Chapter 3

Targeting the Collection of Superior Data for the Estimation of Regional Input-Output Tables

3.1  Introduction

The construction of survey-based input-output (IO) tables is a costly and time-consuming endeavor. Typically, there is a substantial time lag between the actual survey and compilation, and the publication of the IO tables. As a consequence, researchers often have to work with tables that are several years old. Therefore, the earlier literature has paid a lot of attention on updating IO tables via non-survey techniques, where intermediate deliveries between industries from the most recent available table are updated by various modification techniques (Richardson, 1972; Allen and Gossling, 1975; Lecomber, 1975; Round, 1983; Snower, 1990; de Mello and Teixeira, 1993; Bonfiglio and Chelli, 2008). However, most of the methods that are usually adopted, do not take any additional information on the coefficients of the target year into account. Such information may seriously improve the quality of the estimates (Lecomber, 1964, 1975; Allen, 1974; Bullard and Sebald, 1977; Hewings and Janson, 1980; Morrison and Thumann, 1980; Hewings, 1984; Jalili, 2000). Therefore, so-called hybrid techniques have become mainstream for the estimation of the intermediate deliveries in an IO table (see, for example, Brucker et al., 1987, 1990; Lahr, 1993). They combine non-survey techniques with superior data. Such data may be obtained from experts, surveys, and other reliable (primary or secondary) sources.

Since collecting superior data usually requires time and money, a crucial question is for which intermediate deliveries additional information should be gathered. In this respect, a lot of attention has been paid to selecting (sets of) individual cells of the table.

1 This chapter is an extension of Jiang et al. (2010).
matrix with intermediate deliveries (or the matrix of input coefficients). Probably the best-known approach is to identify cells that are most critical to the accuracy of the Leontief inverse or the gross outputs (Jensen and West, 1980; Hewings and Romanos, 1981; West, 1981; see Casler and Hadlock, 1997, for a detailed review).

There is some doubt, however, whether this cell-by-cell approach is the appropriate way to deal with this problem. Surveying cells in the same sector is likely to be more cost-effective, when compared to surveying the same number of cells belonging to different sectors (Lahr, 2001). It has been established that it is more difficult to obtain data for individual cells than for sectors, either row-wise or column-wise (see Miernyk, 1970; Isard and Langford, 1971; Bourque, 1971; Conway, 1975; Afrasiabi and Casler, 1991).

This chapter presents an empirical evaluation of the two approaches (i.e. focusing on individual cells versus focusing on sectors) for selecting cells for superior data collection, which—to our knowledge—has never taken place. For each of 27 provinces in China, the intermediate deliveries matrix of the 2002 regional IO table will be estimated. Because the regional IO tables are available for both 1997 and 2002, we mimic the updating process for each region in 2002. First, the 2002 matrices will be estimated by applying RAS to the corresponding 1997 matrix, given the actual 2002 values for the margins. This is the case with no superior data. Next, we study the case where for a certain number of cells in the intermediate deliveries matrix—namely the ones that have been targeted—superior data is available. That is, the “true” 2002 values for these intermediate deliveries will be inserted and the remaining part of the intermediate deliveries matrix will be estimated using the adjusted RAS procedure, based on the values from the 1997 matrix and given the actual 2002 values for its margins. The accuracy of estimating the intermediate deliveries matrix can be obtained from comparing the estimate for 2002 with the actual 2002 matrix. For the selection of cells for superior data collection, six different approaches will be used and evaluated.
The plan of the chapter is as follows. Section 3.2 summarizes the six approaches to select the cells for collecting superior data. We will use two approaches for identifying “important coefficients” (i.e. focusing on individual cells), two approaches for identifying “key sectors” row-wise, and two approaches for identifying “key sectors” column-wise. Section 3.3 discusses our simulation of the estimation procedure and the criteria that will be employed to compare the six approaches. Section 3.4 presents the results and Section 3.5 provides further discussions.

3.2 Approaches to Select Cells for Superior Data Collection

According to Lahr (2001), the basic steps in applying the hybrid technique are: (i) the preparation of an initial non-survey intermediate deliveries matrix; (ii) identifying sectors or individual cells for superior data collection; (iii) insertion of the superior data; and (iv) balancing using RAS. For the identification of cells for superior data collection, there are broadly two kinds of approaches. The first focuses on individual cells of the intermediate deliveries matrix and we will use two alternative techniques, i.e. selecting the largest direct input coefficients (Section 3.2.1) and selecting the so-called “inverse important coefficients” (Section 3.2.2). In both cases, the actual 2002 intermediate deliveries will be inserted for the selected cells. The second set of approaches focuses on targeting entire sectors, the so-called “key sectors”. Following the literature on linkages, all intermediate deliveries in either the selected sector’s column or its row will be replaced by the actual 2002 values. The techniques we will use are based on the column-sums (Section 3.2.3) and the row-sums (Section 3.2.4) of the Leontief inverse, and on the effect of hypothetically extracting entire columns (Section 3.2.5) or rows (Section 3.2.6) from the direct input matrix.

3.2.1 Large Input Coefficients

According to the standard IO model, we have

\[ x = Ax + f \]  
(3-1)
where the input coefficients in matrix $A$ are defined as $a_{ij} = \frac{z_i}{x_j}$, with $z_i$ the intermediate deliveries (or inputs) of product $i$ (which includes—in Chinese IO tables—imports from other regions and from abroad) to sector $j$. Vectors $x$ and $f$ indicate gross outputs and final demands, respectively. Some researchers (see Jensen et al., 1979; Jensen and West, 1980; Israilevich, 1986) have suggested that the largest input coefficients $a_{ij}$ exert the most influence on the Leontief inverse matrix and hence the IO multipliers. That is, a change in a larger coefficient will result in a larger overall change in output than would a similar change in a smaller coefficient. The logical consequence of this is that one should select the cells corresponding to the largest input coefficients.

Specifically, if $p$ is the number of cells that are selected for superior data collection, this approach simply selects the $p$ largest elements of the matrix $A$. In the results section this approach will be indicated as ‘LARGE’.

It is interesting to note that the regional statistics bureaus of China have used this selection criterion when compiling hybrid IO tables between two survey years. That is, every five years (including 1997 and 2002), the regional statistics bureaus compile IO tables using the survey method. In the mid-period, hybrid or semi-surveyed tables are constructed. Superior data are collected for cells that had the largest input coefficients in the last survey year. For the 2000 table, for example, the number of cells selected for superior data collection (i.e. $p$) was determined such that the corresponding deliveries in 1997 accounted—in total—for 70% of the all the intermediate deliveries (NBS, 2005).

