AgBioForum, 12(3&4): 422-436. ©2009 AgBioForum.

Efficiency Analysis of the US Biotechnology Industry: Clustering Enhances Productivity

Man-Keun Kim and Thomas R. Harris
University of Nevada-Reno

Slavica Vusovic
Department of Motor Vehicles, State of Nevada

This article attempts to identify the factors affecting location of biotechnology firms in the United States. To achieve this goal, the regional efficiency in the biotech industry is measured using the Data Envelopment Analysis (DEA). We investigate the causal structure of the regional biotech industry performance and a set of other locational variables using the Directed Acyclic Graph (DAG). Clustering the biotech industry directly enhances the regional industry efficiency, and the high-tech infrastructure and regional income directly affect the clustering the biotech industry.

***Key words*:** biotechnology, causality, cluster, data envelope analysis, directed acyclic graph.

Introduction

The biotechnology (hereafter biotech) industry[1] is one of the fastest-growing industries in the United States during the past decades (Cortright & Mayer, 2002; Ernst & Young, 2007). The biotech industry realized approximately $105 billion in sales in 1997 (US Census Bureau, 2001) and more than $200 billion in sales in 2002 (US Census Bureau, 2005). It employed approximately 300,000 persons in 1997 (US Census Bureau, 2001) and 500,000 workers in 2002 (US Census Bureau, 2005). Many researchers and analysts regard the biotech industry as "a key driver of modern economic progress" (Battelle Technology Partnership Practice [BTPP], 2008, pp. 19) and as an economic area in which the United States maintains a comparative advantage over other nations. BTPP (2008) reported that most state economic development authorities have the location and attraction of biotech firms as a primary strategic objective.

The biotech industry is mainly concentrated in nine regions,[2] so-called biotech clusters, which account for 75% of all biotech firms, more than 60% of all National Institute of Health (NIH) research spending, and 67% of all biotech-related patents (Cortright & Mayer, 2002). Given the locational concentration of biotech industry, it is natural for researchers to examine the locational factors of biotech firms and identify the factors influencing those locational choices.

Past studies have attempted to identify the locational factors for biotech firms, which are for example, Goetz and Morgan (1995), Gray and Parker (1998), Zucker, Darby, and Brewer (1998), Hall and Bagchi-Sen (2001), Cortright and Mayer (2002), Goetz and Rupasingha (2002), Zucker, Darby, and Armstrong (2002), Sambidi and Harrison (2006), and Vusovic (2006). These studies have tried to answer questions as to what factors affect the location and performance of biotech firms (or high-technology firms in some studies). Similar conclusions are drawn from the studies. The location of biotech firms is attributed to venture capital (R&D funds), proximity to research universities and metropolitan areas, educated populations, and regional income and/or properly-developed (high-technology) infrastructure. A brief summary of these works is presented in the following section.

1. *The definition of biotech or the biotech industry might not be unique. It can be defined as "the application of biological knowledge and techniques pertaining to molecular, cellular, and genetic processes to develop products and services" (Cortright & Mayer, 2002) or as "any technique that uses living organisms to make/modify products, improve plants or animals, or develop microorganisms for a specific use" (Goetz & Morgan, 1995). For more definitions by industry or academia, see Cortright and Mayer (2002) Tables A1 and A2. In this article we follow the definition suggested by Cortright and Mayer (2002). The biotech industry consists of medical and botanical manufacturing (NAICS 325411), pharmaceutical preparation manufacturing (NAICS 325412), in-vitro diagnostic substance manufacturing (NAICS 325413), biological products except diagnostic (NAICS 325414), and research and development in physics, engineering, and life science (NAICS 541710).*

2. *Boston-Worcester-Lawrence (MA-NH-ME-CT), San Francisco-Oakland-San Jose (CA), San Diego (CA), Raleigh-Durham-Chapel Hill (NC), Seattle-Tacoma-Bremerton (WA), New York-Northern New Jersey-Long Island (NY-NJ-CT-PA), Philadelphia-Wilmington, Atlantic City (PA-NJ-DE-MD), Los Angeles-Riverside-Orange County (CA), and Washington DC-Baltimore (DC-MD-VA-WV) (Cortright & Mayer, 2002, Table 4, pp. 11).*

However, these studies might overlook two factors. First, they tried to identify the factors affecting agglomeration of the biotech industry, but they did not answer the question as to *why* biotech firms cluster. Porter (1998) hypothesized that clustering similar or related firms can increase firms' market competition by enhancing their productivity. The questions become "does clustering enhance productivity? Can we prove this empirically?" This article will explore this question and answer why biotech firms cluster. Second, previous research seem to assume *implicitly* that the locational variables (or environmental variables), such as universities and characteristics of population, affect agglomeration of biotech firms. This article will reveal the structure of the set of locational factors and identify direct and indirect factors that influence the location of biotech firms using a rigorous test. In other words, the *causal* flows among variables are of interest.

To answer the above two questions, productivity in the US biotech industry should be measured. Second, locational variables are collected and a causal structure among these variables—including efficiency—is developed. Various approaches are possible in assessing productivity in the biotech industry. One approach is the Data Envelopment Analysis (DEA), which was proposed by Charnes, Cooper, and Rhodes (1978; 1981). DEA is an optimization-based technique for evaluating the relative efficiency of decision-making units (DMUs). It has been widely applied in performance evaluation and benchmarking of schools, hospitals, bank branches, production plants, etc. Gattoufi, Oral, and Resiman (2004) have a comprehensive bibliography of DEA studies. It is expected for this article that states with biotech clusters (e.g., California, Massachusetts, or North Carolina) have higher DEA efficiency scores. Once the state level of efficiency is measured, the Directed Acyclic Graphs (DAG) procedures can be employed to test causality among efficiency, cluster, and other locational variables. The causal flow should be directed to the efficiency from the cluster with positive correlation, which means the cluster enhances the industry efficiency. DAG is an illustration using arrows and variables to represent the causal flow among a set of variables (Pearl, 1995, 2000; Spirtes, Glymour, & Scheines, 2000). The relationship among other locational variables can be also identified.

The goals of this article are two-fold. First, state levels of efficiency in the biotech industry is assessed using DEA. Second, the causal relationship among associated variables will be investigated. This article is divided into four parts to achieve these goals. First, a brief summary for the previous studies is presented. Second, a discussion on the methodologies of DEA and DAG follows. Third, we present empirical results, and finally, conclusions and policy implications from the DEA and DAG analysis are presented.

Literature Review

Goetz and Morgan (1995) developed a formal model to identify factors affecting the location of biotech firms. They found that venture capital and favorable fiscal policies, R&D spending, and raising the number of doctoral degrees awarded were important factors in establishing biotech firms. Goetz (1997) found that high-technology firms, in general, avoid counties with predominantly rural populations, not adjacent to a metropolitan area, and with a low portion of college graduates. Overall, Goetz (1997) concluded that rural areas have limited chances of attracting high-technology firms. Gray and Parker (1998) examined the location and organization of biotech firms based on product life-cycle theory. They insisted that manufacturing capabilities and marketing channels of more established companies in mature regions are major sources of a competitive advantage. Zucker, Darby, and Brewer (1998) added that the growth and diffusion of intellectual human capital (e.g., star scientists) is the main determinant of where and when the biotech industry is developed and localized.

Hall and Bagchi-Sen (2001) examined the relationship between R&D intensity, innovation, and performance in the US biotech industry using surveys and interviews. The purpose of their paper was not limited to locational choice of the biotech firm, but they uncovered collaboration, specifically with university scientists, was important for successful biotech firms. Goetz and Rupasingha (2002) provides similar results, which indicated that the availability of existing high-technology firms, number of college graduates, local property taxes, population, total county income, highway access, and county-amenity scale have a positive and significant impact on the location of high-technology firms.

