*JOURNAL OF REGIONAL SCIENCE, VOL. 45, No. 3, 2005, pp. 539–563*

# A FLEXIBLE MATHEMATICAL PROGRAMMING MODEL TO ESTIMATE INTERREGIONAL INPUT–OUTPUT ACCOUNTS*

**Patrick Canning**

*Economic Research Service, US Department of Agriculture, 1800 M street, NW, Washington, D.C. 20036. E-mail: pcanning@ers.usda.gov*

**Zhi Wang**

*US Department of Commerce, Bureau Of Economic Analysis, 1441 L Street, N.W. (BE-57), Washington, DC 20230. E-mail: zhi.wang@bea.gov*

**ABSTRACT.** This study implements and tests a mathematical programming model to estimate interregional, interindustry transaction flows in a national system of economic regions based on an interregional accounting framework and initial information of interregional shipments. A national input–output (IO) table, regional data on gross output, value-added, exports, imports, and final demand at sector level are used as inputs to generate an interregional IO account that reconciles regional economic statistics and interregional transaction data. The model is tested using data from a multi-regional global IO database and shows remarkable capacity to discover true interregional trade patterns from highly distorted initial estimates.

## 1. INTRODUCTION

A major obstacle in regional economic analysis and empirical economic geography is the lack of consistent, reliable regional data, especially data on interregional trade and interindustrial transactions. Despite efforts by regional economists, data analogous to national input–output (IO) accounts and international trade accounts, which have become increasingly available to the public today, still are generally not available even for well-defined subnational regions in many developed countries. Therefore, economists had to develop various non-survey and semi-survey methods to estimate such data. In the earlier years, quotient-based and regional purchase coefficient-based non-survey methods were popular but lacked logical and theoretical structures, and hence have been deemed 'deficiency methods' (Jensen, 1990).

Since the 1980s, various constrained matrix-balancing procedures have become increasingly popular for estimating unknown data based on limited initial information subject to a set of linear constraints. Attempts have been

---

*The authors thank the editor of this Journal and three anonymous referees for their valuable comments. However, the views expressed in this paper are those of the authors and do not represent the opinions of the institutions with which they are affiliated.

Received: October 2003; Revised: May 2004; Accepted: August 2004.



Blackwell Publishing, Inc., 350 Main Street, Malden, MA 02148, USA and 9600 Garsington Road, Oxford, OX4 2DQ, UK.

made to estimate regional and interregional transactions in a unified national accounting system of economic regions. Batten (1982) extended the earlier work by Wilson (1970) and laid out an optimization model based on information theory and linkages between national and regional IO accounts to simultaneously estimate interregional deliveries in both intermediate and final goods.[1] Batten and Martellato (1985) established a simple hierarchical relationship among five classical models associated with authors such as Isard, Chenery' and Leontief who address interregional trade within an IO system. They found those models could be reduced to a statistical estimation problem based on varying degrees of available interregional trade data and demonstrated that the net effect of additional data or additional theoretical assumptions is similar in reducing the number of unknown variables in the underdetermined estimation problems. They also demonstrated such estimation problems are best undertaken with a closed system, i.e., when all the geographic components of the national or state data are estimated simultaneously. Following this philosophy, Byron et al. (1993), Boomsma and Oosterhaven (1992), and Trendle (1999) have found evidence that the additional accounting constraints imposed by such a closed system are useful as a checking device on individual cell values and hence improve estimation accuracy. Golan, Judge, and Robinson (1994) further generalize such an estimation problem to an ill-posed, underdetermined, pure inverse problem that can be formulated in an optimization context that involves a non-linear criterion function and certain adding up and consistency constraints. They also show that under such a framework, it is easy to take account of whatever initial information and data that exist through the specification of additional constraints. However, they did not pay attention to how such procedures could be used in a multiregional context and thus the potential gain from implementing the procedure in a complete national system of economic regions.

Methods for matrix balancing can be classified into two broad classes—biproportional scaling and mathematical programming. The scaling methods are based on the adjustments of the initial matrix to multiplying its row and column by positive constants until the matrix is balanced. It was developed by Stone and other members of the Cambridge Growth Project (Stone, Bates, and Bacharach, 1963) and is usually known as RAS. The basic method was originally applied to known row and column totals but has been extended to cases where the totals themselves are not known with certainty (Senesen and Bates, 1988; Lahr, 2001). Mathematical programming methods are explicitly based on a constrained optimization framework, usually minimizing a penalty function, which measures the deviation of the balanced matrix from the initial matrix subject to a set of balance conditions.

[1]Wilson (1970) had suggested an entropy maximizing solution for a model which integrated gravity models and multiregional IO equations as constraints to estimate interregional commodity flows. However, his work did not clearly incorporate a complete system of national and regional IO accounts as in Batten (1982).

The scaling methods such as RAS have been one of the most widely applied computational algorithms for the solution of constrained matrix-balancing problems. They are simple, iterative, and require minimal programming effort to implement. However, as pointed out by Ploeg (1982), they are not straightforward to use when including more general linear restrictions and when allowing for different degrees of uncertainty in the initial estimates and restraints. They also lack a theoretical interpretation of the adjustment process.[2] Those aspects are crucial for an adjustment procedure to improve the information content of the balanced estimates rather than only adjusting the initial estimates mechanically. Mohr, Crown, and Polenske (1987) discussed the problems encountered when the RAS procedure is used to adjust trade flow data. They pointed out that the special properties of interregional trade data increase the likelihood of non-convergence of the RAS procedure and proposed a linear programming approach that incorporates exogenous information to override the unfeasibility of the RAS problem.

In recent years, more researchers have tended to formulate constrained matrix-balancing problems as mathematical programming problems (Ploeg, 1988; Nagurney and Robinson, 1989; Bartholdy, 1991; Byron et al., 1993), with an objective function that forces "conservatism" on the process of rationalizing $\mathbf{X}$ from the initial estimate $\bar{\mathbf{X}}$. The theoretical foundation for the approach can be viewed from both the perspectives of mathematical statistics and information theory, and the solution of RAS is equivalent to constrained entropy minimization with fixed row and column totals, as shown by Bregman (1967) and McDougall (1999), thus can been seen as a special case of the optimization methods.[3]

Another important advantage of mathematical programming models over scaling methods is in its flexibility. It allows a wide range of initial information to be used efficiently in the data-adjustment process. Additional constraints can be easily imposed, such as allowing precise upper and lower bounds to be placed on unknown elements, inequality conditions, or incorporating an associated term in the objective function to penalize solution deviations from the initial row or column total estimates when they are not known with certainty. Therefore, it provides more flexibility to the matrix-balancing procedure. This flexibility is very important in terms of improving the information content of the balanced estimates as shown by Robinson, Cattaneo, and El-said (2001).

[2]In a recently published special issue of *Economic System Research* (Vol. 16, No. 2 June 2004) on biproportional scaling techniques and its recent extensions, Lahr and Mesnard (2004) and many others demonstrate that while some researchers have been developing alternative approaches such as mathematical programming to overcome one or more of RAS apparent shortcomings, others have been extending RAS to including those new properties, such as incorporating the reliability of initial data and other known information into RAS.

[3]Using Monte-Carlo simulation, Robinson, Cattaneo, and El-said (2001) show that when updating column coefficients of a social accounting matrix (SAM) is the major concern, the cross-entropy method appears superior, while if the focus is on the flows in the SAM, then the two methods are very close with the RAS performing slightly better.

A mathematical programming approach also permits one to routinely introduce relative degrees of reliability for initial estimates. The idea of including data reliability in matrix balancing can be traced back over half a century to Stone, Champernowne, and Meade (1942) when they explored procedures for compiling national income accounts. Their ideas were formalized into a mathematical procedure to balance the system of accounts after assigning reliability weights to each entry in the system. The minimization of the sum of squares of the adjustments between initial and balanced entries in the system, weighted by the reliabilities or the reciprocal of the variances of the entries, is carried out subject to linear (accounting) constraints. This approach had first been operationlized by Byron (1978) and applied to the System of National Accounts of the UK by Ploeg (1982, 1984). Zenios, Drud, and Mulvey (1989) further extended this approach to balance a large social accounting matrix in a non-linear network-programming framework. Robinson, Cattaneo, and El-said (2001) provided a way to handle measurement error in cross-entropy minimization via an error-in-variables formulation. Although computational burden is no longer a problem today, the difficulty of estimating the error variances in a large data set by such approaches still remains unsolved.

