Interpolating Spatial Interaction Data

Jang, W.[1] and X. Yao[2]

---

[1] Woo Jang
PhD student
Department of Geography, University of Georgia, Room 120C, GG Building, Athens, GA 30602, USA.
Tel: 706-542-2926, Fax: 706-542-2388, E-mail:wjmn98@uga.edu

[2] Xiaobai Yao
Associate Professor
Department of Geography, University of Georgia, Room 204, GG Building, Athens, GA 30602, USA.
Tel: 706-583-0326, Fax: 706-583-0326 (office) or 706-542-2388, E-mail: xyao@uga.edu

Spatial interpolation has been widely used to improve the spatial granularity of data, or to mediate between inconsistent zoning schemes of spatial data. Traditional areal interpolation methods translate values of source zones to those of target zones. Each data record is associated with one spatial unit. These methods have difficulty in dealing with the flow data which is associated with a pair of origin and destination zones.

The purpose of this study is to develop an interpolation method to estimate commuting flow data between spatial units in a target zoning scheme based on such data in a source zoning scheme. This estimate flow data is weighted by the math product of two factors. The first factor is the proportion of the overlapped area of the trip origin spatial unit in the source zoning scheme and that in the target spatial zoning, while the second factor is the similar proportion of overlapped area between two destination spatial zones. To test the estimation accuracy and the application potential of this interpolation method, a case study of Fulton County, Georgia, is conducted. Although commuting flow data are available among census units such as census tracts and traffic analysis zones in decennial census data, the zoning schemes of these census units have changed dramatically over time. By applying the developed interpolation of flow data, time series flow data is produced in an identical spatial zoning scheme. Spatial analysis is then performed to investigate the spatial-temporal changes of house-job relations in the study area.

Key words: spatial interpolation, spatial interaction, commuting flow, urban transportation

1 INTRODUCTION

Due to the fast growing process of suburbanization or urban sprawl, travel patterns become more complex with the increasingly multifaceted ways in which people interact in space and time. Trip data play an important role in our understanding of the spatial interactions in an urban space. Some trips are largely discretionary such as those made for shopping, social or recreational activities, and family or personal matters. Even though commuting travel (journey to work) only accounts for about 25 percent of weekday trips, it is a major contribution to traffic congestion and highway construction due to its concentrated travel time during the day. According to the U.S. Census Bureau (2008), most American workers spend more time commuting to work than in the past, with those driving private vehicles making up 75.5 percent of workers commuting to their workplaces For instance, average travel time to work increased from 22.4 minutes in 1990 to 25.5 minutes in 2008. Moreover, the number of people who start their commute between 5 to 6 am to avoid traffic jams on highways during peak commuting times (6 am – 10 am) has risen significantly (6.4 percent in 1990, 7.6 percent in 2000, and 8.6 percent in 2008).

The Census Transportation Planning Package (CTPP) data, published by the Department of Transportation (DOT), are special tabulations from the U.S. census data designed for transportation planners. It is published as decennial datasets too, along with the census. This data series has been the most popular and authoritative data source for studies of commuting patterns in metropolitan areas in the United States. This decennial data series releases data aggregated by geographical units such as census statistical units (e.g. census tract) and Traffic Analysis Zones (TAZ). However, as the zoning schemes of census statistics units and the TAZ often change in the 10 years between two releases, the zoning schemes become inconsistent. For example, the boundaries of (TAZs) are different in 1990 and in 2000. This creates problems for time-series studies of commuting patterns. To make the even more complicate, CTPP trip data are not always available at levels of spatial aggregation. . Data in some places are available only at county level as the finest spatial units. This is a barrier when researchers need commuting data at finer spatial granularity. The inconsistency of data in the geographic units may expose studies to possible the modifiable area unit problem (MAUP), which is a significant caution for spatial studies which utilize aggregate data sources (Unwin, 1996). Spatial interpolation has been widely used to improve the spatial granularity of data, or to mediate between inconsistent zoning schemes of spatial data. However, traditional areal interpolation methods only deals with point or area data, they are not directly applicable to flow (line) data which involves a pair of points (or areas) with each basic element for interpolation. This study aims to propose the concept of flow-data interpolation and develop a first model of such.

2 LITERATURE REVIEW

Spatial analyses such as spatial interpolation and spatial interaction (SI) modeling are two research fields that are closely related to the topic of this study. Spatial interpolation translates data values from one zoning scheme to another (Lam 1983). Spatial interaction modeling concerns direct or abstract movements between origins and destinations. In urban transportation research, SI models have been widely used to estimate trips between home and workplace locations (Horner 2002, Fotheringham and O’Kelly 1989).

