Exploring the Spatial Patterns and Popularity Determinants of TNC in San Francisco using ArcGIS

CPLN 503 Modeling Geographical Objects
He Zhang

Introduction

Background

TNC, Transportation Network Company or Uber and Lyft in specific, is playing an increasingly critical role in urban mobility. With an special interest in TNC, this project is to summarize TNC activities, and to further explore possible factors that are associated with TNC popularity through spatial analysis on ArcGIS platforms.

The Main Goals:

1) To locate popular TNC pick-up and drop-off areas and to reveal the variations across times.
2) To explore possible factors that are associated with TNC popularity in an area. These could include land use, transportation, demographics, socio-economic aspects, etc.

Factors to consider:

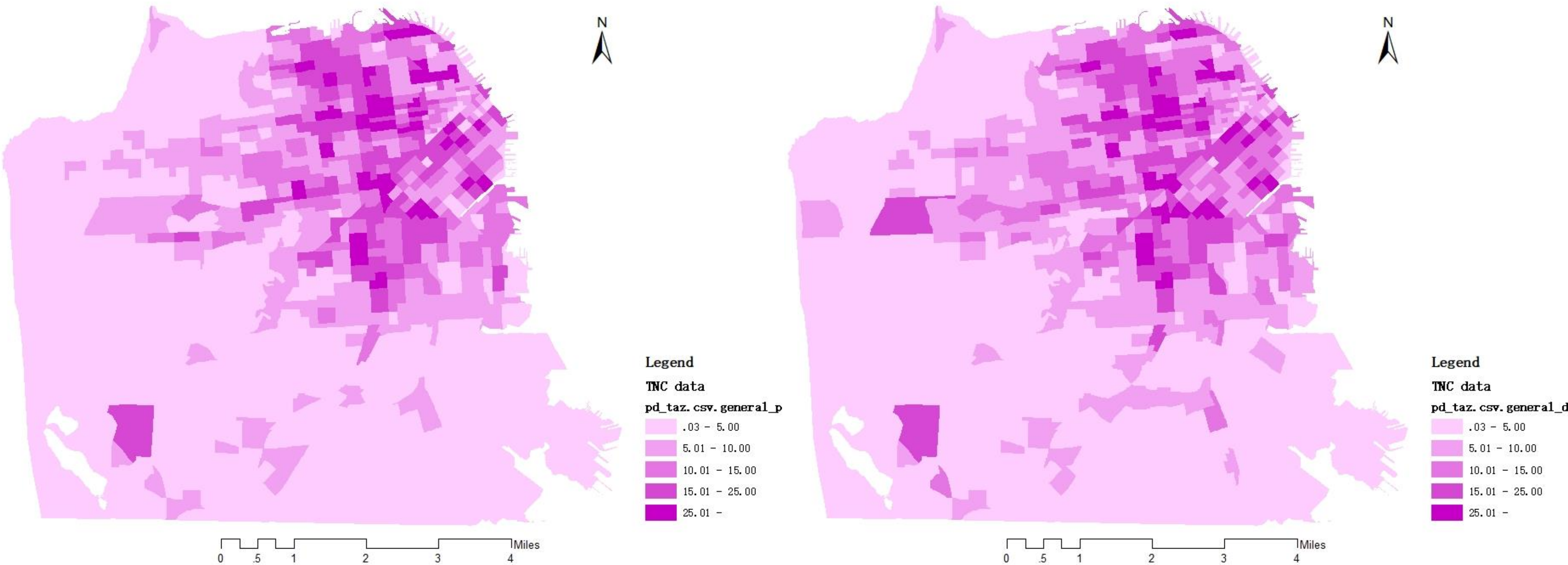
1) Land Use Activities
2) Demographics and Socio-economics
3) Supply of other modal transportation (substitution)

Category	Data input	Data Type	Coordinate system	Source
Base map	San Francisco Travel Analysis Zone.	Shapefile	GCS_North_ American_1983	http://tncstoday.sfcta.org/
	SF Census tracts 2010.	Shapefile	WGS84	https://data.sfgov.org/
	Bay area counties boundaries.	Shapefile	WGS84	https://data.sfgov.org/
TNC data	TNC pickup/dropoff by TAZ.	Csv	/	http://tncstoday.sfcta.org/
Land use/ activity data	Addresses with Units - Enterprise Addressing System (EAS).	Csv	/	https://data.sfgov.org/
	Map_of_Registered_Business_Location	Csv	/	https://data.sfgov.org/
	Land use.	Shapefile	WGS84	https://data.sfgov.org/
Transportation data	Streets (with speed limits).	Shapefile	WGS84	https://data.sfgov.org/
	MTA. Muni_simpleroutes.	Shapefile	WGS84	https://data.sfgov.org/
	SFMTA transit stops.	Csv	/	https://www.sfmta.com/reports/gtfs-transit-data
	Bikeway.	Shapefile	WGS84	https://data.sfgov.org/
Census data	Demographic and Socioeconomics data of SF census tracts.	Csv	/	https://factfinder.census.gov/
Image	Satellite image of part of the city.	JPEG	/	Google Map

Analysis summary

Task	Method& Tools
S1. What are the spatial and temporal variations of different kinds of TNC activities?	Descriptive statistics; 3D analysis tools; Spatial autocorrelation (Moran Index)
S2. Aggregate demo, socio-economic, transit, and land use data to TAZs in preparation for modeling.	Spatial Join/ join by location.
S3. Model the relationship between TNC activities and demo, socio-economic, transit, land use data.	Ordinary Least Square modeling; Geographically Weighted Regression modeling
S4. Data automation: build a model to perform the task in section 2 & 3.	Model builder.

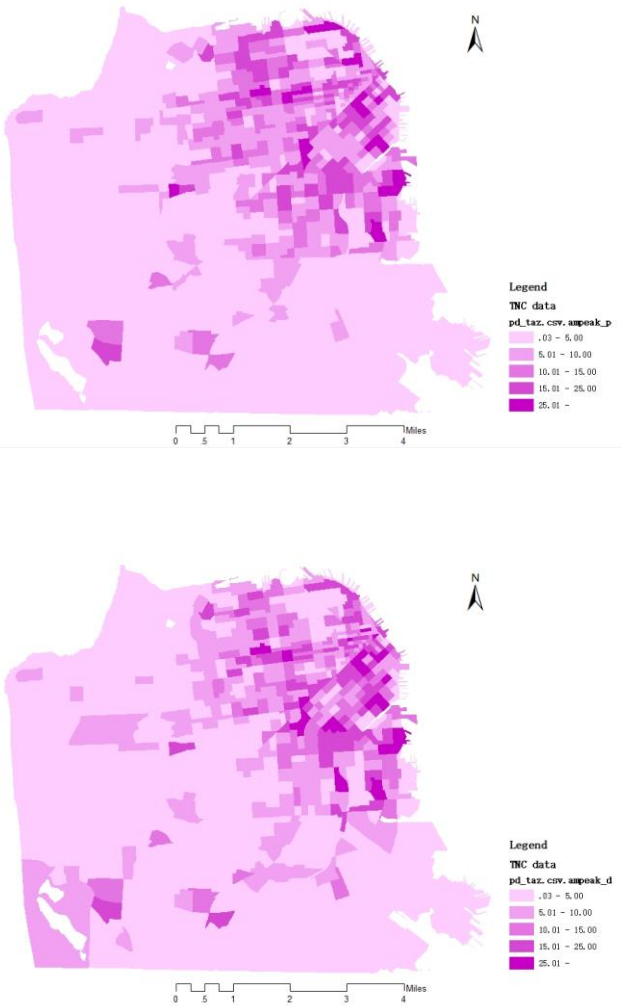
1. The patterns of TNC activities



TNC pick-up (left) and drop-off (right) frequency in general

To enable comparison, the breaks of the TNC data display are set to be the same, with a reference to natural breaks. In general pick-up and drop-off activities largely share the same spatial pattern. The popular locations are mainly in the northeastern grids area, while other areas are of low activity level.

1. The patterns of TNC activities



A.M. peak pick-up (up) and drop-off (down) frequency

P.M. peak pick-up (up) and drop-off (down) frequency

Weekend pick-up (up) and drop-off (down) frequency

P.M. peak is when pick-up and drop-off are most popular; in A.M Peak the frequency is slightly higher than that on weekends. The aggregation is always in the northeastern area, while in P.M. peak the area expands apparently.

1. The patterns of TNC activities

Then, I am comparing the patterns of pickups and drop-offs across different time groups by looking at the descriptive statistics.