Another study, by Cortright and Mayer (2002), reported on biotech research and commercialization activities in the 51 largest US metropolitan areas. As mentioned above, only nine of them have significant biotech activity. Cortright and Mayer (2002) insisted that the most important factors responsible for success and biotech development in certain regions have strong research capacity (universities, NIH funding, and patents) and the ability to convert research into successful commercial activity (e.g., venture capital). Munroe,

Craft, and Hutton (2002) provided similar answers using a survey of biotech companies in California. Their survey showed that proximity to leading research centers and venture capital is the main reason for the firm's location. Sambidi and Harrison (2006) tested the hypothesis of spatial agglomeration economies in the biotech industry and confirmed it using spatial econometrics. Vusovic (2006) included the demand factors of location, such as regional income, family composition, and population growth and showed that these factors were also important for the locational choice of biotech firms.

Methodologies

Data Envelopment Analysis (DEA)

For this article, we employed two analytical procedures: Data Envelopment Analysis (DEA; procedure that estimates efficiency of biotech firms) and Directed Acyclic Graph (DAG; procedure that reveals the causal relationships of biotech clusters).

DEA was developed in the management science tradition, focusing on computing the relative efficiency of different decision making-units (DMUs), such as firms, hospitals, or counties. The DEA approach can recognize the performance of industry between regions such as states or counties. Kim and Harris (2008) and Raab and Lichty (1997, 2002) used DEA to compare the efficiencies between Nevada and Utah counties and Minnesota counties, respectively. In addition, Raab and Kotamraju (2006) utilized DEA to measure the efficiency of the high-technology economy from the state level.

To define DEA efficiency estimates, the following notation is established; let $\mathbf{x}_j \in \Re_+^p$ denote a vector of p inputs and $\mathbf{y}_j \in \Re_+^q$ denote a vector of q outputs for DMU j, where j=1,…, n. The production possibility set is defined by P={($\mathbf{x}$, $\mathbf{y}$)| inputs $\mathbf{x}$ can produce outputs $\mathbf{y}$}. The boundary of P is referred to as the production frontier. Technically inefficient DMUs operate at points that are inferior to the production frontier, while technically efficient DMUs operate somewhere along the frontier. Define an efficiency measure θ for DMU j, $(\mathbf{x}_j, \mathbf{y}_j) \in \Re_+^{p+q}$ such that

$$\theta_j \equiv \sup\{\theta \mid (\mathbf{x}_j, \theta\mathbf{y}_j) \in P, \theta > 0\}, \tag{1}$$

where sup is the supremum. This is the Farrell (1957) measure of output technical efficiency, which is the reciprocal of the Shephard (1970) output distance function. The DEA estimator θ defined in Equation 1 at a specific point (DMU j) can be written in terms of the linear programming (LP) model, which was initially proposed by Charnes et al. (1978, 1981) and extended by Banker, Charnes, and Cooper (1984).

$$\hat{\theta}_j = \max\{\theta > 0 \mid \theta\mathbf{y}_j \le \mathbf{Y}\boldsymbol{\lambda}, \mathbf{x}_j \ge \mathbf{X}\boldsymbol{\lambda}, \boldsymbol{\lambda} \in \Re_+^n\}, \tag{2}$$

where $\mathbf{Y}$=[$\mathbf{y}_1$, $\mathbf{y}_2$, …, $\mathbf{y}_n$], $\mathbf{X}$=[$\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$], and $\boldsymbol{\lambda}$ is $n \times 1$ intensity variables.

It is noteworthy that the DEA formulation differs slightly along with the assumption of returns to scale. Under the constant returns to scale (CRS), the LP formulation is given by Equation 2 and is called the CCR model (which stands for the authors, Charnes, Cooper, and Rhodes; Charnes et al., 1978). The DEA estimator under the assumption of variable returns to scale (VRS) is found by solving the same LP problem in Equation 2 with additional constraint, $\mathbf{i}'\boldsymbol{\lambda}$=1, where $\mathbf{i}$ denotes an $n \times 1$ vector of ones. This model is called the BCC model (which stands for the authors, Banker, Charnes, and Cooper; Banker et al., 1984). The BCC model adds the additional constraint to impose a convexity condition on allowable ways in which the observations for the n DMUs may be combined (Cooper, Seiford, & Tone, 2007). When the above constraint is replaced by $\mathbf{i}'\boldsymbol{\lambda} \le 1$, the production set exhibits non-increasing returns to scale (NIRS).

The LP models in Equation 2, along with additional constraints, are run n times to identify relative efficiencies of all the DMUs. DEA efficiency models derive estimates of efficiency that are equal to or less than 1. DMU is said to be efficient if its calculated DEA estimate is equal to 1. The DEA estimate of less than 1 implies inefficiency. Also, $\theta_j^{crs} \le \theta_j^{nirs} \le \theta_j^{vrs}$ by construction. This is easily shown by looking at Figure 1. Figure 1 shows a typical production possibility set in two dimensions for the single input and single output case. Panel A in Figure 1 has the production frontiers under CRS assumption and Panel B has that of under the VRS. Under the CRS assumption, DMU C is efficient and other DMUs would be inefficient (Panel A). However, under the VRS assumption, DMUs A, C, and F are efficient and DMUs B, D, and E are inefficient (Panel B). Obviously, the DEA estimator under VRS assumption is larger than those under CRS assumption.

The existence of increasing or variable returns to scale is of importance. Simar and Wilson (2002) proposed the bootstrap procedure to test returns to scale. The statistical test for the returns to scale begins with

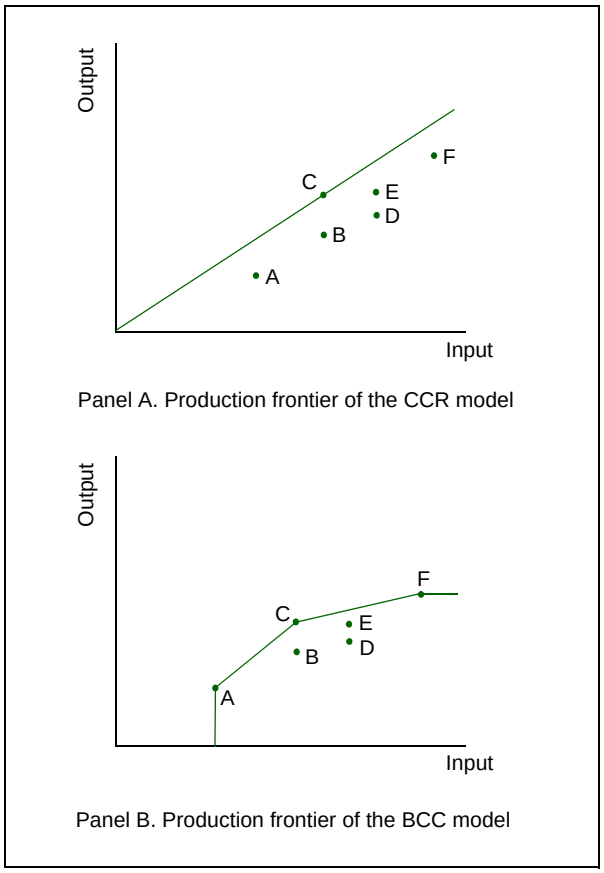

Figure 1. Production frontiers and efficiency measure.

CRS; that is, efficiency scores under CRS ($\hat{\theta}_j^{crs}$) are estimated first using Equation 2. Next, efficiency scores under VRS ($\hat{\theta}_j^{vrs}$) are estimated with adding the additional restriction, $\mathbf{i'\lambda}$=1, to Equation 2. The null hypothesis is that the production set exhibits CRS and the alternative hypothesis is that it shows VRS. Various test statistics are possible; however, the mean of ratios $\hat{\theta}_j^{crs} / \hat{\theta}_j^{vrs}$, that is $t_{crs} = n^{-1} \sum_{j=1}^{n} \hat{\theta}_j^{crs} / \hat{\theta}_j^{vrs}$, will be used as in Simar and Wilson (2002). By construction, $t_{crs} \leq 1$ because $\hat{\theta}_j^{crs} \leq \hat{\theta}_j^{vrs}$. The null hypothesis is rejected when t_{crs} is significantly less than 1. The critical value for deciding if the test statistic is significantly less than 1 can be derived from bootstrapping (Simar & Wilson, 2002). For more information concerning bootstrapping, refer to Simar and Wilson (1998, 2000). When the null hypothesis of CRS is rejected, another test is preformed with a less restrictive NIRS versus VRS. Similarly the

$$x_2 \rightarrow x_3 \rightarrow x_4$$
$$\uparrow$$
$$x_1$$

Figure 2. Example of directed graph and causal relationship.

test statistic is $t_{vrs} = n^{-1} \sum_{j=1}^{n} \hat{\theta}_j^{crs} / \hat{\theta}_j^{vrs}$ and decision is made based on the critical value from bootstrapping again. If t_{vrs} is significantly less than 1, we reject the null hypothesis.