2.1 Spatial Interpolation

Spatial interpolation can be further categorized as point and areal interpolation (Burrough 1986, Flowerdew and Green 1994). Point interpolation derives data at for new points based on a set of existing points with known data. Areal interpolation converts geographic data from one zoning scheme to another (Fisher and Langford 1995). The areal weighting interpolation method is the simplest and arguably most popular interpolation method (Lam 1983; Flowerdew and Green 1992; Goodchild et al. 1993). This interpolation method assumes that data are uniformly distributed in the source zones. Therefore the method contributes a fraction of data in the source zone to a target zone, in proportion to the intersection area between the target zone and the source zone. This method can be mathematically defined in Equations 1 and 2 (following Lam 1983)

$$Y_t = \sum_{s=1}^{n} \frac{Y_s A_{s \cap z}}{A_s}, \tag{1}$$

$$Y_t = \frac{Y_s A_t}{A_s} \tag{2}$$

Where, Y_t = data in the target zone, Y_s = data in the source zone, $A_{s \cap z}$ = areas of the intersection between source and target zone, A_s = areas of the source zone, A_t= areas of the target zone, and n = total number of source zones. One of the major criticisms has been on the oversimplified assumption of uniform distraction of data within source zones. There are indeed other alternative interpolation approaches that are free of such assumption. A notable example is Tobler's Pycnophylactic interpolation (1979).

More recently, a new line of interpolation methods appears is found to significantly improve estimation accuracy. These methods make use of additional (ancillary) information that can reveal more detailed spatial structure of the data of concern. These methods are sometimes called dasymetric mapping or more generally intelligent interpolation. Flowerdew and Green (1992) applied the EM algorithm, which improved interpolation of target zones with available ancillary data. Rather than using the traditional areal interpolation method, which uses arbitrary boundaries (Census tract or TAZ), the dasymetric mapping technique uses ancillary datasets (land use/land cover) to delineate areas of homogeneous values, thus creating areas that are much more realistic and accurate (Fisher and Langford 1995, Eicher and Brewer 2001, Garb *et al* 2007). Langford (2003) argued that the dasymetric method can improve accuracy by as much as 33 percent, as compared to cartographic mapping techniques such as choropleth mapping. Moreover, Fisher and Langford (1995) assert that no sophisticated areal interpolation could be better than the simple dasymetric method.

Spatial interpolation can be used as a solution to some problems associated with the well-known modifiable unit problem (MUAP). Openshaw (1984) demonstrated that because spatial zoning systems of geographic data were modifiable (or can be arbitrarily designed) the results of the same type of analysis vary among different spatial zoning systems. This issue is referred to as the modifiable area unit problem (MAUP). The MAUP can be an outcome of either scale or zoning effects (Openshaw 1984, Amrhein 1995). The scale effect refers to the impact of the level of spatial aggregation. Geographic data that span across different scales (or spatial resolution) may give rise to inconsistent results in analysis. The zoning effect is related to the changes in findings due to different spatial partitioning of zones, even though the scale of the zones remains constant. Spatial interpolation can be useful when dealing with the zoning effect by interpolating data to identical spatial partitioning of zones (or zoning scheme).

2.2 Spatial Interaction

Spatial interaction refers to movements between origins (e.g., home) and destinations (e.g., jobs). As noted by Fotheringham and O'Kelly (1989), spatial interaction model had been applied in various studies of migration, journey-to-work, and retail location. The gravity model is the most widely accepted and applied techniques in modeling spatial interactions. In the gravity model, the quantity of interaction (flow) is proportional to the propelling power of origin and the attraction power of the destination, while it is inversely proportional to the distance (or a function of distance) between the origin and the destination. It is clear that an interaction is highly dependent on distance in SI modeling. According to Tobler (1970), "*Everything is related to everything else, but near things are more related than distant things*." Tobler's First Law of Geography regards distance as impedance (decay). Thus, commuting patterns of shorter distance anticipate much higher values than those of longer distance. Traditional gravity models neglect intrazonal trips where origin and destination locations are the same. Frost *et al* (1998) suggested that intrazonal problems could be solved by maintaining non-zero values of the actual minimum distance of each zone (See equation 2.4 and equation 2.5). This modification of intrazonal distance is most appropriate for disaggregate zoning schemes. Horner (2002) argued that impedance of distance could be optimized with a transportation problem (TP), which minimizes transportation costs such as distance and travel time.