### 3.2.2 Inverse Important Coefficients

The solution to the IO model in Equation (3-1) is given by

$$x = (I - A)^{-1}f = Bf$$

(3-2)
where $b_{ij}$ are elements of the Leontief inverse $B = (I - A)^{-1}$, with $I$ the identity matrix. Some researchers have defined important coefficients as those for which a change affects the inverse matrix $B$ or total output $x$ the most (see, for example, Morrison and Thumann, 1980; Hewing and Romanos, 1981; Jackson, 1986; Bullard and Sebald, 1988; Sonis and Hewings, 1989, 1992). At the basis of the identification of these so-called inverse important coefficients was the Sherman-Morrison equation (Sherman and Morrison, 1950), which addressed directly the changes induced in the elements of an inverse matrix by a change in one element of the original matrix. That is,

$$
\Delta b_{ij} = \frac{b_{ij} \Delta a_{ij}}{1 - b_{ij} \Delta a_{ij}}
$$

(3-3)

where $\Delta a_{ij}$ denotes the change of the input coefficient $a_{ij}$, $\Delta b_{ij}$ denotes the change in Leontief inverse coefficients $b_{ij}$ corresponding to the change $\Delta a_{ij}$.

There are several ways to measure the effect of a single coefficient change $\Delta a_{ij}$ in IO models (see Xu and Madden, 1991, for an overview). In this chapter, we adopt the systemwide perspective and measure the effect of a single coefficient’s change on the totality of the Leontief inverse matrix, i.e. $\Sigma, \Sigma, b_{ij}$. Also Liu (2004) has used this criterion to select the most important 15% of the input coefficients $a_{ij}$. She reported considerable gains in accuracy for Chinese national IO table construction, when these 15% of all coefficients were obtained by using econometric techniques combined with exogenous information.

We select the $p$ cells $(k, l)$ for which a 0.0001 change in $a_{ij}$ (i.e. $\Delta a_{ij} = 0.0001$) generates the largest value of
In the results section this approach will be indicated as ‘INVIMP’.  

It should be mentioned that the outcomes depend on the specific choice of the change. On the one hand, our choice \( \Delta a_{ij} = 0.0001 \) implies that the changes are negligibly small for the larger coefficients. On the other hand, however, much larger changes would be unrealistic for the very small coefficients. Our alternative calculations with \( \Delta a_{ij} = 0.01 \), however, led to a selection of inverse important coefficients that is very close to the one obtained for \( \Delta a_{ij} = 0.0001 \). This is in line with the findings of Xu and Madden (1991) that the set of inverse important coefficients largely remained the same.

3.2.3 Leontief Inverse – Column-sums

Unlike targeting individual cells of the input matrix, the targeting of key sectors raises the question whether to focus on surveying rows or columns. When superior data are to be gathered, this implies a choice between focusing on either patterns of purchases (i.e. column-wise) or patterns of sales (row-wise) within sectors. The first type is usually preferred because, as Qi (2007) pointed out, firms are—as a rule—much better informed about the sectoral origins of their inputs (i.e. their purchases) than about the sectoral destinations (i.e. sales) of their outputs. Most tables constructed in both the USA (see Richardson, 1985; Ralston et al., 1986; Brucker et al., 1990) and in Australia (see West, 1990), use information for purchases data to construct regional input-output tables. Nevertheless, also the use of sales data has been encouraged in the literature (Hansen and Tiebout, 1963; Lee et al., 1973; Schaffer, 1976; Richardson, 1985; West, 1990). That is, it has been argued that, in a spatial context, better information exists for the spatial destination of sales than for the spatial origin of the inputs (Isard and Langford, 1971; Boomsma and Oosterhaven, 1992). In our application to the Chinese regional IO tables, both aspects are included.
Targeting the collection of Superior data

The question “what sector is the most important”, is not a new one. Rasmussen (1957) and Hirschman (1958) were amongst the first to answer this question using the elements of the Leontief inverse. For any sector $j$, $\Sigma b_j$ gives the sum of the elements in the $j$th column (i.e. the $j$th column-sum) of the Leontief inverse $B$. It has become known as the sum of the backward linkages of sector $j$ or as sector $j$’s output multiplier. It gives the total increase in the outputs due to an increase of one unit in the final demand of sector $j$.

If $q$ is the number of sectors, whose columns are selected for superior data collection, this approach simply selects the $q$ sectors with the largest output multipliers (or column-sums of the Leontief inverse). In the results section this approach will be indicated as ‘COLSUM’.

3.2.4 Leontief Inverse – Row-sums

Alternatively, also the row-sums of Leontief inverse have been frequently used to define the key sectors. That is, the $i$th row-sum ($\Sigma j_i$) can be interpreted as the output increase in sector $i$ due to a one-unit increase in the final demands of each sector (which is related to Rasmussen’s “Index of sensitivity of dispersion”). The underlying idea is that a larger row-sum indicates that sector $i$ is more important in the economy in case of a general expansion of final demands. If $q$ sectors are selected for superior data collection, this approach selects the $q$ sectors with the largest row-sums of the Leontief inverse. In the results section this approach will be indicated as ‘ROWSUM’.

3.2.5 Hypothetical Extraction of Columns

The work by Rasmussen (1957) and Hirschman (1958) on linkages, led to a lively discussion on key sectors and many approaches have been proposed afterwards (see, for example, Miller, 1966; Strassert, 1968; Schultz, 1976, 1977; Meller and Marfan, 1981; Loviscek, 1982; Cella, 1984; Szyrmer, 1984, 1986, 1992; Harrigan and
McGilvray, 1988; Clements, 1990; Dietzenbacher and van der Linden, 1997). One of them was the hypothetical extraction method, developed in Paelinck et al. (1965) and revised and refined by many others (see Miller and Lahr, 2001, for an excellent and detailed overview). For the Chinese economy, Andreosso-O’Callaghan and Yue (2004) reported that the hypothetical extraction method provided superior results for the identification of key sectors when compared to traditional methods, such as the column-sums of the Leontief inverse. Therefore, we will also employ the hypothetical extraction method in our case of Chinese regions.

