Estimation and Accuracy of Origin-Destination Highway Freight Weight and Value Flows

Paul Metaxatos  
University of Illinois at Chicago  
412 South Peoria Street, Suite 340  
Chicago, IL 60607  
Tel: (312) 996-4713  
Fax: (312) 413-0006  
E-mail: pavlos@uic.edu

Accepted for presentation at the 82nd Annual Meeting of the Transportation Research Board  
Submitted for consideration for publication in the Transportation Research Record

July 2002  
revised November 2002

Number of words: 4,870  
Number of Tables: 4×250=1,000 words  
Number of Figures: 4×250=1,000 words  
Total: 6,870 words
ABSTRACT

This paper proposes and implements a maximum likelihood estimation of highway freight weight and value flows using the gravity model. The computation of the standard error of the flow estimates provides the basis for measuring the level of accuracy of the estimates. The results provide ample evidence of the suitability of gravity models for freight forecasting given the excellent fit and the small variances.

Keywords: freight origin-destination flow estimation, covariance of estimates
INTRODUCTION

The measurement of freight movements requires tracking freight flows across geographic and political boundaries. This is a particularly challenging task beyond the present state and regional data acquisition capabilities. Various mathematical approaches have been implemented (1,2,3) to circumvent this problem but none, to the best of our knowledge, proposes a measure to assess the accuracy of the computed flows.

This paper proposes to fill this gap using developments in spatial interaction modeling that have not been demonstrated on a large scale to date. The methodology computes maximum likelihood flow estimates and obtains their covariance matrices that, in turn, may be used to obtain confidence intervals and carry out certain tests of hypotheses. The approach can accommodate the large number of origins and destinations that one typically encounters in freight (and passenger) travel forecasting.

The methodology is applied to highway freight weight and value flows of international trade traffic between seaports or border ports and destination states (see (4) for details). The variances of the flow estimates computed are remarkably small. The demand for freight transportation flows can now be estimated within a desired confidence level.

THEORETICAL FRAMEWORK

Commodity shipments in this paper are thought to be realized patterns of spatial interactions that typically result from many independent decisions by individual firms, each constituting a relevant subsystem within the economy as a whole. Hence if the travel behavior of each firm is modeled as a very small interaction process the resultant interaction process can be taken to be the superposition of all these processes. It may be argued that for large collections of small frequency processes, the resulting superimposed process is approximately Poisson and, therefore, completely characterized by its associated mean interaction frequencies (5).

In this light, assuming that the observations $N_{ij}$ of shipment weight and value between origin seaports/border ports of entry $i$ and destination states $j$ can be described by the gravity model, we can write:

$$N_{ij} = T_{ij} + \epsilon_{ij}$$
$$T_{ij} = E(N_{ij}) = A_i B_j F_{ij} \forall i, j$$  \hspace{1cm} (1)

In this paper, $T_{ij}$'s (the stochastic term) are interpreted as the expected international trade traffic flow (in terms of weight and value) carried by highway from external station $i$ to state $j$. The $A_i$'s are factors related to the origin zone $i$ and the $B_j$'s are destination related factors. The $F_{ij}$'s are factors which reflect the separation between $i$ and $j$. A common form which is general enough for most applications is:

$$F_{ij} = \exp\left[\sum_{k=1}^{K} \theta_{k} c_{ij}^{(k)}\right]$$  \hspace{1cm} (2)
This form is called an exponential form and $c_{ij}^{(k)}$ are different measures of separation, while $\theta_k$'s are parameters to be estimated. Potential measures of separation, include travel time, distance, generalized costs, etc.

**Maximum Likelihood Estimation**

The model (1) will be estimated using maximum likelihood. Maximum likelihood estimates have desirable asymptotic properties (consistency, efficiency and asymptotic normality), and are robust to distributional assumptions for realistic departures from the Poisson assumption (note that the multinomial distribution leads to identical estimates with the Poisson distribution) and are essentially unbiased even for a very small sample of flows (5).

Under some mild conditions (5) the ML estimate of $\zeta = (A(1), \ldots, A(I), B(1), \ldots, B(J), \theta_1, \ldots, \theta_q)'$ exist and that of $\theta$ is unique. The estimates of $A$ and $B$ are not unique; however, if one $A(i)$ or $B(j)$ is chosen to be an arbitrary positive number, the remaining $A(i)$'s and $B(j)$'s are unique under the previous mild conditions. It can be proved (5) that the ML estimate of $\zeta$ result as a solution to the following system of equations:

$$T_{is} = N_{is} \quad \forall i \in I$$

$$T_{sj} = N_{sj} \quad \forall j \in J$$

$$\sum_{ij} c_{ij}^{(q)} T_{ij} = \sum_{ij} c_{ij}^{(q)} N_{ij} \quad \forall q \in Q$$

where the operator $*$ indicates summation with respect to the subscript it replaces (e.g. $T_i = \sum_i T_{ij}$, $T_s = \sum_j T_{ij}$). A number of standard numerical methods or more specialized procedures can be used to solve equations (3)-(5) (see 5). The three procedures adapted for the paper are the Deming-Stefan-Furness (DSF) procedure, the linearized DSF (LDSF) procedure and the Modified Scoring procedure. An account of the development of the three procedures is given in (5). For completeness of the presentation, some of the details for each procedure are now discussed.

