USE OF THE GENERAL TRANSIT FEED SPECIFICATION (GTFS) IN TRANSIT PERFORMANCE MEASUREMENT

A Thesis
Presented to
The Academic Faculty

By

James C. Wong

In Partial Fulfillment
of the Requirements for the Degrees
Master of Science in Civil Engineering
School of Civil and Environmental Engineering, and
Master of Science in City and Regional Planning
School of City and Regional Planning

Georgia Institute of Technology
December 2013

USE OF THE GENERAL TRANSIT FEED SPECIFICATION (GTFS) IN TRANSIT PERFORMANCE MEASUREMENT

Approved by:

Dr. Kari E. Watkins, PE, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Catherine Ross
School of City and Regional Planning
Georgia Institute of Technology

Dr. Michael Hunter
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: November 18, 2013

To transit riders everywhere.

ACKNOWLEDGEMENTS

Although a thesis is just one report in an academic career, it represents much more than the specific research tasks that led to it. It is a culmination of the experience of graduate school. To that end, I think this is the right opportunity to share my thanks with everyone who was not only helpful in this research, but who helped me while I have been at Georgia Tech.

My first thank you has to be to my advisor, Dr. Kari Watkins. We met before she began her first faculty job at Tech and I had the distinct pleasure of being her very first student. (Some will argue the veracity of this claim based on graduation dates instead of start dates; they are wrong). Kari provided me guidance and support for two and a half years both inside the classroom and out. Between conferences, presentations, travel and scholarship applications, she was always in my corner. Any student would be lucky to have her as an advisor because of her commitment to students.

My other thesis advisors, Dr. Michael Hunter and Dr. Catherine Ross are also deserving of thanks not only for being on my committee, but because of their contributions to my learning in the classroom. Dr. Lisa Rosenstein should also be called out by name because she is the reason this thesis is a page-turner; clear and concise!

One of the reasons Georgia Tech was such a good fit for me was the collaborative spirit among students. I don't think that much of the work I did would have nearly as much value without the close relationship that I shared with my colleagues and friends. Landon Reed, Aaron Gooze, Melody Butler, Mags Carragher, Candy Brakewood and many others are on that list. My work with Aaron, Landon and his brother, Jack, on a project nicknamed "Transitio dot US" was likewise crucial in the development of ideas in this thesis. Georgia Tech was definitely the right fit and, in fact, two campus facilities are also deserving of my thanks, the Campus Recreation Center's racquetball courts and Cypress Street Pint and Plate – both of which were instrumental in supporting mental health.

My interest in transit goes back nearly eight years to a session for engineering students who hadn't picked a concentration. In 45 minutes, Dr. Vukan Vuchic got me excited about the field that ultimately led to my major, my first job, my graduate degree and I'm sure many things ahead. Thanks for setting me on the right path, Dr. Vuchic.

The research for this thesis was actually begun during an internship with the transportation group at OpenPlans, which has since spun off into a group called Conveyal. Many thanks are given to my colleagues from that internship: Kevin Webb, Matt W. Conway, David Turner, David Emory, Andrew Byrd, and Aaron Ogle. I'm also grateful to Keith Gates at NTD for being available for information.

I also want to thank the many staff at Kittelson & Associates, Inc. that taught me to value research and education as part of my professional development. I left a great job to go to grad school because people like Wayne Kittelson, Brandon Nevers, Ed Myers, Yolanda Takesian, Kevin Lee and Alek Pochowski never doubted that it would be good for me to do so.

Finally, I want to thank my family whose support is unwavering. Mom, Dad and David Thomas – thanks always for the love and support… and for covering a few flights home during the holidays.

DISCLAIMER

Portions of the work contained in this thesis have been published by the author in the Transportation Research Record:

Wong, J. (2013). Leveraging the General Transit Feed Specification for Efficient Transit Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2338*(-1), 11–19. doi:10.3141/2338-02

The literature review in Chapter 2 is based largely on the literature review from the published work, as are the results in Chapter 3. The methodology in Chapter 3 has changed since the publication and now introduces a single branded GTFS Reader tool.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

API	Application Programming Interface
APTA	American Public Transportation Association
AVRH	Annual Vehicle Revenue Hours
AVRM	Annual Vehicle Revenue Miles
FTA	Federal Transit Administration
GTFS	General Transit Feed Specification
ISTEA	Intermodal Surface Transportation Efficiency Act
JSON	Javascript Object Notation
MAP-21	Moving Ahead for Progress in the 21st Century
NTD	National Transit Database
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act A Legacy for Users
TCQSM	Transit Capacity and Quality of Service Manual
TEA-21	Transportation Equity Act for the 21st Century

SUMMARY

Until recently, transit data lacked a common data format that could be used to share and integrate information among multiple agencies. In 2005, however, Google worked with Tri-Met in Oregon to create the General Transit Feed Specification (GTFS), an open data format now used by all transit agencies that participate in Google Maps. GTFS feeds contain data for scheduled transit service including stop and route locations, schedules and fare information. The broad adoption of GTFS by transit agencies has made it a de facto standard. Those agencies using it are able to participate in a host of traveler services designed for GTFS, most notably transit trip planners. Still, analysts have not widely used GTFS as a data source for transit planning because of the newness of the technology. The objectives of this project are to demonstrate that GTFS feeds are an efficient data source for calculating key transit service metrics and to evaluate the validity of GTFS feeds as a data source. To demonstrate GTFS feeds' analytic potential, the author created a tool called GTFS Reader, which imports GTFS feeds into a database using open-source products. GTFS Reader also includes a series of queries that calculate metrics like headways, route lengths and stop-spacing. To evaluate the validity of GTFS feeds, annual vehicle revenue miles and hours from the National Transit Database (NTD) are compared to the calculated values from agencies whose GTFS feeds are available. The key finding of this work is that well-formed GTFS feeds are an accurate representation of transit networks and that the method of aggregation presented in this research can be used to effectively and efficiently calculate metrics for transit agencies. The daily aggregation method is more accurate than the weekly aggregation method, both introduced in this thesis, but practical limitations on processing time favor the weekly method. The reliability of GTFS feed data for smaller agencies is less conclusive than that of larger agencies because of discrepancies found in smaller agencies when their GTFS-generated metrics were compared to those in the NTD. This research will be of particular interest to transit and policy analysts, researchers and transit planners.

CHAPTER 1

INTRODUCTION

Background

Transit planning and decision-making are increasingly data-driven processes that require extensive measurement of various metrics including ridership, on-time performance and hours of service provided. Although the methods of analysis are well documented in the Transit Capacity and Quality of Service Manual *(1)* and textbooks such as Vuchic's Urban Transit: Operations Planning and Economics *(2),* there is a large gap on the subject of obtaining or collecting data for analysis. Even the latest transit guidance documents recommend using printed timetables as viable sources of information about transit service *(1)*, a recommendation that is increasingly outdated. Other documents may simply ignore the task of data acquisition in their guidance.

A trend now codified in federal policy called "open data" calls for the availability of public data in machine readable formats *(3)* and has been discussed and advocated for in the transportation field *(4)*. The open data trend is proving itself to have many indirect benefits to the transportation industry, one of which is the availability of structured data for transit analysis. Having structured transit data available to the public has allowed for the proliferation of apps and user services, but it has also allowed for its use as a data source in transit analysis. A few project specific examples were presented in the last two meetings of the Transportation Research Board where GTFS data was used as part of an analysis *(5–7)*, but there are many more opportunities for using this data that will be discussed in this thesis.

Motivation

This project is motivated in part by proposals made by the National Center for Transit Research which identified GTFS as a potential data source for transit analyses. In that report, Catalá, Downing and Hayward explained in great detail the potential for GTFS to be used as a data source in various business activities including, most significantly, service evaluations and planning *(8)*. GTFS is a standard that is shared by hundreds of transit service providers around the world, therefore any methodology that effectively utilizes data in that format can be applied to a vast number of agencies and services. This provides new opportunities for performance measurement, benchmarking and research. For example, modes can be characterized based on their service frequencies, route lengths and stop densities in a way that was previously impractical due to the non-digitized and un-standardized format of transit information.

Still, static schedule data is limited in its ability to support decision making as most performance measures are concerned with what actually takes place rather than what is scheduled. Another motivating factor in this research is to understand the usefulness and validity of open agency-endorsed datasets in general. Trends in open transit data are fast moving and already include datasets with real-time vehicle location information; future data may even include granular ridership information. The ability to track on-time performance and reliability through open data will happen soon and this research can be used as a basis for evaluating the usefulness of agency-generated information.

Objective

The objectives of this research are to demonstrate that GTFS feeds are an efficient data source for calculating key transit service metrics and to evaluate the validity of published GTFS feeds as a data source by batch processing them and comparing the results to metrics in the NTD. By doing so, future researchers and individuals involved in transit analysis will be better informed on the use and limitations of GTFS data. This thesis documents the capabilities and processes used to generate performance measures,

which will be of interest to researchers and analysts, among others. In addition, a detailed methodology for calculating system-wide performance measures comparable to those in the NTD will be useful for anyone pursuing additional research in this field using open data for performance measurement.

Outline

Chapter 2 of this thesis is a literature review of three main topics relevant to the research: open data in transit, the General Transit Feed Specification, and performance measures in transit. This literature review will form a foundation on which the data methodologies are based. These methods are described in Chapter 3 which explains how data is compiled and processed for use in generating performance measures. Additionally, Chapter 3 contains an analytic demonstration of the power of the data methodology by analyzing the industry's use of the specification, computing stop-level headways and route-level stop-densities for an example agency. It also calculates system-wide headway metrics across 50 large agencies in North America. Finally, Chapter 4 describes an attempt to validate the use of GTFS data by comparing two metrics found in the National Transit Database for a selection of transit agencies in the United States in FY 2012. Chapter 5 discusses the findings and conclusions for the report along with gaps that future research could fill.

CHAPTER 2

LITERATURE REVIEW

This project explores the opportunities for transit analysis using a new data source available to the transit industry and attempts to validate the data source by comparing metrics derived from that data to existing metrics from the National Transit Database (NTD). To that end, this chapter explores the major concepts surrounding open data, a key requirement to capitalize on the opportunities of the data; the General Transit Feed Specification (GTFS), the data standard used in the analysis; and an overview of popular performance measures that may be applicable for calculation using GTFS feeds.

Open Data

Following a trend among public agencies to improve transparency and invite broader participation in the design of citizen services, many transit agencies have begun to publish their schedule data online for public consumption; this approach is referred to as "open data." The open data movement has been influential throughout the last few years as public sector culture has begun to accept the notion that data should be in the public realm. An executive order from May 2013, "Making Open and Machine Readable the New Default for Government Information," laid federal groundwork for how open data should be incorporated into the culture of public agencies. In the implementation guidance of this executive order, open data is described as "publicly available data structured in a way that enables the data to be fully discoverable and usable by end users." *(9)*

The same memorandum recalls the openness associated with weather and GPS data, and how that openness fueled innovation in warning systems, navigation systems and farming tools. That mentality is shared by many who advocate for open data and argue that many kinds of innovation rely on open data to succeed, even if the direct positive benefits for agencies are not readily apparent. Hemerly writes:

> "[Positive] impacts are often one or two steps down the chain from the original decision, event, or policy. It is difficult to say that the 'opening' of transit data is responsible, but it is clear that the information system built on the data, and the entry points they offered to developers, have had a positive effect. In large-scale systems, it is difficult to isolate data as individual variables to effectively measure their impact."*(10)*

This notion is supported by the computer science theory of complementarities, which suggests that coordinated activities yield higher and more efficient returns than uncoordinated activities; that they are greater than the sum of their parts. *(11)* Open data by itself is not going to prove its value, but the digital artifacts that support agencies or constituents in concert with that data have value. The benefits cannot be fully predicted because there is value in data that will only be realized when developers or engaged citizens make use of it and share insights about it. Tim O'Reilly, an influential thought leader on the subject suggests that government should act as a platform on which citizens and developers can build; by releasing data, governments allow citizens to develop user services, research and other benefits that the government agency itself would never pursue because of their narrowly defined missions *(12)*. As it pertains to transit data, one of the primary results of agencies releasing data is a host of new methods for delivering customer information *(8, 13)*.

The magnitude of the public value of open data is widely discussed in non-academic settings with enthusiasm *(14–16)*, although the empirical study that introduced the complementarity theory earlier suggests that the tangible value of open data is usually overstated *(11)*.

The open data trend is strong in the transit sector. A 2013 survey of transit agencies conducted by the American Public Transportation Association (APTA) notes that 88 percent of large agencies and just under half of small agencies surveyed provide static schedule data to third-party app developers - a proxy for open data (a separate question asked about those using Google Maps specifically). About two-thirds of all

agencies in the survey participated on Google Transit *(17)*. The market drive for use of Google Transit has likely had a great impact on the high adoption rate of GTFS and the subsequent opening of that data to third-parties other than Google.

General Transit Feed Specification (GTFS) and Transit Data

History of GTFS

The General Transit Feed Specification (GTFS), first introduced in 2005, is the result of a project between Google and TriMet in Portland to create a transit trip-planner using the Google Maps web application. Because of the collaborative approach to its development, the specification was designed to be simple for agencies to create, easy for programmers to access and comprehensive enough to describe an intricate transit system.*(14)* GTFS identifies a series of comma separated files which together describe the stops, trips, routes and fare information about an agency's service. Google opened the feed for general use in mid-2007 and it propagated widely as agencies translated their transit schedules into the format. The feed is the most used standard for static transit data exchange in the United States today. According to data from the GTFS Data Exchange as of July 2012, just over 25 percent of agencies in the United States published open transit data in GTFS format (6).

Primary Functions of GTFS

According to the specification's documentation, "[GTFS] defines a common format for public transportation schedules and associated geographic information. GTFS 'feeds' allow public transit agencies to publish their transit data and developers to write applications that consume that data in an interoperable way." *(18)* This succinctly describes its purpose and highlights a number of key elements of the specification. The first is that it covers static schedule and map data (as well as fare information), but does not include any real-time vehicle location or prediction information. Secondly, the

description envisions agencies publishing data in a one-way work flow that doesn't require two-way interaction with a potentially large number of developers. Lastly, it highlights the idea of interoperability which has been a key driver in the broad adoption of the specification as apps written for many different agencies are transferrable to others when using GTFS.

The description provided in the documentation doesn't speak to the kind of applications that would be developed, but a look at most apps using GTFS tend to provide travelers with information about various transit systems. The formats and mechanisms for providing that information vary widely among mobile apps, websites and other services, but are generally created to deliver some kind of personalized information to a traveler*(19)*. Exceptions to this include visualizations of transit movement *(20)*, geospatial applications that leverage the geographic information in GTFS feeds for tasks like apartment searching *(21)*, or other general interest applications.

Alternative Uses for GTFS Data

An important precursor to this study is a report produced by Catalá, Downing and Hayward that described the potential for alternative uses of the GTFS while proposing updates to it. In it, they wrote that "GTFS data provides a clear illustration of an agency's service and can be very helpful in understanding [the impact of service changes] *(8)*." They highlighted the wealth of visualization techniques that can help decision-makers understand the impacts of service changes. Additionally, the report describes the challenges of regional or state transportation planning due in part to the disparate data sources of multiple local agencies. The report discussed a case study with the Florida Department of Transportation District 7 office where there was a need to locate high-activity bus stops throughout the region in order to identify pedestrian safety focus areas. Aggregating and keeping their database up-to-date without a standard data feed would have been arduous; instead, their research partners used GTFS feeds and simple scripts to maintain their database. *(22)*

Some public entities are relying on this data for an array of activities including, for example, travel demand modeling. The Delaware Valley Regional Planning Commission (DVRPC) modeling group cites the advantages of GTFS feeds to avoid manual coding errors, ease data integration among multiple providers and improve general data quality. They also emphasized the importance of easily updating transit service information when schedules change, which was previously a manual task *(23)*. Researchers in San Francisco are likewise using GTFS data as part of their transit assignment model for use as a component in other planning models. *(5)* Two research efforts presented at the 92nd Annual Meeting of the Transportation Research Board in 2013 also make use of GTFS data for single-agency studies: a study from École Polytechnique de Montréal used GTFS to build public transit trip-generation models *(7)*; and a study from the University of Arizona used GTFS data to explore transit route restructuring plans *(6)*. In all instances, research focused on the use of a GTFS feed as a data set for one region, rather than the use of multiple feeds to represent multiple agencies whose metrics could be compared as will be shown in this research.

Technical Elements of GTFS

GTFS describes a series of 13 unique text files that, when compressed in a .zip file, form a GTFS feed. Each of the text files is formatted as a comma-separated-value file and the specific header fields in each text file are prescribed by the specification. A GTFS feed viewed in a typical file explorer is shown in Figure 1 along with the text contents of a stops.txt file. Additional tables and fields are allowed in GTFS feeds, but the minimum requirements are provided by the specification. The files are related to one another using certain shared values; for example, a trip in the trip.txt file is related to a route in the route.txt file by sharing the same route_id, a field in both files. This is akin to a relational database, although not called it in the specification, and the text files are often referred to as tables (as they will be for the remainder of this paper).

