User Guide: Estimating a Large OD Drive Time Matrix

A.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 the 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 USA. 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 Maps 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 and county.

4. Extract a subset of OD ZIP code pairs from our calibrated dataset.

A.2 Data and Programs

The folder "`Large_OD_Data_Estimation`" contains data and program tools such as:

1. `ZCTA_PWC.gdb`: It 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_0-3hours.csv`: It represents the calibrated OD drive time matrix at Level 1 in (0,3] hours.

3. `NA_OD_3-6hours.csv`: It represents the calibrated OD matrix at Level 2 in (3,6] hours.[1]

4. `Generate OD Cost Pro.tbx`: It is implemented in ArcGIS Pro to estimate OD drive time and extract calibrated data. Detailed descriptions are provided in the next section.

---

[1] In rare (188) cases, applying the regression models yielded negative values for estimated distances. We used the original values plus the intrazonal distances to adjust the estimated distances.

5. `LargeODcost`: It contains the Python scripts used in `Generate OD Cost Pro.tbx`.

6. `US_OD_Cost_Calibrated_Data`: It contains two folders, `hour03` and `hour36`, to store the OD matrices at Levels 1 and 2 in ".`data`" format for fast retrieval.[2] They will be used in the tool `Generate OD Cost Pro.tbx` to extract a subset of the OD matrices.

The `NA_OD_0-3hours.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.

A.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.

---

[2] Under the two folders, each subfolder is named by the first three digits of the origin ZIP code. Under each subfolder, each file name ended with .`data` is named by the last two digits of the origin ZIP code, and each record contains the five-digit destination ZIP code, travel time in minutes, and travel distance in miles.

In ArcGIS Pro, under Catalog view, right-click Toolboxes and 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 Point 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`. To do this, right-click each tool and 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 A.1). 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.
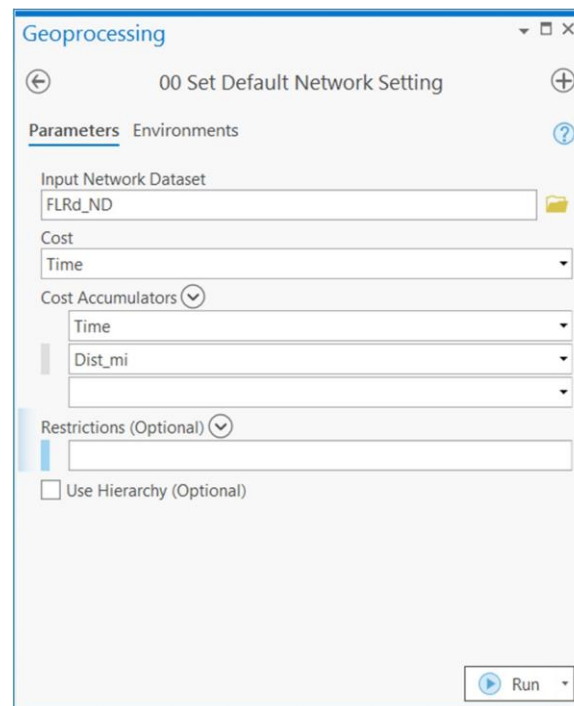
Figure A.1 Interface for default network environment setting.

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 A.2, the Output Folder should be created by users and the optional Max Snap Distance has a default value of 500 meters. This tool will create a temporary geodatabase named `temp.gdb` under the output folder.
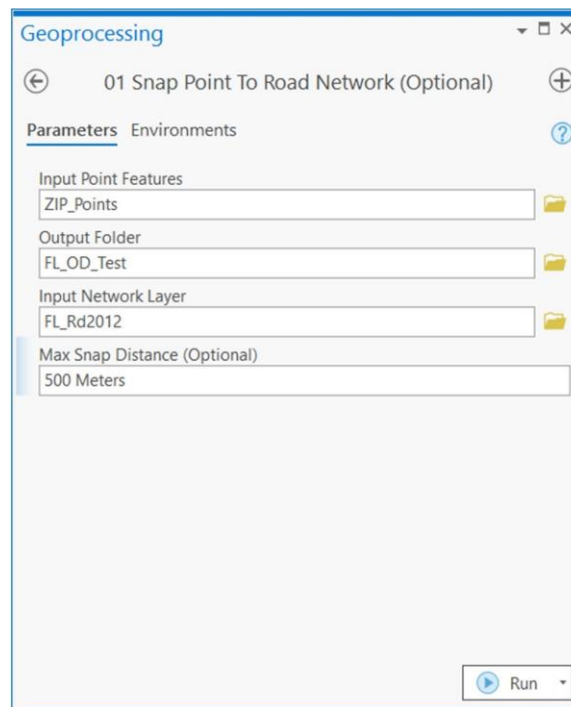

Figure A.2 Interface for "Snap Point to Road Network (Optional)" setting.