**The DSF procedure for Parameters $A_i$ and $B_j$**

The DSF procedure gives values of $T_{ij}$ for any choice of $(\theta')$, and $N_{is}, N_{sj}$, with $\sum_i N_{is} = \sum_j N_{sj}$. Generally speaking, the DSF procedure proceeds by adjusting the rows (columns) of a two-dimensional table in each even (odd) iteration. After choosing an initial value for the column balancing coefficient, $B_j^{(0)} = 1$, say, the DSF procedure iterates as follows:

$$A_i^{(2r-1)} = O_i \sum_{j=1}^{J} B_j^{(2r-2)} F_{ij} \quad \forall i$$

$$B_j^{(2r)} = D_j \sum_{i=1}^{I} A_i^{(2r-1)} F_{ij} \quad \forall j$$

where $O_i = T_{is}$, $D_j = T_{sj}$, and $F_{ij}$ is a function of the separation measures $c_{ij}^{(q)}$. 
Upon convergence (see (5) for a proof of convergence), the values of $T_{ij}$ are given by $T_{ij}^{(2r)} = A(i)^{(2r-1)}B(j)^{(2r)} F_{ij}$. In this paper we chose a fairly stringent criterion for convergence as follows:

$$
\sum_{i=1}^{I} |O_i - T_{ii}^{(2r)}| + \sum_{j=1}^{J} |D_j - T_{ij}^{(2r)}| < \delta
$$

(8)

where $\delta = 10^{-12}$. The algorithm attained this criterion in less than 100 iterations.

The DSF procedure essentially expresses $T_{ij}$ as a function of $\theta$. These values of $T_{ij}$ could then be used in (5) to solve for an updated value of $\theta$. In general, however, as $\theta$ changes, equations (3)–(4) could be violated, unless these changes are very small. This is achieved by using a linearized version of the DSF procedure, called the LDSF procedure (6), which is computationally very attractive.

**The LDSF Procedure for $T_{ij}$**

Let us assume that we have run the DSF procedure and obtained a good set of $T_{ij}$ that solves (3)–(4) for any given $\theta$. This means that $O_i = T_{ii} = N_{ii}$, and $D_j = T_{ij} = N_{ij}$.

Define by, $\Delta O = (\Delta O_1, \ldots, \Delta O_I)'$, and $\Delta D = (\Delta D_1, \ldots, \Delta D_J)'$ to be small changes in the values of $O = (O_1, \ldots, O_I)'$, and $D = (D_1, \ldots, D_J)'$, respectively. Let, also, $\Delta F_{ij}$, be a small change in, $F_{ij} = \exp(\theta c_{ij})$. It can be proved (5) that the corresponding small change, $\Delta T_{ij}$, in each $T_{ij}$, so that $\Delta T_{ii} = \Delta O_i$, and $\Delta T_{ij} = \Delta D_j$ can be obtained by the LDSF procedure, which iterates as follows:

$$
\Delta T_{ij}^{(2r-1)} = \Delta T_{ij}^{(2r-2)} + (T_{ij}/O_i)(\Delta O_i - \Delta T_{ii}^{(2r-2)})
$$

(9)

$$
\Delta T_{ij}^{(2r)} = \Delta T_{ij}^{(2r-1)} + (T_{ij}/D_j)(\Delta D_j - \Delta T_{ij}^{(2r-1)})
$$

(10)

for $i \in I$, $j \in J$ and $r = 0, 1, 2, \ldots$, with initial $T_{ij}$ values given by,

$$
\Delta T_{ij}^{(0)} = (T_{ij}/F_{ij})\Delta F_{ij}
$$

(11)

A proof for the convergence of the procedure is given in (6).

**Changes in $T_{ij}$ as a Function of a Change in $\theta$**

For a small change $\Delta \theta$ in $\theta$ and a small change $0$ so that $\Delta O = \Delta D = 0$, an approximation for the corresponding small change $\Delta T_{ij}$ for each $T_{ij}$. It can be proved (7) that for all $i \in I$ and $j \in J$,

$$
\Delta T_{ij} \approx \Delta \theta \{c_{ij}T_{ij} - \sum_{j} c_{ij}T_{ij}(T_{ij}/O_i) - \sum_{i} c_{ij}T_{ij}(T_{ij}/D_j)
$$

$$
+ \sum_{i} \sum_{j} c_{ij}T_{ij}(T_{ij}/O_i)(T_{ij}/D_j)\} = S_{ij} \Delta \theta
$$

(12)

The $S_{ij(k)}$'s are constants, and $O_i = T_{ii}$, $D_j = T_{ij}$. Therefore, if $T_{ij}$'s are known, the only unknown in (12) is the $\Delta \theta$. The solution for the $\Delta \theta$ will be the issue of the next section.
Estimation of \( \theta \) Using the Modified Scoring Procedure

So far, for an initial value for \( \theta \), we have obtained \( T_{ij}(\theta) \) by using the DSF procedure to solve (3)–(4). We have then changed \( \theta \) to \( \theta + \Delta \theta \) and computed, using the LDSF procedure, with \( \Delta \mathbf{O} = \Delta \mathbf{D} = 0 \), the corresponding change, \( \Delta T_{ij}(\theta, \Delta \theta) \)'s in \( T_{ij}(\theta) \)'s as a function of \( \Delta \theta \).

