Journal of Econometrics 168 (2012) 141–155

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Estimation of market power in the presence of firm level inefficiencies☆

Levent Kutlu[a,*], Robin C. Sickles[b]

[a] School of Economics, Georgia Institute of Technology, United States

[b] Department of Economics, Rice University, United States

ARTICLE INFO

Article history:
Available online 6 November 2011

JEL classification:
C23
C73
L1

Keywords:
Market power
Panel data
Dynamic games
Kalman filter
Airline competition
Suboptimal allocations

ABSTRACT

"The quiet life hypothesis" (QLH) by Hicks (1935) argues that, due to management's subjective cost of reaching optimal profits, firms use their market power to allow inefficient allocation of resources. Increasing competitive pressure is therefore likely to force management to work harder to reach optimal profits. Another hypothesis, which also relates market power to efficiency is "the efficient structure hypothesis" (ESH) by Demsetz (1973). ESH argues that firms with superior efficiencies or technologies have lower costs and therefore higher profits. These firms are assumed to gain larger market shares which lead to higher concentration. Ignoring the efficiency levels of the firms in a market power model might cause both estimation and interpretation problems. Unfortunately, the literature on market power measurement largely ignores this relationship. In the context of a dynamic setting, we estimate the market power of US airlines in two city-pairs by both allowing inefficiencies of the firms and not allowing inefficiencies of the firms. Using industry level cost data, we estimate the cost function parameters and time-varying efficiencies. An instrumental variables version of the square root Kalman filter is used to estimate time-varying conduct parameters.

1. Introduction

"The quiet life hypothesis" (QLH) by Hicks (1935) argues that, due to management's subjective cost of reaching optimal profits, firms use their market power to allow inefficient allocation of resources. Increasing competitive pressure is likely to force management to work harder to reach optimal profits. Another hypothesis that relates market power and efficiency is "the efficient structure hypothesis" (ESH) by Demsetz (1973). ESH argues that firms with superior efficiencies or technologies have lower costs and therefore higher profits. These firms are assumed to gain larger market shares which lead to higher concentration. However, ignoring the efficiency levels of the firms in a market power model may cause both estimation and interpretation problems. We briefly outline these estimation and interpretation problems below.

The first estimation problem involves the effect of measurement error in marginal cost calculations on outcomes from a dynamic game when there is inefficiency that is ignored. For market power measurement in dynamic environments Pindyck (1985) proposes using full marginal cost (FMC) instead of marginal cost (MC). FMC is MC plus the shadow cost of participating in a coalition.[1] This necessitates estimation of the shadow cost. When the reason for dynamics is strategic (i.e., firms are playing a repeated game), shadow costs are affected by the degree of symmetry of firms (increase in and similarity of efficiency levels) since one generally observes higher cooperation among firms as the degree of symmetry increases. If the efficiency levels of firms are observed, then inefficient firms will find it less beneficial to deviate relative to the state in which they are more efficient. For a variety of interesting cases, the higher today's efficiency level the weaker the punishment relative to the gain from deviation. If the efficiency levels are not observable, then firms cannot know if observed outcomes are due to a deviation or to poor management. Hence, during the time periods when we have high levels of inefficiency, firms

☆ This paper is a substantially revised and extended version of "Market Power and Efficiency: A Dynamic Approach" by L. Kutlu. The authors would like to thank the participants of the Texas Econometrics Camp XIV (2009, Rice University Student Workshop (2009), Southern Economic Association Conference (2009), Rice University Econometrics Seminar (2010), Monash University Economics Seminar (2010), UNSW Economics Seminar (2010), North American Productivity Workshop (2010), Asian Pacific Productivity Conference (2010), Georgia Institute of Technology School of Economics Seminar (2010), University of Massachusetts Amherst Dept. of Resource Economics Seminar (2011), University of Cambridge (2011) and University of Leicester (2011), for their comments and criticisms. The authors also thank Prof. Chang-Jin Kim for providing us the draft version of his work. The usual caveat applies.

* Corresponding author.
E-mail address: levent.kutlu@econ.gatech.edu (L. Kutlu).

[1] The reason for the dynamic nature of the optimization problem might be capacity constraints as well. In that case, shadow cost of capacity should be included in the calculation of full marginal cost.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.11.001

may undercut each other to prevent potential deviations. The potential dependence of the shadow cost on the efficiency levels of the firms puts in question the consistency of parameter estimates of the shadow cost function which in turn may contaminate other model parameters. Thus inefficiency levels would appear to play a more important role in dynamic competition models than their static counterparts.

A second key estimation problem may be caused by the introduction of measurement error in the estimation of markups if asymmetry in costs is ignored and thus inefficiency is not explicitly modeled. Ignoring firm inefficiencies may thus also invalidate standard dead-weight-loss (DWL) calculations. Even if parameters are consistently estimated, the standard interpretation of estimated market power would not be valid if the firms actually are inefficient. The problem of interpretation is due to the fact that the calculation of the deadweight loss (DWL) from collusive behavior depends on whether one attributes firm heterogeneities to factors other than inefficiency. If the heterogeneity among the firms is due to factors other than their efficiency levels, then the traditional DWL calculation methods would be valid. On the other hand, if firms exhibit unobservable inefficiency that is misinterpreted as firm heterogeneity in the neoclassical framework of efficient market behavior, then standard calculations of DWL used in evaluating mergers or antitrust actions may not be valid. In such cases we recommend using the efficient full marginal cost (EFMC) for the markup calculation. EFMC is defined as the sum of the shadow cost of participating in a coalition and efficient marginal cost calculated from stochastic frontier analysis (SFA) techniques.

A third potential estimation problem is due to the possibility that the DWL follows a non-monotonic path as a function of traditional market power measures, e.g., the conduct (parameter) of firms. Under the QLH a shift from monopoly to competition not only reduces the monopoly rents but also increases the efficiencies of firms which in turn leads to cost reductions. Thus, the negative effect of market power on social welfare is two-fold. Standard DWL calculations consider only the former effect. In the case when the most efficient firms follow the ESH it is possible to observe a non-monotonic DWL path as the market power as traditional market power measures such as the conduct (parameter) of firms increases. We propose below a market power measure that partially addresses this monotonicity issue. That is, under the ESH the welfare loss due to monopoly rent might be dominated by the gain in welfare due to higher efficiency levels. Unfortunately, the literature on market power measurement largely ignores these issues; and when the issues are considered they are done so in a static setting. [2] In contrast to these studies, we measure the market power in a dynamic setting.

Because the effects of ignored inefficiency are so pronounced in the dynamic framework it may be instructive to briefly focus on this particular issue before presenting our general model. As we have mentioned above and show below as we develop our analytical arguments, the essential difference between the static and dynamic settings is that in the dynamic setting conduct is determined by the efficient full marginal cost (EFMC) rather than the efficient marginal cost (EMC). Ignoring the shadow cost of being in a coalition introduces an omitted variable bias into the static model. Ignoring shadow costs tends to induce a positive bias in the conduct parameter and this in turn induces a positive bias in the estimate of dead-weight-loss (DWL). However, ignoring firm inefficiencies induces a negative bias in the DWL estimates. Hence, the overall bias in the DWL estimates is ambiguous. In practice, underreaction to market power would mean, among other things, that mergers would be approved which should be prohibited from the social welfare point of view, whereas the overreaction to market power would mean that mergers would not be approved although the efficiency gain from the mergers dominates the negative effects of the mergers. To the best of our knowledge, we are the first to consider the efficiencies of the firms when measuring the market power in a dynamic setting.

In our empirical model we estimate the market power of US airlines in two city-pairs by both allowing and not allowing the inefficiencies of the firms. Using industry level cost data, we estimate the cost function parameters and time-varying efficiencies by the fixed effects model proposed by using the model of Cornwell et al. (1990) (CSS). In order to estimate the conduct parameters, we extend a particular Kalman filter procedure dealing with the endogeneity problem proposed by Kim and Kim (2007) to the multivariate case. We also examine the implications of ignoring inefficiencies of firms for the DWL calculations. In order to calculate the DWL for the inefficient firms we calculate the efficient MC levels based on our efficiency estimates. Our results indicate that even in the static case, and using parameter estimates that are consistent under standard neoclassical assumptions of efficiency, traditional DWL calculations maybe very inaccurate. To be more specific, when the efficiency levels of firms are low the size of the DWL can vary substantially depending on whether we assume firms are efficient or not. Moreover, the paths that DWL follows as a function of firm conduct when we ignore inefficiencies and when we allow for inefficiency also differ substantially. We propose a new measure of market power in order to capture this potential non-monotonic behavior of DWL as a function of the conduct parameter.

In the next section we provide a brief discussion of modeling approaches that have been taken in the literature to estimate market power and to measure efficiency. Section 3 provides the details for the technical aspects of our dynamic modeling approach and develops the Kalman filter estimator we implement in Section 4. In Section 3 we provide the empirical model, discuss our data, detail our estimation methodology, and explain our results. Section 4 concludes.

2. Measuring market power and efficiency

Market power is defined as the ability of a firm (or a group of firms) to raise the price of a good or a service above the competitive level. A widely used measure of market power is the Lerner index, proposed by Lerner (1934):

$$L \equiv \frac{P - MC}{P}. \tag{1}$$

This index measures how much market power a firm exercises as opposed to measuring how much market power it has. Encaoua and Jacquemin (1980) derive a link between the Lerner index and the Herfindahl index which is a measure of market concentration. The Lerner index is independent of units of price and marginal cost. Usually it is presumed to be between zero and one.

The Lerner index assumes static profit maximization so that the firm produces MC equal to marginal revenue (MR). In dynamic markets, price and production are determined intertemporally. There are at least two reasons leading to a dynamic market setting: strategic and fundamental.[3] If the firm believes that its

[2] Berg and Kim (1998), Maudos and Fernández de Guevara (2007), Delis and Tsionas (2009), Koetter et al. (2008), and Koetter and Poghosyan (2009) exemplify some papers that consider the efficiencies of firms when measuring the market power. None of these papers concentrate on a solution for the potential non-monotonicity of the DWL as a function of conduct.

[3] See Perloff et al. (2007) for a book-length treatment of this subject.

rivals will respond to its current actions in the future, the reason for the dynamic market setting is referred to as strategic. If the current action affects stock variables that affect future profits, then the reason for the dynamic market setting is referred to as fundamental. A stock variable, for example, might be some amount of goodwill or knowledge, or level of a quasi-fixed output. In a dynamic setting, a risk-neutral firm maximizes discounted expected profits. Indeed, even for the static setting with price greater than MC, one can construct examples so that the Lerner index might not be a reliable measure of market power. Consider a monopolist who produces an exhaustible resource facing an isoelastic demand curve and has zero extraction cost. Although the Lerner index is one, the producer has no market power. In such cases Pindyck (1985) proposes using full marginal cost, which is MC plus user cost, rather than MC to measure the market power.

Another approach for measuring market power is to estimate a conduct parameter rather than the Lerner index. This approach uses a conjectural variations approach and treats the conduct as a parameter to be estimated. One infers the conduct through the responsiveness of price to changes in demand elasticities. For a static setting the conduct is deduced from a generalization of the monopolist's first order condition:

$$P + \theta QP'(Q) = MC \tag{2}$$

where P is the price, Q is the industry output, and θ is the industry conduct parameter. The conduct parameter is equal to one for perfect collusion (or monopoly); is equal to zero for perfect competition; and is equal to the inverse of the number of firms for symmetric Cournot competition. In the conjectural variations approach the conduct parameter derived from the above equation is the demand elasticity adjusted Lerner index. In the case of a high margin, markets with inelastic demand and less competitive markets are distinguished by this demand elasticity adjustment.

One of the problems with the conduct parameter approach, like the Lerner index approach, is that it is static and hence not valid for dynamic oligopoly games.[4] Corts (1999) argues that if the optimization problem of the firms is a dynamic one, then the success of the conduct parameter approach depends on the discount factor and the persistency of the demand. As the discount factor increases and the demand becomes more persistent, the conduct parameter approach becomes more accurate. The conduct parameter approach cannot detect any market power if the discount factor is low and the demand has substantial shocks. A dynamic version of the conduct parameters method would appear to be necessary for correct inference about market power. Sickles et al. (2007) considered a dynamic market power model with conduct fixed over the sample period. However, considering changing market conditions, the assumption of a constant conduct parameter may not be realistic. Röller and Sickles (2000) allowed the conduct parameter to vary at different stages of a two-stage capacity and price game, while time-varying specifications of conduct have been considered in different contexts by Bresnahan (1989), Brander and Zhang (1993), Gallet and Schroeter (1995), Captain and Sickles (1997), and Kim (2005).