Strassert’s (1968) version of the hypothetical extraction method compares the actual production with the production in the hypothetical case where all intermediate deliveries to and from a particular sector are extracted. Dietzenbacher and van der Linden (1997) make a distinction between backward and forward linkages by extracting the cells that correspond to the purchases, respectively sales, of a sector. In the case of the backward linkages, the extraction of the column of sector \( k \)'s purchases leads to a new input matrix \( A^{-k} \) with \( a_k^{i,j} = 0 \) for all \( i \) and \( a_k^{i,j} = a_{ij} \) for all \( i \) and \( j \neq k \). This means that column \( k \) of the input matrix is set equal to zero, implying that sector \( k \) does not depend on any domestic industry for its inputs (which reflects the backward nature). The outputs then become \( x^{-k} = (I - A^{-k})^{-1} f \) and the total backward linkages of sector \( k \) are measured by \( \sum_i [x_i - x_i^{-k}] \). If \( q \) is the number of sectors, whose columns are selected for superior data collection, this approach simply selects the \( q \) sectors with the largest outcome for \( \sum_i [x_i - x_i^{-k}] \). In the results section this approach will be indicated as ‘COLHYP’.

3.2.6. Hypothetical Extraction of Rows

In case of forward linkages, exactly the same procedure is adopted. The only difference is that the matrix \( A^{k-} \) now has its \( k \)th row set to zero (instead of its \( k \)th column). That is, \( a_{ik}^{i,j} = 0 \) for all \( j \) and \( a_{ik}^{i,j} = a_{ij} \) for all \( j \) and \( i \neq k \). Again, the \( q \)
sectors with the largest outcome for $\sum (x_i - x^{(i-1)})$ are the key sectors, but now their rows are selected for superior data collection. In our application, we will indicate this approach as ‘ROWHYP’.

### 3.3 Outline of the Simulation Experiment

China has a long tradition in compiling IO tables, not only at the national level, but also at the provincial level. Yet, Jiang et al. (2007) found that traditional compiling methods, which work very well in developed countries, are not suitable for the Chinese regional tables. Therefore, finding an appropriate way (in terms of accuracy and cost-effectiveness) to select cells for collecting superior data is highly relevant for China in particular.

In our application, we will mimic the real world estimation of the Chinese regional IO tables for 2002. The regional statistics bureaus in China have collected IO tables for 30 out of the 31 provinces except Tibet (Liu and Wu, 1991). For our experiment we were able to use the tables for 27 provinces (the tables for Hainan province and for the autonomous regions of Qinghai and Xinjiang were not available). Our dataset covers tables for 1997 (using a 40-sector classification) and for 2002 (using a 42-sector classification). We aggregated the tables to 31 sectors to arrive at a uniform classification.\(^2\) For each of the 27 provinces we will estimate the intermediate deliveries in 2002, using each of the six methods (described in Section 3.2) to select cells for superior data collection. Section 3.3.1 describes the estimation procedure and Section 3.3.2 discusses how the performance of the six methods will be compared.

---

\(^2\) Note that the aggregated regional IO tables we use in this chapter are identical to the ones we used in Chapter 2. See Appendix 2A for the sector classification.
3.3.1 The Estimation Procedure

For a certain province, we will take the 1997 matrix of intermediate deliveries as a starting point. Applying each of the six methods (described in Section 3.2) to the 1997 table yields a set of cells for which superior data are to be collected.

In practice, the number of cells for superior data collection is usually predetermined subjectively by the researchers or organizations and depends to a large extent on the amount of resources and funds available to carry out the task (Jalili, 2000). The choice for the number of selected cells, no doubt, is a critical part of the estimation process and affects the results, conclusions and recommendations derived from the estimated matrices. Therefore, our application takes several possibilities into account. The IO tables that we use are 31×31 and we would like to compare the performance of selecting entire sectors with the performance of selecting equally sized sets of individual cells. This implies that the numbers of targeted cells for our experiment are proportional to the number of cells in one sector (i.e. 31). In the application, the number of targeted cells equals \( p = 0, 31, 62, 93, \ldots, 248 \), which is 0.0%, 3.2%, 6.5%, 9.7%, \ldots, 25.8% of all cells in the matrix. The case with \( p = 0 \) indicates that no superior data have been collected and that the update has been obtained by applying RAS only. This case is presented as a frame of reference. When selecting sectors, we thus have \( q = 0, 1, \ldots, 8 \) (note \( p = 31q \)). Although the number of selected cells is the same across methods, it should be borne in mind that surveying cells of a given column or row (i.e. a sector) is much easier than surveying individual cells that probably belong to several sectors (Lahr, 2001).

After the cells have been selected, their corresponding intermediate deliveries are replaced by superior data. We assume that the superior data is 100% correct (i.e. the maximum amount of superior information) and use the 2002 deliveries to replace the 1997 figures. That is, if cell \((k, l)\) is selected then the true value for \( z_{kl} \) in 2002 is assumed to be known and inserted. Finally, the remaining part of the 1997 intermediate deliveries matrix is updated by applying the adjusted RAS approach (see
Miller and Blair, 1985). This approach is appropriate to cover cases of additional exogenous information, e.g. if certain intermediate deliveries are known. When applying the RAS approach, we assume fully correct information again and take the margins (i.e. row- and column-sums of the intermediate deliveries matrix) in 2002 as given.

We thus obtain estimates for the 2002 intermediate deliveries matrix for region \( r \) (= 1, ..., 27), for each targeting method \( m \) (= 1, ..., 6), where the number of targeted cells is \( p \) (= 0, 31, 62, 93, ..., 248). This yields 1323 estimated matrices of size 31\( \times \)31, including 27 estimates for which no superior data (i.e. \( p = 0 \)) have been used.

### 3.3.2 Evaluation Criteria

The accuracy of the estimation methods will be evaluated at three levels. First of all, it will be checked how accurate the intermediate deliveries have been estimated since that was the original goal. Second, when IO tables are used in policy studies and impact analyses, all calculations require the use of the Leontief inverse. For applied work, one might argue that an accurate estimation of the Leontief inverse is more relevant than an accurate estimation of the intermediate deliveries themselves. Third, because the Leontief inverse is obtained from the matrix of input coefficients, we will also evaluate the accuracy of the estimation of the input coefficients matrix.

Denote the true matrix of intermediate deliveries in 2002 by \( Z \) and its estimate by \( \hat{Z} \). The overall deviation of the estimate and the true matrix will be measured by the weighted absolute percentage error (WAPE, Leontief, 1966), as same as the indicator which chapter 2 employed. The WAPE gives the average percentage error where errors in large cells receive a larger weight than errors in small cells (Oosterhaven et al., 2008). It is defined as
Here $z_i$ and $\hat{z}_i$ denote the actual and estimated value of the intermediate deliveries. For the accuracy of estimating the input coefficients they are replaced by $a_i$ and $\hat{a}_i$, respectively, where $a_i = z_i / x_i$ and $\hat{a}_i = \hat{z}_i / x_i$ (note that the outputs in 2002 are assumed to be known correctly). For the accuracy of the Leontief inverse, the situation is slightly different because its diagonal elements are never smaller than one by definition. Using the power series expression, i.e. $B = (I - A)^{-1} = I + A + A^2 + A^3 + \ldots$, it becomes clear that a major part of the diagonal elements involves no estimation at all. The estimated part is given by $B - I$.

