Appendix A

User Guide: Estimating a Large OD Drive Time Matrix

1. Background and Purposes

This user guide is to support the release of data and toolkits related to the following paper:

Hu, Y., C. Wang, R. Li, and F. Wang. 2020. Estimating a large drive time matrix between zip codes in the United States: A differential sampling approach. *Journal of Transport Geography* 86, 102770.

The paper illustrated challenges of estimating a large drive time matrix faced by researchers from various fields, such as geography, public health, and transportation, and proposed a feasible and efficient solution to the estimation of a large drive time matrix from 32,840 origin (O) ZIP codes to 32,840 destination (D) ZIP codes in the U.S. The task is broken down to estimation of travel time matrices at three levels. Each begins with a preliminary baseline estimation:

1) Level 1 utilizes a complete road network including interstates, highways, major roads, and local roads to calculate a drive time and/or distance matrix for short-range trips.
2) Level 2 utilizes a simplified road network of only interstates and highways to measure the drive time/or distance matrix for medium-range trips.
3) Level 3 utilizes the geodesic distance to measure the long-range trips.

A subset of OD pairs from each level is then randomly selected, and the drive times for these OD pairs are estimated via the Google Map API. Based on the regression models between the baseline estimates and the Google times from this subset, we derive the final estimates for the full dataset of ZIP-to-ZIP drive time matrix.

This user guide helps users to implement four tasks:

1) Download our calibrated datasets at Level 1 (0-3 hour drive time) and Level 2 (3-6 hour drive time).
2) Replicate the drive time estimation between ZIP codes in different years and/or different regions.
3) Replicate the drive time estimation between other geographic units, such as census tract, county.
4) Extract a subset of OD ZIP code pairs from our calibrated dataset.

2. Data and Programs

The folder "Large_OD_Data_Estimation" contains data and program tool such as:

(1) ZCTA_PWC.gdb contains one feature class ZCTA_PWC_2010. It includes 32,840 ZIP code population-weighted centroids calibrated from the population data at the census block level. It has four important fields, ZCTA5CE10, POINT_X, POINT_Y, and ST_NAME, to represent the unique ID of ZIP code, longitude, latitude, and state.

(2) NA_OD_3hours.csv represents the calibrated OD drive time matrix at Level 1 in (0,3] hours. Note that for 188 records that contain negative distance after applying the regression model, we keep the original values plus the distances within origin ZIP code and within destination ZIP code to finalize the estimated distance.

(3) NA_OD_3-6hours.csv represents the calibrated OD matrix at Level 2 in (3,6] hours.

(4) Generate OD Cost Pro.tbx, implemented in ArcGIS Pro to estimate OD drive time and extract calibrated data. Detailed descriptions are provided in the next section.

(5) LargeODcost contains the python scripts used in Generate OD Cost Pro.tbx.

(6) US_OD_Cost_Calibrated_Data contains two folders, hour03 and hour36, to store the calibrated OD matrix at Levels 1 and 2 in .data format for fast fetch, which are also identical to (3) and (4) accordingly. They will be used in the Generate OD Cost Pro.tbx to extract a small number of OD matrix. Under the two folders, each subfolder is named by the first three digits of the origin ZIP code, and under each subfolder, each file ended with .data is named by the last two digits of the origin ZIP code, and it contains records with five-digit destination ZIP code, travel time in minutes, and travel distance in miles.

The NA_OD_3hours.csv and NA_OD_3-6hours.csv have identical fields such as:

- OZCTA: origin ZIP code, identical to the field ZCTA5CE10 in ZCTA_PWC_2010.
- DZCTA: destination ZIP code, identical to the field ZCTA5CE10 in ZCTA_PWC_2010.
- EstTime: travel time from origin ZIP code to destination ZIP code, unit: minutes
- EstDist: travel distance from origin ZIP code to destination ZIP code, unit: miles.

The OD drive time matrix data at Level 3 is not provided here due to its massive data size and long processing time for downloading. It can be reconstructed by using the regression model reported in the paper and based on geodesic distances that can be quickly derived by the tool "Generate Near Table" in ArcGIS Pro or the ninth tool (05B Write All OD Pairs with Geodesic Distance) to be discussed in the next section.

3. Descriptions of tools under "Generate OD Cost Pro.tbx"

The section illustrates how to use each tool in "`Generate OD Cost Pro.tbx`" to estimate, calibrate, and extract an OD drive time matrix by ArcGIS Pro from a road network and by Google Maps API.

In ArcGIS Pro, go to Contents pane > Catalog > Project > Toolboxes, right click Toolboxes, click Add Toolbox to open the dialog window, select and add the toolkit of "`Generate OD Cost Pro.tbx`" from the folder `Large_OD_Data_Estimation`. Expand the toolkit to display 12 tools:

(1) 00 Set Default Network Setting
(2) 01 Snap Points to Road Network (Optional)
(3) 02 Slice Feature Layer
(4) 03A Generate OD Matrix from Road Network
(5) 03B Generate OD Matrix from Google Maps (Optional)
(6) 04A Calibrate Inter-zonal OD Matrix (Optional)
(7) 04B Calibrate Intra-zonal OD Matrix (Optional)
(8) 05A Merge OD Matrix at the Same Level (Optional)
(9) 05B Write All OD Pairs with Geodesic Distance
(10) 06A Construct OD Pair by States
(11) 06B Construct OD Pair by ZCTA List
(12) 07 Extract Calibrated Travel Cost using OD pairs

