Improved Estimation of Regional Input-Output Tables Using Cross-Regional Methods

Xuemei Jiang, Erik Dietzenbacher, Bart Los

Improved Estimation of Regional Input-Output Tables Using Cross-Regional Methods

Journal:	*Regional Studies*
Manuscript ID:	CRES-2009-0189.R1
Manuscript Type:	Main Section
JEL codes:	D57 - Input–Output Analysis < D5 - General Equilibrium and Disequilibrium < D - Microeconomics, R11 - Regional Economic Activity: Growth, Development, and Changes < R1 - General Regional Economics < R - Urban, Rural, and Regional Economics
Keywords:	non-survey methods, cross-regional methods, regional input-output tables, China

CRIS: note that the Chinese abstract has been provided by the authors

Improved Estimation of Regional Input-Output Tables Using Cross-Regional Methods

XUEMEI JIANG[*], ERIK DIETZENBACHER[**] and BART LOS[***]

[*]Center for Forecasting Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road, No.55, 100190, Beijing, China (E-mail: jiangxuem@gmail.com);

[**]Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV, The Netherlands and Graduate University, Chinese Academy of Sciences, Beijing, China. (E-mail: h.w.a.dietzenbacher@rug.nl);

[***]Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV, The Netherlands (E-mail: b.los@rug.nl)

(Received June 2009: in revised form August 2010)

Abstract

Many regional input-output tables are estimated by means of non-survey methods. Often, information on the margins of the projected table is complemented by full information on intermediate inputs from tables for other regions. This paper compares the performance of four of such 'cross-regional' methods. Two of these were proposed in the literature before, whereas the other two are based on recent advances in regression analysis. The methods are not only tested against each other, but also against traditional methods that do not employ cross-regional information. To this end, 27 regional input-output tables for China in 1997 and 2002 are used.

Keywords

Non-survey methods, cross-regional methods, regional input-output tables, China.

基于多地区集成的地区投入产出表非调查编制方法研究

摘要：基于已有统计数据进行非调查的地区投入产出表的编制一直是投入产出学界的关注焦点。利用其它多个地区的投入产出表以及目标地区的已知中间投入行和、列和向量，可以编制出目标地区的投入产出表。本文首先将这种方法定义为多地区集成编制方法，综述了文献中已有的两种多地区集成编制方法，并结合计量经济学模型的新进展，提出了两种结合计量经济学模型的非调查多地区集成编制方法。基于中国 27 个地区 1997 年和 2002 年的投入产出表，本文不仅验证了这四种方法的预测精度，还将其与传统非调查方法进行了比较。

关键词：非调查方法，多地区集成编制方法，地区投入产出表，中国

JEL Codes

D57; R11

1. Introduction

Fueled by the policy relevance of regional input-output analysis, a vast literature on the construction of regional input-output tables has emerged. Especially hybrid methods, which combine non-survey approaches with superior survey-based data, have become popular.[1] This does not mean, however, that non-survey methods are not being employed anymore. On the contrary, non-survey techniques still receive considerable attention, if only since they are at the heart of the first step of hybrid methods (see e.g. Lahr, 1993, 2001; Okamoto and Zhang, 2007; Bonfiglio and Chelli, 2008).

A number of non-survey techniques to estimate an 'object table' (or, a table for the 'object year') have been developed over the past decades. These techniques, like all methods introduced an analyzed in this study, have in common that row and column totals (like sectoral gross output) are known, but that the block of intermediate inputs has to be estimated.

Updating the latest available survey-based input-output table by iteratively rescaling rows and columns to known margin totals of the object table, i.e. the so-called RAS technique, is still a very popular method. In terms of estimation performance, it is hard to beat if no supplementary information is available (Oosterhaven et al., 1986; Polenske, 1997; Jackson and Murray, 2004). Alternatively, regionalization using location quotients is an often used method if a survey-based national table for the object year is available (see, e.g. Flegg et al., 1995). In case survey-based tables for other regions are available for the object year, substituting input coefficients from a table for the region that is similar according to some yardstick is also widely used (see, for example, Rueda-Cantuche et al., 2009, who use information for Belgium to construct import tables for Luxembourg). These methods have in common that estimated coefficients are based on information contained in a single survey-based table.[2] We feel that much less experience has been gained with regional IO table construction based on information contained in several other regional tables, although some methods have been proposed (see, e.g., Jensen et al., 1988; 1991).

In this study, we aim at providing information to practitioners about how to take full advantage of the information on intermediate inputs included in a cross-section of other

regional tables in estimating a regional object table. Next to methods based on classical linear regression analysis as applied by Jensen et al. (1988; 1991), we also study methods grounded in more recent contributions to regression analysis, such as robust regression (Rousseeuw and Van Zomeren, 1990) and threshold regression (Hansen, 2000). We apply these advanced methods after having shown that data contained in regional input-output tables can have distributional characteristics that render classical regression methods less appropriate. The methods we analyze are empirically compared on the basis of a collection of survey-based input-output tables for Chinese provinces in 1997 and 2002, covering 27 regions and 31 industries.[3,4] Our choice for these Chinese tables is suggested by two considerations (see Appendix 1 for a brief description of the Chinese provincial tables and their compilation). First, the Chinese set of regional IO tables is unique in the sense that it is the largest available set of harmonized tables expressed in one single currency. Second, the well-known characteristic of large geographical disparities in China adds to the attraction of our analysis; the vast majority of regions are clearly not representative for the nation and heterogeneity abounds.

The plan of the paper is as follows. Section 2 briefly reviews 'traditional' approaches of constructing non-survey regional technical tables, which do not rely on the identification of cross-regional patterns. Section 3 proposes the four cross-regional methods that will be employed. Section 4 presents a comparison of the estimation results obtained using the cross-regional approach to those generated by the traditional methods. Section 5 systematically tests the robustness of our comparison results if the available cross-regional sample would be smaller and contain much fewer than 26 tables. These experiments provide guidelines on which method to use in a variety of situations regarding data availability. In section 6 we summarize our findings and conclude.

2. Non-Survey Methods Based on the Coefficients of a Single Input-Output Table

Before starting our review of methods, we should first delve a little bit deeper into the nature of the Chinese regional input-output tables at hand. It should be emphasized that Chinese regional tables only provide information on intermediate deliveries including

imports. This means that the intermediate delivery X_{ij} expresses the total input of products from industry *i* by industry *j* in region *r*, irrespective of the location of industry *i*. This makes that some parts of the literature on the construction of regional IO tables, which focus on estimation of intra-regional inputs only, is not relevant for the situation at hand. Since we focus on what Boomsma and Oosterhaven (1992) coined "technical tables", we do not have to deal with the estimation of location quotients (alternatively called regional purchase coefficients).[5] Location coefficients (Flegg et al., 1995, Flegg and Webber, 2000, Tohmo, 2004, Riddington et al., 2006) indicate what share of a regional industry's inputs are sourced domestically. Sizes of regions and transport costs of specific inputs are just two of the main variables that are often supposed to play an important role in the determination of location coefficients. We can abstain from these issues.

2.1 Intertemporal Updating

The 'RAS' technique developed by Stone and Brown (1962) has been acknowledged as one of the most widely-used ways to update tables, based on the input-output structure of an older survey-based table and information on the margins (such as total intermediate input use and total intermediate inputs supplied by industry) for the object table. Many variations of the original RAS updating techniques exist, however (see e.g. Morrison and Smith, 1974; Sawyer and Miller, 1983; Polenske, 1997; Jalili, 2000; Jackson and Murray, 2004). RAS can be seen as a method that tries to reconcile the old intermedite input structure as well as possible with the new column and row totals. Despite regular complaints about the poor performance of RAS, reviews of empirical results such as Polenske (1997) and Jackson and Murray (2004) tend to conclude that RAS results are seldom outperformed by alternatives using the same type of information.

In the context of the present analysis, information on total interindustry sales and total interindystry purchases taken from a 2002 table and the input coefficients taken from the 1997 table for the same region allows us to apply the RAS method to update all 27 regional tables to 2002. Next, the quality of these estimates by updating can be assesed

by comparing the updated tables to the true 2002 tables, by yardsticks that will be discussed below.

2.2 Regionalization of National Tables

Updating techniques, however, cannot be used if IO tables have not been constructed for the object region before. For regional analyses (as opposed to country-level studies), the literature recognizes many alternative approaches to produce non-survey IO tables, but most of these focus on the domestic sourcing issue that is not relevant to us, as we explained above.

As far as technical tables (the cells of which contain both domestically produced and imported inputs) are concerned, national tables are most often regionalized by RAS methods (Boomsma and Oosterhaven, 1992). The national input coefficients are taken as a starting point and information on the row and column sums of the regional intermediate deliveries matrix is taken as constraints. Iterated rescaling of rows and columns then generates a table with estimated technical coefficients for the object region.

2.3 Exchanging coefficients

Instead of using a national table to reflect the economic characteristics of a particular region *r*, one might use information from an existing table for another region, *r'*. Especially if *r* and *r'* are thought to be economically and technologically similar, the estimation error is likely to be small (Miller and Blair, 1985). Hewings (1977) gave an example of coefficient exchange at the regional level, estimating a table for the state of Kansas 1965 borrowing input coefficients from the table of Washington State for 1963. Finally, RAS was used to balance the Kansas table obtained in this way.

A problem arises if several regional tables are available to choose from: which of the regions is defined to be most similar to *r* (the object region), in particular in a situation in which the input coefficients of the object table are unknown? This issue has hardly been discussed in the literature. In this paper, we propose to use the vector of input coefficients for each sector in 1997 to represent the input technology of the corresponding region and sector.[6] Then the similarity index SI_j^{rk} for 1997 is calculated for a pair of region *r* and *k* for each and every sector *j*:

$$SI_j^{rk} = \frac{\sum_{i=1}^{n} a_{ij}(r) \cdot a_{ij}(k)}{\left[\sum_{i=1}^{n} a_{ij}(r)^2 \cdot \sum_{i=1}^{n} a_{ij}(k)^2\right]^{1/2}} \quad (1)$$

in which a_{ij} denotes the input coefficients for a region. The expression in the right hand side is the cosine between the two input coefficients vectors of r and k. Jaffe (1986) proposed such a measure (which is bounded by zero and one given the nonnegativity of input coefficients) based on shares of technology classes in the patent portfolios of firms.[7] For each industry j, we consider the region k which has the highest SI_j^{rk} with the object region r as the most similar region. Consequently, its coefficients for 2002 have been inserted in the corresponding column of the object table. We repeated this experiment for all sectors, after which application of RAS ensured a balanced estimated table for 2002.