In our paper we assume that firms are playing a dynamic game but we allow the conduct parameter to be time-varying. The strategies of the firms determine the actions, where firms take as a function of state variables known to the firms but only partially observed by the econometrician, in which case we have both observed and unobserved factors in our empirical model that can affect conduct. A common way to model the time-varying conduct parameters is to use some explanatory variables as proxies for conduct. These models are estimated via either three stage least squares (3SLS) or generalized method of moments (GMM).[5] While these studies allow for time-varying conduct, they do not allow for a time-varying relationship between conduct and the explanatory variables which proxy conduct. Moreover, the parameters specified in the structural model are assumed to be constant over time. Ignoring time-varying parameters in the structural model would typically mean ignoring them in estimating the reduced form predictors for the right-hand-side endogenous variables and thus may also lead to the problem of weak instruments. While the firms have a "core conduct" which is constant over time, it is assumed in our study that due in part to unobserved factors the conduct parameter is changing over time. We allow the unobserved factors that affect market power to follow a stationary autoregressive process in order to allow, for example, for the effect of an oil price shock on market power to have some persistence in our empirical model. The Kalman filter method we describe in the next section can deal with these problems.[6]

Productive efficiency is a measure of performance of firms and is an important factor to consider when analyzing the effects of deregulation, mergers, and market structure. Technical efficiency can be used to rank firms according to their performances and improve managerial oversight by identifying "best practices" and "worst practices". Approaches to measure technical efficiency, based on the modification in the error structure of the linear regression model and referred to as the stochastic frontier model, were introduced by Aigner et al. (1977), Battese and Cora (1977), and Meeusen and van den Broeck (1977) for cross-sectional models. Jondrow et al. (1982) provided a way to estimate firm specific technical efficiency in a cross section. Panel data potentially provide more reliable information about the efficiencies of firms and are essential for measuring dynamic firm decision-making. Fully parametric maximum likelihood approaches for estimating a random effects panel stochastic frontier were introduced by Pitt and Lee (1981). Schmidt and Sickles (1984) introduced non-parametric regression-based methods to estimate fixed effects stochastic frontier models and to test for the orthogonality of inefficiency effects and input levels, an assumption of the random effects parametric model. The assumption of time invariance was lifted and estimators for panel stochastic frontiers with time-varying efficiencies were introduced by Cornwell et al. (1990), Kumbhakar (1990), and Battese and Coelli (1992).[7]

In the presence of firm level inefficiencies, the market power estimates should be corrected by using efficiency adjusted marginal costs. The SFA literature provides a wide range of methods to estimate the cost frontiers from which we can easily calculate the frontier marginal cost (full efficiency marginal cost). By utilizing SFA efficiency estimates Koetter et al. (2008) calculate the static version of the efficiency adjusted Lerner index.[8] As we already mentioned, the static version of the Lerner index has a variety of issues such as an omitted variable bias and usage of the incorrect version of the marginal cost. Hence, for the dynamic framework the EFMC seems to be a better marginal cost concept

[4] See, Genesove and Mullin (1998) for some evaluations regarding the success of static oligopoly models in characterizing conduct.

[5] For example, Gallet and Schroeter (1995) use 3SLS and Kim (2005) uses GMM.

[6] The Kalman filter has other advantages, such as handling missing observations in a direct and relatively transparent fashion and allowing one to explicitly model non-stationary stochastic processes, which we do not exploit in our empirical work.

[7] See, Kumbhakar and Lovell (2000) for an extensive survey on stochastic frontier analysis. Also, see Sickles (2005) for comparisons of many efficiency estimators including recent ones.

[8] See also Koetter and Vins (2008) and Koetter and Poghosyan (2009).

for the Lerner index or conduct parameter estimation. However, after such a correction a monotone DWL path as a function of the market power index cannot be guaranteed if the firms follow ESH for some market power level. We examine this issue below.

3. The empirical model and estimations

In this section we outline our empirical model, discuss our data, detail our estimation methodology, and explain our results. We examine the market power of US airlines in two city-pairs [Chicago–San Diego (SAN) and Chicago–Salt Lake City (SLC)]. Using industry level cost data, we also estimate city-pair cost function and marginal cost parameters and time-varying efficiencies using the fixed effects model of CSS. We consider two basic models, one with and one without technical efficiency. Time-varying conduct parameters are estimated using an extended Kalman filter procedure.

3.1. The dynamic competition model

The model we describe in this section is reminiscent of Puller's (2007, 2009) models. Puller (2009) introduces dynamics in order to provide more realism in modeling dynamic decision making in the conduct parameter framework. While the former model differs from the traditional static models due to capacity constraints, the reason for the difference in the latter model is strategic (repeated game). None of these studies allow for inefficiency (thus the asymmetry in costs is not due to inefficiency). In contrast to Puller we allow for costs to differ because of idiosyncratic inefficiencies. We refer to the setting in which firms share the same cost function (and full efficiency) as the symmetric case and its alternative as the asymmetric case. Moreover, in contrast to Puller (2007, 2009) we estimate industry conduct rather than firm level conduct. In our model firms choose output and play an efficient supergame where no structural assumptions are made about the form of the punishment rules for deviations from the coalition strategies. However, these deviations will be punished and this will lead to lower profits, such as those consistent with a Cournot equilibrium. As in Rotemberg and Saloner (1986) and Puller (2009) we assume a full-information environment. More precisely, at the beginning of each period firms know the demand and cost shocks before they make their decisions. Then, the firms simultaneously make their strategic decisions which become common knowledge. The observability of the shocks allows the oligopoly members to adjust their quantity choices and thus dampen profits strategically. If the demand and cost shocks are such that the incentive to deviate is high, firms adjust strategies such that they have lower profits relative to the case in which the incentive to deviate is not high. This is done to prevent deviation. By doing so firms may prevent deviations at times when it would seem most likely to deviate. This looks like a price war and is the general idea of Rotemberg and Saloner (1986) who also assume that shocks are observable. The shocks are observed by the econometrician as well. That is, the econometrician has the data for variables that proxy the demand and cost shocks. We assume that the good is homogenous. Hence, the price is uniform at a given time period.

We specify the linear inverse demand function as:

$$P_t = \beta_0 + \beta_1 Q_t + \beta_2 PCI_t + \sum_{k=1}^{3} \delta_{i,k} Qtr_{kt} + \varepsilon_{it} \tag{3}$$

where P is the price, Q is the quantity, PCI is the per capita income, and Qtr are the seasonal dummies.

Firm i's profit function is given by:

$$\pi_{it} = P_t(Q_t) q_{it} - C(q_{it}; ie_{it}) \tag{4}$$

where ie_i is the inefficiency level and C is the cost function. We assume that firms share the same cost frontier but they have different efficiency levels. Hence, the realized cost is affected by the efficiency of the firms. This introduces asymmetry to our model. Cost depends on the city-pair market but, for the sake of notational simplicity, we suppressed all city-pair subscripts.

The cost function is given by:

$$\ln C(q_{it}; ie_{it}) = \ln \tilde{C}(q_{it}) + v_{it} + u_{it} \tag{5}$$

where $u_{it} \geq 0$ and $v_{it} \sim N(0, \sigma_v^2)$ are mutually independent random variables and $ie_{it} = 1 - \exp(-u_{it})$ is the inefficiency of firm i at time t.

From the above equation the MCs are calculated as follows:

$$MC(q_{it}; ie_{it}) = \frac{\tilde{C}(q_{it})}{1 - ie_{it}} \frac{\partial \ln \tilde{C}(q_{it})}{\partial q_{it}}. \tag{6}$$

Hence, for a fully efficient firm we have:

$$MC(q_{it}, 0) = \tilde{C}(q_{it}) \frac{\partial \ln \tilde{C}(q_{it})}{\partial q_{it}}. \tag{7}$$

The calculation of marginal cost is based on similar ideas used to calculate conduct in the market power literature. In the market power literature one estimates perceived marginal revenue, which depends on conduct. In the case of full market power perceived marginal revenue is the monopoly outcome. Similarly, in our calculation of marginal cost we assume that each firm uses its perceived cost function which is $C_i = \tilde{C}_i \exp(u_i)$, where the inefficiency level $\exp(u_i)$ takes on the role of conduct. We can calculate the corresponding perceived marginal cost by differentiating the perceived cost function. Like the cost curves of firms with different efficiency levels, the MC curves are also parallel shifts of each other. Moreover, the 'distance' between the MC curves (of different firms) are the same as the 'distance' between the corresponding cost curves. The direct implication of this is that we can measure the efficiencies of the firms by utilizing their MC's. Thus, all the following formulas are for both the inefficiency and the full efficiency cases. The optimization problem of the firms is given by:

$$Q_t^*(S_t, \beta) = \arg\max_{Q_t, s_t} \sum_i \pi_{it}(s_{it} Q_t; S_t) \quad \text{st} \tag{8}$$

$$\pi_{it}^b(Q_t; S_t) + \sum_{k=1}^{\infty} \beta^k E_t \left[\pi_{it}^r(S_{t+k})\right] \leq \pi_{it}(s_{it} Q_t; S_t) + \sum_{k=1}^{\infty} \beta^k E_t \left[\pi_{it}^*(S_{t+k})\right] \quad \forall i$$

where s is the market share, Q is the total quantity, π^b is the best response profit, π^r is the profit for the retaliation period, π^* is the profit when collusion is sustained, $S_t = \begin{bmatrix} cs_t & ds_t \end{bmatrix}'$ is the state of the world at time t, and β is the discount factor. The components of the state are as follows: cs is a variable representing the cost shock and ds is a variable representing the demand shock.

The first-order condition for the output is:

$$\sum_i \left[P'(Q_t^*) Q_t^* + P(Q_t^*) - MC_{it}(s_{it} Q_t^*; ie_t)\right] s_{it} - \mu_t^* = 0 \tag{9}$$

$$P'(Q_t^*) Q_t^* + \sum_i MK_{it}(s_{it} Q_t^*; ie_t) s_{it} - \mu_t^* = 0 \tag{10}$$

$$\theta_t P'(Q_t^*) Q_t^* + MK_t - \mu_t^* = 0 \tag{11}$$

$$\theta_t \beta_1 Q_t^* + MK_t - \mu_t^* = 0 \tag{12}$$

where $MK_{it} \equiv P_{it} - MC_{it}$ is the markup for firm i, $MK_t \equiv \sum_i MK_{it} s_{it}$ is the market share weighted markup, and μ_t^* is the dynamic factor which reflects the incentive compatibility constraint.

Appelbaum (1982) defines the industry Lerner index as the market share-weighted Lerner index. He defines the degree of the market power of an industry as the industry Lerner index. Similar to his index, our model involves the market share-weighted markup as the industry markup. We call this markup the industry markup. A traditional model for the industry conduct is:

$$MK = -\theta QP'(Q). \tag{13}$$

Our model resembles the traditional model. The only difference is the μ_t^* term that is introduced due to the dynamic nature of the problem. If $\mu_t^* = 0$ for each t, we can conclude that firms are playing a static game. If $\mu_t^* \neq 0$, then firms are playing a repeated game and not including μ_t^* causes an omitted variable bias. We define θ_t as the industry conduct. If $\theta_t = 0$ and $\mu_t^* = 0$, then we can deduce that the industry conduct is consistent with perfect competition and if $\theta_t = 1$, it is consistent with efficient collusion. After estimating our model one can calculate the dynamic version of the Lerner index as follows:

$$L \equiv \frac{P - MC - \mu^*}{P} = -\theta \frac{\partial P}{\partial Q}\frac{Q}{P} = -\frac{\theta}{E_d} \tag{14}$$

where E_d denotes the price elasticity of demand.

We examine the consequences of not considering the efficiencies of firms on market power analysis for the dynamic game environment. Unfortunately, if the firms are inefficient, then this might lead to substantially inaccurate conclusions about the market power. If μ_t^* is a function of the efficiency levels, we can conclude that the classical repeated game models give invalid inferences by not taking into account the effect of the efficiency on the market power. Even if μ_t^* is not a function of efficiency levels, the optimization model may be irrelevant if the firm-specific cost structures that account for inefficiencies are not utilized. As the "degree" of inefficiency increases, the severity of the bias from this kind of misspecification would also increase. Moreover, the Lerner index formula requires full marginal cost, $FMC = MC + \mu^*$, rather than the marginal cost. Hence, even with consistent parameter estimates, the static version of the Lerner index overestimates the market power if $\mu^* > 0$. The dynamic efficiency adjusted Lerner index proxies the DWL due to socially inefficient allocation of resources associated with monopoly and is defined as follows:

$$L^{SFA} \equiv \frac{P - MC^{SFA} - \mu^*}{P} \tag{15}$$

where MC^{SFA} denotes the fully efficient marginal cost calculated from SFA estimates of the cost function.

In the presence of inefficient firms, this DWL triangle is larger than the traditional DWL triangle and the size difference depends on the extent of inefficiency levels. The efficiency correction captures this difference.

Finally, we estimate a counterfactual model for inefficiency. From the stochastic frontier model of CSS, we find the efficient cost frontier and assume that all firms share the corresponding efficient MC function. Then we estimate the market power of the firms assuming full efficiency. This provides us with a vehicle for examining how much market power firms lose by not exploiting their full-efficiencies.

