# **Multimodal GTFS Measures**

# **GTFS/ARNOLD Snapping Process**

# **3 May 2017**

# Introduction

The goal of this series of scripts is to snap GTFS bus shape points to FHWA’s ARNOLD (All Roads Network of Linear referenced Data) in order to capture Highway Performance Monitoring System (HPMS) data such as Annual Average Daily Traffic (AADT) along transit routes, and calculate transit trip frequency along the road network. The base of analysis is the shapes.txt GTFS file, a tabular representation of all the geospatial points that make up a linear transit network. Each point in the shapes.txt file is associated with a shape ID (essentially a unique transit travel path). A shape ID (hereinafter referred to as a route shape) is in turn associated with a route ID. Typically there is at least one route shape per route direction (e.g. inbound/outbound) and additional route shapes if the route’s path varies depending on time of day, day of week, etc. To understand the level of transit service along a particular road corridor, we need to attach all transit trips that traverse that corridor to one “master” geospatial dataset, which eliminates any inconsistencies in the GTFS route shape data and ensures that route shapes that overlap snap to a consistent geospatial dataset. We choose ARNOLD because it is a national “all roads” dataset that linear references to valuable HPMS data that is submitted by the state DOTs.

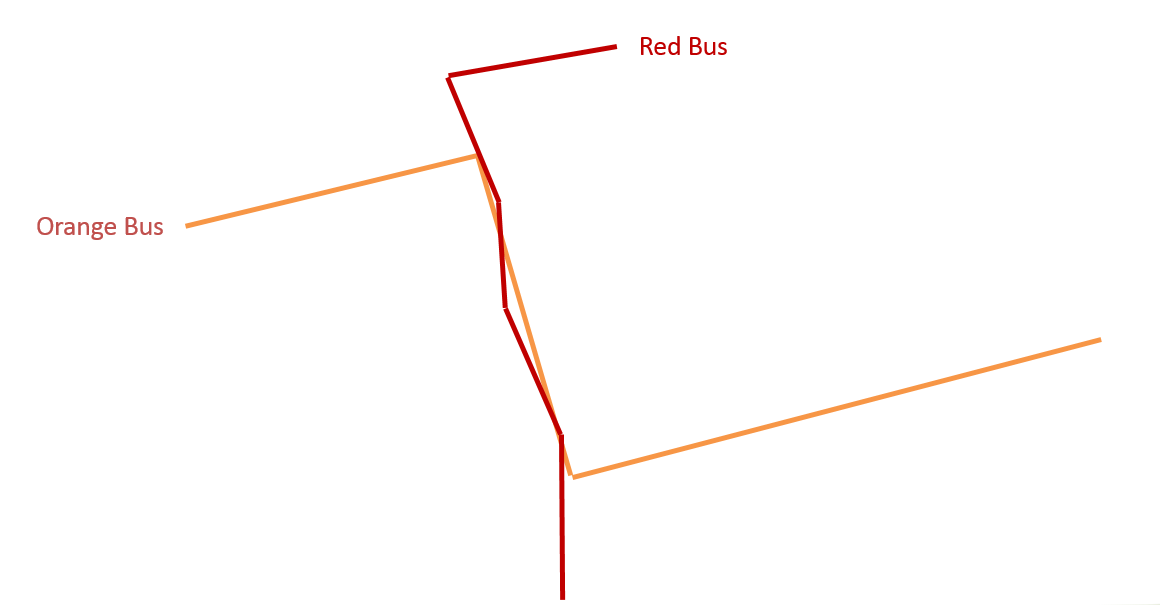


Figure 1 – Hypothetical example of why snapping to a road network (e.g. ARNOLD) is needed

# Technical Documentation

The code developed for this project is designed to work with the 32-bit or 64-bit ArcGIS installation of Python 2.7 and also requires ArcGIS 10.3 (code has also been tested successfully on ArcGIS 10.4). One non-standard Python library is also required and must be installed before use: PyGTFS. PyGFTS documentation is available here: <https://github.com/jarondl/pygtfs>

# Script 0: GTFS Ingest

**Description:** Converts a GTFS Feed into a SQLite database that can be read by ArcGIS and/or easily queried by programs such as DB Browser for SQLite.