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 are large. In other words, if the point feature layer is small, no need to use this tool. Figure A.3 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.
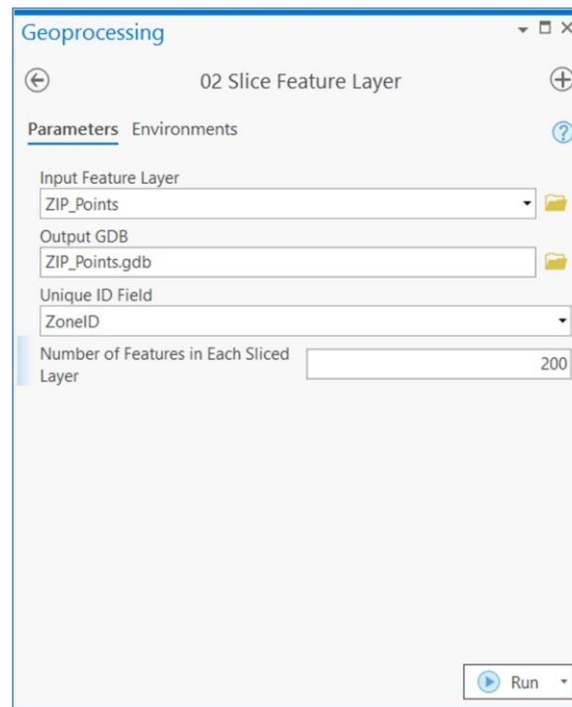
Figure A.3 Interface for "Slice Feature Layer" setting.

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 A.4). 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 at 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.
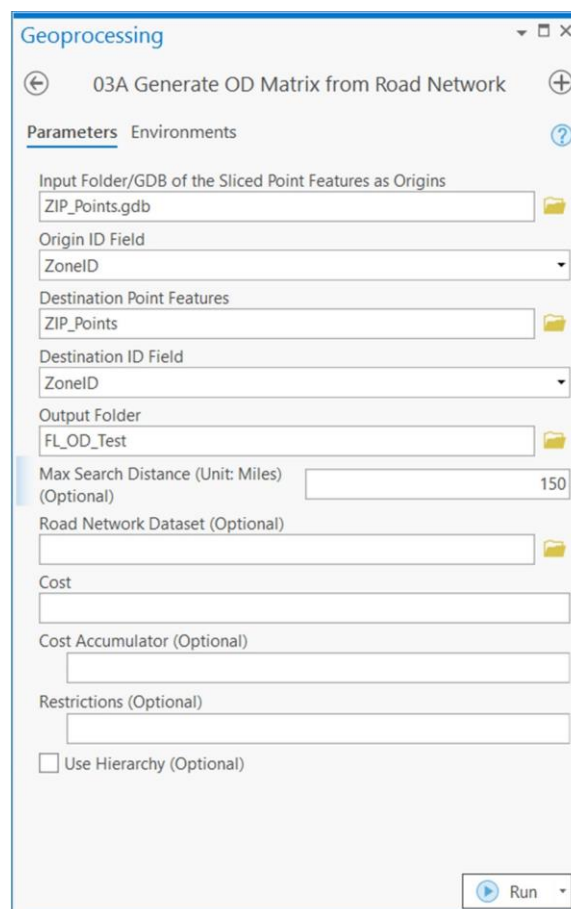
Figure A.4 Interface for "Generate OD Matrix from Road Network".

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 A.5). 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.

Figure A.5 Interface for "Generate OD Matrix from Google Maps (Optional)".

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 A.6, 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.

Figure A.6 Interface for "Calibrate Inter-zonal OD Matrix (Optional)".