Related to further statistical analysis with DEA efficiency estimates, e.g., regression or causal relationship investigation, Simar and Wilson (2007) insisted that the statistical analyses may not be consistent unless the DEA estimates are corrected. They showed the inconsistency using the Monde Carlo experiment, especially the second-stage regression. According to Simar and Wilson (2007), this inconsistency exists because the DEA estimates are complicated, serially correlated, and biased downward by construction. Simar and Wilson (2007) proposed bootstrap procedures to improve statistical properties of DEA estimates such that $\hat{\theta}_j = \hat{\theta}_j - bias(\hat{\theta}_j)$. The bias term is constructed using bootstrapping. The empirical DEA estimates and bias corrected DEA estimates are reported in the following section.

Directed Acyclic Graph (DAG)

The DAG approach attempts to identify the causal relationship among a set of observational or non-experimental data. The DAG is a picture representing causal flow using arrows among a set of variables (Pearl, 1995, 2000; Spirtes et al., 2000) based on a conditional independence relationship as given by the recursive decomposition

$$\Pr(x_1, x_2, x_3, ..., x_n) = \prod_{i=1}^{n} \Pr(x_i \mid pa_i), \quad (3)$$

where Pr(.) is the joint probability of variables x_1, x_2, x_3,…, x_n and pa_i are parents (direct causes) of x_i, the minimal set of x_i's predecessors (the variables that come before in a causal sense) that renders x_i independent of all its other predecessors (Pearl, 2000). Geiger, Verma,

Table 1. Descriptive statistics on output variables in US biotechnology industry.

	1997			2002		
	y_1	y_2	y_3	y_1	y_2	y_3
	Value of shipments[a]	Receipts[b]	Number of patent[c]	Value of shipments[a]	Receipts[b]	Number of patent[c]
	$ in million	$ in million		$ in million	$ in million	
Average	2,203	985	1,777	3,303	1,381	2,004
St. dev	3,150	1,467	2,204	5,372	2,456	3,164
Max	PA – 11,273	CA – 7,346	CA –11,289	PA – 21,966	CA – 14,225	CA –18,829
Min	0	IA – 23	DC – 52	0	WY – 23	AK – 43

[a] *from Pharmaceutical & Medicine Manufacturing sector from US Census Bureau (2001, 2005)*
[b] *from Research & Development in Life Science from US Census Bureau (2001, 2005)*
[c] *Source: US Patent and Trademark Office (2008)*

and Pearl (1990) have shown that there is a one-to-one correspondence between the set of conditional independencies among variables implied by Equation 3 and the graphical expression of variables in a DAG. For example, consider four variables, x_1, x_2, x_3, and x_4. If there is causal relationship such as x_1, x_2 cause x_3 and x_3 causes x_4, then the directed graphs that represent this causal relationship is represented in Figure 2. This directed graph is expressed as the probability distribution product by

$$\Pr(x_1, x_2, x_3, x_4)=\Pr(x_1)\ \Pr(x_2)\Pr(x_3 \mid x_1, x_2)\Pr(x_4 \mid x_3). \quad (4)$$

A Greedy Equivalence Search algorithm (GES) suggested by Meek (1997) and discussed by Chickering (2002, 2003) is used for identifying the causal flow among the variables. The GES algorithm starts from a graphical representation with no edges, which implies that all variables are independent, and it proceeds stepwise searching over causal flow based on Equation 2 using the Bayesian scoring criterion. After score comparison among all possible equivalence classes,[3] the equivalence class with the maximum score is chosen for the next step. Once a local maximum is attained in the first phase, the second phase proceeds by single-edge deletions and compares the scores of DAG in equivalence classes repeatedly until a local maximum is again reached. When the algorithm reaches a local maximum, it obtains the optimal solution and DAG (Chickering, 2003). The GES algorithm and more refined extensions are marketed as the software project TETRAD (http://www.phil.cmu.edu/projects/tetrad/index.html). The DAG analysis for DEA efficiency scores and environmental variables are reported in the following section.

Empirical Results of DEA and DAG

Data and Preliminary Analysis

This section will discuss data and procedures. Using DEA, the performance of the US biotech industry is investigated. For the analysis, US Census data are utilized. Twenty-nine states for 1997 and 40 states for 2002 are selected according to data availability.[4] The years 1997 and 2002 are analyzed, which allows the capture of the variation of efficiencies over time. Charnes, Clark, Cooper, and Golany (1984) proposed a technique called "window analysis" in DEA. Window analysis assesses the performance of states over time by treating them as different entities in each time period. This is also beneficial because the larger data set facilitates further statistical investigations.

To use DEA, inputs and outputs should be defined. The value of shipments (y_1) from the Pharmaceutical and Medicine Manufacturing Sector (NAICS 32541) and the receipts (y_2) from the Research and Development in Life Science Sector (NAICS 54171) are the first type of outputs used for the analysis. The value of shipments includes net selling values of all products shipped, both primary and secondary (US Census Bureau, 2005), and the receipts includes gross receipts from customers or clients for services provided, from the use of facilities and from merchandise sold (US Census Bureau, 2005). Both outputs are collected from the US Census Bureau (2001, 2005). The number of patents (y_3) represents a second type of output activity, which is collected from the US Patent and Trademark Office (2008). Note that the number of patents is considered as the primary output, although it might be interpreted as an intermediate input rather than final output.

Table 1 offers descriptive statistics on each output variable for 1997 and 2002. In addition to this, Figures

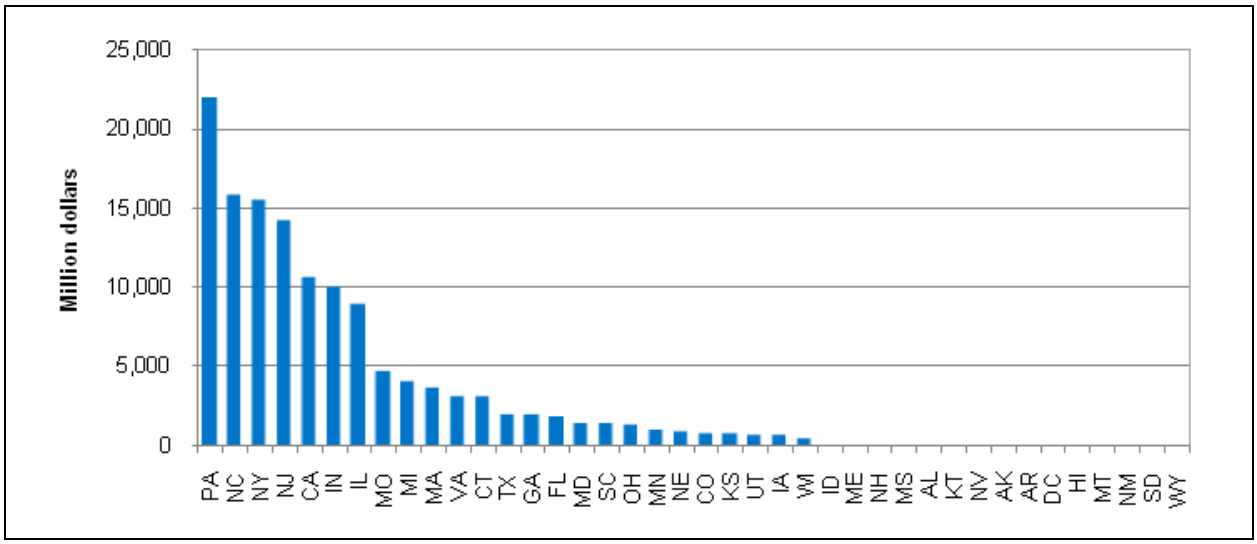

Figure 3. Output of US biotech industry: Value of total shipments in pharmaceutical and medicine manufacturing in 2002.