The objectives of this paper are threefold: (i) to develop and implement a formal model to estimate interregional, interindustry transaction flows in a national system of economic regions based on incomplete statistical information at the regional level; (ii) to evaluate the model's performance against data from the real world, and (iii) finally to discuss the issues when applying this modeling framework to estimate a multiregional IO account containing well-defined subregions.

The paper is organized as follows. Section 2 specifies the modeling framework and discusses its theoretical and empirical properties. Section 3 tests the model by using a four-region, ten-sector data set compiled from a global database documented in McDougall, Elbehri, and Truong (1998). Test results from seven experiments are evaluated against eight mean absolute percentage error indexes. Section 4 discusses some empirical issues involved in applying such a framework to data from a national statistical system. The paper ends with conclusions and direction for future research.

## 2. MATHEMATICAL PROGRAMMING MODELS FOR ESTIMATING INTERREGIONAL TRADE AND INTERINDUSTRIAL TRANSACTION FLOWS

Our model builds upon earlier work by Wilson (1970) and Batten (1982) with two important departures. First, it explicitly incorporates interregional trade flow information into both the accounting framework and initial estimates of an Interregional Input–Output (IRIO) account. We find this greatly enhances the accuracy of estimation results. Second, the IRIO account is simplified to a multiregional input–output (MRIO) account and estimated first, which substantially reduces the possibility of introducing spurious information in lieu of survey data and also diminishes the "dimension explosion" problem in real world applications.

*General Assumptions and Mathematical Notations*

Consider a national economy consisting of $N$ sectors that are distributed over $G$ geographic regions. The sectors use each other's products as inputs for their own production, which is in turn used up either in further production or by final users. Each region exports some of its products to other regions and some to other nations. They also import products from other regions and nations to meet their intermediate and final demand. Assuming a predetermined location of production that defines the structure of the national economic system of regions, the deliveries of goods and services between regions are determined by imbalances between supply and demand inside the different regions.

In this economy, a comprehensive account of annual product and payment flows within and between regions is summarized by an IRIO table. The notation used to describe the elements of a commodity-based IRIO table and its relationship to both a national IO table and a MRIO table are as follows (expressed in annual values):

Regional gross output, final demands, value-added, and international trade

$x_i^r$ = Gross output of commodity "$i$" in region "$r$"

$v_i^r$ = Value added by production of commodity "$i$" in region "$r$"

$y_i^r$ = Final demand (excluding exports) of commodity "$i$" in region "$r$"

$e_i^r$ = Exports of commodity "$i$" from region "$r$" to international market

$m_i^r$ = Imports of commodity "$i$" to region "$r$" from international market

Interregional and international deliveries

$z_{ij}^{sr}$ = Deliveries of domestic commodity "$i$" produced in region "$s$" for use by sector "$j$" in region "$r$"

$y_{ik}^{sr}$ = Deliveries of domestic commodity "$i$" produced in region "$s$" for type "$k$" final use in region "$r$"

$m_{ij}^r$ = International imports of commodity "$i$" for use by sector "$j$" in region "$r$"

$m_{ik}^r$ = International imports of commodity "$i$" for type "$k$" final use in region "$r$"

National IO table

$x_i$ = Gross domestic output of commodity "$i$"

$y_i$ = Final domestic demand (excluding exports) of commodity "$i$"

$e_i$ = International exports of commodity "$i$" from domestic origins of movement

$v_i$ = Value added by domestic production of commodity "$i$"

$z_{ij}$ = Intermediate demand of commodity "$i$" by sector "$j$"

$m_i$ = Imports of commodity "$i$" from international origins of movement

Aggregation variables for linkage to an MRIO table

$d_i^{sr}$ = Deliveries of domestic commodity "$i$" from region "$s$" to region "$r$"

$z_{ij}^{\bullet r}$ = Intermediate demand of commodity "$i$" by sector "$j$" in region "$r$"

*IRIO Account and Estimation Model*

Using notations defined above, the following two accounting identities describe the relationship among elements of each row $(i,r)$ and column $(j,s)$ of the IRIO table for a static national system of economic regions:

(1) $$\sum_{s=1}^{g}\sum_{j=1}^{n} z_{ij}^{rs} + \sum_{s=1}^{g}\sum_{k=1}^{h} y_{ik}^{rs} + e_i^r = x_i^r$$

(2) $$\sum_{r=1}^{g}\sum_{i=1}^{n} z_{ij}^{rs} + \sum_{i=1}^{n} m_{ij}^{s} + v_j^s = x_j^s$$

At each given year, Equations (1) and (2) must hold for all $i,j \in N$, $k \in H$, and $s,r \in G$. In addition, this IRIO account has to be consistent with a national IO account and related regional economic statistics, which requires the following accounting identities also to be satisfied each year:

(3) $$\sum_{h=1}^{k}\sum_{s=1}^{g} y_{ih}^{sr} + \sum_{k=1}^{h} m_{ik}^{r} = y_i^r$$

(4) $$\sum_{j=1}^{n} m_{ij}^{r} + \sum_{k=1}^{h} m_{ik}^{r} = m_i^r$$

(5) $$\sum_{r=1}^{g}\left(\sum_{s=1}^{g} z_{ij}^{sr} + m_{ij}^{r}\right) = z_{ij}$$

(6) $$\sum_{r=1}^{g} x_i^r = x_i$$

(7) $$\sum_{r=1}^{g} v_i^r = v_i$$

(8) $$\sum_{r=1}^{g} y_i^r = y_i$$

(9) $$\sum_{r=1}^{g} e_i^r = e_i$$

(10) $$\sum_{r=1}^{g} m_i^r = m_i$$

Collectively, Equations (1)–(10) define a commodity-based IRIO account within a national system of regions. The economic meanings for each of the 10 equations are straightforward. Equation (1) shows that total gross output of commodity "$i$" in region "$r$" is delivered to domestic intermediate and final users in all regions (including itself) within the nation and what is not delivered to domestic users is exported to international market. Equation (2) defines the value of gross output for commodity "$j$" in region "$s$" as the sum of value from all of its intermediate (domestic plus imported) and primary factor inputs. Equation (3) indicates that each region's total final demand for commodity "$i$" must be met by final goods and services delivered from all regions within the nation plus imports from other nations, while Equation (4) states each region's foreign imports of intermediate and final goods and services have to equal the region's total imports from international markets. Equations (5)–(10) are simply the facts that in a national system of regions, sums of all the region's economic activities must equal the totals from the national account.

Assume a national IO table always exists. There also exists superior statistical data for each regional sector on gross outputs and associated value-added, total final demands, and international exports and imports ($x_i^r$, $v_i^r$, $y_i^r$, $e_i^r$, $m_i^r$). Then all variables on the right side of Equations (1)–(10) listed above can be treated as parameters. With this information, we seek to estimate an IRIO table containing $G \times G$ different intermediate transaction tables ($Z^{rs}$, $r$, $s \in G$), $2 \times G$ different international transaction tables ($\mathrm{MI}^r$, $\mathrm{MY}^r$, $r \in G$), and $G \times G$ different final demand tables ($\mathrm{Y}^{rs}$, $r,s \in G$).

To formulate a mathematical programming model to this problem, one can construct either informed (e.g., survey-based) or uninformed (e.g., data pooling) initial estimates for each endogenous element of the IRIO table—$\bar{z}_{ij}^{sr}$, $\bar{y}_{ik}^{sr}$, $\bar{m}_{ij}^{r}$, and $\bar{m}_{ik}^{r}$—along with reliability measures to weight each initial estimate —$wz_{ij}^{sr}$, $wy_{ik}^{sr}$, $wm_{ij}^{r}$, and $wm_{ik}^{r}$—and specify a cross-entropy (Harrigan and Buchanan, 1984, Golan et al., 1994) or a quadratic objective penalty function subject to Equations (1)–(5) as constraints.[4] In this context, "uninformed" initial estimates are derived in the absence of information about variations in row or column structures in the target account. In such cases, one typically adopts proportional allocation methods and assigns weights in these same proportions. Applying "informed" initial estimates requires the development of a maximum concordance among data sources that support initial estimates. In other words, an informed mathematical programming calibration of an IRIO account requires a classification of sectors and regions that allows using the greatest amount of primary information from multiple sources

[4]When $x_i^r$, $v_i^r$, $y_i^r$, $e_i^r$, and $m_i^r$ are known, they have to be consistent with the national IO account. This implies Equations (6)–(10) have to be pre-satisfied so that the initial data set for the model is internally consistent. This can be achieved by solving a similar quadratic programming model with $x_i^r$, $v_i^r$, $y_i^r$, $e_i^r$, $m_i^r$, and their reliability weights in the objective function and Equations (6)–(10) as constraints. In such cases, the IRIO estimation problem will be solved in two steps and only Equations (1)–(5) remain as constraints with Equation (11) as the objective function in the second step.

that collectively provide consistent descriptions of all row or column structures in the target account. Ideally, the primary information sources include statistical measures of reliability that can be used to weight these initial estimates.

