

Estimating Production Technology for Policy Analysis: Trading off Precision and Heterogeneity

Abstract

This paper develops an approach to select models that can make the best use of limited micro-level data sets to estimate production function parameters. Since production is often the core of the agricultural and environment policy analyses, we evaluate the models using criteria that reflect the objectives of policy analysis. We argue that policy production models should optimize the precision of policy response predictions, but also incorporate sufficient heterogeneity to allow policy makers to consider the distributional consequences of policies. Hence we develop a series of quantitative metrics of both precision and heterogeneity to compare model performance.

Our approach consists of two steps. We first combine the method of Generalized Maximum Entropy and Data Envelopment Analysis and simultaneously estimate the production frontier and technical inefficiency parameters. With a set of household level data, we estimate production models at three different levels. The *province-level* model restricts the production technology parameters to be the same for all households. The *county-level* models allow production technology parameters to vary by county but restrict them to be equal across communities within the same county. The *community-level* models allow production technology parameters to vary by community. In the second step, we use the Disaggregated Information Gain, Percentage Absolute Prediction Error and the Theil's U statistic to evaluate these models.

JEL classification: C14; C52;

Keywords: production frontier; technical efficiency; Data Envelopment Analysis, precision; heterogeneity; generalized maximum entropy

Estimating Production Technology for Policy Analysis: Trading off Precision and Heterogeneity

1. Introduction

In many developed and developing agricultural economies, there is considerable emphasis on the effect of agricultural policies and production on the environment, and conversely, the effect of environmental policies on the agricultural sector. This emphasis has made production functions essential to the analysis of many policy problems. Often at the core of a policy model is the estimation of production function parameters.

At the same time, the importance of heterogeneity in policy analysis has long been recognized. Heterogeneity plays an even more important role in studies in developing countries. One of the important sources of heterogeneity comes from production technology. Shah, Singh and Mukherji (2006) describes how different groups of farmers in South Asia use different production technologies even though they are farming in the same community. Taylor et al. (2005) document the tremendous heterogeneity in the rural economy in Mexico that comes from various production technologies, diverse household resources and activities, and migration.

The high degree of heterogeneity necessitates the consideration of the distributional effects in policy making. Distributional concerns are characterized by the statement that “all politics is local” made by Thomas (Tip) O’Neill (O’Neill and Hymel 1994). Under the heterogeneous environment in developing countries, policy makers must understand the distributive consequences of policies on different groups of producers in order to design policies that can be effective as well as politically feasible.

Disaggregated or micro approaches have been suggested as producing estimates of parameters that characterize the behavior of small actors (e.g., rural households) that are

heterogeneous in their production activities, technologies and physical environment. For example, Just and Antle (1990) promote the use of disaggregated models to capture the heterogeneity of the physical environment and the economic behavior of producers. Taylor et al. (2005) also demonstrate that a disaggregated approach is necessary to accurately model the heterogeneous household responses that shape the outcomes of policy reforms.

However, despite the importance of capturing heterogeneity, given limited data availability, disaggregated models do not always produce the best results. A disaggregated approach requires micro-level data. Detailed micro-level data, however, are always time-consuming and expensive to collect, especially in developing countries (Deaton 1997). Since farmers do not keep any records of their input purchases or output sales in most developing countries (partly because they do not need to file tax returns), lengthy questionnaires are often used in order to collect detailed information on their production activities. This makes data collection time-consuming. It also makes it costly to scale up the sample size. Consequently, time and financial constraints have dictated that most micro-level data sets only have a small number of data points. For example, the 1997/1998 Uttar Pradesh and Bihar Survey of Living Conditions has only 18 households per community (World Bank). There are only 10 households per community in the China national household income and expenditure survey. In this case, if the analysis is carried out at the community level in order to capture differences on a community by community basis, only 10 data points are available.

When researchers face the reality of limited data sets, among other problems, a small sample size reduces the precision of estimates. When parameters are poorly estimated, one cannot obtain accurate predictions of how each group of producers will respond to a policy. Precise and accurate estimates, however, are essential to good policy analyses. One of the main

purposes of policy analyses is to characterize the responses of producers to policy changes.

Analyses based on models with large errors in predicting policy responses are unacceptable.

In summary, two aspects of a policy model are of fundamental importance to policy makers: a). the ability of a policy model to generate accurate and precise predictions of responses to policies under consideration; and b). its ability to capture the distributional effects of policies. Given the limited data availability, these two aspects generate some conflicts in the modeling approach: the first aspect calls for an approach than can use more data points in estimation, while the second aspect requires a disaggregated approach that will unavoidably reduce the sample size.¹ Thus, even when micro-level data are available, researchers must consider the trade-off between lower precision associated with small samples, which in the rest of the paper is called the *small sample problem*, and the failure to capture the full distributional effects associated with ignoring heterogeneity, which is termed the *homogeneity problem*. When balancing this trade-off in policy analyses, the ultimate question is that: Given the limited data set at hand, what is the best policy model?²

The objective of this paper is to develop an approach to select the *best* policy model when micro-level data sets are available but are limited. A policy model is best in the sense that it achieves the objectives of policy analysis in the most effective way; it not only produces predictions of policy responses with the precision possible, it also allows policy makers to assess the distributive consequences of policies. In particular, this study focuses on selecting the best policy model when researchers are operating with limited data set that only has a small number

¹ These two aspects are not necessarily always in conflict. If a policy model can capture more heterogeneity, it means the model recovers more information in the data and thus improves the precision of estimates *ceteris paribus*.

² Taylor et al. (2005) posed a similar question “How disaggregated should rural models be?” Their explanation is that it depends on the diversity of the economy being modeled and the objective of research. Because their research objective is to characterize linkages among diverse households, which dictates household level models, they estimate parameters for households without first deciding whether alternative policy models would be better (e.g., community level models). In our paper, we will focus on selecting the best policy model.

of observations. This is a common situation. Most model builders have no control over the sample size since they often only have access to previously collected data sets. Therefore, it is meaningful and useful to focus on how to make the best use of the limited data available in order to meet the demands of policy makers than to search for the perfect model under the unrealistic assumption that an ideal survey data set with large sample size would always be available.³

Our approach consists of two steps. In the first step we combine the Data Envelopment Analysis (DEA) with the method of Generalized Maximum Entropy (GME) and simultaneously estimate technical inefficiency and frontier parameters. With a set of household level data from Northern China, we estimate three different sets of policy models. The *province-level* model restricts the production technology parameters to be the same for all households. The *county-level* models allow production technology parameters to vary by county but restrict them to be equal across communities within the same county. The *community-level* models allow production technology parameters to vary by community. Choosing among these three models involves tradeoffs between capturing heterogeneity and obtaining precise and accurate prediction. The province-level model pools all observations in sample data but ignores any potential heterogeneity at the county or community level. The county and/or community level models, on the other hand, can capture more heterogeneity but are subject to less precise estimates since fewer data points can be used in estimation. In the second step, we then develop a series of quantitative metrics that measure precision and heterogeneity. These metrics are the criteria we use to select the best mode out of these three sets of models.

³ Of course, if researchers had the time and financial resources to collect their own survey data (which is rare), a different set of issues would be addressed (e.g., what is the appropriate sample size given the objective of the research and the budget and time constraints; whether to use a cluster sampling or a stratified sampling strategy). This is not the focus of our paper. We focus on how to select a best policy model with a data set that has already been collected and thus sample size is fixed.

The contributions of this paper are two-fold. First we develop a new approach to select the model that makes the best use of limited micro-level data when estimating production technology. While the trade-off between precision and heterogeneity has been discussed in the literature, we are not aware of a study that uses quantitative metrics of both precision and heterogeneity to compare model performance. Second, we propose a new alternative to simultaneously estimate technical inefficiency and frontier parameters. The joint estimation of frontier parameters and technical efficiency parameters has been done in several previous papers. In Oude Lansink, Silva and Stefanou (2001), the technical efficiency parameters are estimated in the GME framework by constraining them to be between zero and one. In Campbell, Rogers and Rezek (2008), the technical efficiency is specified as a one-sided component in the composite error term (so it is similar to the specification in the Stochastic Frontier approach) and then the one-side error term is estimated together with other parameters used GME. In our paper, we propose a new alternative. We use DEA to estimate the technical efficiency parameters and use GME to estimate the frontier parameters. The estimation of the two sets of parameters is done simultaneously. We also use the bootstrap technique developed in a series of papers by Simar and Wilson (2007; 2000a; 2000b) to obtain bias-corrected estimates of technical efficiency parameters and construct confidence intervals.

