Cyrus Blankinship
Spatial Analysis – Project 2
Prof. Paul Torrens
11/26/2018

Night-time NYC Spatial Interaction Model

Introduction

What are the main drivers of night-time pedestrian flow between New York City neighborhoods? This paper attempts to answer this question by developing the four spatial interaction models defined by Alan Wilson in his 1971 paper, *A family of spatial interaction models, and associated developments*. I will look at each model in turn and refine each by calibrating the parameters through Poisson based generalized linear regression. Finally, I will select which model best describes my gathered data.

Data Sources

To develop a spatial interaction model, I first needed some form of flow data. Though it is by no means fully representative of the city's population, *taxi trip information* provided by the Taxi and Limousine Commission (TLC) serves this purpose well by containing both origin and destination location information, as well as the number of passengers, fare amount, and distance and time travelled (which could each be used as the cost measure in my models).

Next, I needed geographic divisions to aggregate my flow data. For this, I determined *traffic analysis zones* (TAZs) to be most appropriate. These are the basic spatial units of analysis used by transportation planners to forecast trip volumes and analyze transportation capacities, and is provided in a spatial format by the US Census Bureau.

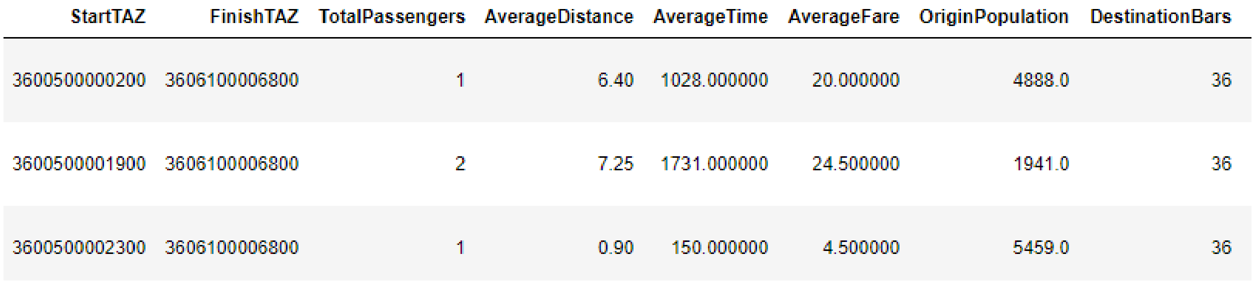
Finally, I gathered origin and destination attributes that would serve as the structural variables in models. For origins, I gathered the *total population* using American Census Survey (ACS) population data. For destinations, I gathered the *locations of every bar in NYC* as determined by onsite liquor licenses.

Data Preparation & Visualization

The TLC dataset needed to be aggregated and transformed into a matrix that had the total flow between each TAZ to TAZ pair, as well as our three predictor variables (distance, total population, and number of bars). The first step was to spatially join the 'GEOID' of each TAZ to both the origin and destination of each individual trip in my TLC dataset. I then aggregated the total number of passengers by each unique TAZ pairing, which gave me my flow count. I similarly performed a spatial join/aggregation for my bar location data.

Because the ACS provides population statistics by census blocks, I performed a spatial intersection with the TAZ boundaries and aggregated based on the total population of each/divided by its percentage of original census block area. The end result was a total population per TAZ.

In order to calculate distance, I created a distance matrix by calculating a simple Euclidean distance measure between each TAZ to TAZ pair. However, I quickly realized that a better measure of distance was already included in the TLC data in the form of fare amount, trip time, and distance travelled. I therefore included those measures in my TLC aggregation instead of the values from my distance matrix.



StartTAZ	FinishTAZ	TotalPassengers	AverageDistance	AverageTime	AverageFare	OriginPopulation	DestinationBars
3600500000200	3606100006800	1	6.40	1028.000000	20.000000	4888.0	36
3600500001900	3606100006800	2	7.25	1731.000000	24.500000	1941.0	36
3600500002300	3606100006800	1	0.90	150.000000	4.500000	5459.0	36

Table 1: The top three rows from my final flow dataset (out of 138,028). Notice that these all go to the same destination Transportation Access Zone, hence the repeating value for 'DestinationBars'

Next, I wanted to visualize the connection between each zone pairing, with graduated line weights for the total amount of flow. However, doing this quickly became cluttered with the sheer number of records (138,028), and I realized it was better to visualize on a TAZ by TAZ basis (See Figure 1). In order to speed up the visualization, I utilized the Kepler GL JavaScript package, which allows users to hover over each TAZ to see the corresponding flow in an interactive environment.