For example, the quadratic objective penalty function for this mathematical programming model is as follows:

$$\min S = \frac{1}{2}\left[\sum_{s=1}^{g}\sum_{r=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(z_{ij}^{sr}-\bar{z}_{ij}^{sr})^2}{wz_{ij}^{sr}} + \sum_{s=1}^{g}\sum_{r=1}^{g}\sum_{i=1}^{n}\sum_{k=1}^{h}\frac{(y_{ik}^{sr}-\bar{y}_{ik}^{sr})^2}{wy_{ik}^{sr}} + \sum_{r=1}^{g}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(m_{ij}^{r}-\bar{m}_{ij}^{r})^2}{wm_{ij}^{r}} + \sum_{r=1}^{g}\sum_{i=1}^{n}\sum_{k=1}^{h}\frac{(m_{ik}^{r}-\bar{m}_{ik}^{r})^2}{wm_{ik}^{r}}\right] \tag{11}$$

A solution to this quadratic programming model provides a complete set of estimates for a full-fledged IRIO table with imports endogenous (Miller and Blair, 1985; Isard, et al., 1998). It is similar in many aspects with the interregional accounting framework proposed by Batten (1982) two decades ago, who used an entropy formulation based on an uninformed data pooling approach for initial estimates where all weights are equal to one. This type of model becomes operational and provides better empirical estimation results on interregional shipments only when interregional trade flow information is explicitly incorporated into both the initial estimates and the underlying accounting framework.

In practice, calibration of such an account directly is hampered by two limitations. First, as combinations of sectors and regions increase, the dimension of this model becomes very large even for a moderate account size. One quickly encounters the problem known as dimension explosion. Related to this, the data requirements of an IRIO account are daunting. The account requires not only knowing the origin and destination of all product flows, but also every intermediate and/or final use must be specified for all such flows. Few national statistical systems can provide such detailed statistics to support the development of informed initial estimates. Therefore, it is not surprising that uninformed initial estimates were used in Battan's approach.

### *MRIO Account and Estimation Model*

The IRIO account described in the previous section can be easily reduced to an MRIO account by forming aggregations of $z_{ij}^{sr}$, $y_{ik}^{sr}$, and $m_{ij}^{r}$ as follows:[5]

[5]The variables $d_i^{sr}$ and $z_{ij}^{\bullet r}$ have no counterparts in Batten's framework, reflecting important departures in the present approach.

$$\sum_{j=1}^{n} z_{ij}^{sr} + \sum_{k=1}^{h} y_{ik}^{sr} = d_i^{sr} \tag{12}$$

$$\sum_{s=1}^{g} z_{ij}^{sr} + m_{ij}^{r} = z_{ij}^{\bullet r} \tag{13}$$

Inserting Equation (12) into Equation (1) gives us Equation (14)

$$\sum_{s=1}^{g} d_i^{rs} + e_i^{r} = x_i^{r} \tag{14}$$

Inserting Equation (13) into Equation (2) results in Equation (15)

$$\sum_{j=1}^{n} z_{ji}^{\bullet r} + v_i^{r} = x_i^{r} \tag{15}$$

It is easy to show that sum Equation (13) by $j$ over $N$ plus Equation (3) equals sum of Equation (12) by $s$ over $G$ plus Equation (4). This linear combination of Equations (3), (4), (12), and (13) produces Equation (16)

$$\sum_{j=1}^{n} z_{ij}^{\bullet r} + y_i^{r} = \sum_{s=1}^{g} d_i^{sr} + m_i^{r} \tag{16}$$

Finally, inserting Equation (13) into Equation (5) results in Equation (17)

$$\sum_{r=1}^{g} z_{ij}^{\bullet r} = z_{ij} \tag{17}$$

Equation (14) indicates that total gross output of commodity "$i$" in region "$r$" is delivered to domestic regions (including its own) and what is left over is exported to other nations. No indication about the type of use is given. Equation (15) indicates the value of gross output of commodity "$i$" in region "$r$" is attributed to the value of all sector "$i$" intermediate purchases (regardless of origin) and to the value of services from sector "$i$" primary factor inputs. Equation (16) indicates total intermediate and final requirements for commodity "$i$" in region "$r$" must be met by deliveries from all regions (including from its own) within the nation plus imports from other nations. Thus, Equations (14)–(17) plus equations (6)–(10) together also consistently define an accounting framework for the national system of economic regions, conventionally called an MRIO table in the literature (Miller and Blair, 1985; Isard, et al., 1998). Such an account stops short of assigning specific intermediate or final uses for inter/intraregional product flows, but guarantees that these flows exactly meet all regional demands. Further, because this alternative formulation [Equations (14)–(17)] is mathematically equivalent to Equations (1)–(5), a solution to the MRIO account will also be consistent with the IRIO account, so that it can be seen as an important intermediate step toward estimating a full-fledged IRIO account. Needless to say, the MRIO account has a much smaller dimension thus significantly reducing the data required and computational difficulties to empirically estimate interregional trade flows and interindustrial transactions.[6] The smaller information requirements make it more plausible to develop an objective function with informed initial estimates and reliability weights. The use of informed initial estimates is another major motivation underlying this alternative formulation.

To demonstrate, suppose, as before, that statistics exist for each regional sector on the gross outputs and value added ($x_i^r$ and $v_i^r$), the origin of exports and destination of imports ($e_i^r$ and $m_i^r$), and the final regional demands $(y_i^r)$; the above MRIO estimation problem can be formally stated as follows:

*Given a $n \times g \times g$ non-negative array $\overline{\mathbf{D}} = \{\bar{d}_i^{sr}\}$ and a $n \times n \times g$ non-negative array $\overline{\mathbf{Z}} = \{\bar{z}_{ij}^{\bullet r}\}$, determine a non-negative array $\mathbf{D} = \{d_i^{sr}\}$ and a non-negative array $\mathbf{Z} = \{z_{ij}^{\bullet r}\}$ that is close to $\overline{\mathbf{D}}$ and $\overline{\mathbf{Z}}$ such that Equations (14) to (17) are satisfied, where $s \in G$ denotes the shipping regions, $r \in G$ denotes the receiving regions, and $i, j \in N$ denotes the make and use sectors respectively.*

In plain English, the estimation problem is to modify a given set of initial interregional and interindustrial transaction estimates to satisfy the above four known accounting constraints. With the account structure known and with predefined parameter values, what remains is the formulation of a criteria for changing the initial estimates in the account to conform to the known linear accounting constraints. As introduced in Equation (11), many have proposed using a mathematical programming approach that employs an objective function which penalizes the deviations of the estimated array D and Z from the initial array $\overline{\mathbf{D}}$ and $\overline{\mathbf{Z}}$. Two types of alternative functional forms are often used. One is the Quadratic function similar to equation (11):

$$(18) \qquad \min S = \frac{1}{2}\left[\sum_{i=1}^{n} \sum_{s=1}^{g} \sum_{r=1}^{g} \frac{(d_i^{sr} - \bar{d}_i^{sr})^2}{wd_i^{sr}} + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{g} \frac{(z_{ij}^{\bullet r} - \bar{z}_{ij}^{\bullet r})^2}{wz_{ij}^{\bullet r}}\right]$$