3. Non-Survey Methods Using Cross-Regional Information

As opposed to the methods described in the previous section, methods using information from a multitude of regional tables have barely been evaluated. The availability of comparable regional input-output tables for as many as 27 Chinese regions allows for a systematic analysis along these lines. We will compare estimated tables against the survey-based tables, as well as to the more traditional estimates based on information from a single region. In this section, we present two commonly used cross-regional approaches. These are based on regression analysis. We find, however, that assumptions essential to classical linear regression are violated in our dataset. Hence, we also propose two novel methods that deal with these problems.

The idea to use information from other available regional tables in constructing regional IO tables is not entirely new. Imansyah (2000), for example, proposed the "averaging" method, which computes the average input coefficients of the other regions, multiplies these with the the industry's gross output level and balances the resulting table using RAS method, to generate the objective table.

Another well-known way to produce a matrix of deliveries for the object region from a cross-regional perspective starts from the notion of the Fundamental Economic

Structure (FES), as proposed by Jensen et al. (1988; 1991). By regressing the intermediate deliveries on an independent variable that represents the regional economic 'size', the concept of FES provides a cross-regional insight into the estimation of intermediate deliveries, as the following equation shows:

$$X_{ij}(r) = \alpha_{ij} + \beta_{ij} X(r) + \varepsilon_{ij}(r) \qquad (2)$$

$X_{ij}(r)$ represents the intermediate deliveries for the *rth* region and $X(r)$ is an indicator of the economic size of the *rth* region. α_{ij} and β_{ij} are cell-specific parameters to be estimated, ε_{ij} is considered to be random noise. Based on a series of IO tables for ten regions of Queensland, Jensen et al. (1988) found highly significant estimates for the parameters for many cells X_{ij}, though not for all. Jensen et al. consider the cells for which Equation 2 has a high explanatory power to be part of the Fundamental Economic Structure and indicated that such a FES could be used in a compilation of regional tables. Van der Westhuizen (1992) and Thakur (2004) actually tried to use the FES technique to compile regional IO tables: They also estimated regression equations that related the intermediate delivery $X_{ij}(r)$ to alternative region-specific variables, such as total population, total value-added, gross output for sector *i* and gross output for sector *j* in the region. Next, they estimated cells for the object table based on the parameters and corresponding independent variables, and applied RAS for balancing.

We can show that Imansyah's (2000) averaging method represents a special case of the FES approach, if we limit Jensen et al.'s (1988) FES method to regressions with regional sectoral gross output levels $X_j(r)$ as the independent variable. Denoting the input coefficients by a_{ij}, Equation 2 can then be written as

$$a_{ij}(r) = \frac{X_{ij}(r)}{X_j(r)} = \frac{\alpha_{ij}}{X_j(r)} + \beta_{ij} + u_{ij}(r)$$

If α_{ij} is set to zero, the averaging method produces identical estimates as this FES equation. The actual differences between the estimates depend on the extent to which returns to scale are non-constant. If regions with large sectors can use their inputs more

efficiently, α_{ij} will be significantly positive. In our analyses below, we will start from an equivalent regression equation that has the advantage of being linear in $X_j(r)$:

$$a_{ij}(r) = \kappa_{ij} + \lambda_{ij} X_j(r) + e_{ij}(r) \quad (3)$$

All four cross-regional methods discussed below have Equation 3 as their point of departure and can therefore be seen as originating from the FES approach.[8]

3.1 Averaging coefficients

Our first cross-regional approach amounts to estimating Equation 3 for all a_{ij}s with the restriction that λ_{ij} is equal to zero. The sample consists of all Chinese regional input-output tables for 2002 in our dataset, with the exception of the object table. Next, the estimated input coefficients are multiplied by the values of X_j of the object region to arrive at estimated intermediate input flows for 2002. These are reconciled with the available margin totals for the object table by means of simple RAS. The method is exactly identical to what Imansyah (2000) proposed.

3.2 Ordinary Least Squares regression

Estimating Equation 3 by means of Ordinary Least Squares (OLS) without any restrictions on the two parameters comes close to the procedures advocated by Jensen et al. (1988). The samples are identical to the samples used for the averaging method. After having obtained the estimates for the a_{ij}s, the remaining steps in the procedure are exactly the same as those for the averaging method.

Figure 1 graphically depicts two situations. In the left panel (which relates to the inputs of "textiles" per unit of gross output of the "wearing apparel" industry), OLS regression yields an almost flat line that nearly coincides with the line produced by the averaging method. Apparently, the use of textiles in the wearing apparel manufacturing industry is not subject to economies of scale.

INSERT FIGURE 1 ABOUT HERE

The right panel of Figure 1 shows an example of an input coefficient for which OLS regression and averaging yield completely different results. In this case, which refers to the inputs of "electronic and telecommunications equipment" in the "wearing apparel" industry, the OLS regression line is clearly upward sloping. Most probably, the input of such high-tech equipment (instead of labor or more traditional equipment) is only commercially attractive when high volumes are produced.

As is well-known, linear regression by means of the method of least squares leads to estimates with desirable properties if a number of assumptions are met. One of these assumptions is 'homogeneity of the data-generating process'. This assumption can be violated in several ways. An example is the occurrence of outliers, which are often generated by differences in parts of the data-generating process related to variables that are omitted from the regression equation. Another type of violation emerges if the true value of parameters included in the regression equation varies with ranges of values of the explanatory variables. Taking Equation 3 as an example, one might think that increasing returns to scale do not play a role for relatively small values of X_j, but might set in for larger values (or the other way round). If so, the relationship between the variables cannot be represented by a single set of parameters and one should allow for parameter heterogeneity. So far, the limited body of literature proposing cross-regional methods has not addressed these potential problems.

The right panel of Figure 1 above suggests that violations of the homogeneity assumption may play a role indeed. The very high input coefficient of almost 0.006 in the upper right corner of the diagram, for example, is either an outlier or points towards a relation between a_{ij} and X_j that is different between low and high values of X_j. Below, we will propose two advanced regression approaches that address these problems. The robust regression approach explicitly deal with the potentially disturbing effects of outliers, whereas the threshold regression approach allows for parameter heterogeneity.

3.3 Robust regression

Outliers can have substantial impacts on OLS estimates of parameters in a regression equation. If such estimates are not accurate, the estimates of the object tables will be inaccurate as well. The potential effects of outliers can be illustrated by means of Figure

1 (right panel). The very high regional input coefficient of just below 0.006 associated with a sectoral total input of approximately 160 millions RMB is a clear outlier. Since this outlier is located at one of the extremes in the horizontal dimension, it tilts the OLS regression line anticlockwise. This single observation (which is called a 'bad leverage point' in the terminology of Rousseeuw and van Zomeren, 1990) has much more impact on the estimated coefficients than the observations that are closer to the centre of the cloud of observations. The second highest regional input coefficient of about 0.004 is located much closer to the center of this cloud. Hence, this outlier does not have much of an impact on the estimated slope. Its effect largely remains limited to the estimated intercept.

In order to reduce the effects of outliers and bad leverage points, several robust regression techniques have been developed. In our robust regression approach to estimating Equation 3, we use the procedure that underlies the *robustfit* algorithm in the Matlab programming language. This algorithm uses an iteratively reweighted least squares sequence.[9] In this algorithm, observations that yield a large residual in the first iteration, get a small weight in the weighted least square estimation in the next iteration. Hence, the impact of outliers is severely reduced. In our application weights are determined according to a bisquare weighting function (see Beaton and Tukey, 1974). After having obtained estimates for the parameters of Equation 3 in this way, an a_{ij} for the object region is predicted based on the total sectoral inputs $X_j(r)$. If the sample for a specific a_{ij} does not contain outliers, the weights in the iteratively reweighted least squares do not deviate much from each other and the estimates using robust regression will not be very different from those obtained using OLS. After all the input coefficients for the object table have been estimated using robust regression, the RAS algorithm is used to align the corresponding table of intermediate input flows to the marginal totals.

3.4 Threshold regression

If the relation between the dependent variable and the explanatory is characterized by strong parameter heterogeneity, estimating parameters as if they were identical for the entire sample is likely to lead to undesirable results. One of the simplest approaches to

avoid such potential problems is threshold estimation, pioneered by Hansen (2000).[10] In the context of this paper, the point of departure is the following set of equations

$$\begin{aligned} a_{ij}(r) &= \kappa_{ij}^{1} + \lambda_{ij}^{1} X_j(r) + u_{ij}(r) \qquad \forall X_j(r) \le \gamma_{ij} \\ a_{ij}(r) &= \kappa_{ij}^{2} + \lambda_{ij}^{2} X_j(r) + u_{ij}(r) \qquad \forall X_j(r) > \gamma_{ij} \end{aligned} \tag{4}$$

where $r = 1,...,m; i, j = 1,...,n$. Equations 4 can be seen as a generalization of Equation 3: For regions with large total inputs of sector j (X_j), the linear relationship between the intermediate input coefficient a_{ij} and X_j is characterized by different values of κ and λ than for regions with small total inputs. γ is the threshold between the two 'regimes'. It is endogenously estimated, by taking the sample value for which the reduction in the sum of squared residuals (SSR) attained by allowing for two sets of parameters is largest (Hansen, 2000).[11] A likelihood ratio test, the outcome of which depends on the degree to which SSR is reduced by allowing for two sets of parameters, leads to the decision about whether the split is significant or not.[12,13] If it is significant, the estimation of a coefficient of the object table depends on the size group to which the corresponding total inputs belong. If not, Equation 3 is estimated for all the observations and the estimate for the input coefficient in the object table is based on the estimates for the coefficients in this equation. Figure 2 describes the entire procedure underlying the threshold approach.