We are ready now to insert the \( [T_{ij}(\theta) + \Delta T_{ij}(\theta, \Delta \theta)] \)'s into the left hand side of (5) and solve the resultant equation for \( \Delta \theta \). Inserting \( T_{ij} + \Delta T_{ij} \) in place of \( T_{ij} \) in (5) and using (12), equations (3)–(4) would remain approximately satisfied, while obtaining the following system of \( Q \) linear equations with \( Q \) unknowns (the \( \Delta \theta_a \)'s):

\[
\sum_{ij} c_{ij}^{(1)} (\Delta T_{ij}) = \sum_{ij} c_{ij}^{(1)} (N_{ij} - T_{ij}) \\
\vdots \\
\sum_{ij} c_{ij}^{(Q)} (\Delta T_{ij}) = \sum_{ij} c_{ij}^{(Q)} (N_{ij} - T_{ij})
\]

(13)

This system of equations can be solved by any standard solution method such as Gaussian elimination.

The current solution for \( \theta \) at iteration \( r \) is updated next using the formula

\[
\theta^r = \theta^{r-1} + \Delta \theta^{r-1}
\]

(14)

If the corrections \( \Delta \theta^{r-1} \) have become negligible the values of \( \theta \) have been stabilized and the MS procedure terminates. Otherwise, new \( T_{ij} \)'s are obtained from the DSF procedure and the MS procedure continues. There is no guarantee that the MS procedure always converges (5) although our computational experience is positive.

Goodness of Fit

Under the previous assumption that observations \( N_{ij} \) are independently Poisson distributed, the (Pearson) \( \chi^2 \) statistic,

\[
\chi^2 = \sum_{ij} \frac{(N_{ij} - \hat{T}_{ij})^2}{\hat{T}_{ij}}
\]

(15)

where \( \hat{T}_{ij} \) is an estimate of \( T_{ij} \), is an appropriate measure of the overall fit of a model. Moreover, when \( \hat{T}_{ij} \) is obtained using maximum likelihood, (15) has a \( \chi^2 \) distribution with \( df = IJ - I - J - K + 1 \) degrees of freedom (8,9).

If \( \hat{T}_{ij} \approx T_{ij} \), then \( \chi^2 = Z^2 \), where

\[
Z^2 = \sum_{ij} \frac{(N_{ij} - T_{ij})^2}{T_{ij}}
\]

(16)
Since $E(N_{ij}) = T_{ij}$ and because $N_{ij}$ have the Poisson distribution,

$$\text{var}(N_{ij}) = E(N_{ij} - T_{ij})^2 = T_{ij}$$

(17)

Therefore, $E(Z^2) = IJ$, where $I$ is the number of origin zones $i$ and $J$ the number of destinations $j$. Equivalently, $E(Z^2/IJ) = 1$. Thus the so-called “$\chi^2$-ratio”, $\chi^2/df$, has an expectation which is asymptotically 1. It can be shown (5) that the variance of the $\chi^2$-ratio is,

$$\text{var}[Z^2/(IJ)] \approx \sum_{ij} [(T_{ij}I^2J^2)^{-1} + 2IJ^{-2}]$$

(18)

Hence, if $T_{ij}$’s are bounded away from zero (which is the case in exponential gravity models with finite parameters $\theta$), the variance of $Z^2/IJ \to 0$, as $IJ \to \infty$. It follows that when $T_{ij} \to T_{ij}$ and $T_{ij}$’s are bounded away from zero, the variance of $\chi^2/df \to 0$.

In practical applications, since the Poisson assumption seldom holds perfectly (as is the case here, where every pound or dollar value of shipment does not travel independently of each other pound or dollar value), a Chi-square ratio less than 2 is a good indication that the gravity model fits the data well (5).