The other is cross-entropy function as follows:

$$(19) \quad \min S = \sum_{i=1}^{n} \sum_{s=1}^{g} \sum_{r=1}^{g} \frac{d_i^{sr}}{wd_i^{sr}} \bullet \ln(d_i^{sr}/\bar{d}_i^{sr}) + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{g} \frac{z_{ij}^{\bullet r}}{wz_{ij}^{\bullet r}} \bullet \ln(z_{ij}^{\bullet r}/\bar{z}_{ij}^{\bullet r})$$

*Properties of the Estimation Model*

There are desirable theoretical properties of the estimation model outlined above. First, it is a separable non-linear programming problem subject to linear constraints. The entropy function is motivated from information theory and is the objective function underlying the well-known RAS procedure with

[6]The aggregate model only has $N(NG + G^2 + 5G)$ variables and $N(3G + N + 5)$ constraints, while the full detailed model has $(N^2G + NHG)(G + 1)$ variables and $N(G^2 + NG + N + 5)$ constraints. It is a much smaller model, having $NG^2(N - 1) + NG(HG - 5)$ less variables and $NG(G + N - 3)$ less constraints.

row and column totals known with certainty (Senesen and Bates, 1988). It measures the information gain by $\mathbf{D}$ and $\mathbf{Z}$ given the initial estimates $\overline{\mathbf{D}}$ and $\overline{\mathbf{Z}}$. The quadratic penalty function is motivated by statistical arguments. There are different statistical interpretations underlying the model by choices of different reliability weights $wd_i^{sr}$ and $wz_{ij}^{\bullet r}$. When the weights are all equal to one, solution of this model gives a constrained least-square estimator. When the initial estimates are taken as the weights, solution of the model gives a weighted constrained least-square estimator, which is identical to the Friedlander solution, and a good approximation of the RAS solution. When those weights are proportional to the variances of the initial estimates and the initial estimates are statistically independent (the variance and covariance matrix of $\overline{\mathbf{D}}$ and $\overline{\mathbf{Z}}$ are diagonal), the solution of the model yields best linear unbiased estimates of the true unknown matrix (Byron, 1978), which is identical to the generalized least-squares estimator if the weights are equal to the variance of initial estimates (Stone, 1984; Ploeg, 1984). Furthermore, as noted by Stone, Champernowne, and Meade (1942) and proven by Weale (1985), in cases where the error distributions of the initial estimates are normal, the solution also satisfies the maximum likelihood criteria.

Second, the quadratic and entropy objective functions are equivalent in the neighborhood of initial estimates, under a properly selected weighting scheme. By taking second-order Taylor expansion of Equation (18) at point $(\bar{d}_i^{sr}, \bar{z}_{ij}^{\bullet r})$, we have

$$
\begin{aligned}
S &= \sum_{i=i}^{n}\sum_{s=1}^{g}\sum_{r=1}^{g}\left[\left(d_i^{sr}-\bar{d}_i^{sr}\right)+\frac{\left(d_i^{sr}-\bar{d}_i^{sr}\right)^2}{2\bar{d}_i^{sr}}\right] \\
&\quad + \sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{r=1}^{g}\left[\left(z_{ij}^{\bullet r}-\bar{z}_{ij}^{\bullet r}\right)+\frac{\left(z_{ij}^{\bullet r}-\bar{z}_{ij}^{\bullet r}\right)^2}{2\bar{z}_{ij}^{\bullet r}}\right] \qquad (20)\\
&= \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{s=1}^{g}\sum_{r=1}^{g}\frac{\left(d_i^{sr}-\bar{d}_i^{sr}\right)^2}{\bar{d}_i^{sr}}+\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{g}\frac{\left(z_{ij}^{\bullet r}-\bar{z}_{ij}^{\bullet r}\right)^2}{\bar{z}_{ij}^{\bullet r}}\right]+R
\end{aligned}
$$

This is the quadratic function (18) plus a remainder term $R$. As long as the posterior estimates and the initial estimates are close and the initial estimates are used as reliability weights, the term $R$ will be small and the two objective functions can be regarded as approximating one another.[7]

Third, as proven by Harrigan (1990), in all but the trivial case, posterior estimates derived from entropy or quadratic loss minimand will always better approximate the unknown, true values than do the associated initial estimates. In this framework, information gain is interpreted as the imposition

[7]The quadratic functional form has a numerical advantage in implementing the model. It is easier to solve than the entropy function in very large models because they can use software specifically designed for quadratic programming.

of additional valid constraints or the narrowing of bounds on existing constraints as long as the true but unknown values belong to the feasible solution set. This is because adding valid constraints or further restricting the feasible set through the narrowing of interval constraints cannot move the posterior estimates away from the true values, unless the additional constraints are non-binding (have no information value). Although the posterior estimates may not always be regarded as providing a "reasonable" approximation to the true value, they are always better than the initial estimates in the sense the former is closer to the true value than the latter, so long as the imposed constraints are true.[8] In other words, the optimization process has the effect of reducing, or at least not increasing, the variance of the estimates. This property is simple to show by using matrix notation. Define $\mathbf{W}$ as the variance matrix of initial estimates $\overline{\mathbf{D}}$, $\mathbf{A}$ as the coefficient matrix of all linear constraints. The least-squares solution (equivalent to the quadratic minimand as noted above) to the problem of adjusting $\overline{\mathbf{D}}$ to $\mathbf{D}$ that satisfies the linear constraint, $\mathbf{A}\bullet\mathbf{D} = \mathbf{0}$ can be written as:

$$\mathbf{D} = (\mathbf{I} - \mathbf{W}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A})\mathbf{D} \tag{21}$$

Thus,

$$\mathrm{var}(\mathbf{D}) = (\mathbf{I} - \mathbf{W}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A})\mathbf{W} = \mathbf{W} - \mathbf{W}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\mathbf{W} \tag{22}$$

Since $\mathbf{W}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{W}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\mathbf{W}$ is a positive semidefinite matrix, the variance of posterior estimates will always be less, or at least not greater than the variance of the initial estimates as long as $\mathbf{A}\bullet\mathbf{D}^{\mathrm{true}} = \mathbf{0}$ holds. This is the fundamental reason why such an estimating framework will provide better posterior estimates. Imposing accounting relationship's (6)–(10) and (14)–(17) will definitely improve, or at least not worsen the initial estimates, since we are sure from economics that those constraints are identities and must be true for any national system of economic regions.

Finally, the choice of weights in the objective function has very important impacts on the estimation results. For instance, using the initial estimates as weights has the property that each entry of the array is adjusted in proportion to its magnitude to satisfy the accounting identities, and that large variables are adjusted more than small variables. However, the adjustment relates directly to the size of the initial estimates $\bar{d}_i^{sr}$ and $\bar{z}_{ij}^{\bullet r}$ and does not force the unreliable initial estimates to absorb the bulk of the required adjustment. Furthermore, this commonly used weighting scheme (under RAS) can obtain best unbiased estimates provided two assumptions are met. One is that the initial estimates for different elements in the array are statistically independent, and the other that each error variance is proportional to the

[8]The minimand objective function reflects the principle that the "distance" between the posterior and initial estimates should be minimized. What we would like is to minimize the "distance" between the posterior estimates and the unknown true values. This "distance" cannot be measured, but a good estimation procedure should have a desirable influence on it.

corresponding initial estimates. However, those assumptions may not hold in many cases. Fortunately, the model is not restricted to use only a diagonal-weighting matrix such as the initial estimates. When a variance–covariance matrix of the initial estimates is available, it can be incorporated into the model by modifying the objective function as follows:

$$\min S = (\mathbf{D} - \overline{\mathbf{D}})^{\mathrm{T}}\mathbf{W}\mathbf{D}^{-1}(\mathbf{D} - \overline{\mathbf{D}}) + (\mathbf{Z} - \overline{\mathbf{Z}})^{\mathrm{T}}\mathbf{W}\mathbf{Z}^{-1}(\mathbf{Z} - \overline{\mathbf{Z}}) \tag{23}$$

The efficiency of the resulting posterior estimator will be further improved if the error structure of the initial estimates is available, because such a weighting scheme makes the adjustment independent of the size of the initial estimates. The larger the variance, the smaller its contribution to the objective function, and hence the lesser the penalty for $d_i^{sr}$ and $z_{ij}^{\bullet r}$ to move away from their initial estimates (only the relative, not the absolute size of the variance affects the solution). A small variance of the initial estimates indicates, other things equal, they are very reliable data and thus should not change by much, whilst a large variance of the initial estimates indicates unreliable data and will be adjusted considerably in the solution process. Therefore, this weighting scheme gives the best unbiased estimates of the true, unknown interregional and interindustrial transaction value under the assumption that initial estimates for different elements in the array are statistically independent. Although there is no difficulty in solving such a non-linear programming problem like this today, the major problem is the lack of data to estimate the variance–covariance matrix associate with the initial estimates.

