This document describes the execution flow of the codes developed for the manuscript titled "An Evaluation of Geo-located Twitter Data for Measuring Human Migration". The code set includes six steps of the data processing and analysis, which is illustrated in Figure 1. Readers should check out the README.txt in each folder before using the code and data.

Step 1: Raw Twitter data extraction

A script that utilizes Apache Pig to extract "Twitter_userID, latitude, longitude, unix_timestamp" from raw Twitter data

Code: "**Code\Step1_tweets_extraction"**

Data: A sample set of raw geo-located tweets in "**Data\step1_input_sample_raw_twitter_data\ 2013_01_31_stream_sample.txt"**

Step 2: Code for assigning each geo-located tweet to a corresponding county

A R script to assign the geo-location of each tweet to its corresponding US county (i.e., point in polygon)

Code: "**Code\Step2_Tweets_to_County\PointInPolygon.R"**

Data: "**Data\step2_input_tweets_to_county\example_input.txt"; US county shapefile: "Data\us_county.zip"**

Step 3: Code for generating a trajectory for each twitter user and determining his/her county-of-residence

Two MapReduce programs that first generate a trajectory for each user and then determine the user's county-of-residence

Code: **1. Generating trajectories, 2. Determining county-of-residence**

1: "Code\Step3_Tweets_to_Trajectory_to_Resident_county\make_trajectories_county"

2: "Code\Step3_Tweets_to_Trajectory_to_Resident_county\Twitter_County"

Data: **"Data\step3_input_tweets_to_trajectory_flow\ example_input.txt"**

Step 4: Code for generating county-to-county migration flows

Python code to generate county-to-county migration flows based on change of resident county over two years.

Code: "**Code\Step4_county_to_conty_flows\ flow_13_14_inflow.py**"

Data: **"Data\step4_input_flows\us_flow_13_example.txt";**

"Data\step4_input_flows\us_flow_15_example.txt"

Step 5: 4D spatial scan statistic approach based on Poisson process model

Implementation of 4D spatial scan based on Poisson process model using C language

Code**: "Code\Step5_4DSpatialScan\poisson\src"**

Data**: "Data\step5_input_spatialScan_data\twitter_IRS_inflow_pairs_coords_proj_1314.csv"**

Final step: Code for data analysis

Python scripts for data analysis: Mapping/visualizing the migration flows and results from the spatial scan statistics

Code: "**Code\Step6_Analysis**"

Data: "**Data\step6_data**"

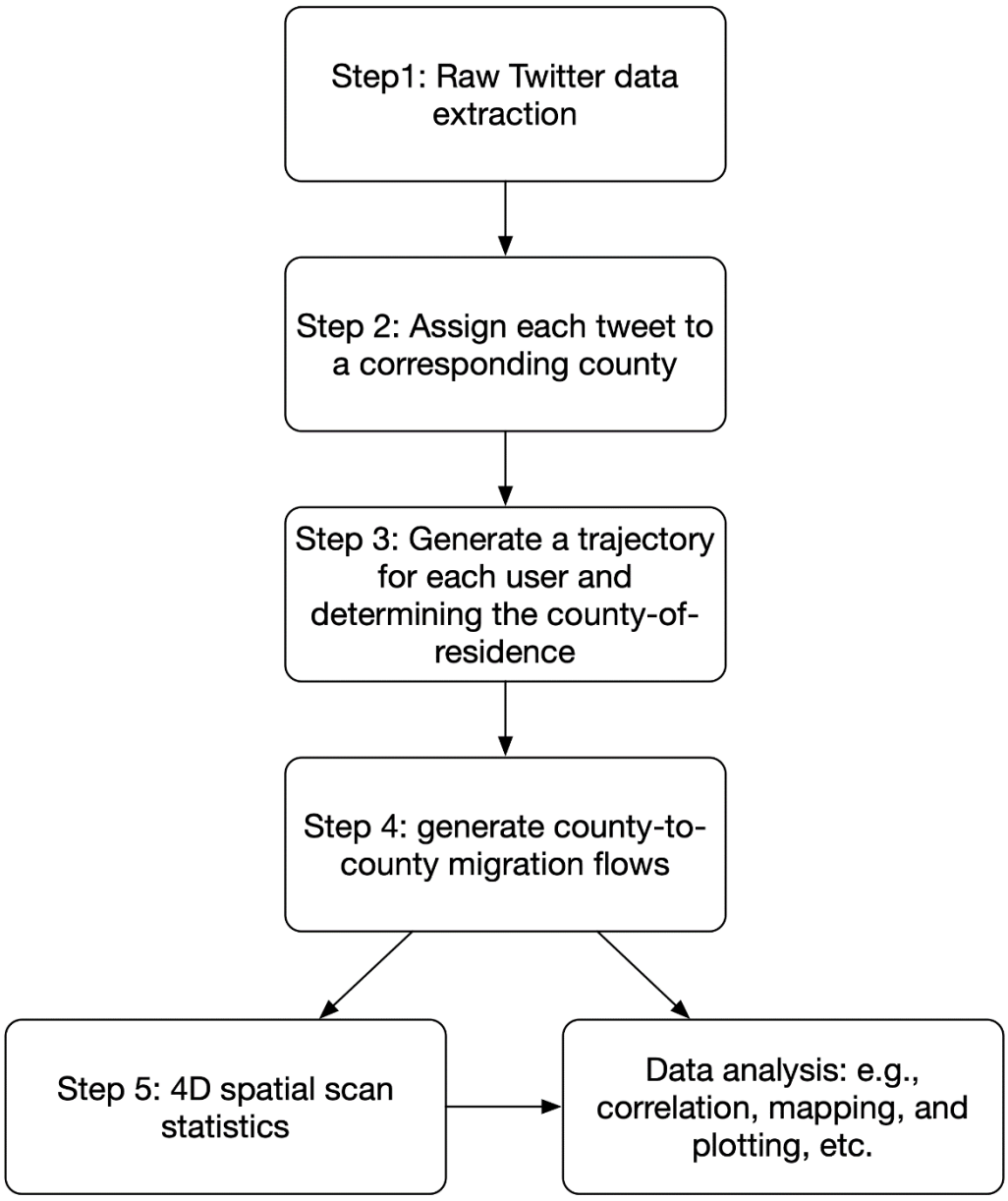

Figure 1: Illustration of the execution flow of the developed code