INSERT FIGURE 2 ABOUT HERE

Figure 3 gives two empirical examples of comparisons between the relationships between input coefficients and total sectoral inputs as found by applying three of our cross-regional estimation methods: OLS, robust regression and threshold regression. The two cases are identical to those depicted in Figure 1, in which the results for the averaging coefficients method and OLS regression were compared.

INSERT FIGURE 3 ABOUT HERE

In the left panel, the results for each of the techniques are very much alike. The absence of outliers leads to results for OLS and robust regression that are virtually identical. The threshold regression approach yields two line segments that are slightly upward sloping, but do not show a clear threshold (it did not turn out to be significant at 10%). Following the procedure depicted in Figure 2, we would use OLS in this case.

A completely different situation emerges from the right panel. The importance of bad leverage points such as the one in the far northeast of the diagram is reduced in the robust regression approach, which leads to a much flatter regression line than in the case of regular OLS regression. This implies that robust regression points towards much less pronounced decreasing returns to scale than OLS, since the required inputs per unit of gross output appear to depend much less on total output levels. The results for the threshold regression approach are also very different from the OLS results. For sectoral total inputs below the estimated threshold of 0.5×10^8 RMB, the regression results are much flatter than for OLS. The estimated intercepts are very different however. For the subsample of four observations above the threshold size, the intercept is considerably higher than for the remaining observations associated with regions with small wearing apparel-manufacturing sectors.

The fact that the results across the three cross-regional methods are very different from each other does not offer proof that adopting more advanced methods is worthwhile. If samples like the one depicted in the right panel of Figure 3 would be very rare in regional input-output tables, not much could be won. Unfortunately, the robust regression analysis procedure does not make a dichotomous distinction between outliers and regular observations. As we explained above, the algorithm recomputes weights for *all* observations. For the threshold regression approach, we can provide more evidence. When considering observations for all 27 regions, we found splits for 112 out of the 961 cells, which amounts to a share of 11.7%.[14] This share of cells seems sufficiently large to warrant further consideration.

The number of cells for which the estimated slopes (λ_{ij} in Equation 3) are significant is small. For about 3% of the cells we find an R^2 of at least 0.25 for the univariate regressions, which indicates that deviations from constant returns to scale are generally not very strong. It should be noted, however, that statistical significance is not our main

concern. Our primary interest lies in accuracy of the projections, for which a comparison of the estimated slopes is much more important. Table 1 provides a comparison of frequencies of classes of slopes as estimated by means of Ordinary Least Squares regression and robust regression.

INSERT TABLE 1 ABOUT HERE

Table 1 clearly shows to what extent corrections for the presence of outliers change the estimation results. The estimated slopes are generally closer to zero. The share of cells with an absolute value of λ_{ij} larger than 0.005 is 53% for Ordinary Least Squares, and 43% for for robust regression.

Table 2 compares the frequencies of estimated slopes for the subset of cells for which threshold regression yielded a split into subsamples corresponding to low and high values of sectoral output X_j (see Equation 4) significant at a level of 10%.

INSERT TABLE 2 ABOUT HERE

The results show that positive slopes larger than 0.005 are found slightly more often for the subsamples associated with large sectoral output levels. The differences are not very marked, though.

In the next section, we will compare the estimating performance of the cross-regional methods introduced above not only to each other, but also to the more traditional methods based on single tables as discussed in Section 2.

4. Comparison of Estimation Results

In this section, we compare the deviations between the survey-based ("true") regional IO tables for 2002 and the estimated tables obtained by applying the procedures outlined in the previous section. Thoughout the empirical analysis, we will employ the WAPE (Weighted Absolute Percentage Error) as our measure of deviation:

$$\text{WAPE} = \frac{\sum_i \sum_j b_{ij} \left| \frac{\hat{b}_{ij} - b_{ij}}{b_{ij}} \right|}{\sum_i \sum_j b_{ij}} = \frac{\sum_i \sum_j \left| \hat{b}_{ij} - b_{ij} \right|}{\sum_i \sum_j b_{ij}} , \quad (5)$$

in which $\hat{b}_{ij}$ and b_{ij} denote the estimated and true values of the Leontief inverse **B** = (**I**-**A**)$^{-1}$, respectively.

The WAPE has been used in a large number of studies, since the weighted average of deviations is taken in such a way that large cells receive a larger weight than small cells (see, for example, Oosterhaven et al., 2007). We decided to compare the deviations for individual cells of the Leontief inverse (instead of, for exampe, the values of intermediate input deliveries or input coefficients), because the cells of the Leontief inverse constitute the building blocks of multipliers used in traditional impact analyses.

INSERT TABLE 3 ABOUT HERE

Table 3 presents the WAPEs for each region and method. The last row ("count") indicates the number of regions for which the methods of the associated columns have the highest accuracy. In a similar vein, the row "average" presents the unweighted averages of WAPEs over regions for the seven methods considered.

The WAPEs documented in Table 3 appear to be high, but it is well-known that applications of unmodified RAS generally lead to inaccurately estimated object tables (see, e.g. Lynch, 1986; Polenske, 1997). For the dataset we study here, most WAPEs would decline sharply to 0.1-0.2 when the 5% most important cells would be replaced by the true, survey-based values (see Jiang et al., 2010). The actual construction of regional input-output tables is often done by means of such 'hybrid' methods.

A first and very important finding is that cross-regional methods yield far better results than single-table methods. The estimations made with the cross-regional models produce overall average WAPEs between 0.291 and 0.295, while the corresponding range is 0.329-0.385 for traditional methods. At the level of individual regions, cross-regional

methods also show a clear superiority over single-table methods, since only for Jilin the minimum WAPE is found for a single-table method. We also observe that for most regions the WAPEs for cross-regional methods are very close to each other. In 21 out of 27 regions the worst cross-regional method still scores better than the best single-table method.

Second, it appears that regionalization based on the national table generates the best estimations among the class of single-table methods, followed by updating, while exchanging coefficients with the most similar region performs worst. This is a rather surprising result since updating a recent table is one of the most popular techniques to compile regional input-output tables. It is also striking that the exchanging coefficients procedure is outperformed by regionalization, in spite of the fact that it uses information from all other regions in selecting the regional production structure that was most similar in 1997. A reasonable explanation for these results might be that input coefficients for regions undergoing rapid development are far less stable than ones for developed countries.[15] Dietzenbacher and Hoen (2006), for example, examined the stability of input coefficients based on a time series of annual input-output tables for the Netherlands, covering the period 1948-1984. They found that 80% of the cells had coefficients of variation below 0.3. For a set of Chinese survey-based national tables covering the period 1987-2002, we find that not a single input coefficients features a coefficient of variation smaller than 0.5, and the proportion of input coefficients with a coefficient of variation below 0.8 is a mere 30%.

Third, turning our attention to the cross-regional methods, we can conclude that the four methods perform very close to each other on average, but that there are some marked differences at the level of individual regions. The robust regression method performs best for 12 regions, while the averaging coefficients method appears superior for 8 regions. Although the WAPEs for ordinary least squares regression are similar to the WAPEs for robust regression if averaged over provinces (see last row of Table 3), OLS and threshold regression score best in substantially smaller numbers of cases. These relative performances are also reflected in the ranking of the accuracies (1 = most accurate; 4 = least accurate) of the four methods, averaged over the 27 regions. These are 2.30, 2.72, 2.26 and 2.72, for averaging, OLS regression, robust regression and threshold regression,

respectively. Apparently, OLS regression as advocated by the proponents of the Fundamental Economic Structure suffers from problems caused by bad leverage points such as in Figure 1 in this empirical application for China. Threshold regression emerges as an approach to this issue that should not be preferred. It yields substantially more accurate estimations for only two Western regions (Guizhou and, particularly, Yunnan). Instead, using the averaging approach (which imposes constant returns to scale) and robust regression turn out to be promising approaches.

The overall average WAPE of the averaging method is slightly lower than the WAPE for the robust regression approach. Robust regression, however, is superior to averaging in the majority of cases (15 out of 27). This paradox is mainly due to two regions with "extreme" results: Shandong and, to a somewhat lesser extent, Zhejiang. For these regions, the robust approach yields far worse accuracies than averaging. The results in Table 4 are based on robust regressions of Equation 3 with all 27 provinces included in the sample. The regression was run 961 times, i.e. for each of the (i,j)-pairs. The colum "numbers of outliers" shows how often an observation for the corresponding region was found to be an outlier.[16] Shandong and Zhejiang are special indeed, in the sense that the numbers of cells considered as outliers is very high. About 12% of the 961 input coefficients in each of these two regions are located very far from the main cloud of observed input coefficients. In the robust regression approach, such observations get a very low weight, as a consequence of which the regression line is relatively often very far away from the observation. Hence, it is not surprising that predictions for regions with large numbers of outliers are relatively bad.

INSERT TABLE 4 ABOUT HERE

5. Robustness test of comparison results

The most important conclusion so far is that cross-regional methods systematically generate better estimations than more traditional methods using information from just one table. It should of course be borne in mind that we obtain these results for a set of

developing regions in China, of which some are undergoing rapid changes in production structure.

If we want to judge our results in a more general context of practitioners in need of regional tables without the funds or time to construct survey-based material, it is a rather strong assumption that as many as 26 tables are available as inputs for cross-regional inputs. In this section, we will investigate whether the superiority of cross-regional methods as reported in the previous section carries over to situations in which less tables can be used.

For reasons of space, we focus on the averaging coefficients and robust regression methods as representatives of the cross-regional methods. Further analyses of Ordinary Least Squares and threshold regression will be omitted, because these were most often outperformed. For similar reasons, we drop the exchanging coefficients method from the set of approaches based on coefficients from a single input-output table. In this class of methods, we scrutinize the performance of intertemporal updating and regionalization of national tables.