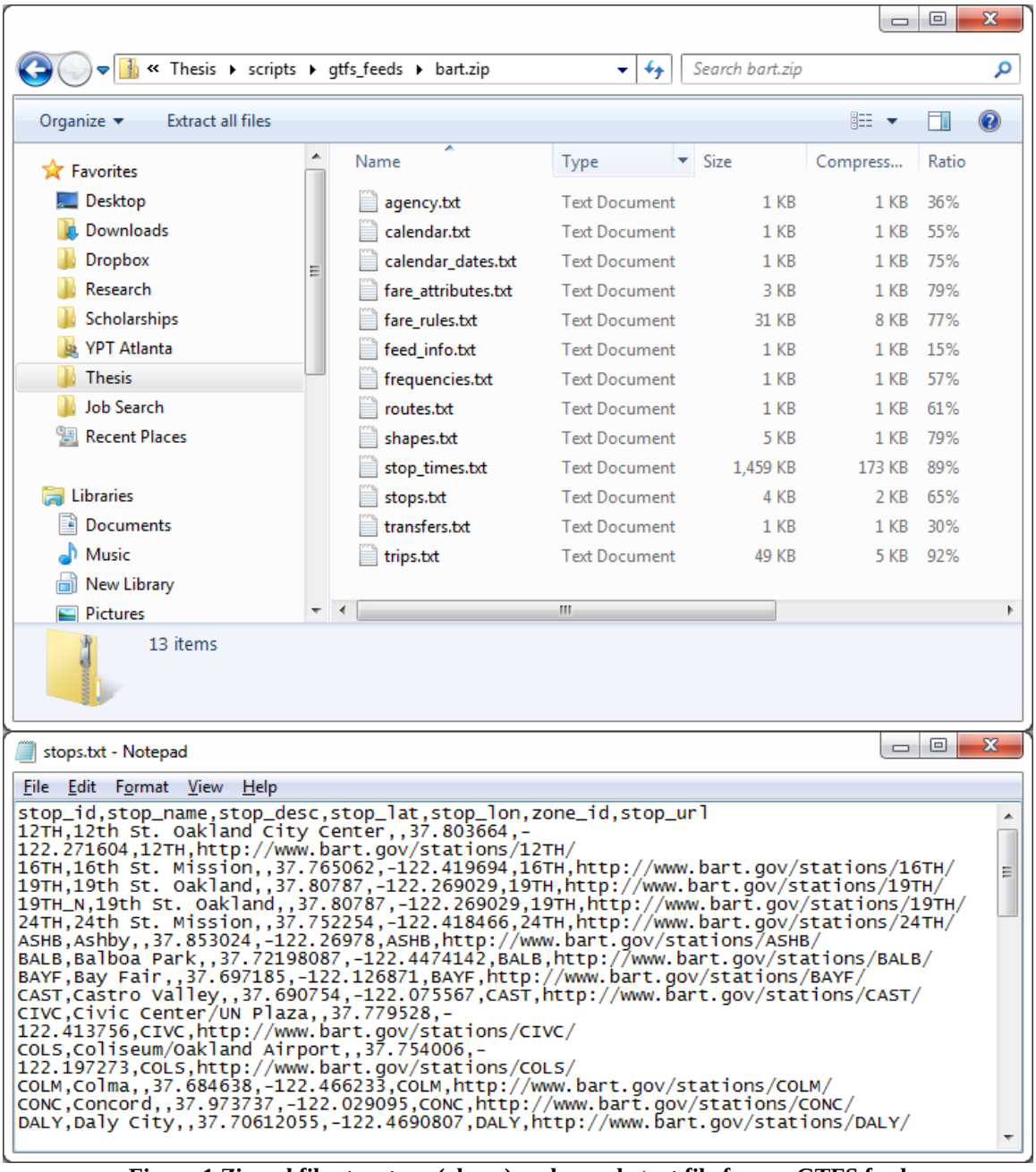


Figure 1 Zipped file structure (above) and sample text file from a GTFS feed (Screenshots from the author's computer)

The author developed a database diagram in that identifies the files from the GTFS *(18)* as database tables and shows the relationships that exist among them. It also shows which tables and fields are required or optional per the specification. The overall structure of the database tries to avoid duplicative information by creating cascading

relationships from the most disaggregated information in the stop_times table to the most aggregated information in the agency table. As an example, a row of data in the stop_times table refers to the scheduled arrival and departure of a transit vehicle on a specific trip; that trip is categorized by a route which is categorized by the agency providing it.

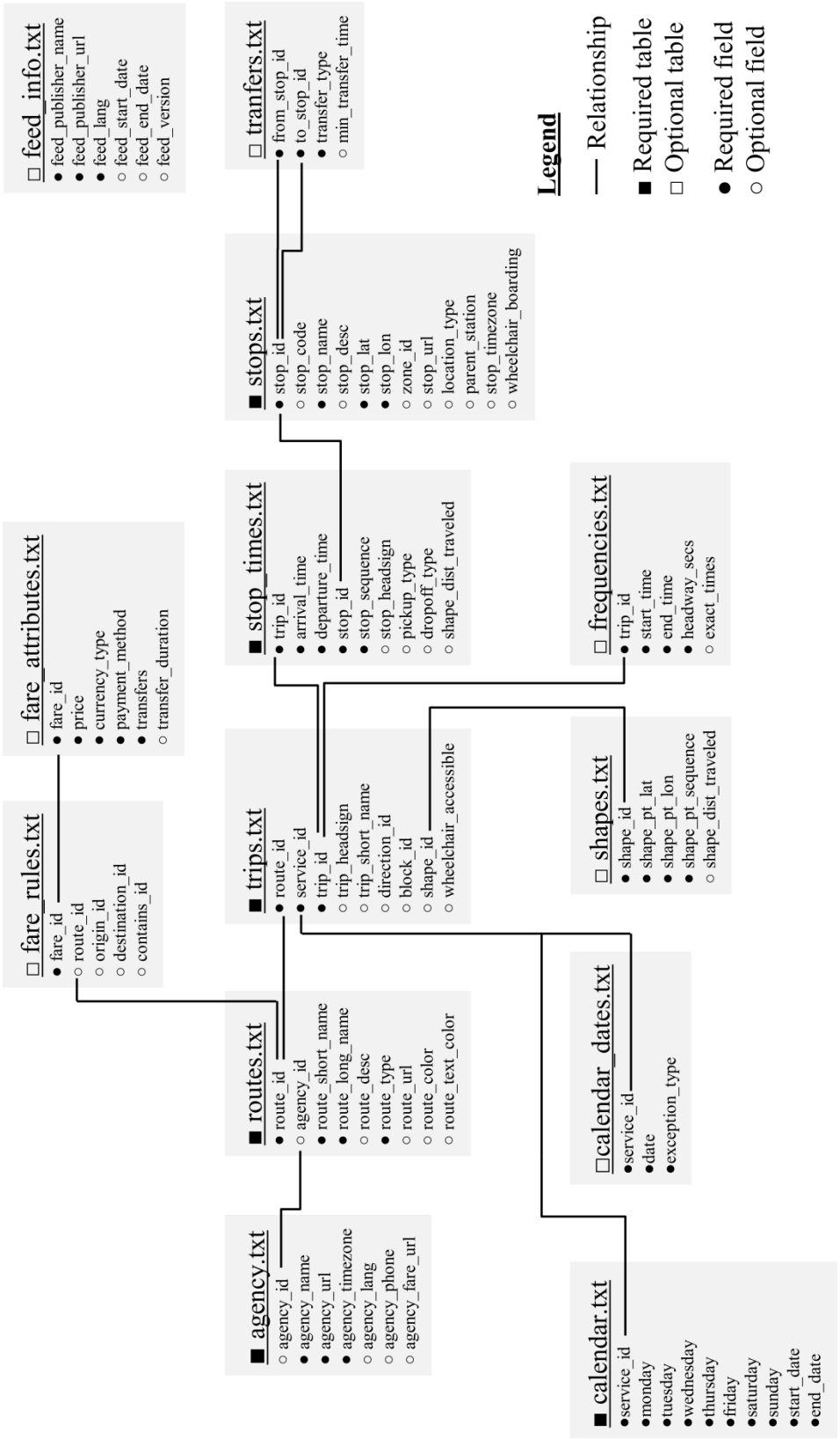


Figure 2 GTFS Specification as a database diagram

When generating a GTFS feed, the most important elements to adhere to in the specification are the use of required fields and files, and the proper relationship of data among the separate files. These structural elements allow a GTFS feed to be read into the many consumer-facing applications that make use of GTFS data. Beyond the structure, however, the creators of GTFS feeds must ensure that the data is internally consistent. For example, stop times (identified by trip_id and stop_id) must be consistently related to trips (identified by trip_id); if the trip_id differs among those two tables, the relationship will not be interpreted by GTFS-reliant applications.

There are many ways that GTFS feeds could be created incorrectly outside the overview discussed here. The Google Feed Validator is a robust open-source tool available from Google[1] that reviews the entries of a GTFS feed and reports errors and warnings including invalid values, duplicate values, unrelated ids between tables and invalid timing (such as transit vehicles that overtake one another) or route/stop placements.

Performance Measurement

Transit planning studies often require a variety of quantitative analyses and metrics to support local decision making and to evaluate a transit system among its peers. Additionally, performance measurement is relevant to agencies because they may be required to collect and report certain information and they may use them to convey the results of changes to the public or third parties. *(24)* In general, their use helps to succinctly characterize the condition of some aspect of an agency, whether it is quantity of service offered, quality of operations or other elements that can be tracked. Some of the performance measures that are reported to the NTD are actually used in the formula grants that provide substantial funding to local transit agencies. As such, the correct

[1] Available at https://code.google.com/p/googletransitdatafeed/

calculation of those measures is critical to the equitable distribution of funding for transit systems around the U.S.

Performance measures are also an increasingly popular policy tool required by legislation such as MAP-21, which requires the use of performance measures and targets as part of the planning process. *(25)* According to industry analysis of the legislation, the planning process will now include "regional surface transportation system performance targets that are coordinated with local transit providers … [and a] new planning process that will establish and use a performance based approach to the national goals [of the legislation]." *(26)* Rulemaking to implement these targets was not yet been finalized at the time of writing *(26)*. The scope of those performance measures are broader than those discussed in this research, but nonetheless support the overall need for performance measurement. The successful use of performance measures is linked to the availability of technical resources to generate those measures *(27)*. A common thread in federal rulemaking discussions is the need for performance measurements that are commensurate with available data. The Center for Transit Oriented Development actually points to the use of GTFS data as a data source for calculating a housing and transportation index for use in the national ridership model. *(28)* These suggestions would have an impact on federal performance measurement requirements, but there are many other reasons that agencies would choose to develop different kinds of performance measurements.

There are myriad types of performance measures and analyses beginning with those documented in the Transit Capacity and Quality of Service Manual (TCQSM), which describes a number of methodologies that aim to provide metrics for service availability and quality of service *(29)*. Many other studies within the past decade have proposed additional transit assessment tools and methods related to reliability *(30)(31)*, service quality *(32)*, and network evaluation *(33)(34)*. In general, however, whether relying on static or real-time data sources, these documents tend to leave data acquisition for the user to determine. As a specific example, a Transit Cooperative Research Program

(TCRP) report on transit performance measurement systems states that "measures developed using [schedule, map, operations and financial] information require little investment in staff time or resources, as the data are already being collected for other purposes and need only be compiled for use in the agency performance-measurement program *(24)*" In practice, however, data acquisition from outside an agency by consultants or researchers can be very challenging.

Like many other guidance documents, the TCQSM provides analytic methods but gives little guidance with respect to data sources. This is likely due to the variety of software solutions and reporting features available in the transit industry. As a result, researchers and analysts who try to compare or aggregate data from one or multiple agencies may face challenges in data acquisition and cleaning. Furthermore, data tools used in one region may not be applicable elsewhere, leading to customized analyses and increased costs for agencies that outsource this kind of work. When the first edition of the TCQSM was released, transit agencies in Florida, especially large ones, found it challenging to use tools that catered to specific data formats *(22)*. Following that experience, a 2008 report with application guidelines for TCQSM methods recommended using data from the National Transit Database (NTD) for some analyses *(35)*, which is challenging given its low resolution with system-level data (information is not provided at the route or stop level).

The NTD is a reporting system required by federal legislation under Title 49 U.S.C. 5335(a):

> (a) NATIONAL TRANSIT DATABASE — To help meet the needs of individual public transportation systems, the United States Government, State and local governments, and the public for information on which to base public transportation service planning, the Secretary of Transportation shall maintain a reporting system, using uniform categories to accumulate public transportation financial and operating information and using a uniform system of accounts. The reporting and uniform systems shall contain appropriate information to help any level of

> government make a public sector investment decision. The Secretary may request and receive appropriate information from any source.
>
> (b) REPORTING AND UNIFORM SYSTEMS — the Secretary may award a grant under Section 5307 or 5311 only if the applicant and any person that will receive benefits directly from the grant, are subject to the reporting and uniform systems.*(36)*

This enabling legislation requires that any agency requesting funding under traditional transit funding mechanisms participate in the NTD. Still, the legislation gives the secretary and through him the Federal Transit Administration significant latitude in the kind of information collected and the manner in which it is collected. The specific requirements of reporting to the NTD are made and amended through the federal rule-making process which provides notices and asks for input from stakeholders through notices in the Federal Register.

The following sections discuss widely used performance measures in transit and the applicability of GTFS in calculating or tabulating those measures.

Existing Measures

Transit Capacity and Quality of Service Manual

The Transit Capacity and Quality of Service Manual (TCQSM) is the leading resource on analytic methods for evaluating transit in the United States. Based on guidance in the second edition of the manual, there are six different performance measures for fixed-route transit pertaining to availability of transit services and the comfort/convenience of those services. These two categories could be analyzed at the system-wide level, encompassing multiple routes and services; the route level, concerning all transit service on a particular route designation; or the stop level, which might contain information for multiple routes or modes that stop at a specific location. Table 1 summarizes the fixed-route transit service measures from the TCQSM *(29)* and identifies those where GTFS feeds can be used as a data source.

Based on previous discussion about GTFS feeds and the methodologies discussed in the TCQSM to calculate the various metrics, the structured data from GTFS could be used to tabulate or calculate some of the measures. A review of the methods shows that two of the six measures can be calculated exclusively with GTFS feeds and the four others can be calculated using GTFS feeds with supplemental data. In general, while GTFS feeds can form part of the data needed for any of the metrics shown in Table 1, the static nature of the GTFS data makes it more effective in availability metrics, and less so for comfort and convenience metrics. As part of this demonstration, the methodology and results in this work use average headway (TCQSM measure of service availability at transit stops) to evaluate the applicability of GTFS feeds as a dataset.

Table 1 Data requirements in TCQSM analyses - Adapted from TCQSM 2nd Ed *(29)*

Quality of Service Category	Resolution	Measure	GTFS Applicable	Additional Data required
Availability	Transit Stops	Average headway	Yes	None
Availability	Route Segments/Corridors	Hours of service	Yes	None
Availability	System	Percent transit-supportive areas covered	Yes	Employment, residential densities
Comfort / Convenience	Transit stops	Passenger Load	Yes	Passenger counts
Comfort / Convenience	Route Segments/Corridors	On-time performance	Yes	Archived actual arrival times
Comfort / Convenience	System	Travel Time Difference	Yes	Traffic network

The third edition of the TCQSM was released in 2013 and expands on the second edition in a number of ways. The most relevant of those to this research is the removal of levels of service in most analyses, and a reorganization of availability concepts that rely less on the system-route-stop analysis designations, instead relying more on the direct concepts of frequency, service span and access. Another addition is the designation of an average system headway which is based on traditional route-level cycle-time calculations using data available in the NTD. *(1)* This is in contrast to the method provided in the

analysis of this work which describes a more directly calculated value based on individual headways within a system.

Most analytic activities in this thesis were conducted prior to the release of the third edition of the TCQSM; while there are few substantive impacts on the methodologies employed, readers should note some of the organizational differences such as the use of the system-route-stop level framework.

National Transit Database

The NTD requires two kinds of reporting, monthly ridership reporting and annual reporting on finances, assets, services provided, resources consumed, employment and federal funding statistics *(37)*. This work is specifically concerned with the Service Module, a set of data related to "transit service supplied by the transit agency and the transit service consumed by passengers." *(38)* The key measures of interest in the services module include metrics such as vehicles operated in maximum service, scheduled vehicle miles, vehicle revenue miles and hours, and train revenue miles and hours. The data required for these metrics are documented in the NTD Reporting Manual which is a large volume providing guidance to reporting agencies.

In general, the NTD requires information that can easily be tabulated in order to reduce the probability of misinterpretation or errors. Most of the values in the service module, for example, are sums of service data such as the time vehicles are in service according to time tables (vehicle revenue hours) or the number of hours of service in which trips for a route or system are occurring. The NTD avoids collection of more nuanced average value metrics such as average headway whose analysis could be misinterpreted easily (such as combining headways for multiple routes along a trunk line instead of considering each one separately).

CHAPTER 3

GTFS READER

This chapter summarizes the purpose and methods employed in the development of a suite of scripts that comprise a tool called the GTFS Reader. It also includes an analytic demonstration of applications for the use of GTFS feeds. The GTFS Reader is used in Chapter 4 for the validation of national performance measures.

