SMALL AREA ESTIMATION OF FREIGHT FLOWS: THE CASE OF GEOGRAPHICALLY REFINED DESTINATIONS

Paul Metaxatos

ABSTRACT. Small area statistics have become increasingly critical for decision making in the public and private sectors of the economy. However, for reasons associated with disclosure of confidential information, data is often released on a fairly coarse geography (e.g., state level) vis-à-vis a much finer geographical level (e.g., county). This has led to extensive research on small area estimation - i.e., estimation at a more detailed geographical level based on data at a coarser level. Most of this work has been single-area-specific or non-flow data. Flows, at a minimum, have origin and destination location specificity which leads to greater complexity. This paper addresses this issue and provides a methodology based on the gravity model which is then demonstrated on freight data in the special case of geographically refined destination ends.

INTRODUCTION

Small area statistics have become increasingly critical for decision making in regard to formulation and administration of policies and programs, allocation of government funds, regional planning, and other purposes in the public and private sectors of the economy. However, for reasons associated with disclosure of personal information, data is often released on a fairly coarse geography (e.g., state level), although they are needed at a much finer geographical level (e.g., county). This has led to extensive research on small area estimation - i.e., estimation at a more detailed geographical level based on data at a coarser level (see, Rao, 2003 and Jiang and Lahiri, 2006 for reviews of the literature).

Most of this work has been single-area-specific or non-flow data (e.g., population surveys at a health service area, economic surveys at a school district, etc.). Flows, at a minimum, have origins and destinations which leads to greater complexity. This paper addresses this issue and provides a methodology which is then demonstrated on available freight flow data. The particular application illustrates the distribution of international-trade

1Research Assistant Professor, Urban Transportation Center, University of Illinois at Chicago, e-mail: pavlos@uic.edu
freight flows within the United States (U.S.) from ports of entry to small-area destinations (counties).

This is not the first effort on small area estimation of freight data. Private vendors provide such estimates, but their methods are proprietary. Moreover, they are synthesized from several data sources and likely to be based on more ‘ad hoc’ type methodologies. In addition, such data are not suitable for freight route planning because the geographic scale is not sufficiently detailed. The methodology presented in this paper has a sound theoretical basis and can be adapted for use in metropolitan freight planning.

An Example

Consider two U.S. states (e.g., New York and Pennsylvania) as seen in Figure 1. Suppose we know the total flow of commodities $T$ (value or tons) from New York to Pennsylvania.

![Figure 1: The states of New York and Pennsylvania.](image)

Since there are 62 counties in New York and 67 counties in Pennsylvania, a simple minded approach would be to divide $T$ by 62 and call them estimates of flow between counties. This is too naive. Undoubtedly, the flows are related to some measure of population and economic activity. Call such a measure $a_i$ for the originating state and $b_j$ for the destination state — where $i$ and $j$ index the counties. Then a less naive estimate of flow from $i$ to $j$ would be proportional to $a_ib_j$ and the proportionality can be made into an equality by scaling to add to the total $T$. However, this too is naive. We know that flows to more difficult to reach destinations would be less. Thus an adjustment is needed and we can estimate the flows to be

$$a_ib_jf_{ij}$$

(1)

where $f_{ij}$ reflects the level of difficulty of shipping from $i$ to $j$ and a constant of proportionality is assumed absorbed into one of the factors. But (1) is the gravity model (see below). The discussion is akin to that given by Carol and Bevis (see Sen and Smith, 1995).

In (1), $f_{ij}$ was used in a multiplicative way, i.e., $f_{ij}$ is multiplied with the other terms.
Intuitively, this is reasonable, since one would expect that doubling activity levels \(a_i\) or \(b_j\) would double flows irrespective of \(f_{ij}\), so long as it does not change.