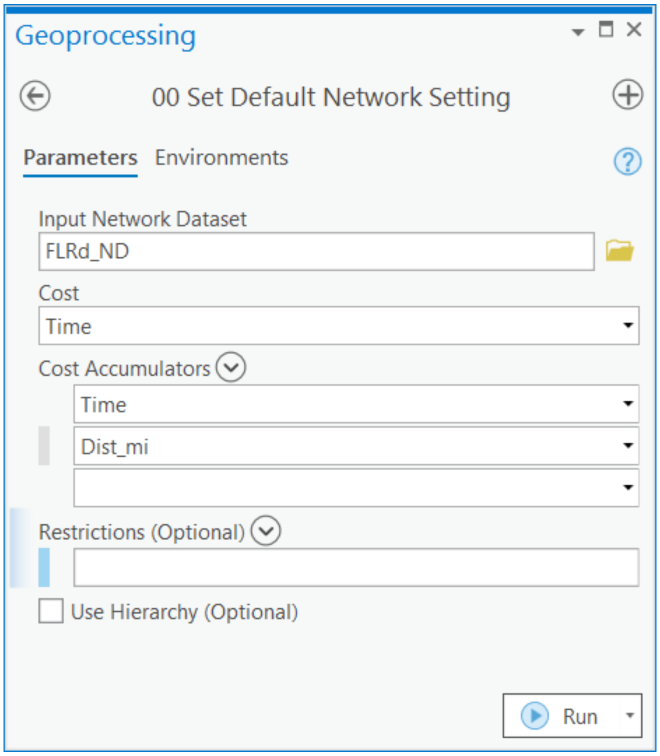
Tools (1)-(9) are used to generate a large OD travel cost matrix (including travel time, distance, or both). Tools (10)-(12) are mainly used to extract the travel cost matrix of the input OD pairs from the calibrated ZIP-to-ZIP pairs illustrated in our paper.

Two important tasks should be done before using these tools. One is to ensure the Script File of each tool points to the same python script under the folder `LargeODcost`. Specially, right click each tool, select Properties to open the dialog window of Tool Properties. In General tab, verify the path of Script File. Another is to create and build a network dataset to ensure it has two costs to represent the travel time and distance. Detailed descriptions are provided in sub-section 2.2.2 of Chapter 2 or by clicking this link: https://pro.arcgis.com/en/pro-app/latest/help/analysis/networks/how-to-create-a-usable-network-dataset.htm.

Metadata for each tool is included so that users can browse the description of each item in each tool to learn more details. The following provides a brief description of each tool.

(1) 00 Set Default Network Setting

This tool is to set default network environment (Figure A1). It is necessary when estimating distance and/or travel time at Level 1 and Level 2 with a complete road network for short-range trips and only major highways for medium-range trips, respectively. In other words, it should be run twice with two different network datasets if both levels are used for estimation. For Level 3 to calculate the geodesic distance for the long-range trips, no need to use this tool.



Geoprocessing

00 Set Default Network Setting

Parameters Environments

Input Network Dataset: FLRd_ND

Cost: Time

Cost Accumulators:
- Time
- Dist_mi

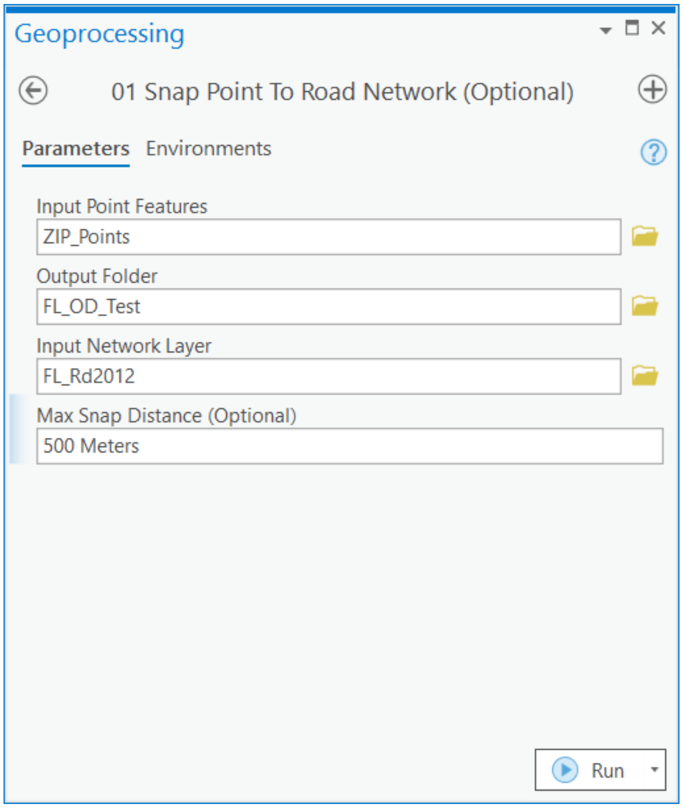
Restrictions (Optional):

□ Use Hierarchy (Optional)

Figure A1 Interface for default network environment setting

(2) 01 Snap Point To Road Network (Optional)

The optional tool is used for snapping ZIP code point features to the nearest edge of the provided polyline road network within a specific distance. It automatically projects two features into the identical coordinate system. As shown in Figure A2, the Output Folder should be created by users and the optional Max Snap Distance has a default value 500 meters. This tool will create a temporary geodatabase named temp.gdb under the output folder.