Purpose

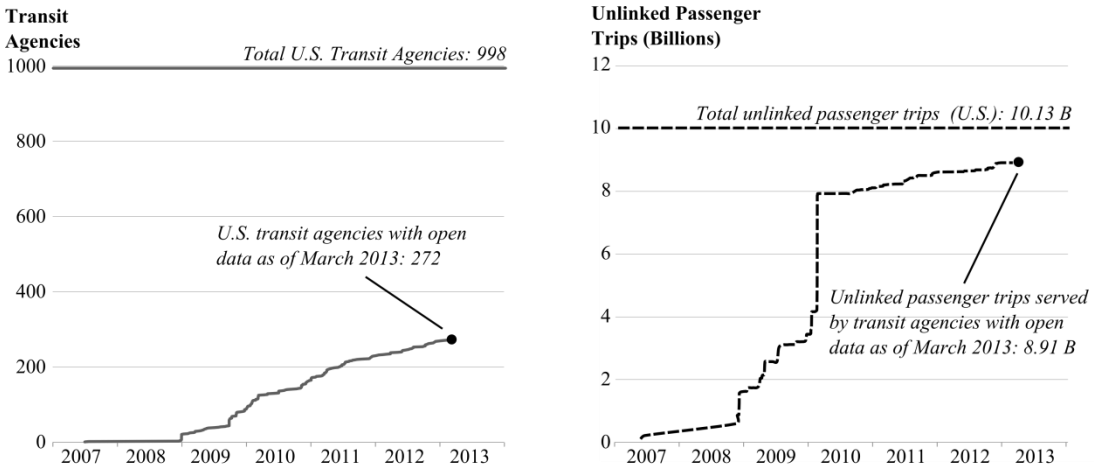
The purpose of this section is to develop a tool that can efficiently calculate performance measures from timetable and map data in GTFS feeds and demonstrate the broad capabilities of the method for expansive analysis. To do so, the author developed the GTFS Reader, a tool to read and analyze GTFS feeds in bulk. This chapter explores the availability of GTFS feeds, documents the methodologies used in the GTFS Reader, and presents the demonstrated capabilities for three kinds of analyses: an evaluation of how agencies are using GTFS feeds; in-depth headway and route-stop-density analyses for the SEPTA bus system; and a multi-agency headway comparison by mode.

Data Sources

GTFS feeds were originally produced by many agencies in order to get their transit information to display on Google Maps; Google would only accept data formatted according to the GTFS. The open data movement discussed in Chapter 2 led many agencies to posts those same feeds in publicly accessible locations. The GTFS Data Exchange (http://gtfs-data-exchange.com) is an informal but reliable website that aggregates and notifies users about updates and releases of GTFS data; it is the best source for these open feeds. The website's use of an application programming interface (API) is also useful as it provides easy access to the data on the site in JSON format.

In addition to the actual feeds, the website provides meta data about each feed including the name and location of the agency reporting, a flag for whether or not it is an official agency-provided feed, the username of the person uploading it, a referral link to any licensing requirements, the date of original feed release and the date of the latest feed update. This information is important as it helps to classify and filter those feeds that are important to the analysis.

Based on an analysis of data from the GTFS Data Exchange and the National Transit Database, although only 27 percent of the agencies in the United States have open GTFS data, these agencies represent approximately 88 percent of the unlinked passenger trips traveled nationally. The plots in Figure 3 show the rapid growth in use of GTFS based on the growing number of agencies with open data and the number of unlinked passenger trips served by those agencies. The trend is shown based on when the agency first released data according to the GTFS Data Exchange and is scaled using 2011 ridership statistics from the NTD. Such a widely adopted standard shows promise for use by researchers and analysts in areas other than trip planners and customer service tools.



(a) U.S. transit agencies with open data

(b) Unlinked passenger trips served by agencies with open data

Note: Data indexed using 2011 NTD ridership. and agency statistics

Data Source: National Transit Database 2011, City-go-Round (http://citygoround.org)

Figure 3 (a) Number of transit agencies and (b) passenger miles served by agencies with open data (as of March 2013).

GTFS feeds are .zip files made up of several individual text files. Consumer-grade computers can extract the individual text files and read them using any text editor. In this format, however, the data is not useful for an end-user as shown in Figure 1. Because of the structure of the data described in previous chapters, the easiest way to interact with and analyze a GTFS feed is to use a database manager and import the data. To that end, the primary functions of the GTFS Reader are the automation of database imports using Python and PostgreSQL, and the automation of analytic tasks using SQL queries and recording the output.

An important caveat to the analysis in this thesis is the reliance on unknown entities to validate data. The API for the GTFS Data Exchange has a flag for whether or not the feed comes from an official data source; it is unclear who authorizes the use of this flag. It is important to recognize that information about the feeds and the information in the feeds themselves are rarely endorsed officially by an agency; agencies often post their data with disclaimers about not being responsible for errors or inaccuracies. Presumably, agencies are very thorough with these datasets because they are used to guide passengers who plan trips on those systems, but errors may still occur.

Application Framework

The overall framework of the GTFS Reader involves source GTFS feeds which are used as data inputs, Python scripts which validate and import those feeds into a PostgreSQL/PostGIS database, and additional Python scripts that run a series of manipulations to data in order to calculate or tabulate performance measure outputs from those feeds. The final outputs of the GTFS Reader are recorded in CSV output files. This work flow, shown in Figure 4, was employed because it allowed for the Python scripts to send SQL queries to the PostgreSQL database, but also because it allowed the Python scripts to read back some of the results and adjust the process accordingly. For example,

an element of one import script identifies those modes which are represented in the feed and runs subsequent queries for only those specific modes.

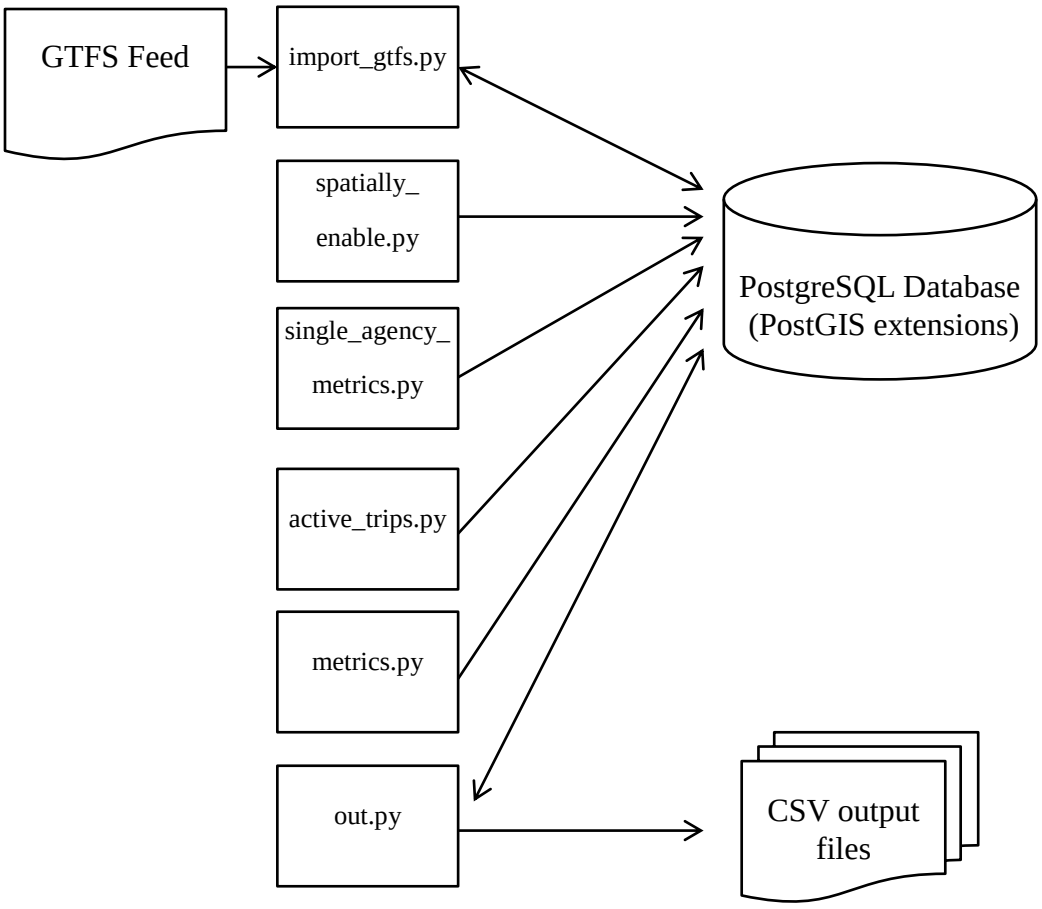


Figure 4 Application framework for GTFS Reader

The Python scripts are separated into six files: import_gtfs.py, spatially_enable.py, active_trips.py, single_agency_metrics.py, metrics.py and out.py. The main functions and analytic steps for each of these files are discussed in Table 2.

Table 2 Description of Python files and main functions in GTFS Reader

Python File	Main Functions
main.py	This is a wrapper file with one function that passes all variables needed to the functions of all selected Python modules.
import_gtfs.py	Imports GTFS data into a PostgreSQL/PostGIS database • Connects to a user-identified database • Drops any previous tables and views that may exist • Loops through the files in the zipped GTFS Feed; for those that are identified as part of the GTFS specification, creates a corresponding table in the database and inserts the data in that table • Creates additional tables that include times formatted as seconds past midnight
spatially_enable.py	Translates data into a format on which geographic queries can be run (projections, measurement, proximity…). • Creates PostGIS geographic point data from latitude/longitude of stop locations • Creates PostGIS geographic poly-line data from latitude/longitude of each point in shapes file
active_trips.py	A separate module to generate active trips by time of day. • Uses the start and end time of each individual trip to create binary indicators every five minutes of whether a trip is active; the sum of these by service_ids allows a user to see the active trips by time of day
single_agency_metrics.py	Calculates various metrics based on now-accessible schedule data. • Feed statistics: creation of a table (feed_stats) based on whether or not valid data is found in each field in the GTFS • Daily average headway: for each route-stop, the time between consecutive departures of a specific route are recorded and averaged to generate the route-stop daily average headway. • Route length/num stops: for each route, the length and number of stops are recorded in a separate table for presentation.

Table 2 Description of Python files and main functions in GTFS Reader (Continued)

Python File	Main Functions
metrics.py	Calculates various metrics based on now-accessible schedule data. • Vehicle revenue miles: geographic analysis for length of routes tabulated based on service_ids and trip departures to determine total vehicle revenue miles scheduled. • Vehicle revenue hours: duration of each trip tabulated based on service_ids and trip departures to determine total vehicle revenue hours scheduled.
out.py	Copies output of previous queries and writes them to CSV files that are saved in a local directory.

Analytic Demonstrations for Single and Multi-Agency Analysis

The three analytic demonstrations provided in this chapter are an evaluation of the fields used by US agencies with open GTFS feeds, an agency specific analysis of the SEPTA bus system and a comparison of headways among the 50 largest transit agencies with GTFS feeds available on the GTFS Data Exchange.

Evaluation of GTFS Usage

GTFS uses a data structure designed for easy generation by transit providers and practical use by programmers. Many fields are optional, providing flexibility to agencies with different service patterns, scheduling procedures and technical staff availability. Programmers that develop software based on GTFS data quickly realize that agencies may or may not use certain fields which will impact the design of transit rider tools. Likewise, to use these datasets for comparative research among multiple agencies, it is useful to understand how many agencies use each field. As of November 2012, there were 211 distinct feeds available from agencies and transit providers in the United States from the GTFS Data Exchange (this does not include approximately twenty transit

service providers not represented on that website but that show up on Google's own list of agencies).

The Python script (feedstats.py) discussed in Table 2 is designed to parse GTFS feeds and report which of the required and optional fields are being utilized. The information in Table 3 reflects usage statistics from the feeds**.** Those tables and fields that the GTFS documentation calls "required" (shown in the table with a "●") should generally have a 100 percent usage rate. In some cases, the GTFS documentation allows required fields to be omitted (see table notes). For optional tables, required fields are only needed when the table is used. An important caveat is that while the GTFS documentation specifies how to write these files, there is no guarantee that the feed developed by an agency and provided for public consumption conforms to that format. Researchers, like programmers, should be sure to validate feeds to ensure the fields needed for their analysis are utilized correctly.

Many of the optional fields have very low usage rates which imply that future research design that uses multiple GTFS feeds as a data source should be cautious in the use of these fields as many agencies do not use them. In particular, those fields associated with the fare_attributes, fare_rules, frequencies, transfers and feed_info tables have low usage rates. In some cases, recent changes to the specification resulted in new fields that lead to low indications of low usage (wheelchair_boarding and wheelchair_accessible are two examples). The usage of these fields will rise as agencies update their feeds to conform with the latest changes to the specification.

Table 3 GTFS table and field usage for open GTFS feeds

File Name	Field Name	Usage
● agency.txt	○ agency_id	83%
	● agency_name	100%
	● agency_url	100%
	● agency_timezone	100%
	○ agency_lang	51%
	○ agency_phone	80%
	○ agency_fare_url	17%
● stops.txt	● stop_id	100%
	○ stop_code	30%
	● stop_name	100%
	○ stop_desc	42%
	● stop_lat	100%
	● stop_lon	100%
	○ zone_id	43%
	○ stop_url	8%
	○ location_type	46%
	○ parent_station	9%
	○ stop_timezone	<1%
	○ wheelchair_boarding[2]	3%
● routes.txt	● route_id	100%
	○ agency_id	73%
	● route_short_name[3]	72%
	● route_long_name[3]	95%
	○ route_desc	33%
	● route_type	100%
	○ route_url	55%
	○ route_color	55%
	○ route_text_color	48%
● trips.txt	● route_id	100%
	● service_id	100%
	● trip_id	100%
	○ trip_headsign	85%
	○ trip_short_name	12%
	○ direction_id	60%
	○ block_id	60%
	○ shape_id	80%
	○ wheelchair_accessible[2]	1%
● stop_times.txt	● trip_id	100%
	● arrival_time	100%
	● departure_time	100%
	● stop_id	100%
	● stop_sequence	100%
	○ stop_headsign	16%
	○ pickup_type	71%
	○ dropoff_type	69%
	○ shape_dist_traveled	44%

File Name	Field Name	Usage
● calendar.txt[4]	● service_id	96%
	● monday	96%
	● tuesday	96%
	● wednesday	96%
	● thursday	96%
	● friday	96%
	● saturday	96%
	● sunday	96%
	● start_date	96%
	● end_date	96%
○ calendar_dates.txt[4]	● service_id	84%
	● date	84%
	● exception_type	84%
○ fare_attributes.txt	● fare_id	54%
	● price	54%
	● currency_type	54%
	● payment_method	54%
	● transfers	35%
	○ transfer_duration	20%
○ fare_rules.txt	● fare_id	45%
	○ route_id	32%
	○ origin_id	19%
	○ destination_id	17%
	○ contains_id	3%
○ shapes.txt	● shape_id	83%
	● shape_pt_lat	83%
	● shape_pt_lon	83%
	● shape_pt_sequence	83%
	○ shape_dist_traveled	48%
○ frequencies.txt	● trip_id	26%
	● start_time	26%
	● end_time	26%
	● headway_secs	26%
	○ exact_times	22%
○ transfers.txt	● from_stop_id	26%
	● to_stop_id	26%
	● transfer_type	25%
	○ min_transfer_time	4%
○ feed_info.txt	● feed_publisher_name	34%
	● feed_publisher_url	34%
	● feed_lang	34%
	○ feed_start_date	2%
	○ feed_end_date	2%
	○ feed_version	3%

Note 1: ● = Required, ○ = Optional.

Note 2: These fields were added to the specification within six months before the analysis.

Note 3: In some cases, feeds may use either route_short_name *or* route_long_name.

Note 4: Calendars.txt may be omitted in certain feeds that use calendar_dates.txt.

Single Agency Analysis

Daily Average Headway

The TCQSM directs practitioners to evaluate average headway at transit stops and stations, separately for each route *(29)*. This is accomplished by calculating the time difference between arrivals at a stop for each successive arrival of a particular route. The calculation is repeated for each route at each stop. Per guidance in the TCQSM, headways less than three minutes[2] (typical of school dismissal times) were ignored, as were headways longer than 90 minutes which researchers assumed was a break in service.

The histogram in Figure 5 illustrates the frequency distribution of daily headways for each route-stop in the SEPTA bus system, evaluated on typical weekdays in five-minute increments. The axis along the top of Figure 5 shows the level of service guidelines from the TCQSM for fixed-route service frequency. As a method of aggregation for multi-agency comparisons, the headways for each route-stop were recorded and averaged for a typical weekday. The average of this selection of headways is 31.3 minutes (the remainder of this chapter will refer to this statistic as an agency-average headway).