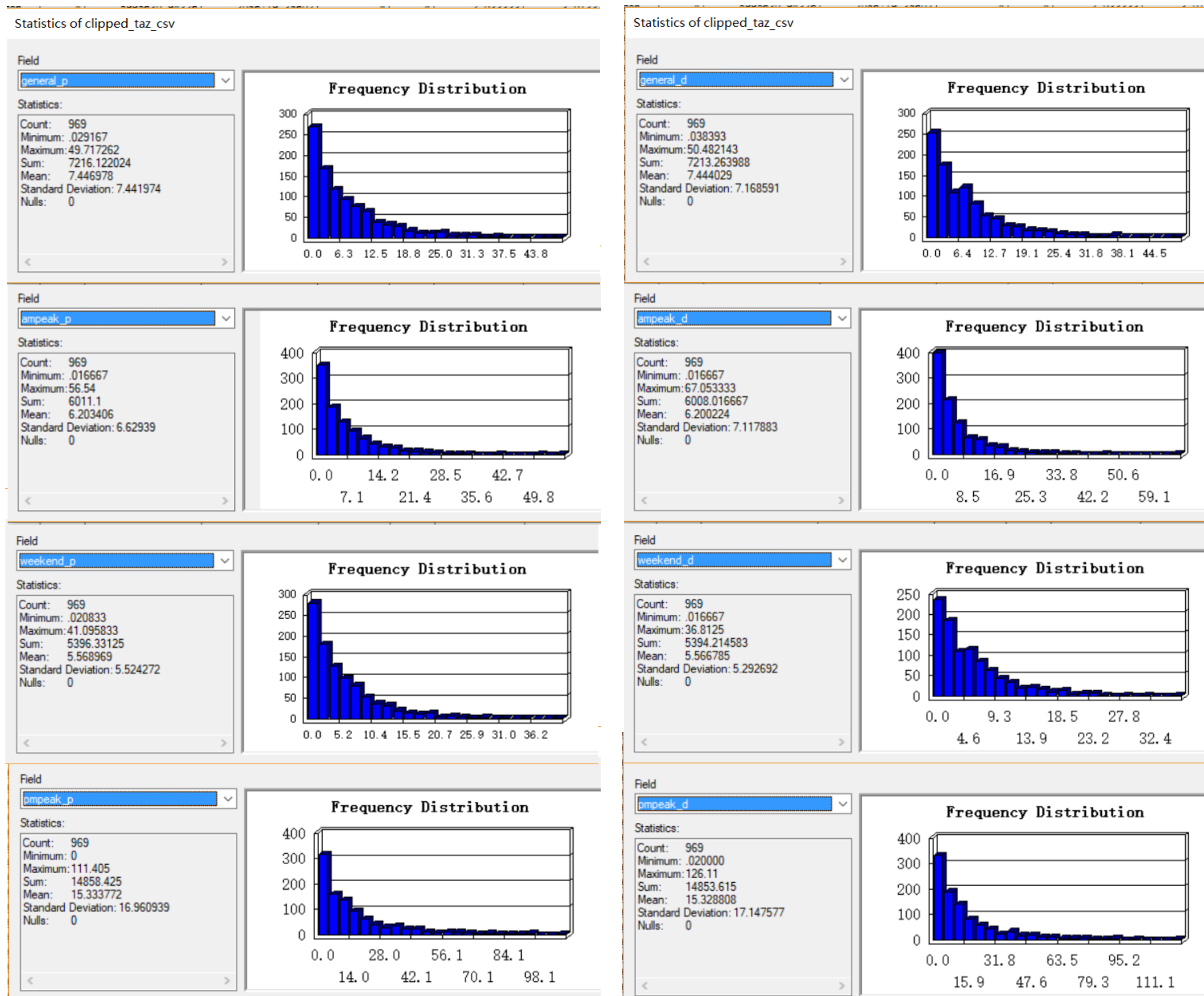
Findings:

1. Largest activity count in p.m. peak, larger than twice the count of activities in a.m. peak, despite that a.m. peak has three hours (6,7,8) yet p.m. peak only has two (18,19).

Hence, TNC activities are definitely <u>most popular in p.m. peak hours.</u> Smallest activity count on weekends.

2. Greater variances of pickup than of drop-off data in general and on weekends. The other way around in peaks.

Hence, <u>in peak hours drop-off locations are more subject to change while pick-up locations are more consistent.</u> This could be due to the consistency of people's residency and work place.



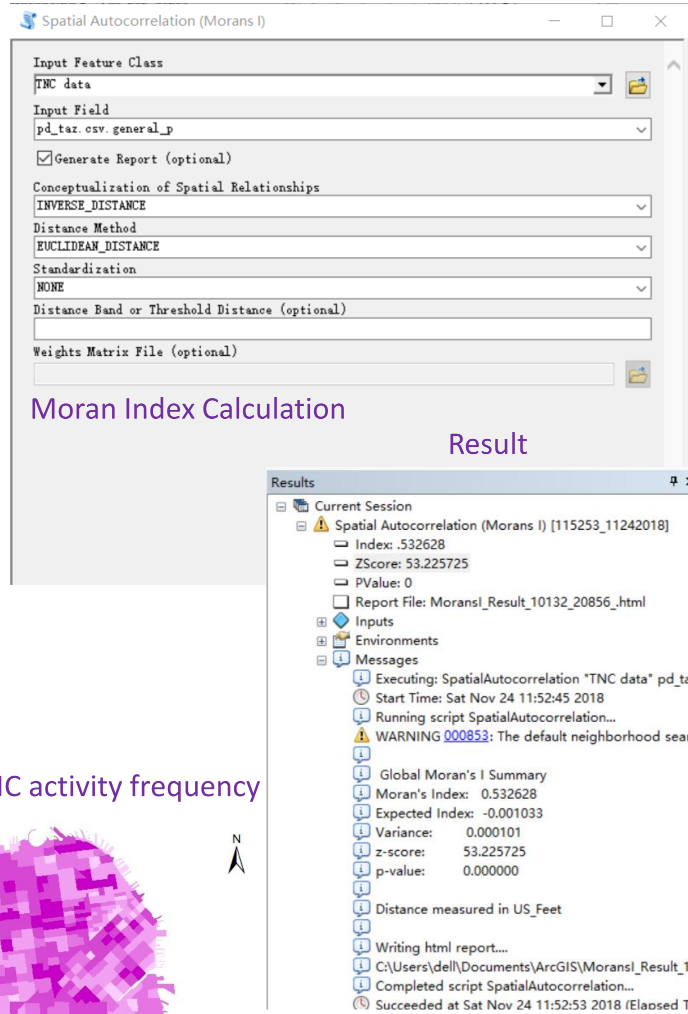
1.1. The spatial patterns: autocorrelation

I want to detect the evenness/ unevenness distribution pattern of TNC activities. This is reflected by spatial auto-correlation coefficients.

I here calculate the global **Moran Index** to measure spatial auto-correlation. I use inverse_ distance to generate the weight matrix. A Moran I close to 1 means positive autocorrelation and unevenness, while a Moran I close to -1 means negative autocorrelation and evenness.

Result: Uber/ Lyft pick-up and drop-off frequency has significant positive spatial auto-correlation across different time in a week. Moreover, the drop-off activity shows greater auto-correlation than the pick-up activity.

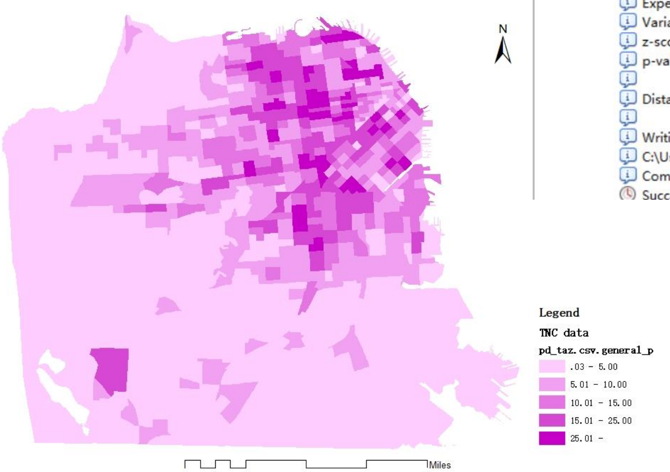
Implication: The spatial distribution of TNC activity is very uneven, so there exists an aggregation pattern. This pattern is more obvious for drop-off data than for pick-up data. When I refer to the TNC activity map, this auto-correlated pattern is obvious: there are clusters of higher frequency in the northeast area and sparse distribution in the southwest area.