Geoprocessing

01 Snap Point To Road Network (Optional)

Parameters Environments

Input Point Features: ZIP_Points

Output Folder: FL_OD_Test

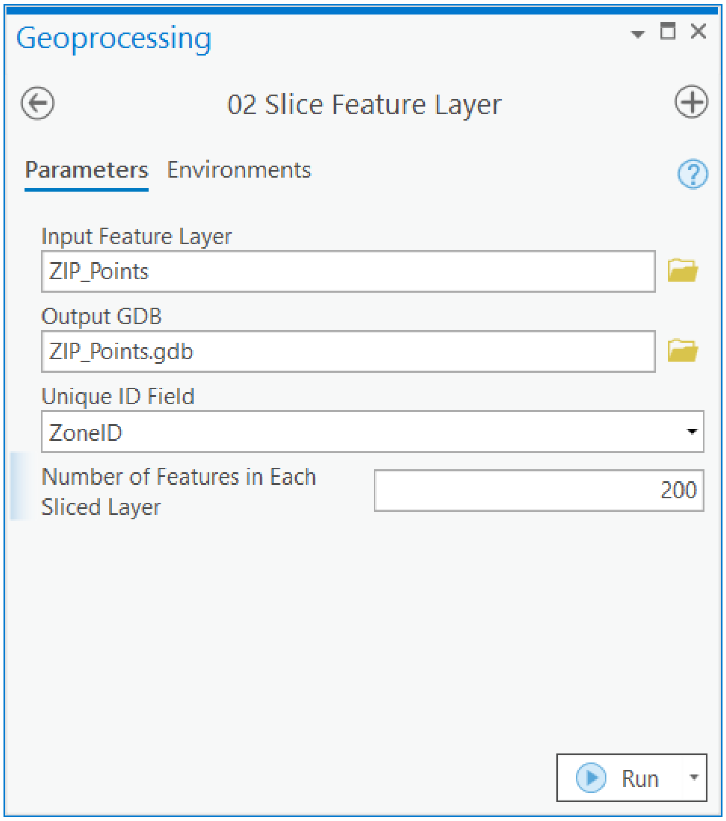
Input Network Layer: FL_Rd2012

Max Snap Distance (Optional): 500 Meters

Figure A2 Interface for “snap points to road network” setting

(3) 02 Slice Feature Layer

This tool is to slice the total number of the input point features into multiple features with the same field attributes. The purpose is to speed up the data processing, particularly when the data is large. In other words, if the point feature layer is small, no need to use this tool. Figure A3 shows the interface. Note that the output GDB should be an empty geodatabase and each generated feature class ends with, for example, _0, _200, and _400 if the number of features in each is 200 for an input feature of 600 points.



Input Feature Layer: ZIP_Points

Output GDB: ZIP_Points.gdb

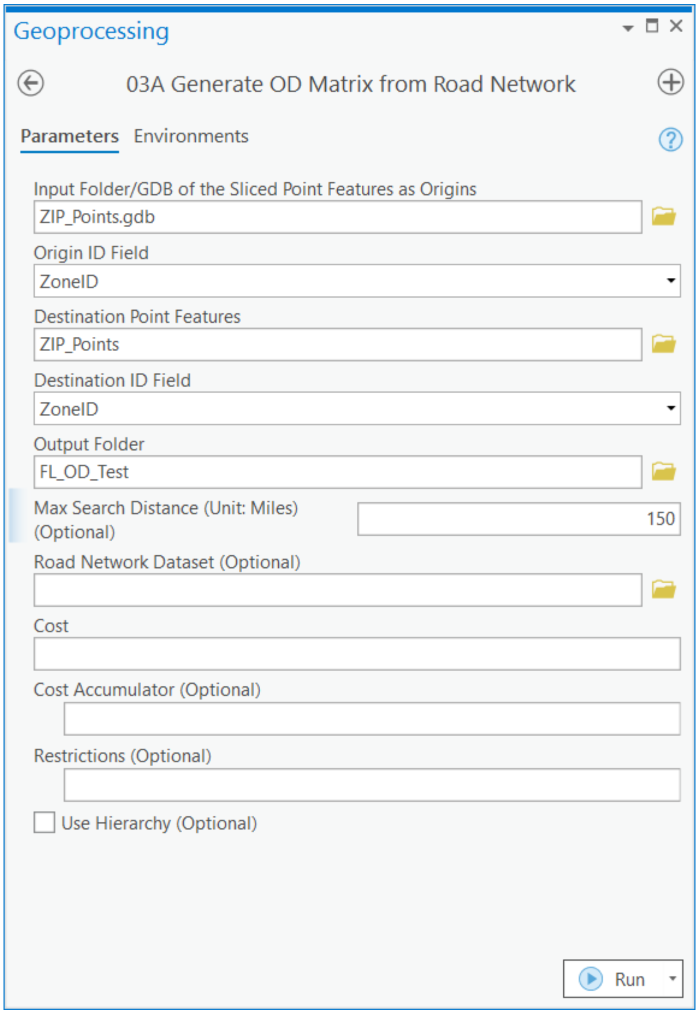
Unique ID Field: ZoneID

Number of Features in Each Sliced Layer: 200

Figure A3 Interface for “slice feature layer” setting

(4) 03A Generate OD Matrix from Road Network

This tool is to generate an OD cost matrix between points based on the default network setting in the first tool (Figure A4). If the first tool has not been used, do not leave the last four items blank in this tool. If the input road network dataset is a complete road network with all levels of roads, input 150 for Max Search Distance (equivalent to 180 minutes under the speed of 50 mph), this tool will then generate OD cost matrix at Level 1 (0-3hours); if the input road network dataset is a simplified road network with major highways, input 300 for Max Search Distance (=6 hours), this tool will generate OD cost matrix at Level 2 (0-6hours); if users want to calculate the geodesic distances between OD pairs at Level 3, no need to use this tool. Note that Level 1 and Level 2 may have some overlapping records in terms of the composite form of the origin and destination ID fields with non-zero travel times and distances. When using the “`05B Write All OD Pairs with Geodesic Distance`,” only the records at Level 1 will be retained. If the OD cost matrix is computed between the input point features, the Destination Point Features should be a full layer of the input point features without slicing. For example, `ZIP_Points.gdb` for Input Folder/GDB contains all sliced feature classes and `ZIP_Points` for Destination Point Features refers to the same one without slicing.