Stone (1984) proposed to estimate the variance of $\bar{z}_{ij}^{\bullet r}$ as $\mathrm{var}(\bar{z}_{ij}^{\bullet r}) = (\theta_{ij}^{*r}\bar{z}_{ij}^{\bullet r})^2$, where $\theta_{ij}^{*r}$ is a subjectively determined reliability rating, expressing the percentage ratio of the standard error to $\bar{z}_{ij}^{\bullet r}$. Weale (1989) had used time-series information on accounting discrepancies to infer data reliability. The similar methods can be used to derive variances associated with those initial estimates in our model.

Despite the difficulties in obtaining data for the best weighting scheme, advantages of such a model in estimating interregional trade flows and interindustrial transactions are still obvious from an empirical perspective. First, it is very flexible regarding the required known information. For example, it allows for the possibility that the region total of output, value-added, exports, imports and final demands are not known with certainty. In the real world, these regional statistics typically have substantial gaps and inconstancies with the national total. Incorporating associated terms similar to $\overline{\mathbf{D}}$ and $\overline{\mathbf{Z}}$ in the objective function to penalize solution deviations from the initial estimates from statistical sources allows the estimation of those regional totals, together with entries in the interregional delivery and interindustrial transaction array. With the use of upper and lower bounds, this fact can also be modeled by specifying ranges rather than precise values for the linear constraints (14)–(17). In addition, the estimation of $\mathbf{D}$ or $\mathbf{Z}$ will be a special case of the framework when only one set of additional data is available.

Second, it permits a wider variety and volume of information to be brought into the estimation process. For example, the ability of introducing upper and/or lower bounds on those regional totals is one of the flexibilities not offered by commonly used scaling procedures such as RAS. The gradient of the entropy function tends to infinity as $d_i^{sr}$ and $z_{ij}^{\bullet r} \to 0$, and hence restricts the value of the posterior estimates to non-negative. This is a desirable property of estimating interregional trade data.[9]

Third, the weights in the objective function reflect the relative reliability of a given set of initial estimates. The interpretation of the reliability weights is straightforward. Other things equal, entries with higher reliability should be changed less than entries with a lower reliability. The choice of those weights is also very flexible. They will use the best available information to ensure that reliable data in the initial estimates are not being modified by the optimization model as much as unreliable data. In practice, such reliability weights can be put into a second array that has the same dimension and structure as the initial estimates. The inverted variance–covariance matrix of the initial estimates is statistically interpreted as the best index of the reliability for the initial data.

Finally, solution of this estimation problem exactly provides the data needed to construct a so-called MRIO model. This model was pioneered by Professor Polenske and her associates at MIT in the 1970s (Polenske, 1980), and is still widely used in regional economic impact analysis today.

## 3. EMPIRICAL TEST OF THE MODEL AND EVALUATION MEASURES

### *The Testing Data Set*

How does the model specified above perform when applied to data from the real world? To evaluate the model's performance, a benchmark data set from the real world is needed. Because good interregional trade data is quite rare and very difficult to obtain in any country of the world, a natural place to find such data sets are existing global production and trade databases such as the GTAP (global trade analysis project) database. For instance, version 4 GTAP database contains detailed bilateral trade, transportation, and individual country's IO data covering 45 countries and 50 sectors (McDougall, Elbehri, and Truong, 1998). For our particular purpose, version 4 GTAP database was first aggregated into a four-region, 10-sector data set. Then three of the four regions (the United States, European Union, and Japan) were further aggregated into a single open economy which engages in both interregional trade among its three internal regions and international trade with the rest of the world. We use this partitioned data set as the benchmark

[9] Zeros can become non-zeros and vice versa under a quadratic penalty function. However, a side effect for the cross-entropy function is that if there are too many zeros in the initial estimates, the whole problem may become infeasible.

for a hypothetical national economy, and attempt to use our model to replicate the underlying intercontinental trade flows among Japan, EU, and the United Sates as well as the individual country's IO accounts.

*Experiment Design*

In the first experiment, we do this without the use of the region-specific IO coefficients as the situation encountered in the real world, where only the national IO table is available to economists (it is the three regions' weighted average in our experiment and are defined as $z_{ij}^{\bullet r} = [z_{ij}/(x_j - v_j)] \times (x_j^r - v_j^r)$ to make full use of the known information). Initial estimates of interregional commodity flows are from the "true" interregional trade data in the GTAP database but was distorted by a normally distributed random error term with zero mean and the size of standard deviation as large as five times the "true" trade data. The solution from the model is compared with the benchmark data set for both the interregional shipment and intersector transaction flows.

In the second experiment, we use the region-specific IO coefficients as constant in the model. We re-estimate the interregional shipment data as the first experiment, and compare the model solution with the benchmark data set for the interregional trade data only.

In the third experiment, we assume the interregional shipment pattern is known with certainty and we use the three regions' weighted average IO coefficients as initial estimates to estimate the region-specific IO accounts.

In the fourth experiment, Batten's model was used to estimate the interregional shipment and individual region's IO flows. In the fifth to the seventh experiments, experiments 1–3 were repeated by using the detailed model. Solutions from both models are compared with the "true" interregional trade and intersector IO flow data in the aggregated GTAP data set. The assumptions, initial estimates, and expected model solution are summarized in Table 1.

*Measures to Evaluate Test Results*

Each experiment produces a different set of estimates, and it is desirable to know how much each set of estimates differs from the true, known data. However, it is difficult to use a single measure to compare the estimated results. Since there are so many dimensions in the model solution sets, a particular set of estimates may score well on one region or commodity but badly on others. It is meaningful to use several measures to gain more insight into the model performance in different experiments. Generally speaking, it is the proportionate errors and not the absolute errors that matter; therefore, the "mean absolute percentage error" with respect to the true data will be calculated for different commodity and regional aggregations. Consider the following aggregate index measure for intra/interregional trade flows:

TABLE 1: Experiment Design

| Experiment Number | Data Known with Certainty | Initial Estimates | What is Estimated by the Model |
|---|---|---|---|
| 1 | None | $d_i^{sr}$ is distorted from the "true" data $\bar{d}_i^{sr}$<br>$z_{ij}^{\bullet r} = (z_{ij}/(x_j - v_j)) \times \left(x_j^r - v_j^r\right)$ | Z and D |
| 2 | $\mathbf{Z} = \bar{\mathbf{Z}}$ | D is distorted from the "true" data $\bar{\mathbf{D}}$ | D only |
| 3 | $\mathbf{D} = \bar{\mathbf{D}}$ | $z_{ij}^{\bullet r} = \left[z_{ij}/(x_j - v_j)\right] \times \left(x_j^r - v_j^r\right)$ | Z only |
| 4 | None | $\bar{z}_{ij}^{sr} = \frac{x_i^s + m_i^s - e_i^s}{x_i + m_i - e_i} \times \frac{x_j^r - v_j^r}{x_j - v_j} \times z_{ij}$<br>$\bar{y}_i^{sr} = y_i^r \times \left[x_i^s + m_i^s - e_i^s\right] / [x_i + m_i - e_i]$<br>[Equations (16) and (17) in Batten (1982)] | Z and D |
| 5 | None | $\bar{z}_{ij}^{sr} = d_i^{sr} \times z_{ij}^{\bullet r} / \left[\sum_j z_{ij}^{\bullet r} + y_i^r\right]$ $\quad \bar{y}_i^{sr} = d_i^{sr} - \sum_j \bar{z}_{ij}^{sr}$ | Z and D |
| 6 | $\mathbf{Z} = \bar{\mathbf{Z}}$ | $\bar{z}_{ij}^{sr} = d_i^{sr} \times \bar{z}_{ij}^{\bullet r} / \left[\sum_j \bar{z}_{ij}^{\bullet r} + y_i^r\right]$ $\quad \bar{y}_i^{sr} = d_i^{sr} - \sum_j \bar{z}_{ij}^{sr}$ | D only |
| 7 | $\mathbf{D} = \bar{\mathbf{D}}$ | $\bar{z}_{ij}^{sr} = \bar{d}_i^{sr} \times z_{ij}^{\bullet r} / \left[\sum_j z_{ij}^{\bullet r} + y_i^r\right]$ $\quad \bar{y}_i^{sr} = \bar{d}_i^{sr} - \sum_j \bar{z}_{ij}^{sr}$ | Z only |