3, 4, and 5 show the distributions of each output for 2002 (from largest to smallest). As shown in Table 1 and Figures 3, 4, and 5, Pennsylvania, North Carolina, New York, New Jersey, California, and Indiana are the large pharmaceutical and medicine manufacturing regions (Figure 3). California, Virginia, Massachusetts, New York, Maryland, and Texas are main states where R&D in the life science sector is located (Figure 4). California has close to 20,000 patents in 2002, which makes California dominant in the number of patents (Figure 5). Other regions such as New York and Texas have more than 6,000 patents in 2002 (Figure 5).

Primary inputs are the number of paid employees (x_3 and x_4) from both sectors from US Census (US Census Bureau, 2001, 2005), and the venture capital (x_1), and R&D funds (x_2) from the various sources.[5] Paid employees consist of full- and part-time employees, including salaried officers and executives who were on

3. *Two directed acyclic graphs (DAGs) are in the same equivalence class when they are equivalent. Both DAGs are equivalent when both are distributionally equivalent and independence equivalent (Chickering, 2003). Distributionally equivalent DAGs have the same Bayesian networks (that is, Equation 2). Consider the following DAGs.*

DAG_1	DAG_2
$x_1 \rightarrow x_2$	$x_1 \leftarrow x_2$
$\downarrow$	$\downarrow$
x_3	x_3

Based on Equation 2, we can construct the joint probability distributions such that
DAG_1: $Pr(x_1,x_2,x_3)=Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_1)$, and
DAG_2: $Pr(x_1,x_2,x_3)=Pr(x_2)Pr(x_1|x_2)Pr(x_3|x_1)$.

Because $Pr(x_2|x_1)=Pr(x_2 \cap x_1)/Pr(x_1)$, DAG_1 can be rewritten as
$Pr(x_1,x_2,x_3)=Pr(x_1)Pr(x_2|x_1)Pr(x_3|x_1)=Pr(x_1 \cap x_2)\ Pr(x_3|x_1)$.

With the same logic, DAG_2 is
$Pr(x_1,x_2,x_3)=Pr(x_2)Pr(x_1|x_2)Pr(x_3|x_1)=Pr(x_1 \cap x_2)\ Pr(x_3|x_1)$.

As a result, both joint probability distributions are identical (distributionally equivalent).

Two DAGs are independence equivalent if the independence constraints are identical (Chickering, 2003). The independence constraint for DAG_1 is $x_2 \perp x_3|x_1$ (the symbol $\perp$ indicates independence and | denotes conditioning on). The independence constraint for DAG_2 is $x_2 \perp x_3|x_1$. Thus two DAGs have the same independent constraints, and in turn, two DAGs are equivalent and in an equivalent class. The GES algorithm computes the score of both DAGs, compares them and picks the DAG whose score is larger systematically.

4. *Data for some states from the US Census are not available to avoid disclosing data of individual companies (US Census Bureau, 2001; 2005).*
5. *The venture capital is collected from Price Waterhouse Cooper/Thomson Venture Economics/National Venture Capital Association Moneytree Survey (available at http://www.pwcmoneytree.com/moneytree/nav.jsp?page=historical). The R&D fund is an aggregation of the following sets: (1) university R&D funds from National Science Foundation Survey of R&D expenditure at universities and colleges (http://caspar.nsf.gov), (2) federal R&D funds from National Science Foundation, (3) SBIR awards, and (4) STTR awards compiled from Small Business Administration (http://www.sba.gov).*

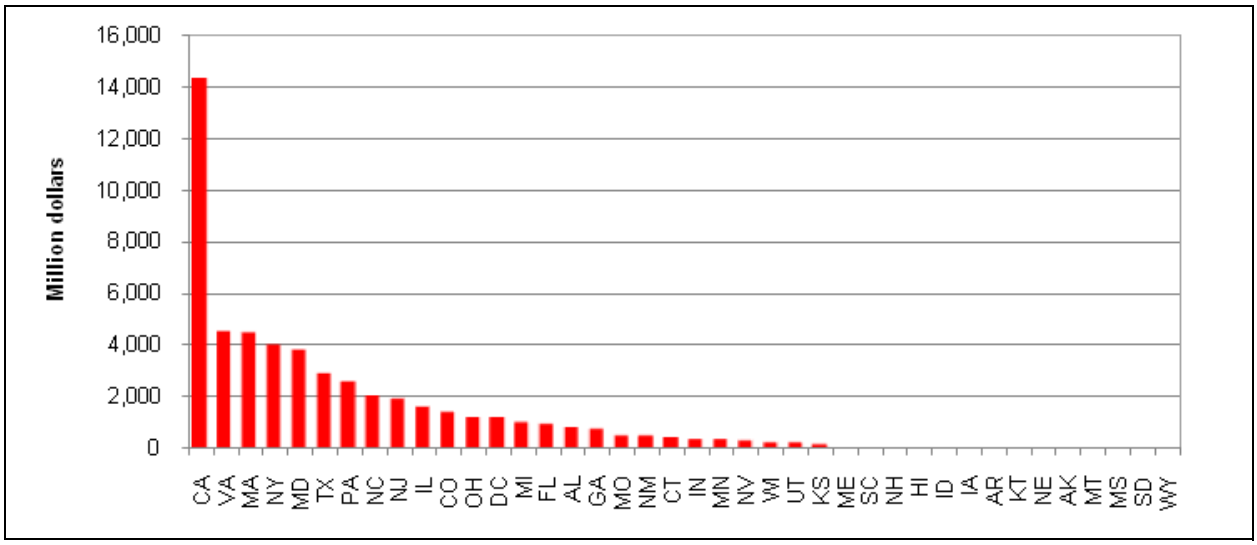

Figure 4. Output of US biotech industry: Receipt in research and development in life science in 2002.

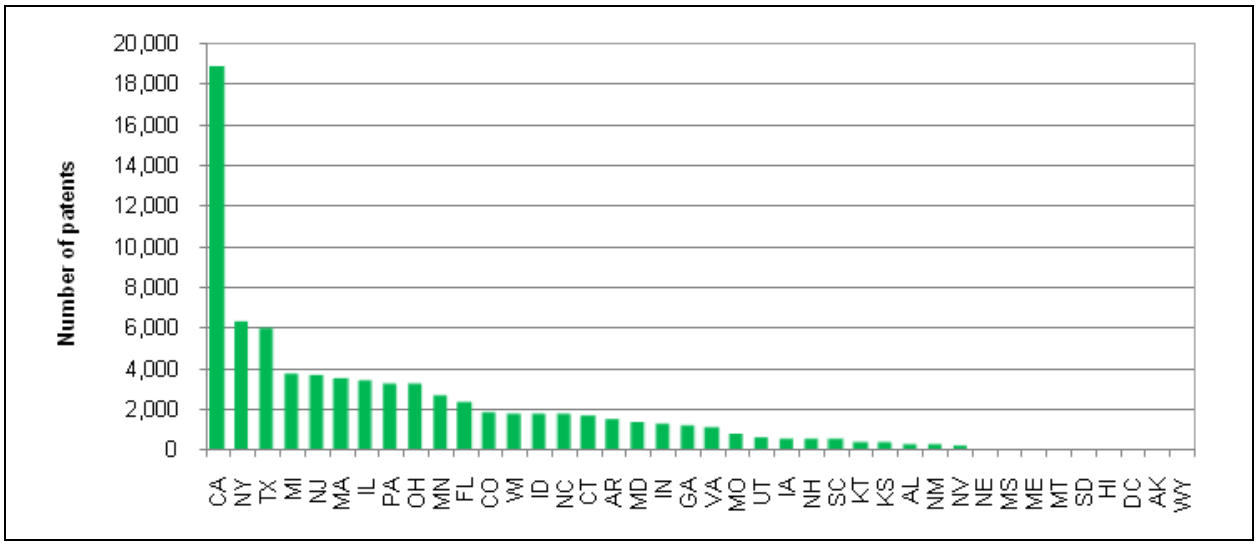

Figure 5. Output of US biotech industry: Number of patent in 2002.

the payroll (US Census Bureau, 2005). The R&D fund is an aggregation of university R&D funds, federal R&D funds, SBIR awards, and STTR awards. The venture capital and R&D funds encourage the biotech industry output and production of patents. Raab and Kotamraju (2006) showed that there exist lagged or feed-back relationships with these variables in the case of high-technology industry. They believed that lagged or feedback effects may exist in the biotech industry, but the effects are ignored in this article. Table 2 offers descriptive statistics on each input variable and Figures 6, 7, 8, and 9 provide the rough distribution of each input. Most of the venture capital was invested in the states of California, Massachusetts, and New Jersey (Figure 6). R&D funds have a similar pattern (Figure 7). The number of employee has a similar pattern of output as we expected (Figures 8 and 9).