5.1 Experiments with sets of random samples

The relative performance of the estimation methods under consideration is likely to depend on the regional tables making up the sample. In the previous section, randomness did not play any role, because all 26 tables (27 minus the object table) were automatically included in the sample. Now, we intend to have a closer look at how the estimation methods perform if, for example, the averaging method is based on just 10 observations. In principle, we could study results for all 26!/(10!·16!)=5,311,735 possible distinct samples, but we decided for a different approach. In an experiment with strong similarities to bootstrapping, we randomly drew 1,000 samples for each region and sample size studied.[17] Next, we computed WAPEs for each sample. We summarize the empirical distribution of WAPEs as obtained in this way by means of the most straightforward statistic: the average WAPEs for the methods (as computed over 1,000 WAPEs). Finally, to facilitate bilateral comparisons of methods, we computed the percentage of random samples for which one method yielded lower WAPEs than for the other.

5.2 Comparison between averaging and robust method

First, we study the relative performance of the two cross-regional methods for situations with samples of 10, 15 and 20 observations, respectively. The results are presented in Table 5. The average WAPEs for the averaging and robust regression methods are listed in the first two columns, for each number of observations. The percentages in the rightmost columns denote the percentage of random samples for which averaging yielded a higher accuracy (lower WAPE) than robust regression.[18] For example, 34% in the first row and third column indicates that for the Anhui region, averaging coefficients outperformed robust regression for only 34% of the random samples.

INSERT TABLE 5 ABOUT HERE

First, we observe that WAPEs increase when less observations are available, irrespective of the method adopted. This implies that as many tables as possible should be used when applying cross-regional methods. We also find, however, that the WAPEs increase remarkably slowly when sample sizes are reduced, which is a reassuring result.

With respect to the comparison between the averaging coefficients and robust regression methods, we find that the advantage of the latter over the first in terms of the number of regions for which it performs better (see Table 3) switches to a disadvantage when the numbers of observations in the sample decline. For the case of 20 observations, the robust regression method performs better in 13 regions according to average WAPE, while these numbers drop to only 10 and 5 in the cases of 15 and 10 observations, respectively (see the bottom line of Table 5). In a similar vein, we find that the advantage of the averaging coefficients method in terms of the unweighted average of average WAPEs as documented in Table 5 also grows if fewer observations are available (from 0.004 for 20 observations to 0.011 for 10 observations). Analysis of the percentages of samples for which averaging coefficients performs better than robust regression tells a similar story. Only for Jilin with a sample size of 10, we find that the results for the majority of random samples favor robust regression, while the average WAPE is smaller

for averaging. Apart from this case, the average WAPE appears to be a statistic that captures the entire empirical distribution well, at least for the purposes of our analysis.

The result that the performance of the robust regression method worsens with lower numbers of observations can be explained by the fact that robust regression is less capable to identify outliers, if the cloud of "regular observations" is small (see Rousseeuw and Van Zomeren, 1990). Consequently, observations that are outliers for the entire set of 27 provinces as counted in Table 4 are often treated as almost regular observations if sample sizes are small and the empirical differences between robust regression and Ordinary Least Squares regression vanish. In Table 3, we already found that OLS regression performs systematically worse than the averaging method. An important intermediate conclusion we draw is that if only a few regional tables are available, the use of the averaging coefficients method is recommended.

5.3 Comparison of averaging against traditional methods

In this section, we compare the performance of the averaging coefficients method (a cross-regional method) to those of the intertemporal updating and regionalization of national tables techniques. If only few regional tables are available, one might expect that the clear advantage of the cross-regional methods (as presented in Table 3) disappears and that the use of single table methods should be favored. Like in the previous subsection, we will first compare the accuracies based on average WAPEs over 1,000 random samples. The results are documented in Table 4.

INSERT TABLE 6 ABOUT HERE

For all regions except Jilin and the cities of Beijing and Shanghai, we find that the cross-regional method beats both single-table methods as long as the number of available regional tables is eight or more. Apparently, small numbers of contemporaneous tables already yield sufficient information to compensate for the fact that the tables relate to regions different from the object region. Only if regions have very special structures, such as city-provinces, regionalization of national tables (and to a lesser extent) intertemporal updating methods may prove superior also for larger samples.

Again, we also present results for another statistic (the percentage of random samples for which regionalization of national tables leads to higher accuracies than averaging), which provides more insight into the empirical distribution of relative WAPEs.[19] The results are presented in Table 7.

INSERT TABLE 7 ABOUT HERE

Table 7 by and large confirms the results obtained for the average WAPEs. For a sample size as small as seven, we find that in as many as 16 of 27 regions regionalization is more accurate than averaging for less than 10% of the random samples. For smaller samples, the percentage of samples for which regionalization beats averaging coefficients increases. For a sample size of five, for example, we find that this happens in eight regions for more than half of the randomly drawn samples.

Again, Beijing and Shanghai are the main exception to the rule of superiority of cross-regional methods like averaging. That is, regionalization not only performs better for moderately small samples, but also for large samples. The production structures of these metropolitan cities are apparently better reflected in national tables (which incorporate the structures of Shanghai and Beijing) than in regional tables for other regions. We also find that some other coastal and central regions with a highly developed manufacturing sector, such as Fujian, Hebei, Hubei, Jilin, Liaoning and Zhejiang, tend to have lower accuracies of the averaging coefficients method in a relatively large fraction of the set of samples.

6. Conclusions

This paper presents four cross-regional non-survey methods to estimate regional IO tables (using as much data for other regions as possible) and tests these methods against three more traditional non-survey methods that rely on information contained in a single regional table: intertemporal updating, regionalization of a national table and exchanging coefficients with a table for the most similar region in the previous period. The empirical

analysis is done on the basis of two series of Chinese survey-based regional input-output tables for 1997 and 2002. They cover 27 regions and 31 industries.

We first argued that Imansyah's (2000) averaging coefficients method can be seen as a special case of the Ordinary Least Squares regression approach that Jensen et al. (1988) advocated (inspired by the notion of the Fundamental Economic Structure). Next, we introduced two alternative regression-based methods, which deal with outliers and bad leverage points. In the present context, these are regions with a production structure that is very different from the production structures in many other regions (a situation that is frequently encountered in China). The robust regression method assumes that there is a single 'law' governing the size of input coefficients and gives low weight to observations that do not appear to obey this relationship. Threshold regression, however, supposes that two 'laws' could prevail, one of which relates to small sectors and the other to regions in which the sector is large. If evidence for two laws is found, the sample is split into two and subsample-specific estimates are obtained.

We find that cross-regional methods have systematically better performance than the traditional methods based on a single table. This result carries over to situations in which fewer regional tables are available. In most cases the availability of seven or eight regional tables is sufficient to render averaging coefficients more accurate than regionalization of national tables and intertemporal updating. Among the group of cross-regional methods, averaging coefficients and robust regression generally turn out to be slightly more accurate than OLS and threshold regression. The accuracy of the robust regression technique as compared to averaging is relatively weak if the number of available regional tables is rather small. In such cases, simple averaging of coefficients appears to be the preferred method.

The results obtained in this paper should be considered carefully, because they cannot be generalized to all situations practitioners might face. The Chinese data we used are attractive for the purpose of this study because sets of 27 harmonized, survey-based regional input-output tables are very rare. This dataset allowed us to compare the estimation performance of a number of techniques as if samples of different sizes were available. One should take into account, however, that this dataset is also rather specific in at least two respects. First, the Chinese economy is both very heterogeneous and

dynamic. Some regions are very backward, while other regions (especially those in the coastal zone) have been developing very rapidly. Regionalization of national tables might perform much better for regions that are part of a country without the differences in production structures associated with the Chinese regional inequality. A similar argument goes for the bad estimation performance of intertemporal updating in our study. If input-output tables are estimated for regions that do not develop as quickly as many of the coastal and central regions in China, production structures as reflected in input coefficients are likely to be much more stable over time. This would enhance the quality of estimates obtained by intertemporal updating significantly.

Secondly, given the nature of our Chinese data, the study focuses on the estimation of technical coefficients, which are defined as intermediate inputs (both domestically produced and imported) divided by gross output. Often, however, practitioners are interested in estimating input coefficients, defined as domestically produced intermediate inputs divided by gross output levels. If cross-regional methods would be used to estimate input coefficients, some additional steps seem to be necessary, including the estimation of location quotients to correct for differences in economic size of regions: large regions will purchase relatively much from domestic sources, while small regions will import relatively much. This will be reflected in different sets of input coefficients, even if the production technologies would be identical. An account of the relative qualities of the (adapted) cross-regional methods discussed in this study and more traditional methods based on information contained in a single table if input coefficients rather than technical coefficients are to be estimated is a subject for future work.

Acknowledgements

The authors thank two anonymous referees for their constructive comments.

References

Beaton, A.E. and Tukey, J.W. (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics*, 16, 147-185.

Bonfiglio, A. and Chelli, F. (2008) Assessing the behaviour of non-survey methods for constructing regional input-output tables through a Monte Carlo simulation, *Economic Systems Research*, 20, 243-258.

Boomsma, P. and Oosterhaven, J. (1992) A double-entry method for the construction of bi-regional input-output tables, *Journal of Regional Science*, 32, 269-284.

Dietzenbacher, E. and Hoen, A.R. (2006) Coefficient stability and predictability in input-output models: a comparative analysis for the Netherlands, *Construction Management and Economics*, 24, 671-680.

Durlauf, S.N. and P.A. Johnson (1995), Multiple regimes and cross-country growth behaviour, *Journal of Applied Econometrics*, 10, 365-384.

Flegg, A.T. and Webber, C.D. (2000), Regional size, regional specialization and the FLQ formula, *Regional Studies*, 34, 563-569.

Flegg, A.T., Webber, C.D. and Elliot, M.V. (1995), On the appropriate use of location quotients in generating regional input-output tables, *Regional Studies*, 29, 547-561.

Greenstreet, D. (1989) A conceptual framework for constructing of hybrid regional input-output models, *Socio-Economic Planning Sciences*, 23, 283-289.

Hansen, B.E. (2000) Sample splitting and threshold estimation, *Econometrica*, 68, 575-603.

Hansen, B.E. (2001) The new econometrics of structural change: Dating breaks in U.S. labor productivity, *Journal of Economic Perspectives*, 15, 117-128.

Hewings, G.J.D. (1977) Evaluating the possibilities for exchanging regional input-output coefficients, *Environment and Planning A*, 9, 924-944.