It should be noted that we do not aggregate data in our analysis. The same set of household level data is used to estimate the province-level, county-level and community-level models. We would like to emphasize that this paper does not address the issue of data aggregation or aggregation bias, a topic often seen in the literature (e.g., Stoker 1993; Wu and Adams 2002).

The rest of the paper is organized as follows. In the first section we define the alternative policy models and the metrics we use to measure their predictive accuracy and their ability to capture heterogeneity in the data. In the next section, we introduce the approach that we use to estimate the different models. In the following section estimation results are used to calculate the metrics, which are in turn used to compare the performance of the policy models. The final section concludes.

2. Policy Models, Heterogeneity and Predictive Accuracy

In this section, we first introduce several alternative model specifications that are used to estimate production functions, but differ in their predictive accuracy and ability to capture heterogeneity. We then discuss the formal metrics of precision and heterogeneity that we use to measure the performance of the different models. Because we use the criteria that are often used to assess policy analyses to evaluate these different specifications of models, in the rest of the paper, we will call them policy models.

2.1. Policy models

Most rural households are engaged in producing multiple crops. In our study area, the major crops are wheat, maize and cotton. There are two cropping seasons. Wheat is planted during the late fall/early winter of the previous year and harvested in spring of the current year. Maize and cotton are planted in summer and harvested in the fall. Five inputs are used in production: land (x_L), labor (x_l), water (x_w), fertilizer (x_f) and material (x_m). Material costs include expenditures on

machinery, seed, plastic sheeting, herbicides and pesticides. In our analysis, a quadratic function frontier is used:⁴

$$\begin{aligned} Y_{jn} &= \theta_n f_j(x_{ijn}) + e_{jn} \\ &= \theta_n \left(\sum_i \alpha_{ij} x_{ijn} - \sum_i \sum_{i'} x_{ijn} z_{ii'j} x_{i'jn} \right) + e_{jn} \end{aligned} \quad (1)$$

The observed output of crop j produced by household n is denoted by Y_{jn} . The function f_j is the production frontier for crop j . The frontier is a function of x_{ijn} , which is the i th input needed to produce crop j . The error term, e_{jn} , captures variation in outputs due to random events such as weather. The crop-specific parameters, α_{ij} and $z_{ii'j}$, are associated with the linear and quadratic/interaction terms of inputs respectively.

The parameter, θ_n , is the technical inefficiency parameter that captures the degree of deviation of each household's actual production from the maximum possible production, the production frontier. More importantly, since technical inefficiency is often the result of a lack of managerial ability (Farrell 1957; Leibenstein 1966), θ_n reflects the inter-household differences in managerial ability.⁵ Accounting for technical inefficiency is important since it can help us overcome a common problem associated with estimating a production function—the potential bias due to the existence of omitted variables. In particular, household managerial ability is often omitted because it is not directly observable to econometricians. Since the managerial ability affects both output level and the producer's choice of input, omitting it will bias the estimates of production function parameters (Griliches 1957). Since the estimated technical inefficiency parameters can capture unobservable heterogeneity that is relevant in our analysis

⁴ Or in matrix notation, the production function is $Y_{jn} = \theta_n (\mathbf{A}'_j \mathbf{X}_{jn} - \mathbf{X}'_{jn} \mathbf{Z}_j \mathbf{X}_{jn}) + e_{jn}$, where $\mathbf{X}_{jn}$ is the vector of all inputs, $\mathbf{A}$ is the vector of linear coefficient and $\mathbf{Z}$ is the matrix of quadratic coefficients in the production frontier.

⁵ Greene (2005) proposed a method to disentangle unobservable heterogeneity from technical inefficiency using panel data methods. We did not do so in our analysis for two reasons. First, we only have cross-sectional data. Second and more importantly, the main purpose of the paper does not require the separation of unobservable heterogeneity from technical inefficiency.

(that is, managerial ability), we believe our estimates of production function parameters will be less affected by omitted variable bias.⁶

In this paper, we focus mainly on the heterogeneity that comes from the differences in the parameters of the production function frontier.⁷ Since we have a set of household level data, we can estimate the production function at three different levels. We will call these three sets of models: *the community-level models*, *the county-level models* and *the province-level models*.⁸ At the province level, we estimate a single set of crop-specific production frontier parameters, assuming that all households in the province use the same the production technology. The province-level model is represented by equation (1) above.

In the second set of policy models, we relax the constraint that the production technology parameters in all communities are the same. We allow production technology to vary by county but restrict them to be equal across communities within the same county. In other words, we estimate one set of crop specific production frontier for each county. The county-level models are expressed as follows:

$$Y_{jn} = \theta_n \left(\sum_i \alpha_{ij,c} x_{ijn} - \sum_i \sum_{i'} x_{ijn} z_{ii'j,c} x_{i'jn} \right) + e_{jn} \quad (2)$$

Note in equation (2), a new subscript c is added to the linear parameter, $\alpha_{ij,c}$, and to the quadratic parameter, $z_{ii'j,c}$, to denote that they vary by counties. In our case, c ranges from one to three since we have three counties in total.

⁶ Conventional panel data models, such as fixed-effects or random-effects models, have been employed to account for unobserved heterogeneity. Unfortunately, for most households in our sample, we do not have more than one observation for a single crop. Therefore, it is not possible to use a household fixed effects approach.

⁷ Heterogeneity could also come from differences in farm size. However, unlike other countries (such as India), land is relatively equally allocated among households in rural China both in terms of land size and soil quality (Benjamin and Brandt, 2002; Kong, 1996). As a result of the equal land distribution, farm size does not vary significantly within the communities or across our sample communities (since all communities are within the same province). So we do not account for size heterogeneity in our analysis.

⁸ Other approaches can be used to incorporate heterogeneity into estimation. For example, a random coefficient model could be used. Alternatively, a joint probability distribution of the heterogeneous factors could be assumed. However, the main conclusions of our paper would not change even if these alternative approaches were used.

In the third set of policy models, we estimate one set of crop specific production function parameters for each community within the province, indexed by v . The community-level models are expressed as follows:

$$Y_{jn} = \theta_n \left(\sum_i \alpha_{ij,v} x_{ijn} - \sum_i \sum_{i'} x_{ijn} z_{ii',v} x_{i'jn} \right) + e_{jn} \quad (3)$$

2.2. Trade-off in the policy models

All three sets of policy modes are linear in nature and can be estimated using the same method. In addition, all three policy models can be used to predict household level activities such as crop production and input uses. However, these models are distinctively different. The pooled estimation approach used in the province-level model (equation 1) ignores any potential heterogeneity in production technology at the county or community level. In contrast, the county-level or the community-level models (equation 2 and 3) can capture the heterogeneity that is ignored by the province-level models.

In addition to the differences in the ability of the alternative models to capture heterogeneity in the data, the policy models also differ in the sizes of the samples that are used during estimation. In the province-level model, all the wheat, maize and cotton producing households in the sample are included to generate estimates of production frontier parameters for each crop. In total, 63 households grew wheat, 86 households grew maize and 18 households grew cotton in 2004 (Table 1, row 1). In the county-level models, the sample sizes included in the estimation are less than half that used in the province-level estimation. The sample size in Xian County, for example, is 26 for wheat; 28 for maize and eight for cotton (row 2). The sample sizes drop even further in the community-level models (row 5 and 6). On average, the sample size is 2.6 for wheat; 3.6 for maize; and 0.75 for cotton. Since the sample size drops as we move away from the province-level model, estimation of the county- or community-level

models potentially suffer more from the small sample problem although they can account for more of the heterogeneity across sample units.

2.3. Two measures for choosing the best policy models

Policy models need to be able to provide accurate prediction as well as allow analysis of distributional effects. In the rest of this section, we introduce the two criteria that can help us assess how policy models help policy makers achieve these objectives.