Once all the output and input data are collected, DEA efficiency for the biotech industry in each state for each year is estimated. FEAR software (Wilson, 2006) is used to estimate DEA efficiency; the software allows us to compute DEA estimates, implement the homogenous bootstrap algorithm described by Simar and Wilson (1998, 2000), and correct biased DEA estimates as in Simar and Wilson (2007). Table 3 contains DEA scores with VRS assumption. Testing returns to scale follows in the next section. After obtaining state-level efficiency scores for the US biotech industry, the causal

Table 2. Descriptive statistics on input variables in US biotechnology industry.

	1997				2002			
	x_1 Venture capital[a] $ in million	x_2 R&D funds[a] $ in million	x_3 No. of employee[b] Persons	x_4 No. of employee[c] Persons	x_1 Venture capital[a] $ in million	x_2 R&D funds[a] $ in million	x_3 No. of employee[b] Persons	x_4 No. of employee[c] Persons
Average	44	3	5,189	7,724	77	39	5,943	12,937
St. dev	117	3	6,705	10,955	221	64	9,239	18,888
Max	CA – 587	CA – 17	CA – 27,022	CA – 51,967	CA –1,270	CA – 332	CA – 40,009	CA – 99,048
Min	0	0	0	AZ – 208	0	0	0	WY – 175

[a] *See footnote #5*
[b] *from Pharmaceutical & Medicine Manufacturing sector from US Census Bureau (2001, 2005)*
[c] *from Research & Development in Life Science from US Census Bureau (2001, 2005)*

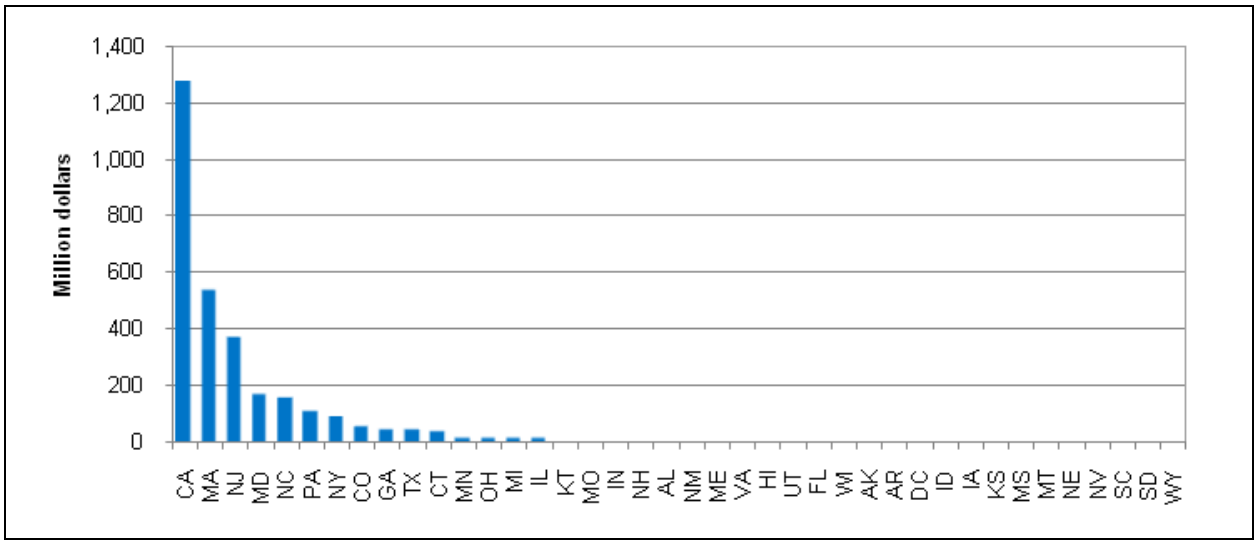

Figure 6. Input of US biotech industry: Venture capital in 2002.

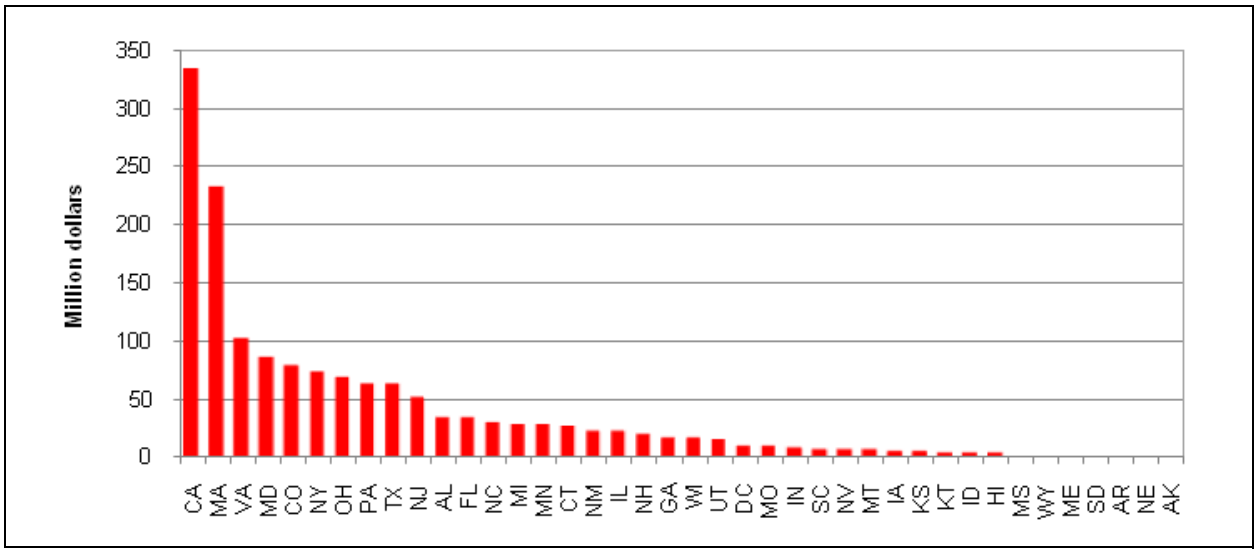

Figure 7. Input of US biotech industry: R&D funds in 2002.

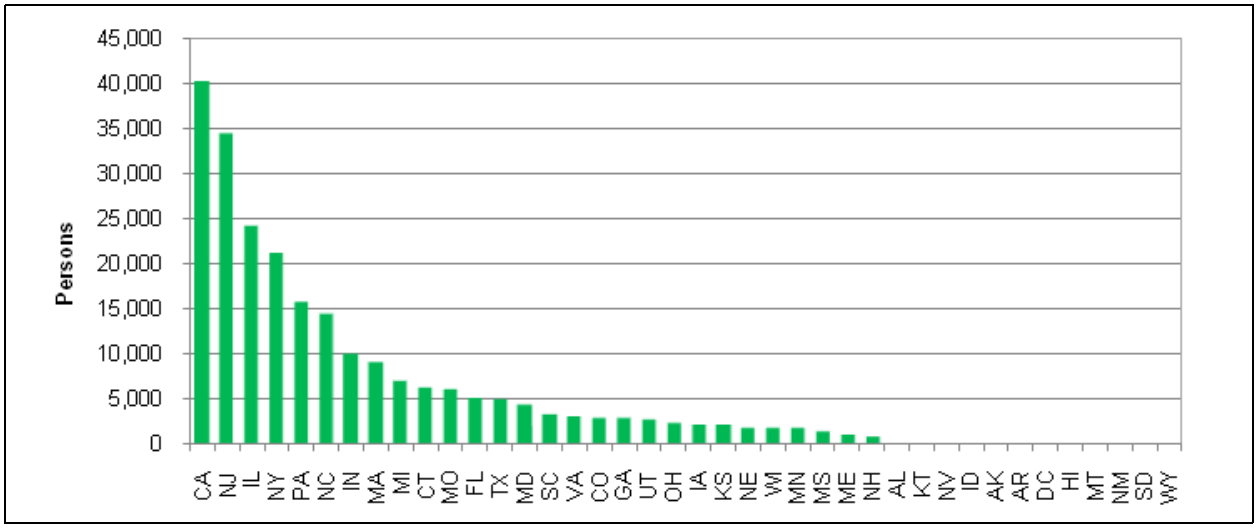

Figure 8. Input of US biotech industry: Number of employee in pharmaceutical and medicine manufacturing in 2002.