Holland, P.H. and Welch, R.E. (1977) Robust regression using iteratively reweighted least-squares, *Communications Statistics: Theory and Methods*, 6, 813-827.

Imansyah, M.H. (2000) An efficient method for constructing regional input-output table: a horizontal approach in Indonesia. Paper presented at *the 13th International Conference on Input-Output Techniques*, Macerata (Italy).

Jackson, R.W. (1998) Regionalizing national commodity-by-industry accounts, *Economic Systems Research*, 10, 223-238.

Jackson, R.W. and Murray, A.T. (2004) Alternative input-output matrix updating formulations, *Economic Systems Research*, 16, 135-148.

Jaffe, A.B. (1986) Technological opportunity and spillovers of R&D: evidence from firms' patents, profits, and market value, *American Economic Review*, 76, 984-1001.

Jalili, A. R. (2000) Comparison of two methods of identifying input-output coefficients for exogenous estimation, *Economic Systems Research*, 12, 113-129.

Jensen, R.C., Dewhurst, J.H.Ll., West, G.R. and Hewings, G.J.D. (1991) On the concept of fundamental economic structure. In J.H.Ll. Dewhurst, G.J.D. Hewings, and R.C. Jensen (eds) *Regional Input-Output Modelling: New Developments and Interpretations* (Aldershot: Avebury), 228-249.

Jensen, R.C., Mandeville, T.D. and Karunaratne, N.D. (1979) *Regional Economic Planning: Generation of Regional Input-Output Analysis* (London: Croom Helm).

Jensen, R.C., West, G.R. and Hewings, G.J.D. (1988) The study of regional economic structure using input-output tables. *Regional Studies*, 22, 209-220.

Jiang, X., Dietzenbacher, E. and Los, B. (2010), Targeting the collection of superior data for the estimation of regional input-output tables, *Environment and Planning A*, forthcoming.

Lahr, M.L. (1993) A review of the literature supporting the hybrid approach to constructing regional input-output models, *Economic Systems Research*, 5, 277-293.

Lahr, M.L. (2001) A strategy for producing hybrid regional input-output tables, In: M.L Lahr and E. Dietzenbacher (eds.), *Input-Output Analysis: Frontiers and Extensions* (London: Palgrave), 211-242.

Leontief, W. (1989) Input-output data base for analysis of technological change, *Economic Systems Research*, 1, 287-295.

Los, B. (2000) The empirical performance of a new inter-industry technology spillover measure, in: P.P. Saviotti and B. Nooteboom (eds.), *Technology and Knowledge; From the Firm to Innovation Systems* (Cheltenham UK: Edward Elgar), 118-151.

Los, B. (2006) A non-parametric method to identify nonlinearities in global productivity catch-up performance, in: A. Pyka and H. Hanusch (eds.) *Applied Evolutionary*

Economics and the Knowledge-Based Economy (Cheltenham UK: Edward Elgar), 231-253.

Lynch, R.G. (1986) An assessment of the RAS method for updating input-output tables, in Sohn, I. (ed.), *Readings in Input-Output Analysis; Theory and Applications* (New York: Oxford University Press), 271-284.

Macgill, S.M. (1977) Theoretical properties of biproportional matrix adjustments, *Environment and Planning A*, 9, 687-701.

Madsen, B. and Jensen-Butler C. (1990) Make and use approach to regional and interregional accounts and model, *Economic Systems Research*, 11, 277-299.

Midmore, P. (1991) Input-output in agriculture: a review, in Midmore, P. (ed.), *Input-Output Models in the Agricultural Sector* (Aldershot: Avebury), 5-20.

Miller, R.E. & Blair, P.D. (1985) *Input-Output Analysis: Foundations and Extensions*, (Englewood Cliffs, NJ: Prentice-Hall).

Morrison, W.I. and Smith, P. (1974) Nonsurvey input-output techniques at the small area level: an evaluation, *Journal of Regional Science*, 14, 1-14.

Okamoto, N. and Zhang, Y. (2007) Non-survey methods for estimating regional and interregional input-output multipliers, in: Okamoto, N. and Ihara, T. (eds.) *Spatial Structure and Regional Development in China: Interregional Input-Output Approach*, IDE Development Perspective Series No.5, 25-45.

Oksanen, E.H. and Williams, J.R. (1992) An alternative factor-analytic approach to aggregation of input-output tables, *Economic Systems Research*, 4, 245-256.

Oosterhaven, J., Piek, G. and Stelder D. (1986) Theory and practice of updating regional versus interregional interindustry tables, *Papers of The Regional Science Association*, 59, 57-72.

Oosterhaven J., Stelder D. and Inomata S. (2008) Estimating international interindustry linkages: Non-survey simulations of the Asia-Pacific economy, *Economic Systems Research*, 20, 395-414.

Polenske, K.R. (1997) Current uses of the RAS technique: a critical review, in Simonovits, A. and Steenge, A.E. (eds.) *Prices Growth and Cycles* (London: Macmillan), 58-88.

Qi, S.C. (2007) Zhongguo Diqu Touru Chanchu Biao Bianzhi Qingkuang Jieshao [Introduction to the Compiling System of the Chinese Regional Input-Output Tables], Paper presented at the Seventh Chinese Input-Output Conference, August 2007, Nanjing, China (in Chinese).

Riddington, G., Gibson, H. and Anderson, J. (2006), Comparison of gravity model, survey and location quotient-based local area tables and multipliers, *Regional Studies*, 40, 1069-1081.

Round, J.I. (1983) Nonsurvey techniques: a critical review of the theory and the evidence, *International Regional Science Review*, 8, 189-212.

Rousseeuw, P.J. and B.C. van Zomeren (1990), Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, 85, 633-639.

Rueda-Cantuche, J.M., Beutel, J., Neuwahl, F., Mongelli, I. and Loeschel, A. (2009), A symmetric input-output table for EU27: Latest progress, *Economic Systems Research*, 21, 59-79.

Sawyer, C. and Miller, R.E. (1983) Experiments in regionalization of a national input-output table, *Environment and Planning A*, 15, 1501-1520.

Stone, R. and Brown, A. (1962) *A Computable Model of Economic Growth*. (London: Chapman and Hall).

Thakur, S.K. (2004) Structure and structural changes in India: a fundamental economic structure approach, Ph.D. dissertation, Ohio State University, Columbus, US.

Tohmo, T. (2004), New developments in the use of location quotients to estimate regional input-output coefficients and multipliers, *Regional Studies*, 38, 43-54.

van der Westhuizen, M. (1992) Towards developing a hybrid method for input-output table compilation and identifying a fundamental economic structure, Ph.D. dissertation, University of Pennsylvania, Philadelphia, US.

Yamakawa, A. and Peters, G.P. (2009), Environmental input-output analysis: Using time-series to measure uncertainty, *Economic Systems Research*, 21, 337-362.

West, G.R. (1990) Regional trade estimation: a hybrid approach, *International Regional Science Review*, 13: 103-118.

Appendix 1: Construction of Chinese Provincial Input-Output Tables

The construction of survey-based provincial input-output tables became a regularly activity since 1987, at five-year intervals. During each survey year, the national statistical agency (NBS) establishes methods of conducting the survey and explains these to the provincial statistical bureaus. The guidelines include the survey forms, the tabulation method and the industry classifications.[20] In each region, provincial officials firstly train accountants of enterpirses in the methods of filling out the forms. Enterprises provide information of intermediate consumption for production and the generation of output by means of these forms. All large-scale enterprises are surveyed, while random sampling is done for medium-scale and small-scale enterprises. For 1997 and 2002, the proportions of sampled firms were set at 20% and 8% for the medium-scale and small-scale enterprises respectively.[21]

Subsequently, the forms are submitted to provincial statistical offices, who forward the data to NBS for their national tabulation and adjustments, whereas they also use the data to tabulate their own regional tables. The standardized method of data survey ensures the quality of data collection over space. For the purposes of this study, it is essential that provincial tables are not derived from a national table (which would imply that production technologies of a province would be considered as similar to those of the country), but purposefully constructed from data at the provincial level.

INSERT TABLE A1 ABOUT HERE

It should be noted that the Chinese input-output survey defines intermediate inputs as including both domestically produced materials and materials that have been imported (either from an other province or from abroad). This procedure results in a subtle but important difference between Chinese provincial input-output tables and the internationally more common regional tables. Input coefficients derived from Chinese provincial input-output tables could be considered as "technical coefficients", which represent production technologies well. For studies of the impact of policy measures or changes in consumption or investment behavior on provincial economies (e.g. via

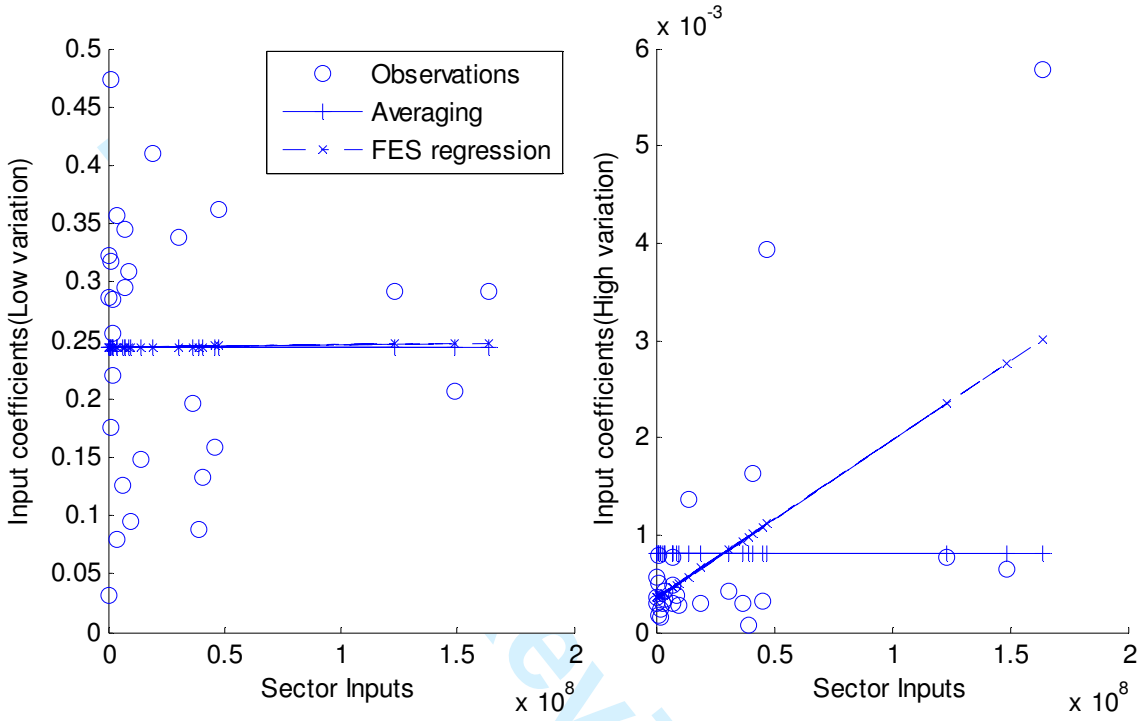
multiplier analyses), Chinese provincial tables should not be used. In such cases, assumptions about the origin of inputs are needed to adapt the tables in a suitable way.