2.3.1. Quantitative metric that measures heterogeneity

Since the homogeneity problem is caused by the heterogeneity among sample units not captured in the analysis, we first measure quantitatively the relative ability of the three different sets of models to recover the heterogeneity embodied in the data. To do so, a quantitative measure of the proportion of heterogeneity that is recovered by each model is needed. In defining such a measure, we construct an index that will be higher when a set of models captures more of the heterogeneity in the sample and smaller when capturing less. The measure that we use in this paper is based on the *disaggregated information gain* (DIG) developed by Howitt and Reynaud (2003). The DIG measure is constructed from the concept of Kullback-Leibler information or cross-entropy, a well-known measure of information. In the multi-crop environment, heterogeneity can be seen in differences in production shares of crops. The cross-entropy between the estimated distribution and the true distribution of production share is defined as:

$$\hat{CE} = \sum_v \sum_j \hat{s}_j^v \cdot \ln\left(\frac{\hat{s}_j^v}{s_j^v}\right) \quad (4)$$

In equation (4), the symbol s_j^v denotes the actual production share of crop j in community v .

Greater differences in the actual values of s_j^v across the sample communities indicate greater heterogeneity. The estimated production share of crop j using the three sets of models (the

community- or the county- or province-level models) in community v is $\hat{s}_j^v$. Equation (4) seeks to evaluate how well each set of models ($\hat{s}_j^v$) can capture the differences in the values of s_j^v that are actually observed across communities and thus capture the heterogeneity in the data.

The DIG measure is defined in a way that it represents the proportion of community-level heterogeneity that is recovered. Mathematically, it is expressed as:

$$\text{DIG} = 1 - \frac{\sum_v \sum_j \hat{s}_j^v \cdot \ln\left(\frac{\hat{s}_j^v}{s_j^v}\right)}{\sum_v \sum_j s_j \cdot \ln\left(\frac{s_j}{s_j^v}\right)} \quad (5)$$

When we use the province-level model, the production share of crop j is at the province level and s_j is used to estimate the community-level production shares. As can be seen from equation (5), when the province-level share is used, $\hat{s}_j^v = s_j$ absent any prediction errors, the DIG measure is equal to zero (or zero percent of the heterogeneity in the data is captured). In the case of community-level models, if $\hat{s}_j^v = s_j^v$ in the absence of prediction errors, the DIG measure is equal to 100% (or all of the community-level heterogeneity in the data is captured). When the county-level model is used, the DIG measure will be between zero and 100 percent.

2.3.2. Quantitative metrics that measure out-of-sample prediction

To measure the out-of-sample prediction performance, we use a “leave-one-out” cross validation procedure (Stone 1974). Implementing this procedure requires several steps. In the first step, we estimate the crop-specific production frontier coefficients. During the estimation, one household is removed from the sample. This household is reserved for out-of-sample prediction in the second step. The estimates of the production function parameters are then used in conjunction with the values of inputs of the household that is not used in estimation to predict its crop

production. This process is repeated for each household in the sample. For example, when estimating the production frontier at the province level, we re-estimate the province-level model 88 times since there are 88 households in total and the estimates of each round is used to predict the production of the corresponding out-of-sample household.

We use three different measures to calculate the out-of-sample prediction error. The *Percentage Absolute Prediction Error* (PAPE) measure is calculated as the absolute difference between the predicted and observed household-level crop production divided by the observed household-level crop production (i.e., the prediction error is measured in terms of the proportion of the observed household crop production level). Another popular measure is the mean squared prediction errors (MSPE). Mathematically, PAPE and MSPE are defined as:

$$\text{PAPE} = \frac{1}{N_j} \sum_{n=1}^{N_j} \frac{|Y_{jn} - \hat{Y}_{jn}^{(-n)}|}{Y_{jn}}; \quad \text{MSPE} = \frac{1}{N_j} \sum_{n=1}^{N_j} (Y_{jn} - \hat{Y}_{jn}^{(-n)})^2; \quad (6)$$

where $\hat{Y}_{jn}^{(-n)}$ is the out-of-sample prediction of Y_{jn} and N_j is the number of households that grow crop j . The superscript $-n$ is used because the data on the n th household is not used in the estimation of the parameters.

It should be noted that MSPE is comparable to two other classical model selection criteria, namely the Akaike Information Criterion (AIC - Akaike 1973) and Schwarz's Bayesian Information Criterion (BIC - Schwarz 1978). We show in Appendix 1 that the objective functions of AIC, BIC and MSPE are essentially equivalent when the sample size is substantially larger than the number of parameters to be estimated. Leeb (2008) shows that if the sample size is small relative to the number of parameters to be estimated, then AIC and BIC do not perform well as measures of the out-of-sample predictive accuracy. This is true partly because AIC and BIC do not primarily aim at estimating the out-of-sample squared prediction errors. In addition,

in the case of an ill-posed estimation problem (that is, the sample size is smaller than the number of parameters to be estimated—which is the case in some of the policy models in this study), the AIC and BIC measures are not well-defined. Because of these reasons, we do not include these model selection criteria in our analysis.

In addition to PAPE and MSPE measures, we also use a third statistic that is similar in structure to the well-known Theil's U statistic (U - Theil 1966). The Theil's U statistic is calculated as the square root of MSPE from an alternative model (that is being considered) divided by the root MSPE from a naïve model. The rationale for the Theil's U statistic is that a model should perform at least as well as the simplest models. The Theil's U statistic is often used in time series analysis and the naïve model is a random walk forecast (or a no-change model): $\hat{Y}_{t+1} = Y_t$, where t denotes time period. Since our data is cross-sectional, we define the naïve out-of-sample model that predicts the output of crop j of household n as: $\hat{Y}_{jn}^{Naive} = x_{Ljn} \bar{y}_{jh}$, where x_{Ljn} is the observed cultivated area of crop j of household n and $\bar{y}_{jh}$ is the average yield of all households (excluding household n). A U statistic that is less than one indicates that predictions that come out of our models (the province-level, the county-level or the community-level models) are better than the predictions from the naïve model. If U is more than one, predictions from the naïve model are better than those from our models. The Theil's U in this study is defined as:

$$U = \frac{\sqrt{\frac{1}{N_j} \sum_{n=1}^{N_j} (Y_{jn} - \hat{Y}_{jn}^{(-n)})^2}}{\sqrt{\frac{1}{N_j} \sum_{n=1}^{N_j} (Y_{jn} - \hat{Y}_{jn}^{Naive})^2}} \quad (7)$$

Since the enumerator of U does not vary in different policy models and the numerator is the square root of MSPE, MSPE and Theil's U contain exactly the same information. Therefore, for brevity we only report the Theil's U statistics in the results section since they are scale-free. Both PAPE and Theil's U statistics are calculated by crop for each household.

We also test statistically the equality of predictive accuracy of different policy models. Two tests are used. The first test is the asymptotic loss differential test proposed in Diebold and Mariano (1995). The null hypothesis of this test is that the sample mean of the loss differential series is zero. For example, when we use the PAPE measure to compare the predictive accuracy of the county-level models against the predictive accuracy of the community-level models, we define the loss differential series, $\{d_n\}_{n=1}^N$, as $d_{n,County} = \hat{PAPE}_{n,County} - \hat{PAPE}_{n,Community}$, and then proceed to the test.

The second test is the Wilcoxon's Signed-Rank Test (Wilcoxon 1945). This test is a non-parametric alternative to the usual t test for differences in predictive accuracy of the same outcome variable from two different sets of policy models.

3. Estimation Approach

3.1. Generalized Maximum Entropy estimation of the production frontier parameters

Due to the small sample size, estimating a production frontier at the community or county level makes the use of some estimation methods difficult or even infeasible. Since five inputs are used in production and a quadratic functional form is specified, there are 20 parameters to be estimated for each production frontier. In this case, if classical econometric methods (e.g., maximum likelihood estimation method) were used, the estimation problem would be ill-posed due to an insufficient number of data points. For example, when we estimate a production

frontier for wheat in Tang and Ci County, there are only 18 and 19 data points respectively (Table 1, row 3 and 4, column 2). This means that there are fewer observations than the number of parameters.

As a solution, we choose to use the method of GME that was developed by Golan et al. (1996). The GME estimator allows for estimation with any sample size. The GME estimator emphasizes both prediction and precision in its objective function and thus has the properties of being both subject to limited bias and minimum variance (page 109, Golan, et al. 1996). One of the main advantages of the GME method is the ease to impose inequality constraints, equality constraints, and nonlinear constraints (Golan, Judge and Perloff 1997; Golan, Perloff and Shen 2001). The GME estimator has come into increasing use by researchers who seek to estimate highly disaggregated models in the face of limited data (e.g., Howitt and Reynaud 2003; Miller and Plantinga 1999). It should be noted that the approach we develop that helps select the best policy models is not necessarily associated with the use of GME. If the number of data points were more than the number of parameters even when a researcher was estimating parameters at the community level, classical econometric methods could be used in this study. The use of GME ensures that the same method is used to estimate all three sets of policy models and so any difference in the performance of these models cannot be attributed to differences in estimation methods.