**Covariance of Maximum Likelihood Estimates**

*Covariance of $\hat{\theta}_q$’s*

Let small case letters stand for the logarithms of corresponding capital letters (e.g., $t_{ij} = \log[T_{ij}], a(i) = \log[A(i)], b(j) = \log[B(j)]$. The model (1) may be written as:

$$t_{ij} = a(i) + b(j) + \sum_{q} \theta_q c_{ij}^{q} \quad \forall i \in I, j \in J, q \in Q$$

(19)

Let $M$ denote the coefficient matrix of the right side of the system of equations (19). The matrix $M$ is not of full rank (5). However, the matrix $M_{(2)}$ obtained by deleting one of the first $I + J$ columns of $M$ is of full rank and has dimension $IJ \times (I + J + Q - 1)$. Let $\text{diag}(\cdot)$ stands for a diagonal matrix, the diagonal elements of which are given within the parentheses. Then compute the matrix $M_{(2)}' \cdot \text{diag}(\mathbf{T}) \cdot M_{(2)}$ from the equation:

$$M_{(2)}' \cdot \text{diag}(\mathbf{T}) \cdot M_{(2)} = \begin{pmatrix} U_1 & U_2 \\ U_2' & U_3 \end{pmatrix}$$

(20)

where

$$U_1 = \begin{pmatrix} V_1 & V_2 \\ V_2 & V_3 \end{pmatrix}$$

(21)

$$U_2 = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$$

(22)
\[ U_3 = ((u_{pq}) \text{ with } u_{pq} = \sum_{ij} c_{ij}^{(p)} c_{ij}^{(q)} T_{ij}, \quad W_1 = ((w_{iq}^{(1)}) \text{ with } w_{iq}^{(1)} = \sum_j c_{ij}^{(q)} T_{ij}, \quad W_2 = ((w_{jq}^{(2)}) \text{ with } w_{jq}^{(2)} = \sum_i c_{ij}^{(q)} T_{ij}, \quad V_1 = \text{diag}(T_1, \ldots, T_{1s}), \quad V_2 = ((T_{ij})) \text{ and } V_3 = \text{diag}(T_{i1}, \ldots, T_{iJ-1}). \]

Notice that the subscript \( j \) in each of the matrices above goes only up to \( J - 1 \).

Matrix \( M'_2 \cdot \text{diag}(T) \cdot M_2 \), a square matrix of dimension \((I + J + Q - 1)\), is the covariance matrix of \( M'_2 \cdot \text{diag}(N) \). This is because the \( N_{ij} \)'s have independent Poisson distributions and the covariance matrix \( \text{Cov}(N) \) of \( N \) is \( \text{diag}(T) \). It can be shown (5) that the covariance matrix of 
\[ (\hat{A}(1), \ldots, \hat{A}(I), \hat{B}(1), \ldots, \hat{B}(J-1), \hat{\theta}_1, \ldots, \hat{\theta}_Q)' \]

is:
\[ \Phi^{-1} \cdot M'_2 \cdot \text{diag}(T) \cdot M_2 \cdot (\Phi^{-1})' \]  
(23)

where
\[ \Phi = M'_2 \cdot \text{diag}(T) \cdot M_2 \cdot \\
\text{diag}(1/A(1), \ldots, 1/A(I), 1/B(1), \ldots, 1/B(J-1), 1, \ldots, 1) \]  
(24)

and
\[ \Phi^{-1} = \text{diag}(1/A(1), \ldots, 1/A(I), 1/B(1), \ldots, 1/B(J-1), 1, \ldots, 1) \cdot \\
(M'_2 \cdot \text{diag}(T) \cdot M_2)^{-1} \]  
(25)

Notice that using (23), (25) another expression for the covariance matrix of 
\[ (\hat{A}(1), \ldots, \hat{A}(I), \hat{B}(1), \ldots, \hat{B}(J-1), \hat{\theta}_1, \ldots, \hat{\theta}_Q)' \]

can be written by as follows:
\[ \text{diag}(1/A(1), \ldots, 1/A(I), 1/B(1), \ldots, 1/B(J-1), 1, \ldots, 1) \cdot \\
(M'_2 \cdot \text{diag}(T) \cdot M_2)^{-1} \cdot \\
\text{diag}(1/A(1), \ldots, 1/A(I), 1/B(1), \ldots, 1/B(J-1), 1, \ldots, 1) \]  
(26)

The bottom right \( Q \times Q \) submatrix of matrix (26) is the estimated covariance matrix of \( \hat{\theta} \). The bottom right \( Q \times Q \) submatrix of the inverse of (20) is (9)
\[ (U_3 - U_2^2 U_1^{-1} U_2)^{-1} \]  
(27)

From (26), it follows that (27) is the covariance matrix of \( \hat{\theta} \).

\textit{Covariance of } \( T_{ij} \)'s

Having obtained \( \text{Cov}(A, B, \theta) \) from (23) or using (26), and since \( B(J) \) is set equal to a constant and its variance and covariances involving it are zeros, it can readily be seen that the covariance matrix of \( \hat{T} \), denoted by the symbol \( \text{Cov}(\hat{T}) \), is
\[ \text{Cov}(\hat{T}) = \Psi_1 \cdot \text{Cov}(A, B, \theta) \cdot \Psi'_1 \]  
(28)

with
\[ \Psi_1 = \text{diag}(T) \cdot M \cdot \\
\text{diag}(1/A(1), \ldots, 1/A(I), 1/B(1), \ldots, 1/B(J), 1, \ldots, 1) \]  
(29)
and $M$ the coefficient matrix of the right side of the system of equations (19).