$$Y_{ij} = \frac{Y_i^{min}}{Y_{ij}} * Y_i^{min} \tag{2.6}$$

Subject to:

$$Y_{ij} = \sqrt{\frac{R_i}{\pi}} \qquad (i = j) \tag{2.7}$$

Where Y_{ij} = new intrazonal distance, Y_i^{min} = min of row i, R_i = area of i, and π = 3.1416

Fotheringham *et al* (2000) distinguished spatial interaction as four stages: (1) social physics – Newtonian gravity model, Lowry model (Lowery 1964), and extensions of the Lowry model, (2) statistical mechanics – entropy model (Wilson 1967 and 1971), (3) aspatial information processing – discrete choice model (McFadden 1974), and (4) spatial information processing – nested logit and competing destination model (Fotheringham 2000, Roy and Thill 2004). The first three steps did not deal with spatial issues because simply adopting the concepts of physics and statistical mechanics imposed the limitation of human decision making in terms of spatial processes. Instead, the last step, called the competing destination model, focused on individual behavior in seeking locations, and thus considered the likelihood of an alternative option in the true spatial choice set, incorporating such variables as individual decisions, behavior, and spatial awareness.

Much scholarly work has been done on the topic of how aggregated/disaggregated data play an important role in the use and analysis of spatial and temporal trends (Horner and Murray 2002). Pirie (1979) saw that there was a significant disparity between aggregated and disaggregated data in terms of individual activities based on zonal aspects. In his work, detailed analysis of the spatial distribution of activities was ignored at the aggregate level; instead he focused on disaggregated data that provided a more realistic representation of travel time and mode of transport by different population subgroups. O'Kelly and Lee (2005) continued to argue that the current aggregated data analysis had some difficulties depicting worker behavior, in comparison to disaggregated socio-economic factors (race, sex, occupation type). Fotheringham and O'Kelly (1989) noted that individual decisions could also be analyzed in disaggregated flows.

3 RESEARCH DESIGN

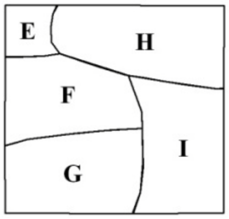
Based on the review of spatial interpolation, there is no existing interpolation model for spatial interaction data. The purpose of this research is to propose a theoretical framework for it and to develop a first model under the theoretical framework. For illustration purpose, we will use commuting flow as sample spatial interaction data in discussions hereafter.

Definition Flow line interpolation: a process of converting spatial flows among spatial units in one zoning scheme to those in another scheme.

Figure (1) shows two zoning schemes, namely TAZ and census tract. If the magnitudes of flows (numbers of trips) between TAZs (a) are known and that between TAZs are unknown, flow line interpolation can be applied to derive flows between tracts, and vice versa. Flows can be represented either graphically or in a matrix form. Figure 1shows commuting trips in a matrix form with each zoning scheme. In the matrix representation , the cell value at row i and column j refers to the magnitude of flow (e.g. number of trips) from zone i to zone j. Commuting trips between five different pairs of TAZs have 20 possible directions (inbound and outbound flows).

For example, zone E has outbound flows ($\frac{\rightarrow}{EF}$, $\frac{\rightarrow}{EG}$, $\frac{\rightarrow}{EH}$, and $\frac{\rightarrow}{EI}$) and inbound flows ($\frac{\leftarrow}{EF}$, $\frac{\leftarrow}{EG}$, $\frac{\leftarrow}{EH}$, and $\frac{\leftarrow}{EI}$) respectively. In contrast, the JTW in 2000 Census tracts have possible commuting directions: $\frac{\rightarrow}{AB}$, $\frac{\rightarrow}{AC}$, $\frac{\rightarrow}{BA}$, $\frac{\rightarrow}{BC}$, $\frac{\rightarrow}{CA}$, and $\frac{\rightarrow}{CB}$. Each pair of the JTW data has a commuting value: 100 in $\frac{\rightarrow}{EG}$ and 90 in $\frac{\rightarrow}{AC}$ (See Figure 1).