Hence, $z_i$ and $\hat{z}_i$ are replaced by $b_i - \delta_i$ and $\hat{b}_i - \delta_i$, respectively, where $\delta_i$ is the Kronecker delta (i.e. $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise).

### 3.4 The Results for Estimating the Chinese Regional Input-Output Tables in 2002

#### 3.4.1 Average WAPEs

Our calculations yield values for $WAPE_m(p)$, where subscript $m (=1, \ldots, 6)$ represents the method to determine the cells for superior data collection, superscript $r (=1, \ldots, 27)$ indicates the region, and $p (=0, 31, 62, 93, \ldots, 248)$ gives the different numbers of cells for superior data collection. Figure 3.1 gives the average WAPE, where the average is taken over all 27 regions. That is, $\overline{WAPE}_m(p) = \frac{\sum_{m} WAPE_m(p)}{27}$. We distinguish between the average WAPE of the intermediate deliveries matrices (in graph A), of the input coefficients matrices (in graph B), and of the Leontief inverses (in graph C). When comparing the different methods, a lower curve exhibits a higher accuracy of the estimates.
Targetting the collection of Superior data

Figure 3.1: Average WAPEs of the estimation

A. Accuracy in terms of the intermediate deliveries matrix

B. Accuracy in terms of the input coefficients matrix

C. Accuracy in terms of the Leontief inverse
In terms of estimation accuracy, the graphs in Figure 3.1 show that LARGE invariably performs the best. The two methods that focus on the rows when searching for key sectors perform second best (except in graph A), where ROWSUM performs slightly better than ROWHYP. Graphs B and C indicate that the other three methods (INVIMP, COLSUM and COLHYP) perform the least. In particular the poor results of INVIMP in graph C are surprising. Recall that the true values have been used for cells for which a small error in its coefficients had the largest effect on the Leontief inverse. One thus would have expected INVIMP to perform better than it does now, at least in estimating the Leontief inverse.

Graphs B and C sketch a very clear distinction of performance into three categories and COLHYP belongs to the category with the worst performance (it even is the worst of all three methods in that category). In contrast, however, COLHYP performs second best when estimating the intermediate deliveries (in graph A). We argue that this is caused by the fact that COLHYP is the only method that takes sector size (in terms of output) into account, which matters for the intermediate deliveries (in graph A) but not for input coefficients or the Leontief inverse (in graphs B and C). Consider the definition of the WAPE in Equation (3-5). It can be written as the weighted average of the WAPEs within a column. That is, for the WAPE of the intermediate deliveries $z_{ij}$, we have

$$WAPE(z_{ij}) = 100 \times \frac{\sum j (\sum _{i} \frac{\hat{z}_{ij} - z_{ij}}{z_{ij}})}{\sum j (\sum _{i} \frac{\hat{z}_{ij}}{z_{ij}})} = 100 \times \frac{\sum j (\sum _{i} \frac{\hat{z}_{ij} - z_{ij}}{z_{ij}})}{\sum j (\sum _{i} \hat{z}_{ij})} = \frac{\sum j (\sum _{i} \hat{z}_{ij}) \text{COLWAPE}(z_{ij})}{\sum j (\sum _{i} \hat{z}_{ij})}$$

(3-6)

with
Targetting the collection of superior data

\[
COLWAPE(z_j) = 100 \times \frac{\sum_i z_i - \bar{z}_j}{\sum_i z_i}
\]

Note that \(COLWAPE(z_j) = COLWAPE(a_j)\), because \(z_j = a_j x_j\) and thus \(\sum_i z_i = x \sum_i a_i\). This means that we can write expression (3-6) as

\[
WAPE(z_j) = \frac{\sum_i x_i (\sum_i a_i) COLWAPE(a_i)}{\sum_i (\sum_i a_i)} \tag{3-7}
\]

whereas the WAPE of the input coefficients is given by

\[
WAPE(a_j) = \frac{\sum_i (\sum_i a_i) COLWAPE(a_i)}{\sum_i (\sum_i a_i)} \tag{3-8}
\]

The implication is that columns corresponding to sectors with a large output \(x_j\), receive a much larger weight in Equation (3-7) than in Equation (3-8). Dietzenbacher and van der Linden (1997) observed that the standard hypothetical extraction method favors large sectors (in terms of output) when finding the key sectors that have the largest backward linkages. This is exactly what happens when COLHYP is used to select the columns for which the actual values are inserted. That is, for the sectors whose columns receive the largest weight, the columns are free of errors. Therefore, it should come as no surprise that COLHYP performs relatively well in graph A (i.e. for intermediate deliveries, where output size does play an important role).

A further observation is that the results for graph B (for input coefficients) and graph C (for the elements of the Leontief inverse) are very similar. Recall that graph C is based on the WAPEs for the matrices \(B - I = A + A^2 + A^3 + \ldots\). For many countries and regions, it has been found that the behavior of \(B - I\) is largely dominated by the behavior of \(A\), because the higher order terms \((A^2, A^3, \ldots)\) decline
very fast. Consequently, the accuracy of $B - I$ will to a large extent be determined by the accuracy of $A$.

A final observation is with respect to the slopes of the curves in Figure 3.1. As expected, all graphs are monotonically declining. The speed of decline in average error (or increase in accuracy) may vary though. For example, in the case of ROWSUM, the increase in accuracy slows down when the number of cells with superior data is increased from 93 to 124, but rises again when the number of selected cells is further increased to 155. For LARGE, it is observed that its slopes are larger than the slopes for the other methods. The differences between the methods, however, decline. That is, for the gain in accuracy when going from 217 to 248 cells with superior data, it does not matter very much which method is used to select the cells for superior data insertion. This is not a surprising result, since the average WAPE will be zero if superior information is inserted for all cells (i.e. $p = 961$), irrespective of the targeting method used. The interesting part of the graphs is on the left-hand side, which shows that the curves for LARGE are the steepest. It is clear that the major advantage of LARGE is in selecting the first cells for which superior data are collected.