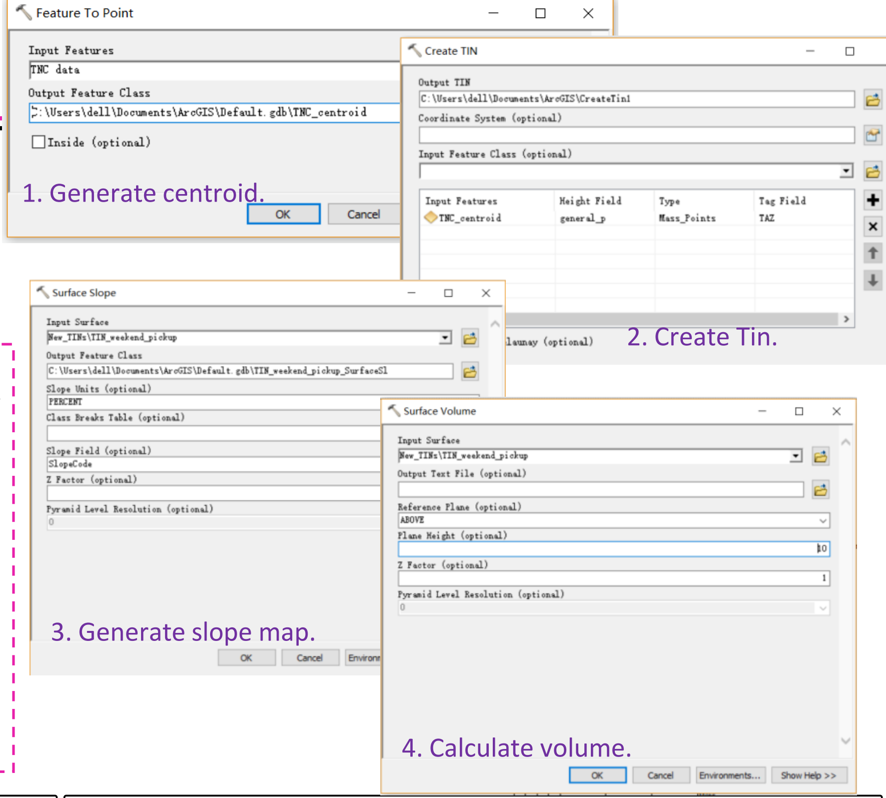
Moran Index For TNC activities by TAZs

Time period	Moran I	Z-score
General pick-up data	0.53	53.23
General drop-off data	0.54	54.04
Weekday A.M. peak pick-up data	0.49	48.62
Weekday A.M. peak drop-off data	0.63	63.30
Weekday P.M. peak pick-up data	0.50	50.47
Weekday P.M. peak drop-off data	0.52	51.83
Weekend pick-up data	0.51	51.30
Weekend drop-off data	0.52	52.18

Data input:
TAZ- aggregated TNC activity frequency



1.2. The patterns of TNC activities by 3D analysis tools



I want to 1) compare the patterns of general pick-ups and drop-offs.
2) compare the pattern (e.g. of pick-ups) of weekdays and weekends.

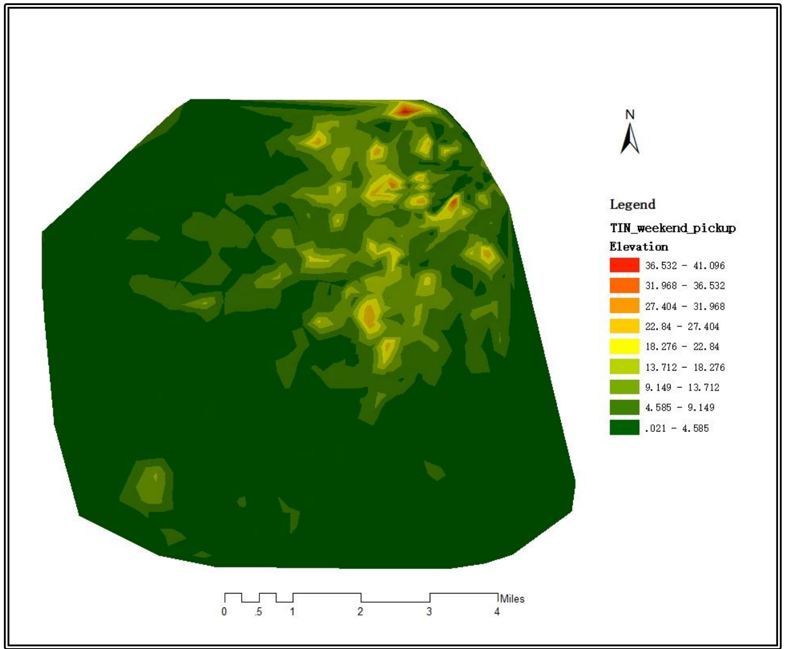
1. Create centroids from TNC data shapefile, by Feature to Point.
2. Generate TIN from the centroid, using TNC activity data as heights, by Create TIN-> Mass point. Generate TINs for generate pick-up/ drop-off, A.M. peak pick-up/ drop-off, weekend pick-up/ drop-off, by specifying the height fields.
3. Measure TIN slope, by Triangulated Surface > Surface Slope. Measure TIN volume of height above 10, by Functional Surface > Surface Volume.

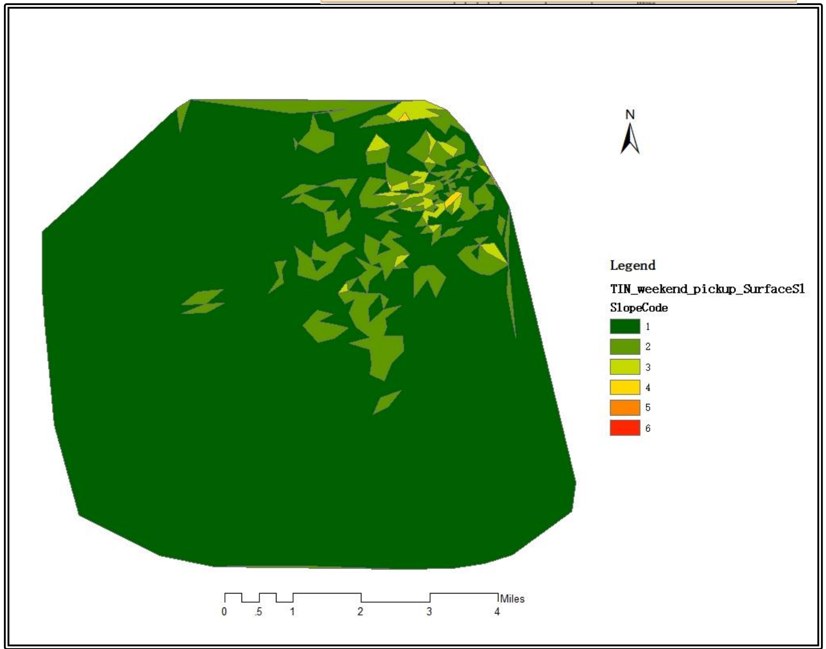
Surface Volume for weekend pickup data (above 10)

Area_2D: 142094024

Area_3D: 142114333

Volume: 615472584





1.2. The patterns of TNC activities by 3D analysis tools

4. Compare. First, display the three pick-up slope maps using the same symbology. The slope ranking is therefore: general pickup> a.m. peak pickup > weekend pickup. This implies that general pickup data has larger spatial variances, at least in the northeast areas, while weekend pickups have the smaller spatial variances, consistent with statistics above.

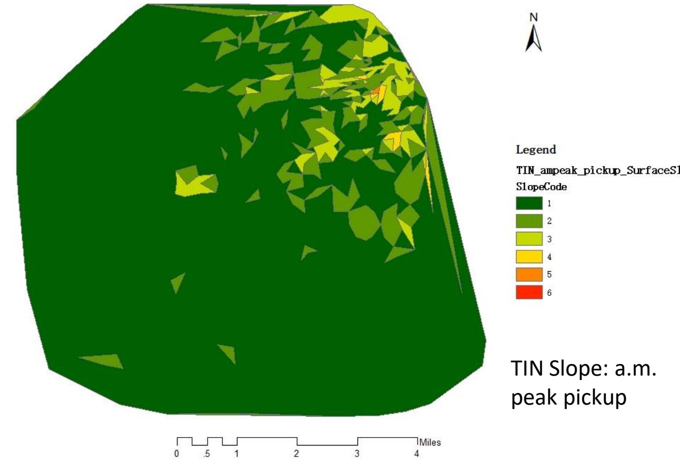
Second, display general pickup and drop-off data using the same symbology. Overall the pickup data could have larger spatial variances compared to the drop-off data.