3.2. The data

The data we use for the cost estimations is a quarterly panel data 1980I–1993IV. These data are constructed from the Department of Transportation's (DOT) Form 41/T100 and are discussed in more detail in Wingrove et al. (1997), Alam and Sickles (2000), and Ahn et al. (2000). There are four main inputs: labor, energy, flight capital, and a residual category called materials. Materials include supplies, outside services, and non-flight capital. Quantity and price data are calculated by the multilateral Tornqvist–Theil index number procedure. Flight capital is disaggregated into short haul capital and long haul capital. We include two aircraft attributes to describe flight capital: average size (measured in seats) and fuel efficiency. The data set for the cost includes information for 11 airlines: American Airlines (AA), Continental Airlines (CO), Delta Airlines (DL), Frontier Airlines (FL), Northwest Airlines (NW), Ozark Air Lines (OZ), Piedmont Airlines (PI), Republic Airlines (RC), Trans World Airlines (TW), USAir (US), and United Airlines (UA).

The labor input was composed of 93 separate labor accounts aggregated into five employment classes: flight deck crews, flight attendants, mechanics, passenger/cargo/aircraft handlers, and other personnel. Since we do not have the number of hours worked by each working class, we could not correct for different utilization rates. In 1977, Schedule P10 was changed from quarterly data to annual data. Hence, after 1977 we only know fourth quarter values of employee numbers for specific categories. Missing periods were calculated by interpolation. After the 1987 modification in Form 41, many expense accounts were eliminated. In order to preserve the compatibility, relevant modifications were made to the data. For example, trainees and instructors moved to another personal category. The monthly personal data was converted to quarterly data by averaging the number of full-time employees plus one-half of the part-time employees over the corresponding quarter. After obtaining the relevant head count information for each employment category, the multilateral Tornqvist–Theil index number procedure is used to derive the aggregate labor input.

The energy input is meant to capture aircraft fuel only. Fuel that is used for ground operations and electricity are included in the materials index. The energy input was developed by combining the information on aircraft fuel gallons used with fuel expense data per period. For normalization a multilateral Tornqvist–Theil index number procedure is used to derive the final energy input.

The materials input consist of 69 expenditure accounts aggregated into 12 classes. Since the carrier specific price or quantity deflators were not available, industry-wide price deflators are used. In 1987, Schedules P6 and P7 changed. This led to the elimination of many account categories. The data is adjusted to preserve the consistency.

DOT Form 41, Schedule T2 contains relevant information about the number of aircraft for each different model of aircraft. Data for technological characteristics for aircraft that are in significant use were collected from Jane's All the World's Aircraft (1945 through 1982 editions), henceforth JATWA. The average number of aircraft in service is constructed by dividing the total number of aircraft days for all aircraft types by the number of the days in the quarter.

In order to adjust this measure of capital, average equipment size is used. For each aircraft type, the highest density single-class seating configuration that is listed in JATWA was used. The fleet-wide average is calculated by taking a weighted average of each aircraft type where the weights are the number of aircraft of each type. In some cases the actual number of seats was substantially less than described by this configuration. This is because airlines sometimes reconfigure aircraft for their need of first-class and business-class seats.

We use the average number of months since the Federal Aviation Administration's type certification of aircraft designs as our measure of fleet vintage. It is assumed that technology for an aircraft does not change unless its design is recertified for its type. This only captures significant innovations. Hence, our model does not fully capture the deterioration in capital and increased maintenance costs caused by use.

The output data consist of two components: scheduled output and non-scheduled output. Non-scheduled output includes cargo and charter operations. We used revenue and output data from DOT form 41. From these data seven different outputs

produced by a typical airline are identified. The price of the output is constructed by dividing the revenue generated by the corresponding category by its output quantity. Some carriers offered only one type of service. In such cases, the service was redefined to be coach class. Charter operations for cargo and passenger service outputs were combined into a single category. Since their output units are different, the average passenger is assumed to weigh 200 pounds including baggage. Also, changes in DOT form 41 in 1985 led to the elimination of the distinction between express cargo and air freight. Hence, two categories were combined. All aggregations were done via a multilateral Tornqvist–Theil index number procedure. The prices were normalized to 1.0 in the baseline period.

Two characteristics of airline output are calculated. These include load factor and stage length. Load factor provides a measure of service quality and is a widely used proxy for service competition in most airline transportation studies. This is found by dividing revenue passenger miles by available seat miles. Stage length provides a measure of the length of individual route segments in the carrier's network. Generally, the shorter the flight, the higher the proportion of ground services required per passenger mile. This implies that, in general, shorter flights have a higher cost per mile than longer flights. The average stage length is calculated by dividing the total revenue aircraft miles flown by total revenue aircraft departures.

The costs of airlines differ largely because of economies of density. The cost reduction is attributable to increasing output on an unchanged network. For example, this can be achieved by flying the same number of frequencies with larger aircraft. This is why airlines increasingly try to exploit economies of density by building hub-and-spoke route networks. With the help of this network system, larger aircraft are utilized more than otherwise could have been justified. We use the average size of the fleet to capture the effect of economies of density on cost.

The data we use for the conduct estimations is a quarterly panel data 1980I–1988IV. These data are constructed from the Department of Transportation (DOT) DB1A data set, which includes a one in ten sample of all tickets issued from January 1980 through December 1988, discussed in more detail in Weiher (2002) and Good et al. (2008). Our data is obtained by aggregating this monthly data to quarterly data. Although the original data set reported tickets up to twenty three segments, in our data set we allow only for six segments. This eliminates only a little more than 1% of the data. The data we use for the conduct estimation includes information for 6 airlines: AA, CO, DL, NW, TW, and UA.

3.3. Estimating the marginal cost

As mentioned earlier, we allow inefficiency in our cost function estimation. We also estimate the cost function under the assumption of symmetric firms. We calculate MC from our cost function estimates. Our cost data come from the DOT Form 41 which has detailed accounts on system wide airline expenses and a variety of quantity measures that Good and Sickles have used to construct a quarterly panel of price and quantity data for a panel of airlines. Similar methods were employed by Baltagi et al. (1995) and Caves et al. (1983) in their construction of annual panel data sets. Unfortunately, this data set is for the entire US system and not for specific city-pair routes. We solve this problem by incorporating a specific number of enplanements for each airline, a specific distance of relevant city-pairs as well as airline fixed effects. Here we assume that the firm specific effects are the same over distinct city-pairs. The city-pair specific cost differences are captured by incorporating corresponding distances. The distance between two city-pairs is calculated as an average itinerary distance between these city-pairs. We used a similar methodology employed by Röller and Sickles (2000), and Good et al. (2008). We constructed the marginal costs of a particular route by estimating a panel total cost system using the DOT Form 41 data and deriving the marginal costs of a route segment for a particular carrier, distance traveled, and time period. Terminal costs (enplanement marginal costs) were estimated based on the marginal cost of enplanements and these were combined with the marginal cost of a route segment to construct the full marginal costs of a trip between the city-pairs considered. We assume the following cost structure[9]:

$$\begin{aligned}\ln\frac{C_{it}}{MP_{it}} &= \beta_{it} + \beta_1 \ln\frac{LP_{it}}{MP_{it}} + \beta_2 \ln\frac{KP_{it}}{MP_{it}} + \beta_3 \ln\frac{EP_{it}}{MP_{it}} \\ &\quad + \beta_4 \ln Enp_{it} + \beta_5 \ln Proq_{it} + \beta_6 \ln Nroq_{it} \\ &\quad + \beta_7 \ln SL_{it} + \beta_8 \ln LF_{it} + \beta_9 \ln Aves_{it} \\ &\quad + \beta_{10} \ln Fuel_{it} + \sum_{k=1}^{3} \delta_k Qtr_{kt} + Dummies \end{aligned} \tag{16}$$

where $\beta_{it} = b_{i0} + b_{i1}\frac{t}{t_{\max}} + b_{i2}(\frac{t}{t_{\max}})^2$ is the time-varying effects term that is used to calculate the efficiencies of the firms (for the symmetric case $\beta_{it} = b_0 + b_1\frac{t}{t_{\max}} + b_2(\frac{t}{t_{\max}})^2$), LP is the labor price, KP is the capital price, EP is the energy price, MP is the materials price, Enp is the number of enplanements, Proq is the passenger revenue output quantity, Nroq is the non-scheduled revenue output quantity, SL is the stage length, LF is load factor, Aves is the average size of the airline fleet, Fuel is fuel efficiency, *Qtr* are seasonality dummies, Dummies are firm and period specific dummies (Iran–Iraq war; Gulf war; air traffic strike; AA and CL merger; CO and EA buyout; CO and FR merger; CO and TI merger; RC and HA merger; NW and RC merger; TW and OZ merger; US and PSA merger; DL and WN merger; UA pilot strike; and CO pilot strike).

We want to estimate not only the MCs of the firms but also their time-varying efficiencies. As mentioned earlier, we used the fixed effects model of CSS for this purpose. This estimator approximates the time-varying effects term by a second degree time polynomial. We imposed homogeneity restriction in our estimations. Hence, estimated cost functions are homogenous of degree one in prices. The Cobb–Douglas cost estimates are given in Table 1.

Column # 1 estimates based on the CSS estimator statistically dominate those from the estimates based on the assumption of symmetric costs (column #2). Symmetry is rejected by the Wald test with a $p\text{–}value$ of .000 ($\chi^2(30) = 511$). The use of Column 1 estimates allow us to carry out a counterfactual exercise to assess the impact of moving all firms to the frontier. Based on the estimates from Column # 1 we calculated carrier-specific and time-specific marginal costs for the passenger's flight. The terminal costs are the marginal costs of enplanements while the per mile costs are equal to the marginal costs associated with a revenue passenger mile. Total segment costs equal the marginal costs of enplanements added to the product of the marginal costs of a passenger revenue mile and the number of miles flown on the particular segment. Defining the output margin as the flight segment flown by a passenger then allows us to calculate the marginal costs of a flight segment[10]:

$$MC = MC_{enp} + MC_{proq} \times Miles\ Flown. \tag{17}$$

[9] We also tried full and restricted translog functional forms. In all of our attempts the concavity condition failed at almost all sample points. This result was robust to which estimator that we used. Battese–Coelli estimator, Battese–Coelli estimator combined with input share equations, and a first order Taylor series approximation to the Kumbhakar's (1997) exact model for allocative inefficiency, which introduces the allocative inefficiency in a consistent way, are a few examples for such estimators.

[10] More specifically, $MC_{enp} = \hat{\beta}_4\frac{\hat{C}}{enp}$ and $MC_{proq} = \hat{\beta}_5\frac{\hat{C}}{proq}$ where $\hat{\beta}_4$ and $\hat{\beta}_5$ are the parameter estimates for β_4 and β_5 respectively; and $\hat{C}$ is the prediction of cost.

Table 1
Cost estimates.

tc	Inefficient		Symmetric	
lp	0.2370***	(0.0520)	0.2794***	(0.0588)
kp	0.2786***	(0.0297)	0.2006***	(0.0280)
ep	0.0627*	(0.0320)	0.0799*	(0.0351)
enp	0.0489*	(0.0194)	0.1648***	(0.0246)
nroq	0.1159***	(0.0208)	0.1107***	(0.0115)
proq	0.5196***	(0.0339)	0.6420***	(0.0262)
sl	−0.0179	(0.0453)	−0.0962**	(0.0348)
lf	−0.5522***	(0.0590)	−0.5825***	(0.0560)
aves	−0.3260**	(0.1010)	−0.6280***	(0.0844)
fuel	−0.2457***	(0.0336)	−0.3457***	(0.0378)
qtr1	0.0093	(0.0069)	0.0067	(0.0087)
qtr2	−0.0022	(0.0080)	−0.0244*	(0.0099)
qtr3	−0.0013	(0.0087)	−0.0327**	(0.0103)
iraniraq	−0.0184	(0.0112)	−0.0165	(0.0129)
gulfwar	−0.0036	(0.0126)	0.0115	(0.0151)
airtrfc	0.0170	(0.0111)	0.0075	(0.0129)
aamrgcl	0.0818***	(0.0224)	−0.0222	(0.0156)
coandea	−0.0000	(0.0536)	−0.0270	(0.0597)
comrgfr	0.3156***	(0.0544)	0.2104***	(0.0580)
comrgti	−0.2309***	(0.0390)	−0.1671***	(0.0293)
rcmrgha	0.2630***	(0.0406)	0.0115	(0.0195)
nwmrgrc	0.1475***	(0.0339)	−0.0611***	(0.0169)
twmrgoz	0.0934***	(0.0274)	−0.0370*	(0.0161)
usmrgpsa	0.2591***	(0.0406)	0.1820***	(0.0208)
dlmrgwn	0.0800***	(0.0208)	−0.0592***	(0.0155)
uapilot	−0.0065	(0.0492)	0.0352	(0.0552)
copilot	−0.0552*	(0.0263)	−0.1010***	(0.0283)
t			−0.8251***	(0.0984)
t^2			0.7525***	(0.0862)
cons			1.4159***	(0.2482)
N	500		500	

Standard errors in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

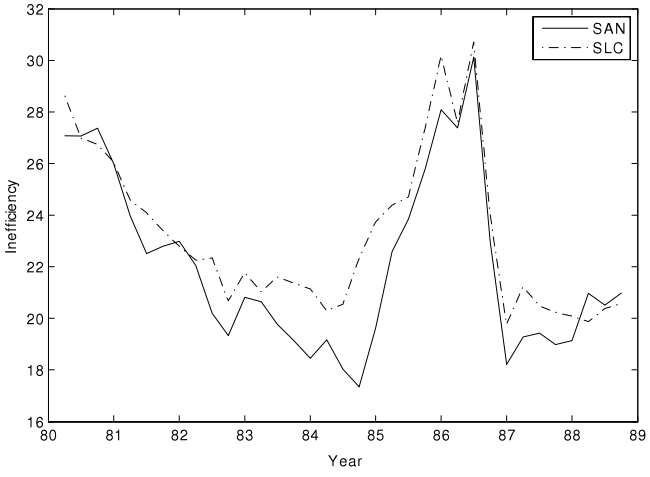

Fig. 1. City-pair inefficiency.