[2] Guidance in the second edition instructs users to ignore headways less than three minutes for the purpose of determining service frequency level of service *(29)*. This was removed in the third edition where the method of calculating service frequency is left for the user to determine. *(1)*

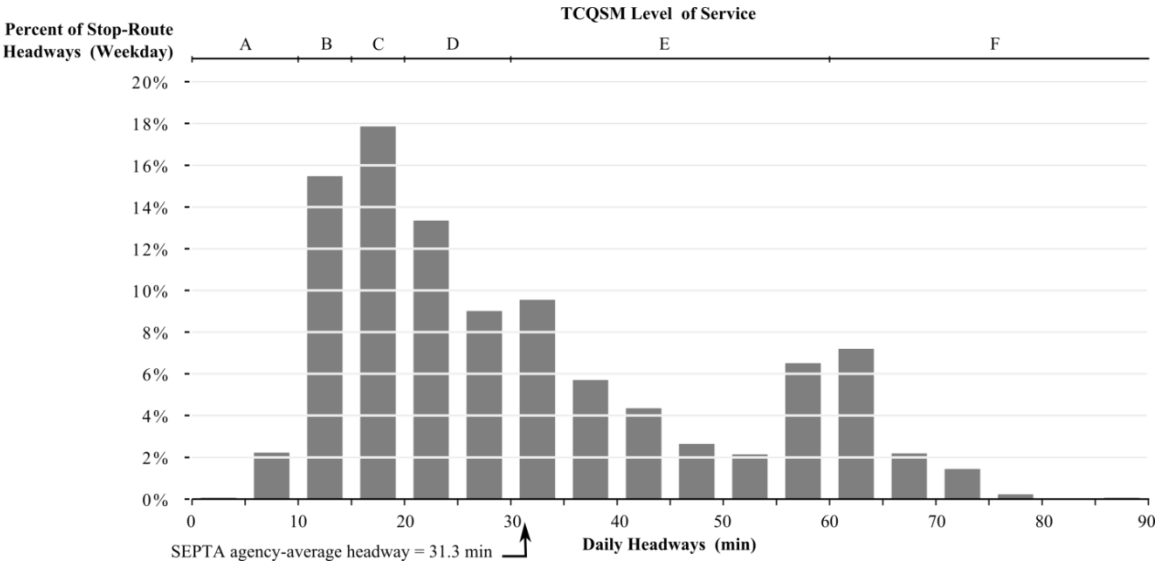


Figure 5 Distribution of stop-route level daily headways for the SEPTA bus system.

Route Length and Stop Density

If an analyst were asked to report the length and number of stops for all the routes in a transit system, it would be an unclear or poorly stated question. There could be routes with express configurations, routes that only serve particular stops on weekends and routes with several branches off a trunk line, all with the same route name. In this example of route-level analysis, special attention is paid to the intricacies of working with GTFS feeds that include these different configurations. The schema used in GTFS has certain flexibility so that a single route_id might represent different configurations of stops. To overcome these intricacies, each data point represents the average of the length and number of stops for every trip sharing a single route ID. The author recognizes that this method of aggregation hides certain details, but chose to do so as an example of one method to summarize data using aggregation. It is important that any analyst engaging in use of GTFS data analysis become familiar with the different coding permutations that agencies choose before writing queries or reconfiguring data to represent operational summary statistics.

The results of the route length and stop frequency analysis are shown in Figure 6. Agencies can use information like this to quickly identify stops with abnormal trip patterns like very dense stop placement or excessively long routes. Notice that some of the routes have long lengths but very few stops; this suggests the presence of commuter routes which may have several stops near the beginning and end with express service along freeways.

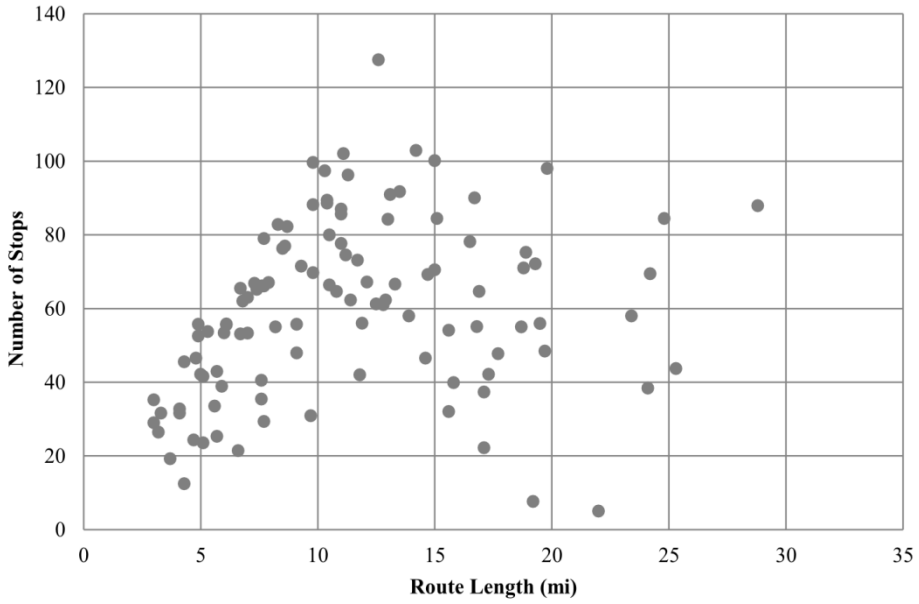


Figure 6 Length and number of stops for SEPTA bus routes.

In addition to this format, the data can also be categorized based on the derived distance-between-stops (calculated as the quotient of route length and number of stops) as shown in Figure 7. This histogram identifies the bulk of routes that have stops spaced less than a quarter-mile apart, common in dense urban cores such as in Philadelphia. Basic visualizations like these are the result of data insights that can be made efficiently once GTFS feeds are put into an accessible database format.

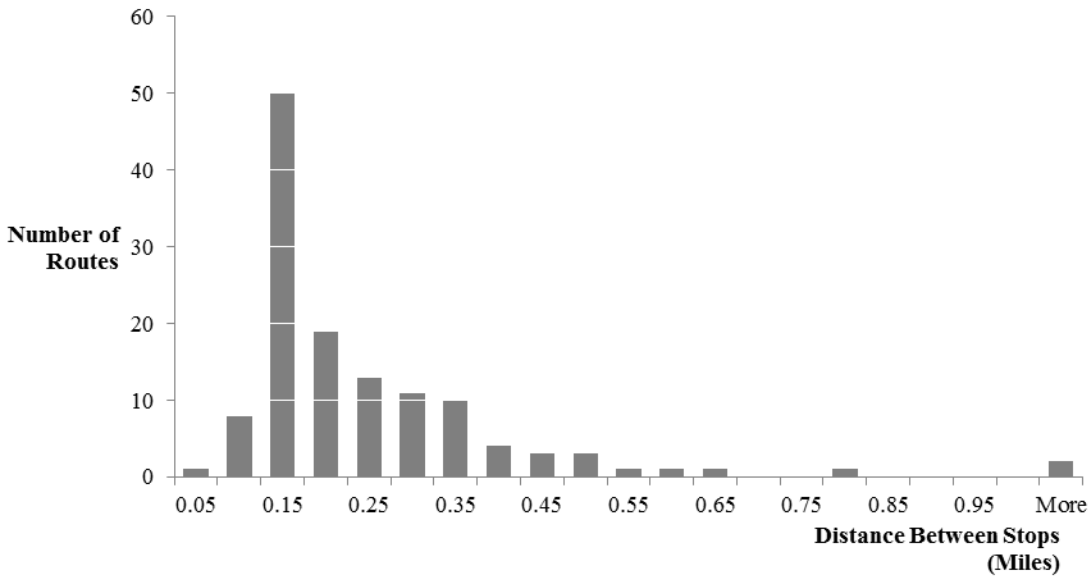


Figure 7 Histogram of Route-level Distance Between Stops

Multi-Agency Analysis

While agencies are more often interested in the details of their own services, researchers and national policy experts will find it useful to have the ability to efficiently compare data from multiple transit providers. The application chosen for this project is thought to be of interest to those considering revisions to the TCQSM LOS methodology. In this analysis, the agency average headway, discussed earlier, is calculated for four fixed-route mode categories of 50 large agencies in North America that provide open transit data. A list of those agencies and the modes available are shown in Table 4. GTFS defines these mode categories as a user perceives them rather than using their operational and traction characteristics as suggested by Vuchic *(2)*. The four mode categories used in this analysis are taken from the description of GTFS:

- Light rail, Tram, Streetcar. Any light rail or street level system within a metropolitan area.
- Subway, Metro. Any underground rail system within a metropolitan area.
- Rail. Used for intercity or long-distance travel.
- Bus. Used for short- and long-distance bus routes. *(18)*

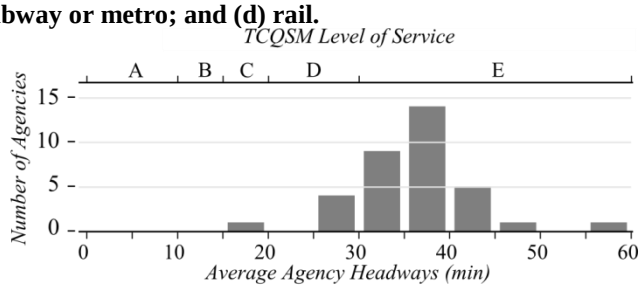
Using data from agencies as available in Table 4, the author ran each of the disaggregated feeds through the data processes described in Figure 4. The output was a series of reports for each agency which were then aggregated using R. The agency-average headway for each feed was recorded and is shown in the frequency distributions in Figure 5. In the end, the simplified histograms in Figure 5 represent in-depth analysis with a data point for every time a transit vehicle arrives at any stop in every one of the 50 agencies analyzed. This demonstrates the value of batch processing using the methods from this thesis because until now, there has been no efficient way to analyze these statistics quickly among multiple transit providers without significant labor requirements.

Table 4 Availability of GTFS Feeds at 50 Large North American Transit Agencies by Mode (July 2012).

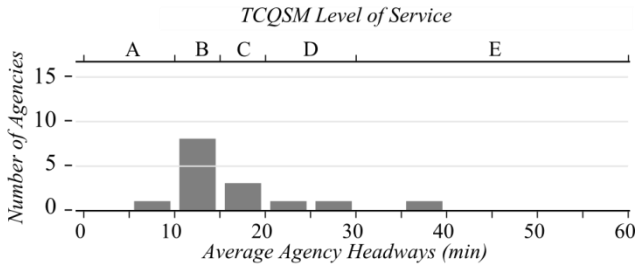
Agency	City, State	Bus	Light Rail	Subway	Rail
Alameda-Contra Costa Transit District	Oakland, CA	●			
San Francisco Bay Area Rapid Transit District	Oakland, CA	●		●	
Broward County Transportation Dept.	Pompano Beach, FL	●			
Peninsula Corridor Joint Powers Board	San Carlos, CA				●
Capital Metropolitan Transportation Authority	Austin, TX	●			●
City of Detroit Dept. of Transportation	Detroit, MI	●			
Chicago Transit Authority	Chicago, IL	●		●	
Dallas Area Rapid Transit	Dallas, TX	●	●		●
The Greater Cleveland Regional Transit Authority	Cleveland, OH	●	●	●	
Transportation District Commission of Hampton Roads	Hampton, VA	●	●		
Hillsborough Area Regional Transit Authority	Tampa, FL	●			
Kansas City Area Transportation Authority	Kansas City, MO	●			
King County Dept. of Transportation	Seattle, WA	●	●		
Lane Transit District	Eugene, OR	●			
MTA Long Island Bus	Garden City, NY	●			
Long Island Railroad	Jamaica, NY				●
Maryland Transit Administration	Baltimore, MD	●	●	●	●
Massachusetts Bay Transportation Authority	Boston, MA	●	●	●	●
Northeast Illinois Regional Commuter Railroad Corp.	Chicago, IL				●
Southern California Regional Rail Authority	Los Angeles, CA				●
L.A. County Metropolitan Transportation Authority	Los Angeles, CA	●	●	●	
Metro-North Commuter Railroad Company	New York, NY				●
Metropolitan Transit Authority of Harris County, Texas	Houston, TX	●			
Bi-State Development Agency	St. Louis, MO	●			●
Metro Transit	Minneapolis, MN	●	●		●
Madison County Transit District	Granite City, IL	●			
Miami-Dade Transit	Miami, FL	●	●		
Milwaukee County Transit System	Milwaukee, WI	●			
Ride-On Montgomery County Transit	Rockville, MD	●			
MTA New York City Transit	New York, NY			●	●
San Diego Metropolitan Transit System	San Diego, CA	●	●		
New Jersey Transit Corp.	Newark, NJ		●		●
Niagara Frontier Transportation Authority	Buffalo, NY	●	●		
North County Transit District	Oceanside, CA	●			●
Orange County Transportation Authority	Orange, CA	●			
Pace - Suburban Bus Division	Arlington Heights, IL	●			
Port Authority of Allegheny County	Pittsburgh, PA	●	●		
Pinellas Suncoast Transit Authority	St. Petersburg, FL	●			
Regional Transportation Commission of S. Nevada	Las Vegas, NV	●			
Denver Regional Transportation District	Denver, CO	●	●		
Sacramento Regional Transit District	Sacramento, CA	●	●		
San Francisco Municipal Railway	San Francisco, CA	●	●		
Southeastern Pennsylvania Transportation Authority	Philadelphia, PA	●	●	●	
Central Puget Sound Regional Transit Authority	Seattle, WA	●	●		
Spokane Transit Authority	Spokane, WA	●			
City & Co. of Honolulu Dept. of Transportation Svcs.	Honolulu, HI	●			
Tri-County Metro. Transportation District of Oregon	Portland, OR	●	●		●
Utah Transit Authority	Salt Lake City, UT	●	●		●
VIA Metropolitan Transit	San Antonio, TX	●			
Washington Metropolitan Area Transit Authority	Washington, DC	●		●	

Figure 8 Distribution of agency-average headways for (a) bus; (b) light rail; tram or streetcar; (c) subway or metro; and (d) rail.

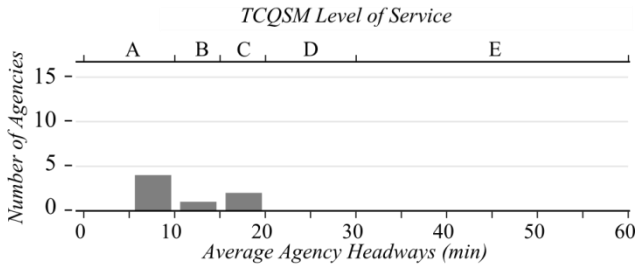
(a) Bus



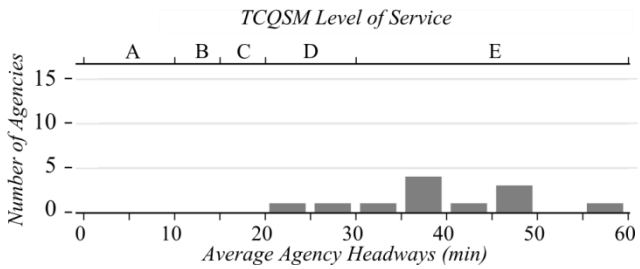
(b) Light rail, Tram, Streetcar



(c) Subway, Metro



(d) Rail



Lessons Learned Working with GTFS Data

Perhaps some of the most important findings from this work are identifying the intricacies of working with GTFS data so that future researchers are aware of their existence.

Information in the GTFS feed is provided at a granular level with comprehensive coverage of an entire system all the way down to the stop times for each scheduled trip. GTFS feeds have far greater resolution than the NTD which only provides summary data for each agency. Using GTFS is helpful for in-depth analysis of specific metrics, but can be cumbersome for analysts awash in data about a transit system. Particular attention should be paid to avoiding misrepresenting aggregation procedures which will quickly accumulate when building statistics that use stop-level data to summarize system-level metrics. Because of the many ways that data can be summarized, this poses a challenge for those generating or interpreting performance measures which are intended to be clear, concise representations of information. As an example, an 'average route headway' for a route might be a summary of each individual interarrival time of all trips at all stops, or it might be a summary of the interarrival times at one representative stop along a route (ignoring the effects of route branches). In both instances, the nuance of calculation should be better described in the metric than 'average route headway.'

GTFS feeds are usually provided by agency, rather than by region or geography. Depending on the requirements of a user, it is important to take this organization into account. For example, a single transit agency might be interested in evaluating operations within its own service area which can be effectively evaluated using their own GTFS feed; a metropolitan planning organization might be more interested in the regional coverage of transit service which would be best served by combining the feeds of multiple agencies in the region and evaluating them without regard for the specific agency providing the service.

GTFS feeds can typically be kept current as there is an expiration date coded into the calendars.txt file. Some agencies may choose to make this sufficiently far in the future that it effectively doesn't expire. Analysts and researchers must decide whether or not the data is current based on those dates as well as the communication channels that should be in place for consumers of the data. Additionally, the rate of updates to the feed should be kept in consideration as some agencies actually release their GTFS feeds on a daily basis while others may only do so bi-annually or less. There may be a conflict between the time of feed publication and the validity of the feeds, causing either a lapse in valid data or confusion about which is more accurate. An extensive evaluation of the historic availability of data on the GTFS Data Exchange including a discussion of when feeds are valid can be found in Chapter 4.

Building tools for multiple agencies should be done carefully by individuals who are familiar with GTFS feeds. For example, a query may work for one agency's GTFS feed because it uses the agency_id field; that field, however, is optional and may not work for agencies that do not use that field. The information in Table 3 will be helpful to those developing applications for multiple users.