Volume (above surface of 10, from the txt reports)

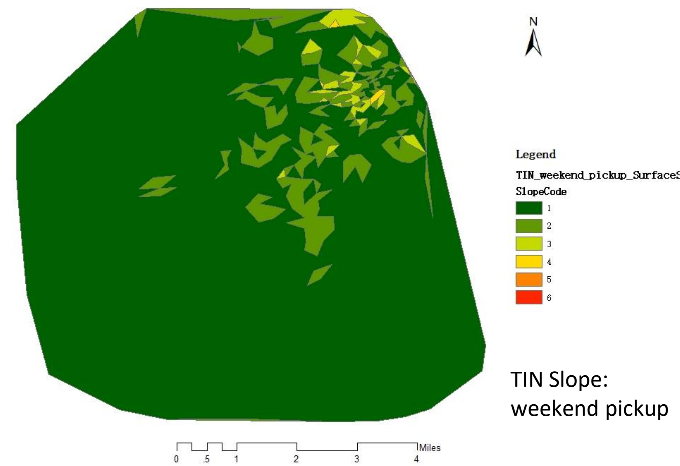
	A.M. Pickup	A.M. Dropoff	Weekend Pickup	Weekend Dropoff	General Pickup	General Dropoff
Area_2D	1.74E+08	1.62E+08	1.42E+08	1.39E+08	2.26E+08	2.32E+08
Area_3D	1.74E+08	1.62E+08	1.42E+08	1.39E+08	2.26E+08	2.32E+08
Volume	8.72E+08	9.24E+08	6.15E+08	5.66E+08	1.43E+09	1.37E+09

Third, compare the volume data Among the more popular areas ,there are more pickup than drop-offs across a week.

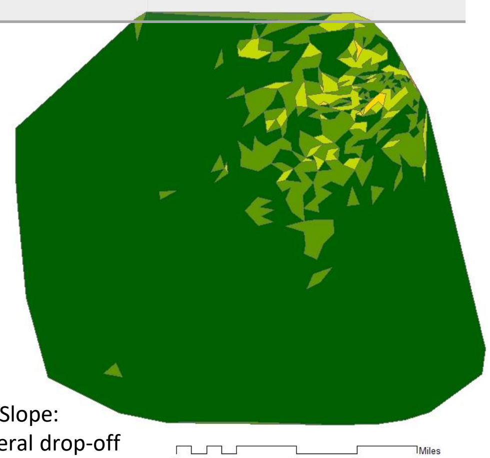
Therefore, the pick-ups can be more concentrated or aggregated, and the drop-offs more dispersed, in terms of spatial distribution.



TIN Slope: a.m. peak pickup



TIN Slope: weekend pickup

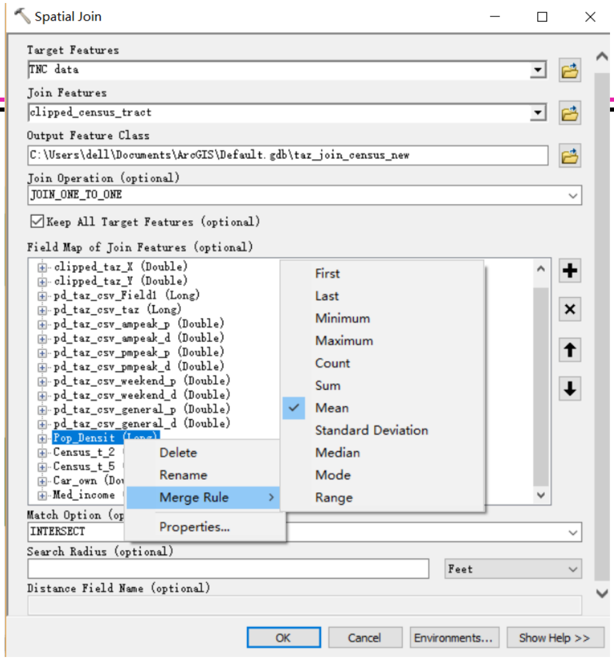


TIN Slope: general drop-off

TIN Slope: general pickup

2. Aggregating data to TAZs 3.1 Census data

- Assign the average value of census tracts that overlaps with a TAZ. The method is illustrated by the chart at the right bottom.
- Spatial Join-> Intersect -> Select fields-> Merge rule-> Mean.
- Keep the factors that could be related to TNC activity. Delete the unwanted columns.

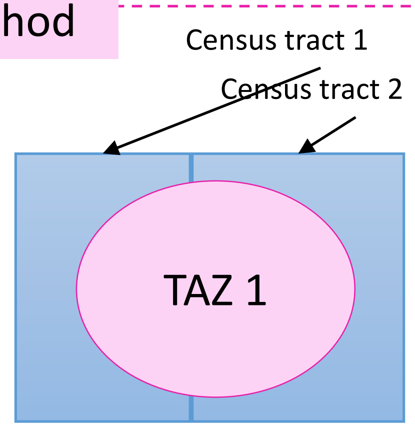


Therefore, the new output shapefile bears both the TNC activity data and the census data of:

Population density	Average household size	Average commute	Car ownership ratio	Median household income
Mean	Mean	Mean	Mean	Mean

general_d	Pop_Densit	Census_t_2	Census_t_5	Car_own	Med_income
.611012	31133	3.7	35	.806667	55189.67
1.621131	24148	3.45	35	.910000	71787.5
.896429	22107	3.375	36	.680000	41573.25
4.127976	23930	3.533333	35	.893333	67687.34
.522321	33047	3.7	36	.743333	41376
1.359524	24762	3.525	35	.835000	66809
.362202	24120	3.466667	35	.603333	29969
.807440	22738	3.5	36	.900000	79716.5
2.975595	26185	3.8	36	.880000	72890.34
4.811905	23930	3.533333	35	.893333	67687.34
.764881	31713	3.65	34	.865000	74336
.678274	25134	3.475	36	.650000	35150.25
1.619048	27530	3.85	36	.890000	79592
1.769048	26498	3.466667	35	.880000	75019.34

Method



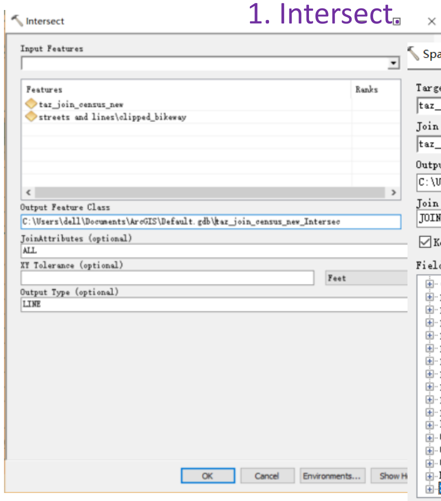
For example, the population density of TAZ 1 will be the mean of population density of census tract 1 and 2.

2. Aggregating data to TAZs 3.2 transportation- bikeway

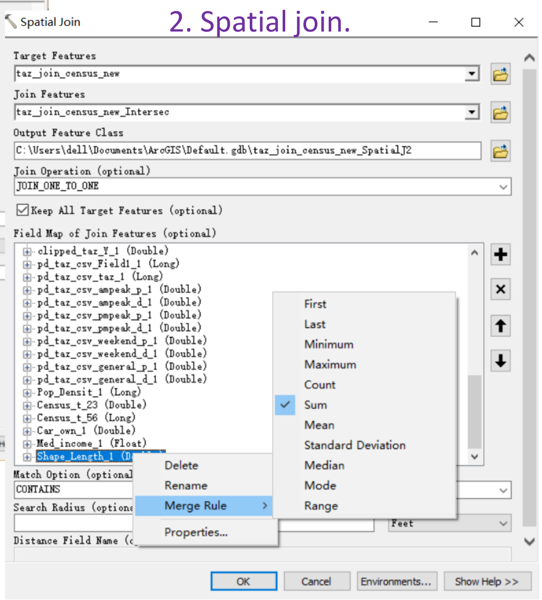
To calculate the total length of bikeway within every TAZ.

- Perform Overlay-> Intersect to generate a “sliced” bikeway line shapefile. The new field Shape_length is the length of each individual sliced. A total of 8900 pieces are generated.
- Conduct Spatial Join-> Contains to calculate the sum of the shape lengths.
- Replace the “null” value with 0 through field calculator programming (python), which means there is no bikeway within these TAZs.