Our MC estimates correspond very well with those of Perloff et al. (2003). Fig. 1 gives the market share weighted efficiencies.

In our data period the behavior of the airlines is affected by at least three important events: The Airline Deregulation Act (1978); the second oil crisis (1979); and 1980's oil glut (1980–1986). The 1980's oil glut is the period in which the price of the oil fell after the second oil crisis. Although the price of oil fell for about six years, it never reached its pre-oil crisis levels. Deregulation increased the competitiveness which put a downward pressure on the market power of airlines. The oil crisis have two opposing effects on the market power: An increase in cost due to increase in input prices and a (possible) decrease in the dynamic cost due to binding incentive compatibility constraints. The overall effect of the oil crisis depends on the functional forms of demand and supply as well as the punishment scheme of the dynamic game that the airlines are playing. If the former effect dominates, then the oil crisis would put a downward pressure on the market power of the airlines. Hence, both the deregulation and the oil crisis decrease the market power. Once the effects of the shocks are settled down, the efficiency level go back to its stationary level. The inefficiency level between 1980 and 1986 accords with this type of behavior and is consistent with the quiet life hypothesis. Since a positive cost shock (might) increases the incentive to deviate from a coalition, the sudden decrease in the inefficiency level following the 1986 oil price collapse might be due to a breakdown in coalition.[11]

3.4. Estimating the supply–demand system

In Section 3.1, we described our general model. This section provides further details about our empirical example and the estimation procedure that we used in order to estimate this model.

3.4.1. The supply–demand system

In this section, we provide further details about the demand and supply equations. Recall that we assume that the good is homogenous. Hence, there is only one price in our model. In our theoretical description, made in Section 3.1, we did not specify how we calculated the market price. In our empirical example, we use the market share weighted price in order to calculate the aggregate market price. This aggregation of prices accords with the way in which we define the Lerner index as well. For each city-pair market, the inverse demand for firm i is assumed to be as follows:

$$P_t = \beta_0 + \beta_1 Q_t + \beta_2 PCI_t + \sum_{k=1}^{3} \delta_{i,k} Qtr_{kt} + \varepsilon_t \tag{18}$$

where P_t is the market share weighted price, PCI is the population weighted per capita income for relevant city-pairs, Qtr are seasonality dummies, and ε_{it} is the error term.

Firms' dynamic behaviors are influenced by current demand levels, expected future demands, current costs, and expected future costs (see, for example, Borenstein and Shepard, 1996). The sustainability of a collusion depends on the gain from deviation and expected future loss due to the punishment. For a variety of interesting cases[12] the higher today's demand is relative to the expected future demand, the weaker the punishment becomes relative to the gain from deviation. Similar arguments hold for the cost shocks. As we mentioned earlier the shock variables are observable by the econometrician. We use industry market output divided by expected industry market output for the next period as our demand shock variable. We proxy expected industry market output with future output. Input price indices are geometric means of the expenditure share-weighted input prices. The weights in this index are expenditure shares of each cost component. The supply shock is given by the ratio of this index to its expectation for the next period. Again we use the next period value of this index as a proxy to its future expectation. We use demeaned shock variables

[11] As a check on our results we have estimated cost efficiencies for the US airlines based on the annual data of Baltagi et al. (1995, 1998). We find a similar pattern of efficiency for the period 1980–1989 using both the data in Baltagi et al. and the Good–Sickles quarterly panel data (Wingrove et al., 1997).

[12] See Rotemberg and Saloner (1986).

in our estimations. More specifically, we calculate the shocks as follows:

$$cs_t = \frac{ICI_t}{ICI_{t+1}} - mean\left(\frac{ICI_t}{ICI_{t+1}}\right) \tag{19}$$

$$ds_t = \frac{IQ_t}{IQ_{t+1}} - mean\left(\frac{IQ_t}{IQ_{t+1}}\right) \tag{20}$$

where LP_t, KP_t, EP_t, and MP_t are the labor, capital, energy and materials prices, $ICI_t = LP_t^{a_{1t}} KP_t^{a_{2t}} EP_t^{a_{3t}} MP_t^{a_{4t}}$ is the industry cost index, IQ_t is the industry quantity, a_{it}'s are expenditure shares, cs_t is the cost shock variable, and ds_t is the demand shock variable.

The dynamic factor, μ_t^*, measures the shadow cost of coalition and represents the incentive compatibility constraint in Eq. (8). We model μ_t^* as a linear function of the demand and supply shocks and the inefficiency:

$$\mu_t^* = \mu_0 + \mu_1 cs_t + \mu_2 ds_t + \mu_3 ie_t$$

where *cs* and *ds* are the shock variables defined in Eqs. (19) and (20); and *ie* is the city-pair inefficiency level.

In Section 3.1 we derived our supply equation which was given in Eq. (12). This equation is:

$$\theta_t \beta_1 Q_t^* + MK_t - \mu_t^* = 0 \tag{21}$$

where $MK_{it} \equiv P_{it} - MC_{it}$ is the markup for firm i, $MK_t \equiv \sum_i MK_{it} s_{it}$ is the market share weighted markup, and μ_t^* is the dynamic factor which reflects the incentive compatibility constraint.

Conduct, θ_t, is modeled as an unobserved time-varying state whose evolution is generated by $AR(1)$ shocks. The augmented demand–supply system thus becomes:

$$P_t = \beta_0 + \beta_1 Q_t + \beta_2 PCI_t + \sum_{k=1}^{3} \delta_{i,k} Qtr_{kt} + \varepsilon_{1t} \tag{22}$$

$$\begin{aligned} MK_t &= -\alpha_t \beta_1 Q_t - \theta \beta_1 Q_t + \mu_0 + \mu_1 cs_t + \mu_2 ds_t \\ &\quad + \mu_3 ie_t + \varepsilon_{2t} \end{aligned} \tag{23}$$

$$\alpha_{t+1} = \rho \alpha_t + \eta_t \tag{24}$$

where $\theta \equiv E[\theta_t|\Psi]$, $\alpha_t \equiv \theta_t - \theta$, $\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim N(0, H)$, $\eta_t \sim N(0, Q)$, and $\alpha_1 \sim N(0, \frac{Q}{1-\rho^2})$.

As we will describe in the following section the calculation of markup, *MK*, depends on the way in which we calculate the marginal cost, *MC*. For example, the traditional market power measures, *L* (the Lerner index), and the analysis uses the observed marginal cost whereas L^{SFA} uses the efficient marginal cost, MC^{SFA}, that we described earlier.

3.4.2. Econometric procedure

We estimate the supply–demand system via the square root Kalman filter. The Kalman (1960) filter is a very useful technique for estimating time-varying parameter models.[13] In this approach the time-varying parameters, $\alpha_1, \alpha_2, \ldots, \alpha_n$, are assumed to be slowly changing unobserved states that are associated with observations, $y_1, y_2, \ldots, y_n$. A Kalman filer model consists of two equations: (1) Measurement and (2) Transition. In the measurement equation the relationship between α_t and y_t is modeled; and in the transition equation the relationship between α_t and α_{t+t} is modeled. In order to give some intuition about the Kalman filter, consider the following univariate model:

$$y_t = \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H) \tag{25}$$

$$\alpha_{t+1} = \alpha_t + u_t, \quad u_t \sim N(0, Q) \tag{26}$$

$$\alpha_1 \sim N(a_1, P_1) \tag{27}$$

where ε_t's and u_t's are mutually independent and are independent of α_1.

This simple Kalman filter model is called a local level model. The first equation is the measurement equation and the second equation is the transition equation. Let $a_{t+1} = E[\alpha_{t+1}|Y_t]$ be the prediction of α_{t+1} conditional on the information at time t and $P_{t+1} = \mathrm{Var}[\alpha_{t+1}|Y_t]$ be the conditional variance of α_{t+1} where $Y_t = \{y_1, y_2, \ldots, y_t\}$. The one step ahead forecast error, $v_t = y_t - a_t$, and its variance, F_t, are very important for the Kalman filter estimation. Given a_t and P_t the Kalman filter recursions enable us to calculate $a_{t+1} = a_t + K_t v_t$ and $P_{t+1} = P_t(1 - K_t) + Q$ where $K_t = \frac{P_t}{F_t}$ is so called the Kalman gain. At time $t - 1$ we predicted α_t by using Y_{t-1}. Then, at time t we can update our prediction for α_t by using the additional information, i.e., y_t. Our prediction for α_{t+1} at time t (i.e., a_{t+1}) is the same as our prediction of α_t at time t (i.e., $a_{t|t} = E[\alpha_t|Y_t]$). Hence, in $a_{t+1} = a_t + K_t v_t$ equation, the K_t term is the optimal weight between a_t and v_t. If the uncertainty in our earlier prediction is high (i.e., P_t is large), then more weight is assigned to the new observation. Similarly, if the variance in the forecasting error is large (i.e., F_t is large), then the new data is not reliable and its weight should be small. Calculation of the Kalman filter requires knowledge or estimation of the initial values a_1 and P_1. Most of the time we do not know these values and we should estimate them. The linear structure described here implies that $y = (y_1, y_2, \ldots, y_n)$ is normally distributed and the maximum likelihood estimation (MLE) method can be used to estimate the system parameters (including the initial values).[14] While this procedure enables us to write the likelihood function for y, it does not use future observations when predicting the unobserved states. For more reliable estimates for the unobserved states one should use full information. This procedure is called smoothing. After getting the MLE parameter estimates, one can get the smoothed estimates of the unobserved states and its variance, i.e., $E[\alpha_t|Y_n]$ and $\mathrm{Var}[\alpha_t|Y_n]$.

Unfortunately, for many econometric applications the Kalman filter provides invalid estimates due to the problem of endogeneity. In order to solve this problem some researchers[15] use the fitted values of the endogenous regressors rather than the variables themselves in the Kalman filter estimation. This procedure resembles two-stage least squares (2SLS) but, as Kim and Kim (2007) mention, it has no theoretical justification. Kim (2006) proposes a Heckman-type two-step MLE procedure[16] that deals with the endogeneity problem for single equation time-varying parameter models.[17] The first stage is similar to the first stage of 2SLS. The only difference is that instead of using OLS to predict the expected value of the endogenous right-hand-side variables, instruments for the right-hand-side endogenous variables are constructed from the traditional Kalman filter in which the coefficients are time-varying. Kim (2006) assumes that the

[13] See Harvey (1989) and Durbin and Koopman (2001) for detailed treatments of the Kalman filter in econometric applications. See also Koopman et al. (2007) for a basic introduction to the Kalman filter.

[14] In our example the unobserved state is not stationary. Hence, the initialization requires some caution. Whenever the model contains stationary unobserved states, one can use the initial values that are consistent with the stationarity. (See, Durbin and Koopman (2001) for more details about initialization.)

[15] For example, McKiernan (1996), Bacchetta and Gerlach (1997), and Peersman and Pozzi (2004).

[16] See Heckman (1976).

[17] See Kim and Nelson (2006) for an application of Kim (2006).

prediction error for the first stage is correlated with the error term from measurement equation. Using the error terms from the first stage, a new measurement equation is specified that corrects the endogeneity bias. Kim and Kim (2007) criticize this approach as it does not specify a direct correlation between the first stage error term and the error term from the measurement equation. Moreover, Kim's approach fails to correct Pagan's (1984) generated regressors' problem in the second step. Kim and Kim (2007) provide a joint estimation method[18] as well as a two-stage method for dealing with endogeneity. In order to estimate our empirical model, we extended the joint estimation method of Kim and Kim (2007) to the multivariate case. Moreover, the traditional Kalman filter estimation is known to be numerically unstable due to rounding errors which might cause variances to be non-positive definite during the update process (Durbin and Koopman, 2001). A solution to this issue, which has been viewed as impractical due to computational complexity, uses the square root Kalman filter. The square root Kalman filter is based on the Givens transformation of the underlying variance matrices.