Coding practices for GTFS vary among agencies. While GTFS has specific field names and data formats, the way that agencies use those fields still varies considerably. The following observations are important to, and best understood by, individuals working closely with GTFS data:

- Schedule configurations, represented by service_id, are neither mutually exclusive nor exhaustive. They are defined by the day of the week that they are active and a date range for validity. There may be multiple active service_ids at any one time. For any attempt to recreate actual service scheduled on certain days as in vehicle revenue hours per year, it is best to design applications as a user on each specific date in question and pull the relevant information for that day (as opposed to using date ranges and validity options).
- Since GTFS is not strictly designed as a relational database, the concept of primary and foreign keys is not preserved. One-to-one or one-to-many

relationships can exist between fields like route_id and shape_id which will affect file size and consistency of data.

- Depending on how the feed was generated, it may include only information for time points from a schedule, or it may include specific times for every stop. Calculations of headway or other statistics must be sensitive to the fact that data may appear missing. According to the GTFS documentation, agencies should not interpolate schedules where they have no data, but some still do.
- Agencies can use either schedule-based or frequency-based coding and will use different tables accordingly; the queries used in this project were designed for schedule-based systems.
- Stops and stations may be coded at the intersection level or more precisely by location and direction. Consider a northbound route that crosses an eastbound route; this is usually coded as one stop for rail systems with transfer points, but may be coded as one or two separate stops for bus routes where they are separate facilities in close proximity to one another.
- Transit modes are defined in GTFS based on user-oriented categories rather than operational and traction characteristics (for example, light rail and streetcars are coded as the same).
- Different text encoding in the .txt files of a GTFS feed (using UNICODE or UTF-8, for example) can pose challenges for some scripting languages.

CHAPTER 4

VALIDATING GTFS FEEDS FOR TRANSIT ANALYSIS USING THE NATIONAL TRANSIT DATABASE

The previous chapter discussed a process for importing GTFS feeds into a database and the set of queries that calculated metrics at the stop, route and system level. The GTFS Reader was shown to be an efficient process to provide insight into how agencies schedule and supply their services. The validation exercise documented in this chapter uses the GTFS Reader framework to process multiple feeds and to calculate two metrics that agencies already report annually to the National Transit Database (NTD): annual vehicle revenue hours (AVRH) and annual vehicle revenue miles (AVRM). The calculation of those metrics requires a thorough understanding of both the internal structure of GTFS and the process for aggregating metrics from the trip-level to the system-level over time. These concepts and the resulting analysis are presented in this chapter.

Purpose

The purpose of the research in this chapter is to compare metrics calculated from raw GTFS feeds to those reported in the NTD.

Methodology

To calculate the metrics from GTFS feeds for comparison to NTD metrics, the comparison process employs a modified version of the GTFS Reader (presented in the previous chapter). Instead of calculating headways and stop density, the GTFS Reader is used to calculate AVRM and AVRH. The GTFS Reader can process one GTFS feed at a time, but can be quickly scaled to analyze and generate outputs for a series of GTFS feeds. The revised GTFS Reader is shown in Figure 9 where NTDmetrics.py is used in

lieu of single_agency_trips.py, active_trips.py and metrics.py. The application imports a GTFS feed, adapts it for spatial analysis, calculates NTD metrics (AVRH and AVRM) and finally saves the output.

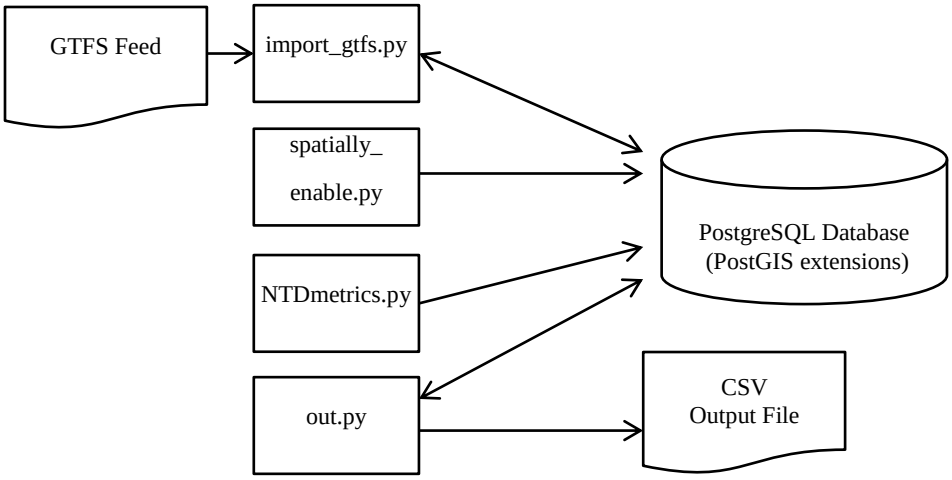


Figure 9 GTFS Reader Framework using NTD Metrics Module

According to the NTD Reporting Manual, revenue service includes both running time and layover/recovery time, which typically ranges from 10 to 20 percent of running time. *(38)* The general impact of layover/recovery time on vehicle revenue *hours* is thus 10 to 20 percent of running time; the impact on vehicle revenue *miles*, however, is negligible under the assumption that vehicles do not traverse a significant distance during a layover. For example, a transit vehicle that lays over at the end of a linear route by waiting at the last stop and turning around will accrue additional time in vehicle revenue hours during the layover time, but it will only accrue the distance to physically turn around for vehicle revenue miles. The consequence of this difference on the validation methodology is that AVRM from the NTD can be compared directly to the AVRM calculated from GTFS, but AVRH of the NTD are expected to fall 10 to 20 percent higher than the AVRH calculated from GTFS.

Calculating Daily NTD Metrics

The NTDmetrics.py module implements the process of calculating vehicle revenue hours and miles from the raw timetable information contained in GTFS feeds. The methodology is such that the NTD metrics are compared to values calculated from GTFS feeds. By setting it up this way, the methodology is actually checking both the GTFS data and the method of aggregation employed. To that end, the following includes a thorough discussion of the method employed to calculate AVRH and AVRM for one mode of an agency using a single GTFS feed.

The diagram in Figure 10 summarizes the process of calculating daily vehicle revenue hours (DVRH) and daily vehicle revenue miles (DVRM). It shows the GTFS tables used in raw format, the queries that transact with the database, and intermediate tables that store values for use in other steps.

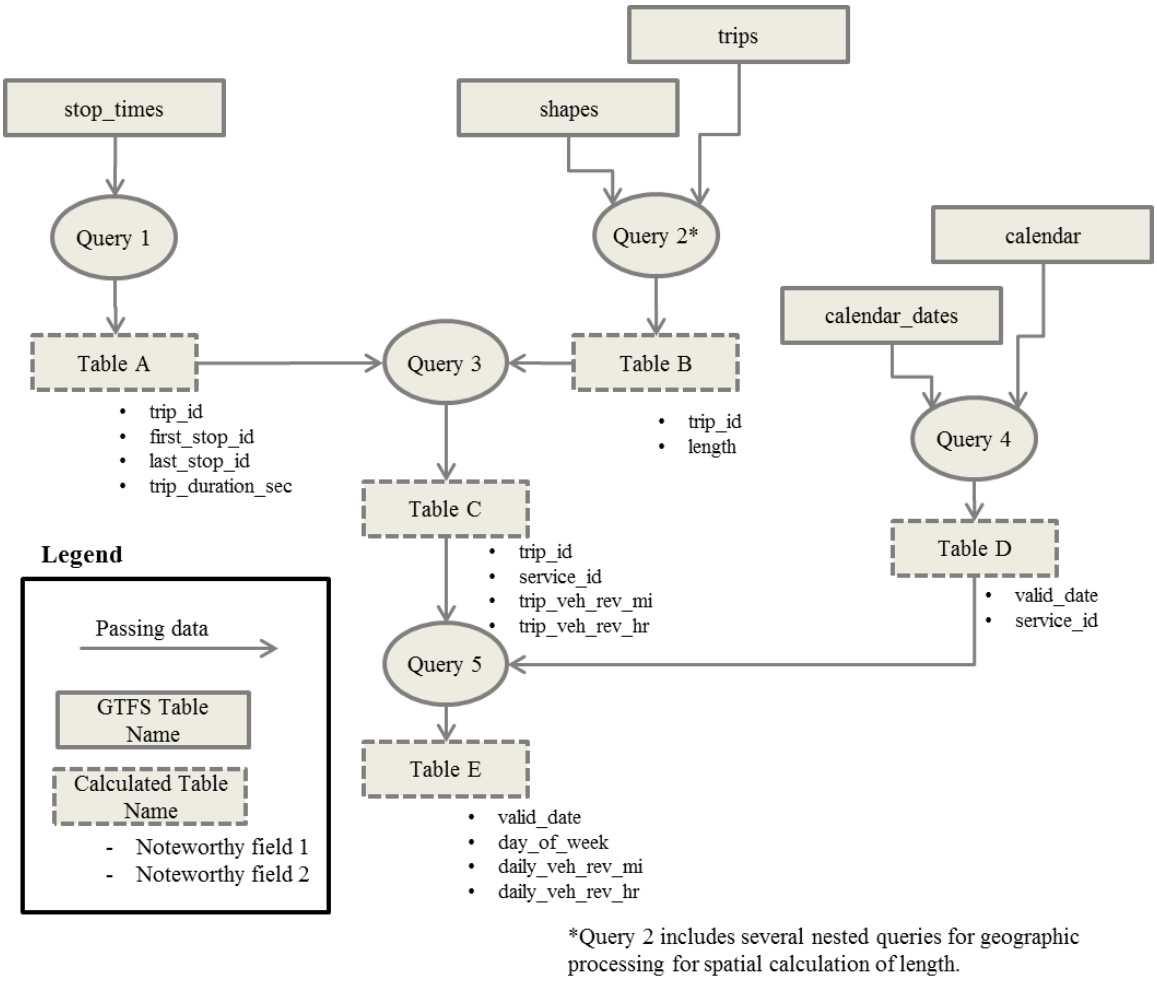


Figure 10 Daily NTD Metric Calculation

The overall process calculates vehicle revenue miles and hours for each trip, then aggregates those trip characteristics as appropriate on a set of specific dates synthesized based on the feed's validity period. Query 1 calculates vehicle revenue hours for each trip as the difference in seconds between the first departure and last arrival of that trip. This is considered the time that the vehicle was scheduled to be in revenue service, although the previous discussion clarifies that this is actually running time. Query 2 calculates vehicle revenue miles using the shapes table to generate geographic poly-lines for each shape_id. The length of each shape is calculated using the spatial extension of PostgreSQL, PostGIS. Because this analysis is designed for general application in any location, the

global WGS84 (World Geodetic Survey of 1984) is used without a projection; distance is measured assuming Earth is a spheroid and was scaled from meters to miles.

The actual mileage for each trip is counted as the length along each trip shape between a point on the line closest to the first stop and a point on the line closest to the last stop (shown in Figure 11). This resolves potential over-counting that can result from route shapes that extend beyond the first and last stops of the route. In the event that a stop exists beyond the end of a shape, the maximum length of the shape is used. The result is a conservative estimate of vehicle mileage that errs on the side of fewer miles per route. A brief review of the impact of this on bus routes operated by Portland's TriMet shows that the average shortening of the trip shape is 4.3 percent of the trip shape length. This is a known source of potential error in the completed aggregation of AVRM.

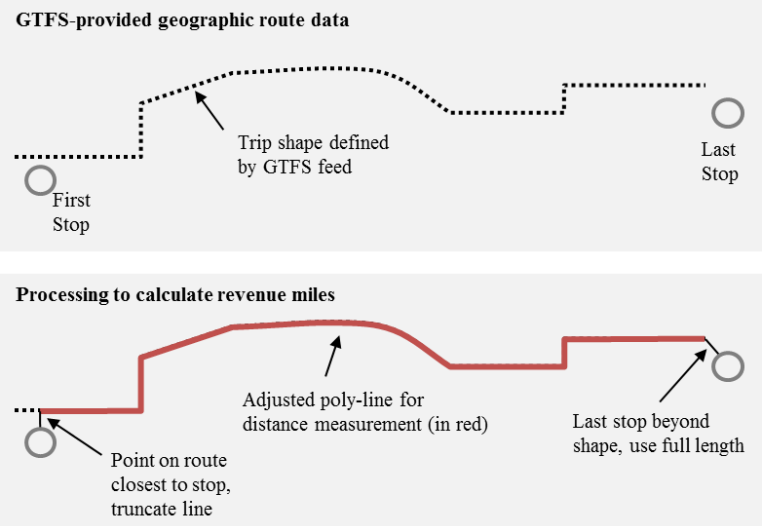


Figure 11 Process for adjusting routes for revenue mile calculation

At this point in the process, each trip is associated with both revenue hours and revenue miles leading into Query 3. Google describes the calendar.txt file as: "Dates for service IDs using a weekly schedule. Specify when service starts and ends, as well as days of the week where service is available." *(18)* A service_id represents a typical weekday schedule; it is defined by the days of the week on which it operates and two

dates between which the service_id is valid. Trips are uniquely associated with service_ids. This allows applications to know which trips to invoke on particular days of the week within a service_id's valid date range. The relationship between the calendar.txt and trips.txt tables is shown in Figure 12.

In the example shown here, an agency has two schedules included in one GTFS feed: Winter and Spring. Each schedule has weekday and weekend service. This information is shown in the calendar.txt table. On special holidays, like the Fourth of July, the agency will run weekend service as shown in the calendar_dates.txt table. Both calendar.txt and calendar_dates.txt are taken directly from the GTFS feed. At this point in the larger process, Table C contains information for each trip and its vehicle revenue miles and hours. Service_id and trip_id share a one-to-many relationship; many trip_ids may have the same service_id, but each trip_id is associated with only one service_id. The combination of the calendar.txt, calendar_dates.txt and Table C yields the output in Table E. (This is a summary of the processes shown in Queries 4 and 5). Notice that the schedules change between June 30 and July 1, that weekend and weekday schedules are respected, and that the weekend spring service operates on the Fourth of July.

calendar.txt

service_id	mon	tues	wed	thur	fri	sat	sun	start_date	end_date
MF-Winter	1	1	1	1	1	0	0	Jan 1, 2012	June 30, 2012
SaSu-Winter	0	0	0	0	0	1	1	Jan 1, 2012	June 30, 2012
MF-Spring	1	1	1	1	1	0	0	July 1, 2012	Dec 31, 2012
SaSu-Spring	0	0	0	0	0	1	1	July 1, 2012	Dec 31, 2012

Table C

trip_id	service_id	veh_rev_mi	veh_rev_hr
001	MF-Winter	2.5	.75
002	MF-Winter	2.5	.75
003	SaSu-Winter	2.5	.5
004	SaSu-Winter	2.5	.5
101	MF-Spring	2.5	.75
102	MF-Spring	2.5	.75
103	SaSu-Spring	2.5	.5

calendar_dates.txt

service_id	date	Exception_type
MF-Spring	July 4, 2012	Remove
SaSuSpring	July 4, 2012	Add

Table E (shown with explanatory columns in italics)

Synthesized Date	Day of Week	*Active service_id(s)*	*Trip(s) Served*	daily_veh_rev_mi	daily_veh_rev_hr
…	…	…	…	…	…
June 29, 2012	Friday	*MF-Winter*	*001, 002*	5	1.5
June 30, 2012	Saturday	*SaSu-Winter*	*003, 004*	5	1
July 1, 2012	Sunday	*SaSu-Spring*	*103*	5	1
July 2, 2012	Monday	*MF-Spring*	*101,102*	5	1.5
July 3, 2012	Tuesday	*MF-Spring*	*101,102*	5	1.5
July 4, 2012	Wednesday	*SaSu-Spring*	*103*	5	1
July 5, 2012	Thursday	*MF-Spring*	*101,102*	5	1.5
…	…	…	…	…	…

Figure 12 Aggregation method for daily metrics on specific synthesized dates

Internal GTFS Data Consistency

The documentation provides limited guidance to users about how to handle multiple schedules. Because of that, some feeds are created in ways that abide by the specification's format, but incorrectly describe a schedule that doesn't reflect actual transit operations. In the example discussed in Figure 12, the service_ids begin and end on adjacent dates. In the example, the service_ids are mutually exclusive. If all service_ids are shown for a transit service that the GTFS feed is supposed to represent, then it is also collectively exhaustive. This is not always the case, however. Figure 13 describes how the calendar.txt table and use of service_ids can lead to inconsistent data.

Scenario A is considered the ideal format; an agency has a May and June schedule, and has different weekday and weekend service. This combination yields four service_ids. The table in Scenario A is how it would be represented in a GTFS feed's calendar.txt file. The reader should note that these are mutually exclusive (no overlap) and collectively exhaustive (assuming there are only weekday and weekend service types). In Scenario B, the MF-May schedule extends through June 1 (instead of May 31) leading to duplicate data on June 1. The effective result describes a day in which both the May and June weekday schedules are active; this would lead to twice as many trips as there should be. This data is not mutually exclusive. Lastly, Scenario C occurs when service_ids are active for different lengths of time. In Scenario C, the weekend schedule runs throughout the full year but the May schedule only runs during May. The result is that days in June are only partially represented, failing to be collectively exhaustive.