In our GME estimation, in addition to the *data-consistent constraints*, which are specified in equation (1-3), two additional sets of constraints are specified that will aid us in: a.) making our estimates consistent with economic theory (the *theoretical constraints*); and b.) implementing the GME estimation (the *numeric estimation constraints*). In the rest of the

section, we will only present relevant equations for the province-level models. Similar presentation apply, mutatis mutandis, to the county-level and the community-level models.

Theoretical Constraints

Two sets of theoretical constraints are used so that the estimated production technology is consistent with profit maximization behavior of households. Our assumption of profit maximization is supported by other studies on production behavior in China (e.g., Huang and Rozelle 1996; Lin 1992).⁹ The *optimality condition constraints* state the first order condition of households' profit maximization problem:

$$x_{ijn} = \frac{1}{2z_{ii}} \left[\alpha_{ij} - 2\sum_{i' \neq i} x_{i'n} z_{ii'} - \frac{c_{in}}{p_j \theta_n} \right] + v_{ijn} \quad \forall i = \text{Fertilizer, Capital} \quad (8)$$

where the output price for crop j is p_j and the cost for input i that household n faces is c_{in} .¹⁰ It should be noted that only the optimality conditions for variable inputs are included. The optimality conditions for fixed allocatable inputs include the shadow values, which are not directly estimatable.

The second set of theoretical constraints is the *curvature constraint*, which requires that the quadratic matrix $\mathbf{Z}$, the matrix of all parameters on the squared and interaction terms of inputs ($z_{ii'}$ s) be positive (semi)definite. The curvature constraint is imposed using the Cholesky decomposition of $\mathbf{Z}$ (Paris and Howitt 1998). The Cholesky decomposition is defined by $\mathbf{Z} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is an $I \times I$ lower triangular matrix, I is the total number inputs. The positive semidefinite property of $\mathbf{Z}$ is guaranteed by constraining the diagonal elements of $\mathbf{L}$ to be nonnegative ($L_{ij} \geq 0$).

⁹ Households are assumed to be risk neutral. In our preliminary analysis, we assume households are risk averse. The risk aversion parameter, however, did not come out statistically different from zero in estimation results. Therefore, without loss of generality, we decided to assume risk neutrality in the maximization problem.

¹⁰ Households are assumed to be risk neutral. In our preliminary analysis, we assume households are risk averse. The risk aversion parameter, however, did not come out statistically different from zero in estimation results. Therefore, without loss of generality, we decided to assume risk neutrality in the maximization problem.

The Cholesky decomposition also ensures the symmetry of the matrix $\mathbf{Z}$. In addition, we also impose the monotonicity constraints: $p_j \theta_n [\alpha_{ij} - 2 \sum_{i'} x_{i'n} z_{ii'}] - c_{in} > 0$.¹¹

Numeric estimation constraints

When the GME method is used, instead of directly estimating the mean and variance of the coefficient, a probability distribution is estimated for each coefficient and the error term. During estimation, a range of feasible values is chosen as the *support values* of the probability distribution and then a probability (usually uniform) is assigned to each value.¹² The coefficients and the error terms are then reparameterized in terms of unknown probabilities over the support values. This set of *reparameterization constraints* are defined as $\alpha_{ij} = \sum_m p_{\alpha_{ij}}^m \bar{\alpha}_{ij}^m$,

$$z_{ii'j} = \sum_m p_{z_{ii'j}}^m \bar{z}_{ii'j}^m, v_{ijn} = \sum_m p_{v_{ijn}}^m \bar{v}_{ijn}^m \text{ and } e_{jn} = \sum_m p_{e_{jn}}^m \bar{e}_{jn}^m, \text{ where } m \text{ is the index of the support}$$

values and the p s are the unknown probabilities to be estimated. Symbols with upper bars denote the support values. The unknown probabilities satisfy properties of probabilities: they are positive and all probabilities associated with the same coefficient or error term add up to one (the *adding up constraints*).

3.2. The use of Data Envelopment Analysis to estimate technical inefficiency parameters

We use the classic Farrell definition of the output distance function to measure θ_n . The output distance function for the n th household that produces a single output, $D_0(\mathbf{x}_n, y_n)$, is defined as $\inf_{\theta_n} \{\theta_n > 0 : (y_n / \theta_n) \in P(\mathbf{x}_n)\}$. The parameter θ_n denotes the inverse of the factor by which the output y_n could be increased while still remaining within the feasible production set, $P(\mathbf{x}_n)$, for the given input level $\mathbf{x}_n$. The value of $D_0(\mathbf{x}_n, y_n)$ is less than or equal to one if y_n is an element of $P(\mathbf{x}_n)$. If $D_0(\mathbf{x}_n, y_n)$ is one, then the observation $(\mathbf{x}_n, y_n)$ is on the production frontier.

¹¹ In our empirical estimation, the monotonicity constraints hold for all of the observations.

¹² We follow Golan, Judge and Miller (1996) and choose five support points both for the parameters to be estimated and for the error term.

Data Envelopment Analysis (DEA) involves measurement of efficiency for a given observation in the sample data relative to the boundary (that is, the production frontier) of the convex hulls of the data intersected with the free-disposal hull. The distance function, $D_0(\mathbf{x}_n, y_n)$, is estimated by solving a Linear Programming (LP) problem (Färe and Primont 1995; Farrell 1957). The LP problem can be expressed as:

$$\begin{aligned} & \text{Max}_{\theta_n, z_1, z_2, \dots, z_H} \quad 1/\theta_n \\ & \text{Subject to:} \quad \begin{cases} \sum_{h=1}^H z_h y_{jh} \geq y_{jn} / \theta_n, j=1, 2, \dots, J \\ \sum_{h=1}^H z_h x_{ih} \leq x_{in}, i=1, 2, \dots, I \\ \sum_{h=1}^H z_h \leq 1 \\ z_h \geq 0, h=1, \dots, H \end{cases} \end{aligned}$$

where the variable, z_h , represents the intensity level of the production activity of household h .

The restriction, $\sum_{h=1}^H z_h \leq 1$, ensures that efficiency is measured against a Nonincreasing Returns to Scale (NIRS) technology, which is consistent with the quadratic functional form of the production frontier specified in equations (1) – (3). Both h and n are indices for households in the sample.

3.3. Summary of the estimation approach

In GME estimation, the estimates of probabilities given their support values are obtained through maximizing the sum of the entropy values for the distributions of the coefficients and the error terms, conditional on the data-consistent constraints, the theoretical constraints and other constraints.¹³ Estimates of technical efficiency are obtained through maximizing the inverse of

¹³ We believe that our approach of estimating the production frontier does not suffer from the fundamental identification problems raised by Marschak and Andrews (1944) for several reasons. First of all, a system of equations is estimated including both the production frontier equation and the optimality condition equations. Second, in a separate set of analysis, we instrumented for inputs using a set of variables (input prices, whether there is a production shock or not, distance from house to plots, etc.), the results do not differ much from the case using the raw input uses. Third, since levels of output used in estimation (after correcting for the impacts of production

the output distance function. In the GME estimation, it is possible to treat differently some of the parameters. The technical inefficiency parameter, θ , is not reparameterized using the probability support vector. Instead of estimating probabilities distribution for θ , we obtain only point estimates of those unknowns. The empirical distribution of estimated of θ is obtained through bootstrap in our paper. The objective functions of the LP Problems of all the households in our sample can be easily added since it is also an optimization problem. In summary, the estimation problem can be expressed as:

$$\begin{aligned} \text{Max}_{p_{\alpha_i}^m, p_{z_{ii'}}^m, p_{e_n}^m, p_{v_{jn}}^m} & H(p_{\alpha_i}^m, p_{z_{ii'}}^m, p_{e_n}^m, p_{v_{jn}}^m) \\ & = -\sum_{j=1}^J \sum_i \sum_m p_{\alpha_{ij}}^m \ln p_{\alpha_{ij}}^m - \sum_{j=1}^J \sum_i \sum_{i'} \sum_s p_{z_{ii'}}^m \ln p_{z_{ii'}}^m - \sum_{n=1}^N \sum_m p_{e_n}^m \ln p_{e_n}^m - \sum_{n=1}^N \sum_i \sum_m p_{v_{jn}}^m \ln p_{v_{jn}}^m \\ & \quad + \sum_{n=1}^N (1/\theta_n) \end{aligned}$$