Appendix 2: Relative Performance of Averaging Coefficients and Intertemporal Updating

INSERT TABLE A2 ABOUT HERE

FIGURES AND TABLES

Figure 1. Two cases: Averaging coefficients vs. OLS regression*



*Both panels contain observations for all 27 provinces for which input-output tables are available. In the empirical analyses, Equation 3 is estimated on the basis of only 26 observations, since the object table is assumed to be unknown. The parameter estimates are used to estimate the input coefficients of the object table. Since each of the 27 tables can arbitrarily be selected as the object table, this figure depicts observations for the entire population of tables.

Figure 2: Threshold estimation procedure

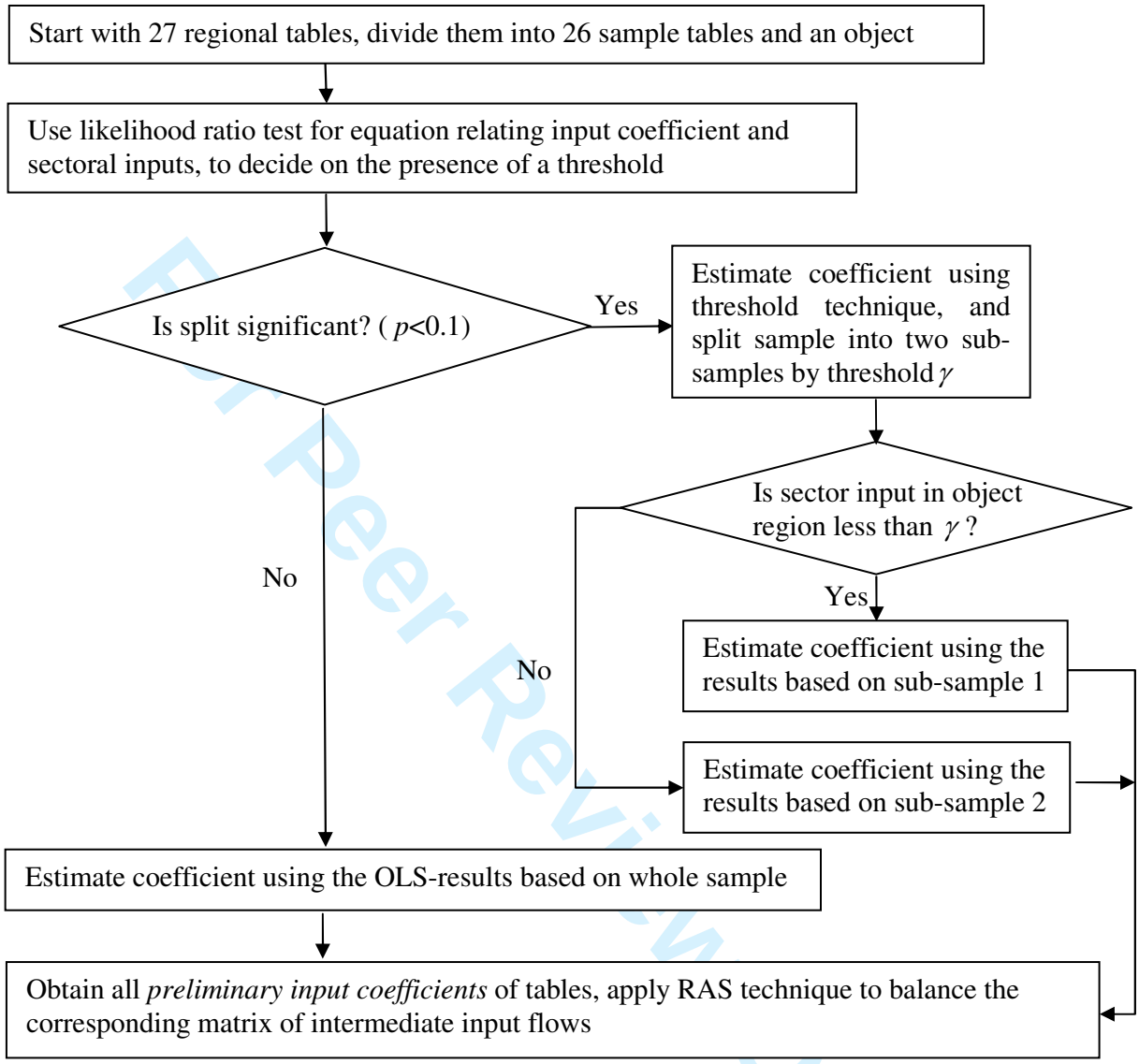
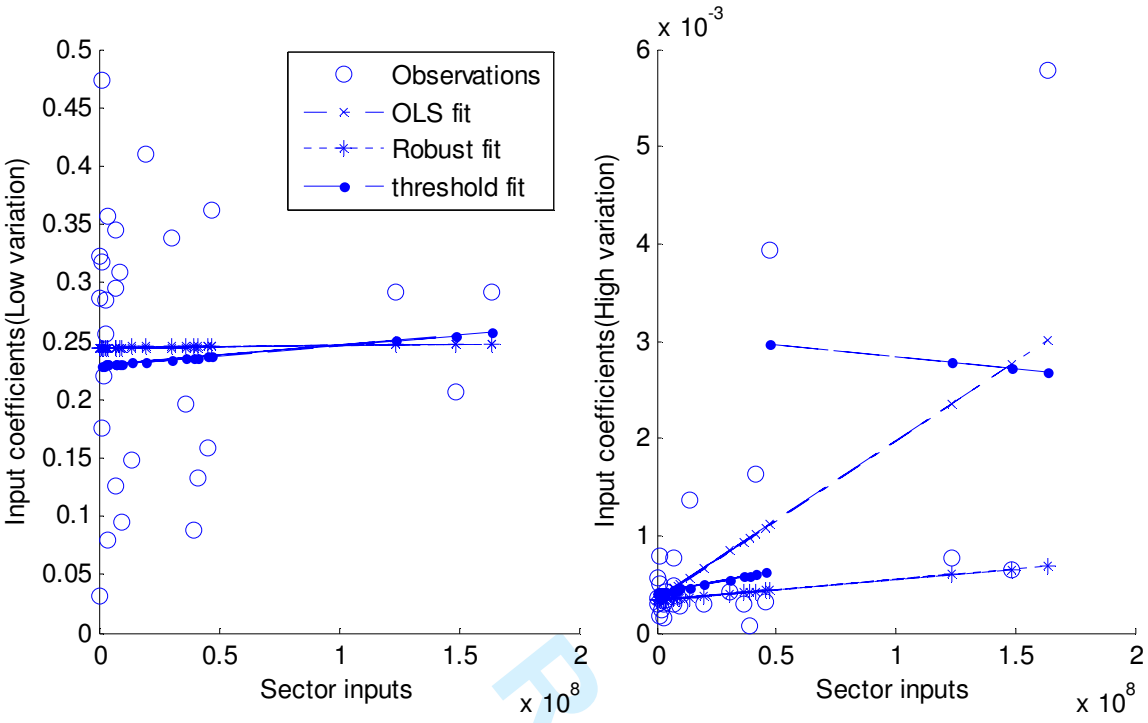


Figure 3: The two cases revisited: OLS, robust regression and threshold regression compared*



*See the note to Figure 1.

Table 1. Frequencies of estimated slopes for OLS and robust regression

Estimated slope	Frequency (Ordinary Least Squares)	Frequency (Robust Regression
$\lambda < -0.005$	2	0
$-0.005 \leq \lambda < 0$	21	44
$0 \leq \lambda < 0.0025$	277	382
$0.0025 \leq \lambda < 0.005$	156	124
$0.005 \leq \lambda < 0.01$	133	114
$0.01 \leq \lambda < 0.025$	155	125
$0.025 \leq \lambda < 0.05$	114	90
$\lambda \geq 0.05$	103	82
Total	961	961

Table 2. Frequencies of estimated slopes for various regression approaches (subset of cells with threshold significant at 10%)

Estimated slope	Ordinary Least Squares	Robust Regression	Threshold regression (small sectoral output)	Threshold regression (small sectoral output)
$\lambda < -0.005$	0	0	6	3
$-0.005 \leq \lambda < 0$	1	2	13	10
$0 \leq \lambda < 0.0025$	31	35	27	25
$0.0025 \leq \lambda < 0.005$	15	16	12	8
$0.005 \leq \lambda < 0.01$	8	13	10	11
$0.01 \leq \lambda < 0.025$	28	23	15	26
$\lambda \geq 0.025$	29	23	29	29
Total	112	112	112	112

Table 3. Accuracies of estimation methods by region (WAPEs of cells in estimated regional Leontief inverse matrices)