1. Intersect

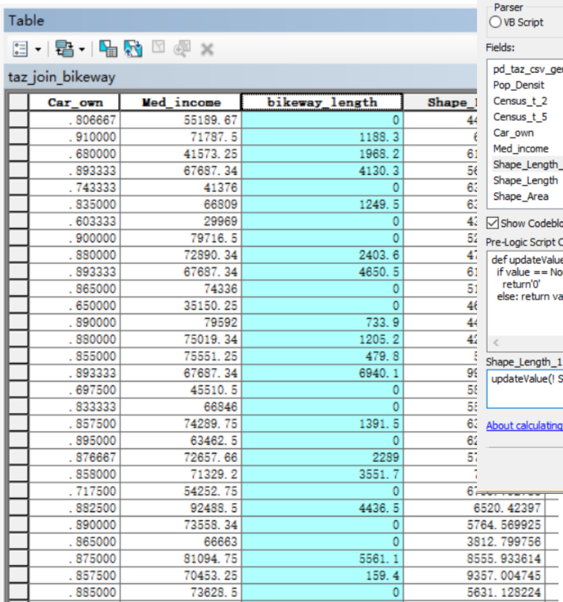


2. Spatial join.



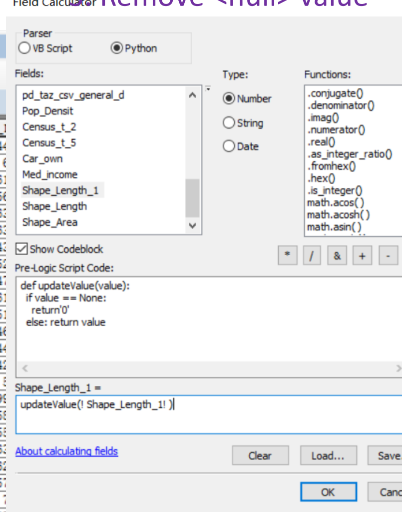
Bikeway length of TAZs, San Francisco

Result



Car_own	Med_income	bikeway_length	Shape_l
.806667	55189.67	0	44
.910000	71787.5	1188.3	6
.680000	41573.25	1968.2	61
.893333	67687.34	4130.3	56
.743333	41376	0	63
.835000	66809	1249.5	63
.603333	29969	0	43
.900000	79716.5	0	52
.880000	72890.34	2403.6	47
.893333	67687.34	4650.5	61
.865000	74336	0	51
.650000	35150.25	0	46
.890000	79592	733.9	44
.880000	75019.34	1205.2	42
.855000	75551.25	479.8	5
.893333	67687.34	6940.1	99
.697500	45510.5	0	58
.833333	66846	0	55
.857500	74289.75	1391.5	63
.895000	63462.5	0	62
.876667	72657.66	2289	57
.858000	71329.2	3551.7	7
.717500	54252.75	0	67
.882500	92488.5	4436.5	6520.42397
.890000	73558.34	0	5764.569925
.865000	66663	0	3812.799756
.875000	81094.75	5561.1	8555.933614
.857500	70453.25	159.4	9357.004745
.885000	73628.5	0	5631.128224

3. Remove <null> value



2. Aggregating data to TAZs 3.2 transportation- transit stops 3.3 activity- business address

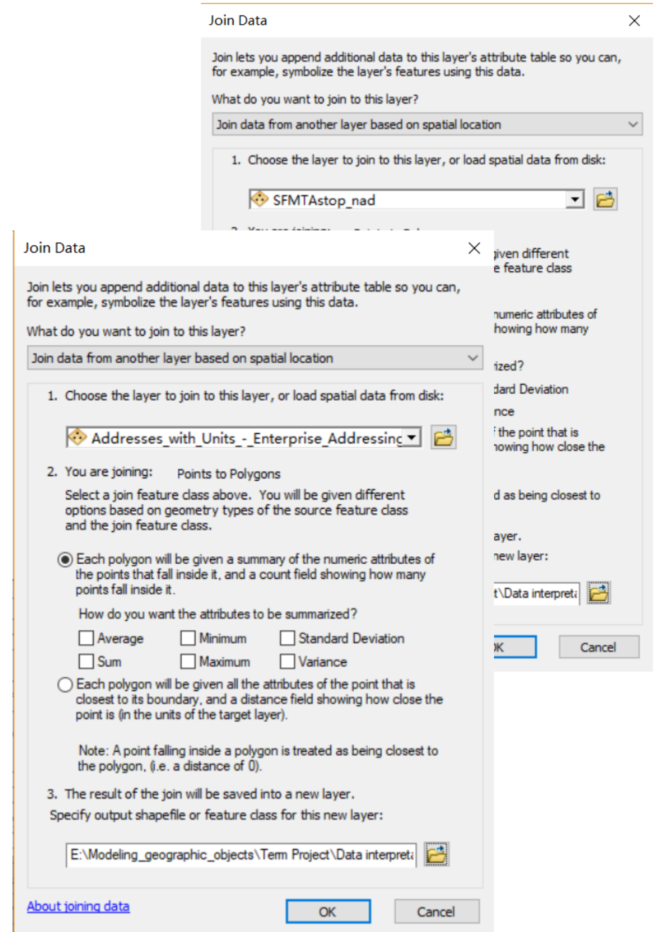
- Conduct Join Data by Location to obtain the count of transit stops and active business addressed of each TAZ.
- Display the business count using color ramp & the transit stop count by gradient symbols.

Result-Table:

Table

Join_t_b

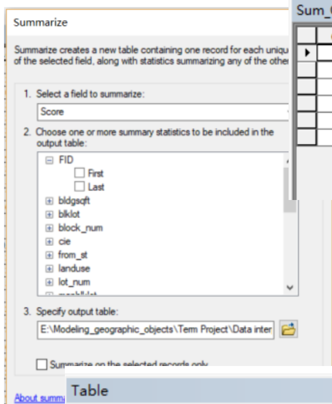
Med_income	Shape_Leng	Shape_Le_1	Shape_Area	Count_transit	Count_busines
55189.7	0	4419.669999	931009.351706	3	310
71787.5	1188.261204	6826.33586	1793394.36497	0	412
41573.3	1968.15348	6158.242434	1360854.94394	5	231
67687.3	4130.319706	5686.615127	1593115.19186	2	353
41376	0	6316.868605	1063485.60551	5	360
66809	1249.546268	6309.865612	2230085.74848	14	463
29969	0	4325.102085	795796.498379	3	12
79716.5	0	5229.434024	1527501.17314	11	300



2. Aggregating data to TAZs 3.3 activity- land use

- To interpret the level of activity from the land use map, I assigned an " Activity Score" for each use on a 0-9 base.
- I added a field to the land use shapefile called "score" and assigned scores by Selecting by Attributes and Field Calculator.
- I summarized the score field to see the general pattern and to double check with my input, by right click on field-> summarize.
- As an estimation, I calculated the mean land use score by Spatial Join-> Intersects -> Merge-> Mean.
- Let's compare the original score and TAZ-based score. To fully display the scores, I first dissolved land use shapefile by land use category, by Dissolve.

Land use category:
7' CIE = Cultural, Institutional, Educational
5' MED = Medical
8' MIPS = Office
9' MIXED = Mixed Uses (Without Residential)
8' MIXRES = Mixed Uses (With Residential)
2' PDR = Industrial
9' RETAIL/ENT = Retail, Entertainment
5' RESIDENT = Residential
9' VISITOR = Hotels, Visitor Services
0' VACANT = Vacant
5' ROW = Right-of-Way
5' OPENSPACE = Open Space