Geoprocessing

03A Generate OD Matrix from Road Network

Parameters Environments

Input Folder/GDB of the Sliced Point Features as Origins: ZIP_Points.gdb

Origin ID Field: ZoneID

Destination Point Features: ZIP_Points

Destination ID Field: ZoneID

Output Folder: FL_OD_Test

Max Search Distance (Unit: Miles) (Optional): 150

Road Network Dataset (Optional):

Cost:

Cost Accumulator (Optional):

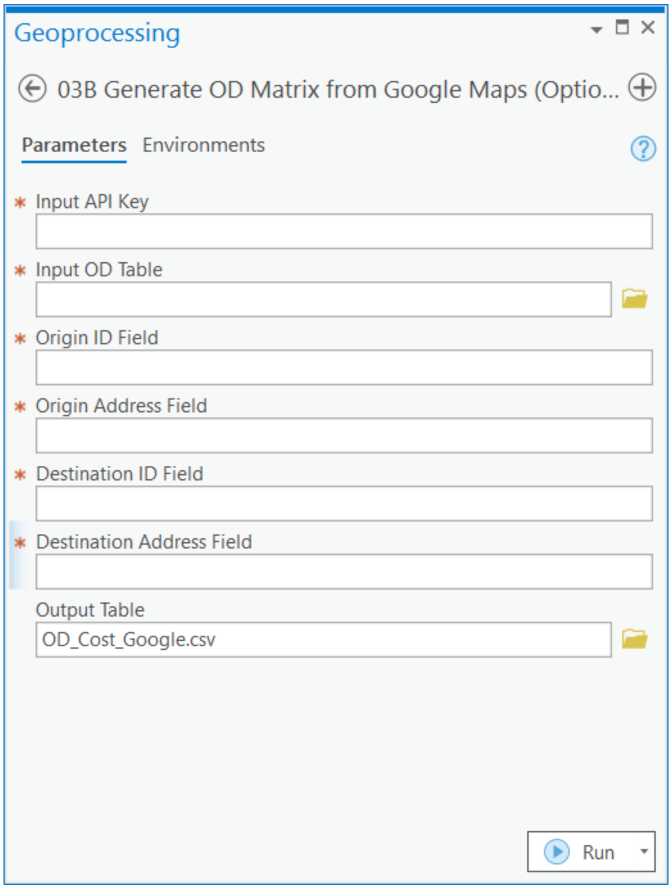
Restrictions (Optional):

□ Use Hierarchy (Optional)

Figure A4 Interface for "generate OD matrix from road network"

(5) 03B Generate OD Matrix from Google Maps (Optional)

This tool is to generate OD cost matrix based on the Google Maps Distance Matrix API that provides travel time and distance for a set of origins and a set of destinations (Figure A5). There might be two scenarios to use this tool. One is to obtain OD cost matrix from Google Maps based on the addresses of origins and destinations and to use the derived travel costs directly. Another is to use an existing OD cost matrix from Google Maps to adjust the one preliminarily derived from a road network. Users need to run related bivariate regression models and sampling and input the two coefficients from the regression models in the calibration tool. If users prefer to use the default coefficients calibrated from Hu et al. (2020), no need to use this tool.


Geoprocessing

03B Generate OD Matrix from Google Maps (Optio...

Parameters Environments

* Input API Key

* Input OD Table

* Origin ID Field

* Origin Address Field

* Destination ID Field

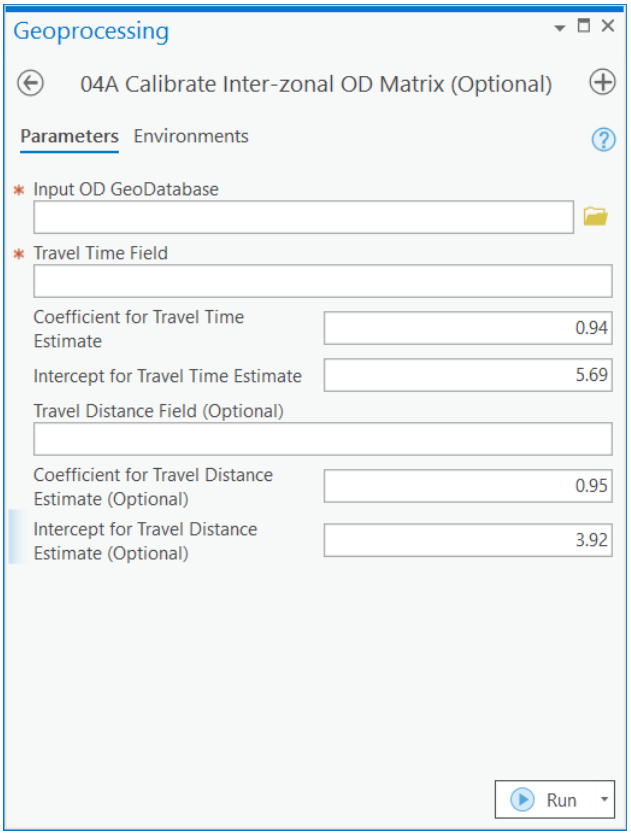
* Destination Address Field

Output Table: OD_Cost_Google.csv

Figure A5 Interface for “generate OD matrix from Google Maps”

(6) 04A Calibrate Inter-zonal OD Matrix (Optional)

The optional tool is to calibrate the preliminary inter-zonal OD cost matrix at Level 1 and Level 2 by the travel cost matrix from Google Maps, respectively. As shown in Figure A6, the default values for coefficients and intercepts of travel time and distance are from Hu et al. (2020). Users can also define these parameters based on their own regression models. One scenario for the usage is to calibrate the travel time and/or distance at Level 1 from the tool of “`03A Generate OD Matrix from Road Network`” by default values. Users are required to set the coefficients and intercepts if the OD cost matrix at Level 2 is calibrated.