**Inputs Required**

* Directory or Zip containing a transit system’s GTFS Data

**Caveats:** The code requires PyGTFS to be installed before use. The code does not currently work with the Bureau of Transportation Statistics (BTS) National Transit Map (NTM) data as NTM does not use the same column names as the GTFS standard used by PyGTFS plug-in. Agency-based GTFS feeds must be used instead.

# Script 1: GTFS-ARNOLD Merge

**Description:** Associates GTFS route data with the ARNOLD Road network.

* Fetches and processes route shapes data from GTFS file
* Prepares, subsets, and cleans up ARNOLD Data—ensuring network connectivity.
  + “Enhanced” ARNOLD data which has been manually digitized is added to the existing ARNOLD dataset
  + ARNOLD is subset to the area immediately surrounding the transit system (five miles)
  + AADT data is linear-referenced with the ARNOLD data
  + Overlapping ARNOLD segments (a common issue in Massachusetts and perhaps other states) are eliminated while retaining the AADT
  + Flaws in ARNOLD data mean that not all roads are properly connected to neighboring roads. These ‘dangles’ are artificially linked to the nearby roads.
  + Network is temporarily loaded into an ArcGIS geometric network to detect any remaining disconnected features (orphan roads disconnected from the rest of the road network) and delete them to prevent routing failure later in the process.
* Prepares a blank network dataset from a template and copies ARNOLD Data into network dataset
* Calculate locations along road network for each unique GTFS point
* Setup and route solve (using ArcGIS Network Analyst)
* Reports and compares solved routes to raw shape routes (as created by Esri geoprocessing tool)

**Inputs Required**

* **sqlite\_file**: SQLite file containing the transit system’s GTFS Data (created in Step 0)
* **arnold**: ARNOLD shapefile representing the state’s road network where the transit system is located. Script is not currently set up to handle multi-state transit systems.
* **hpms**: CSV file representing the AADT data for the state where the transit system is located. This file is not currently leveraged in the script as there is a bug caused by coincident ARNOLD segments which can have varying levels of AADT. If this file is used in any capacity, ensure that data is comma-delimited (not pipe-delimited) before use.
* **flag**: A point located along the transit area's ARNOLD network. This file is required and facilitates a check for disconnected features that must be deleted before the ArcGIS Network Analyst route solve. Must be manually created for each transit agency.
* **template\_network**: A blank network dataset needed to programmatically build networks for multiple cities. Provided with script.
* **enhanced\_arnold**: A shapefile that contains road segments missing from ARNOLD. Must be created manually and maintained as a companion to each state's ARNOLD data.
* **calc\_location\_tol\_1**: Tolerance used to snap GTFS route shape points to an ARNOLD node (adjustable, 10 meters default)
* **calc\_location\_tol\_2**: Tolerance used to snap GTFS route shape points to an ARNOLD edge (adjustable, 5 meters default)

**Caveats:** Does not currently function with transit systems that traverse more than one state. Code is also limited to bus routes and excludes rail and other transit routes which do not run along roads.

# Script 2: GTFS-ARNOLD Stops

**Description:** Extracts stops from GTFS feed, converts to geospatial format, linear references stops to ARNOLD network, and splits ARNOLD network at stops.

**Inputs Required**

sqlite\_file: SQLite file containing the transit system’s GTFS Data (created in Step 0)

# Script 3: GTFS-ARNOLD Reflow

**Description:** Repeats a significant portion of Step 1—this time incorporating each route’s series of stops and the road network that is now split at each stop.