Subject to

$$\text{Data consistent constraints: } Y_{jn} = \theta_n \left(\sum_i \alpha_{ij} x_{ijn} - \sum_i \sum_k x_{ijn} z_{ikj} x_{kjin} \right) + e_{jn}$$

Theoretical constraints

$$\text{First order conditions: } x_{ijn} = \frac{1}{2z_{ii'}} \left[\alpha_{ij} - 2 \sum_{i' \neq i} x_{i'n} z_{ii'} - c_{in} / p_j \theta_n \right] + v_{ijn}$$

$$\text{Curvature conditions: } \mathbf{Z} = \mathbf{L}\mathbf{L}'; \quad L_{ij} \geq 0$$

$$\text{Monotonicity: } p_j \theta_n \left[\alpha_{ij} - 2 \sum_{i'} x_{i'n} z_{ii'} \right] - c_{in} > 0$$

$$\begin{aligned} \text{Reparameterization: } \alpha_{ij} &= \sum_m p_{\alpha_{ij}}^m \bar{\alpha}_{ij}^m, \quad z_{ii'} = \sum_m p_{z_{ii'}}^m \bar{z}_{ii'}^m, \\ v_{ijn} &= \sum_m p_{v_{jn}}^m \bar{v}_{ijn}^m \text{ and } e_{jn} = \sum_m p_{e_{jn}}^m \bar{e}_{jn}^m \end{aligned}$$

$$\begin{aligned} \text{Adding up constraints: } \sum_m p_{\alpha_{ij}}^m &= 1; \quad \sum_m p_{z_{ii'}}^m = 1; \\ \sum_m p_{e_{jn}}^m &= 1; \quad \sum_m p_{v_{jn}}^m = 1 \end{aligned}$$

$$\begin{aligned} \text{Output distance function } \sum_{h=1}^H z_h (Y_{jh} - e_{jh}) &\geq (Y_{jn} - e_{jn}) / \theta_n; \quad h = 1, \dots, H \\ \sum_{h=1}^H z_h (x_{ijh} - v_{ijh}) &\leq (x_{ijn} - v_{ijn}) \\ z_h &\geq 0; \quad \sum_{h=1}^H z_h \leq 1 \end{aligned}$$

shocks) are close to the expected levels of outputs, we believe there is no simultaneous equation bias pointed out by Hoch (1958), which is associated with using actual output instead of expected output in estimation. Last but not least, the reasonable range out-of-prediction errors (reported in the next section) further shows that there is no serious bias in our estimates. This confirms what Golan (1996) has stated (page 194): “this formulation (of GME estimation) may lead to parameter estimates that are slightly biased but have excellent precision.”

Solving our estimation problem is analogous to estimating a system of equations in classical econometrics. For example, when Full Information Maximum Likelihood (FIML) is used, the objective function is the likelihood function that represents joint probability density of data from multiple equations. Similar to the reason that FIML is more efficient than a single equation estimation method, our strategy of simultaneously estimating the two sets of parameters also generates more efficient estimates. Note that the production frontiers of all crops are estimated jointly. In the set of constraints imposed on the output distance function, we subtract e_{jn} from Y_{jn} and v_{ijn} from x_{ijn} . This subtraction also avoids attributing any statistical noise to deviations from the frontier, a weakness of DEA that is pointed out by many researchers.¹⁴

The use of DEA together with GME allows us to estimate technical inefficiency and frontier parameters simultaneously. This is true for three reasons. First, constraints used in the DEA approach are consistent with the theoretical constraints (concavity and monotonicity) in the GME estimation approach. The constraints in the LP problem specify disposability, which corresponds to the monotonicity of the production frontier, and the convexity of the feasible production set, which corresponds to the concavity of the frontier. Second, the simultaneous estimation of the technical efficiency parameters and the frontier parameters can be done with relative ease. This is true because the objective function of the LP problem and its constraints are easily incorporated into that of the GME framework since GME is also solved through a maximization problem. Third, bootstrapping can be used to obtain confidence intervals of technical efficiency and frontier coefficients simultaneously. By construction of the problem of

¹⁴ Having a set of panel data can also overcome since it enables separating the efficiency part and the stochastic part of the composed error term. In the case when error term is not independent of the inputs, having a set of panel data helps too. In general, panel data or time series data would produce better estimation results than cross-sectional data. In some cases, even panel data cannot solve the problem. One typical cause of the correlation between error term and inputs is that inputs are not measured perfectly. It is widely believed in the literature that panel data methods exacerbate measurement error bias, especially when measurement error is positively serially correlated (Wooldridge 2002).

estimating technical efficiency parameters, estimates of technical efficiency for some households will be at the boundary. The boundary nature of the estimator introduces some problems. To overcome the boundary problem, Simar and Wilson (2007; 2000a; 2000b) used a smoothed bootstrap to resample from the original efficiency estimates. In particular, they adopted the reflection method proposed by Silverman (1986). The procedure first reflects each of the n observations through the boundary at unity and then estimate the unknown density of the $2n$ unbounded points. We used a bootstrapping procedure that combines algorithms developed in Simar and Wilson (2007; 2000b) to obtain bias-corrected estimates of technical efficiency parameters and construct confidence intervals. It should be noted that this is different from the bootstrapping technique that is often used in GME estimation to obtain standard errors of coefficients. Details of bootstrapping are not presented here but are available from authors upon request. The General Algebraic Modeling System (GAMS) software is used. STATA is used to generate the bootstrapping samples.

4. Data Description

The data used in the study come from the 2004 China Water Institutions and Management (CWIM) Survey.¹⁵ The survey was run by the authors jointly with researchers from the Center for Chinese Agricultural Policy in Beijing. We collected data in 24 communities in Hebei province, a province on the eastern coast of northern China that covers most of the Hai River Basin and produces about 12 percent of the nation's grain (Ministry of Agriculture of China 2004).

¹⁵During the CWIM survey, in total we collected data in 80 communities in three provinces in north China. Because the objective of this paper is to find policy models that can recover the heterogeneity at micro-level such as the county level or the community level, so we chose to only focus on one province.

The scopes of the CWIM survey were quite broad. In each community, enumerators interviewed community leaders, surface water irrigation managers and/or groundwater irrigation management and households. Community leader survey questionnaire, water manager survey questionnaire and household survey questionnaire are used for different type of respondents. We collected information on the nature of water resources, type of water management institutions and water management activities, government policies, adoption of water saving technology and the condition of village irrigation infrastructure.

The key section of the survey for this paper came from household level production activities we collected. We asked lots of questions during the survey in order to obtain detailed and accurate production information. For example, for labor input alone, we asked about 30 questions for each crop and by production activities (e.g., plow, harvest, apply fertilizer). On average, a household interview took two or three hours to complete. We also interviewed the community leader and several well operators within each community to supplement information from the household interviews. This makes the survey even more time-consuming and also makes it much more costly to scale up the sample size. As a result, we were only able to interview four households in each community.

The major crops in Hebei province are wheat, maize and cotton. For each crop, the enumerator also asked each household about the cultivated area (irrigated and rainfed areas), the yield and the cost and quantity of each type of input: fertilizer, labor (by activity), machinery (use of own equipment or rent), pesticide and plastic sheeting etc. To construct a measure of the volume of water applied, enumerators asked the farmer to report for each crop the average length of irrigation time, the total number of irrigations during the entire growing season and the

average volume of water applied per irrigation. In addition, households reported their payment for irrigation water for each crop, which is the cost of water.