**EMPIRICAL ANALYSIS**

**Freight Shipments**

The Oak Ridge National Laboratory (ORNL) made available two sets of origin-destination flow data for freight shipment weight and value between origin seaports/border ports of entry $i$ and destination states $j$ from the following sources: (a) *The Transborder Surface Freight Database*. This database (see http://www.bts.gov/transborder/prod.html) is distributed by the Bureau of Transportation Statistics (BTS) and contains freight flow data by commodity type and by surface mode of transportation (rail, truck, pipeline, or mail) for U.S. exports to and imports from Canada and Mexico. (b) *The Port Import Export Reporting Service (PIERS) database*. This is a commercial database (see http://www.piers.com/about/default.asp) prepared by the 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. Two matrices, $N_{ij}$, were developed: one for shipment weight and one for shipment value for each of the two databases with dimensions $144\times50$ and $128\times50$ for the PIERS and Transborder database, respectively.

**Separation Measures**

ORNL made available two sets with two separate impedance matrices, $c_{ic}^{(k)}$, $k = 1, 2$ from origin point $i$ to destination county $c$. The first set included impedances between origin seaports and destination counties; the second set between origin border ports and destination counties. Both measures were computed from the National Highway Network (NHN). The first of the two measures is the route (network) distance in miles; the second computes a function of travel time for different functional classifications of highway segments and adds these time penalties together while tracing the previous network routes. The previous impedance matrices presented two problems: (a) more than 50% of the cells were empty since not all seaports/border ports of entry are connected to each county through the NHN; (b) occasionally, freight shipments were realized between origin-destination pairs with a missing connection (separation measure).

In order to deal with both problems, we devised and implemented the following stepwise procedure for the distance matrix (the first separation measure, $c_{ic}^{(1)}$).

1. County centroid coordinates were computed for all counties nationwide.

2. Spherical distances in miles from each county to every other county (ignoring the earth’s flatness was not considered more onerous than estimating airline distances in lieu of network distances).

3. For every destination county with a missing connection to an origin point (seaport/border port of entry), the closest county with an existing connection to an
origin point was estimated based on the previous county-to-county distances.

4. The missing connection from an origin point to a destination county was finally computed to be the distance between the same point of origin and the closest destination county augmented by 130% the (airline) distance between the two destination counties (as a proxy to the actual road distance). This simplification was made under the assumption that a missing connection would imply that the destination county is off the NHN and thus would require additional time to be accessed from its closest county on the NHN.

The procedure above resulted in the construction of a synthesized separation measure comprised of both route distance (ORNL estimate) and multiples of airline distance (our estimate). It is important to take a closer look at those two components of the (re-estimated) first separation measure $r_{k1}$.

The airline distance between origin point $i$ and destination county $c$, even augmented by 30%, is only an approximation of the actual miles traveled and it only represents a surrogate for the complex set of factors that express the difficulty of overcoming separation. On the other hand, the route distance for the same origin-destination pair is the total over-the-road distance on a realistic route. As a measure of separation, the route distance would provide better accuracy in the eastern half of the limited access highway system of the United States that has a higher level of complexity resulting in less circuitous routes than in the western half. Note that the route distance is usually the shortest route distance obtained by an application of a route assignment model.

Whereas the distance measure developed above is a measure that depends more or less on the physical characteristics of the road link, the impedance measure, a function of link travel time as estimated by ORNL, depends on the special roadway type or certain conditions encountered on the link. ORNL’s estimating procedure could not be replicated in this paper for the missing impedances. A surrogate value was computed based on the average speed obtained by ORNL’s distance and time estimates, and the synthesized distances.

The procedure classified the ORNL’s distance and time estimates into ten deciles. The stratum average speed was computed by the ratio of the average distance and average time for that stratum as seen in Table 1. Each missing time impedance was then estimated by the ratio of its corresponding distance estimate and the average speed of the stratum the distance estimate belongs to.

Clearly, this procedure imposes some inconsistency in the imputation of travel times because the average speed estimate is based on decile groupings of the initial estimates. The consistency of the procedure could be improved if in a second iteration the decile groupings are based on all distances (initial and imputed), but we did not attempt this here.
The previous steps resulted in two re-estimated sets of separation measures: (a) a distance matrix and a travel time matrix between origin seaports and destination counties; and (b) a distance matrix and a travel time matrix between origin points of entry and destination counties. The final step in preparing the separation measures for model estimation was to estimate these measures between origin points and destination states rather than counties. This was made possible simply by computing for each origin point \( i \) and all destination counties \( c \in C_j \) in state \( j \) the average distance and travel time. That is, \( \forall i \in I, \)
\[
c_{ij}^{(1)} = \frac{\sum c_{k,i}}{n}, \quad \text{and} \quad c_{ij}^{(2)} = \frac{\sum c_{n,i}}{n}, \quad c_n \in C_j,
\]
where, \( n \) is the cardinality of \( C_j \). Note that had we avoided the intermediate step of origin-to-destination county distance estimation, say, by computing distances to state geographical centers, origin-destination pairs with both ends in the same state would have been indistinguishable in terms of their separation.