### 3.4.2 Coefficients of Variation

In order to obtain some insight into the stability of the accuracy across regions, we have calculated—for each of the methods—the coefficients of variation. However, this may yield misleading results. Because of strong regional disparities across China, the accuracy varies also in the case without superior data (i.e. if $p = 0$). A large variation of accuracies in cases with $p 
eq 0$ then raises the question whether this should be attributed to regional disparities or to the specific method (which is what we would like to know). In order to filter out the first source of variation, we have calculated relative WAPEs. That is, $RW\text{APE}_r(p) = \frac{WAPE_r(p)}{WAPE_r(0)}$ and it expresses the gain in accuracy due to superior data (relative to the accuracy in the case without superior data). Note that $RW\text{APE}_r(0) = 1$ and the relative WAPE is...
Targeting the collection of Superior data

expected to decline if $p$ increases. Moreover, the lower $RWAPE'_{s}(p)$ is, the larger is the gain in accuracy due to adding superior data. The stability is given by

$$CV_{RWAPE'_{s}(p)} = \frac{\text{StDev}[RWAPE'_{s}(p)]}{RWAPE'_{s}(p)}$$

(3.9)

Note that $CV_{RWAPE'_{s}(p)} = 0$ and is expected to increase when $p$ becomes larger. A “small” $CV$ indicates that the gain (from adding superior data) in accuracy is more or less the same for all regions, while a large $CV$ indicates that this gain differs largely across regions. In practical work, the ideal is to work with a method that is not only very accurate (i.e. a “small” $WAPE_{s}(p)$), but that is also very reliable. That is, ideally its performance should be good and it should be good in each and every region (and hence exhibit a “small” $CV_{RWAPE'_{s}(p)}$).

Figure 3.2 gives the coefficients of variation corresponding to the six methods, for the intermediate deliveries (graph A), the input coefficients (graph B), and the Leontief inverse (graph C). There are many similarities with Figure 3.1, such as the clear distinction between graph A on the one hand and graphs B and C on the other hand. Just like in Figure 3.1, the graphs B and C are very much alike. In the graphs B and C, we find that the largest CVs are found for LARGE, that ROWSUM and ROWHYP show smaller CVs, while the smallest CVs are found for COLSUM, COLHYP and INVIMP. This is the same partitioning of methods as we found in Figure 3.1. This suggests that poor performance is invariably poor across regions (i.e. the lowest accuracy is coupled to the largest stability). Vice versa, the best performance in terms of accuracy goes together with the largest variability across regions.
Figure 3.2: Variation of the WAPEs across regions

A. Estimating the intermediate deliveries matrix

B. Estimating the input coefficients matrix

C. Estimating the Leontief inverse
Graph A shows that also COLHYP exhibits large CVs (which are only slightly less than those for LARGE). This corresponds to the finding in Figure 3.1 that COLHYP exhibits the second largest accuracy. So also in this case a larger accuracy is coupled to a lower stability. Another observation in graph A of Figure 3.2 is that the difference in stability between ROWSUM and ROWHYP has become pretty large (for certain percentages of superior data, ROWHYP’s CVs are even the largest). In graph A of Figure 3.1 we found that ROWSUM performs a little bit better than ROWHYP and Figure 3.2 shows that ROWSUM is also more stable in its performance across regions than ROWHYP is (and sometimes the difference is substantial). Typically, even if the average performance is the same for two methods, one would prefer the method that is most stable. So, ROWSUM is clearly preferred to ROWHYP when estimating the intermediate deliveries matrix.

3.4.3 Evaluation of Results

In evaluating the results, the first relevant finding is that using the large coefficients (LARGE) to select cells that are replaced by their actual 2002 values is the most successful. Although this is an extremely simple method to apply, it invariably performs best. A much more sophisticated method that selects cells on the basis of inverse important coefficients (INVIMP) performs in the worst category. This poor performance holds also if the estimation of the Leontief inverse is evaluated. This is somewhat surprising because INVIMP is based on the sensitivity of the Leontief inverse with respect to a change (of 0.0001) in a single coefficient. Inverse important coefficients are those that generate the largest effect once they are changed. The implication of our empirical findings is that the inverse important coefficients apparently are estimated relatively well by their 1997 values, whereas other coefficients (to which the Leontief inverse may be less sensitive) are estimated with much greater error.

The second finding is that for the estimation of the input coefficients matrix and the Leontief inverse, focusing on the rows corresponding to the key sectors
(ROWSUM and ROWHYP) yields much better results than focusing on the columns of the key sectors (COLSUM and COLHYP). For determining the key sectors, the simple row-sums of the Leontief inverse outperform the more complex hypothetical extraction method (exhibiting a larger accuracy and stability). For the estimation of the intermediate deliveries, however, the best estimation is obtained if superior data are inserted into the columns that have been selected by the hypothetical extraction method (i.e. COLHYP). In most applied IO work, in particular policy studies and impact analyses, the intermediate deliveries themselves are of little interest, what matters is the Leontief inverse which is at the heart of all calculations.

All in all, selecting the large coefficients (LARGE) works best if a set of individual cells is selected for superior data collection. If entire sectors are selected for superior data collection the simple row-sums of the Leontief inverse (ROWSUM) is the best method of selecting the key sectors, provided that one is interested in the input coefficients matrix or the Leontief inverse. In case one is interested in the intermediate deliveries themselves, selecting columns by means of the hypothetical extraction method (COLHYP) works best. In comparing these two sets of results, let us focus on the accuracy of the Leontief inverse, given its importance in practical work. Although LARGE clearly outperforms ROWSUM, it should be remembered that collecting superior data for individual cells in different sectors is much more expensive than data collection for entire sectors. From graph C in Figure 3.1 it follows that the average WAPE of LARGE for a given number of cells \( p \) with superior data is more or less the same as the average WAPE of ROWSUM for \( 3p \). This means that both methods perform roughly the same if LARGE is three times as expensive as ROWSUM and the budget for superior data collection is fixed. If LARGE is more than three times as expensive, selecting the key sectors from ROWSUM is the preferred method.
3.5 Further Discussions

3.5.1 Links between the Methods of Selecting Superior Data and their Results

An interesting aspect about our results is that the row methods work clearly better than the column methods (unless the intermediate deliveries are the ultimate goal of the estimation). The question is why this is the case. Table 3.1 gives the average number of overlapping cells, which have been selected for superior data collection. The amount of overlap is calculated for each pair of methods, setting the number of selected cells at 248 (or 8 sectors). For example, 61 in row 1 and column 2, means that LARGE and INVIMP have 61 identical cells in their sets of 248 cells that have been selected. The numbers in Table 3.1 are averages over all the 27 regions.