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 A.7, 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).
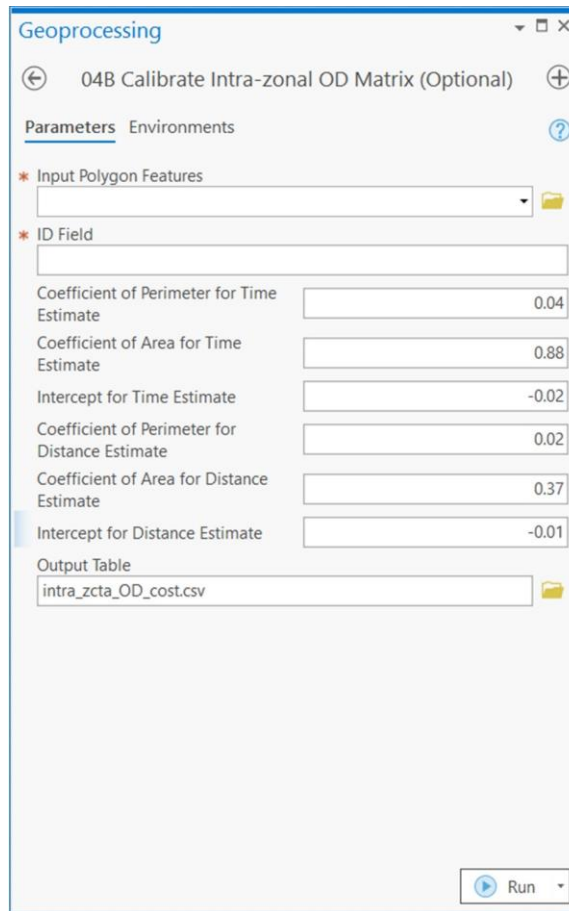
Figure A.7 Interface for "Calibrate Intra-zonal OD Matrix (Optional)".

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 A.8). 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.
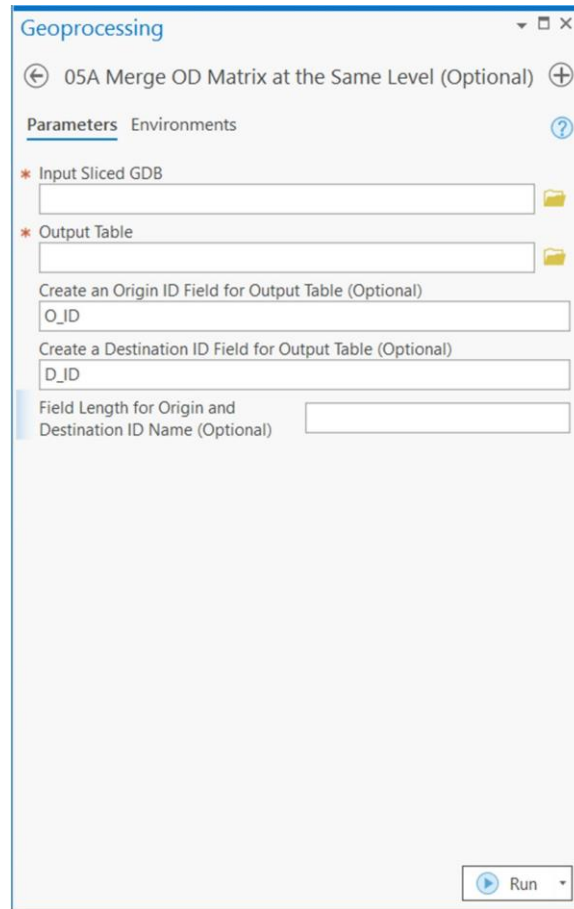
Figure A.8 Interface for "Merge OD Matrix at the Same Level (Optional)".

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 to Level 3 to extract the travel cost. As shown in Figure A.9, 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 Level 2, respectively. Note that each feature class in the two geodatabases must have the identical fields representing the origin ID field and destination ID field, respectively. The