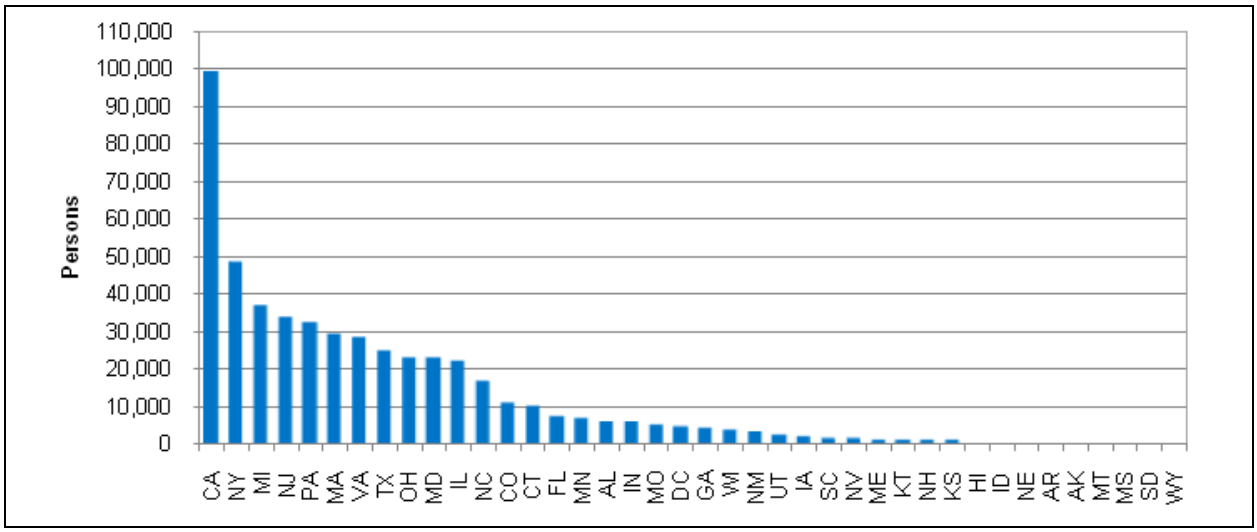

Figure 9. Input of US biotech industry: Number of employee in research and development in life science in year 2002.

structure among a set of variables is identified using DAG. Ten variables are selected for this analysis based on previous studies. These variables include (1) DEA efficiency scores, (2) cluster dummy, (3) regional income, (4) population, (5) top-ranked bioscience universities, (6) education level, (7) number of advanced degrees awarded in bioscience, (8) biotech firm size, (9) biotech average payroll, and (10) high-tech infrastructure.

The cluster dummy represents the existence of biotech clusters in the region. States are divided into three categories based on the number of biotech firms in the region. Table 3 contains information on each state. The cluster dummy has a value of 2 if the number of firms in the state is over 350 and 1 if the number of firms is between 260 and 350. Otherwise, the cluster dummy has a value of 0. The values are determined and assigned based on the slope of cumulative distribution function (CDF) with respect to the number of biotech firms. Figure 10 shows the CDF, and each segment has a different value from 0 to 2. Thirteen states have a value of 2 as a biotech cluster region, which are California, states along the eastern shore (Massachusetts, New York, New Jersey, Pennsylvania, Maryland, Virginia, North Carolina, and Florida), states in the Midwest (Illinois and Ohio), and states in the south (Texas and Colorado). This is consistent with the biotech cluster map in Cortright and Mayer (2002) and BTPP (2008).

Regional income is included as in the study by Vusovic (2006) from the Regional Economic Information

Table 3. DEA scores; Output orientation with VRS assumption.

	Number of firms (year 2002)	Efficiency rank (year 2002)	DEA estimates		Corrected DEA estimates	
			1997	2002	1997	2002
Alabama	171	29	-	0.669	-	0.606
Alaska	47	35	-	0.528	-	0.460
Arkansas	47	20	-	1.000	-	0.729
Arizona	164*		1.000	-	0.797	-
California	**3012**	**23**	**1.000**	**1.000**	**0.715**	**0.722**
Colorado	**417**	**17**	**0.926**	**0.824**	**0.846**	**0.763**
Connecticut	194	32	1.000	0.584	0.817	0.530
DC	130	22	1.000	1.000	0.708	0.724
Florida	**615**	**2**	**1.000**	**1.000**	**0.713**	**0.887**
Georgia	282	10	0.884	0.908	0.785	0.836
Hawaii	59	34	-	0.546	-	0.495
Idaho	50	12	-	1.000	-	0.805
Illinois	**408**	5	1.000	1.000	0.710	0.871
Indiana	166	21	-	1.000	-	0.725
Iowa	113	28	1.000	0.676	0.701	0.606
Kansas	103	4	-	1.000	-	0.882
Kentucky	85	40	-	0.320	-	0.290
Maine	71	38	-	0.448	-	0.403
Maryland	**632**	3	-	1.000	-	0.886
Massachusetts	**853**	**1**	1.000	0.980	0.874	0.897
Michigan	333	7	1.000	1.000	0.825	0.857
Minnesota	254	6	1.000	0.957	0.780	0.859
Mississippi	61	36	-	0.490	-	0.449
Missouri	223	27	1.000	0.786	0.729	0.692
Montana	60	37	1.000	0.477	0.762	0.428
Nebraska	57	11	-	1.000	-	0.811
Nevada	95	19	1.000	0.844	0.749	0.758
New Hampshire	90	31	1.000	0.581	0.821	0.538
New Jersey	**694**	**26**	**-**	**0.785**	**-**	**0.698**
New Mexico	166	30	1.000	0.647	0.720	0.565
New York	**841**	13	1.000	1.000	0.737	0.794
North Carolina	**449**	8	1.000	1.000	0.734	0.844
Ohio	**426**	9	1.000	0.945	0.796	0.843
Oregon	154*		1.000	-	0.849	-
Pennsylvania	**549**	**18**	**1.000**	**1.000**	**0.815**	**0.761**
Rhode Island	39*		0.574	-	0.515	-
South Carolina	104	14	-	1.000	-	0.788
South Dakota	25	39	-	0.366	-	0.329
Texas	**738**	**16**	**1.000**	**1.000**	**0.835**	**0.767**
Utah	180	33	0.746	0.579	0.687	0.530
Virginia	**557**	**24**	**1.000**	**1.000**	**0.740**	**0.717**
Washington	287*		1.000	-	0.832	-
Wisconsin	201	15	0.868	0.851	0.769	0.770
Wyoming	24	25	-	1.000	-	0.709

*Note: Numbers in bold indicates the states where biotech industry cluster (no. of firms ≥ 350); * 1997 number*

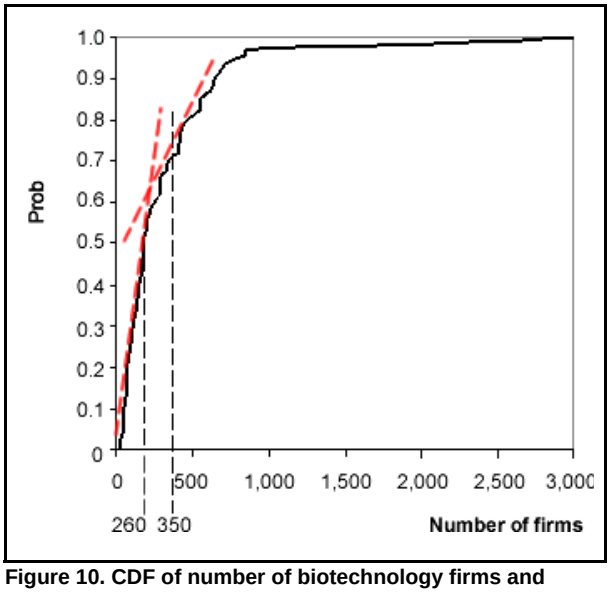

Figure 10. CDF of number of biotechnology firms and assigning dummy values.