5. Selecting the best policy models

Crop-specific production frontiers were estimated using the approach described in the previous section. The GME estimation using the community-, county- and province-level models all perform satisfactorily and produce reasonable estimates of the production frontier coefficients. For the sake of brevity, we only report the results for the wheat production frontier in Xian County from the county-level model in Table 2. The 95% confidence intervals are constructed from the empirical bootstrap distributions of estimated parameters. Most of the estimated coefficients are statistically significant since their 95% confidence intervals do not contain the value zeros. For example, the linear coefficients (the α_i s) are all positive and statistically significant, which is a necessary but not sufficient condition for the property that production is increasing with the quantity of inputs. The quadratic coefficients on the interaction terms of material and labor, z_{ml} , are negative and statistically significant for Xian and Tang counties. This indicates that labor and material are substitutes (which we know because from the optimality condition, we can derive that the sign on the cross price elasticity of input i and i' is the opposite of the sign on $z_{ii'}$). The substitutability of labor and material is a result that is consistent with the finding that in most developing countries labor and capital (material cost in our case can also be considered as capital cost) are substitutes (e.g., Garcia-Penalosa and Turnovsky 2005; Khandker and Binswanger 1992). Additional details on the GME estimation are reported in Appendix 2 and 3.

Despite the fact that the performance of all the estimation procedures was fairly good for each of the models (the province-level, county-level and community-level models), comparing communities within counties (or across counties with the province), there are marked differences in the specific realized values of the parameter estimates. Importantly, the result of entropy ratio test, which is a test that is similar in nature to the more familiar likelihood ratio tests, rejects the null hypothesis that the frontier parameters are same across counties (Appendix 4). The differences in the results from the three models means that policy recommendations generated from each different set of estimates are likely to vary.

When applying the estimates from the community-, county- and province-level to calculate the value of the DIG measure, as expected, we can explain more of the heterogeneity in the data using models that are less constrained. The set of community-level models recovers the highest proportion of the heterogeneity in the data. The value of the DIG measure is 85 percent. It is not 100 percent because there is a slight deviation between the estimated production share, $\hat{s}_j^v$, and the observed production share, s_j^v . For the same reason, the value of the DIG measure for the province-level model is 54 percent. Most importantly, when using the county-level model, the value of the DIG measure is 71 percent. Hence the loss of information is fairly small when moving up from the community to county levels. In our model, and using our data set (in northern China), the set of county-level models captures 83.5 percent (71/85) of the heterogeneity that is explained by the community-level model.

When explaining the accuracy of predictions, the value of the PAPE measure is the highest for the set of county-level models; second for the set of province-level models and lowest for the set of community-level models. The ranking of performance can almost certainly be explained by the trade-off between the homogeneity problem and the small sample problem.

Although the set of community-level models captures most of the heterogeneity, according to our analysis, the benefit (to policy makers) does not outweigh the cost of introducing the small sample problem. Since for most of the community-level models the sample size is only three or four observations, it is not surprising that losing one observation in the estimation stage (for the purpose of generating out of sample predictions) has a relatively large impact on the accuracy of the estimates. The estimates from the set of community-level models suffer severely from the small sample problem. For example, the average value of the PAPE measures generated by the set of community-level models is as large as 51 percent for cotton (Table 3, row 3, column 3). At the other end, although the set of province-level models have much larger sample sizes, they still generate large prediction errors. This is because the same production frontier parameters are applied to different counties in the out of sample prediction. The average values of the PAPE measures generated by the set of province-level models, while smaller than those generated by the set of community-level models, are all more than 20 percent (row 1-3, column 1). Such large errors may lead to highly inaccurate estimates in policy analysis.

In contrast, the set of county-level models recovers a relatively large part of the heterogeneity in the data and thus their estimates do not suffer as much from the homogeneity problem as the estimates of the set of province-level models do. In addition, the set of estimated parameters from the county-level models depend on a larger sample size than those from the set of community-level models. Because of this, it follows that the estimates of the county-level models do not suffer so severely from the small sample problem as the estimates of the community-level models. The values of the PAPE measures generated by the estimates of the county-level models are all smaller than 20 percent. It is only 10.5 percent for wheat, 13 percent for maize and 16 percent for cotton (row 1-3, column 2). Hence, it is reasonable in our case to

conclude that the set of county-level models can be said to perform the best in terms of out-of-sample prediction.¹⁶

The comparison using Theil's U statistics generates similar conclusions in terms of relative performance of the alternative models. The county-level models still generate the same results. The Theil's U statistics from the county-level models are all smaller than one (Table 3, row 4-6, column 2). The Theil's U statistics from the community-level models, however, are larger than one for wheat and cotton (row 4 and 6, column 2), which indicates that wheat and cotton production could be better predicted by just using the average yield of the in-sample observations and the observed cultivated area (prediction from the naïve model). The province-level model also yields a U statistic that is larger than one for cotton (row 3, column 1).

The differences in the predictive accuracy of alternative policy models are also statistically significant. Results of the asymptotic loss differential tests show that when we compare the predictive accuracy of the county-level and the community-level models, the difference is statistically significant at 1% for wheat and maize and 10 percent for cotton (Table 3, row 1-3, column 2). When we compare the predictive accuracy of the province-level and the community-level models, the difference is statistically significant at five percent for wheat and maize and 10 percent for cotton (Table 3, row 1-3, column 2). The results from the Wilcoxon's signed-Rank Test are similar but with much smaller p-values for most of the tests (not reported in the table). Results of these tests support our conclusion that the county-level models should be used.

¹⁶ It would be desirable to construct a metric that combines these two measures that assess the ability to recover heterogeneity (DIG) and the predictive accuracy (PAPE or MSPE or U). Such a measure, however, will depend on the specific policy question and the relative importance policy makers put on the ability of policy models to look at distributive consequences and the precision of predictions. So we will leave this to the choice of policy analysts.

6. Conclusion

In this paper, we develop an approach to select the policy model that makes the best use of limited micro-level data to achieve the objectives of policy analysis: to optimize the precision of policy response predictions, and also to incorporate sufficient heterogeneity to allow policy makers to consider the distributional consequences of policies. To do so, we first estimate different policy models at the community-, county- and province-levels. We then develop a series of quantitative metrics to compare the performance of these models. We use the Disaggregated Information Gain metric to measure the proportion of heterogeneity embodied in the data that is recovered by each of the policy models. We use the Percentage Absolute Prediction Error and Theil's U statistic to measure the accuracy of the out-of-sample predictions generated by each policy model.

These quantitative metrics help us evaluate the trade-off between precision and heterogeneity existing in different policy models and select the best ones. Specific to the data set we use in this study, the county level models are chosen as the best policy models since they generate high values of DIG and low values of prediction errors. Intuitively, the set of county-level models strike a balance between precision and heterogeneity. They capture a relatively more of the heterogeneity within the data (compared to the province-level model) and also use a large enough sample size so that their estimates do not suffer so severely from the small sample problem (compared to the community-level models).

Since production function is often the core of many agricultural and environmental policies, the policy models focus on the estimation of production function parameters. We proposed a new alternative to the existing literature to estimate production function parameters. In particular, our innovation is to combine the GME and DEA to jointly estimate the production

frontier parameters and the technical efficiency parameters. The use of GME allows estimation at any sample size, so we can use GME to conveniently estimate all three policy models. The use of DEA to measure technical inefficiency helps us overcome the potential bias due to the existence of omitted variables in production function estimation. Moreover, the joint use of GME and DEA is easy to implement using GAMs or other statistical and mathematical software packages.