<table>
<thead>
<tr>
<th></th>
<th>INVIMP</th>
<th>LARGE</th>
<th>COLSUM</th>
<th>COLHYP</th>
<th>ROWSUM</th>
<th>ROWHYP</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVIMP</td>
<td>248</td>
<td>61</td>
<td>66</td>
<td>151</td>
<td>63</td>
<td>119</td>
</tr>
<tr>
<td>LARGE</td>
<td>248</td>
<td>67</td>
<td>65</td>
<td>138</td>
<td>64</td>
<td>116</td>
</tr>
<tr>
<td>COLSUM</td>
<td>248</td>
<td>88</td>
<td>64</td>
<td>116</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COLHYP</td>
<td>248</td>
<td>64</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROWSUM</td>
<td></td>
<td>248</td>
<td>230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROWHYP</td>
<td></td>
<td>248</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 3.1, it follows that both row methods (ROWSUM and ROWHYP) have an overwhelming 93% overlap in cells. On the one hand, this explains why their performance is very close (note that only the right endpoints of the graphs are relevant). On the other hand, it suggests that the remaining 18 cells that they do not have in common are not extremely influential in terms of affecting the estimation. This latter implication may also explain why ROWHYP does not perform better than ROWSUM, although the overlap between LARGE and ROWHYP is clearly larger than the overlap between LARGE and ROWSUM. In any case, these two pairs exhibit large overlaps which is line with their second best performance. Note that the other three methods (INVIMP, COLSUM, and COLHYP) all have relatively few overlapping cells with LARGE, which explains their poor performance. Recall that
we had observed (in graph A of Figure 3.1) that for the intermediate deliveries themselves, COLHYP performs surprisingly well. Also INVIMP performs much better than in the other two cases (i.e. graphs B and C). This corresponds to the finding of a large overlap in cells between COLHYP and INVIMP. Finally, it should be mentioned that ROWHYP has a pretty large overlap with all other methods (where those with LARGE and ROWSUM are outstanding), an observation for which we have no explanation.

Figure 3.3: Pattern of important coefficients based on LARGE (national table for 1997)

Figures 3.3 and 3.4 provide the distribution of the 248 individual cells that have been selected for superior data collection by LARGE and by INVIMP. It should be stressed that these figures are obtained from calculations with the national table for 1997, using the same sector classification as we used for the regional tables. It was found that the 27 distributions from the regional tables are very similar to those in Figures 3.3 and 3.4, indicating that the use of the national table for the present
Targetting the collection of Superior data

purpose is quite representative. Hence, we only report the distribution for the national table and not for each of the 27 regions separately.

Figure 3.4: Pattern of important coefficients based on INVIMP (national table for 1997)

It is very clear that INVIMP in Figure 3.4 selects cells that can be grouped into columns. For example, only 26 of the 248 cells do not belong to the group that consists of sectors 1, 3, 11, 13, 15, 18, 21, 23, and 24. Observe that for three of these sectors (11, 13, 23) all elements in their columns have been selected individually (and 30 of the 31 column elements of sector 24). Figure 3.3 with the selection of cells obtained by applying LARGE is somewhat less outspoken. Still, it is clear that the selected cells often are in the same row. Sectors for which the corresponding row hosts 15 or more selected cells are 11, 21, 23, and 24 (note the overlap with the columns in Figure 3.4). There is no column that covers 15 or more selected cells. The central conclusion from Figures 3.3 and 3.4 is that cells selected by LARGE exhibit a pattern of rows, while those selected by INVIMP show a grouping into columns. This
finding perfectly matches our earlier observations on the overlap between the methods. Finally, observe also that LARGE selects 28 out of 31 diagonal elements, while these are not extremely relevant for INVIMP (which selects only 6 diagonal elements).

3.5.2 Robustness of the Results

In this chapter we have presented the results for six methods, using the WAPE for the evaluation of results. An interesting issue is to what extent the results are robust when alternative methods are used or when different evaluation criteria are employed. First, to determine the important coefficients, several researchers have used a different approach based on the concept of “tolerable limits” (see, for example, Schintke and Stäglin, 1985, 1988; Tarancón et al., 2008). In estimating the Chinese regional IO tables, the “tolerable limits” approach reports accuracies that are clearly larger than those of INVIMP, but smaller than those of LARGE.

Second, the row-sums of the Leontief inverse have been used as a measure for forward linkages in this chapter. Several authors (see, for example, Beyers, 1976; Jones, 1976; Dietzenbacher, 1992, 1997; Oosterhaven, 1981) have argued that forward linkages are better reflected by the row-sums of the inverse matrix in Ghosh’s supply-driven IO model (Ghosh, 1958). This Ghosh inverse is based on output coefficients rather than on input coefficients. For the present purpose of selecting sectors for row-wise superior data collection, we find that this “Ghoshian alternative” leads to less accurate estimates than using ROWSUM.

Third, we have discussed already that COLHYP performed well in estimating the intermediate deliveries (see graph A of Figures 3.1 and 3.2), because it favors large sectors when searching for key sectors. This was the reason for Dietzenbacher and van der Linden (1997) to introduce the so-called relative backward linkages, which correct for output size. Also in the present study we have run calculations with relative linkages (both in the backward case for columns and in the forward case for

3 See Appendix 3A for detailed information on this approach.
Targetting the collection of Superior data

101 rows). The results are similar to those for COLHYP and ROWHYP, but show a somewhat lower performance.

Fourth, instead of focusing on replacing either a full column or a full row in the intermediate deliveries matrix, it might be relevant to consider the option to replace (for targeted sector $k$) both its column and its row. In some cases, it seems reasonable that information on both sales and purchases is available. To this end, we have used the original hypothetical extraction method (see Strassert, 1968). When sector $k$ is extracted we have $A^{(-k)}$ with $a_{ik}^{(-k)} = 0$ for all $i$, $a_{ik}^{(-k)} = 0$ for all $j$, and $a_{ik}^{(-k)} = a_{ij}$ for all $i$ and $j \neq k$. Note that replacing the intermediate deliveries for $q (= 1, 2, 3, 4)$ sectors, implies that the number of superior data amounts to $q(2n-q)$ ($= 61, 120, 177, 232$, respectively). The results for this method (say HYP) are in the lowest performing category (comparable to COLHYP, COLSUM, and INVIMP) for the input coefficients matrix and for the Leontief inverse. For estimating the intermediate deliveries matrix, HYP is in the middle category and its performance is comparable to ROWHYP (and slightly worse than COLHYP). Hence, adding superior data for the column and row of one sector improves the accuracy of the estimated intermediate deliveries just as much as adding superior data for the rows of two separate sectors. In this case, it clearly is a matter of costs to determine which choice to make. 4

Fifth, next to using WAPE, all results have also been evaluated by the weighted absolute difference (WAD), which is defined as follows.

$$WAD = \frac{\sum_i \sum_j |z_{ij} - \hat{z}_{ij}|}{\sum_i \sum_j z_{ij}} \tag{3-10}$$

where $z_{ij}$ and $\hat{z}_{ij}$ are defined as in Section 3.3. A typical feature of the WAD is that it is in the same unit as the original data, e.g. in 10 thousands of RMB for the intermediate deliveries of the regional Chinese IO tables. The WAPE gives a mean...