In contrast to the Kalman filter, the square root Kalman filter updates the lower triangular parts of the relevant variance matrices. This is achieved by updating a specific UU' decomposition of a square matrix consisting of the variance of innovations (F), the variance of state variables (P), and the Kalman gain matrix (K) times F. The lower triangular decompositions F and P; and K may be calculated via the Givens rotations. A Givens rotation is a rotation in the plane spanned by two coordinates axes that can be used to zero out one element of U matrix. After a successive application of Givens transformations the U matrix can be transformed into a lower triangular rectangular matrix.[19]

In what follows we describe our method to extend the procedure of Kim and Kim (2007) to the multivariate square root Kalman filter framework. In addition to a vector of dependent variables, y_t, we model a vector of endogenous variables, x_t, via the Kalman filter. We assume that the error term from the measurement equation of the dependent variable, ε_t, and the error term from the measurement equation of the endogenous variables, e_t, are jointly normally distributed, which allows ε_t to be decomposed into two components. One is correlated with the endogenous regressors and the other is not. After this decomposition one can calculate the relevant log-likelihood function via two separate Kalman filter runs.

Consider the following model with endogenous explanatory variables[20]:

$$y_t = X_t\alpha_{1,t} + \varepsilon_t, \quad \varepsilon_t \sim \mathbf{N}(0, H_1) \tag{28}$$

$$\alpha_{1,t+1} = \tau_1\alpha_{1,t} + R_1u_{1,t}, \quad u_{1,t} \sim \mathbf{N}(0, Q_1) \tag{29}$$

$$\alpha_{1,1} \sim \mathbf{N}(a_{1,1}, P_{1,1}) \tag{30}$$

$$x_t = Z_t\alpha_{2,t} + e_t, \quad e_t \sim \mathbf{N}(0, H_2) \tag{31}$$

$$\alpha_{2,t+1} = \tau_2\alpha_{2,t} + R_2u_{2,t}, \qquad u_{2,t} \sim \mathbf{N}(0, Q_2) \tag{32}$$

$$\alpha_{2,1} \sim \mathbf{N}(a_{2,1}, P_{2,1}) \tag{33}$$

where y_t is a $p \times 1$ vector of observations, x_t is a $m \times 1$ vector of endogenous regressors; $Z_t = I_m \otimes z_t'$ where z_t is a $l \times 1$ (with $l \geq m$) vector of exogenous variables (instruments); and $X_t = I_p \otimes x_t'$, τ_i is a transition matrix, R_i is a selection matrix and determines whether a state will be stochastic or not. We assume that ε_t, e_t, $u_{1,t}$, and $u_{2,t}$ are serially independent and independent at all other time periods. Moreover, error terms are independent with $\alpha_{1,1}$ and $\alpha_{2,1}$. Exogenous variables are not explicitly introduced in order to streamline the notation. They can be added later to the model (and are) in a straightforward manner consistent with the identifiability of the structural parameters.

Let $\tilde{e}_t$ be the standardized version of e_t. Thus, the variance of $\tilde{e}_t$ is the identity matrix. The joint distribution of $\begin{bmatrix}\tilde{e}_t & \varepsilon_{t1} & \varepsilon_{t2} & \dots & \varepsilon_{tp}\end{bmatrix}'$ is given by:

$$\begin{bmatrix} \tilde{e}_t \\ \varepsilon_{t1} \\ \varepsilon_{t2} \\ \vdots \\ \varepsilon_{tp} \end{bmatrix} \sim \mathbf{N}(0, \Omega) \tag{34}$$

where

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} I_m & \rho_1\sigma_{\varepsilon_1} & \rho_2\sigma_{\varepsilon_2} & \cdots & \rho_p\sigma_{\varepsilon_p} \\ \rho_1'\sigma_{\varepsilon_1} & \sigma_{\varepsilon_1}^2 & \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \cdots & \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} \\ \rho_2'\sigma_{\varepsilon_2} & \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \cdots & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_p'\sigma_{\varepsilon_p} & \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} & \cdots & \sigma_{\varepsilon_p}^2 \end{bmatrix}$$

$$\Omega_{11} = I_m$$

$$\Omega_{12} = \begin{bmatrix} \rho_1\sigma_{\varepsilon_1} & \rho_2\sigma_{\varepsilon_2} & \cdots & \rho_p\sigma_{\varepsilon_p} \end{bmatrix}$$

$$\Omega_{21} = \begin{bmatrix} \rho_1'\sigma_{\varepsilon_1} \\ \rho_2'\sigma_{\varepsilon_2} \\ \vdots \\ \rho_p'\sigma_{\varepsilon_p} \end{bmatrix}$$

$$\Omega_{22} = H_1 = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \cdots & \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} \\ \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \cdots & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} & \cdots & \sigma_{\varepsilon_p}^2 \end{bmatrix}.$$

Hence we have:

$$Proj\left(\begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \vdots \\ \varepsilon_{tp} \end{bmatrix} | \tilde{e}_t \right) = \Omega_{21}\Omega_{11}^{-1}\tilde{e}_t = \begin{bmatrix} \rho_1'\sigma_{\varepsilon_1} \\ \rho_2'\sigma_{\varepsilon_2} \\ \vdots \\ \rho_p'\sigma_{\varepsilon_p} \end{bmatrix} \tilde{e}_t \tag{35}$$

and see Box I.

This implies that:

$$\begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \vdots \\ \varepsilon_{tp} \end{bmatrix} = \Gamma\tilde{e}_t + \varpi_t \tag{37}$$

where $\varpi_t \sim N(0, MSE)$ and $\Gamma \equiv \begin{bmatrix} \rho_1'\sigma_{\varepsilon_1} \\ \rho_2'\sigma_{\varepsilon_2} \\ \vdots \\ \rho_p'\sigma_{\varepsilon_p} \end{bmatrix}$.

Hence, we can write the measurement equation as:

$$\begin{aligned} y_t &= X_t\alpha_{1,t} + \Gamma\tilde{e}_t + \varpi_t \\ &= X_t\alpha_{1,t} + \Gamma H_2^{-1/2}(x_t - Z_t\alpha_{2,t}) + w_t. \end{aligned} \tag{38}$$

[18] For a similar approach in the stochastic frontier analysis framework see Kutlu (2010).

[19] Note that the individual Givens transformation matrices can be multiplied together to construct the requisite orthogonal matrix triangularization. For more detail about Givens transformations, see Durbin and Koopman (2001) and Golub and Van Loan (1996).

[20] See Jin and Jorgenson (2010) for a special case where the parameters in the instrument equation (i.e. Eq. (28)) do not vary.

$$
\begin{aligned}
MSE &= \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \\
&= \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \cdots & \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} \\ \rho_{12}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \cdots & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p}\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} & \rho_{2p}\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} & \cdots & \sigma_{\varepsilon_p}^2 \end{bmatrix} - \begin{bmatrix} \rho_1'\sigma_{\varepsilon_1} \\ \rho_2'\sigma_{\varepsilon_2} \\ \vdots \\ \rho_p'\sigma_{\varepsilon_p} \end{bmatrix} \begin{bmatrix} \rho_1\sigma_{\varepsilon_1} & \rho_2\sigma_{\varepsilon_2} & \cdots & \rho_p\sigma_{\varepsilon_p} \end{bmatrix} \\
&= \begin{bmatrix} (1-\rho_1'\rho_1)\sigma_{\varepsilon_1}^2 & (\rho_{12}-\rho_1'\rho_2)\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \cdots & (\rho_{1p}-\rho_1'\rho_p)\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} \\ (\rho_{12}-\rho_1'\rho_2)\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & (1-\rho_2'\rho_2)\sigma_{\varepsilon_2}^2 & \cdots & (\rho_{2p}-\rho_2'\rho_p)\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} \\ \vdots & \vdots & \ddots & \vdots \\ (\rho_{1p}-\rho_1'\rho_p)\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} & (\rho_{2p}-\rho_2'\rho_p)\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} & \cdots & (1-\rho_p'\rho_p)\sigma_{\varepsilon_p}^2 \end{bmatrix}
\end{aligned} \tag{36}
$$

Box I.

Now, consider the following joint density function:

$$
\begin{aligned}
f(Y,X) &= \prod_t f(y_t, x_t | Y_{t-1}, X_{t-1}) \\
&= \prod_t f(y_t, |x_t, Y_{t-1}, X_{t-1}) f(x_t | X_{t-1})
\end{aligned} \tag{39}
$$

where $Y_t \equiv \{y_1, y_2, \ldots, y_t\}$ and $X_t \equiv \{x_1, x_2, \ldots, x_t\}$.

We have:

$$\hat{y}_t \equiv E[y_t | x_t, Y_{t-1}] = X_t a_{1,t} + \Gamma H_2^{-1/2}(x_t - Z_t a_{2,t}) \tag{40}$$

$$
\begin{aligned}
F_{1,t} &\equiv \mathrm{Var}[y_t | x_t, Y_{t-1}] \\
&= X_t P_{1,t} X_t' + \Gamma H_2^{-1/2} Z_t P_{2,t} Z_t' H_2^{-1/2} \Gamma' + T
\end{aligned} \tag{41}
$$

$$\hat{x}_t \equiv E[x_t | X_{t-1}] = Z_t a_{2,t} \tag{42}$$

$$F_{2,t} \equiv \mathrm{Var}[x_t | X_{t-1}] = Z_t P_{2,t} Z_t' + H_2 \tag{43}$$

where $a_{1,t} \equiv E[\alpha_{1,t} | x_t, X_{t-1}, Y_{t-1}]$, $P_{1,t} \equiv \mathrm{Var}[\alpha_{1,t} | x_t, X_{t-1}, Y_{t-1}]$, $a_{2,t} \equiv E[\alpha_{2,t} | X_{t-1}]$, $P_{2,t} \equiv \mathrm{Var}[\alpha_{2,t} | X_{t-1}]$, and see Box II.

The joint density function at time t in Eq. (39) becomes:

$$
\begin{aligned}
&f(y_t, x_t | Y_{t-1}, X_{t-1}) \\
&= (2\pi)^{-(p+m)/2} |F_t|^{-1/2} \exp\left(-\frac{1}{2} \upsilon_t' F_t^{-1} \upsilon_t\right)
\end{aligned} \tag{44}
$$

where $\upsilon_t \equiv \begin{bmatrix} y_t - \hat{y}_t \\ x_t - \hat{x}_t \end{bmatrix}$ and $F_t \equiv \begin{bmatrix} F_{1,t} & 0 \\ 0 & F_{2,t} \end{bmatrix}$.

In order to calculate Eq. (44) we begin with the state space model:

$$
\begin{aligned}
\begin{bmatrix} y_t \\ x_t \end{bmatrix} &= \begin{bmatrix} \Gamma H_2^{-1/2} x_t \\ 0 \end{bmatrix} \\
&\quad + \begin{bmatrix} X_t & -\Gamma H_2^{-1/2} Z_t \\ 0 & Z_t \end{bmatrix} \begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{bmatrix} + \begin{bmatrix} \varpi_t \\ e_t \end{bmatrix}
\end{aligned} \tag{45}
$$

$$\begin{bmatrix} \varpi_t \\ e_t \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} T & 0 \\ 0 & H_2 \end{bmatrix}\right) \tag{46}$$

$$\begin{bmatrix} \alpha_{1,t+1} \\ \alpha_{2,t+1} \end{bmatrix} = \begin{bmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{bmatrix} \begin{bmatrix} \alpha_{1,t} \\ \alpha_{2,t} \end{bmatrix} + \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \tag{47}$$

$$\begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}\right) \tag{48}$$

or more compactly as:

$$\tilde{y}_t = A_t + B_t \tilde{\alpha}_t + \tilde{\varepsilon}_t, \quad \tilde{\varepsilon}_t \sim \mathbf{N}(0, H) \tag{49}$$

$$\tilde{\alpha}_{t+1} = \tau \tilde{\alpha}_t + R\tilde{u}_r, \quad \tilde{u}_r \sim \mathbf{N}(0, Q). \tag{50}$$

One can use the traditional Kalman filter on equation system (49) and (50). Thus we have transformed our initial Kalman filter model into another Kalman filter model where the endogeneity is not a problem. This implies that all of the usual properties of the Kalman filter and the traditional updating equations for the Kalman filter can be applied to our equation system (49) and (50). For the sake of completeness we present the corresponding Kalman filter recursion equations:

$$\upsilon_t = \tilde{y}_t - A_t - B_t \tilde{a}_t \tag{51}$$

$$K_t = \tau P_t B_t' F_t^{-1} \tag{52}$$

$$\tilde{a}_{t+1} = \tau \tilde{a}_t + K_t \upsilon_t \tag{53}$$

$$L_t = \tau - K_t B_t \tag{54}$$

$$P_{t+1} = \tau P_t L_t' + RQR'. \tag{55}$$

From these we can compute the corresponding square root Kalman filter equations based on the Givens rotations. Let:

$$U_t = \begin{bmatrix} B_t \tilde{P}_t & \tilde{T} & 0 \\ \tau \tilde{P}_t & 0 & R\tilde{Q} \end{bmatrix} \tag{56}$$

where $P_t = \tilde{P}_t \tilde{P}_t'$, $T = \tilde{T}\tilde{T}'$, $Q = \tilde{Q}\tilde{Q}'$ in which matrices $\tilde{P}_t$, $\tilde{T}_t$, and $\tilde{Q}$ are lower triangular matrices.