**Results**

The only decision necessary to apply the procedures described above is the choice of the flow unit. In the case of 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 or dollars. This is not surprising given that the bulk of flows are long-distance shipments that are usually performed by large (80,000 pound) trucks or rail (on a limited scale for the particular data) with an average shipment value of $1.39 per pound for the PIERS data and $2.03 per pound for the Transborder data.

The procedures described above were run to a tight convergence. For each iteration of the modified scoring procedure, the DSF procedure attained a \( 10E - 12 \) convergence as defined by (8) in less than 100 iterations. The modified scoring procedure itself attained a \( 10E - 06 \) convergence as defined by the right-hand side of (13) in less than 20 iterations.

**Parameter Estimates and Model Fit**

Separate gravity models were estimated for both seaport-to-state (seaport data) and border port-to-state (transborder) weight and value flows. In particular, for the seaport data, several model specifications were tested based on transformations of the distance and time separation measures, \( c_{ij}^{(1)} \) and \( c_{ij}^{(2)} \), respectively, after a careful examination of residuals that showed the presence of outliers.

A consequence of the Poisson distribution is the unequal variances of residuals. As a result, instead of the plain residuals, \( N_{ij} - T_{ij} \), we examined the components

\[
\sqrt{N_{ij}} - \sqrt{T_{ij}}
\]

The investigation signaled the need for transformation of the \( c_{ij}^{(k)} \)'s. We found the square
root transformation to provide adequate fit as seen in Table 2 for both weight flow and value flow estimates. The model provides an excellent fit for the data, as the Chi-square statistic hovers around its expected value of 1. In addition, the cell-to-cell weight flow estimates correlate very well with the data (Pearson correlation coefficient, \( r = 0.89, p < 0.01 \)). The cell-to-cell value flow estimates also correlate very well with the data (\( r = 0.98, p < 0.01 \)). Moreover, the weight and value length distributions in Figures 1 (1% corresponds to 2,292,872,822 pounds) and 2 (1% corresponds to 3,198,123,011 dollars), respectively, seem to corroborate the previous results.

Of interest in Table 1 is the appearance of \( \hat{\theta}_1 \) with a positive value. This 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 above in collinearity situations, all available impedance measures were retained. After all, the sign of \( \theta_1 \) 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). The proposed parameterization was deemed adequate for the intent of the paper.

In the end, the model specifications for the weight and value flows, respectively, are:

\[
T_{ij} = A_i B_j \exp(\theta_1 \sqrt{c_{ij}^{(1)}} + \theta_2 \sqrt{c_{ij}^{(2)}})
\]

\[
T_{ij} = A_i B_j \exp(\theta_1 c_{ij}^{(1)} + \theta_2 \sqrt{c_{ij}^{(2)}})
\]

(31)

In the case of transborder data, we conducted a similar investigation and observed the presence of a few remaining outliers, despite using transformed \( c_{ij}^{(k)} \)'s. These outliers were mainly states receiving unusually large share of shipments compared to other states. The discovery was made after squaring the residuals in (30) and adding them together for all origin points of entry for each destination state. Destination states with large share of shipments had much larger sum of squares of residuals than other destinations. This was understandable since these large destination states attract more shipments from larger distances, and suggested the use of an additional impedance variable \( c_{ij}^{(3)} = \delta \sqrt{c_{ij}^{(2)}} \) for the weight flows, and \( c_{ij}^{(3)} = \delta \log(c_{ij}^{(2)}) \) for the value flows, where \( \delta \) is 1 for destinations with more than 10% share in weight (4% in value) and 0 otherwise. In effect, we sought a different parameter estimate for the square root/log of distance for destination states with large shares.

The previous use of the indicator (dummy) variable \( \delta \) is typical in spatial analysis for the treatment of residuals (5) and introduces the effects of spatial structure on flow patterns (10, 11, 12) which is not the central intent in this paper. More complex measures of relative location have been developed and tested empirically (13, 14). In the end, the
model specifications for the weight and value flows, respectively, are:

\[
T_{ij} = A_i B_j \exp(\theta_1 \sqrt{c_{ij}^{(1)} + \theta_2 \sqrt{c_{ij}^{(2)}} + \theta_3 \delta \sqrt{c_{ij}^{(2)}}})
\]

\[
T_{ij} = A_i B_j \exp(\theta_1 c_{ij}^{(1)} + \theta_2 c_{ij}^{(2)} + \theta_3 \delta \log|c_{ij}^{(2)}|)
\] (32)

The previous steps removed all remaining outliers. Parameter estimates and goodness-of-fit statistics are shown in Table 3. The previous observation regarding the sign of \(\theta_1\) applies.