**The Gravity Model**

A modern version of the gravity model (see Sen and Smith, 1995) is

\[
N_{ij} = T_{ij} + \epsilon_{ij}
\]

where \(N_{ij}\) is the flow from \(i\) to \(j\), \(\epsilon_{ij}\) an error term and \(T_{ij} = E[N_{ij}]\), the expectation of \(N_{ij}\), is written as

\[
T_{ij} = A_i B_j F_{ij}.
\]

In (3), \(A_i, B_j\) reflect activity at \(i\) and \(j\) and \(F_{ij}\) reflects the difficulty of getting from \(i\) to \(j\) or of shipping from \(i\) to \(j\).

The gravity model has been intensively studied and has been given a sound theoretical basis. In Sen and Smith (1995), the model has been derived on the basis of a very small number of axioms. Since these axioms are few and intuitively easily believable and also because they are shown to be both necessary and sufficient, some (including the author) find the theory of the gravity model more complete than the usual theories cited for the logit model. However, because of the relationship between the gravity and logit models, it is most likely that the theory behind the gravity model can be extended to cover the logit models — although this has not occurred so far as we know. Unfortunately all this is not too well known perhaps because the gravity model theory is mathematically rather demanding.

For convenience in estimation, \(F_{ij}\) is written, without loss of generality, as

\[
F_{ij} = \exp[\phi_{ij}],
\]

where \(\phi_{ij} = \log[F_{ij}]\). It is even more convenient to write

\[
\phi_{ij} = \sum_{k=1}^{K} \theta_k c_{ij}^k,
\]

where \(\theta_k\) are parameters to be estimated and \(c_{ij}^k\) are variables measuring separation between \(i\) and \(j\). From a practical standpoint, this specification also does not lose much generality since it includes polynomials which are known to be dense in space of continuously differentiable functions. Thus the model we will consider is

\[
T_{ij} = A_i B_j \exp[\phi_{ij} = \sum_{k=1}^{K} \theta_k c_{ij}^k].
\]

**Freight and the Gravity Model**

The gravity model has been used before for freight flow estimation (Sen and Pruthi, 1983; Ashtakala and Murthy (1993); Smadi and Maze, 1996; Black, 1999; Cheu et al., 2003; Metaxatos, 2004; Matsumoto, 2007). However, most theoretical derivations of the gravity
model are confined to people movements. While it is possible that these theories can be extended to freight (Sen and Smith, 1995; Metaxatos, 2004), for the purposes of this paper, we find the discussion above adequately persuasive.

Note also the close connection between the gravity model and linear programming. If any of the $\theta$s in (6) goes to $-\infty$, while other terms stay the same, $T_{ij}$ becomes a linear programming solution (Evans, 1973; Senior and Wilson, 1974; Erlander and Stewart, 1990; Sen and Smith, 1995). This fact also might yield a theoretical basis for the gravity model.

Extending to freight a model originally justified for passenger flows also creates an estimation problem. Maximum Likelihood (ML) procedures for estimation are based on the Poisson or multinomial distribution. It is unlikely that freight flows, which combine flows of coal (which move in train-loads) and much smaller deliveries, have either kind of distribution. The number of deliveries might be Poisson (although that is not clear) but the size is something else which would depend on origin and destination location (e.g., consider power stations vs. a retail outlet). The combined distribution would be very complex.

There are two possible simple remedies. One is to use least squares, which does not require knowledge of distributions provided the Gauss-Markov conditions are met (Sen and Srivastava, 1990, p.35). The key conditions in this context is that of equality of variance for which empirical diagnostic procedures can be used and combined with weighting. Alternatively, we could depend on the robustness of ML Procedures as discussed in Sen and Smith (1995).

Estimation of Origin/Destination Activity Levels

Since origin and destination activity levels, $T_{i+}^{(s)} = \sum_j T_{ij}$ and $T_{+j}^{(s)} = \sum_i T_{ij}$, respectively, are at the level of individual origins and destinations, the current literature on small area estimation (e.g., Rao, 2003; Jiang and Lahiri, 2006) can be brought to bear on them. In cases where it is not entirely clear as to how these activity levels are defined, we could also use more traditional multiple regression for this purpose — a method which will help identify them, as well as make relevant estimates. We will report on such methods in future papers. In this paper, we will focus instead on the estimation of small-area bi-directional flows when the geography of destination ends is refined.