* Input OD GeoDatabase

* Travel Time Field

Coefficient for Travel Time Estimate 0.94

Intercept for Travel Time Estimate 5.69

Travel Distance Field (Optional)

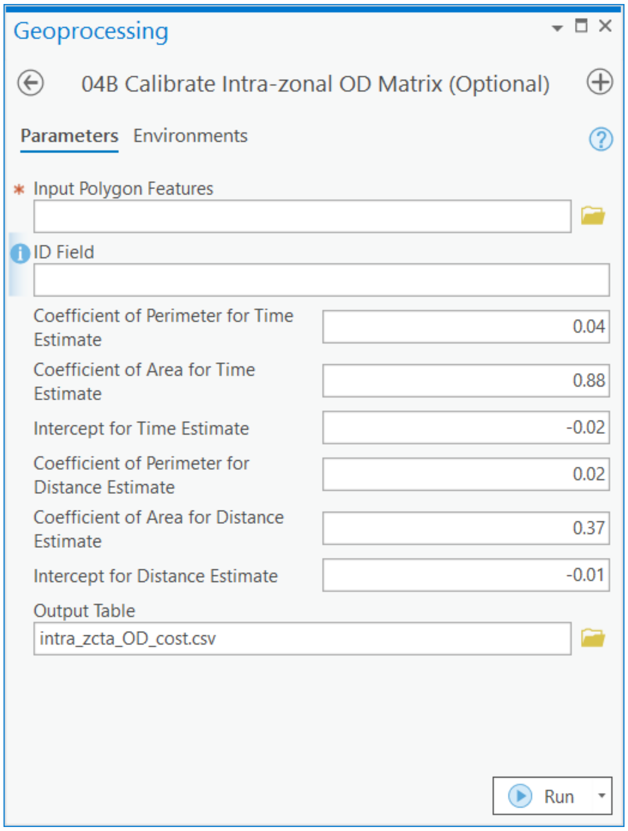
Coefficient for Travel Distance Estimate (Optional) 0.95

Intercept for Travel Distance Estimate (Optional) 3.92

Figure A6 Interface for “calibrate inter-zonal OD matrix”

(7) 04B Calibrate Intra-zonal OD Matrix (Optional)

The optional tool is to calibrate the intra-zonal OD matrix by the perimeter and area size of the input polygon features. As shown in Figure A7, the default values for the coefficients and intercepts of travel time and distance are derived from Hu et al. (2020). Users can append this intra-zonal OD matrix to the inter-zonal OD matrix derived from the tool of “`05B Write All OD Pairs with Geodesic Distance`” to obtain a more accurate estimate of the OD drive time matrix, especially for those short-range OD pairs (Hu et al. 2020, p.5).



04B Calibrate Intra-zonal OD Matrix (Optional)

* Input Polygon Features

ID Field

Coefficient of Perimeter for Time Estimate: 0.04

Coefficient of Area for Time Estimate: 0.88

Intercept for Time Estimate: -0.02

Coefficient of Perimeter for Distance Estimate: 0.02

Coefficient of Area for Distance Estimate: 0.37

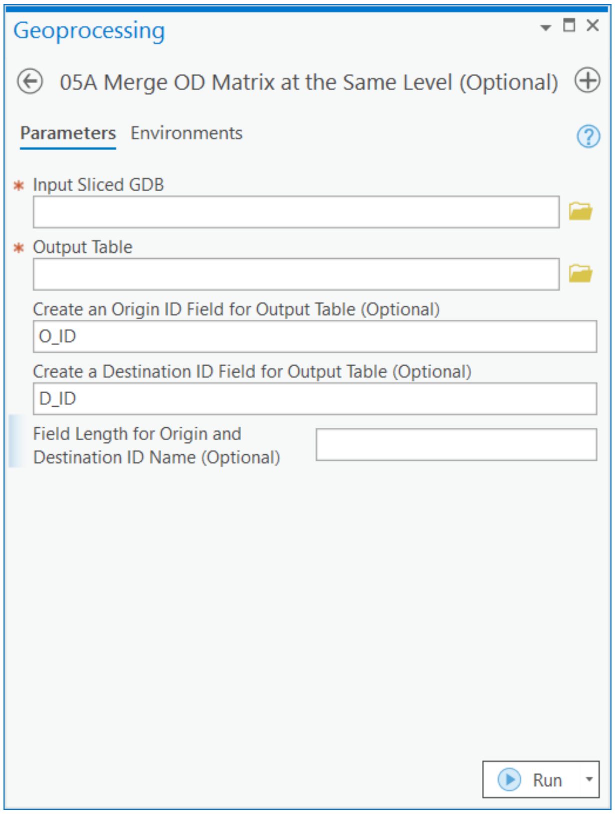
Intercept for Distance Estimate: -0.01

Output Table: intra_zcta_OD_cost.csv

Figure A7 Interface for “calibrate intra-zonal OD matrix”

(8) 05A Merge OD Matrix at the Same Level (Optional)

The optional tool is to merge all previously sliced OD cost matrices to one matrix at the same level (Figure A8). For example, there are three sliced OD matrices estimated at Level 1, use this tool to merge them to one matrix at Level 1; if the sliced OD matrices are at Level 2, use this tool to merge them to one at Level 2.