Figure 13 Potential scenarios for calendar.txt and service_id usage

GTFS feed publication on GTFS Data Exchange

Table E in Figure 12 is the last table shown in the process in Figure 10. It has a list of all the dates during which the GTFS feed self-identifies as valid (based on the service_id start and end dates). In an ideal setting, a single GTFS feed will be valid for an entire fiscal year in order to compare the sum of daily metrics during that fiscal year to the AVRH and AVRM found in the NTD. Unfortunately, the general practice among those releasing GTFS feeds is inconsistent from one agency to another and often includes other procedures that complicate the process. As a reminder, all the GTFS feeds used in this analysis are from the GTFS Data Exchange. Each agency contributes voluntarily and there are no endorsements by the website itself that information is being published according to the GTFS standard; it is simply a clearinghouse where self-identified agencies can make their data available for developers and the public to access.

Despite the high number of contributing agencies and the ease of access from a programming perspective, a significant hurdle had to be overcome in determining which GTFS feeds were considered usable. As discussed earlier, the feeds may have meta data embedded in them to provide a valid date range, but as shown earlier in Table 3 GTFS table and field usage for open GTFS feeds), those fields are almost never used. Instead of relying on that meta data, a proxy analysis was developed to assess how current a GTFS feed is using upload frequency to the GTFS Data Exchange.

Each time a GTFS feed is uploaded to the GTFS Data Exchange, the timestamp is noted and saved on the site's archives. By reviewing the consecutive upload dates by each agency, it was clear that some agencies have not been revising their GTFS feeds at regular intervals. Recall that while the data in GTFS feeds represents 'static' information like schedules, these may change over time as schedules themselves are updated. Figure 14 shows each agency's average number of days between updates with a maximum update interval of one week (assuming repeated updates within that time were to fix errors or as part of a daily upload protocol). There are 27 agencies that were eliminated

from this figure because their last update was before January 1, 2012 and are assumed defunct. With those exceptions noted, agencies provide GTFS feed updates to this website every 81.3 days or about every three months. This represents an active community and an ongoing commitment to ensure that GTFS data is up-to-date. It is also consistent with anecdotal knowledge that large agencies review and update transit schedules on a quarterly basis.

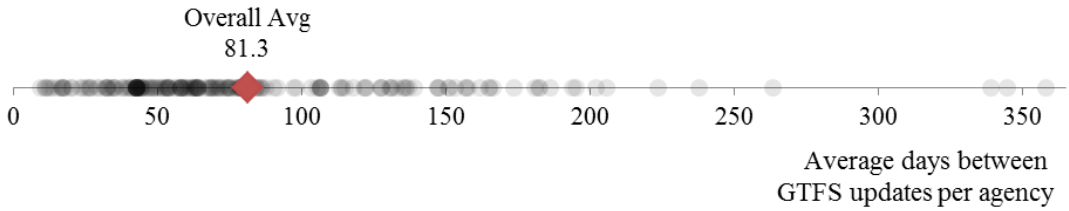


Figure 14 Rate of GTFS feed update by agency from GTFS Data Exchange

With GTFS feeds updated regularly, an analyst may benefit from a rich history of changes to service (assuming that each subsequent feed publication represented a change in service). As an example, the daily aggregation method discussed in Figure 10 was applied to a sample of 26 GTFS feeds that TriMet, the transit agency in Portland, Oregon, published on the GTFS Data Exchange. This translates to a new feed approximately every two weeks. Using the latest available data for each date in history, the author constructed a view of daily vehicle revenue hours shown in Figure 15.

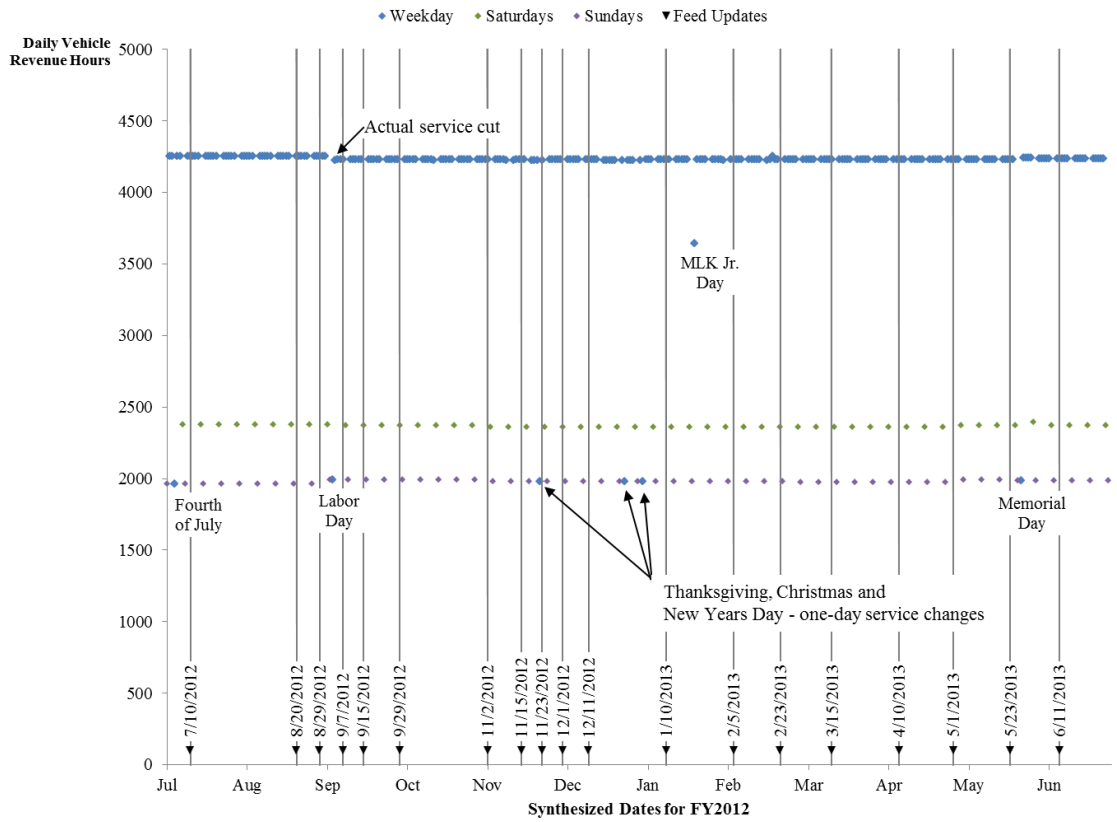


Figure 15 Daily Vehicle Revenue Hours for TriMet Buses in FY2012

In this figure, the dots represent weekday, Saturday and Sunday daily revenue hours; there is a consistent weekday-weekend pattern throughout the year with selected holidays highlighted in the chart. The data for each synthesized date is calculated from the most recent feed published (shown as grey vertical lines). For example, in early August 2012, the latest feed was from July 10, 2012; even if data from a previous feed published in June were valid for August, the July data was considered more accurate. If a feed from September had been published with valid data for August (would have occurred before its publication date), that data was discarded. An actual service cut occurred in early September 2012, which is seen in the data.

Although this daily aggregation method of building a composite dataset for a full fiscal year is ideal as shown in Figure 15, there are a number of complicating factors in

the way agencies release data that prove challenging for their use in this way. A number of scenarios are shown in Figure 16 to illustrate types of issues that were encountered while preparing this thesis. In the hypothetical situation shown, a transit agency has three distinct schedules throughout the year. Scenario I is the ideal feed release schedule where each feed is released shortly before a schedule becomes active and the feed is valid only during that specific schedule. The balance of publication and validity is more often related to the agency's approach to open data; if an agency trusts developers to check daily for updates, the agency may only need to publish on the dates when something changes. Agencies that don't trust developers to check for updates may put a tentative Schedule B in the feed when they release Schedule A. The balance identified earlier can also be thought of as an agency's preference for developers to have no data on a specific date or out-of-date data.

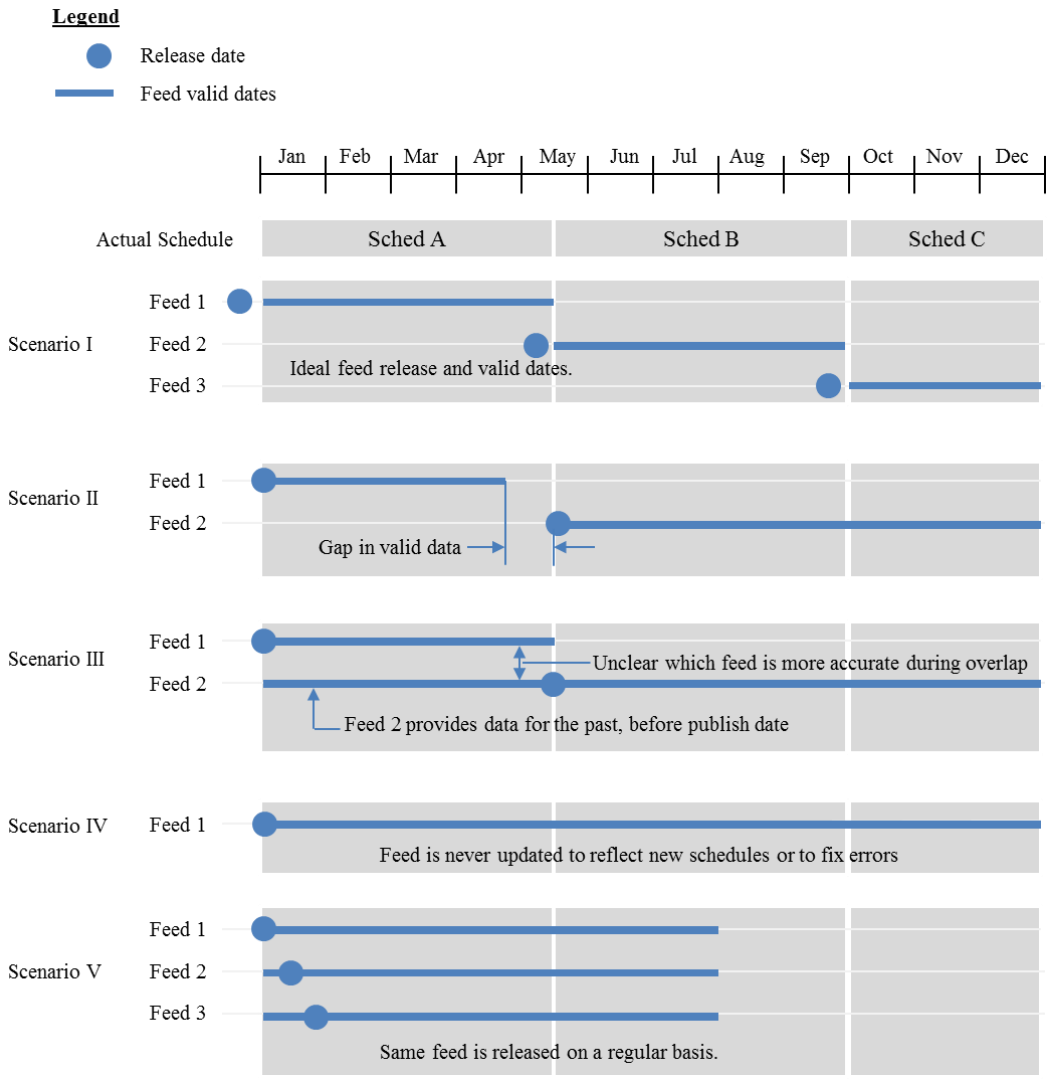


Figure 16 Potential scenarios in sequential GTFS feed releases on GTFS Data Exchange

The following scenarios pose challenges for analysis, but are not always considered "wrong" per the GTFS guidelines. Scenario II occurs when a feed is valid until a certain point in time, and the next feed is not released until after that date. There is no valid data between the end of Feed 1 and the beginning of Feed 2 in this scenario (although it may be easy to "extend" Feed 1 by replicating it until a new feed is published). Scenario III is when a feed is valid on dates before it was published. In a situation where errors are fixed in a schedule, it may impact dates in the past and for convenience, an agency won't change the previous information. It is unclear in these

situations if data before the publication date is actually valid, even though the feed may say it is. Scenario IV happens when an agency publishes once and never updates their feed or does so with such little frequency that the feed's validity is questioned. Small agencies may rarely have schedule changes, but analysts are skeptical of data as it ages. The author already identified 27 agencies that were eliminated from the calculation in Figure 14 because they had not updated their data on the website in almost two years.

The final scenario, Scenario V, occurs when an agency publishes feeds daily or at some regular interval. It may be the same actual file, but the agency updates it so often that it is unclear if and when any actual substantive changes are made to fix the feed or if it is just to maintain the update schedule. Since it is primarily designed for traveler information, GTFS feeds that are uploaded daily are likely considered the most up-to-date for the following actual day; it doesn't matter that a feed in the past was inaccurate because the data is used by riders want to know about trips they will take now or in the future, not the past. This caveat is raised because the author does not necessarily recommend that these scenarios are eliminated. If more rules or restrictions are put on GTFS feeds, it may hinder the adoption or maintenance of the data in the industry. As a reminder, this analysis is a secondary use of GTFS and should not pose changes that would inhibit its primary purpose as a traveler information data format.

Although the GTFS Data Exchange is useful because of its ease of programming access, as discussed earlier, it led to challenges stemming from the loosely defined rules inherent in GTFS feed generation and publication. The scenarios shown in Figure 16 highlight some of the scenarios that were encountered early on while trying to use multiple feeds. Agencies that do follow a release schedule like that in Figure 16 (Scenario A) will find their data easier to use by third-party developers and transit analysts.

Because of the various challenges posed by the feed update schedules discussed in this section, the analysis procedure was revised from an original plan to accumulate multiple feeds' calculations and create a composite fiscal year for an agency. If a small

number of feeds existed for each agency that emulated the ideal scenario in Figure 16, then the compilation of the daily metrics for each date in an agency's fiscal year could be executed. Because of increased processing time, however, doing so was impractical. A revised methodology to aggregate from daily values to annual values is presented in the following section.

Weekly Aggregation Method

The weekday aggregation method uses all available data from a single feed and extrapolates it to a generic year of 52 full weeks. In doing so, it assumes that the schedules that appear in a single GTFS are representative of an entire year. This assumption will have impacts for agencies with seasonally based schedules, but the magnitude of seasonal service changes is assumed to be small in the calculation of system-wide AVRM or AVRH. Since FY2012 NTD data is used for comparison, which could exist for an agency anywhere between January 1, 2012 and September 30, 2013, the feed selected was the last feed released in 2012 (roughly the middle of this time period).

The weekly aggregation method is shown in Figure 17 as an extension of the Daily NTD Metric Calculation (introduced in Figure 9). Using the output from Table E, Query 6 calculates the average value of each metric grouped by each day of the week, storing it in Table F. Using the average of each day of the week, regardless of what schedule it ran that specific date, accounts for holidays and exceptions. For example, holidays often fall on Mondays in the United States where agencies run weekend service for the holiday; an average value for all Mondays would thus be less than the average value for all Tuesdays (assuming holidays didn't fall disproportionately on Tuesdays). Query 7 multiplies it by 52 to generate an annual total for a generic year, which is actually representative of 52 full weeks. It also appends agency identifying information so that the output can be merged after processing multiple GTFS feeds.

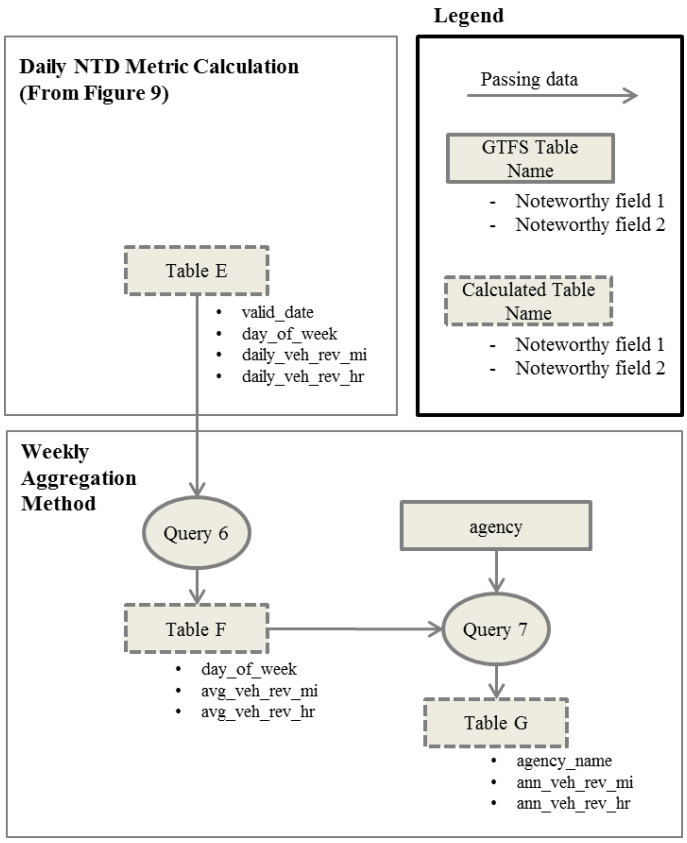


Figure 17 Weekly Aggregation Method

The output from the whole process is saved as a one-row CSV file to describe the agency's metrics. Those are then combined to create a dataset with each agency and its GTFS-calculated AVRM and AVRH included.