Table

Sum_Output

OID	Score	Count_Score
0	0	5832
1	2	2090
2	5	116285
3	7	1365
4	8	24178
5	9	4938

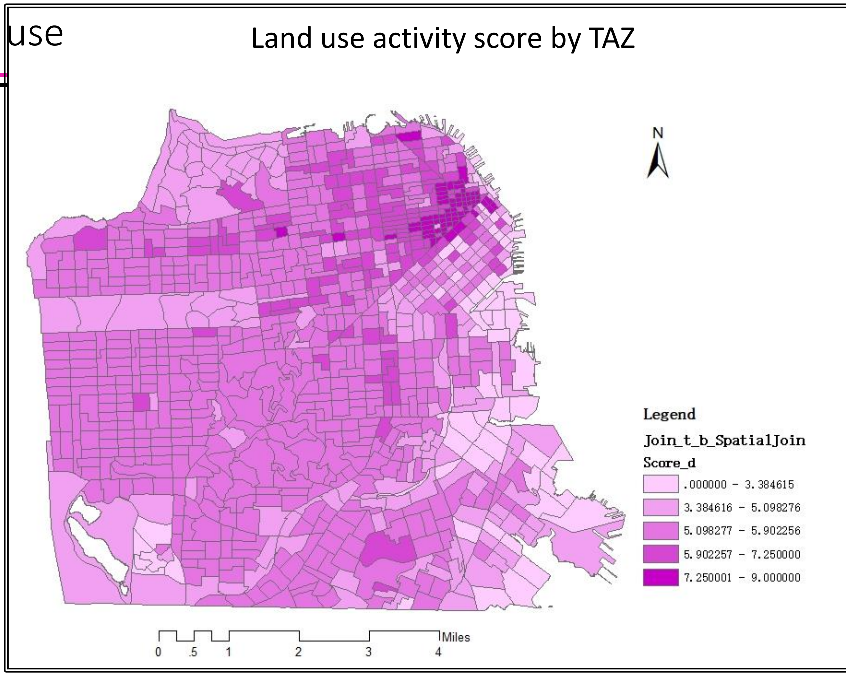
Summarize & result

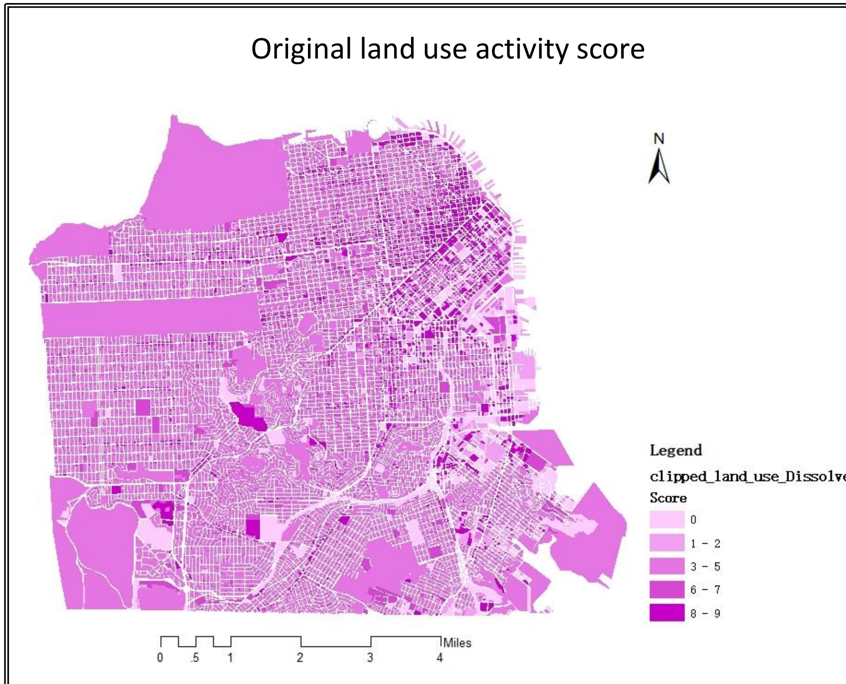
Dissolve result

Table

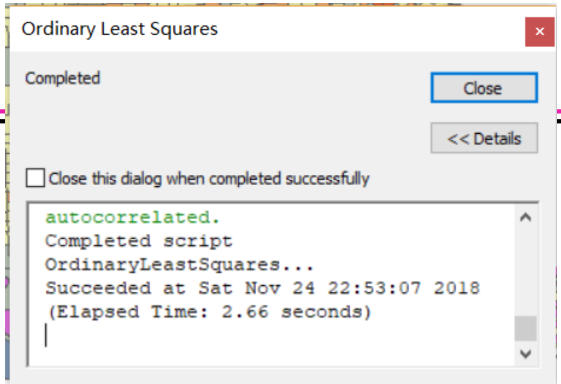
clipped_land_use_Dissolve1

FID *	Shape *	landuse	SUM_shape_area	Shape_Length	Shape_Area	Score
1	Polygon	CIE	40348678. 323566	764193. 598706	40347862. 538797	7
2	Polygon	MED	1711271. 048366	70926. 851821	1711271. 049033	5
3	Polygon	MIPS	14009211. 936183	515222. 468124	14009212. 275757	8
4	Polygon	MISSING DATA	42169418. 029972	1021237. 914398	38729075. 95137	0
5	Polygon	MIXED	24578723. 844782	785899. 16928	24578679. 505225	8
6	Polygon	MIXRES	86166622. 591953	6298608. 823226	86157809. 872333	7
7	Polygon	OpenSpace	312839344. 500851	1193878. 204366	293072885. 609827	5
8	Polygon	PDR	40518890. 346037	976449. 82844	39046439. 363961	2
9	Polygon	RESIDENT	356764300. 346358	30022423. 753705	355093083. 11704	5
10	Polygon	RETAIL/ENT	23497262. 541941	867852. 512363	22635956. 837239	9
11	Polygon	Right of Way	716865. 137766	6933. 375759	716865. 137785	2
12	Polygon	VACANT	60671817. 421637	1655583. 047531	60138298. 23013	0
13	Polygon	VISITOR	4118605. 760041	122892. 302876	4076401. 547843	9





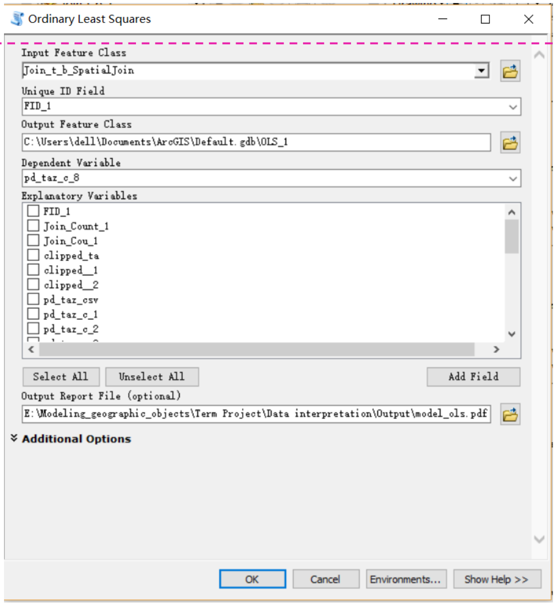
3. Model building- 3.1 OLS modeling

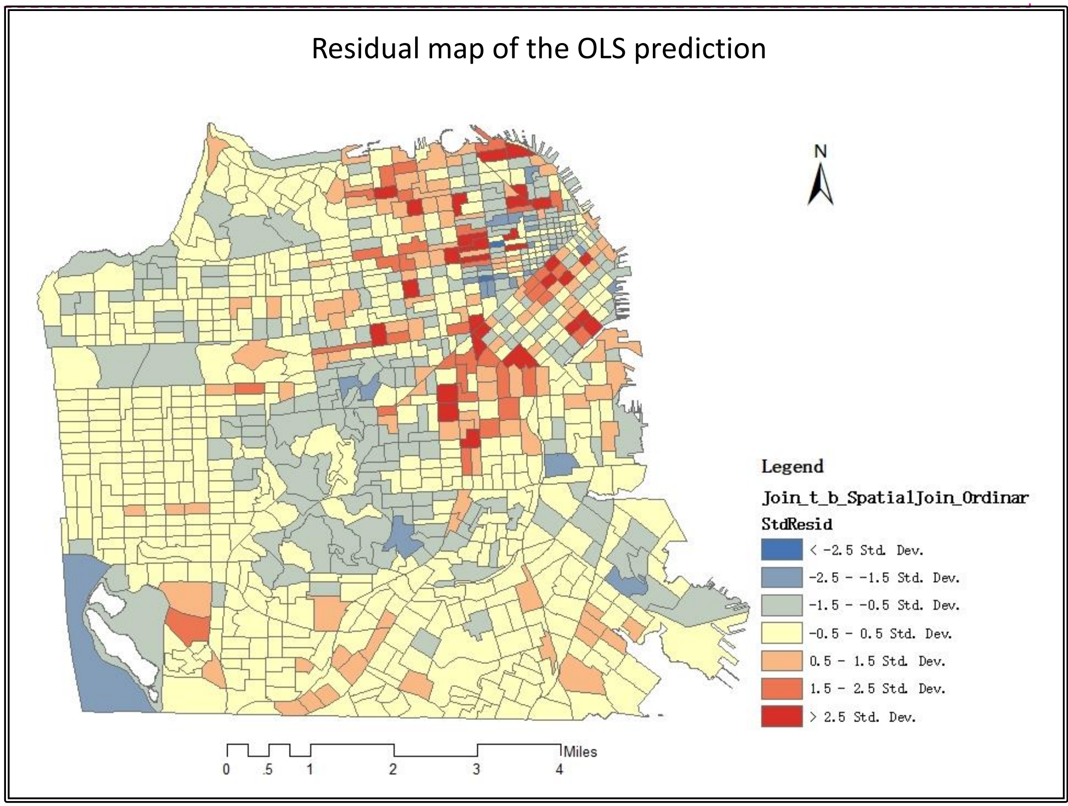


As I have detected significant auto-correlation in the TNC data in section 1, OLS is not proper for modeling this data. I still tried OLS regression at first, and "autocorrelated" is reported.