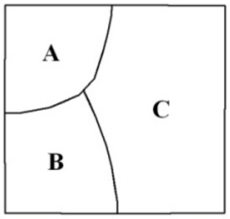
TAZ



TAZ Flows

	E	F	G	H	I
E	10	10	100	50	30
F	5	50	30	10	5
G	20	120	80	10	0
H	0	10	30	35	0
I	0	100	50	15	75

Tract



Tract Flows

	A	B	C
A	50	110	90
B	70	100	50
C	80	180	190

Figure 1: TAZs in 1990 and Census Tracts in 2000

For this study, we will present a flow line interpolation method based on the assumption that is comparable to that of the areal weighting interpolation. It assumes that total trips between i and j (T_{ij}) has origins uniformly distributed in zone i and has destinations uniformly distributed in zone j. We can call it areal weighted flow line interpolation. With this assumption, Equation (3) presents the model to calculate trips between i and j in the target zoning scheme based on the flow information in source zoning scheme.

$$F_{ij}^{t} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[P^{a} * \ P^{b} * \ T_{ij}^{s}\right] \tag{3.1}$$

Subject to:

$$P^{a} = \frac{A_{iO}^{s \cap t}}{A_{i}^{s}} \tag{3.2}$$

$$P^{b} = \frac{A_{jD}^{s \cap t}}{A_{j}^{s}} \tag{3.3}$$

Where, F_{ij}^{t} = commuting flow of target zone scheme, i = origin unit of flow in source scheme, j = destination unit of flow in source scheme, O = number of origin location, D = number of destination location, T_{ij}^{s} = commuting flow of source zone scheme, P^{a} = proportion of origin areas of the intersection between origin unit in source and origin area in source, P^{b} = proportion of destination areas of the intersection between destination unit in source and destination area in source, $A_{iO}^{s \cap z} =$ areas of origin flow unit in the intersection between source and target zone, A_{i}^{s} = areas of source zone in original unit of flow, $A_{jD}^{s \cap z} =$ areas of destination flow unit in the intersection between source and target zone, and A_{j}^{s} = areas of source zone in destination unit of flow.

To estimate flow data between two different boundaries in empirical study areas, the proportion rates of the overlapped areas of origin and destination zones are calculated by an association rule and the percentage rates of their correspondence to the overlapped areas. The association rule, which discovers a certain rule between two variables, is created from an intersection between source zones (TAZs) and target zones (Tract). For instance, Zone A in Tract 2000 intersects with Zones E, F, and H in TAZ 1990. The commuting trips of $\frac{\rightarrow}{AC}$ in 2000 associated with the 10 directions of the JTW in 1990 depended on the intersection of Zone A (E, F, and H) and Zone C (F, G, H, and I). Thus, the new commuting trips of $\frac{\rightarrow}{AC}$ in 1990 can be generated with flows, including pairs of $\frac{\rightarrow}{EF}$, $\frac{\rightarrow}{EG}$, $\frac{\rightarrow}{EH}$, $\frac{\rightarrow}{EI}$, $\frac{\rightarrow}{FG}$, $\frac{\rightarrow}{FH}$, $\frac{\rightarrow}{FI}$, $\frac{\rightarrow}{HF}$, $\frac{\rightarrow}{HG}$, and $\frac{\rightarrow}{HI}$.

The next step is to calculate the percentage of each feature in target zones where the polygon contains the overlayed features of source zones using the association rule. The percentage rates are implemented using ArcObjects with Microsoft Visual Basic for Application (VBA). For example, to create a year 1990 flow with census tract zones, a base layer (TAZ) intersects with a source layer (Tract). Zone F in 1990 covers 43% of Zone A, 29% of Zone B, and 28 % of Zone C in 2000 (See Table 1).

Table 1 Association and Proportion of TAZs (1990) Compared to Tract (2000)

Source	Target			Total
E	A			
	100%			100%
F	A	B	C	
	43%	29%	28%	100%
G	B	C		
	77%	23%		100%
H	A	C		
	25%	75%		100%
I	C			
	100%			100%

The commuting flows of $\frac{\rightarrow}{BC}$ are associated with eight directions in the source zones (F, G, H, and I). The commuting flow, $\frac{\rightarrow}{GF}$ in $\frac{\rightarrow}{BC}$ is 120. Thus, a new flow of $\frac{\rightarrow}{GF}$ in TAZ 1990 for a new flow of $\frac{\rightarrow}{BC}$ in Tracts 2000 is calculated by 120 * 77% * 28% = 26. The remaining flows assign values to 4 $(\frac{\rightarrow}{FF})$, 2 $(\frac{\rightarrow}{FG})$, 2 $(\frac{\rightarrow}{FH})$, 1 $(\frac{\rightarrow}{FI})$, 14 $(\frac{\rightarrow}{GG})$, 6 $(\frac{\rightarrow}{GH})$, and 0 $(\frac{\rightarrow}{GI})$. A total flow of the proportion $\frac{\rightarrow}{BC}$ is 56 (See Table 2).