The weekly aggregation method introduces known differences between GTFS-calculated metrics and NTD-reported ones. To understand the magnitude of the change in this methodology, the same sample data from TriMet was used to compare the daily and weekly aggregation methods. Table 5 shows the results of using the daily and weekly aggregation methods. The feed chosen for use in the analysis (last feed published in 2012) is shown and compared to the daily aggregation method. In this example, the weekly aggregation method using a feed from December 2012 is 0.8 percent less than the value from the daily aggregation method. Not all implementations of the weekly

aggregation method will be as successful, but this result provides confidence in the use of the weekly aggregation method for the bulk analysis of agency data from the GTFS Data Exchange.

Table 5 Comparison of Weekday and Daily Aggregation Methods

Aggregation Method	Average Weekday Revenue Hours							Annual VRH	Percent Diff
	Mon	Tues	Wed	Thurs	Fri	Sat	Sun		
Daily Aggregation (FY12)	--	--	--	--	--	--	--	1,314,766	--
Weekly Aggregation (Feed:Dec 11, 2012)	4,184	3,883	4,229	4,229	4,229	2,361	1,978	1,304,752	0.8%

Analysis Results

Sample

An original list of 93 distinct GTFS feeds with bus modes[3] were identified for use with the methodology discussed earlier. These were transit service providers in the United States that had uploaded to the GTFS Data Exchange, were identified by the site as official uploads, and had uploaded GTFS data at the time of analysis (December 2012). The NTD Metrics module was run for all feeds. Of the original list of 93 feeds, 10 feeds returned errors, 20 feeds were associated with organizations not found in the NTD (private or small systems, for example), and six feeds had ambiguous agency designations making it difficult to compare to a specific organization within the NTD[4]. In the end, the 55 GTFS feeds that could be reliably matched to specific agencies within the NTD included:

- Arlington Transit
- Asheville Transit Service
- Capital Metro
- Capital District Transportation Authority
- Montgomery County MD Ride On
- Mountain Line
- San Diego Metropolitan Transit System (MTS)
- Niagara Frontier Transportation Authority

[3] According to GTFS, used for short- and long-distance bus routes

[4] These were first visually identified as extreme outliers; if an outlier in the analysis was determined to contain multiple agencies or had an ambiguous organizational designation, it was purposefully removed. The rationale is that a feed with known differences in organizational distribution than the NTD agencies cannot be reasonably compared to NTD data. Outliers whose agency designation within the feed corresponded with the NTD were not removed from the analysis.

- Champaign Urbana Mass Transit District
- Charlottesville Area Transit
- Chicago Transit Authority
- Corona Cruiser
- Corvallis Transit System
- Dallas Area Rapid Transit
- Fort Worth Transportation Authority
- Golden Empire Transit District
- Golden Gate Transit
- Greater Cleveland Regional Transit Authority
- Intercity Transit
- Island Transit
- Kitsap Transit
- Lane Transit District
- Lehigh and Northampton Transportation Authority
- Transit Authority of Lexington (LexTran)
- Maryland Transit Administration
- Massachusetts Bay Transportation Authority
- Metro St. Louis
- Metropolitan Atlanta Rapid Transit Authority
- Metropolitan Transit Authority of Harris County
- Miami Dade Transit
- Milwaukee County Transit System
- Modesto Area Express
- North County Transit District
- Orange County Transportation Authority
- Port Authority of Allegheny County
- Pinellas Suncoast Transit Authority (PSTA)
- Pioneer Valley Transit Authority of Western Massachusetts
- Redding Area Bus Authority
- Regional Transportation Commission of Southern Nevada
- Regional Transportation District
- Roseville Transit
- Sacramento Regional Transit
- San Francisco Municipal Transportation Agency
- San Joaquin Regional Transit District
- Santa Cruz Metro
- Southwest Ohio Regional Transit Authority
- Spokane Transit Authority
- Transit Authority of Northern Kentucky
- TriMet
- Unitrans (Davis)
- University of Michigan Transit Services
- Utah Transit Authority
- VIA Metropolitan Transit
- Yakima Transit
- Yuba-Sutter Transit

Comparison of Metrics

Using the methodology described earlier and the set of agencies listed above, two system-level metrics were calculated for each agency's bus service: annual vehicle revenue miles and annual vehicle revenue hours. These were then compared to values in the NTD's preliminary FY2012 data on transit operations in the United States.

Annual Vehicle Revenue Miles

GTFS revenue miles are expected to be the same as NTD revenue miles (since layover time is assumed to have a negligible impact on vehicle revenue miles). The

difference between the NTD-reported and GTFS-calculated values represents discrepancies that would need to be resolved before using GTFS to calculate NTD data in an official capacity. The results are shown in Figure 18; each point represents the NTD-reported and GTFS-calculated value for one agency's bus system. If GTFS-calculated values were equal to NTD-reported values, all points would lie on the black equivalence line, which represents a 1:1 ratio. If points lay above that line, it implies that the NTD-reported values are *greater* than GTFS-calculated values; if it lies below the line then NTD-reported values are *less* than GTFS-calculated values.

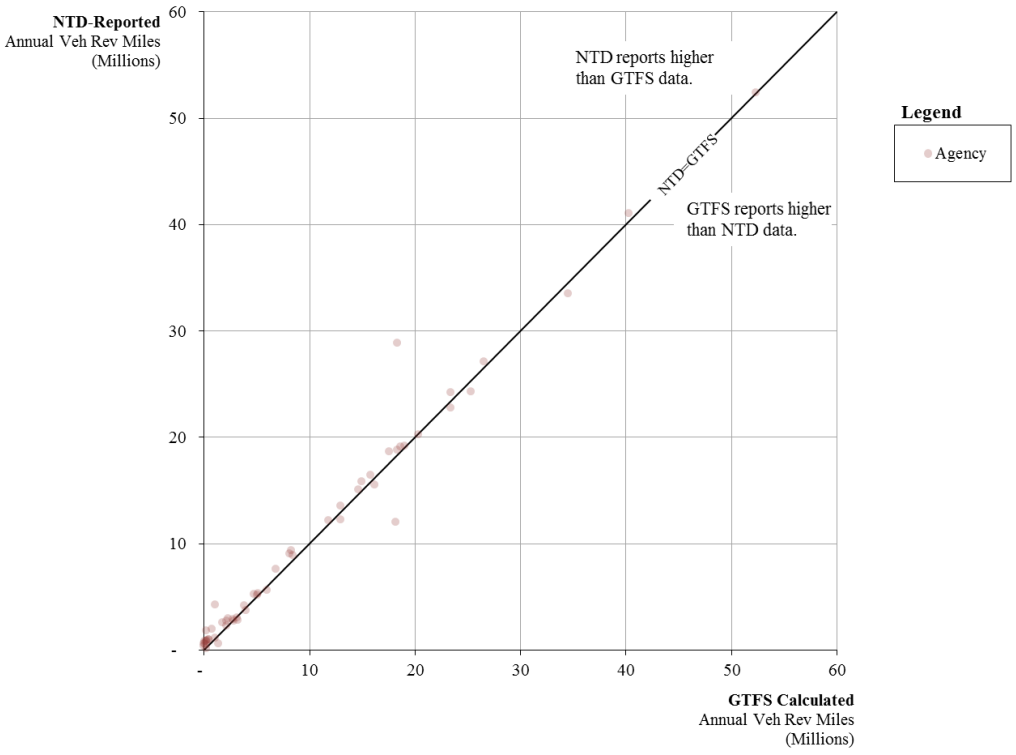


Figure 18 Comparison of NTD-Reported and GTFS-Generated Annual Vehicle Revenue Miles

The data results of the comparison are shown in Table 6, which presents each evaluated agency with its AVRM using both methods, their differences and the percent difference using the NTD values as a base. With the exception of only a few outliers, the

GTFS-calculated values are tightly distributed about the equivalence line. This means that the combination of the method employed in this research and the GTFS-feeds arrive at generally the same conclusions as those generated by transit agencies themselves when they report to the NTD. Initial inspection of the underlying GTFS feeds for the outliers in this graph did not give clear reasons for their differences.

Table 6 Comparison of NTD-Reported and GTFS-Calculated Annual Vehicle Revenue Miles for Bus Systems

Agency Name	NTD-Reported	GTFS-Calculated	Difference	Percent Difference (NTD Base)
Arlington Transit	1,128,974	1,060,748	68,226	6%
Asheville Transit Service	808,629	82,441	726,188	90%
Capital Metro	13,576,900	12,958,749	618,152	5%
Capital District Transportation Authority	7,608,400	6,848,357	760,043	10%
Champaign Urbana Mass Transit District	3,057,585	3,152,663	(95,078)	-3%
Charlottesville Area Transit	951,548	459,087	492,461	52%
Chicago Transit Authority	52,427,711	52,343,924	83,787	0%
Corona Cruiser	167,690	185,480	(17,790)	-11%
Corvallis Transit System	373,522	5,151	368,371	99%
Dallas Area Rapid Transit	27,144,101	26,502,878	641,223	2%
Fort Worth Transportation Authority	4,214,600	3,833,579	381,021	9%
Golden Empire Transit District	3,735,670	3,954,464	(218,794)	-6%
Golden Gate Transit	5,209,200	5,075,470	133,730	3%
Greater Cleveland Regional Transit Authority	12,224,802	12,970,594	(745,792)	-6%
Intercity Transit	2,725,700	2,818,180	(92,480)	-3%
Island Transit	623,600	1,336,491	(712,891)	-114%
Kitsap Transit	1,964,675	716,156	1,248,519	64%
Lane Transit District	2,786,100	3,186,048	(399,948)	-14%
Lehigh and Northampton Transportation Authority	2,611,912	1,727,886	884,026	34%
Transit Authority of Lexington (LexTran)	2,268,839	2,150,672	118,167	5%
Maryland Transit Administration	24,274,200	25,335,172	(1,060,972)	-4%
Massachusetts Bay Transportation Authority	24,222,300	23,370,571	851,729	4%
Metro St. Louis	18,635,163	17,552,856	1,082,307	6%
Metropolitan Atlanta Rapid Transit Authority	22,803,997	23,362,011	(558,014)	-2%
Metropolitan Transit Authority of Harris County	41,074,000	40,316,750	757,250	2%
Miami Dade Transit	28,838,300	18,318,732	10,519,568	36%
Milwaukee County Transit System	15,509,683	16,145,030	(635,347)	-4%
Modesto Area Express	1,833,780	225,760	1,608,020	88%
Montgomery County MD Ride On	12,207,982	11,808,015	399,967	3%
Mountain Line	637,171	19,423	617,748	97%
San Diego Metropolitan Transit System (MTS)	16,424,300	15,810,821	613,480	4%
Niagara Frontier Transportation Authority	9,028,514	8,111,682	916,833	10%
North County Transit District	5,237,788	4,710,239	527,549	10%
Orange County Transportation Authority	19,087,600	18,670,761	416,839	2%
Port Authority of Allegheny County	18,829,161	18,312,459	516,702	3%

Table 6 Comparison of NTD-Reported and GTFS-Calculated Annual Vehicle Revenue Miles for Bus Systems (Continued)

Agency Name	NTD-Reported	GTFS-Calculated	Difference	Percent Difference (NTD Base)
Pinellas Suncoast Transit Authority (PSTA)	8,877,800	8,393,392	484,409	5%
Pioneer Valley Transit Authority of Western Massachusetts	4,286,349	1,085,953	3,200,396	75%
Redding Area Bus Authority	578,433	62,881	515,552	89%
Regional Transportation Commission of Southern Nevada	15,104,687	14,604,800	499,887	3%
Regional Transportation District	33,521,000	34,554,319	(1,033,319)	-3%
Roseville Transit	502,100	207,172	294,928	59%
San Francisco Municipal Transportation Agency	12,066,127	18,207,990	(6,141,863)	-51%
San Joaquin Regional Transit District	2,705,000	2,124,723	580,277	21%
Santa Cruz Metro	2,991,700	2,263,979	727,721	24%
Southwest Ohio Regional Transit Authority	9,351,070	8,302,308	1,048,762	11%
Spokane Transit Authority	5,313,529	5,100,102	213,427	4%
Transit Authority of Northern Kentucky	2,885,892	2,722,633	163,259	6%
TriMet	19,169,232	19,008,484	160,748	1%
Unitrans (Davis)	803,164	205,412	597,752	74%
University of Michigan Transit Services	1,009,846	476,871	532,975	53%
Utah Transit Authority	15,865,000	14,936,819	928,181	6%
VIA Metropolitan Transit	20,275,073	20,310,111	(35,038)	0%
Yakima Transit	800,854	190,292	610,562	76%
Yuba-Sutter Transit	877,900	313,807	564,093	64%

The goal of this work is to have a data methodology that yields results from GTFS feeds that are consistent with data reported from the NTD. A paired t-test for matched samples evaluates the null hypothesis that these two methods are the same. Under this test, the null hypothesis is H_0: $\mu_D = 0$ and the alternative is H_A: $\mu_D \neq 0$, where μ_D is the mean of differences between paired values. Because the two-tailed P-value (0.069) is not less than alpha (0.050), one cannot reject the null hypothesis at a 95 percent confidence level. The result is that the GTFS feed methodology does not produce results that are statistically different from the results reported in the NTD.

Although the data is promising and the mean difference between paired values is not statistically significant, the high standard deviation and percent differences found in Table 6 call attention to very different results from one agency to another. The graph in Figure 20 shows the percent difference from Table 6 based on size of agency. Smaller agencies are far more likely to have calculated results over 50 percent different from their

NTD-reported values. The percent deviations among the larger agencies to the right are much smaller[5]. The majority of large differences between calculation methods is among those agencies with less than 5 million AVRM. With smaller agencies, discrepancies are likely to be more pronounced since the denominator is small in the percent difference calculation, but these show evidence of a major discrepancy.

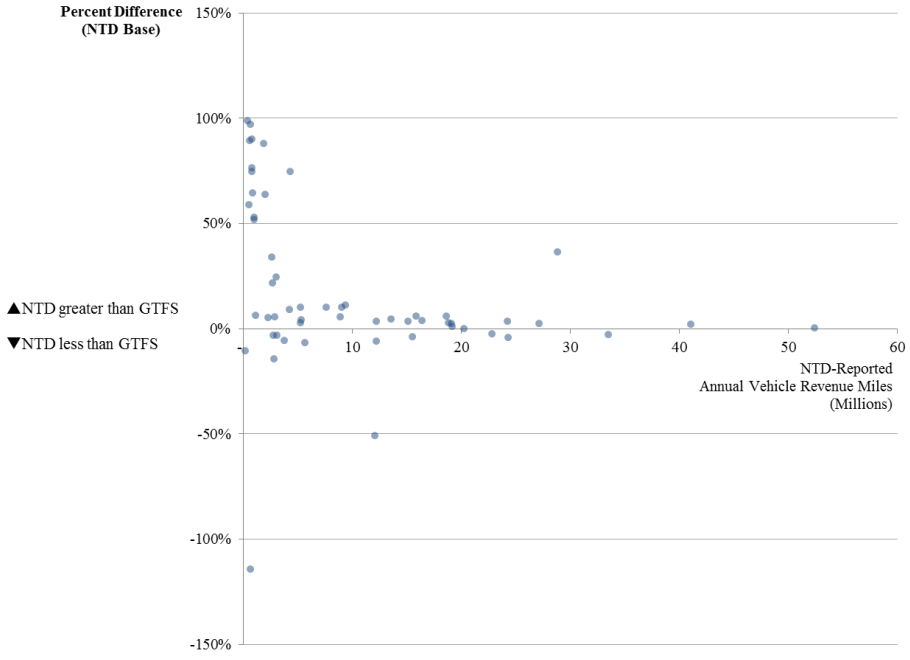


Figure 19 Percent difference between NTD-provided and GTFS-calculated methodologies by agency size

Closer inspection did not reveal specific reasons that these smaller agencies were represented differently (such as the case where specific services were represented in the NTD that were not represented in the GTFS feeds, like commuter buses). However, the current methodology does not seem to reliably capture NTD-reported AVRM for small agencies. Future efforts will need to explore possible source for these differences.

[5] The data point around 30 million AVRM shown with nearly 40% difference is Miami-Dade transit. On closer inspection, the GTFS feed chosen was found not to be collectively exhaustive (as discussed earlier) and is a poorly-formed GTFS feed. An analysis of the next published feed in January is 1.2 percent different from values reported in the NTD, as opposed to 36 percent different with the existing feed. This was not eliminated because part of the exercise is to evaluate the readiness of open data for use in transit analysis. This is an example of a failure where this feed, if it were chosen for analysis, would result in errors.

Annual Vehicle Revenue Hours

Layover time is included in the NTD calculation of revenue hours but not in the GTFS calculation, which only includes running time. Because of this major discrepancy in calculation methods, statistical testing and additional quantitative analysis were not pursued. Since the NTD reporting manual says that layover time is usually equivalent to 10 to 20% of the running time, isolines are shown in to identify where points would fall if the NTD data were adjusted to GTFS data with an additional 10 and 20%. NTD data is expected to lie between the two isolines where NTD equals GTFS+10% and GTFS+20%. The trend line for this dataset is within that range for larger agencies and is generally consistent with results discussed earlier. Layover time varies considerably from route to route and among agencies; for that reason the best-fit line is a more appropriate visual representation of the aggregate trend of the GTFS-calculated values.