Geoprocessing

05A Merge OD Matrix at the Same Level (Optional)

Parameters Environments

* Input Sliced GDB

* Output Table

Create an Origin ID Field for Output Table (Optional)

O_ID

Create a Destination ID Field for Output Table (Optional)

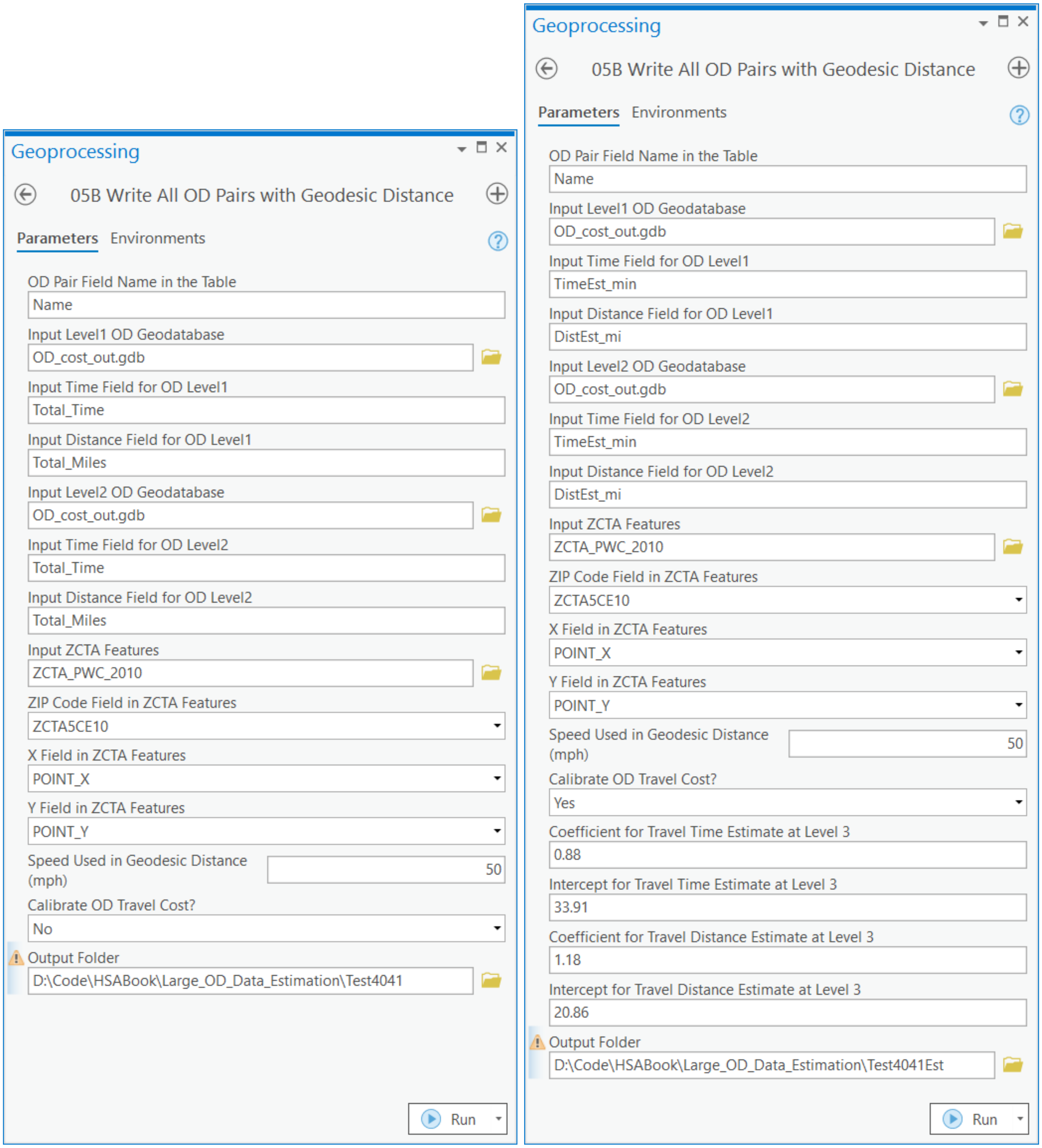
D_ID

Field Length for Origin and Destination ID Name (Optional)

Figure A8 Interface for “merge OD matrix at the same level”

(9) 05B Write All OD Pairs with Geodesic Distance

This tool is to merge the inter-zonal OD cost matrix at Level 1 and Level 2 and calculate the geodesic distances for remaining OD pairs at Level 3. As designed in Hu et al. (2020), if the travel time and/or distance for an OD pair is not found at Level 1, it goes to Level 2, and then Level 3 to extract the travel cost. As shown in Figure A9, the default value Name for the first item is generated by the tool of “03A Generate OD Matrix from Road Network” and corresponds to the composite form of Origin ID Field and Destination ID Field with a connection symbol “ - ”. Users need to input two geodatabases of OD matrices at Level 1 and 2, respectively. Note that each feature class in two geodatabases must have the identical fields represent the origin ID field and destination ID field, respectively. The remaining OD pairs not found at Level 1 and 2 will be computed based on the latitude (Y Field) and longitude (X Field) of the input ZCTA point or polygon features, i.e., the feature class ZCTA_PWC_2010 under ZCTA_PWC.gdb. The ZIP code field should be identical to the origin ID field and destination ID field in the two geodatabases. If users only want to obtain OD cost matrix at Level 1, 2, and 3 without any calibration, select the original estimated travel time and distance for Level 1 and 2, and select the default value No for the Calibrate OD Travel Cost (see Figure A9a). If users want to obtain a calibrated OD cost matrix at three levels, select the calibrated travel time and distance for Level 1 and 2, and select Yes for the Calibrated OD Travel Cost. Then the default values for the coefficients and intercepts in Hu et al. (2020) will be shown in this tool (see Figure A9b).