System (REIS) in the Bureau of Economic Analysis (BEA, http://www.bea.gov/regional/). Vusovic (2006) used per-capita income as a locational variable; however, total regional income was used in this analysis to reflect the purchasing power of biotech products. State-level population from the REIS in the BEA is used as another variable to represent the market size. The number of top-ranked bioscience universities in each state is constructed based on a Milken Institute University biotech publication ranking from DeVol et al. (2006). These variables are expected to have a positive effect on industry performance directly or indirectly. Narin, Hamilton, and Olivastro (1997) suggested that science-technology linkages are stronger in life sciences than in other field, and in turn, Xia and Buccola (2005) showed that both bioscience research and graduate training in US universities are the sources of bioscience productivity. For example, New York has six, California has five, and Massachusetts has three top-ranked bioscience universities.

Both the state education level (the percentage of the population with bachelor's degrees or higher, regardless of major) and the number of advanced degrees awarded in the biosciences were collected from US Census data (US Census Bureau, 2001, 2005). Both education level and the number of advanced degrees awarded are the indication of the quality of the labor supply to the biotech industry in the region. Average payroll in the biotech industry and biotech firm size are calculated based on US Census data (US Census Bureau, 2001, 2005). These variables portray the incentive for job seekers toward the biotech industry. These variables are important because employees with higher degrees might migrate to other states and enhance other states' efficiency score. Sanderson and Dugoni (2002) suggested that doctorate recipients are very mobile. Raab and Kotamraju (2006) insisted that emigration of high-technology-degree employees can be the source of inefficiency in the high-technology industry.

The high-technology infrastructure index is built using the New State Economy Index (NSEI) score. NSEI ranks the structural transformation of the 50 states through five categories, including knowledge jobs, globalization, economic dynamism and competition, transformation into a digital economy, and technological innovation indicators (Atkinson, 2002; Atkinson & Court, 1998). The score is the weighted average of 21 indicators (various high-tech related elements), e.g., employment of IT professionals, globalization, economic dynamism and competition, and so on. In short, the NSEI score is interpreted as the high-technology infrastructure in the state. It is expected that the DEA efficiency score is high for the states that have high NSEI score. Massachusetts has the highest score (90 in 2002), while West Virginia has the lowest score (40.7 in 2002). For the entire ranking in 2002, refer to http://www.neweconomyindex.org/states/2002/overview.html. Note that the 1997 index is not available so 1998 scores are used for 1997 data.

Once all the variables are collected, the directed acyclic graph as specified by the GES algorithm in TETRAD IV (version 4.3.9-0) is applied.

Returns to Scale

Returns to scale and the impact of returns to scale are not thoroughly understood for the biotech industry. This article attempts to identify returns to scale in the US biotech industry. Returns to scale are tested using the DEA efficiency estimates and bootstrapping procedure as in Simar and Wilson (2002). As discussed earlier, the test begins with the null hypothesis of CRS. The test statistic is t_{crs}=0.873. If this statistic is significantly less than 1 (or critical value), the null hypothesis is rejected. From bootstrapping, the critical value at 5% significance level is 0.880; therefore, the null hypothesis is rejected. When the null hypothesis of CRS is rejected, another test is preformed with a less restrictive NIRS versus VRS. The test statistic is $t_{nirs} = n^{-1} \sum_{j=1}^{n} \hat{\theta}_j^{nirs} / \hat{\theta}_j^{vrs} = 0.988$ and 5% critical value is 0.989. Thus, we fail to the reject the null

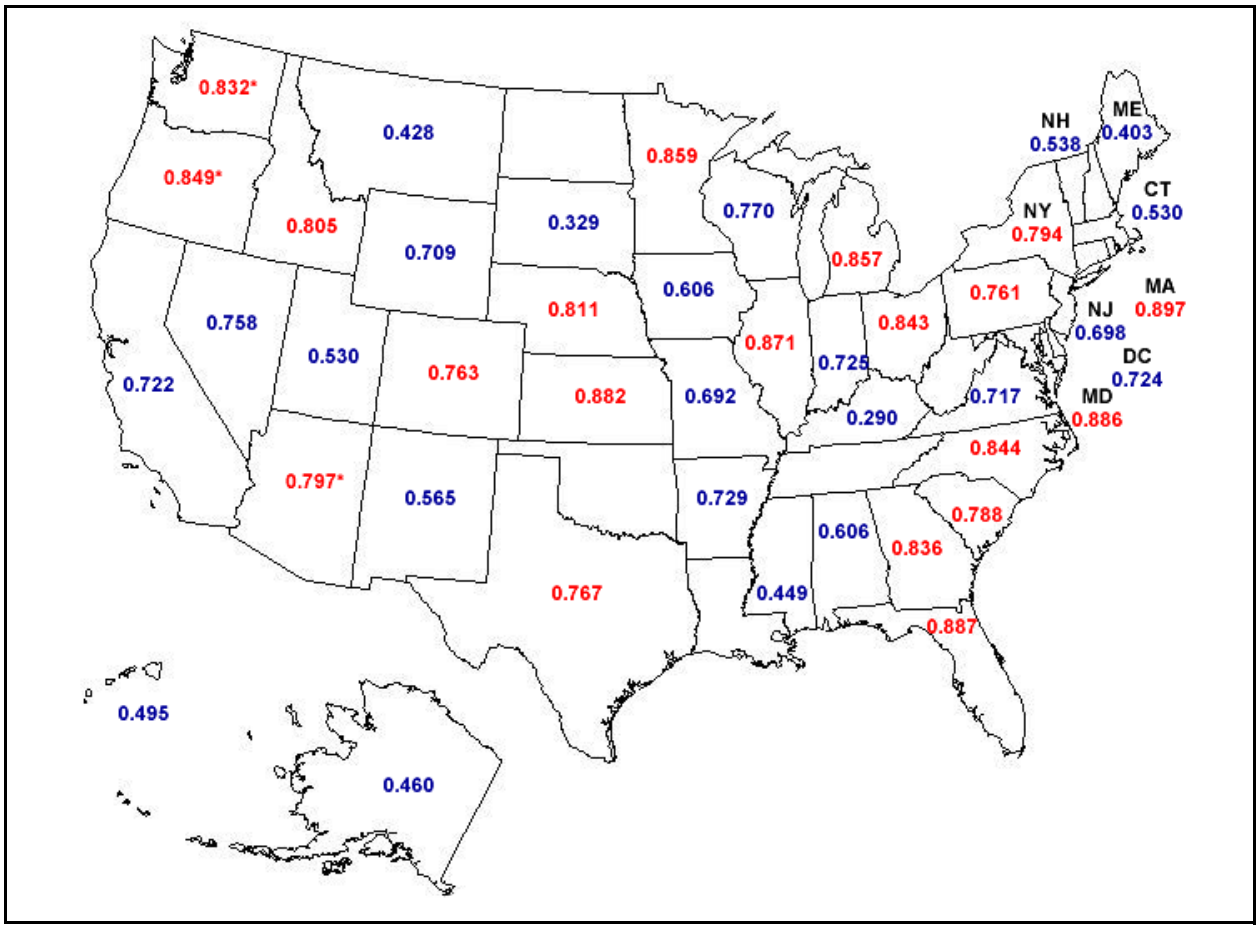

Figure 11. Geographical pattern of DEA scores (2002).

hypothesis. This implies that the US biotech industry exhibits the variable return to scale (VRS) and, in turn, the BCC model (Banker et al., 1984) is appropriate.