Table 2 Commuting flow matrix

			A	A	A	B	B	C	C	C	C
			E	**F**	**H**	**F**	**G**	**F**	**G**	**H**	**I**
			1.00	0.43	0.25	0.29	0.77	0.28	0.23	0.75	1.00
A	**E**	1.00	10	10	50	10	100	10	100	50	30
A	**F**	0.43	5	50	10	50	30	50	30	10	5
A	**H**	0.25	0	10	35	10	30	10	30	35	0
B	**F**	0.29	5	50	10	50	30	50	30	10	5
B	**G**	0.77	20	120	10	120	80	120	80	10	0
C	**F**	0.28	5	50	10	50	30	50	30	10	5
C	**G**	0.23	20	120	10	120	80	120	80	10	0
C	**H**	0.75	0	10	35	10	30	10	30	35	0
C	**I**	1.00	0	100	15	100	50	100	50	15	75

VBA for MS Excel and Structure Query Language (SQL) in MS Access are used to create Table 2. First, we need to make a table based on two proportion rates. The table requires six columns, such as F00 (From zone 2000), F90 (From zone 1990), T00 (To zone 2000), T90 (To zone 1990), P1 (proportion in origin zone scheme), and P2 (proportion in destination zone scheme). A pseudo-code for the purpose of creating the table is as follows (See Figure 2):

Figure 2: Pseudo-code and Description

```
Code                                        Description
Count = 1                                   'Initialize Count as 1

  For mRow = 1 To m                         ' Set maximum row
          For mCol = 1 To n                 ' Set maximum column
                  a = Cells(mRow + 1, 1)    ' Get values from cells using nested loop
                  b = Cells(mRow + 1, 2)

                  c = Cells(mCol + 1, 1)
                  d = Cells(mCol + 1, 2)

                  e = Cells(mRow + 1, 3)
                  f =  Cells(mCol + 1, 3)
```

```

            Write #1, a, b, c, d, e, f      ' Write a, b, c, d, e, and f
            Count = Count + 1               ‘ Increment
        Next mCol
    Next mRow
```

By applying the pseudo-code in Figure 2, VBA for MS Excel creates a txt file format, and then is imported into MS Access to estimate final flow data. A new table is called NTRT90 and is merged with a table (FlowTAZ90) from TAZ 1990. Structure Query Language (SQL) deals with a multiple joining operation, which combines fields from two tables. Whereas the FlowTAZ90 table has three columns (F90, T90, and FLOW), the NTRT90 table has six columns: F00 (From zone 2000), F90 (From zone 1990), T00 (To zone 2000), T90 (To zone 1990), P1 (proportion in origin zone scheme), and P2 (proportion in destination zone scheme). For instance, two conditions need to be set: F90 in FlowTAZ90 is connected with F90 in NTRT90 and also T90 in FlowTAZ90 with T90 in NTRT90 (See Figure 3).

Figure 3 A SQL Code

```
SELECT N.F00, N.T00, Sum([P1]*[P2]*[FLOW]) as NFLOW
FROM FlowTAZ90 F, NTRT90 N,
WHERE (F.F90 = N.F90) AND (F.T90 = N.T90)
        GROUP BY N.F00, N.T00;
```

By applying the new method, new tract-based JTW in 1990 and TAZ-based JTW in 2000 are estimated. The total values in the new tract record keep the same total value of observed flows (845). In addition, the total values in the new TAZs are the same as original values of source flow (920) as well (See Table 3 and Figure 1).

Table 3: A New Generated JTW in 1990 and JTW in 2000

New Tract in 1990

	A	**B**	**C**
A	43	103	117
B	65	85	56
C	82	120	176

New TAZ in 2000

	E	**F**	**G**	**H**	**I**
E	3	16	24	12	10
F	13	53	76	41	31
G	16	56	74	39	24
H	12	55	83	48	41
I	9	44	65	40	35

4 Model Evaluation and Case Study

Several software tools for the abovementioned methodology have been developed in the study. First, the percentage rates are implemented using ArcObjects with Microsoft Visual Basic for Application (VBA), Second, VBA for MS Excel creates a txt file format (.txt), comprised of six columns: FS (from source zone), FT (from target zone), TS (to source zone), TT (to target zone), P1 (proportion in origin zone scheme for source zone), and P2 (proportion in destination zone scheme for source zone). Finally, MS Access calculates the new flows, i.e., FT (from target zone), TT (to target zone), and NF (new flow), using Structure Query Language (SQL) based on the template table and the source flow data (from source zone, to source zone, original flow). In order to evaluate the validity and accuracy of the proposed method, I conducted a case study as explained below.