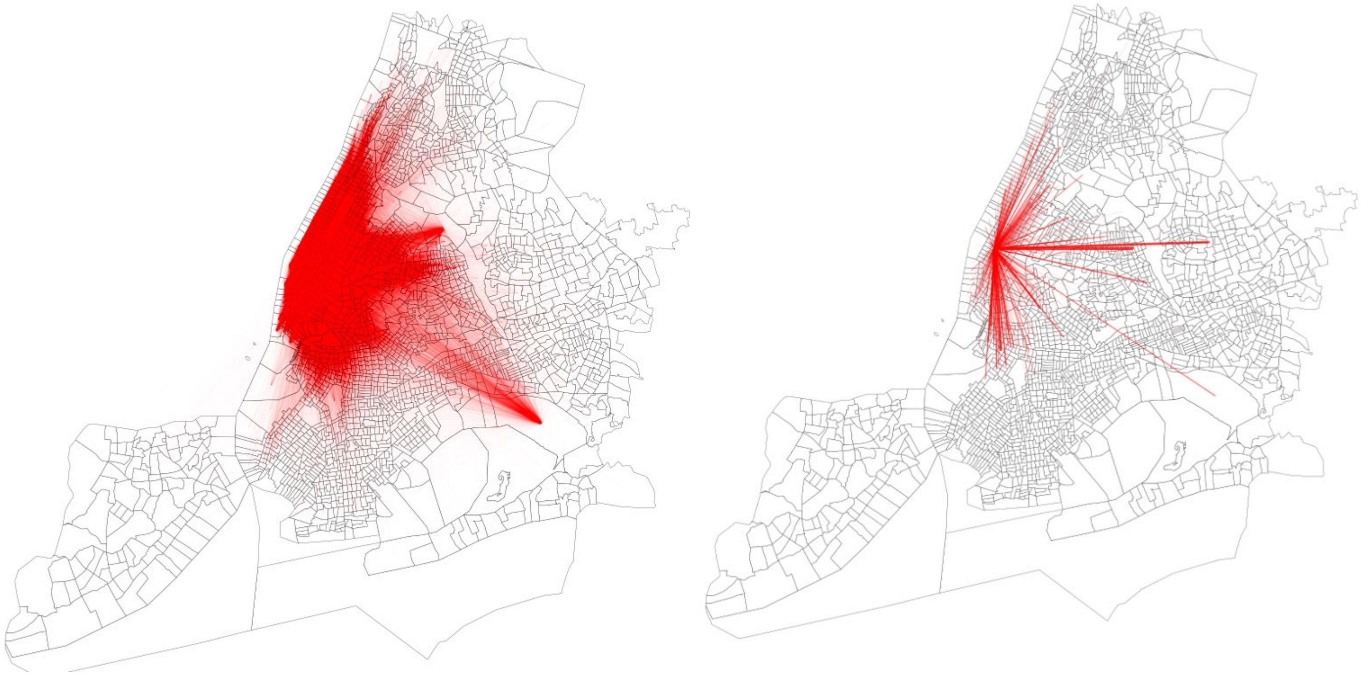


Figure 1: Flow Data visualized for the entire city (left) vs. a single TAZ (right). Line weights vary by total number of passengers of each.

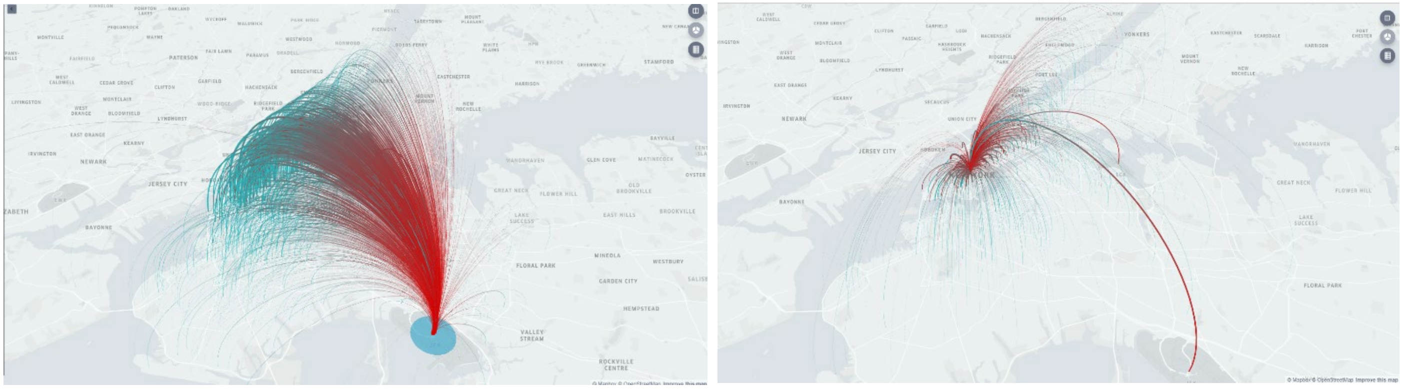


Figure 2: Kepler GL Visualization. JFK TAZ (left), East Village TAZ (right). Origins are blue, Destinations are red, and line weights increase with more Flow. Notice the sheer volume that goes to both JFK and La Guardia Airports. This will likely impact the capability of our spatial interaction models.

Spatial Interaction Modelling

Unconstrained Gravity Model

Finally, the data was ready for spatial interaction modelling using the derived values for flow, distance, origin population, and destination bars. I began with the simple *unconstrained Gravity Model.*

(1)

$$T_{ij} = k * V_i^{\mu} * W_j^{\alpha} * d_{ij}^{-\beta}$$

Equation 1: Here T is the flow between origins i and destinations j. V and W represent the structural variables ability to generate and attract trips; in our case this is the origin population and destination bars, respectively. D is the matrix of costs relating to flows between i and j: in our case this can be either average distance, fare, or time travelled. k, μ, α and β are all model parameters that we will need to calibrate.