Object Region	Single-table Methods			Cross-regional Methods			
	Updating	Regiona-lization	Exchan-ging coefficients	Averaging	OLS regression	Robust regression	Threshold regression
Anhui	0.347	0.281	0.324	0.240	0.246	0.233	0.251
Beijing	0.346	0.339	0.390	0.337	0.338	0.324	0.341
Chongqing	0.452	0.423	0.449	0.374	0.377	0.366	0.376
Fujian	0.444	0.370	0.382	0.350	0.353	0.362	0.357
Gansu	0.377	0.365	0.404	0.301	0.296	0.286	0.297
Guangdong	0.328	0.309	0.329	0.284	0.294	0.284	0.301
Guangxi	0.407	0.381	0.393	0.317	0.311	0.326	0.316
Guizhou	0.429	0.399	0.419	0.332	0.326	0.344	0.321
Hebei	0.264	0.228	0.351	0.202	0.210	0.205	0.213
Henan	0.385	0.335	0.408	0.306	0.323	0.299	0.330
Heilongjiang	0.297	0.263	0.284	0.214	0.220	0.212	0.223
Hubei	0.239	0.236	0.296	0.207	0.208	0.225	0.213
Hunan	0.364	0.290	0.335	0.256	0.260	0.256	0.258
Jilin	0.362	0.405	0.374	0.394	0.390	0.376	0.388
Jiangsu	0.337	0.306	0.314	0.272	0.265	0.269	0.268
Jiangxi	0.352	0.301	0.345	0.260	0.269	0.243	0.263
Liaoning	0.304	0.239	0.262	0.218	0.221	0.214	0.217
Neimeng	0.455	0.401	0.376	0.334	0.327	0.345	0.325
Ningxia	0.389	0.362	0.429	0.313	0.309	0.308	0.320
Shaanxi	0.335	0.313	0.390	0.283	0.288	0.277	0.292
Shandong	0.391	0.401	0.508	0.348	0.384	0.430	0.389
Shanxi	0.383	0.407	0.419	0.373	0.377	0.387	0.375
Shanghai	0.256	0.226	0.298	0.237	0.233	0.214	0.232
Sichuan	0.329	0.306	0.321	0.248	0.248	0.258	0.247
Tianjin	0.433	0.349	0.432	0.330	0.335	0.331	0.332
Yunnan	0.434	0.365	0.446	0.285	0.276	0.280	0.265
Zhejiang	0.320	0.275	0.348	0.254	0.258	0.286	0.254
Average	0.361	0.329	0.385	0.291	0.294	0.294	0.295
Count	0	0	1	8	2	12	4

* Shaded cells indicate the method with the highest accuracy for a region.

Table 4. Numbers of outliers and differences in accuracy between averaging and robust regression.

Region	Number of Outliers*	WAPEa - WAPEr**	Region	Number of Outliers	WAPEa - WAPEr	Region	Number of Outliers	WAPEa - WAPEr
Anhui	49	0.006	Heilongjiang	63	0.002	Ningxia	65	0.005
Beijing	62	0.014	Henan	62	0.007	Shaanxi	52	0.006
Chongqing	58	0.007	Hubei	91	-0.018	Shandong	116	-0.082
Fujian	97	-0.012	Hunan	61	0.000	Shanghai	59	0.023
Gansu	38	0.016	Jiangsu	62	0.003	Shanxi	90	-0.015
Guangdong	67	0.000	Jiangxi	46	0.016	Sichuan	70	-0.010
Guangxi	62	-0.009	Jilin	73	0.018	Tianjin	50	0.000
Guizhou	41	-0.012	Liaoning	59	0.004	Yunnan	57	0.005
Hebei	67	-0.003	Neimeng	48	-0.011	Zhejiang	108	-0.032

* Outliers are defined as observations receiving a weight smaller than 0.00005 in the final stage of the iteratively reweighted least squares program as reported by Matlab's *robustfit* routine.

** Positive values point at more accurate estimates by robust regression (WAPEa = weighted average percentage error for averaging; WAPEr = weighted average percentage error for robust regression).

Table 5. Comparison of the accuracy of the averaging coefficients and robust regression methods, with different numbers of observations*

Object Region	Number of observations								
	20			15			10		
	Average WAPE of averaging coeff.	Average WAPE of robust regress.	Percent of WAPEa < WAPEr	Average WAPE of averaging coeff.	Average WAPE of robust regress.	Percent of WAPEa < WAPEr	Average WAPE of averaging coeff.	Average WAPE of robust regress.	Percent of WAPEa < WAPEr
Anhui	0.242	0.240	34%	0.246	0.246	49%	0.252	0.256	68%
Beijing	0.339	0.336	31%	0.341	0.344	56%	0.345	0.356	74%
Chongqing	0.375	0.367	4%	0.378	0.370	13%	0.382	0.377	31%
Fujian	0.352	0.363	99%	0.355	0.366	94%	0.360	0.371	86%
Gansu	0.304	0.290	0%	0.306	0.294	4%	0.312	0.304	21%
Guangdong	0.285	0.287	51%	0.288	0.301	73%	0.291	0.327	94%
Guangxi	0.319	0.327	92%	0.322	0.329	82%	0.329	0.336	78%
Guizhou	0.335	0.343	95%	0.338	0.344	80%	0.344	0.350	75%
Hebei	0.204	0.207	62%	0.208	0.216	73%	0.214	0.231	81%
Henan	0.308	0.305	31%	0.311	0.312	50%	0.317	0.327	70%
Heilongjiang	0.217	0.216	33%	0.222	0.220	42%	0.230	0.231	56%
Hubei	0.210	0.226	100%	0.215	0.230	99%	0.223	0.239	95%
Hunan	0.258	0.260	76%	0.262	0.265	74%	0.268	0.273	78%
Jilin	0.396	0.382	19%	0.400	0.392	36%	0.405	0.410	45%
Jiangsu	0.273	0.279	71%	0.276	0.290	82%	0.280	0.308	95%
Jiangxi	0.262	0.248	0%	0.265	0.254	5%	0.269	0.266	36%
Liaoning	0.221	0.217	20%	0.224	0.222	32%	0.230	0.234	56%
Neimeng	0.337	0.344	87%	0.341	0.346	71%	0.347	0.352	67%
Ningxia	0.315	0.313	28%	0.318	0.320	61%	0.324	0.333	79%
Shaanxi	0.285	0.279	9%	0.287	0.283	20%	0.292	0.291	44%
Shandong	0.350	0.429	100%	0.353	0.428	100%	0.358	0.429	100%
Shanxi	0.374	0.385	91%	0.376	0.385	83%	0.380	0.387	73%
Shanghai	0.238	0.223	3%	0.240	0.234	26%	0.245	0.254	61%
Sichuan	0.250	0.259	99%	0.254	0.263	93%	0.261	0.273	90%
Tianjin	0.331	0.332	58%	0.332	0.334	62%	0.335	0.341	74%
Yunnan	0.288	0.283	18%	0.292	0.287	31%	0.299	0.296	42%
Zhejiang	0.256	0.290	100%	0.258	0.298	100%	0.264	0.308	100%
Average**	0.293	0.297	52.3%	0.297	0.303	58.9%	0.302	0.313	69.2%
Count	14	13	13#	17	10	10#	22	5	6#

* Shaded cells indicate the method with the highest accuracy for a region.

** Unweighted averages over regions.

Number of regions with percentage of WAPEa<WAPEr smaller than 50%.

Table 6. Comparison of averaging with different numbers of observations to single-table methods (1,000 random samples).*

Object Region	Num of Obs. for Averaging							Update	Regionalization
	26**	20	15	10	8	5	3		
Anhui	0.240	0.242	0.246	0.252	0.256	0.267	0.286	0.347	0.281
Beijing	0.337	0.339	0.341	0.345	0.348	0.356	0.368	0.346	0.339
Chongqing	0.374	0.375	0.378	0.382	0.385	0.394	0.408	0.452	0.423
Fujian	0.350	0.352	0.355	0.360	0.363	0.372	0.384	0.444	0.370
Gansu	0.301	0.304	0.306	0.312	0.317	0.327	0.345	0.377	0.365
Guangdong	0.284	0.285	0.288	0.291	0.295	0.302	0.317	0.328	0.309
Guangxi	0.317	0.319	0.322	0.329	0.333	0.345	0.364	0.407	0.381
Guizhou	0.332	0.335	0.338	0.344	0.348	0.360	0.379	0.429	0.399
Hebei	0.202	0.204	0.208	0.214	0.218	0.231	0.251	0.264	0.228
Henan	0.306	0.308	0.311	0.317	0.320	0.329	0.345	0.385	0.335
Heilongjiang	0.214	0.217	0.222	0.230	0.235	0.250	0.271	0.297	0.263
Hubei	0.207	0.210	0.215	0.223	0.228	0.243	0.265	0.239	0.236
Hunan	0.256	0.258	0.262	0.268	0.272	0.283	0.301	0.364	0.290
Jilin	0.394	0.396	0.400	0.404	0.405	0.426	0.451	0.362	0.405
Jiangsu	0.272	0.273	0.276	0.280	0.284	0.291	0.304	0.337	0.306
Jiangxi	0.260	0.262	0.265	0.269	0.274	0.284	0.302	0.352	0.301
Liaoning	0.218	0.221	0.224	0.230	0.235	0.247	0.264	0.304	0.239
Neimeng	0.334	0.337	0.341	0.347	0.352	0.363	0.381	0.455	0.401
Ningxia	0.313	0.315	0.318	0.324	0.328	0.337	0.353	0.389	0.362
Shaanxi	0.283	0.285	0.287	0.292	0.295	0.304	0.318	0.335	0.313
Shandong	0.348	0.350	0.353	0.358	0.362	0.371	0.387	0.391	0.401
Shanxi	0.373	0.374	0.376	0.380	0.383	0.390	0.401	0.383	0.407
Shanghai	0.237	0.238	0.240	0.245	0.247	0.257	0.272	0.256	0.226
Sichuan	0.248	0.250	0.254	0.261	0.265	0.278	0.299	0.329	0.306
Tianjin	0.330	0.331	0.332	0.335	0.336	0.344	0.352	0.433	0.349
Yunnan	0.285	0.288	0.292	0.299	0.303	0.314	0.333	0.434	0.365
Zhejiang	0.254	0.256	0.258	0.264	0.268	0.279	0.295	0.320	0.275

* Shaded cells indicate that the averaging coefficients method outperforms both single table-based methods.