I first put all possible factors studied in the preliminary model:

General pick up frequency = β0+ β1*population density+ β2*average household size+ β3*average commute time+ β4*car ownership percentage+ β5*median household income+ β6*bikeway length+ β7*transit stop count+ β8* business activity count+ β9*land use activity score+ ε





>- From the residual map, the areas with higher TNC activity levels tend to have residuals with greater absolute values, as opposed to a random distribution.

3. Model building- 3.1 OLS modeling

Information from the model result:

- All independent variables are significantly correlated with general TNC pick-up frequency, except *land use activity score.*
- The adjusted R-square is 0.484, showing a moderate model fit.

Let's go on to use GWR (Geographically weighted regression) to model the relationship.

Model result of the OLS regression

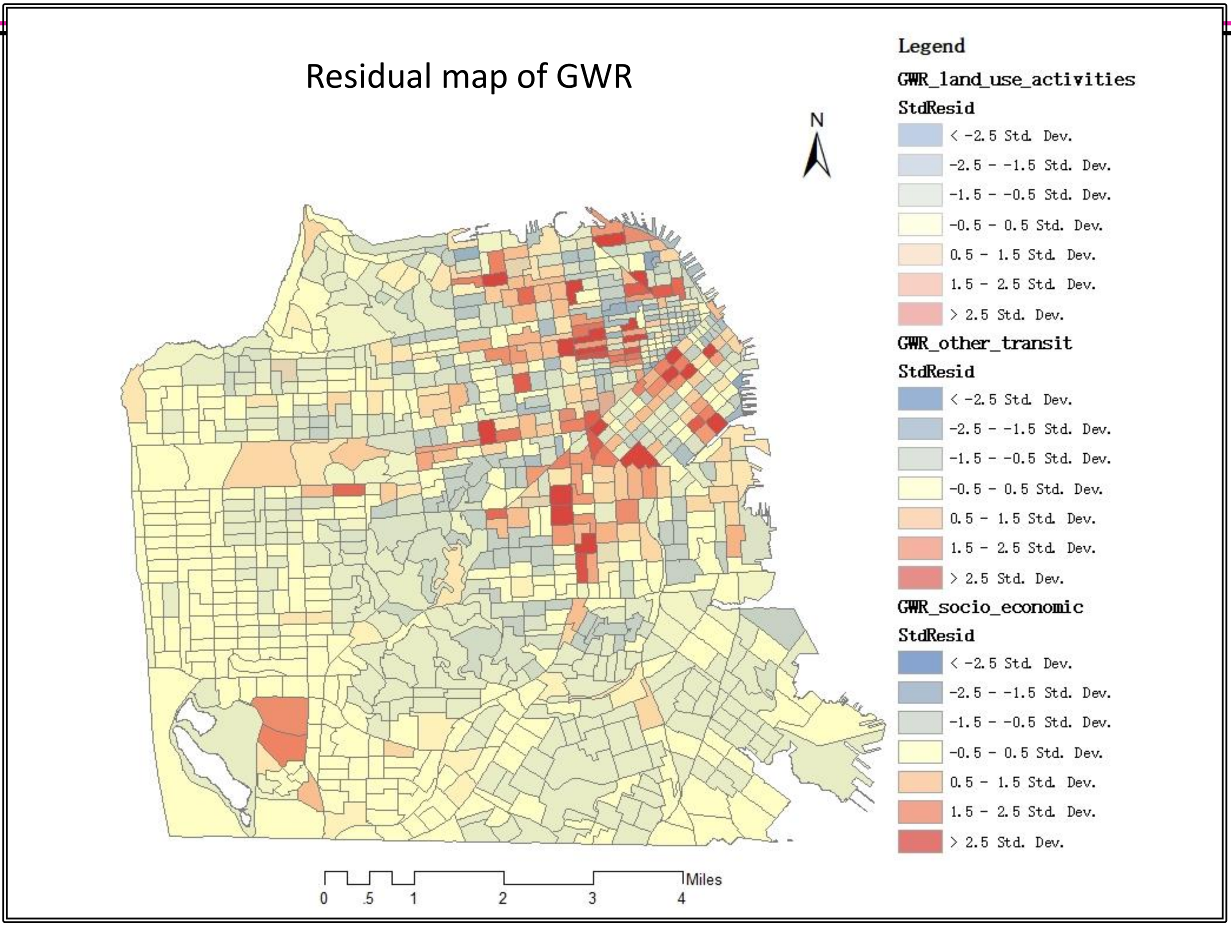
	Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]
	Intercept	25.086081	2.691085	9.321920	0.000000*
population density	POP_DENSIT	0.000064	0.000015	4.293717	0.000023*
average household size	CENSUS_T_2	-3.460991	0.561996	-6.158393	0.000000*
average commute time	CENSUS_T_5	-0.183616	0.077545	-2.367865	0.018075*
car ownership percentage	CAR_OWN	-15.643792	2.672674	-5.853236	0.000000*
median household income	MED_INCOME	0.000038	0.000009	4.162122	0.000040*
bikeway length	SHAPE_LENG	0.000131	0.000045	2.925851	0.003524*
transit stop count	COUNT_	0.295714	0.060300	4.904082	0.000002*
business activity count	COUNT_1	0.007003	0.000882	7.941635	0.000000*
land use activity score	SCORE_D	-0.192577	0.147528	-1.305356	0.192094

3. Model building- 3.2 GWR

As GWR refused to run with many variables (could be due to computing work load), I went on to model the relationship between TNC activity and three groups of independent variables:

1) demography and socio-economics attributes;
2) alternative transit provision;
3) land use and activities.

On the right I displayed the residual of the three models in one map by <u>adjusting transparency.</u> It shows that the residual pattern is different from OLS, although the autocorrelation still exist.



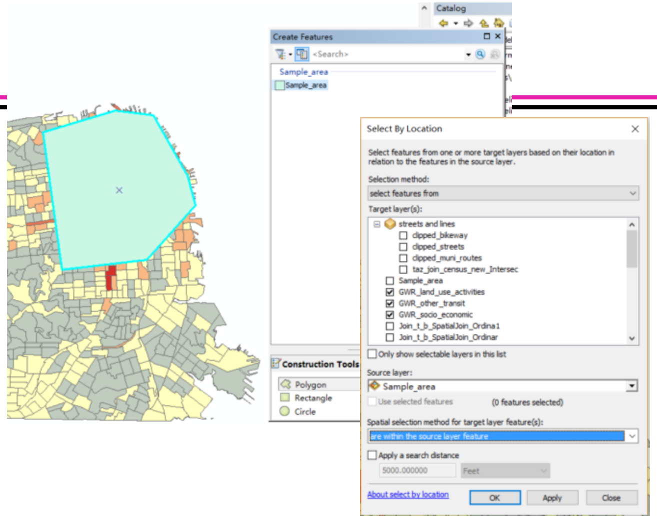
From the R2 comparison table, it is clear that given the same set of independent variables, GWR models have dramatically better fit than OLS models.

Comparison of Adjusted R-square of GWR and OLS model

	OLS	GWR
M1: Demography and socio-economics attributes	0.399	0.465
M2: Alternative transit provision	0.011	0.559
M3: Land use and activities	0.074	0.401

3. Model building- 3.2 GWR& Model Interpretation

- As GWR is Local Spatial Regression, I looked at local level coefficients.
- I drew a new layer of a sample area polygon. The sample is the northeast area where TNC activities are most concentrated.
- Then I select the GWR results of TAZs within it, by Selecting by Location.
- I created three new layers from the three selections.
- Finally, I summarized the local model result. The significance is determined by calculating t-statistics (β/standard error) by Field Calculator.



Modelling Summary (local)

Variable	Correlation
Population density	Positive
Median household income	Negative
Median household size	Negative
Bikeway length	Not significant
Transit stops	Not significant
Business count	Positive
Land use activity score	Positive

GWR local result (Model 2)

GWR_other_transit selection

Local R2	Predicted	Coefficient Intercept	Coefficient #1 Shape_Leng	Coefficient #2 Count_	Residual
.073456	7.809645	8.571866	.000842	-.291024	6.664463
.082483	9.138353	9.554084	.001030	-.318274	6.780694
.091157	10.781975	8.648537	.000942	-.287113	12.240942
.062098	13.333189	10.51934	.001022	-.073842	20.023656
.058674	13.267412	10.711243	.000895	.134234	-1.021281
.065065	13.345892	9.522875	.000681	.258664	-4.629523
.082147	14.810858	10.265444	.001128	-.211314	-2.611751
.068063	12.904534	10.778861	.001027	.036722	20.021359
.064440	15.803462	10.87887	.000893	.206213	6.073622
.074121	11.906935	10.953483	.000806	.375581	-.421518
.075879	11.019161	10.553048	.001075	-.036584	-2.498328
.069462	12.089482	10.854948	.000948	.186328	2.842363

Interpreting the result:

In the northeast-the most economically-vibrant area, the higher the population density, the higher the TNC activity. It is reasonable because more people is going to take TNC in the area. Larger household size and higher median income means lower TNC activity. It can be due to that these households are more likely to have cars and drive as opposed to single persons.