AN APPLICATION

The application presented in this paper focuses on the estimation of (international-trade) freight flows within the United States (U.S.) between U.S. ports of entry and individual U.S. counties. We have borrowed some of the data needed from an earlier study (Metaxatos, 2004) but the methodology below and the results are new.

Methodology

The methodology we are proposing for obtaining small-area synthetic origin-destination flow estimates is conceptualized in Figure 2. The steps of the procedure are:
Figure 2: Conceptualization of a SAE methodology for origin-destination flows.

- Step 1: Obtain origin-destination flows, \( N_{ij}^{(L)} \), and separation measures, \( c_{ij}^{(L)} \), for large (e.g., state to state) areas;
- Step 2: Estimate the impedance parameter vector, \( \theta \), that corresponds to each separation measure \( c_{ij}^{(L)} \), using the ML procedure presented in Metaxatos (2004);
- Step 3: Obtain separation measures, \( c_{ij}^{(s)} \), for small areas (e.g., county to county);
- Step 4: Estimate exogenously for small areas commodity flow productions and attractions, \( T_{i+}^{(s)} \) and \( T_{+j}^{(s)} \), respectively; and
- Step 5: Using \( \theta \) (from Step 2), \( c_{ij}^{(s)} \) (from Step 3), and \( T_{i+}^{(s)} \) and \( T_{+j}^{(s)} \) (from Step 4), apply the DSF procedure (described in Metaxatos, 2004), a particular algorithmic implementation of an iterative proportional fitting procedure, to obtain small-area flow estimates, \( T_{ij}^{(s)} \).

For completeness of the presentation, the ML procedure in Step 2 solves the following system of equations:

\[
T_{is} = N_{is} \quad \forall i \in I \tag{7}
\]
\[
T_{sj} = N_{sj} \quad \forall j \in J \tag{8}
\]
\[
\sum_{ij} c_{ij}^{(k)} T_{ij} = \sum_{ij} c_{ij}^{(k)} N_{ij} \quad \forall k \in K \tag{9}
\]

where, \( I \) and \( J \) are the origin and destination sets, respectively, and the operator * indicates summation with respect to the subscript it replaces (e.g. \( T_{is} = \sum_j T_{ij} \), \( T_{sj} = \sum_i T_{ij} \)).
In addition, the DSF procedure in Step 5 iterates as follows (the index \( r \) denotes the iteration number):

\[
A_i^{(2r-1)} = \frac{O_i}{\sum_{j=1}^{J} B_j^{(2r-2)} F_{ij}} \quad \forall i \in I
\]

\[
B_j^{(2r)} = \frac{D_j}{\sum_{i=1}^{I} A_i^{(2r-1)} F_{ij}} \quad \forall j \in J
\]

where \( O_i = T_i^s \), \( D_j = T_j^s \), and \( F_{ij} \) is a function of the separation measures \( c_{ij}^{(k)} \). Upon convergence (see Sen and Smith, 1995 for a proof of convergence), the values of \( T_{ij} \) are given by \( T_{ij}^{(2r)} = A(i)^{(2r-1)}B(j)^{(2r)}F_{ij} \).

Note that in Step 4 above if we exogenously estimate productions and attractions at a small enough geographic scale — say, at the Traffic Analysis Zone (TAZ) level — then the proposed methodology can be adapted for use in metropolitan freight route planning. In addition, further refinement can be achieved if we disaggregate the freight flows by commodity type and industry given data availability.

Figure 3: Application of the SAE methodology for port-to-county freight flows.