remaining OD pairs not found at Level 1 and Level 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 Levels 1, 2, and 3 without any calibration, select the original estimated travel time and distance for Level 1 and Level 2 and select the default value `No` for the Calibrate OD Travel Cost (see Figure A.9a). If users want to obtain a calibrated OD cost matrix at three levels, select the calibrated travel time and distance for Level 1 and Level 2 and select `Yes` for the Calibrate 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 A.9b).
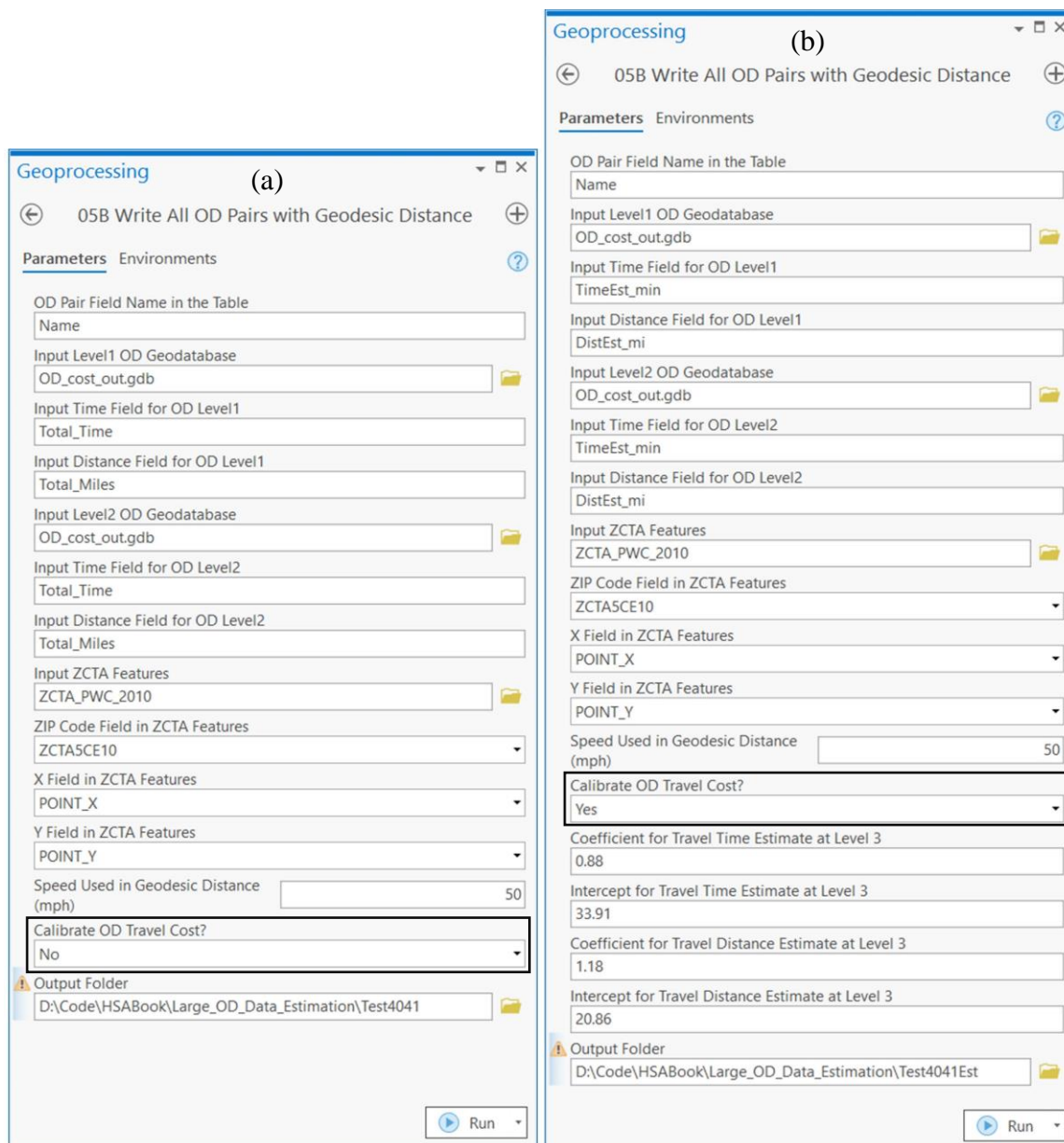
Figure A.9 Interfaces for writing (a) all uncalibrated OD pairs and (b) all calibrated OD pairs.