Note: In all experiments, national totals: $z_{ij}$, $x_i$, $y_i$, $v_i$, $e_i$, and $m_i$ are known with certainty, i.e. they enter the model as constant. It is not necessary for the state totals–$x_i^r, y_i^r, v_i^r, e_i^r, m_i^r$ –be known with certainty in the model, however, in all experiments reported in this paper, they enter the model as constant. In experiments 5–7, we did not distinguish different final demand types in the detailed model.

$$\text{MAPE}^{\text{D}} = \frac{100 \bullet \sum_{i=1}^{n} \sum_{s=1}^{g} \sum_{r=1}^{g} \left| d_i^{sr} - \bar{d}_i^{sr} \right|}{\sum_{i=1}^{n} \sum_{s=1}^{g} \sum_{r=1}^{g} \bar{d}_i^{sr}} \tag{24}$$

Alternating the removal of summations over $i$, $s$, and $r$ in Equation (24) produces MAPE estimates on shipments by commodities, shipping regions, and receiving regions respectively. For regional intermediate transactions, the aggregate MAPE index is defined as:

$$\text{MAPE}^{\text{Z}} = \frac{100 \bullet \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{g} \left| z_{ij}^{\bullet r} - \bar{z}_{ij}^{\bullet r} \right|}{\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{g} \bar{z}_{ij}^{\bullet r}} \tag{25}$$

Alternating the removal of summations over $i$, $j$, and $r$ in Equation (25) produces MAPE estimates on intermediate transactions by inputs, using sectors, and regions respectively. The model and all test experiments are implemented in GAMS and the complete GAMS program and related data set are available from the authors upon request.

*Test Results*

Table 2 summarizes all the eight measurement indexes from the seven experiments listed in Table 1. The accuracy of the estimates is judged by their closeness to the true interregional trade and individual region's IO flows aggregated from the GTAP database.

Generally speaking, the model has remarkable capacity to rediscover the true interregional trade flows from the highly distorted data. The estimated shipment data are very close to the true data, as judged by the eight MAPE measurements, in all testing experiments except the Batten model. Most of the mean absolute percentage errors are about 4–7 percent of the true data value, which implies the model has great potential in the application of estimating interregional trade flows. In contrast, recovering the individual region's IO flows from weighted average national values only obtained limited success, indicating national IO coefficients in detailed sectors may be the best place to start in building regional IO accounts if there is no additional prior information on regional technology or cost structure available.[10]

Comparing estimates from different test experiments, there are several interesting observations. First, when there is no additional information that can be incorporated into the estimation framework, a more detailed model may not perform better than a simpler model. Comparing results from Experiments 1 and 5, the more sophisticated model actually brings less accurate estimates overall because of additional numbers of unknown variables without additional known data. However, as results in Experiments 2–3 and 6–7 show, the estimation accuracy does improve by a more detailed model when more useful data become available. Second, the marginal accuracy gained from actual individual regional IO flows is significant in estimating interregional trade flows using the detailed model, but very small in the aggregate version. In contrast, the marginal value of accurate interregional shipment data is rather small in estimating individual regional IO coefficients under both versions of the model. Finally, Batten's model performed poorly in interregional shipment estimation, but obtained similar estimates on individual regional IO flows as our model, providing further evidence that there may be no high dependency between individual regional IO coefficients and interregional trade flows. However, this is not a firm conclusion because the particular data set used to test the model in this paper may be part of the problem. Since the United States, EU, and Japan are all large economies, their intermediate demands are largely met by their own production.

[10]Following the product mix method outlined in Miller and Blair (1985), initial estimates of IO coefficients for each of the 10 aggregated industries are unique for each region. They are weighted averages of the three-region detailed (50-industry) IO coefficients where the weights are the gross regional outputs of the relevant detailed industries. Experiment results show that a "product mix" approach improves the accuracy of the true regional IO flow estimates compared to an approach that directly uses the three-region average IO coefficients, although the differences are small in our particular model aggregation.

TABLE 2: Mean Absolute Percentage Error from the True Data

| Experiment Number | Distorted priors | | Experiment 1 | | Experiment 2 | Experiment 3 | Experiment 4 Batten model | | Experiment 5 | | Experiment 6 | Experiment 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Indexes | $d_i^{sr}$ | Average IO | $d_i^{sr}$ | $z_{ij}^{\bullet r}$ | $d_i^{sr}$ | $z_{ij}^{\bullet r}$ | $d_i^{sr}$ | $z_{ij}^{\bullet r}$ | $d_i^{sr}$ | $z_{ij}^{\bullet r}$ | $d_i^{sr}$ | $z_{ij}^{\bullet r}$ |
| Total MAPE | 399.75 | 21.72 | 5.92 | 18.22 | 5.69 | 17.40 | 126.13 | 18.54 | 7.02 | 19.54 | 2.05 | 15.65 |
| *Receiving region MAPE* | | | | | | | | | | | | |
| United States | 265.83 | 17.28 | 8.75 | 19.03 | 8.68 | 15.41 | 129.88 | 16.49 | 10.46 | 24.12 | 3.90 | 13.82 |
| European Union | 447.06 | 20.94 | 3.97 | 15.31 | 3.61 | 15.72 | 111.73 | 16.51 | 4.93 | 14.74 | 0.74 | 14.22 |
| Japan | 494.73 | 28.51 | 5.57 | 22.47 | 5.34 | 22.83 | 145.59 | 24.68 | 6.12 | 22.60 | 1.86 | 20.43 |
| *Sector MAPE I* | | Inputs | | | | | | | | | | |
| Primary agriculture | 304.53 | 25.48 | 5.37 | 25.61 | 5.19 | 24.61 | 125.51 | 34.92 | 7.51 | 27.43 | 1.67 | 23.16 |
| Processed agriculture | 319.40 | 14.18 | 9.99 | 15.73 | 10.67 | 11.82 | 129.42 | 13.06 | 9.74 | 18.23 | 2.97 | 10.81 |
| Resource-based sectors | 392.24 | 53.70 | 3.16 | 20.06 | 5.52 | 21.76 | 135.00 | 13.28 | 4.10 | 15.17 | 2.15 | 16.90 |
| Non-durable goods | 312.28 | 15.85 | 4.46 | 9.03 | 3.85 | 10.04 | 127.87 | 11.44 | 5.82 | 10.72 | 3.36 | 9.38 |
| Durable goods | 413.91 | 13.69 | 4.81 | 12.74 | 4.36 | 12.02 | 121.60 | 14.06 | 5.24 | 12.91 | 3.38 | 10.43 |
| Utility | 774.76 | 22.36 | 5.29 | 22.56 | 1.40 | 22.62 | 121.86 | 24.73 | 5.93 | 23.30 | 0.95 | 24.08 |
| Construction | 484.64 | 44.19 | 3.34 | 21.58 | 2.61 | 21.16 | 133.12 | 22.53 | 3.63 | 23.87 | 0.01 | 18.45 |
| Trade and Transport | 406.12 | 21.53 | 12.24 | 22.47 | 12.68 | 22.11 | 130.52 | 20.83 | 13.04 | 26.37 | 3.08 | 23.83 |
| Private services | 245.15 | 20.86 | 4.47 | 20.56 | 5.07 | 19.35 | 126.71 | 20.30 | 5.83 | 21.55 | 1.17 | 17.31 |
| Public services | 539.32 | 30.69 | 2.48 | 29.30 | 1.30 | 27.49 | 118.65 | 29.77 | 6.01 | 30.08 | 0.62 | 16.12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Shipping region MAPE | | | | | | |
| United States | 264.78 | 9.17 | 9.08 | 130.65 | 9.92 | 2.90 |
| European Union | 445.56 | 3.83 | 3.64 | 111.83 | 5.30 | 1.57 |
| Japan | 495.24 | 5.28 | 4.80 | 144.28 | 6.22 | 1.75 |
| *Sector MAPE II* | Use | | | | | |
| Primary agriculture | 13.54 | 12.98 | 11.04 | 12.03 | 13.22 | 9.31 |
| Processed agriculture | 15.42 | 20.90 | 15.61 | 18.90 | 27.60 | 16.17 |
| Resource-based sectors | 42.54 | 18.91 | 18.45 | 21.81 | 17.67 | 17.24 |
| Non-durable goods | 14.22 | 9.83 | 10.65 | 12.32 | 11.35 | 11.68 |
| Durable goods | 19.07 | 11.37 | 11.73 | 12.40 | 11.25 | 11.31 |
| Utility | 33.77 | 25.90 | 27.60 | 29.16 | 24.46 | 22.75 |
| Construction | 42.75 | 43.54 | 41.74 | 46.29 | 43.43 | 41.60 |
| Trade and Transport | 21.89 | 22.42 | 20.04 | 20.88 | 29.75 | 18.02 |
| Private services | 16.81 | 17.75 | 16.61 | 16.68 | 18.19 | 15.88 |
| Public services | 51.25 | 46.73 | 46.64 | 50.94 | 40.98 | 16.26 |