References

- Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, and F. Csaki, ed. *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiadó, pp. 267–281.
- Campbell, R., K. Rogers, and J. Rezek (2008) Efficient frontier estimation: a maximum entropy approach. *Journal of Productivity Analysis* DOI: 10.1007/s11123-008-0103-9.
- Deaton, A. (1997) *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: The Johns Hopkins University Press.
- Diebold, F. X., and R. S. Mariano (1995) Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 13 (3): 253-263.
- Färe, R., and D. Primont (1995) *Multi-output Production and Duality: Theory and Applications*. Boston/London/Dordrecht: Kluwer Academic Publisher.
- Farrell (1957) The measurement of production efficiency. *Journal of the Royal Statistics Society A* 120 253-281.
- Fraser, I. (2000) An Application of Maximum Entropy Estimation: The Demand for Meat In the United Kingdom *Applied Economics* 32 (1): 45-59.
- Garcia-Penalosa, C., and S.-J. Turnovsky (2005) Production Risk and the Functional Distribution of Income in a Developing Economy: Tradeoffs and Policy Responses. *Journal of Development Economics* 76 (1): 175-208.
- Golan, A. (2003) An information theoretic approach for estimating nonlinear dynamic models. *Studies in Nonlinear Dynamics and Econometrics* 7 (4): 26.
- Golan, A., G. Judge, and J. Perloff (1997) Estimation and inference with censored and ordered multinomial response data. *Journal of Econometrics* 79 (1): 23-51.
- Golan, A., G. G. Judge, and D. Miller (1996) *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley & Sons.
- Golan, A., E. Moretti, and J. M. Perloff (1999) An Information-Based Sample-Selection Estimation Model of Agricultural Workers' Choice between Piece-Rate and Hourly Work. *American Journal of Agricultural Economics* 81 (3): 735-41.
- Golan, A., J. M. Perloff, and E. Z. Shen (2001) Estimating a demand system with nonnegative constraints: Mexican meat demand *Review of Economics and Statistics* 83 (3): 541-550.
- Greene, W. (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics* 126 (2): 269-303.
- Griliches (1957) Specification Bias in Estimation of Production Function. *Journal of Farm Economics* 39 (1): 8-20.
- Hoch, I. (1958) Simultaneous equation bias in the context of the Cobb-Douglas production function. *Econometrica* 26 (4): 566-578.
- Howitt, R., and A. Reynaud (2003) Spatial Disaggregation of Agricultural Production Data Using Maximum Entropy. *European Review of Agricultural Economics* 30 (3): 359-387.
- Huang, J., and S. Rozelle (1996) Technological change: Rediscovering the engine of productivity growth in China's rural economy. *Journal of Development Economics* 49 (2): 337-369.
- Just, R. E., and J. M. Antle (1990) Interaction between Environmental and Agricultural Policies: Opportunities for Coordination and Limitations for Evaluation. *American Economic Review* 80 (2): 197-202.
- Khandker, S., and H. Binswanger (1992) The impact of formal finance on the rural economy of India. World Bank Policy Research Working Paper Series No. 949.

- Leeb, H. (2008) Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli* Forthcoming.
- Leibenstein, H. (1966) Allocative efficiency versus X-efficiency. *American Economic Review* 56 392-415.
- Lin, J. Y. (1992) Rural Reforms and Agricultural Growth in China. *American Economic Review* 82 (1): 34-51.
- Marschak, J., and W. H. Andrews, Jr. (1944) Random simultaneous equations and the theory of production. *Econometrica* 12 (3/4): 143-205.
- Miller, D., and A. J. Plantinga (1999) Modeling Land Use Decisions with Aggregate Data *American Journal of Agricultural Economics* 81 (1): 180-94.
- Ministry of Agriculture of China (2004) *China Agriculture Yearbook*. Beijing, China: China Agriculture Publishing House.
- O'Neill, T., and G. Hymel (1994) *All Politics Is Local: And Other Rules of the Game*. Holbrook, MA: Adams Media Corporation.
- Oude Lansink, A., E. Silva, and S. Stefanou (2001) Inter-Firm and Intra-Firm Efficiency Measures. *Journal of Productivity Analysis* 15 (3): 185-199.
- Paris, Q., and R. E. Howitt (1998) An analysis of ill-posed production problems using Maximum Entropy. *American Journal of Agricultural Economics* 80 (1): 124-138.
- Park, B. U., S. O. Jeong, and L. Simar (2010) Asymptotic Distribution of Conical-Hull Estimators of Directional Edges. *Annals of Statistics* 38 (3): 1320-1340.
- Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6 (2): 461-464.
- Shah, T., O. P. Singh, and A. Mukherji (2006) Some aspects of South Asia's groundwater irrigation economy: analyses from a survey in India, Pakistan, Nepal Terai and Bangladesh. *Hydrogeology* 14 (3): 286-309.
- Simar, L., and P. W. Wilson (2007) Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* 136 (1): 31-64.
- Simar, L., and P. W. Wilson (2000a) A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics* 27 (6): 779-802.
- Simar, L., and P. W. Wilson (2000b) Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis* 13 (1): 49-78.
- Stoker, T. M. (1993) Empirical Approaches to the Problem of Aggregation Over Individuals. *Journal of Economic Literature* 31 (4): 1827-1874.
- Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (2): 111-147.
- Taylor, J. E., G. A. Dyer, and A. Yunez-Naude (2005) Disaggregated Rural Economywide Models for Policy Analysis. *World Development* 33 (10): 1671-1688.
- Theil, H. (1966) *Applied Economic Forecasting*. Amsterdam: North-Holland.
- Wahba, G. (1990) *Spline Models for Observational Data*. Vol. 59. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics.
- Wilcoxon, F., 1, (1945) Individual comparisons by ranking methods. *Biometrics* 1 (6): 80-83.
- Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.
- World Bank. Living Standards Measurement Survey Dataset Finder, <http://iresearch.worldbank.org/lsm/lsmssurveyFinder.htm>, accessed: May 10, 2008.

Wu, J., and R. Adams (2002) Micro versus Macro Acreage Response Models: Does Site-Specific Information Matter? *Journal of Agricultural and Resource Economics* 27 (1): 40-60.

Table 1. Number of Households in Different Policy Models

	(1) Total number of households	(2) Number of households that grew wheat	(3) Number of households that grew maize	(4) Number of households that grew cotton
Province-level model				
1	88	63	86	18
County-level models				
2 Xian County	30	26	28	8
3 Tang County	29	19	29	1
4 Ci County	29	18	29	9
Community-level models^a				
5 Mean	3.6	2.6	3.6	0.75
6 Range	[2, 4]	[0, 4]	[2, 4]	[0, 4]

a. Since there are more than 20 communities, we have estimated more than 20 set of community-level models. For the sake of brevity, we only report the mean and range of the number of households.

Data source: Authors' survey in 2004 (CWIM data).

Table 2. The County-level Models: Estimation Results of the Wheat Production Frontier in Xian County

$$f() = \sum_i \alpha_i x_i - \sum_i \sum_{i'} x_i z_{ii'} x_{i'}$$

Input	α_i	$z_{ii'}$				
		Land	Water	Labor	Fertilizer	Material
Land	34.7379 [27.2592, 42.2909]	66.3825 [42.1052, 87.2175]				
Water	1.7589 [1.5160, 2.0236]	-0.2111 [-0.3880, -0.1015]	0.0012 [0.0009, 0.0016]			
Labor	0.9125 [0.3986, 1.4223]	-0.1297 [-0.4345, 0.1522]	0.0005 [-0.0004, 0.0017]	0.0039 [0.0011, 0.0062]		
Fertilizer	1.3989 [0.8229, 1.8740]	-0.1046 [-0.4076, 0.1525]	-0.0004 [-0.0013, 0.0004]	-0.0001 [-0.0018, 0.0016]	0.0028 [0.0015, 0.0042]	
Material	1.3232 [0.6458, 1.9960]	-0.2267 [-0.6085, 0.1129]	-0.0000712 [-0.0011, 0.0014]	-0.0027 [-0.0056, -0.0000]	-0.0012 [-0.0032, 0.0006]	0.0101 [0.0093, 0.0124]

Notes:

a. For the sake of brevity, only estimation results of the county -level model for wheat production in Xian County are reported.

b. When performing the GME estimation, we change the unit of land to be squared meter so that the magnitude of land is in range with that of other inputs. Because of the rescaling, coefficients on land are large in magnitudes.

c. The 95% confidence intervals, which are constructed from the empirical bootstrap distributions, are reported in brackets.

Table 3. Comparison of the Predictive Performance of Different Policy Models

		Percentage Absolute Prediction Error (PAPE)		
		(1) Province-level Model	(2) County-level Models	(3) Community-level Models ^a
1	Wheat	20% **	10.5% ***	27%
2	Maize	24% **	13% ***	26%
3	Cotton	30% *	16% *	51%

		Theil's U-Statistic (U)		
		Province-level Model	County-level Models	Community-level Models
5	Wheat	0.9524	0.6512	1.2504
6	Maize	0.7379	0.4681	0.8293
7	Cotton	1.2889	0.7866	1.7942

a. Since there are more than 20 communities, we have estimated more than 20 set of community-level models. For each set of community-level model, we have calculated the value of its PAPE. For the sake of brevity, we only report the means of the PAPE.

b. The asterisks indicate the significance level of the test statistics of the asymptotic loss differential test (when comparing the province-level model or the county-level models with the community-level models. Asterisk (*), double asterisk (**) and triple asterisk (***) denote variables significant at 10%, 5% and 1% respectively.