In this paper, the methodology has been adjusted to the particular application at hand as seen in Figure 3. The objective here is to estimate international trade freight flows from ports of entry to destination counties. Therefore, in Step 4 of the procedure above we do not need to estimate \( T_{i+}^{(s)} \) because ports of entry are already refined geographically. Regarding estimation of \( T_{i+}^{(s)} \) in Step 4, we borrowed the method developed in Metaxatos (2004). The method simply assigns the total freight tonnage that comes into a state \( j \), \( N_{ij}^{(L)} = \sum_i N_{ij}^{(L)} \), to counties in the state, \( T_{ij}^{(s)} \), based on total employment levels (in all
industries) in each county of each state using the U.S. Census Bureau’s County Business Patterns data (see, http://www.census.gov/epcd/cbp/view/cbpview.html). The latter is an annual census of all business establishments with one or more paid employees including information on payroll, employment and industry.

Data

The particular application in this paper requires the availability of freight shipments from ports of entry (origins) to destination states. We used a data set made available to us by the Oak Ridge National Laboratory (ORNL). The data are from the Port Import Export Reporting Service (PIERS) database. This is a commercial database (see http://www.piers.com/about/default.asp) prepared by the U.S. Department of Commerce and offers statistics on global cargo movements transiting seaports in the U.S., Mexico and South America to companies around the globe. A matrix, $N_{ij}$, between 144 origin seaports $i$ and 50 destination states $j$ with dimensions $144 \times 50$ was developed based on this data. Note that a limitation with using PIERS is that reported origins and destinations may be billing addresses rather than shipment points.

In identifying specific types of spatial separation that tend to impede or enhance the likelihood of interactions between ports of entry and destination states, the most obvious type involves physical space as exemplified by travel distance and travel time which are quantifiable in terms of meaningful units of measurement. In this application, we used the two (one for travel distance and one for travel time) separation measures, $c^{(L)}_{ij}$'s from each origin seaport $i$ to each destination state $j$ discussed in Metaxatos (2004). These two separation measures are matrices with dimensions also $144 \times 50$. From the same study we also borrowed two other separation measures (again, one for travel distance and one for travel time), $c^{(s)}_{ij}$'s from each origin seaport $i$ to each destination county $c$. The last two matrices have dimensions $144 \times 3140$.

Results

In passenger transportation, a flow unit of 1 (consistent with a Poisson or multinomial distribution assumption) would be reasonable. In the case of freight shipments of goods, a basic unit of flow would appear to be a trainload (for shipments by rail) or a truckload (for shipments by truck). In the absence of mode-specific information as well as information related to the variation in modal size, we experimented with different values and determined an ‘optimal’ (with regard to providing the best model fit) basic unit of flow of 100,000 pounds (50 short tons). This is not surprising given that the bulk of flows are long-distance shipments that are usually performed by large trucks (currently, the gross weight limit for a 6-axle combination truck is 80,000 pounds (40 short tons) or rail (on a limited scale for the particular data). Interestingly, our results appeared to be quite insensitive to the choice of the flow unit, largely because the flows were inordinately large. This is in agreement with previous work (Sen and Pruthi, 1983).

Note that if we are interested in estimating the impact of goods movement on the transportation network and the environment (e.g., congestion effects, vehicle miles (kilometers) of travel, etc.) we would need to perform a traffic assignment. In that case, we would have to convert the freight flows into truck trips using, for example, the
Vehicle Inventory and Use Survey (VIUS) produced by the Census Bureau (U.S. Census Bureau, 2004). In this regard, the previous adjustment is only approximate and can be refined using the weight classes in VIUS. Alternative methods are also available (see, Chin and Hwang, 2006).

The sheer size of the matrices involved in this application called for the use of specialized routines first demonstrated for large-scale freight forecasting in Metaxatos (2004). Both the ML procedure (in Step 2 of the algorithm) and the DSF procedure (in Step 5) run to a tight convergence (see Metaxatos, 2004 for details).

A square-root transformation of travel distance and travel time adequately addressed issues with outliers and resulted in excellent model fit as seen in Figure 4. More formal statistics regarding model fit gave a Chi-square ratio value close to 1, and a Pearson correlation coefficient between observed and estimated cell-to-cell flows with a value of 0.89 (see Metaxatos, 2004 for details).