The model fits the data very well as the Chi-square ratios for both weight and value remain under 2. In addition, the cell-to-cell weight flow estimates correlate very well with the data (Pearson correlation coefficient, \(r = 0.91, p < 0.01\)). The cell-to-cell value flow estimates also correlate very well with the data \((r = 0.94, p < 0.01)\). Moreover, the weight and value length distributions in Figures 3 (1% corresponds to 767,519,815 pounds) and 4 (1% corresponds to 1,565,313,371 dollars), respectively, seem to corroborate the previous results.

**Computation of Covariance of Estimates**

The previous point estimates of the \( \hat{\theta} \) parameters and the flows \( \hat{T} \) were used in the methodology described earlier to obtain the covariance of these estimates. The procedure can accommodate any reasonable number of separation measures, \( c_{ij}^{(k)} \), and the large number of origins and destinations typically encountered in practice.

The computational requirements of the procedure are no longer prohibitive. We were able to obtain the covariance matrix of a 144x50 flow matrix, a 7200x7200 matrix, in a little more than an hour on a Pentium III, 800 MHz, 512MB RAM laptop computer running a FORTRAN 77 compiler.

Replacing \( T_{ij} \)'s by their estimates in (27), the covariance matrix of \( \hat{\theta} \) for the weight flows from the PIERS data set was found to be

\[
\begin{pmatrix}
\hat{\theta}_1 \\
2.62830473 & -2.81464286 \\
\hat{\theta}_2 \\
-2.81464292 & 3.02464367
\end{pmatrix}
\] (33)

The correlation between \( \hat{\theta}_1 \) (the parameter estimate for the distance measure) and \( \hat{\theta}_2 \) (the parameter estimate for the travel time measure) is readily apparent from (33). The negative sign of the covariances should not be disconcerting. It shows that, if for some small shift in the observations, \( \theta_1 \) were to increase, \( \theta_2 \) would decrease to 'compensate'.

Similarly, the covariance matrix of \( \hat{\theta} \) for the value flows from the PIERS data set was found to be

\[
\begin{pmatrix}
\hat{\theta}_1 \\
1.53388134 & -1.66648732 \\
\hat{\theta}_2 \\
-1.66648735 & 1.81284692
\end{pmatrix}
\] (34)

The correlation between \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) is similarly apparent from (34). Moreover, the covariance matrices of \( \hat{\theta} \) from the Transborder data set for the weight and value flows,
respectively, were found to be

\[
\begin{bmatrix}
\hat{\theta}_1 & \hat{\theta}_2 & \hat{\theta}_3 \\
9.26053517 & -9.88856511 & 0.08654230 \\
-9.88856442 & 10.60290046 & -0.15624519 \\
0.08654113 & -0.15624389 & 0.67484586
\end{bmatrix}
\]

(35)

and

\[
\begin{bmatrix}
\hat{\theta}_1 & \hat{\theta}_2 & \hat{\theta}_3 \\
2.87729568 & -3.13541449 & 0.01302693 \\
-3.13541410 & 3.42693579 & -0.02059522 \\
0.01302676 & -0.02059503 & 0.01040623
\end{bmatrix}
\]

(36)

These last two covariance matrices in (35) and (36) show relatively high correlation between \( \hat{\theta}_1 \) (the parameter estimate for the distance measure) and \( \hat{\theta}_2 \) (the parameter estimate for the travel time measure) and relatively low correlation between \( \hat{\theta}_3 \) (the parameter estimate for the transformed travel time measure) and \( \hat{\theta}_1 \) or \( \hat{\theta}_2 \). The interpretation of the negative signs is as above.

For illustration purposes, we present an example of the usefulness of the computed covariance matrices for \( T_{ij} \)'s in computing confidence intervals in Table 4. The table shows a few \( N_{ij} \)'s, \( T_{ij} \)'s, percent difference between \( N_{ij} \) and \( T_{ij} \), and 90 percent confidence intervals constructed by adding and subtracting 1.65 times the standard error available from the variances.

CONCLUSIONS

The movement of freight shipments can now be estimated within a desired confidence level as a result of maximum likelihood estimation of Poisson gravity models. The freight transportation modeler has now two procedures available for computing reliable information within a predetermined accuracy: one for computing freight flow estimates and one for computing covariance matrices. These procedures can accommodate any reasonable number of separation measures, \( c_{ij}^{(k)} \), and the large number of origins and destinations typically encountered in practice.

REFERENCES


List of Tables

1  Average distance, time and speed by decile for PIERS and Transborder databases ........................................... 15
2  Parameter estimates for PIERS data ............................................... 16
3  Parameter estimates for Transborder data ..................................... 17
4  An illustration of $N_{ij}, \hat{T}_{ij}$ and 90% confidence intervals .............. 18