** For a sample size of 26, only one sample can be constructed, containing all regional tables except the object table. Thus, the reported WAPE is the same as in Table 1.

Table 7. Comparison of accuracies of averaging and regionalization methods for different numbers of observations*

Num of obs.	20	15	10	8	7	6	5	3
Anhui	0%	0%	0%	0%	1%	5%	14%	63%
Beijing	53%	64%	70%	75%	78%	80%	84%	91%
Chongqing	0%	0%	0%	0%	0%	0%	0%	14%
Fujian	0%	0%	10%	19%	28%	41%	51%	81%
Gansu	0%	0%	0%	0%	0%	0%	2%	16%
Guangdong	0%	0%	5%	13%	19%	25%	38%	62%
Guangxi	0%	0%	0%	0%	0%	0%	1%	15%
Guizhou	0%	0%	0%	0%	0%	0%	2%	16%
Hebei	0%	0%	8%	23%	31%	42%	58%	87%
Henan	0%	0%	5%	13%	21%	26%	39%	66%
Heilongjiang	0%	0%	0%	0%	2%	5%	13%	66%
Hubei	0%	0%	3%	17%	30%	51%	69%	96%
Hunan	0%	0%	0%	0%	3%	8%	26%	74%
Jilin	25%	35%	44%	47%	48%	50%	52%	60%
Jiangsu	0%	0%	0%	2%	4%	9%	16%	49%
Jiangxi	0%	0%	0%	0%	1%	6%	14%	44%
Liaoning	0%	0%	16%	33%	46%	56%	68%	89%
Neimeng	0%	0%	0%	0%	0%	0%	1%	17%
Ningxia	0%	0%	0%	0%	1%	5%	11%	38%
Shaanxi	0%	0%	1%	4%	8%	13%	23%	58%
Shandong	0%	0%	0%	0%	0%	1%	4%	22%
Shanxi	0%	0%	0%	2%	4%	7%	14%	36%
Shanghai	99%	96%	95%	94%	93%	96%	96%	100%
Sichuan	0%	0%	0%	0%	0%	0%	1%	35%
Tianjin	0%	0%	8%	16%	18%	27%	33%	53%
Yunnan	0%	0%	0%	0%	0%	0%	1%	13%
Zhejiang	0%	0%	12%	28%	35%	46%	65%	88%

* Percentages indicate the share of random samples for which the regionalization method yielded a higher accuracy than averaging coefficients. Shaded cells represent shares less than 10%.

Table A1. Industry classification

	Sector		Sector
01	Agriculture	17	Electric equipment and machinery
02	Coal mining, Crude petroleum and natural gas extraction	18	Electronic and telecommunication equipment
03	Metal ore mining	19	Instruments, meters, cultural and office machinery
04	Nonmetal mineral mining	20	Other manufacturing products
05	Manufacture of food products and tobacco processing	21	Electricity, gas and water production and supply
06	Textile goods	22	Construction
07	Wearing apparel, leather, furs, down and related products	23	Transport and storage, post and telecommunication
08	Sawmills and furniture	24	Wholesale and retail trade, catering trade
09	Paper and products, printing and record medium reproduction	25	Finance and insurance
10	Petroleum processing and coking	26	Real estate
11	Chemicals	27	Social services
12	Nonmetal mineral products	28	Health services, sports and social welfare
13	Metals smelting and pressing	29	Education, culture and arts, radio, film and television
14	Metal products	30	Scientific research and general technical services
15	Machinery and equipment	31	Public administration and other sectors
16	Transport equipment		

Table A2. Comparison of accuracies of averaging and intertemporal updating methods for different numbers of observations (1,000 random samples)[*]

Num of obs.	20	15	10	8	7	6	5	3
Anhui	0%	0%	0%	0%	0%	0%	0%	0%
Beijing	8%	23%	42%	55%	59%	64%	72%	84%
Chongqing	0%	0%	0%	0%	0%	0%	0%	0%
Fujian	0%	0%	0%	0%	0%	0%	0%	0%
Gansu	0%	0%	0%	0%	0%	0%	0%	9%
Guangdong	0%	0%	0%	0%	2%	3%	7%	30%
Guangxi	0%	0%	0%	0%	0%	0%	0%	1%
Guizhou	0%	0%	0%	0%	0%	0%	0%	1%
Hebei	0%	0%	0%	0%	0%	0%	2%	25%
Henan	0%	0%	0%	0%	0%	0%	0%	4%
Heilongjiang	0%	0%	0%	0%	0%	0%	0%	10%
Hubei	0%	0%	1%	7%	19%	36%	59%	93%
Hunan	0%	0%	0%	0%	0%	0%	0%	0%
Jilin	100%	96%	92%	89%	87%	88%	89%	92%
Jiangsu	0%	0%	0%	0%	0%	0%	0%	4%
Jiangxi	0%	0%	0%	0%	0%	0%	0%	1%
Liaoning	0%	0%	0%	0%	0%	0%	0%	3%
Neimeng	0%	0%	0%	0%	0%	0%	0%	0%
Ningxia	0%	0%	0%	0%	0%	0%	0%	13%
Shaanxi	0%	0%	0%	0%	0%	0%	1%	17%
Shandong	0%	0%	0%	1%	2%	4%	11%	38%
Shanxi	0%	11%	39%	48%	52%	61%	65%	84%
Shanghai	0%	2%	18%	28%	36%	41%	49%	71%
Sichuan	0%	0%	0%	0%	0%	0%	0%	5%
Tianjin	0%	0%	0%	0%	0%	0%	0%	0%
Yunnan	0%	0%	0%	0%	0%	0%	0%	0%
Zhejiang	0%	0%	0%	0%	0%	0%	0%	10%

[*] Percentages indicate the share of random samples for which the intertemporal updating method yielded a higher accuracy than averaging coefficients. Shaded cells represent shares less than 10%.

[1] See, e.g., Jensen et al. (1979), Greenstreet (1989), West (1990), Midmore (1991), Jackson (1998), Madsen and Jensen-Butler (1999) and Lahr (2001).

[2] Purists would be right in arguing that exchanging or substituting coefficients first uses information from multiple regional tables to identify the most similar region. Next, however, it disregards all information contained in tables for regions that might have a high degree of similarity to the object region, but are not the most similar.

[3] The mainland of China is administratively divided into 31 regions, including 22 provinces, 5 autonomous regions and 4 centrally administrative municipalities. We do not have data for Hainan province and the autonomous regions Tibet, Qinghai and Xinjiang.

[4] In order to make the tables comparable for 1997 and 2002, the industries in our data set were aggregated into 31 industries. See Table A1 in Appendix 1 for the classification.

[5] See Round (1983) and Lahr (1993; 2001) for thorough overviews of these techniques.

[6] This approach is inspired by Leontief (1989), who viewed columns of input coefficients as lists of ingredients for sectoral 'cooking recipes'.

[7] See Oksanen and Williams (1992) and Los (2000) for applications of the cosine measure as a similarity measure of two vectors of input coefficients.

[8] We conducted many estimation experiments using other explanatory variables than regional sectoral gross output and nonlinear forms (for example using quadratic forms or allowing for parameter heterogeneity between "poor" regions and "rich" regions). Equation 3, however, consistently led to the estimated tables that most closely resembled the true object tables. Results obtained with other explanatory variables and other functional forms are not included in this paper, but are available upon request.

[9] An early overview of alternative robust regression techniques was given by Holland and Welch (1977).

[10] Hansen (2001) presents a very accessible introduction to a strongly related approach used to identify structural breaks in time series. In an input-output context, Yamakawa and Peters (2009) apply both robust regression and (slightly different) sample-splitting techniques to study input coefficient stability over time.

[11] In principle, more than two sets of parameters might govern the relationship between the input coefficients and the total sectoral inputs. In this study, we focus on a situation with two subsamples only. We have two reasons for this decision. First, the estimation theory for multiple sample splits has not been developed thoroughly, and second, the numbers of observations in ours samples are not very high, as a consequence of which we would lose many degrees of freedom when estimating multiple splits.

[12] In this study, we adopt a significance level of 10%.

[13] The threshold regression approach advocated by Hansen (2000) requires the minimum size of both subsamples to be set exogenously ('trimming'). We followed Hansen's (2000, 2001) convention to set this value to 10-15% of the sample size. This implies that the minimum size of each subsample is 3 observations.

[14] This result obtained for Chinese regions does not necessarily generalize to other sets of national or regional input-output tables, particularly because the Chinese economy is characterized by a strong variation of regional specialization patterns.

[15] Note that the similarity index (Equation 1) is based on information for 1997.

[16] Formally, the iteratively reweighted least squares procedure does not yield a clear distinction between outliers and regular observations. However, if the algorithm leads to observations with a very small weight after the final iteration, this is a sign that the corresponding observation is located well above or below the robust regression line. Here, we rather arbitrarily denote observations with a final weight smaller than 0.00005 as outliers.

[17] In theory, it is possible that the sets of 1,000 samples contained duplicates, since we drew samples with replacement. Please note that a single sample could not contain a regional table more than once, since we drew regions, in each sample, without replacement.

[18] The percentages depend on the draw of the 1,000 samples and are therefore random variables themselves.

[19] For the large majority of Chinese regions, regionalization of national tables performs better than intertemporal updating. Hence, we benchmark averaging to regionalization. Readers interested in a comparison of the averaging coefficients method and intertemporal updating are referred to Table A2 in Appendix 2. From a practitioner's point of view, the results in Table A2 might be very relevant, because national tables are sometimes unavailable while an old regional table might exist.

[20] The 31-industry classification used in this study (see Table A1) is slightly more aggregated than the input-output tables published for 1997 and 2002, which contain 40 and 42 industries, respectively. Aggregation was needed to have tables with an identical industry classification.

[21] There might be minor adjustments in terms of the sample percentages, made by local regional statistic bureaus. It might be the case, for example, that some regions do not have large-scale enterprises in a certain industry. In that case, the regional statistic bureau tend to increase the sample percentages for medium-scale and small-scale enterprises in the industries (Qi, 2007).