Note that:

$$U_t U_t' = \begin{bmatrix} F_t & B_t P_t \tau' \\ \tau P_t B_t' & \tau P_t \tau' + RQR' \end{bmatrix}. \tag{57}$$

One can transform U_t by Givens transformations so that $U_t U_t' = \tilde{U}_t \tilde{U}_t'$ where $\tilde{U}_t = \begin{bmatrix} \tilde{U}_{1,t} & 0 & 0 \\ \tilde{U}_{2,t} & \tilde{U}_{3,t} & 0 \end{bmatrix}$ is a lower triangular rectangular matrix. We deduce that:

$$
\begin{aligned}
\tilde{U}_{1,t} &= \tilde{F}_t \\
\tilde{U}_{2,t} &= K_t \tilde{F}_t \\
\tilde{U}_{3,t} &= \tilde{P}_{t+1}
\end{aligned} \tag{58}
$$

where $F_t = \tilde{F}_t \tilde{F}_t'$ and $\tilde{F}_t$ is lower triangular.

Thus by updating $\tilde{U}_t$ we obtain square root updated version of P_{t+1}. Update for a_{t+1} is also straightforward and given as:

$$\tilde{a}_{t+1} = \tau \tilde{a}_t + \tilde{U}_{2,t} \tilde{U}_{1,t}^{-1} \upsilon_t. \tag{59}$$

We also provide the smoothing and the corresponding square root equations. The smoothing equations are not affected apart from the way in which relevant variables are computed in the Kalman filter step and given as:

$$r_{t-1} = B_t' F_t^{-1} \upsilon_t + L_t r_t \tag{60}$$

$$\widehat{\alpha}_t = a_t + P_t r_{t-1} \tag{61}$$

$$N_{t-1} = B_t' F_t^{-1} X_t + L_t N_t L_t' \tag{62}$$

$$V_t = P_t - P_t N_{t-1} P_t \tag{63}$$

$$T \equiv \begin{bmatrix} (1-\rho_1'\rho_1)\sigma_{\varepsilon_1}^2 & (\rho_{12}-\rho_1'\rho_2)\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \cdots & (\rho_{1p}-\rho_1'\rho_p)\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} \\ (\rho_{12}-\rho_1'\rho_2)\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & (1-\rho_2'\rho_2)\sigma_{\varepsilon_2}^2 & \cdots & (\rho_{2p}-\rho_2'\rho_p)\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} \\ \vdots & \vdots & \ddots & \vdots \\ (\rho_{1p}-\rho_1'\rho_p)\sigma_{\varepsilon_1}\sigma_{\varepsilon_p} & (\rho_{2p}-\rho_2'\rho_p)\sigma_{\varepsilon_2}\sigma_{\varepsilon_p} & \cdots & (1-\rho_p'\rho_p)\sigma_{\varepsilon_p}^2 \end{bmatrix}$$

Box II.

Table 2
Dynamic game estimates.

	SAN				SLC			
	Inefficient	Efficient	Symmetric	Static	Inefficient	Efficient	Symmetric	Static
β_0	−179.7954	−91.1969***	−136.0509***	−90.1739***	−94.6148***	−110.8948***	−125.9973***	−141.1308**
	(99.8867)	(26.7361)	(4.4794)	(6.3614)	(17.5149)	(22.0313)	(26.4067)	(50.0440)
β_1	−10.1364***	−8.5972***	−9.4504***	−8.8341***	−6.9566***	−7.0955***	−7.1875***	−4.7394***
	(1.5578)	(0.9531)	(0.6683)	(0.7387)	(1.1435)	(0.9739)	(0.9887)	(1.3928)
β_2	60.7070***	50.8623***	56.1100***	51.3590***	51.1167***	52.3596***	54.2783***	47.3161***
	(10.2063)	(3.1854)	(1.6731)	(1.7671)	(3.0968)	(3.5308)	(3.6462)	(5.4448)
δ_1	−17.8680*	−18.1089*	−17.5196*	−17.5025*	15.4983	18.9987	13.4377	33.9590
	(8.5293)	(7.6959)	(8.3244)	(7.9652)	(12.2538)	(15.7154)	(13.0270)	(17.7662)
δ_2	32.8212*	21.4005*	26.2222**	24.0242**	−14.5617	−5.0397*	−17.8894	6.6223
	(14.0222)	(10.7864)	(9.4196)	(7.6494)	(17.1960)	(14.6144)	(14.3152)	(13.3151)
δ_3	46.8010***	35.0231**	38.5866***	35.9539***	−0.0746	5.3498	−5.3553	−3.2223
	(13.6080)	(11.2296)	(11.1186)	(9.2927)	(16.7587)	(13.9064)	(18.4702)	(12.7022)
μ_0	49.5252	109.6428*	0.0000	0.0000	202.2628***	173.9227***	166.6795*	0.0000
	(45.5529)	(51.0647)	(–)	(–)	(60.5038)	(31.7741)	(72.7441)	(–)
μ_1	−41.7642**	−55.2878*	0.0000	0.0000	−115.1264***	−122.8321*	−88.2581	0.0000
	(13.1419)	(26.5138)	(–)	(–)	(32.5559)	(48.9401)	(60.2404)	(–)
μ_2	95.0603**	63.6637	0.0000	0.0000	160.0243	133.6866	163.4894	0.0000
	(35.6776)	(103.9800)	(–)	(–)	(90.4186)	(143.9166)	(86.1330)	(–)
μ_3	−2.6598	0.0000	0.0000	0.0000	−2.1856	0.0000	0.0000	0.0000
	(1.4532)	(–)	(–)	(–)	(1.3397)	(–)	(–)	(–)
θ	0.1167	−0.0187	−0.0837*	0.0975***	−0.1501	−0.0491	−0.4498	0.5822***
	(0.1281)	(0.1668)	(0.0332)	(0.0260)	(0.1652)	(0.1272)	(0.3154)	(0.1143)
ρ	0.7899***	0.8360***	0.8407***	0.8466***	0.8219***	0.6638*	0.9530***	0.8555***
	(0.2103)	(0.2501)	(0.1451)	(0.1649)	(0.1197)	(0.3382)	(0.1139)	(0.1961)
ρ_1	0.5483*	0.1623	0.3018	0.2226	0.3943**	0.4158*	0.4084*	−0.1863
	(0.2673)	(0.3158)	(0.2201)	(0.2452)	(0.1876)	(0.1888)	(0.1834)	(0.5205)
ρ_2	−0.8595***	−0.8374***	−0.8435***	−0.8611***	−0.0226	−0.0448	0.0161	−0.6463
	(0.0366)	(0.0984)	(0.0605)	(0.0521)	(0.3009)	(0.3785)	(0.3530)	(0.5049)

Standard errors in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

where $r_n = 0$, $N_n = 0$, and $\widehat{\alpha}_t$ and V_t are the smoothed state and variance, respectively.

For the square root version of these equations we only need to concentrate on N_t. Hence, we just explain the way in which N_{t-1} is updated.

Let:

$$N_{t-1}^* = \begin{bmatrix} B_t'\tilde{U}_{1,t}^{-1} & L_t'\tilde{N}_t \end{bmatrix} \tag{64}$$

where $N_t = \tilde{N}_t\tilde{N}_t'$ and $\tilde{N}_t$ is lower triangular.

Then it follows that by transforming the matrix N_{t-1}^* to a lower triangular matrix via Givens transformations we obtain $\tilde{N}_{t-1}$.

3.4.3. Estimation

In this section we present our estimates for the demand–supply system. The instrumental variables for our model are the industry output and the industry average price as well as the exogenous variables *PCI*, *Qtr*, *cs*, *ds*, and *ie*. Our estimates for the demand–supply system are given in Table 2. The inefficient column assumes that the airlines are playing a dynamic game and are allowed to be inefficient; the efficient column assumes that firms are playing a dynamic game and are fully efficient. In this counterfactual model we calculate the full efficiency frontier and assume that all firms share the corresponding efficient MC function; for the symmetric column we first determine whether the firms are in a static or dynamic environment, and then depending on this finding we assume that airlines are playing the corresponding game (SAN: static; SLC: dynamic); and the static column assumes that firms are playing a static game and are allowed to be inefficient. We used the likelihood ratio test in order to determine whether the firms are playing a static game or a dynamic game. For SAN at a 5% significance level, we conclude that the game is static. In contrast to SAN, for SLC the game turned out to be dynamic. In our empirical example, following Rotemberg and Saloner (1986), a boom in demand increases the incentive to deviate.[21] The bias in the market power estimates is larger

[21] Under the full-information structure we use, yet in a different quantity choice model, Rotemberg and Saloner (1986) show that whenever the demand and the cost are linear, the incentive to deviate from collusion increases as demand increases. For the non-linear case they show that this result does not necessarily hold.

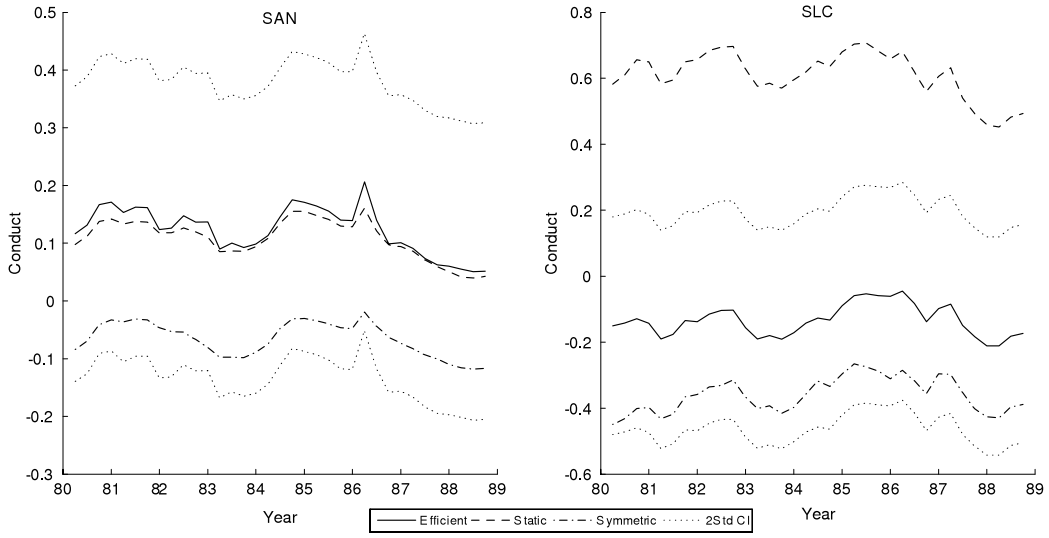

Fig. 2. City-pair conducts.

whenever there is a demand boom or a positive cost shock. Note that for both SAN and SLC the cost shock and the demand shock have opposite signs as expected. For both city-pairs the inefficiency terms in the dynamic factors are significant at the 10% level. When colluding an efficient firm has more incentive to deviate, both because it might gain more from undercutting its rivals and because it has less to fear from its inefficient rivals' retaliations. Hence, an increase in efficiency leads to a higher shadow cost of participating in a coalition. In our study for both city-pairs the coefficient of the *ie* term is negative which is in line with the idea that more efficient firms are more likely to deviate from the coalition.

In order to gauge the extent to which market power estimates change when firms are assumed to utilize their full efficiencies, we compare the market powers of the inefficient MC and efficient MC cases. For both city-pairs we did not observe a significant change in market power. The reason for this seems to be that even in the fully efficient case the firms do not have market power. Hence, in the inefficient case they did not have market power as well. Also, Delis and Tsionas (2009) mention that those firms that are close to the frontier obey the efficient structure hypothesis. Thus the increase in inefficiency decreases the market power. In our case it seems that they are statistically quite similar. However, when we use another version of full efficiency assumption (assuming symmetry) the fully efficient version underestimated the market power. Here the reason was that in this case our MC estimates are higher than the symmetric ones, most probably because of misspecification.