4.1 Study Area

The study area is Fulton County, located in the middle of the Atlanta, Georgia metropolitan area. As of the 2008 Census, the population was 1,014,932, up from 816,006 in

2000 and 648,951 in 1990. This means that the county is the most populous in the state of Georgia. The population increase was 20.5 percent from to 2000, and 24.4 percent from 2000 to 2008. The total number of workers in Fulton County was 315,336 in 1990 and 385,442 in 2000, an increase of 18.2%. Figure 4 shows a map of Fulton County. There are 167 census tracts and 492 small places listed by the Metropolitan Planning Organization (MPO). The total number of trips taken in 2000 was 264,100 within census tract boundaries and 262,021 within small places.

The main goal of this section is to match commuting flow data based on different boundaries. Flow and boundary data are obtained from the CTPP part 3. Census tract and small places in MPO are the smallest zones available in 2000 CTPP. Using equation 3.1, the final output of commuting flow data can be produced in an identical spatial zoning scheme by either new census tracts or new small places in MPO boundaries.

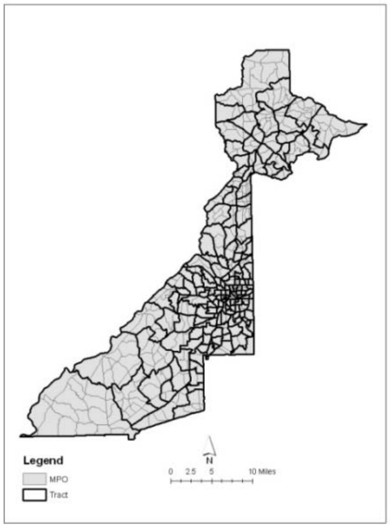


Figure 4: Fulton County, GA

4.2 Accuracy Evaluation

Statistical techniques such as Root Mean Square Error (RMSE) and F test are utilized to examine and validate the results of the data. Whereas RMSE is a statistical test commonly used to validate how closely expected values correspond to their actual values, the F test is used to check for difference among variances. First, the RMSE is the square root of the variance of the residuals between actual and expected values. Lower values of the RMSE represent better goodness of fit. The RMSE is a good measure of how close the model generates estimated values based on actual values. It can be calculated with the following equation:

$$RMSE = \sqrt{\frac{\sum(X - X')^2}{N}} \tag{3.4}$$

Where X is the actual values, X' is the observed values and N is the total number.

In this study, we calculate the estimated values by either an aggregated or disaggregated approach. In Fulton County, there are 167 census tracts and 492 small places within Metropolitan Planning Organization (MPO) boundaries. Total trips in Fulton County are 264,100 within census tract boundaries and 262,021 within small places boundaries. Newly estimated trips are 261,813 within census tracts and 264,096 within small places. The corresponding percent changes are -1% within census tracts and 1% in small places. The RMSE of newly estimated commuting data from census tracts to small places is 4.996 and the RMSE from small places to census tracts is 1.2871 (See Table 4). Both the RMSE of census tracts and small places are relatively low, which shows that pairs of newly estimated commuting data are sufficiently close to the actual values of the pairs. However, there is no standard scale of RMSE, so we do not know whether the results of the estimated values are valid for their corresponding actual values. In other words, how accurate is the model's estimated flow? In general, the RMSE is a good measure of a goodness of fit model, but not a solid statistical approach to validate accuracy evaluation.

Table 4: Summary of Accuracy Analysis

	Census Tract	**Small Places**
Actual Flows	264100	262021
Expected Flows	261813	264096
% Change	-1%	1%
RMSE	1.2871	4.9960

An F test is a statistical test, which is an analysis of the variance of a regression to test whether a group of variables is significant or not. The F test equation is as follows.

$$F = \frac{S_1^2}{S_2^2} \tag{3.5}$$

Where S_1^2 is the larger sample variance and S_2^2 is the smaller sample variance