**Inputs Required:**

* **sqlite\_file**: SQLite file containing the transit system’s GTFS Data (created in Step 0)
* **template\_network**: A blank network dataset needed to programatically build networks for multiple cities. Provided with script
* **calc\_location\_tol\_1**: Tolerance used to snap GTFS route shape points to an ARNOLD node (adjustable, 10 meters default)
* **calc\_location\_tol\_2**: Tolerance used to snap GTFS route shape points to an ARNOLD edge (adjustable, 5 meters default)

# Script 4: GTFS-ARNOLD Arc Level

**Description:** Allows user to calculate frequency along ARNOLD segments for a particular day in a GTFS feed, using trip frequency. Ultimate output in GIS format is called “flow\_result”.

**Inputs Required**

* **sqlite\_file**: SQLite file containing the transit system’s GTFS Data (created in Step 0)
* **date**: Must identify a date in ‘yyyy-mm-dd’ format that represents a service date of interest. User must examine GTFS feed calendar to choose an appropriate date that is covered by feed.
* **day\_of\_week**: Must identify a day of week that represents a service day of interest (e.g. Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, or Sunday’
* **actual\_ridership\_csv:** CSV file containing stop-pair level ridership data from transit agency
* **modeled\_ridership\_csv:** CSV file containing stop-pair level ridership data output by the model

**Caveats:** Note that the user must properly set the date used in the code. The user must examine the GTFS feed to determine which day represents a typical weekday (or other service of interest) within the transit system’s GTFS feed.

# Outputs

* **gtfs\_arnold\_prelim.gdb**
  + **network:** A feature dataset containing the feature classes necessary to route transit along the ARNOLD network. arnold\_split\_nw is the ARNOLD network processed and used for routing in Step 1. arnold\_split\_stops\_nw is the ARNOLD network processed is Step 2 and used for routing in Step 3. network\_ND is the network dataset generated when the network is built in ArcGIS. Network\_ND\_Junctions are the network junctions generated when the network is built in ArcGIS.
  + **all\_bus\_shape\_id\_pt:** A point feature class containing all bus route shape ID points from the GTFS feed.
  + **bus\_route\_shapes raw:** A line feature class containing the bus route shapes built using the raw point shape ID data (e.g. not snapped to ARNOLD).
  + **bus\_shape\_id\_pt:** A point feature class containing all the *unique* bus route shape ID points from the GTFS feed.
  + **problem\_shape\_points:** A point feature class containing route shape ID points that occur outside a certain distance from route\_results feature class—they may identify areas where there is missing ARNOLD data or problematic GTFS data.
  + **problem\_shape\_points\_summary:** A summary of the number of route shape ID points from the problem\_shape\_points feature class by route shape ID.
  + **route\_results:** Route shapes as snapped to the ARNOLD network
  + **stops\_lrs:** GTFS stops snapped and linear-referenced to the ARNOLD network
  + **TraversedEdges:** A comprehensive feature class of the ARNOLD route segments traversed by each route shape.
* **gtfs\_arnold\_final.gdb:** This geodatabase contains many of the same layers as the first geodatabase. The main difference is that all of these layers now reflect the new ARNOLD network that is split at transit stops
  + **flow\_result:** a line feature class containing the frequency of transit routes along an ARNOLD route segment, along with the AADT (if available) and other calculated measures based on the GTFS feed/ridership data

# Potential Future Improvements

* Coordinate with FHWA to improve ARNOLD data where gaps and deficiencies exist
  + Ensure consistency in inclusion/exclusion of dual carriageways
  + If dual carriageways are not included in ARNOLD submission, connectivity to neighboring roads must be maintained.
  + In states such as Florida and Pennsylvania, ARNOLD is incomplete (does not include all roads).
  + Ideally, all ARNOLD data should be routable without the series of processing/cleanup steps undertaken in Script 1.
* Develop code to allow for the processing of multi-state systems (ARNOLD data is currently released at a state-by-state level and is not set up to be networked across state borders)
* Develop code to snap rail-based transit routes to a standardized rail dataset