Geoprocessing

05B Write All OD Pairs with Geodesic Distance

OD Pair Field Name in the Table: Name

Input Level1 OD Geodatabase: OD_cost_out.gdb

Input Time Field for OD Level1: Total_Time

Input Distance Field for OD Level1: Total_Miles

Input Level2 OD Geodatabase: OD_cost_out.gdb

Input Time Field for OD Level2: Total_Time

Input Distance Field for OD Level2: Total_Miles

Input ZCTA Features: ZCTA_PWC_2010

ZIP Code Field in ZCTA Features: ZCTA5CE10

X Field in ZCTA Features: POINT_X

Y Field in ZCTA Features: POINT_Y

Speed Used in Geodesic Distance (mph): 50

Calibrate OD Travel Cost?: No

Output Folder: D:\Code\HSABook\Large_OD_Data_Estimation\Test4041

Geoprocessing

05B Write All OD Pairs with Geodesic Distance

OD Pair Field Name in the Table: Name

Input Level1 OD Geodatabase: OD_cost_out.gdb

Input Time Field for OD Level1: TimeEst_min

Input Distance Field for OD Level1: DistEst_mi

Input Level2 OD Geodatabase: OD_cost_out.gdb

Input Time Field for OD Level2: TimeEst_min

Input Distance Field for OD Level2: DistEst_mi

Input ZCTA Features: ZCTA_PWC_2010

ZIP Code Field in ZCTA Features: ZCTA5CE10

X Field in ZCTA Features: POINT_X

Y Field in ZCTA Features: POINT_Y

Speed Used in Geodesic Distance (mph): 50

Calibrate OD Travel Cost?: Yes

Coefficient for Travel Time Estimate at Level 3: 0.88

Intercept for Travel Time Estimate at Level 3: 33.91

Coefficient for Travel Distance Estimate at Level 3: 1.18

Intercept for Travel Distance Estimate at Level 3: 20.86

Output Folder: D:\Code\HSABook\Large_OD_Data_Estimation\Test4041Est

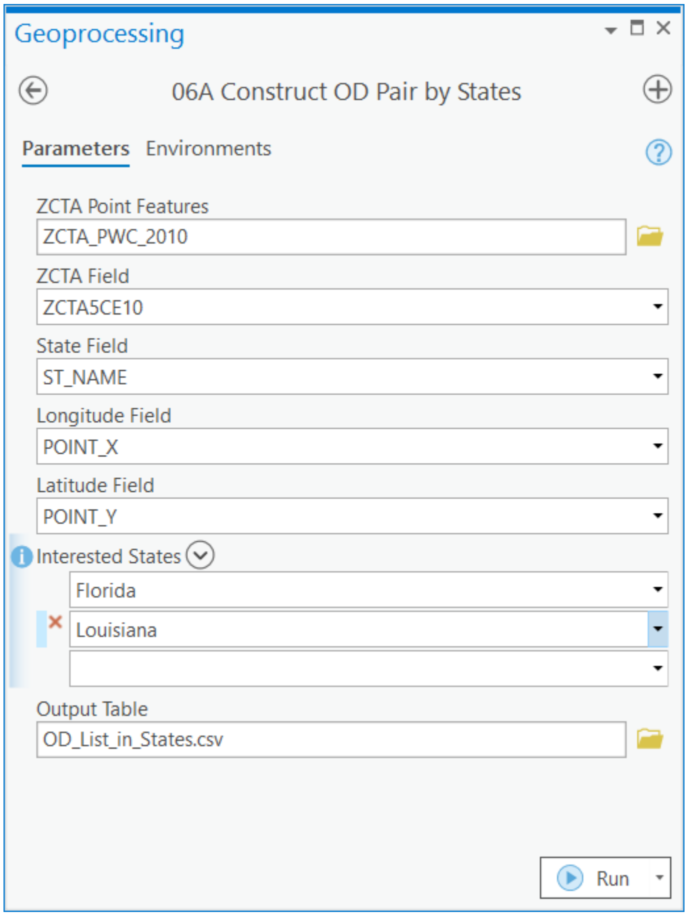
Figure A9 Interfaces for writing (a) all uncalibrated OD pairs and (b) all calibrated OD pairs

4. Extracting ZIP-to-ZIP OD cost matrix

(10) 06A Construct OD Pair by States

This tool is to generate OD pairs with origin ZIP codes and the associated geographic coordinates, destination ZIP codes and the associated geographic coordinates in selected states.

The derived OD pairs will be input in the tool of “`07 Extract Calibrated Travel Cost using OD pairs`” to obtain the calibrated travel time and distance (see Figure A12). As shown in Figure A10, the `ZCTA_PWC_2010` (N=32,840) contains four fields `ZCTA5CE10`, `ST_NAME`, `POINT_X`, and `POINT_Y`, and say, users want to create OD pairs between ZIP codes in `Florida` and `Louisiana`. The output table will have all OD ZIP code pairs in these two states.