List of Figures

1  PIERS data: Weight length frequency distribution. ......................... 19
2  PIERS data: Value length frequency distribution. .......................... 20
3  Transborder data: Weight length frequency distribution. .................. 21
4  Transborder data: Value length frequency distribution. ..................... 22
<table>
<thead>
<tr>
<th>Decile</th>
<th>Average Distance (miles)</th>
<th>Average Time (min.)</th>
<th>Estimated Speed (m/h)</th>
<th>Average Distance (miles)</th>
<th>Average Time (min.)</th>
<th>Estimated Speed (m/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>230.69</td>
<td>241.56</td>
<td>57.30</td>
<td>381.11</td>
<td>400.32</td>
<td>57.12</td>
</tr>
<tr>
<td>2</td>
<td>469.65</td>
<td>465.70</td>
<td>60.38</td>
<td>708.18</td>
<td>711.44</td>
<td>59.73</td>
</tr>
<tr>
<td>3</td>
<td>642.53</td>
<td>625.89</td>
<td>61.59</td>
<td>929.82</td>
<td>920.37</td>
<td>60.61</td>
</tr>
<tr>
<td>4</td>
<td>797.46</td>
<td>769.51</td>
<td>62.17</td>
<td>1121.06</td>
<td>1093.26</td>
<td>61.52</td>
</tr>
<tr>
<td>5</td>
<td>954.44</td>
<td>915.38</td>
<td>62.56</td>
<td>1291.37</td>
<td>1251.02</td>
<td>61.93</td>
</tr>
<tr>
<td>6</td>
<td>1127.45</td>
<td>1077.43</td>
<td>62.78</td>
<td>1457.91</td>
<td>1405.62</td>
<td>62.23</td>
</tr>
<tr>
<td>7</td>
<td>1324.50</td>
<td>1261.04</td>
<td>63.01</td>
<td>1632.94</td>
<td>1569.44</td>
<td>62.42</td>
</tr>
<tr>
<td>8</td>
<td>1596.70</td>
<td>1512.70</td>
<td>63.33</td>
<td>1850.64</td>
<td>1771.00</td>
<td>62.69</td>
</tr>
<tr>
<td>9</td>
<td>2109.35</td>
<td>1979.77</td>
<td>63.92</td>
<td>2166.76</td>
<td>2048.56</td>
<td>63.46</td>
</tr>
<tr>
<td>10</td>
<td>2958.11</td>
<td>2963.42</td>
<td>59.89</td>
<td>2757.87</td>
<td>2641.23</td>
<td>62.65</td>
</tr>
<tr>
<td>Parameter</td>
<td>Weight Flows</td>
<td>Value Flows</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sample Mean</td>
<td>Standard Deviation</td>
<td>Sample Mean</td>
<td>Standard Deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>11.1227</td>
<td>1.1646</td>
<td>3.7854</td>
<td>0.3963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-16.7447</td>
<td>1.7533</td>
<td>-11.1437</td>
<td>1.1668</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$-ratio</td>
<td>0.7253</td>
<td>0.0000</td>
<td>1.1760</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TABLE 3 Parameter estimates for Transborder data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Weight Flows</th>
<th>Value Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>14.8743</td>
<td>1.5575</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>-21.4296</td>
<td>2.2439</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>-2.1181</td>
<td>0.2217</td>
</tr>
<tr>
<td>$\chi^2$-ratio</td>
<td>1.9642</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
TABLE 4 An illustration of $N_{ij}$, $T_{ij}$ and 90% confidence intervals

<table>
<thead>
<tr>
<th>$N_{ij}$</th>
<th>$T_{ij}$</th>
<th>$N_{ij} - T_{ij}$ % Diff.</th>
<th>90% Confidence Interval of $T_{ij}$ Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>4850905</td>
<td>4764382.12</td>
<td>1.78</td>
<td>4764179.19</td>
<td>4764585.05</td>
</tr>
<tr>
<td>87454</td>
<td>88012.62</td>
<td>0.63</td>
<td>87934.47</td>
<td>88090.77</td>
</tr>
<tr>
<td>39922216</td>
<td>38847527.51</td>
<td>2.69</td>
<td>38845275.15</td>
<td>38849779.87</td>
</tr>
<tr>
<td>16002283960</td>
<td>15556162969.96</td>
<td>2.78</td>
<td>15555972990.43</td>
<td>1556352949.49</td>
</tr>
<tr>
<td>216787500</td>
<td>211385682.81</td>
<td>2.49</td>
<td>211374117.32</td>
<td>211397248.30</td>
</tr>
<tr>
<td>4174427</td>
<td>4266148.17</td>
<td>2.19</td>
<td>4266017.84</td>
<td>4266278.50</td>
</tr>
<tr>
<td>3076999</td>
<td>3100768.36</td>
<td>0.77</td>
<td>3100346.82</td>
<td>3101189.90</td>
</tr>
<tr>
<td>2029122669</td>
<td>2108940894.95</td>
<td>3.93</td>
<td>2108896284.86</td>
<td>2108985505.04</td>
</tr>
<tr>
<td>693089</td>
<td>664724.67</td>
<td>4.09</td>
<td>664638.45</td>
<td>664810.89</td>
</tr>
</tbody>
</table>
Figure 1: PIERS data: Weight length frequency distribution.
Figure 2: PIERS data: Value length frequency distribution.
Figure 3: Transborder data: Weight length frequency distribution.
Figure 4: Transborder data: Value length frequency distribution.