First, null hypothesis can be defined as describing a case in which "*the variances of two flow data are equal.*" If an F value is smaller than an F critical value and a P value is greater than the level of significance, then we fail to reject the null hypothesis. Thus, the F-ratio indicates that the newly estimated values are similar to the actual values. Otherwise, we reject the null hypothesis by saying that there is a significant discrepancy between their variances. In this study, we test to see if the variance of the estimated values of commuting data varies significantly from the variance of the actual values of commuting data. The aggregated approach (Small Places to Census Tracts) in F test shows that the F value (1.0037) is smaller than the F critical value (1.0200), and that the P value (0.3789) is greater than the level of significance (0.05). Thus, we conclude that the variance of the estimated commuting trips is not significantly different from that of the number of actual trips. In contrast, the disaggregated approach (Census Tracts to Small Places) represents that the F value (1.9946) is greater than the F critical value (1.0067) and the P value (0) is smaller than the level of significance (0.05), indicating that there is evidence to support the conclusion that there is a significant variance difference between the actual and the expected values.

In summary, both results of the RMSE are relatively low, showing that the actual and the expected are similar to each other. But the RMSE cannot tell us how good the model is because it is just a diagnostic for the model, which doesn't represent a standard scale for goodness of fit. In contrast, a Chi-squared test can assess the model's goodness of fit. In our scenarios, the Chi-squared test is not valid. One of the reasons is that if expected values lower than 5 comprise more than 10% of the commuting data, then the Chi-square test is not normally acceptable. Expected commuting data below 5 in our study comprise about 76% of the trips within census tracts and 95% of those in small places. Thus, we use the F test to examine whether the variances of two trips are equal. The F test shows that the new data for commuting within census tracts are only acceptable for the purpose of validating the accuracy evaluation (See Table 5). The problem of the disaggregated approach indicates that though the percent change of total trips based on the actual and the expected is very small, some pairs of commuting trips do not match each other between them. It is true that there are several sources of uncertainty when estimating flows from larger zones to smaller zones. One of the problems is that the model we present calculates the newly estimated commuting flows based on the proportion rates of origin and destination schemes, without differentiating land use type. Commuting trips of higher values mostly relate to urban areas and road networks. To get higher accuracy of results in the model, we need to assign higher proportional values to urban areas (high/low density residential, commercial and industrial), compared to lower proportional values on agriculture or forest land.

Table 5: F-test Two-Sample for Variances

Census Tract

	Actual	*Expected*
Mean	9.469683388	9.38768
Variance	1494.228938	1488.719
Observations	27889	27889
df	27888	27888
F	1.003701394	
P(F<=f) one-tail	0.378855764	
F Critical one-tail	1.019894851	

Small Places in MPO

	Actual	*Expected*
Mean	1.08158586	1.091016
Variance	41.9662805	21.04042
Observations	242064	242064
df	242063	242063
F	1.9945554	
P(F<=f) one-tail	0	
F Critical one-tail	1.00670883	

4.3 Interpretation and Discussion

Due to inconsistent geographical units of the decennial commuting data, problems occur when spatial-temporal analysis is implemented. For example, the boundaries of TAZs and census tracts are different in 1990 and 2000. Traditional area interpolation methods can identify two different zones, but have some difficulties in dealing with pairs of commuting data. The new method can possibly match and calculate flow data at various spatial aggregation levels. As discussed earlier, the newly estimated flow data at the aggregate level fit with the actual flow data. The F test shows that there is no significant difference between the actual and the expected. Thus, by applying this method, time series flow data can be produced in an identical spatial zoning scheme at the aggregate level and then finally can perform investigations of spatial-temporal changes.

Furthermore, the American Community Survey (ACS) program launches more frequent commuting data at coarser spatial granularity (e.g., county level). This method may allow us to interpolate the ACS data into higher levels of spatial granularity. In contrast, the estimated flow data at the disaggregate level must be treated with caution. The method evenly assigns the proportion of the overlapped area of the trip origin and the trip destination zones. Thus, it may impact the accuracy of the estimated flow in certain zones. For example, it is reasonable to assume that urban areas have higher traffic flows than forest and barren areas. To distinguish the land use type and assign different proportional rates of the overlapped areas in origin and destination zones will greatly improve the accuracy of the new flow data, especially with regard to commuting flow in metropolitan areas.

5 CONCLUSIONS AND FUTURE STUDY

The main goal of the study is to estimate commuting data based on the two different boundaries. By applying the method, spatial-temporal analysis can be performed in the study area. However, this commuting data only deals with homogenous characteristics like total commuting flows in metropolitan areas. This aggregated approach of homogenous commuting flow is appropriate for predicting overall flow patterns in a metropolitan area, but may mislead in light of these heterogeneous commuting patterns. To deal with this problem, future research is needed to develop a disaggregated approach which aims to extract data from subgroups, rather than the aggregate, thus making it possible to analyze the heterogeneous commuting patterns of individual workers.