ZCTA Point Features: ZCTA_PWC_2010

ZCTA Field: ZCTA5CE10

State Field: ST_NAME

Longitude Field: POINT_X

Latitude Field: POINT_Y

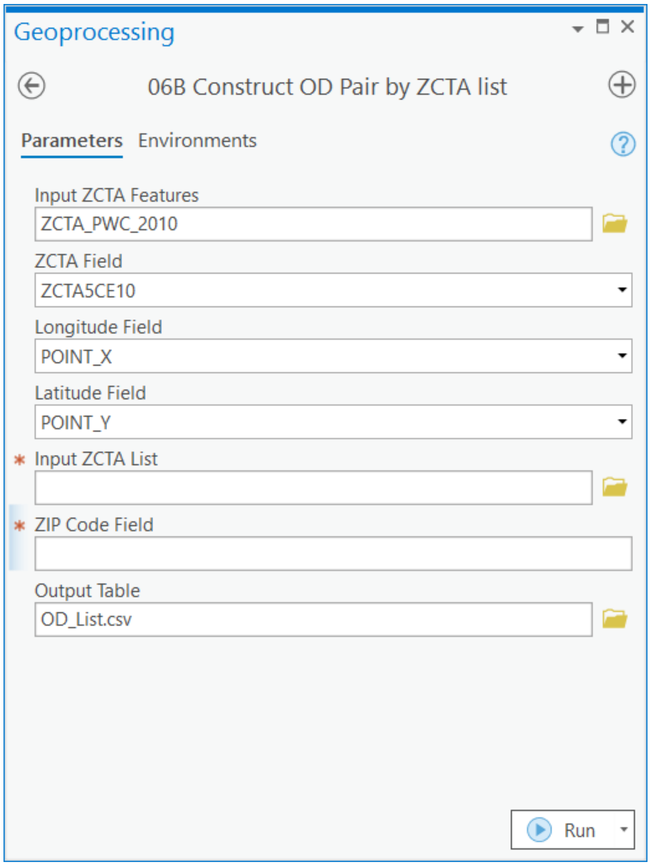
Interested States: Florida, Louisiana

Output Table: OD_List_in_States.csv

Figure A10 Interface for “construct OD pairs by states”

(11) 06B Construct OD Pair by ZCTA list

This tool is to generate OD pairs with origin ZIP codes and the associated geographic coordinates, destination ZIP codes and the associated geographic coordinates based on a list of ZIP codes. As shown in Figure A11, the input ZCTA list should have a ZIP code field that corresponds to the ZCTA field in the input ZCTA features. Only the matched OD pairs will be generated, and users can then use the tool “`07 Extract Calibrated Travel Cost using OD pairs`” to obtain the calibrated travel time and distance (see Figure A12). For those unmatched ZIP codes, refer to Hu et al. (2020, p.8).



Geoprocessing

06B Construct OD Pair by ZCTA list

Input ZCTA Features: ZCTA_PWC_2010

ZCTA Field: ZCTA5CE10

Longitude Field: POINT_X

Latitude Field: POINT_Y

* Input ZCTA List:

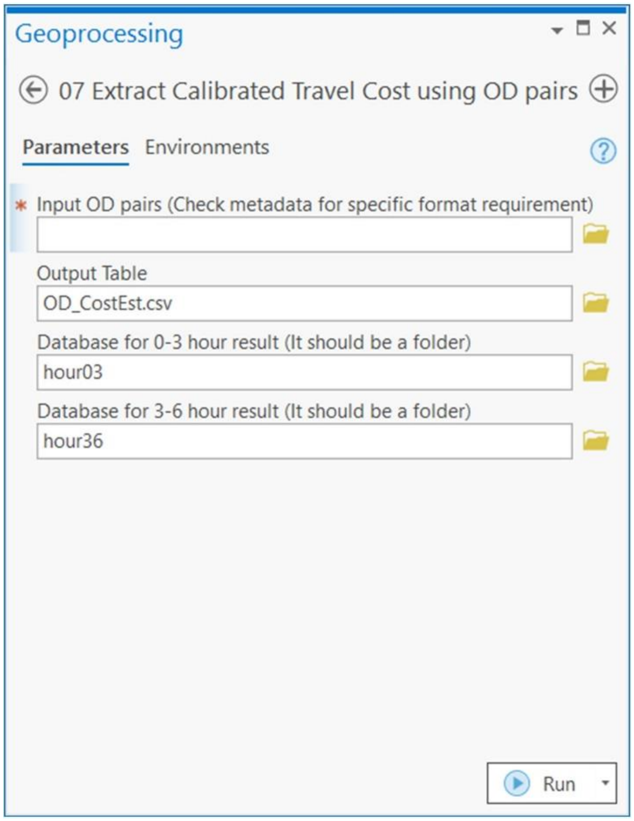
* ZIP Code Field:

Output Table: OD_List.csv

Figure A11 Interface for “construct OD pair by ZCTA list”

(12) 07 Extract Calibrated Travel Cost using OD pairs

This tool is to extract the calibrated OD travel time and distance from our provided dataset. For those not found at Level 1 or Level 2, this tool will regard them to be OD pairs at Level 3 and automatically compute the geodesic distance and travel time under the speed of 50 mph, and then calibrate both by the coefficients and intercepts reported in Hu et al. (2020, p.8). As shown in Figure A12, the Databases for 0-3 hours and 3-6 hours correspond to the subfolder `hour03` and `hour36` under the folder `US_OD_Cost_Calibrated_Data`. The input OD pairs can be those generated from the tools of “`06A Construct OD Pair by States`” or “`06B Construct OD Pair by ZCTA list`” or pre-existed OD data that contain fields `OZCTA`, `DZCTA`, `olat`, `olong`, `dlat`, and `dlong` accordingly. For those unmatched OD pairs, refer to Hu et al. (2020, p.8). Note that it may be time consuming for a large OD pair data (e.g., for Florida and Louisiana, our experiment took 1 hour and 7 minutes).



* Input OD pairs (Check metadata for specific format requirement)

Output Table: OD_CostEst.csv

Database for 0-3 hour result (It should be a folder): hour03

Database for 3-6 hour result (It should be a folder): hour36

Figure A12 Interface for “extract calibrated travel cost using OD pairs”