 Feb 1;24(2):283-304.
- Luong TT, Houston D. Public opinions of light rail service in Los Angeles, an analysis using Twitter data. *iConference 2015 Proceedings*. 2015 Mar 15.
- Marcus M, Santorini B, Marcinkiewicz MA. Building a large annotated corpus of English: The Penn Treebank. 1993.
- Mai E, Hranac R. Twitter interactions as a data source for transportation incidents. 2013.
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 2013* (pp. 3111-3119).
- Maghrebi M, Abbasi A, Rashidi TH, Waller ST. Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time. In *2015 IEEE 18th international conference on intelligent transportation systems 2015 Sep 15* (pp. 208-213). IEEE.
- Miller HJ. Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures. *Geographical analysis*. 1999 Jan;31(1):187-212.
- O'Sullivan D, Morrison A, Shearer J. Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science*. 2000 Jan 1;14(1):85-104.
- Owen A, Levinson DM. Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice*. 2015 Apr 1;74:110-22.
- Polzin SE, Pendyala RM, Navari S. Development of time-of-day-based transit accessibility analysis tool. *Transportation Research Record*. 2002;1799(1):35-41.
- Sang EF, Buchholz S. Introduction to the CoNLL-2000 shared task: Chunking. arXiv preprint cs/0009008. 2000 Sep 18.

- Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web 2010 Apr 26 (pp. 851-860). ACM.
- Schweitzer L. Planning and social media: a case study of public transit and stigma on Twitter. *Journal of the American Planning Association*. 2014 Jul 3;80(3):218-38.
- Schrank D, Eisele B, Lomax T, Bak J. 2015 urban mobility scorecard.
- Scheurer J, Curtis C. Accessibility measures: Overview and practical applications. Department of Urban and Regional Planning, Curtin University. 2007;52.
- Social Exclusion Unit. Making the connections: final report on transport and social exclusion: summary, 2003.
- Steur RJ. *Twitter as a spatio-temporal source for incident management* (Master's thesis). 2015.
- Tasse D, Hong JI. Using social media data to understand cities. 2014.
- Tian Y, Zmud M, Chiu YC, Carey D, Dale J, Smarda D, Lehr R, James R. Quality Assessment of Social Media Traffic Reports-A Field Study in Austin, Texas. 2016.
- Ukkusuri SV, Zhan X, Sadri AM, Ye Q. Use of social media data to explore crisis informatics: study of 2013 Oklahoma Tornado. *Transportation Research Record*. 2014 Jan;2459(1):110-8.
- Utah Department of Transportation (UDOT) et al. Utah's unified transportation plan 2015-2040, 2015.
- Van Wee B, Hagoort M, Annema JA. Accessibility measures with competition. *Journal of Transport geography*. 2001 Sep 1;9(3):199-208.
- Vision Zero. High Injury Network. <http://visionzero.lacity.org/high-injury-network/>. 2016.
- Wei R, Liu X, Mu Y, Wang L, Golub A, Farber S. Evaluating public transit services for operational efficiency and access equity. *Journal of transport geography*. 2017 Dec 1;65:70-9.
- Yin Z, Fabbri D, Rosenbloom ST, Malin B. A scalable framework to detect personal health mentions on Twitter. *Journal of medical Internet research*. 2015;17(6):e138.