Estimates of the parameters $\theta$’s corresponding to travel distance and travel time separation measures are, respectively, $\hat{\theta}_1 = 11.1227$ and $\hat{\theta}_2 = -16.7447$. The unusual positive sign of the travel distance parameter estimate is due to collinearity between the used impedances, distance and travel time, phenomenon that is well known to adversely affect the sign of parameter estimates. Dropping one of the impedances would bias the parameter estimates left in the model. Given the robustness of the maximum likelihood procedure used in collinearity situations, all available impedance measures were retained (see Sen and Smith, 1995, Chapter 5). After all, the sign of $\hat{\theta}_1$ would have changed to a negative value had we reparameterized the model and considered instead of travel time, as the first impedance measure, the difference between distance and travel time (in appropriate units).

All the data items as well as the parameters are now available for Step 5 of the SAE procedure:

- the freight flow productions for each seaport $i$, from the original data since we did
not have to refine the geographic scale — i.e., \( T_{i+}^{(s)} = \sum_j N_{ij}^{(L)} \),

- the freight flow attractions, \( T_{+j}^{(s)} \), for each county \( j \); and

- the separation measures, \( c_{ij}^{(s)} \)'s, for travel distance and travel time from each port \( i \) to each county \( j \).

Applying the DSF procedure, an iterative proportional fitting algorithm, results in the final estimates of freight flows, \( T_{ij}^{(s)} \), between each port \( i \) and each county \( j \). A test of reasonableness of the latter estimates is developed as follows:

- Compute the freight flow length distribution for seaport-to-state origin-destination pairs using the estimated flows \( T_{ij}^{(L)} \)'s, and separation measures, \( c_{ij}^{(L)} \)'s;

- Aggregate the estimated seaport-to-county flows, \( T_{ij}^{(s)} \), into seaport-to-state flows;

- Compute the freight flow length frequency distribution for seaport-to-state origin-destination pairs using the newly aggregated flows, and separation measures, \( c_{ij}^{(L)} \)'s; and

- Compare the two distributions, either visually or using a more formal statistical method.

As seen in Figure 5 the two distributions are very close. To formally compare the two distributions we used the exact Wilcoxon two-sample test (since the sample size is small, the normal approximation may not be completely accurate, and it is appropriate to compute the exact test). The Wilcoxon statistic (see Agresti, 1992) equals 98.5. The one-sided exact p-value equals 0.323, while the normal approximation yields a one-sided p-value of 0.3276, neither of them significant; the two distributions are practically indistinguishable.
CONCLUSIONS

The unavailability of small-scale origin-destination freight flow data continues to negatively impact the proper management of goods movements at the national, regional and local levels. This paper has presented a method to synthesize such data at a small geographical level. It is based on a sound theoretical basis and can be adapted for use in metropolitan freight route planning.

In addition, the method was demonstrated using available international trade freight flow data and estimated origin-destination freight flows from seaports to counties within the United States. A preliminary verification of the methodology is apparently promising. Understandably, a more robust verification effort would have compared the port-to-county estimated flows with actual data. Regrettably, non-proprietary small-scale data were not conveniently available as of this writing.

ACKNOWLEDGEMENTS

This research has been made possible with continuous support of the Metropolitan Transportation Research Initiative (METSI) at the University of Illinois at Chicago (UIC) Urban Transportation Center (UTC). We are also grateful to Ashish Sen, Professor Emeritus in the College of Urban Planning and Public Affairs at UIC, who commented on an earlier version of this paper.

REFERENCES

Weight and Value Flows”, Journal of Transportation and Statistics. 6(2/3), 67-80.
Applications. Springer-Verlag, New York.
Verlag, New York.
Programming and Spatial Interaction Models of Residential Location”, Geographical
Analysis, 6, 209-237.
and Case Study (With Discussion)”, Transportation Research Record, 1522, 55-63.
U.S. Census Bureau (2004). Vehicle Inventory and Use Survey. U.S. Department of
Commerce, Washington, D.C.