Fig. 2 gives the dynamic conducts, two standard error confidence intervals for the dynamic conducts, the symmetric cost conducts, and the static conducts. The biases are upwards under the assumption of a static game and downwards for the symmetric cost case. The upward bias is due to ignored dynamic factors and the downward bias is due to over-estimation of MCs.

We can also examine the reduced form relationship between the estimates of inefficiency and conduct. The OLS estimates for the inefficiency-conduct slopes are given in Table 3. We observe a negative relationship between efficiency and market power, a result that mirrors that found in the banking industry by Berger and Hannan (1998) and which supports the QLH, but counters the findings of Maudos and Fernández de Guevara (2007) who find a positive relationship between market power and cost efficiency. Delis and Tsionas (2009) show that the QLH is supported on average but that those banks with more efficient management have relatively more market power.

Table 3
Conduct versus inefficiency.

Inefficiency	Coeff.	SE
SAN		
Conduct	36.3337**	(10.7063)
Constant	17.5995***	(1.7082)
SLC		
Conduct	33.8719***	(8.5081)
Constant	27.8025***	(1.1162)

Robust errors in parentheses
*$p < 0.05$.
** $p =< 0.01$.
*** $p < 0.001$.

3.5. Welfare analysis

In this section we consider the welfare implications of our model. For this purpose we use the SAN estimates as an illustration. We assume that MC is constant and airlines are playing a static game. In Fig. 3 we provide the estimated conduct as well as two equilibrium paths. Our equilibrium estimate is very close to the sample mean of (Q, P). In this figure: EMC is the efficient marginal cost; MC is the full marginal cost; P is the price; MR is the marginal revenue; PMR is the perceived marginal revenue for the estimated equilibrium; mean Q is the sample mean of (Q, P); Eqm path QLH and Eqm path ESH are two equilibrium paths that are consistent with QLH and ESH, respectively; and 2 std CI PMR is the two standard error confidence intervals for PMR.

We have at least two types of problems if we do not consider the inefficiencies of the firms for a market power analysis. First, ignoring inefficiencies of the firms from the analysis might lead to inconsistent parameter estimates. Second, even if the parameter estimates are consistent the DWL estimates would be inaccurate. In the inefficiency context, a shift from monopoly to competition not only lowers the price but also changes the MC.[22] Hence, in contrast to the traditional market power and DWL analysis, we have to calculate the EMC. By stochastic frontier techniques one can easily calculate the EMC.

In this section, we consider the second problem. We start by discussing Fig. 3–4. Assume that the econometrician estimates

22 See Comanor and Leibenstein (1969) and Parish and Ng (1972) for more detailed arguments about DWL calculation under inefficiency.

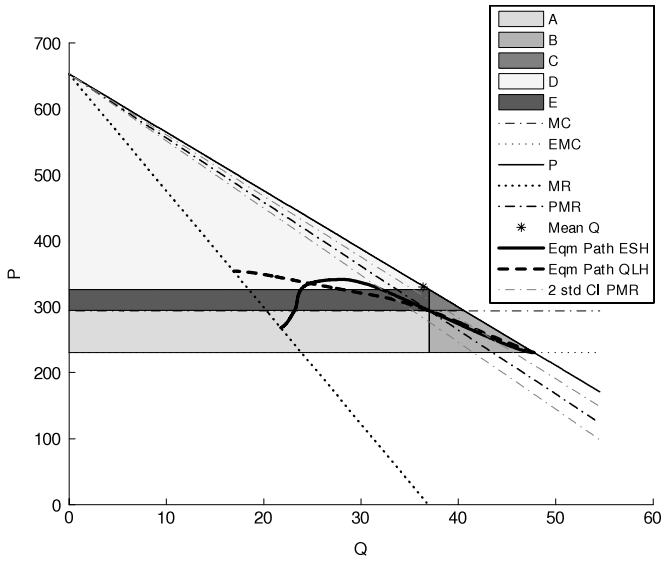

Fig. 3. Equilibrium analysis.

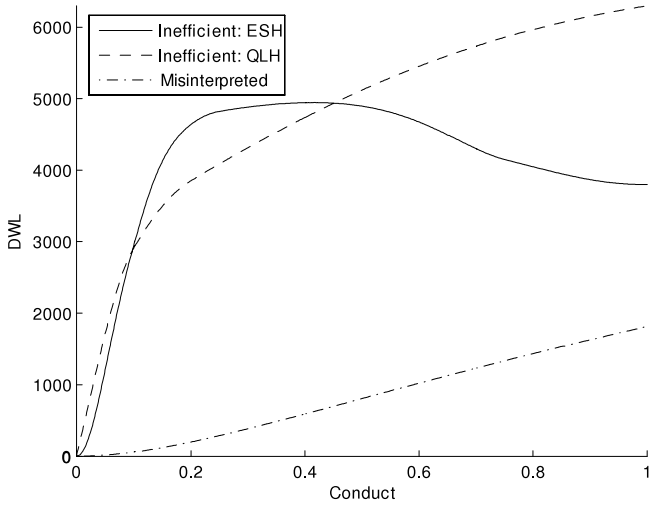

Fig. 4. DWL comparison.

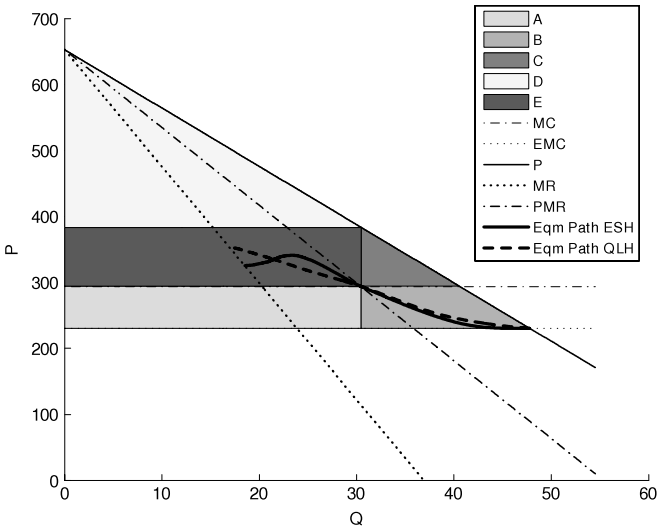

Fig. 5. Equilibrium analysis.

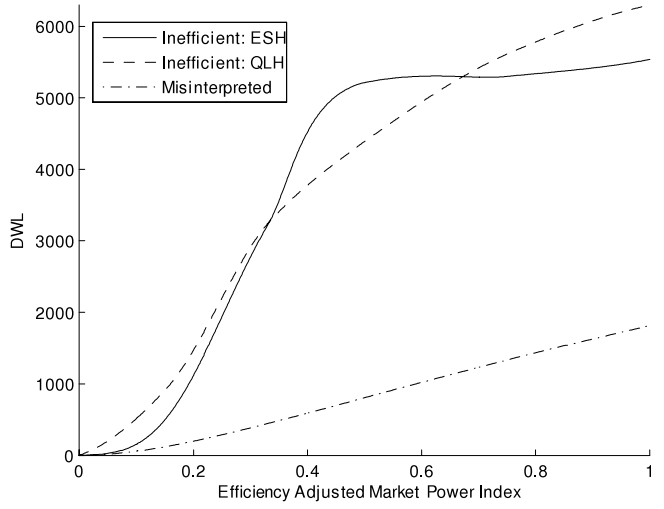

Fig. 6. DWL comparison.

the relevant parameters consistently but believes that the heterogeneity among airlines is due to firm specific differences rather than the difference in the efficiencies of these airlines. Here we assume that the econometrician consistently estimates μ_t as well. When generating Fig. 3 we abstract ourselves from seasonal shifts by using the data for the fourth quarter. In order to predict the inverse demand curve, we use the mean of PCI for the relevant time periods. The aggregate MC and EMC are constructed by summing market share weighted firm specific MC and EMC, respectively. The PMR is constructed by using $PMR(Q_t) = \hat{\theta}\hat{\beta}_1 Q_t + \hat{P}(Q_t)$ formula where hats represent predictions. The equilibrium paths are chosen such that they pass through the intersection of PMR and MC curves. Moreover, we assume that under perfect competition firms are fully efficient. Hence, these paths pass through the intersection of inverse demand and EMC curves. For both cases the maximum inefficiency level that is reached is approximately equal to 0.35. The DWL values for Fig. 4 are calculated by numerical integration techniques. Note that although the MCs are constant, as the conduct changes the efficiency levels of the firms are changing as well. Hence, from the equilibrium paths we can see the corresponding MC values for a given conduct value. The following set of figures (Figs. 5–6) are generated similarly. The econometrician would conclude that the DWL is equal to the area C. On the other hand if we consider the inefficiencies of the firms, then the DWL is given by the area $A+B+C$. In our case $A+B+C \gg C$. Unfortunately, the problem is more severe than just having a size difference for DWL estimates. While the traditional DWL is a monotone function of the conduct, for the inefficiency case this might not be true. If we believe that for some efficiency levels ESH holds, the equilibrium path will not be monotone. So the DWL would not be monotone as well. For our ESH equilibrium path unless the antitrust authorities could enforce very high levels of competition, it is preferable to have high market power levels. Finally, the misinterpreted measure does not find significant DWL for $\theta \in [0.1, 0.2]$. This is obviously not true if we take the inefficiencies of the airlines into account.

We examine one other counterfactual with our structural model and display our results in Fig. 5–6. The figures are based on the same parameter values that we used for Figs. 3 and 4 except that the value for the conduct parameter corresponds to that of symmetric Cournot competition with three firms, i.e., 1/3. In this counterfactual example we assume that the inefficiency level at the equilibrium is the same as that of Fig. 3. Again the equilibrium paths are chosen such that they pass through the intersection of the PMR and MC curves and that under perfect

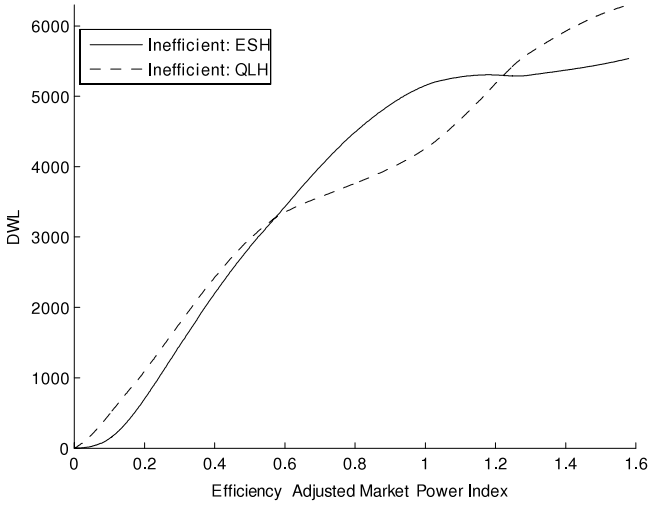

Fig. 7. DWL comparison.

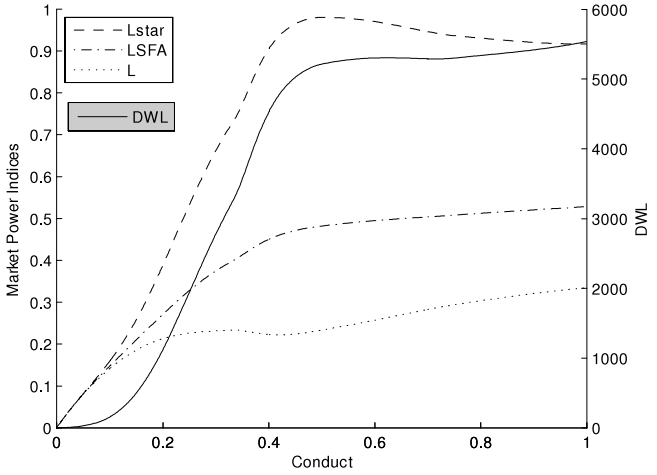

Fig. 8. Market power indices.

competition firms are fully efficient. The vital difference between the equilibrium paths for Figs. 3 and 5 is that for Fig. 3 even a small increment (from the competitive level) in the market power decreases the efficiency level dramatically. In Fig. 5, the change in the inefficiency level is relatively slower. As a result the "error" in DWL for the misinterpreted model is relatively lower for low levels of conduct. Hence, if a small deviation from perfect competition causes high efficiency loses, then the DWL calculations using the misinterpreted model would be less accurate. In such a case, if the antitrust authorities favor enforcement of low DWL levels, then they would be forced to settle on relatively low levels of conduct. Hence, unless we know the relationship between conduct and efficiency, it is very hard to evaluate the effects of market power.

In this section we have only considered the static version of the DWL calculation. Of course this analysis is not valid if we are in a dynamic framework. However, results of this section are instructive and point out that even in the simplest case the DWL calculation can be problematic. An examination of comparable issues using the dynamic version of the DWL calculation requires knowledge of the FMC curve as well as the MC and EMC curves.