To extend the model in order to examine commuting patterns by socio-economic variables, researchers can investigate commuting patterns by different population subgroups such as those categorized by race, gender, and income, and finally come to understand the wide differences in commuting behaviors between different groups of people. Thus, future study on disaggregated work trips with the interpolating commuting data we presented earlier will provide more realistic representations of correlations between jobs and housing in terms of race, gender and income level over time.

REFERENCES

Amrhein, C. (1995). Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment and Planning A*, (27): 105-119.

Blalock, H. (1964). Causal Inferences in Nonexperimental Research. New York: Seminar Press.

Bookout, L. (1990). Jobs and housing: the search for balance. *Urban Land*, (8): 5 - 9.

Burrough, P. (1986). Principles of GIS for Land Resources Assessment. Clarendon, Oxford.

Cervero, R. (1989). Jobs-Housing Balancing and Regional Mobility. *Journal of the American Planning Association*, 55: 136-150.

Eicher, C. and Brewer, C. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28(2): 125-138.

Fisher P. and Langford M. (1995). Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environmental and Planning A*, 27: 211-224.

Flowerdew, R. and Green, M. (1992). Developments in areal interpolation methods in GIS. *The Annals of Regional Science*, 26:67-78.

Fotheringham, A. and O’Kelly, M. (1989). *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic, Amsterdam.

Fotheringham, A., Brunsdon, C., & Charlton, M. (2000). Quantitative Geography: Perspectives on Spatial Data Analysis. SAGE Publications.

Frost, M, Linneker, B, and Spence, N. (1998). Excess or wasteful commuting in selection of British Cities. Transportation Research Part A: Policy and Practice, 32(7): 529-538.

Garb, J. and Cromley, R. and Wait, R. (2007). Estimating Populations at Risk for Disaster Preparedness and Response. *Journal of Homeland Security and Emergency Management*, 4(1): 1 – 17.

Goodchild, M. and Lam, L. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo – processing*, 1: 297 - 312.

Horner, M. (2002) and Murray, A. Excess Commuting and the Modifiable Areal Unit Problem. *Urban Studies*, 39(1): 131-139.

Horner, M. (2002). Extensions to the concept of excess commuting. *Environment and Planning A*, 43: 543- 566.

Horner, M. (2004). Spatial Dimensions of Urban Commuting: A Review of Major Issues and Their Implication for Future Geographic Research. *The Professional Geographer*, 56(2): 160-173.

Kain, J. (1968). Housing segregation, Negro employment and metropolitan decentralization. *Quarterly Journal of Economics*, 82: 175-197.

Lam, N. (1983). Spatial Interpolation Methods: A Review. *The American Cartographer*, 10(2): 129 -149.

Langford, M. (2003). Refining methods for dasymetric mapping using satellite remote sensing. Remotely Sensed cities, London: Taylor and Francis, 31: 19 - 32.

Lowry, I. (1964). *A model of metropolis*, RM-4035 -RC, Rand Corporation, Santa Monica.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in *Frontiers in Econometrics, Zarembka, P., ed. New York: Academic* Press, 105-142.

O'Kelly, M. and Lee, W. (2005). Disaggregate journey-to-work data: implications for excess commuting and jobs – housing balance. *Environment and Planning A*, 37(2): 2233-2252.

Openshaw, S. (1983). The modifiable area unit problem. *Concepts and Techniques in Modern geography*, (38): Norwich: Geobooks.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Pirie, G. (1979). Measuring accessibility: a review and proposal. *Environment and Planning A*, 11: 299-312.

Roy, J. Thill, J. (2004). Spatial Interaction modeling. *Regional Science*, 83: 339-361.

Sadahiro, Y. (1999). Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographic Systems*, 1:323-346.

Taaffe, E., & Gauthier, H., & O'Kelly, M. (1996). *Geography of Transportation*. Prentice Hall.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240.

U.S. Census Bureau (2008). Retrieved October 2008 from http://www.census.gov/.

Wilson, A. (1967). A statistical theory of spatial distribution model. Transportation Research, 1: 253-269.

Wilson, A. (1971). A family of spatial interaction models and associated developments. Environment and Planning, 3: 1-32.