With equation 1 in hand, I first used default values for my model parameters (k, μ, α and β). The calculation was incredibly easy, but when I tested the goodness of fit of my model, I had a very low R Squared value of 0.000048. In other words, my model had almost no predictive ability to determine flow.

(2)

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij}$$

Equation 2: This is a simple algebraic log of Equation 1. What this does is allows us to perform n linear regression in order to calibrate the parameters of our Gravity Model.

The next step was to calibrate the parameters using a generalized linear regression model (GLM), using equation 2 above. The GLM optimized the parameter values by using an ‘iteratively re-weighted least squares’ algorithm (Dennett 2018). This is slightly different than a standard regression model that’s underpinned by normal, Gaussian distributions. Instead, I specified an underlying Poisson model since the flow data was composed of non-negative integer values (i.e. non-continuous). The coefficients of the models fitted output corresponded to my model parameters, and after plugging in the new values to equation 1, I received a much-improved R Squared Value of **0.053**. Still, there was much room for improvement.

Production (Origin) Constrained Model

Next, I moved on to the next spatial interaction model defined by Wilson, the *production constrained model*

(3)

$$T_{ij} = A_i O_i W_j^{\alpha} d_{ij}^{-} \beta$$

Equation 3: Here, Ai refers to a balancing factor, similar to the k value in Equation 1. Oi refers to the total flow leaving our origin TAZ value. Wj & dij are the same as in Equation 1, along with the α and β parameters.

As with the unconstrained model, I first took the log of equation 3 in order to fit a GLM, using an underlying Poisson distribution. By constraining the origin flow, the sum of our models predicted values is the same as our actual flow data, and the output R Squared value is much higher at **0.501.**

Attraction (Destination) Constrained Model

(4)

$$T_{ij} = D_j B_j V_i^{\mu} d_{ij}^{-} \beta$$

Equation 4: This is essentially the same as Equation 3, but Dj replaces Ai and represents the balancing factor of destination flow. Bj refers to the total flow coming into our destination TAZ, and the Vj and dij values are the same as in Equation 1.

Using the same methodology as my production constrained model, I created a GLM model with optimized parameters using the log of equation 4. The corresponding R Squared value ended up being much lower in this case, with a value of **0.108**. A discussion of why this may be is included in the next section of the paper.

Doubly (Origin-Destination) Constrained Model

(5)

$$T_{ij} = A_i O_i B_j D_j d_{ij}^{-} \beta$$

Equation 5: This is simply a combination of equations 3 & 4, where we are now constraining the origin and destination values.

Finally, I tested the last spatial interaction model defined by Wilson, the *doubly constrained model.* Before I created the model, I assumed that it would be the preferable for our dataset since we have the total flow out of origins and into destinations. As predicted, the output R Squared Value was much higher at **0.812**

Conclusions

Spatial Interaction Model	R Squared
Unconstrained	0.053
Production Constrained	0.501
Attraction Constrained	0.108
Doubly Constrained	0.812

Table 2: Goodness of fit for each spatial interaction model

Table 2 shows that the three *constrained* models were a dramatic improvement over our original *unconstrained* gravity model. However, the performance of our *attraction constrained* model is significantly worse than the other two. This most likely points to the fact that our origin information data (Total population per TAZ) had a low correlation to our flow totals. In other words, by not constraining our known origin flow totals, our model becomes much weaker. In the future, I would find some other structural variable for my origins, perhaps median household income.

The highest performing model was by far the *doubly constrained* model. This makes sense because we knew the entire flow in and out of each TAZ. This model would be perfect for answering out initial question, as it allows us to compare where we get more pedestrian flow than predicted in our model. We could then evaluate what’s unique about that TAZ. The doubly constrained model also allows us to investigate how night-time flow patterns change over time.

Sources:

Dennett, Adam; Modelling Population Flows Using Spatial Interaction Models, March 2018; Link: https://rpubs.com/adam_dennett/376877

Oshan, Taylor; A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL); November 2016; http://openjournals.wu.ac.at/region/paper_175/175.html

Wilson, AG; A family of spatial interaction models, and associated developments; 1971; Link: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.310&rep=rep1&type=pdf

Data Sources:

TAZ: https://catalog.data.gov/dataset/tiger-line-shapefile-2011-2010-state-new-york-2010-census-traffic-analysis-zone-taz-state-based29d50

Green Taxi: https://data.cityofnewyork.us/Transportation/2016-Green-Taxi-Trip-Data/hvrh-b6nb

Yellow Taxi: https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t

ACS Population Data: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Bars: https://public.enigma.com/datasets/new-york-quarterly-list-of-active-licenses/3534251a-b423-4434-a0e7-b04352c1f851?filter=%2B%5B%3ELicense_Type_Name%5B%22ON-PREMISES+LIQUOR%22%5D,%3EAgency_Zone_Office_Name%5B%22New+York+City%22%5D%5D