Finally, although our conduct estimate θ is necessary for proper DWL and market power estimations, it is not efficiency adjusted. A possible efficiency adjusted market power index derived from the conduct estimates is given by:

$$L^{SFA} \equiv \frac{P - MC^{SFA} - \mu^*}{P}. \tag{65}$$

For our second example (DWL paths corresponding to Fig. 6) the DWL paths as a function of L^{SFA} are shown in Fig. 7. This index only proxies the DWL due to a socially non-optimal level of production, i.e., area $B + C$. The pure cost inefficiency effect on the DWL is captured in the rectangular area A. Depending on how A changes as a function of the conduct, θ, we can still have non-monotonic DWL paths as a function of L^{SFA}. Hence, L^{SFA} only proxies the DWL due to a socially non-optimal level of production and only partially captures the pure cost inefficiency effect. We propose the following novel measure as a proxy for the full DWL, i.e., area $A + B + C$:

$$L^*(a) = L^{SFA} + a\frac{MC - MC^{SFA}}{P}$$

where $a \geq 0$ weights the importance of the efficiency component.

Hence, this measure is a weighted average of two markups: (1) Price-MC markup and (2) Inefficient MC-efficient MC markup. Our preference for a is 2.[23] We choose $a = 2$ because the area A is rectangular and the area $B + C$ is triangular. Fig. 8 compares L, L^{SFA}, and $L^*(2)$ for the ESH equilibrium path of our second example. From Fig. 8 we can see that $L^*(2)$ performs better than other market power measures in terms of capturing the non-monotonic behavior of DWL as a function of conduct. Moreover, it captures the size of the DWL much better than the other measures.

4. Conclusion

The purpose of this paper was to point out the theoretical issues as well as the modeling issues which appear to conflict in the literature on market power by explicitly considering the role that distorted allocations may on inferential market power measures as well as on estimates of the dead weight loss suffered by the consumer when market power and collusion exist, while at the same time introducing an econometric method to address dynamics and time varying parameters in the dynamic game. More specifically, we have used a dynamic model of conduct in order to examine the relationship between market power and efficiency. We applied our model to two city-pairs in the US airline industry. Although for one of the city-pairs (SAN) we did not find evidence for a dynamic game, for the other city-pair (SLC) we concluded that the game is a dynamic one. We observe that the static conduct is biased upwards and the symmetric conduct is biased downwards. Moreover, a negative relationship is identified between the market conduct and the average efficiency of the market. This result accords with the quiet life hypothesis of Hicks (1935). Finally, we conclude that even if we can estimate the conduct consistently and make an efficiency adjustment to construct a market power measure, it is not easy to make inferences about the DWL. In order to solve this problem one has to identify the relationship between conduct and efficiency. In this paper, while we did not fully address this issue, by using stochastic frontier techniques we calculated the point DWL for the estimated equilibrium. Moreover, we provided a novel measure of market power, $L^*(2)$, that can proxy DWL relatively well. The biggest advantage of this measure is that the econometrician does not need to estimate the equilibrium path which in our experience appears to be a rather complicated exercise. We leave the case where the MC is endogenous to the dynamic game model as a future research. A more extensive empirical study is warranted but we consider this also outside the scope of this paper.

[23] Note that $L^*(-1) = L$ and $L^*(0) = L^{SFA}$.

References

Ahn, S., Good, D., Sickles, R.C., 2000. Estimation of long-run inefficiency levels: a dynamic frontier approach. Econometric Reviews 19, 461–492.

Alam, I., Sickles, R.C., 2000. A time series analysis of deregulatory dynamics and technical efficiency: the case of the US airline industry. International Economic Review 41, 203–218.

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production functions. Journal of Econometrics 6, 21–37.

Appelbaum, E., 1982. The estimation of the degree of oligopoly power. Journal of Econometrics 19, 287–299.

Bacchetta, P., Gerlach, S., 1997. Consumption and credit constraints: international evidence. Journal of Monetary Economics 40, 207–238.

Baltagi, B.H., Griffin, J.M., Rich, D.P., 1995. Airline deregulation: the cost pieces of the puzzle. International Economic Review 36, 245–259.

Baltagi, B.H., Griffin, J.M., Vadali, S.R., 1998. Excess capacity: a permanent characteristic of US airlines. Journal of Applied Econometrics 13, 645–657.

Battese, G.E., Cora, G.S., 1977. Estimation of a production frontier model: with application to the pastoral zone of eastern Australia. Australian Journal of Agricultural Economics 21, 169–179.

Battese, G.E., Coelli, T.J., 1992. Frontier production functions, technical efficiency and panel data with application to paddy farmers in India. Journal of Productivity Analysis 3, 153–169.

Berg, S.A., Kim, M., 1998. Banks as multioutput oligopolies: an empirical evaluation of the retail and corporate banking markets. Journal of Money, Credit, and Banking 30, 135–153.

Berger, A.N., Hannan, T.H., 1998. The efficiency cost of market power in the banking industry: a test of the quiet life and related hypotheses. Review of Economics and Statistics 454–465.

Borenstein, S., Shepard, A., 1996. Dynamic pricing in retail gasoline markets. The RAND Journal of Economics 27, 429–451.

Brander, J.A., Zhang, A., 1993. Dynamic oligopoly in the airline industry. International Journal of Industrial Organization 11, 407–435.

Bresnahan, T.F., 1989. Studies of industries with market power. In: Schmalensee, Richard, Willig, Robert D. (Eds.), The Handbook of Industrial Organization. North-Holland, Amsterdam.

Captain, P., Sickles, R.C., 1997. Competition and efficiency in the European airline industry: 1976–1990. Managerial and Decision Economics 18, 209–225.

Caves, D.W., Christensen, L.R., Trethway, M.W., 1983. Productivity performance of the US trunk and local service airlines in the era of deregulation. Economic Inquiry 21, 312–324.

Comanor, W.S., Leibenstein, H., 1969. Allocative efficiency, X-efficiency and the measurement of welfare losses. Economica 36, 304–309.

Cornwell, C., Schmidt, P., Sickles, R.C., 1990. Production frontiers with time-series variation in efficiency levels. Journal of Econometrics 46, 185–200.

Corts, K.S., 1999. Conduct parameters and the measurement of market power. Journal of Econometrics 88, 227–250.

Delis, M.D., Tsionas, E.G., 2009. The joint estimation of bank-level market power and efficiency. Journal of Banking & Finance 33, 1842–1850.

Demsetz, H., 1973. Industry structure, market rivalry, and public policy. Journal of Law and Economics 16, 1–9.

Durbin, J., Koopman, S.J., 2001. Time series analysis by state space methods. In: Oxford Statistical Series 24. Oxford University Press.

Encaoua, D., Jacquemin, A., 1980. Degree of monopoly, indices of concentration and threat of entry. International Economic Review 87–105.

Gallet, A.G., Schroeter, J.R., 1995. The effects of the business cycle on oligopoly coordination: evidence from the US rayon industry. Review of Industrial Organization 181–196.

Genesove, D., Mullin, W., 1998. Testing static oligopoly models: conduct and cost in the sugar industry, 1890–1914. The RAND Journal of Economics 29, 355–377.

Golub, G., Van Loan, C., 1996. Matrix Computations, third ed. The Johns Hopkins University Press, London.

Good, D., Sickles, R.C., Weiher, J., 2008. A hedonic price index for airline travel. Review of Income and Wealth 54, 438–465.

Harvey, A.C., 1989. Forecasting Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.

Hicks, J.R., 1935. Annual survey of economic theory: the theory of monopoly. Econometrica 3, 1–20.

Jin, H., Jorgenson, D.W., 2010. Econometric modeling of technical change. Journal of Econometrics 157, 205–219.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. Journal of Econometrics 19, 233–238.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, Transactions ASMA, Series D 82, 35–45.

Kim, C.J., 2006. Time-varying parameter models with endogenous regressors. Economics Letters 91, 21–26.

Kim, C.J., Nelson, C.R., 2006. Estimation of a forward-looking monetary policy rule: a time-varying parameter model using ex post data. Journal of Monetary Economics 53, 1949–1966.

Kim, D., 2005, Measuring market power in a dynamic oligopoly model: an empirical analysis, Working Paper.

Kim, Y., Kim, C.J., 2007, Dealing with endogeneity in a time-varying-parameter model: joint estimation and two-step estimation procedures, Working Paper.

Koetter, M., Kolari, J., Spierdijk, L., 2008, Efficient Competition? Testing the quiet life of us banks with adjusted Lerner indices, In: Proceedings 44th Bank Structure and Competition Conference, Federal Reserve Bank of Chicago.

Koetter, M., Poghosyan, T., 2009. The identification of technology regimes in banking: implications for the market power-fragility nexus. Journal of Banking and Finance 33, 1413–1422.

Koetter, M., Vins, O., 2008, The quiet life hypothesis in banking — evidence from German savings banks, Working Paper Series: Finance and Accounting 190, Department of Finance, Goethe University Frankfurt am Main.

Koopman, S.J., Harvey, A.C., Doornik, J.A., Shephard, N., 2007. STAMP: Structural Time Series Analyser, Modeller and Predictor. Timberlake Consultants Press, London.

Kumbhakar, S.C., 1990. Production frontiers, panel data, and time-varying technical inefficiency. Journal of Econometrics 46, 201–211.

Kumbhakar, S.C., 1997. Modeling allocative inefficiency in a translog cost function and cost share equations: an exact relationship. Journal of Econometrics 76, 351–356.

Kumbhakar, S.C., Lovell, C.A.K., 2000. Stochastic Frontier Analysis. Cambridge University Press.

Kutlu, L., 2010. Battese–Coelli estimator with endogenous regressors. Economics Letters 109, 79–81.

Lerner, A.P., 1934. The concept of monopoly and measurement of monopoly power. Review of Economic Studies 1, 157–175.

Maudos, J., Fernández de Guevara, J., 2007. The cost of market power in banking: social welfare loss vs. cost efficiency. Journal of Banking and Finance 31, 2103–2125.

McKiernan, B., 1996. Consumption and the credit market. Economics Letters 51, 83–88.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb–Douglas production functions with composed error. International Economic Review 18 (2), 435–444.

Pagan, A., 1984. Econometric issues in the analysis of regressions with generated regressors. International Economic Review 25, 221–247.

Parish, R., Ng, Y.K., 1972. Monopoly, X-efficiency and the measurement of welfare loss. Economica 39, 301–308.

Peersman, G., Pozzi, L., 2004, Determinants of consumption smoothing, Working Paper.

Perloff, J.M., Karp, L.S., Golan, A., 2007. Estimating Market Power and Strategies. Cambridge University Press, Cambridge.

Perloff, J., Sickles, R.C., Weiher, J., 2003. In: Slottje, D. (Ed.), An Analysis of Market Power in the US Airline Industry, With J. Perloff and J. Weiher, in Measuring Market Power. North-Holland, Amsterdam, pp. 309–323.

Pindyck, R.S., 1985. The measurement of monopoly power in dynamic markets. Journal of Law and Economics 28, 193–222.

Pitt, M.M., Lee, L.F., 1981. The measurement and sources of technical inefficiency in Indonesian weaving industry. Journal of Development Economics 9, 43–64.

Puller, S.L., 2007. Pricing and firm conduct in California's deregulated electricity market. The Review of Economics and Statistics 75–87.

Puller, S.L., 2009. Estimation of competitive conduct when firms are efficiently colluding: addressing the Corts critique. Applied Economics Letters 1497–1500.

Röller, L.H., Sickles, R.C., 2000. Capacity and product market competition: measuring market power in a puppy–dog industry. International Journal of Industrial Organization 18, 845–865.

Rotemberg, J.J., Saloner, G., 1986. A supergame-theoretic model of price wars during booms. American Economic Review 76, 390–407.

Schmidt, P., Sickles, R.C., 1984. Production frontiers and panel data. Journal of Business & Economic Statistics 2, 367–374.

Sickles, R.C., 2005. Panel estimators and the identification of firm-specific efficiency levels in semi-parametric and non-parametric settings. Journal of Econometrics 126, 305–324.

Sickles, R.C., Captain, P., Good, D.H., Ayyar, A., 2007. What if the European Airline Industry had deregulated in 1979?: A counterfactual dynamic simulation, In: Lee, Darin (Ed.), The Economics of Airline Institutions, Operations and Marketing, vol. 2, Amsterdam: Elsevier, North Holland (Chapter 5)..

Weiher, J.C., 2002, Pricing issues in the United States airline industry, Ph.D. Thesis, Rice University.

Wingrove, E.R.III, Johnson, J.P., Sickles, R.C., Good, D.H., 1997, The ASAC air carrier investment model (second generation), National Aeronautics and Space Administration, Langley Research Center, National Technical Information Service.