Note: IO, input–output; MAPE, mean absolute percentage error.

Therefore, the correlation between individual interindustrial flow and interregional shipments may be particularly low.

The detailed model only provides better estimates of interregional shipments when regional IO data are available, and hence the aggregate version of the model specified in this paper may be the best practitioner's tool in estimating interregional trade flows because of the lack of subnational IO data in the real world. It demands less statistical information and has a smaller model dimension, which facilitates the implementation and computation process.

## 4. IMPLICATIONS FOR APPLYING THE MODEL

Results in the previous section offer some guidance for applying the framework outlined in this paper to real world statistics. It was found that initial estimates of regional commodity trade flows based on survey data over very high statistical variability are highly preferable (in the experiments) to a widely used non-survey approach for producing initial estimates.[11] This finding holds promise for opportunities to use other survey data to recover unobserved regional economic accounts. It was also found that solving an aggregate account (e.g., an MRIO or MR-SAM) as an intermediate step is at least as accurate (in the experiments) as producing a direct solution to a detailed account (e.g., IRIO or IR-SAM) when superior data unique to the latter are not widely available. This finding is useful when working with regional economic accounts of considerable sector and region details. Results also support the "product mix" approach, whereby the most feasible sector detail for regional gross output estimates are used to derive weighted average national technical coefficients for more aggregated regional sectors.

Statistical systems vary by nation and no one-size fits all rules exist that tell us how to use each country's statistical data system to its best advantage.[12] However, there are general guidelines for implementing the optimization framework presented in this paper to a large dimension multiregional account. To facilitate discussions of implementation, we assume that a detailed national account always exists and regional sector statistics are also available in a variety of details. Then the implementation process may be classified into three broad phases as discussed below.

[11]A random normal distortion of the "true" trade data by an average of 400 percent was produced in the previous section to simulate a well-designed but poorly sampled transportation survey of annual commodity flows.

[12]Comprehensive studies by West (1990) and Lahr (2001) consider how to identify and use superior data in a regional accounting system context.

*Develop Independent Estimates for Major Components of a Multiregional Account*

It has been stressed as far back as Wilson (1970) that information used to produce parameters and initial estimates of a regional economic system should be estimated independently. While this produces unbalanced initial accounts, it avoids introducing spurious information that can lead to biased estimates (McDougall, 1999). A useful approach is to partition the multiregional account into components that coincide or are related to known statistical survey series published regularly in the nation under study.

For the MRIO account outlined in Equations (6)–(10) and (14)–(17), the major components are gross regional output $\left(x_i^r\right)$, final demand $\left(y_i^r\right)$, primary factor payments $\left(v_i^r\right)$, international trade ($e_i^r$ and $m_i^r$), interindustry transactions ($z_{ij}^{\bullet r}$), and inter/intraregional trade flows $\left(d_i^{sr}\right)$. In many cases, data for several of these components are available from a single major statistical survey series—for example, in the United States $x_i^r$ and $v_i^r$ are available from an Economic Census conducted every five years. Other components, for example $y_i^r$, may themselves require multiple disparate data sources to compile. While the strategic groupings may differ by country, it is likely that for large dimension ($N \times G$) multiregional accounts, primary data for individual regional sectors become sparse.

When the best available data are not consistent with the model structure, it may be necessary to restructure the adding up requirements in the model to accommodate the data. For example, in Equations (14) and (16) of our model, the accounting identities require data for international exports $\left(e_i^r\right)$ and imports $\left(m_i^r\right)$ on an origin of movement and destination of use basis respectively. However, in many countries such as the United Sates, port of entry/exit data are far more reliable. Therefore, different formulation of the corresponding accounting identities should be used.

For certain elements of the multiregional account, very often only a purely theoretical inference is available to produce informed guesses about the initial estimates. A common example is the information about service trade flows within and between regions. In using a theory-based alternative to data, a case must be made for a prevailing empirical model that calibrates the unobserved activities to some other statistics or available survey data.

*Determine Model Dimensions based on Maximum Concordance Among Different Components*

In compiling different components of the multiregional account, the volume and nature of data available for each component can greatly vary. Detailed and survey-based data may be obtained on, for example, gross regional output and incomes, but survey data on the inter/intraregional trade flows of this output may be far less detailed. Interindustry transactions may only be available at the national level, and international trade data may be very detailed, but based on a different product classification system. The notion of conservatism, both in the information theoretic sense and in terms of computational burden, should be the primary guiding principle in reconciling this information.

Robinson, Cattaneo, and El-said (2001) interpret conservatism by the rule of using “only, and all” information in the estimation problem. Considering this rule in the present context, the fact that a component such as gross regional outputs are available from highly detailed and reliable statistics suggests all this information should be used. However, if the associated intra/interregional trade flow account has more general product aggregations than the output account, it appears that one is faced with an “only or all” decision. Although the specific situation often guides the approach one takes, it is worth noting that there are usually many opportunities to introduce all information available to the estimation process.

In practice, conserving on computational burden may also become an issue. When employing a more general estimation framework such as the model presented in this paper, the use of iterative techniques that diminish computational burden may not be readily available.[13] Both computer hardware and software available to the researchers may become binding in many such instances. For example, access to special solvers or greater programming finesse becomes a more prominent issue when computational burdens grow tremendously as model dimension increase. In addition, while conventional personal computers have improved dramatically, limits on current 32-bit operating systems to manage sufficient memory on PCs may become a binding constraint for very large models. Solutions to these issues can become expensive.

*Adding Additional Constraints to Use All Available Information*

The greatest opportunities to use all relevant information are in the form of additional binding linear constraints, beyond the adding up and consistency requirements, on any selected groups of variables in the aggregate or detailed model. Information deemed “superior” and that is related to any group of elements in either the aggregate or detailed accounts is a candidate for a linear constraint. Since both interregional and multiregional economic accounts are comprehensive and detailed, there are many opportunities to introduce such constraints. A few general guidelines are notable.

Both the detailed and aggregate accounts describe flows of payments and products in the form of a matrix with known adding-up and consistency requirements. Any information used to formulate new constraints—either equality or inequality linear constraints—can greatly diminish the feasible

[13]For example, by allowing both regional technical coefficients and intra/interregional flows to adjust, the optimal solution to the cross-entropy or quadratic formulations in Section 2 must be jointly solved.

solution set of the calibration procedure. However, new constraints that are non-binding add no information to the problem, but do increase the computational burdens.

Where and how information is used to formulate constraints depends on many factors. For example, the U.S. Government has published state measures of farm productivity that include estimates of purchased farm inputs by state for broad input categories. A pro-rated version of this data could form the basis for additional linear constraints for agricultural sector IO flows in the model. Other restrictions could be designed to replicate certain highly reliable economic statistics that can be formed by special groupings of certain flow statistics contained in the account being estimated. Although such information must be carefully compiled, their incorporation in the form of constraints will improve the estimation accuracy greatly.