Data source: Authors' survey in 2004 (CWIM data).

Appendix 1. AIC, BIC and MSPE when the sample size is large.

Since all the policy models are linear, we can express them in general as: $\mathbf{Y} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Note the index for crop and the technical inefficiency parameters are suppressed here for simplicity. In the general model, $\mathbf{D}$ is the data matrix used to estimate production frontier for crop j . $\mathbf{D}$ is an $N \times K$ matrix, where N is the size of sample used and K is the number of parameters to be estimated.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are often expressed in logarithms. We express them in exponential terms so that they are comparable to squared errors: $AIC = 1/N \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 e^{2K/N}$ and $BIC = 1/N \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 N^{K/N}$.

When N is much larger than K , K/N is close to zero. This means $\lim_{N \rightarrow \infty} e^{2K/N} = 1$ and $\lim_{N \rightarrow \infty} N^{K/N} = 1$, then

$$AIC \rightarrow 1/N \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 \quad \text{and} \quad BIC \rightarrow 1/N \sum_{n=1}^N (Y_n - \hat{Y}_n)^2. \quad (\text{A1})$$

Mean Squared Prediction Errors (MSPE) is defined as $1/N \sum_{n=1}^N (Y_n - \hat{Y}_n^{(-n)})^2$. To show that when the sample size is large, the objective function of MSPE is equivalent with that of AIC and BIC, we use three facts:

Fact 1. $Y_n - \hat{Y}_n^{(-n)} = (Y_n - \hat{Y}_n)/(1 - h_{nn})$, where h_{nn} is the n th diagonal element of the projection matrix $\mathbf{H} = \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ and $\hat{Y}_n$ is the in-sample predicted value (or fitted value when all data points including the n th observation are used in estimation). The proof of Fact 1 can be found in Wahba (1990).

Fact 2. $\sum_{n=1}^N h_{nn} = K$. Fact 2 can be shown by the following: $\text{tr}(\mathbf{H}) = \text{tr}(\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}') = \text{tr}((\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{D}) = \text{tr}(\mathbf{I}_{K \times K}) = K$, that is, $\sum_{n=1}^N h_{nn} = K$.

Fact 3. The value of h_{nn} is between 0 and 1. Fact 3 is derived from the observation that $\mathbf{H}$ is symmetric and idempotent, and so $h_{nn} = \sum_{m=1}^N h_{nm}h_{mn} = \sum_{m=1}^N h_{nm}^2 = h_{nn}^2 + \sum_{m \neq n}^N h_{nm}^2$. That is, $h_{nn} - h_{nn}^2 = \sum_{m \neq n}^N h_{nm}^2 \geq 0$, which only holds if and only if h_{nn} is between 0 and 1.

Using Fact 1 to 3 and assuming that $\mathbf{D}'\mathbf{D} = O(N)$ and $\lim_{N \rightarrow \infty} h_{nn} = 0$, it follows that:

$$\text{MSPE} = 1/N \sum_{n=1}^N \left[(Y_n - \hat{Y}_n)/(1 - h_{nn}) \right]^2 \xrightarrow{\text{when } N \text{ is large}} 1/N \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 \quad (\text{A2})$$

Combining (A1) and (A2), it is clear that the objective functions of AIC, BIC and MSPE are essentially equivalent when the sample size is much larger than the number of parameters to be estimated.

Appendix 2. Information content of the GME estimates

The information content of estimates is often used to examine the performance of the GME estimation. We use a measure called the normalized entropy, $S(\hat{p})$ (Golan, et al. 1996). The measure, $1-S(\hat{p})$, can be seen as a counterpart of the goodness of fit measure, R^2 , in the estimation of OLS models.

The normalized entropy is defined as the proportion of the unexplained variation in the probability distribution of coefficients after the model is estimated using the method of GME. Mathematically, it is:

$$S(\hat{p}) = \left[-\sum_i \sum_m (\hat{p}_{\alpha_i}^m \ln \hat{p}_{\alpha_i}^m) - \sum_i \sum_{i'} \sum_m (\hat{p}_{z_{ii'}}^m \ln \hat{p}_{z_{ii'}}^m) \right] / K \ln(M) \quad m=1, \dots, M \quad (2A)$$

In equation (2A), the total number of coefficients that are to be estimated for the production frontier is denoted by K . Before adding any constraints in the GME estimation, for each coefficient, a uniform distribution would give the maximum value of the entropy. That is, p_m takes the value of $1/M$ and the entropy of its probability distribution, $-\sum_m p_m \ln p_m$, is $\ln(M)$.

The maximum value of the joint entropy of all coefficients (i.e., the α_i s and z_{ii} s) is, then, $K \ln(M)$. Since entropy measures the uncertainty about the probability distribution of the coefficients, $K \ln(M)$ measures the total uncertainty without adding any constraints in the GME estimation. This is the information content before estimation.

After adding the data consistent and theoretical constraints in GME estimation (the process of adding information), the maximized value of the joint entropy is $-\sum_i \sum_m (\hat{p}_{\alpha_i}^m \ln \hat{p}_{\alpha_i}^m) - \sum_i \sum_{i'} \sum_m (\hat{p}_{z_{ii'}}^m \ln \hat{p}_{z_{ii'}}^m)$. Therefore, the value of $S(\hat{p})$ is the proportion of the unexplained part of the variation after the model is estimated. Since $K \ln(M)$ is the maximum value that the joint entropy can take and the entropy is always positive (p is positive and $\ln p$ is always negative), the value of $S(\hat{p})$ lies between 0 and 1.

According to the measure, the GME estimation can be shown to increase the information content in estimates significantly. The value of $S(\hat{p})$ is 0.83 in the province level model and ranges from 0.7 to 0.85 in the county-level models. On average, the GME estimation reduces the uncertainty of the parameters by about 20 percent. The values of $S(\hat{p})$ are well within the ranges of results in other studies that also use the normalized entropy to measure the performance of the GME estimation (e.g., Fraser 2000; Golan, Moretti and Perloff 1999).

Appendix 3. The County-level models: Estimation Results of Technical Efficiency
— Wheat Production

N	θ_n	n	θ_n	n	θ_n
Xian	County	Tang	County	Ci	County
1	0.9998	1	1	1	1
2	0.9999	2	1	2	0.7905
3	0.9993	3	0.9998	3	0.9993
4	1	4	0.9994	4	0.9979
5	0.9988	5	1	5	0.7714
6	0.7927	6	0.9435	6	1
7	0.9998	7	0.57	7	1
8	1	8	0.7969	8	1
9	0.7053	9	0.9995	9	0.9997
10	0.9995	10	0.9996	10	0.9985
11	1	11	0.9999	11	0.9993
12	0.9996	12	0.9995	12	0.9996
13	0.7664	13	0.9994	13	0.9999
14	0.8513	14	0.9999	14	0.9991
15	0.9374	15	0.9525	15	1
16	0.9999	16	0.9999	16	0.861
17	1	17	0.9998	17	0.9997
18	1	18	0.9998	18	0.8325
19	0.7606	19	0.9999		
20	1				
21	0.8418				
22	0.9993				
23	0.8227				
24	0.9733				
25	0.9785				
26	1				

Note: n is household index.

Appendix 4. Entropy ratio test

Since one of the main purposes of choosing the best policy model is to identify the empirical model that allows analysts to capture the differences in behavior across sample units, before we proceed to select the best model, a more fundamental question is whether there is heterogeneity across places. Following Golan et al. (2001), we use an entropy-ratio test to examine whether households in different counties are using the same production technology. An entropy-ratio statistic can be constructed by taking the difference between the maximum values of the objective functions for the restricted and unrestricted models (that is, the difference between the values of maximized entropies). Two times this difference has a limiting χ^2 distribution with the degrees of freedom equal to the number of restrictions imposed. This test is similar to the traditional log-likelihood test and the empirical likelihood-ratio test.

In our case the null hypothesis is that households in different counties face the same frontier (that is, the province-level model). Since we have five inputs and 60 parameters for each county (20 parameters for each crop), the number of constraints is 120. The entropy-based tests reject the null hypothesis at the one percent significance level for all three crops: the entropy-ratio statistic is 288, which is larger than the corresponding critical value (the χ^2 critical value is only 163 for a p value of 0.005). The test results show that there is a significant amount of heterogeneity across places.