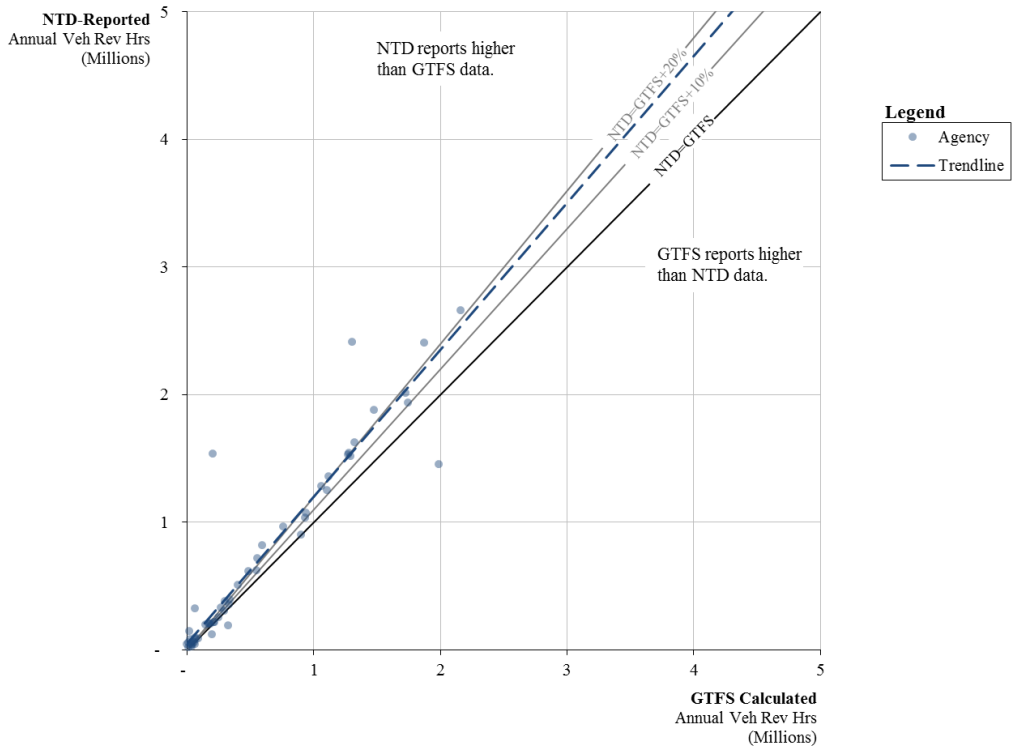


Figure 20 Comparison of NTD-Reported and GTFS-Generated Annual Vehicle Revenue Hours

Discussion

The tight distribution of values in Figure 18**Error! Reference source not found.** demonstrates that GTFS feeds are capable of generating the same system-level revenue miles that are reported in the NTD. The distribution of percent differences, however, identifies a significant challenge for direct comparison of NTD-reported and GTFS-calculated values among small agencies (less than 5 million AVRM). This warrants further analysis to identify whether discrepancies are the result of data in the GTFS feeds or the methodology used to aggregate that data.

Although the direct comparison is unavailable for revenue hours because of layover time, the comparison of NTD-reported to GTFS-generated vehicle revenue hours still suggests results consistent with that of the vehicle revenue miles comparison.

Well-formed GTFS feeds are a good source of data and the method employed in this section is an accurate way of aggregating trip-level metrics to the system level. The more consistent results among large agencies support the notion that the aggregation method is adequate. The comparison of the aggregation methodology to the NTD data is important because Chapter 3 highlighted potential pitfalls of working with granular trip or stop-level data that could misrepresent information when aggregated incorrectly; if the aggregation method were flawed, it would be seen for all results in the last few figures in this chapter. This result implies that the GTFS Reader can be used for other analyses with well-formed GTFS feeds.

A major challenge for analysts is the rate of GTFS feed publication as discussed earlier in Figure 16 (Potential scenarios in sequential GTFS feed releases on GTFS Data Exchange). The high frequency of updates by some agencies required that the analysis in this research only use one feed when it ideally would have incorporated multiple ones. The use of multiple feed may have resulted in a more accurate comparison of fiscal year data than the generic year generated from the single feed analysis. The complications in feed publishing reinforce the notion that GTFS Data Exchange is less suited for historic analysis than it is for the traveler information applications for which it is designed. More consistent upload rules and GTFS feed validation[6] by the GTFS Data Exchange would improve its potential use for historic transit performance analysis.

Two system-level metrics were calculated in this analysis, but others from the NTD could have likewise been included. One such metric is the number of vehicles operated in maximum service, a straightforward calculation of the number of active trips at all times of day. Another is directional route miles, foregone in this analysis because of the more significant geospatial calculations that are needed (the existing GTFS Reader framework does have the capability to use more functions from PostGIS to accomplish

[6] Validation here refers to the format of the GTFS feed and its internal consistency, not its agreement with data in the NTD.

this calculation). Other metrics are straightforward tabulations such as number of stations for rail systems, or even number of bus stops (not currently in the NTD). The lack of real-time information in this dataset limits the metric development to these kinds of scheduled availability metrics, but future use of real-time information could address others like on-time performance.

One caution about GTFS data validity is that since individual transit agencies are responsible for both their own NTD submissions and their own GTFS feeds, it stands to reason that any errors in an agency's raw data would cascade into both the NTD-reported and GTFS-generated metrics. This analysis assumed the correctness of NTD data because it is used by federal agencies for distributing funding, but a successful comparison of these two values only indicates internal consistency within an agency. Still, since riders would likely catch major errors in GTFS feeds, this is a promising finding.

CHAPTER 5

CONCLUSION

As a publicly-provided service, transit service often faces scrutiny by policy-makers and the public. Reporting on the current level of transit service is important in order to demonstrate the value of transit as a public service, to garner support for investment in transit, and to provide policy-makers and the public insight into the operations of a public agency. One way to provide information about an agency's service is to develop and report performance measures that accurately represent different factors of the service. The advent of open data has allowed for citizen action in the form of data analysis and has opened the door to greater transparency in public agency operations. The result of this research is an endorsement of GTFS data and the GTFS Reader framework for use in substantive performance measurement.

One of the original reasons that the author first explored this area was that, as a performance measurement data source, GTFS data seemed inherently reliable; agencies use this data to convey to riders the actual service provided so that trips could be planned and transit could be consumed. For that reason, the information should be right. Of course errors and omissions may occur, but the GTFS data could be trusted to provide accurate performance measurement as much as the general public trusts Google Transit to give accurate scheduled transit data. Getting from standardized digital timetable information to meaningful performance measures required more nuance and understanding of data than originally anticipated, but the resulting methodology is documented here for future researchers to apply elsewhere.

The real impact of this thesis is that future researchers have the opportunity to incorporate better transit analysis going forward. Planning processes are limited in their treatment of transit service most often because transit lacks data and the United States prioritizes quantitative planning processes. The long history of traffic counts, roadway

inventory and other automobile data led to research and policy that addresses automobile concerns. As practitioners and policy-makers are better able to identify and point to specific deficiencies among transit service, there will be a better opportunity to address those deficiencies. Although the author doesn't suspect that many other datasets beyond real-time vehicle location will soon be in the mainstream open data movement, the standardization and use of other transit data could likewise improve the attention paid to transit. These might include passenger boarding and alighting, train-car specific loading, fare payment, vehicle maintenance and right-of-way maintenance. Again, these are not all well suited in the context of open data, but if they are standardized and available to researchers and analysts, the stories that they tell can be addressed.

Future Work

The analytic demonstration and comparison to NTD values from this research is largely a precursor to future practice in the use of GTFS data for transit performance measurement. GTFS-generated metrics for large agencies are good alternatives to the NTD-reported values, but small agencies are more susceptible to differences between their NTD-reported and GTFS-calculated values. The most important near-term work is in identifying the source of discrepancy to determine if it there is a systematic error in either the aggregation presented here or the GTFS feeds themselves. If this is resolved, it will suggest that open data generated by agencies *in general* can be considered a reliable description of the service provided by that agency. Researchers and practitioners could comfortably use open transit data for other applications.

In addition to work like the DVRPC inclusion of GTFS in their travel demand modeling *(23)*, regional planners could use GTFS to identify gaps in mobility for seniors; they could run transit availability analyses that make use of temporal distributions, not just spatial ones; and they could use it for alternatives analysis of transit improvements.

Beyond the applications of GTFS to the planning and analysis process, the other logical step in this arena is to apply one of the real-time transit data standards to calculate reliability measures such as on-time performance. There are currently three main competing standards for real-time vehicle location information: GTFS-realtime, a data standard for transmitting live data about vehicle locations and transit arrival predictions; the Transit Communications Interface Profile (TCIP), a complex standard covering all hardware and software interactions for transit systems in the United States, including a customer information module; and SIRI, the Service Interface for Realtime Information, which includes similar functionality to both GTFS and GTFS-realtime. While none of these three standards has as broad acceptance as GTFS does in the United States, it is possible that agencies will gravitate towards one of these standards as over the next few years. As more and more data becomes standardized and readily available, analysts and researchers will be able to better understand the functions of transit agencies.

One useful project would be to create a modified version of the GTFS Data Exchange that has a more thorough reporting and validating mechanism. A federal repository might look similar to the GTFS Data Exchange except that it would include agency-endorsed GTFS feeds, it would limit the rate of updates to only relevant or changed schedules (or fixed errors), it would tie GTFS feeds to specific agencies, and it would ideally have a policy lever that encouraged agencies to participate. The effect of this clean data repository would be that transit data would be available at a much more granular level than the current annually-reported datasets that ask for system-level characteristics. Federal guidance could support the establishment of this kind of clearinghouse, even if agencies themselves weren't specifically compelled to participate but did so on their own.

If the clearinghouse can store all the valid GTFS feeds (and eventually real-time data connections), it opens up the possibility that new data to describe transit agency modes, routes, and stops could also be generated. The GTFS Reader could, for example,

read every feed when it is uploaded to the clearinghouse and then report all the stop, route and system level characteristics related to headways, stop densities and other metrics discussed in this paper. The result would be actual metrics rather than just timetables could be stored in publicly accessible locations. These are projects whose results would invite broader participation in transit research.

REFERENCES

1. Ryus, P., A. Danaher, M. Walker, F. Nichols, W. Carter, E. Ellis, L. Cherrington, and A. Bruzzone. *Transit Capacity and Quality of Service Manual, 3rd Edition*. Washington, D.C., 2013.

2. Vuchic, V. R. *Urban Transit: Operations, Planning and Economics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

3. Obama, B. Executive Order 13642 - Making Open and Machine Readable the New Default for Government Information. Federal Register, Government Printing Office, Washington, D.C., 2013.

4. Wong, J., Reed, L., Watkins, K. E., & Hammond, R. (2013). Open Transit Data: State of the practice and experiences from participating agencies in the United States. In *92nd Annual Meeting of the Transportation Research Board*. Washington, D.C.: Transportation Research Board.

5. Khani, A., Sall, E., Zorn, L., & Hickman, M. (2013). Integration of the FAST-TrIPs Person-Based Dynamic Transit Assignment Model, the SF-CHAMP Regional, Activity-Based Travel Demand Model, and San Francisco's Citywide Dynamic Traffic Assignment Model. In *92nd Annual Meeting of the Transportation Research Board* (p. 16p). Washington, D.C.: Transportation Research Board.

6. Lee, S. G., Tong, D., & Hickman, M. (2013). A Comparative Study of Alternative Methods for Generating Route-level Mutually Exclusive Service Areas. In *92nd Annual Meeting of the Transportation Research Board* (pp. 1–18). Washington, D.C.: Transportation Research Board.

7. Nazem, M., Trepanier, M., & Morency, C. (2013). Integrated Intervening Opportunities Model for Public Transit Trip Generation-Distribution: A Supply-Dependent Approach. In *92nd Annual Meeting of the Transportation Research Board* (p. 20p). Washington, D.C.: Transportation Research Board.

8. Catalá, M., S. Downing, and D. Hayward. *Expanding the Google Transit Feed Specification to Support Operations and Planning*. Tampa, FL, 2011, p. 64.

9. Burwell, S. M., VanRoekel, S., Park, T., & Mancini, D. J. (2013). Memorandum: Open Data Policy - Managing Information as an Asset. Washington, D.C.

10. Hemerly, J. Public Policy Considerations for Data-Driven Innovation. *Computer*, Vol. 46, No. 6, 2013, pp. 25–31.

11. Kuk, G., & Davies, T. (2011). The Roles of Agency and Artifacts in Assembling Open Data Complementarities. In *Thirty second International Conference on Information Systems* (pp. 1–16). Shanghai.

12. O'Reilly, T. Government as a Platform. In *Open Government: Collaboration, Transparency, and Participation in Practice* (D. Lathrop and L. Ruma, eds.), O'Reilly Media, Inc., Sebastopol, California, pp. 13–40.

13. Rojas, F. M. (2012). Transit Transparency : Effective Disclosure through Open Data. Cambridge, MA.

14. Roth, M. How Google and Portland's TriMet Set the Standard for Open Transit Data. Streetsblog San Francisco, San Francisco, CA, Jan, 2010.

15. Miller, C. C. Local Governments Offer Data to Software Tinkerers. *The New York Times*, Dec 06, 2009.

16. Wonderlich, J. Open Data Creates Accountability. http://sunlightfoundation.com/blog/2012/07/06/open-data-creates-accountability/. Accessed Nov. 4, 2013.

17. Grisby, D. APTA Surveys Transit Agencies on Providing Information and Real-Time Arrivals to Customers. September 2013.

18. General Transit Feed Specification Reference. https://developers.google.com/transit/gtfs/reference. Accessed Jul. 1, 2012.

19. Front Seat Software. Transit App Gallery. http://www.citygoround.org/apps/. Accessed Oct. 24, 2013.

20. Jiang, Y., & Walker, A. (2012). STLTransit - YouTube. http://www.youtube.com/user/STLTransit. Accessed Nov. 1, 2013.

21. Walk Score. Apartments and Rentals - Find a Walkable Place to Live. http://walkscore.com/apartments/. Accessed Jul. 1, 2012, .

22. Perk, V., B. Thompson, and C. Foreman. *Evaluation of First-Year Florida MPO Transit Capacity and Quality of Service Reports*. Tampa, FL, 2001, pp. 1–51.

23. Puchalsky, C., Joshi, D., & Scherr, W. (2011). Development of a Regional Forecasting Model Based on Google Transit Feed. In *13th TRB Planning Application Conference* (pp. 1–17). Reno, Nevada: Transportation Research Board.

24. Ryus, P., M. Connor, S. Corbett, A. Rodenstein, L. Wargelin, L. Ferreira, Y. Nakanishi, and K. Blume. *A Guidebook for Developing a Transit Performance-Measurement System (TCRP Report 88)*. Washington, D.C., 2003.

25. *Moving Ahead for Progress in the 21st Century*. Public Law 112-141, 2012.

26. American Public Transportation Association. *MAP-21: A Guide to Transit-Related Provisions*. Washington, D.C., 2012.

27. Julnes, P. de L., and M. Holzer. Promoting the Utilization of Performance Measures in Public Organizations: An Empirical Study of Factors Affecting Adoption and Implementation. *Public Administration Review*, Vol. 61, No. 6, Nov. 2001, pp. 693–708.

28. Smith, J. R., D. Belzer, and S. Bernstein. Center for Transit-Oriented Development Response to Notice of Proposed Rulemaking and proposed Policy Guidance for the New Starts and Small Starts programs. http://www.regulations.gov/#!documentDetail;D=FTA-2010-0009-0254.

29. Danaher, A., P. Ryus, E. Ellis, M. C. Walker, and K. Hunter-Zaworski. *Transit Capacity and Quality of Service Manual, 2nd Edition*. Transportation Research Board, Washington, D.C., 2003.

30. Chen, X., L. Yu, Y. Zhang, and J. Guo. Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation Research Part A: Policy and Practice*, Vol. 43, No. 8, Oct. 2009, pp. 722–734.

31. Saberi, M., Zockaie K, A., Feng, W., & El-geneidy, A. (2012). Definition and Properties of Alternative Bus Service Reliability Measures at the Stop Level. In *91st Annual Meeting of the Transportation Research Board* (pp. 1–15). Washington, D.C.: Transportation Research Board.

32. Eboli, L., and G. Mazzulla. A New Customer Satisfaction Index for Evaluating Transit Service Quality. *Journal of Public Transportation*, Vol. 12, No. 3, 2009, pp. 21–38.

33. Cambridge Systematics, UMD Center for Advanced Transportation Technology, and Resource Systems Group. *Measuring Transportation Network Performance (NCHRP Report 664)*. Washington, D.C., 2010.

34. Laporte, G., J. A. Mesa, F. A. Ortega, and F. Perea. Planning rapid transit networks. *Socio-Economic Planning Sciences*, Vol. 45, No. 3, Sep. 2011, pp. 95–104.

35. Kittelson & Associates. *Transit Quality of Service Applications Guide*. Orlando, Florida, 2008.

36. *National Transit Database and Apportionment of appropriations for formula grants*. Title 49 USC 5335(a).

37. Federal Transit Administration. NTD Data. http://www.ntdprogram.gov/ntdprogram/data.htm. Accessed Jul. 15, 2012, .

38. National Transit Database. *2012 Reporting Manual*. Washington, D.C., 2012.