With regard to other transit provision, the relationship is not obvious. It might be that busy areas also have bikeways and good transit – high demand and high supply always coexist.

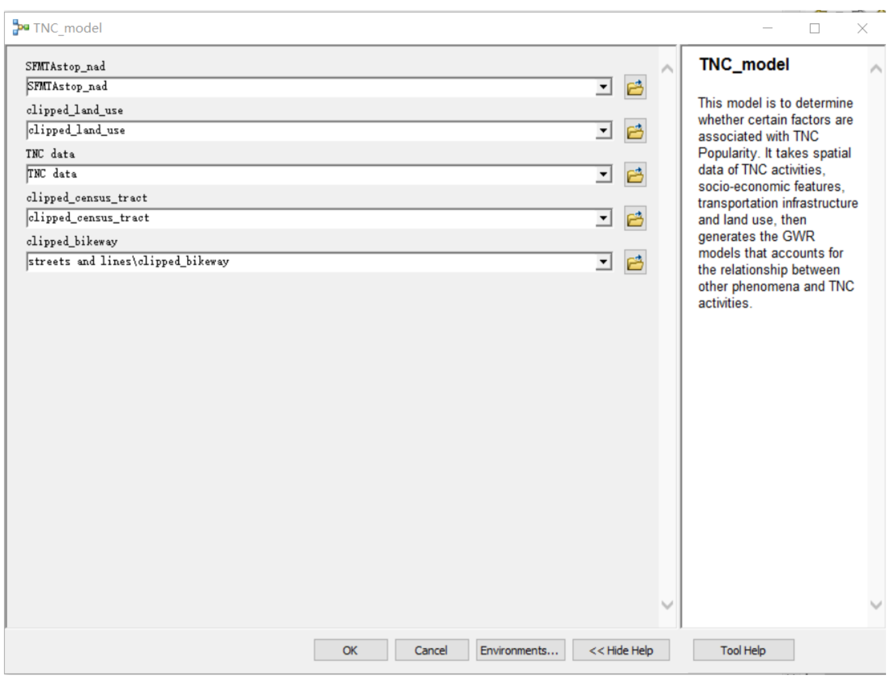
Land use or activities are positively correlated to TNC activity frequency. This echoes the transport-land use interaction. Activities generate transport demand.

4. Data automation

I am presenting a model automates the data analysis in part 2 & 3.

This model is to determine whether certain factors are associated with TNC Popularity. It 1) takes spatial data of TNC activities, socio-economic features, transportation infrastructure and land use, 2) aggregates other data to TAZs, and 3) generates the GWR models based on TAZ units to account for the relationship between other phenomena and TNC activities.

Model Overview



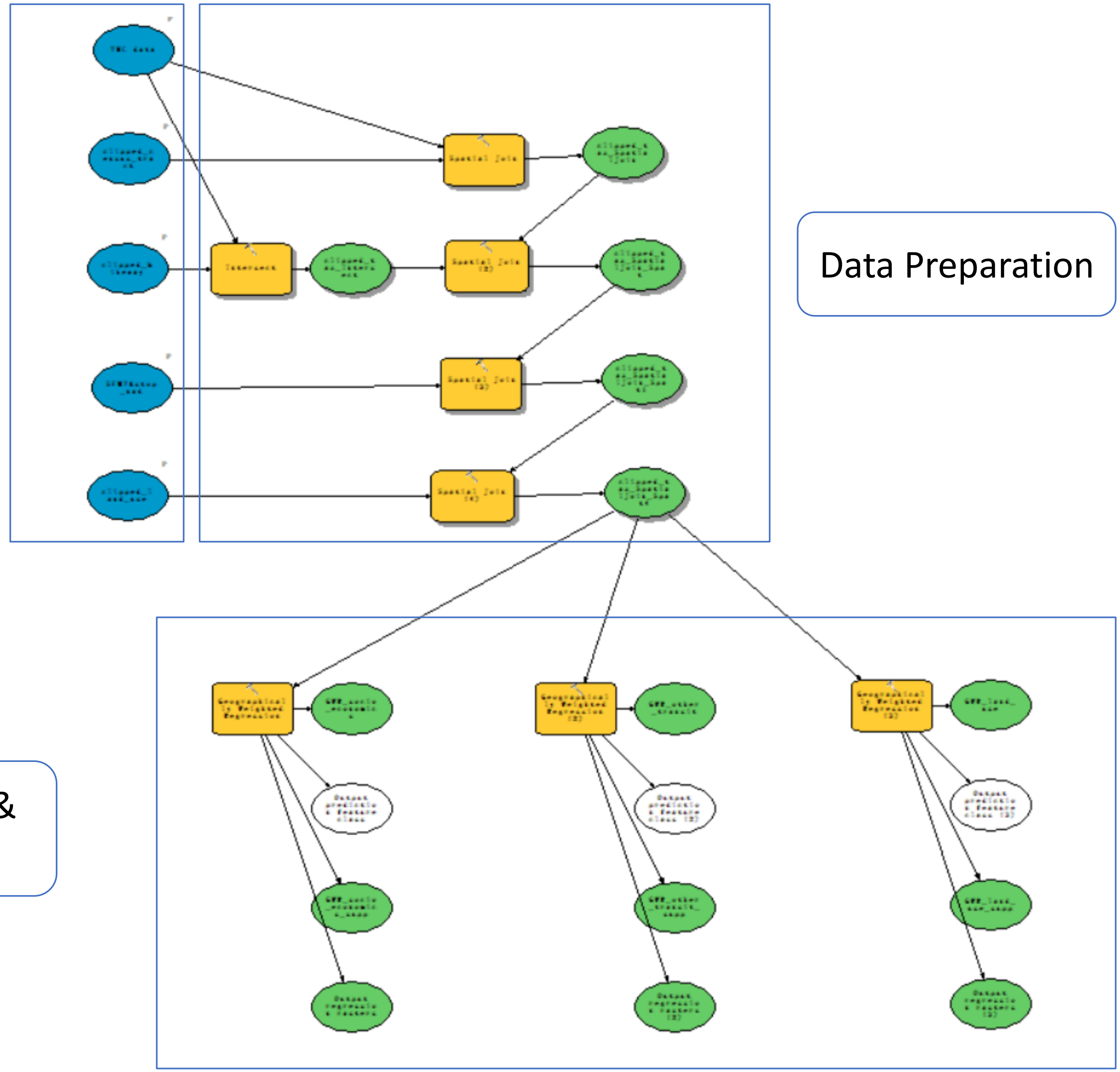
Data input and output

Data Input	**TNC frequency data (aggregated to TAZs)**
	Population density (by census tracts)
	Average household size (by census tracts)
	Median household income (by census tracts)
	Bikeway
	Transit stop
	land use activity score (aggregated to TAZs)
Data Output	GWR model 1: TNC pick-up frequency and socio-economic features.
	GWR model 2: TNC pick-up frequency and alternative transportation infrastructure.
	GWR model 3: TNC pick-up frequency and land use activity scores.

4. Data automation- Model overview

Data Input

Data Preparation



Data analysis&
Data output

Summary: Key Findings

Questions	Key Findings
S1. What is the TNC activity spatial distribution pattern?	Significantly and positively auto-correlated. Uneven spatial distribution with visible clusters.
S2. What are the variations of TNC activity? (pick-up vs drop off; weekday vs weekend…)	TNC activities are definitely most popular in p.m. peak hours. In peak hours drop-off locations are more subject to change while pick-up locations are more consistent. In terms of spatial distribution, the pick-ups can be more concentrated or aggregated, and the drop-offs more dispersed,
S4. Model the relationship between TNC activities and demo, socio-economic, transit, land use data	This significantly auto-correlated data is not quite suitable for OLS, and GWR is a more reasonable tool in this case. In the northeast-the most economically-vibrant area, the higher the population density, the higher the TNC activity. Larger household size and higher median income means lower TNC activity. In regard to other transit provision, the relationship is not obvious. Land use or activities are positively correlated to TNC activity frequency.
S5. Post-modeling: relationship between TNC activities and transit stops/ bikeways (which are non-significant in the model) ?	Bikeway and transit services co-exist with Uber/Lyft; they might be more of complimentary to TNC services instead of a competition.

Thanks for viewing!

He Zhang
CPLN 503 Modeling Geographical Objects