A.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 A.12). As shown in Figure A.10, 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.
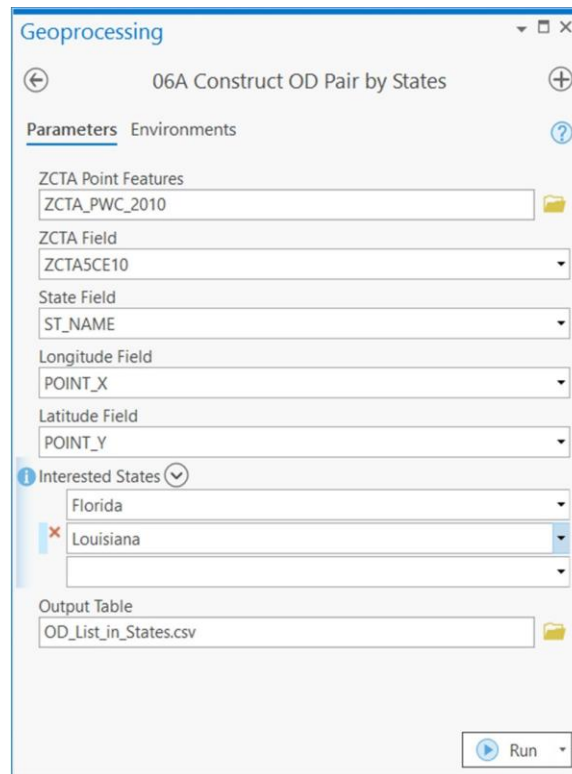

Figure A.10 Interface for "Construct OD Pair by States".

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 A.11, 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 A.12). For those unmatched ZIP codes, refer to Hu et al. (2020, p. 8).
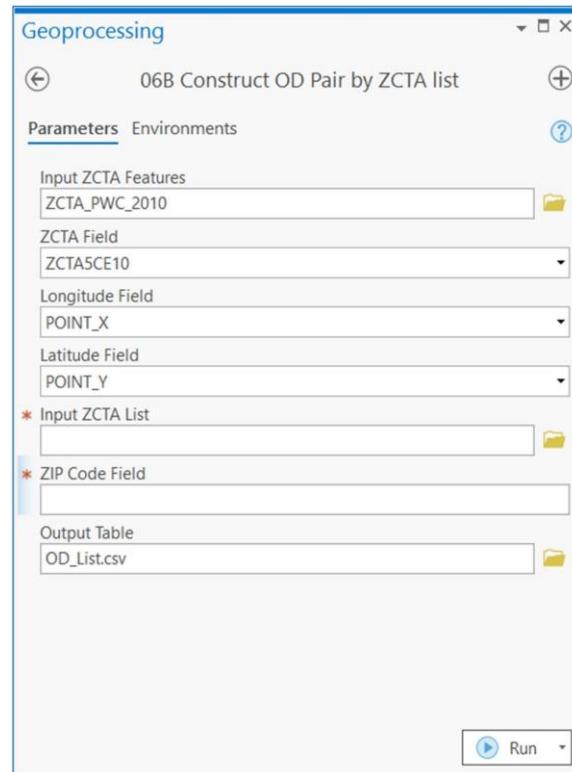

Figure A.11 Interface for "Construct OD Pair by ZCTA list".

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 at the speed of 50 mph, and then calibrate them by both the regression models reported in Hu et al. (2020, p. 8). As shown in Figure A.12, 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-existing 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 dataset.

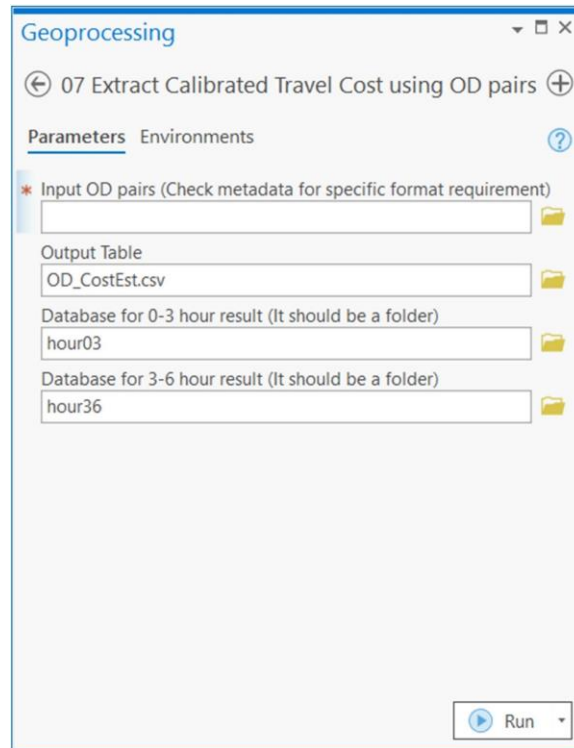For example, for Florida and Louisiana, our experiment took 1 hour and 7 minutes.

Figure A.12 Interface for "Extract Calibrated Travel Cost using OD pairs".