## 5. CONCLUSIONS AND DIRECTION FOR FUTURE RESEARCH

This study constructed a mathematical programming model to estimate interregional trade patterns and IO accounts based on an interregional accounting framework and initial estimates of interregional shipments in a national system of economic regions. The model is quite flexible in its data requirement and has desirable theoretical and empirical properties. An empirical test of the model using a four-region, 10-sector example aggregated from a global trade database shows that the model performed remarkably well in discovering the true patterns of interregional trade from highly distorted initial estimates on interregional shipments. It shows the model may have great potential in the estimation and reconciliation of interregional trade flow data, which often is the most elusive data to assemble. In addition, solutions from the aggregated model exactly provide the data needed for an MRIO model and solution from the detailed model exactly provide the data needed for an IRIO model. This will greatly reduce the data-processing burden in such analysis. Therefore, application of the model will further facilitate quantitative economic analysis in regional sciences.

Lessons from the experiments in this study shaped our view on approaches for applying the model to real data from a particular nation's statistics. A logical conclusion is that widely available and disparate survey data on the economy, including commodity flows data and incomplete geographic data, can effectively be used to substantially narrow the margins for error in obtaining feasible solutions to interregional IO systems. It is also evident that data on region-to-region commodity flows represent a limiting factor in determining the optimal sector dimensions to be solved in the modeling framework.

However, there are important questions not yet answered by the current study. First, test results from the data set aggregated from GTAP also show that our model's ability to improve the IO transaction estimates of individual regions from national averages may be limited. Continuing research on the real underlying causes and means of improvement are needed to further enhance the model's capacity as an estimating and reconciliation tool in building interregional production and trade accounts. Second, the relative importance of regional sector output, value-added, exports, imports and final demand as model input in the accuracy of a model solution is also not analyzed, and could be addressed with minor changes of the current model. Third, the approach employed in this study draws primarily from regional science and constrained matrix-balancing literatures. How insights into economic geography theory can help define a bounded solution needs to be explored. Finally, the robustness of the model performance should be further tested by using other data sets.

## REFERENCES

Bartholdy, Kasper. 1991. "A Generalization of the Friedlander Algorithm for Balancing of National Accounts Matrices," *Computer Science in Economics and Management*, 2, 163–174.

Batten, David F. 1982. "The Interregional Linkages Between National and Regional Input–Output Models," *International Regional Science Review*, 7, 53–67.

Batten, David F., and Dino Martellato. 1985. "Classical Versus Modern Approaches to Interregional Input–Output Analysis," *Annals of Regional Science*, 19, 1–15.

Boomsma, Piet, and Jan Oosterhaven. 1992. "A Double-Entry Method for the Construction of Bi-regional Input–Output Tables," *Journal of Regional Sciences*, 32(3), 269–284.

Bregman, Lev M. 1967. "Proof of the Convergence of Sheleikhovskii's Method for a Problem with Transportation Constraints," *USSR Computational Mathematics and Mathematical Physics*, 1(1), 191–204.

Byron, Ray P. 1978. "The Estimation of Large Social Account Matrix," *Journal of Royal Statistical Society*, A, 141(Part 3), 359–367.

Byron, Ray P., Peter J. Crossman, Jim E. Hurley, and S. C. Smith. 1993. "Balancing Hierarchial Regional Accounting Matrices," Paper presented to the International Conference in memory of Sir Richard Stone, National Accounts, Economic Analysis and Social Statistics, Siena, Italy.

Golan, Amos, George Judge, and Sherman Robinson. 1994. "Recovering Information From Incomplete or Partial Multisectoral Economic Data," *The Review of Economics and Statistics*, LXXVI(3), 541–549.

Harrigan, Frank J. 1990. "The Reconciliation of Inconsistent Economic Data: the Information Gain," *Economic System Research*, 2(1), 17–25.

Harrigan, Frank J., and Iain Buchanan. 1984. "A Quadratic Programming Approach to Input–Output Estimation and Simulation," *Journal of Regional Science*, 24(3), 339–358.

Isard, Walter, Iwan Azis, Matthew P. Drennan, Ronald E. Miller, Sidney Saltzman, and Erik Thorbecke (eds.). 1998. *Methods of Interregional and Regional Analysis*. New York: Ashgate Publishing Company.

Jensen, Rodney C. 1990. "Construction and Use of Regional Input–Output Models: Progress and Prospects," *International Regional Science Review*, 13(1 & 2), 9–25.

Lahr, Michael L. 2001. "A Strategy for Producing Hybrid Regional Input–Output Tables," in M. Lahr and E. Dietzenbacher (eds.), *Input–Output Analysis: Frontiers and Extensions*. Basingstoke, UK: Palgrave, pp. 211–242.

Lahr, Michael L., and Louis de Mesnard. 2004. "Bi-proportional Techniques in Input–Output Analysis: Table Updating and Structural Analysis," *Economic System Research*, 16(2), 115–134.

McDougall, Robert A. 1999. "Entropy Theory and RAS are Friends," Paper presented at the 5th conference of Global Economic Analysis, Copenhagen, Denmark.

McDougall, Robert A., Aziz Elbehri, and Truong P. Truong. 1998. "Global Trade Assistance and Protection: The GTAP 4 Database," Center for Global Trade Analysis, Purdue University.

Miller, Ronald E., and Peter D. Blair. 1985. *Input–Output Analysis: Foundations and Extensions*. Englewood Cliffs, NJ: Prentice Hall.

Mohr, Malte, William H. Crown, and Karen. R. Polenske. 1987. "A Linear Programming Approach to Solving Infeasible RAS Problems," *Journal of Regional Sciences*, 27(4), 587–603.

Nagurney, Anna, and Alan G. Robinson. 1989. "Equilibration Operators for the Solution of Constrained Matrix Problems," Working Paper, OR 196–89, Operations Research Center, MIT.

Ploeg, van der F. 1982. "Reliability and the Adjustment of Sequences of Large Economic Accounting Matrices," *Journal of the Royal Statistical Society*, A, 145, 169–194.

———. 1984. "General Least Squares Methods for Balancing Large Systems and Tables of National Accounts," *Review of Public Data Use*, 12, 17–33.

———. 1988. "Balancing Large Systems of National Accounts," *Computer Science in Economics and Management*, 1, 31–39.

Polenske, Karen R. 1980. *The U.S. Multiregional Input–Output Accounts and Model*. Lexington, MA: Lexington Books.

Robinson, Sherman, Andrea Cattaneo, and Moataz El-Said. 2001. "Updating and Estimating a Social Accounting Matrix Using Cross Entropy Methods," *Economic System Research*, 13(1), 47–64.

Senesen, Gulay, and John M. Bates. 1988. "Some Experiments with Methods of Adjusting Unbalanced Data Matrices" *Journal of the Royal Statistical Society*, A, 151(Part 3), 473–490.

Stone, Richard. 1984. "Balancing the National Accounts. The Adjustment of Initial Estimates: A Neglected Stage in Measurement," in A. Ingham and A.M. Ulph (eds.), *Demand, Equilibrium and Trade*. London: Macmillan.

Stone, Richard, John M. Bates, and Michael Bacharach. 1963. *A Program for Growth, Vol. 3 Input-Output Relationship 1954–1966*, London: Chapman and Hall.

Stone, Richard, David G. Champernowne, and James E. Meade. 1942. "The Precision of National Income Estimates," *Review of Economic Studies*, 9(2), 110–125.

Trendle, Bernard. 1999. "Implementing A Multi-Regional Input–Output Model – The Case of Queensland," *Economic Analysis & Policy*, Special Edition, 17–27.

Weale, Martin R. 1985. "Testing Linear Hypotheses on National Account Data," *Review of Economics and Statistics*, 67, 685–689.

———. 1989. "Asymptotic Maximum-Likelihood Estimation of National Income and Expenditure," mimeo, Cambridge. World Bank Policy Research Working Paper 3359, July 2004.

West, Guy R. 1990. "Regional Trade Estimation: A Hybrid Approach," *International Regional Science Review*, 13, 103–118.

Wilson, Alan G. 1970. "Inter-regional Commodity Flows: Entropy Maximizing Approaches," *Geographical Analysis*, 2, 255–282.

Zenios, Stavros A., Arne Drud, and John M. Mulvey. 1989. "Balancing Large Social Accounting Matrices with Nonlinear Network Programming," *NETWORKS*, 19, 569–585.