4 In Appendix 3B, we report and discuss the WAPEs for five alternative methods.
error percentage and is thus independent of the unit of measurement. Another aspect is that the WAD is also dependent on the size of the economy that it evaluates. This happens for example when estimating the intermediate deliveries themselves. In that case, it may happen that the outcomes for the WAD may be substantially larger for a large region than they are for a small region. Therefore, we have chosen to focus on the WAPE, when presenting our results. The outcome of comparing the methods on the basis of the WAD when evaluating the input coefficients matrix and the Leontief inverse (the two cases where the WADs are not affected by size) was very similar to the outcome on the basis of the WAPEs.\(^5\)

Sixth, in addition we have also determined the degree of approximation, which we defined as the number of cells for which the absolute deviation is no more than 10% of the true value. That is, \(0.9 \times z_{ij} \leq \delta_{ij} \leq 1.1 \times z_{ij}\) for the intermediate deliveries and \(0.9 \times a_{ij} \leq \delta_{ij} \leq 1.1 \times a_{ij}\) for the input coefficients (both of which the same). In case of the Leontief inverse we have counted the number of cells for which \(0.9 \times (b_{ij} - \delta_{ij}) \leq \hat{b}_{ij} - \delta_{ij} \leq 1.1 \times (b_{ij} - \delta_{ij})\). The results for the intermediate deliveries (or the input coefficients) indicate that the number of well approximated cells increases almost by 31 in each step with the addition of superior data for the next 31 important cells (or for the next key sector). This suggests that the original (68) well approximated cells (i) are relatively unimportant and (ii) are not concentrated in either the columns or the rows of the key sectors. The results for the Leontief inverse, sketch the same picture as we have seen in Figures 3.1 and 3.2. That is, LARGE performing the best, followed by ROWSUM and ROWHYP (where, in contrast to earlier findings, ROWHYP does a tiny little bit better than ROWSUM), whereas the column methods (COLSUM and COLHYP) and INVIMP perform the worst.\(^6\)

The six robustness checks described above did not give us reason to modify the main conclusions derived in section 3.4.3. We found that using the large coefficients

\(^5\) Appendix 3C presents and discusses the results when WADs are used for the evaluations.

\(^6\) Appendix 3D gives the findings for the degrees of approximation.
Targeting the collection of Superior data

(LARGE) to select cells that are replaced by their actual 2002 values is the most
successful if a set of individual cells is selected for superior data collection. If entire
sectors are selected for superior data collection, the choice of targeting method
depends on the type of matrix in which the analyst is most interested. If the focus is
on estimating the matrix of intermediate input flows, selecting columns by means of
the hypothetical extraction method (COLHYP) works best. If the analyst is most
interested in producing an accurate matrix of input coefficients or Leontief inverse
matrix (containing output multipliers that are important for impact analyses), superior
data can best be gathered on the basis of the simple row-sums of the Leontief inverse
(ROWSUM).
References


Targetting the collection of Superior data


Rasmussen, P.N. (1957) Studies in Inter-Sectoral Relations (Amsterdam: North Holland).


Appendix

Appendix 3A. Calculation of Tolerable Limits

The tolerable limits approach identifies for each input coefficient how much it can change (as %) at most, under the restriction that no output changes more than $p\%$.

This change is the maximum variation margin for the input coefficient and is calculated as (see Schintke and Stäglin, 1985, p. 130):

$$r_{kl} = \frac{p}{a_{kl}} \left( \frac{b_{kl} p}{100} + \max_{i \in \Lambda} \frac{b_{ki} x_i}{x_k} \right) = \frac{p}{a_{kl}} \left( \frac{b_{kl} p}{100} \right)$$

(3A-1)

where $a_{kl}$ represents the input coefficient (input of sector $k$ per unit of output in sector $l$), $b_{kl}$ is the element $(k, l)$ of the Leontief inverse, $x_l$ and $x_k$ are the total output for sectors $l$ and $k$. So, once input coefficient $a_{kl}$ is changed by more than $r_{kl} \%$, there is at least one sector for which the output increases by more than $p\%$. The lower the maximum variation margin $r_{kl}$ is, the more important is the input coefficient $a_{kl}$. In the empirical application (as reported in Appendix 3B), we have chosen $p = 0.01$. 


Appendix 3B. The Results of Alternative Methods

In this appendix, we present the results for five alternative identification methods. The first uses important coefficients based on the tolerable limits approach as outlined in Appendix 3A (indicated as “LIMIT” hereafter). The second alternative selects rows based on forward linkages in Ghosh’s supply-driven IO model (indicated as “ROWGOH” hereafter). The third alternative identifies key sectors based on the relative linkages of in the hypothetical extraction method for columns. Recall that “COLHYP” selected sectors with the largest outcomes for the linkages as measured by $\sum (x_i - x_i^{'-1})$. Dietzenbacher and van der Linden (1997) have argued that the effect of the sector size should be removed and that relative linkages $\sum (x_i - x_i^{'-1})/x_i$ should be used. The method based on these relative linkages is indicated as “COLRHYP”. Similarly, the fourth alternative uses relative linkages from the hypothetical extraction of rows (“ROWRHYP”). The fifth alternative uses the original hypothetical extraction method (indicated as “HYP” hereafter), which extracts both row $k$ and column $k$. In order to compare the results of these five alternative methods with the results in the main body of this chapter, we also include “LARGE” and “COLHYP”, i.e. the best and the worst performing method in graphs B and C of Figure 3.1. Due to the fact that HYP identifies the cells of a row and corresponding column for $q (= 1, 2, 3, 4)$ sectors simultaneously, the numbers of superior data in this section amount to $q(2n-q)$ (= 61, 120, 177, 232, respectively).

Similar to Figure 3.1, Figure 3B.1 presents the average WAPEs of the intermediate deliveries (in graph A), of the input coefficients (in graph B), and of the Leontief inverse (in graph C). A first observation is that LARGE still outperforms all the alternative methods whereas COLHYP is still the worst one (except in graph A). Graphs B and C in Figure 3B.1 sketch a very clear distinction of performance into three categories: the two methods that focus on individual cells perform the best, where LARGE outperforms LIMIT; the two methods that focus on the rows when searching for key sectors perform second best, where ROWRHYP performs invariably better than ROWGOH; the other three methods (COLRHYP and COLHYP...
Targetting the collection of Superior data focusing on columns and HYP focusing on the combination of a row and its corresponding column) perform comparably poor.

With respect to graph A, as we have seen in Section 3.4.1, COLHYP may yield relatively good estimates for the intermediate deliveries because it is one of the methods that takes sector size (in terms of output) into account. This also explains the difference between graph A and graphs B and C in Figure 3B.1. In graph A, the ordering is slightly different from that in graphs B and C: the two methods that focus on individual cells still perform the best, and LIMIT generates equally good estimates as LARGE; COLHYP and COLRHYP yield very similar results and perform second best; ROWRHYP and HYP follow in the third category, where ROWRHYP is slightly better; ROWGOH performs invariably the worst.

Figure 3B.2 gives the coefficients of variation corresponding to the seven methods, for the intermediate deliveries (graph A), for the input coefficients (graph B), and for the Leontief inverse (graph C). When linked to Figure 3B.1, a general observation is still that poor performance appears to be invariably poor across regions (i.e. the lowest accuracy in Figure 3B.1 is coupled to the largest stability in Figure 3B.2) while the best performance in terms of accuracy goes together with the largest variability across regions. HYP might be the only exception in graphs B and C (coupling a relatively poor accuracy in Figure 3B.1 to a relatively large variability in Figure 3B.2). In this sense, HYP is less favorable than the row-wise methods (which select the same amount of cells for superior data collection), especially when intermediate deliveries are concerned.
Figure 3B.1: Average WAPEs of the estimation

A. Accuracy in terms of the intermediate deliveries matrix

B. Accuracy in terms of the input coefficients matrix

C. Accuracy in terms of the Leontief inverse
Targetting the collection of Superior data

Figure 3B.2: Variation of the WAPEs across regions

A. Estimating the intermediate deliveries matrix

B. Estimating the input coefficients matrix

C. Estimating the Leontief inverse
Appendix 3C. Results of Alternative Evaluation Criterion in WADs

Figure 3C.1 gives the average WADs for the estimates of the intermediate deliveries (graph A), of the input coefficients (graph B), and of the Leontief inverse (graph C). The results for the WADs are very similar to the results for the WAPEs (see Figure 3.1). The ordering of the methods in terms of their accuracy is exactly the same when estimating the input coefficients (compare graphs B in Figures 3.1 and 3C.1) and the Leontief inverse (graphs C). This also holds for the accuracies of estimating the intermediate deliveries (compare graphs A), except for ROWSUM and ROWHYP that change position. When it comes to the estimation of the intermediate deliveries, ROWSUM yields slightly better accuracies than ROWHYP according to the WAPEs whereas ROWHYP outperforms ROWSUM according to the WADs.

Another observation is with respect to the slope of the curves. In Figure 3.1, it was found that for a given number $p$ of cells with superior data the average WAPE with LARGE is more or less equal to the average WAPE with ROWSUM for $3p$. This magnification, however, seems to be larger when WADs are used. From graph C in Figure 3C.1, it follows that the average WAD with LARGE for a given number (e.g. 31) of cells with superior data, approximately equals the average WAD with ROWSUM for $4p$ (124). When applying WADs to measure and compare the accuracies, this implies that both methods perform roughly the same if LARGE is four times as expensive as ROWSUM and the budget for superior data collection is fixed.

Figure 3C.2 depicts the coefficients of variation of the WADs across regions, just like Figure 3.2 did for WAPEs. Again, the performance of the methods is very similar, except for ROWHYP. According to the WAPEs and compared to other methods, ROWHYP generates quite a lot of variation for small numbers of superior cells (less than 124) when the input coefficients or the Leontief inverse (see graphs B and C) are estimated. When WADs are used, however, ROWHYP produces a significantly lower variation than LARGE all the time. Hence ROWHYP is found to be more recommendable when the errors are evaluated by WADs then by WAPEs.
Targetting the collection of Superior data

Figure 3C.1: Average WADs of the estimation

A. Accuracy in terms of the intermediate deliveries matrix

B. Accuracy in terms of the input coefficients matrix

C. Accuracy in terms of the Leontief inverse
Figure 3C.2: Variation of the WADs across regions

A. Estimating the intermediate deliveries matrix

B. Estimating the input coefficients matrix

C. Estimating the Leontief inverse
Appendix 3D. Results for the Degrees of Approximation

This appendix gives the degrees of approximation as a measure for the accuracy of the estimates (as an alternative to the WAPEs). Figure 3D.1 presents the results, with graph A for estimating the intermediate deliveries (or the input coefficients) and graph B for the Leontief inverse.

**Figure 3D.1: Average degrees of approximation**

A. Accuracy in terms of the intermediate deliveries matrix or the input coefficients matrix

B. Accuracy in terms of the Leontief inverse

Graph A is somewhat intriguing, with the results lying almost on a straight line and the results being the same for all methods. In the matrix with 961 intermediate
deliveries, we found on average 60 zero entries in 2002. In the case where no superior data are inserted, on average 68 of the 901 positive cells were estimated with an error no larger than 10% (i.e. were approximated well). Inserting the true values in the 31 cells that have been selected as most important cells or cells of key sectors, adds another 25 cells that are approximated well if LARGE is used (and 29 cells if ROWSUM is applied for the selection). Each and every time that the next 31 cells in the intermediate deliveries matrix are filled with their true 2002 values, the number of well approximated cells increases by 26-28. The fact that the number of well approximated cells increases almost by 31 in each step with the addition of superior data for the next 31 important cells (or for the next key sector), suggests that the original 68 well approximated cells (i) are relatively unimportant and (ii) are not concentrated in either the columns or the rows of the key sectors.

Graph B for the Leontief inverse, sketches the same picture as we have seen in Figures 2.1 and 2.2. That is, LARGE performing the best, followed by ROWSUM and ROWHYP (where, in contrast to earlier findings, ROWHYP does a tiny little bit better than ROWSUM), whereas the column methods (COLSUM and COLHYP) and INVIMP perform the worst.