Biotechnology Industry Efficiencies of States

The DEA efficiencies by state are computed for the biotech industry assuming VRS. Table 3 shows estimated DEA efficiency. The column for number of firms reports the number of biotech firms in a state in 2002. The DEA-estimates column shows the conventional DEA estimates under VRS assumption. The corrected-DEA-estimates column includes the DEA estimates after correcting biased DEA estimates using the procedure proposed by Simar and Wilson (2007). Figure 11 shows the geographical pattern of DEA estimates. Note that some efficiency values in the biotech industry are 1997 estimates. From Table 3 and Figure 11, the larger states in terms of number of biotech firms tend to have higher efficiency scores. This implies *implicitly* that clustering the biotech industry is favorable to the overall performance of the state biotech industry. States that have a higher efficiency are Massachusetts, Florida, and Maryland. We expect that California, Pennsylvania, New York, and New Jersey are efficient states, but they are ranked in the middle (Table 3).

Causal Relationship

After estimating biotech industry efficiencies by state, the causal relationship among a set of variables is identified using DAG analysis (Pearl, 1995, 2000; Spirtes et al., 2000). As discussed earlier, 10 variables were selected and constructed. These variables were DEA estimates (*dea*), cluster dummy (*cluster*), regional income (*income*), population (*pop*), bioscience universities (*univ*), education level (*edu*), number of advanced degrees awarded in bioscience (*phd*), biotech firm size (*firmSz*), biotech average payroll (*pay*), and high-tech infrastructure (*NSEIs*). The DAG analysis result is

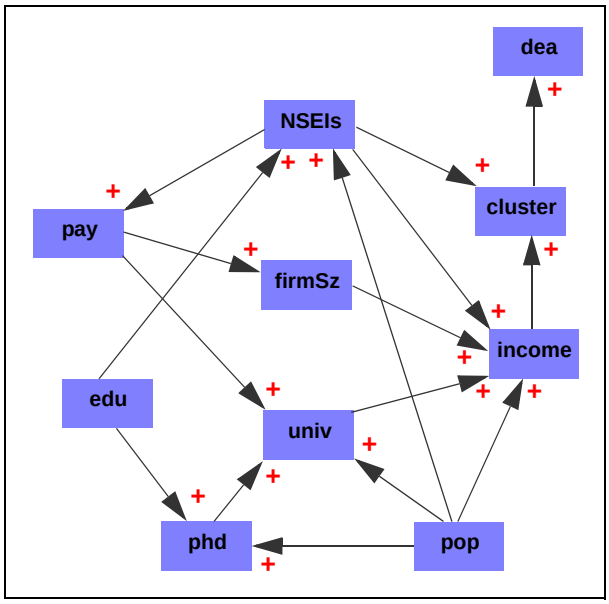

Figure 12. Causal relationship.

Note: Variables

1. ***dea**: efficiency score under variable return to scale assumption (Table 3)*
2. ***cluster**: **dummy variable** for the biotechnology cluster; this variable is constructed based on the slope of the CDF for number of firms; cluster=2 if Pr(noFirm > x)=80% or no. of firms > 350, cluster=1 if Pr(noFirm> x)=62% or 350 ³ no. of firms > 250 and cluster=0 otherwise;*
3. ***income**: personal income in million dollars from BEA*
4. ***phd:** number of PhD and MS biotechnology degrees awarded in 1998*
5. ***univ**: number of top 50 worldwide biotechnology universities based on DeVol et al. (2006)*
6. ***edu**: % of the population with bachelor's degree or higher*
7. ***firmSz**: average biotechnology firm size in number of employee*
8. ***pop**: state population in millions persons from REIS*
9. ***pay**: average payroll in thousand dollars*
10. ***NSEIs**: New State Economy Index Score; it measures the states' high-tech (new economy) infrastructure based on Atkinson and Court (1998) and Atkinson (2002)*

shown in Figure 12. A plus sign (+) in Figure 12 indicates the positive correlation between variables.

The cluster dummy and DEA estimates are key variables of interest. As shown in Figure 12, clustering has a positive impact on the biotech industry efficiency scores. This implies that biotech clustering enhances the performance of a state's biotech industry. High-technology infrastructure and regional income are the direct sources of clustering and indirectly impacts the DEA efficiency score through the cluster dummy. This implies that investment for high-technology infrastructure and improving regional income may be key factors to attract biotech firms. High-technology infrastructure is affected by population size and education level. Regional income is affected by population, biotech firm size, universities, and high-technology infrastructure.

The university variable is affected by average payroll and the number of advanced degrees awarded, which is not expected. However, it is plausible because the university variable is constructed based on bioscience publications, which is affected directly by the number of persons with higher degrees. In addition, higher payrolls attract higher-degree recipients from other states and have an effect on the university variable via publications. Education and population are considered as information roots; that is, all information starts from them. The DEA score is the information sink; that is, all information eventually flows to it (Figure 12). From this, it might be deduced that investment in education and policies for population growth are an important option to attract the biotech firms and improve their productivity. These results also show the impacts of agglomeration on biotech clustering.

Conclusion and Policy Implications

Many states are interested in attracting the biotech industry because it is believed to be a key driver for future economic development. The biotech industry usually clusters because it increases market competition by enhancing performance. The relationship between cluster and performance in the biotech industry is examined and confirmed that the causal flow is directed to performance from cluster. Other factors affecting location of the US biotech industry are identified. As shown in Figure 12, performance of the biotech industry is the information sink, and population and education level are the information roots.

The direct causes in forming a biotech cluster are high-tech infrastructure and regional income. The indirect factors in forming a cluster via regional income are biotech firm size, biotech-oriented universities, the number of advanced degrees awarded in the biotech major, and population. The indirect sources to attract clusters through high-tech infrastructure are average payroll in biotech industry and overall education level in the region, biotech firm size, and universities. All information flows to the biotech cluster and then the biotech industry productivity. The overall education level and population are the information source, and policies related to them are important to creating the cluster and the industry performance.

For states to encourage biotech firm development, they must invest in public goods that enhance the pro-

ductivity of biotech firms and enhance linkages as shown by the DAG analysis. Those states that have strong investments in these factors can see continued growth in biotech firms. Those states with insufficient investments must decide if they fiscally can compete with states already realizing biotech clustering and if investments are fiscally feasible.

References

Atkinson, R.D. (2002). *The 2002 state new economy index: Benchmarking economic transformation in the states*. Washington, DC: Progressive Policy Institute, Technology and New Economy Project.

Atkinson, R.D., & Court, R.H. (1998). *The new economy index: Understanding America's economic transformation*. Washington, DC: Progressive Policy Institute, Technology, Innovation, and New Economy Project.

Banker, R.D., Charnes, A., & Cooper, W.W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*(9), 1078-1092.

Battelle Technology Partnership Practice (BTPP). (2008). *Technology, talent and capital: State bioscience initiatives 2008* (Prepared for Biotechnology Industry Organization [BIO]). Washington, DC: Author.

Charnes, A., Clark, T., Cooper, W.W., & Golany, B. (1984). A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in US air forces. *Annals of Operational Research*, *2*(1), 95-112.

Charnes, A., Cooper, W.W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operation Research, 2*, 429-444.

Charnes, A., Cooper, W.W., & Rhodes, E. (1981). Program evaluation and managerial efficiency: An application of data envelopment analysis to program follow through. *Management Science*, *27*(6), 668-697.

Chickering, D.M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research, 2*, 445-498.

Chickering, D.M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507-554.

Cooper, W.W., Seiford, L.M., & Tone, K. (2007). *Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software* (2nd ed.). New York: Springer.

Cortright, J., & Mayer, H. (2002). *Signs of life: The growth of biotechnology centers in the United States.* Washington, DC: The Brookings Institution Center on Urban and Metropolitan Policy.

DeVol, R., Bedroussian, A., Babayan, A., Frye, M., Murphy, D., Philipson, T.J., et al. (2006). *Mind to market: A global analysis of university biotechnology transfer and commercialization.* Santa Monica, CA: Milken Institute.

Ernst & Young. (2007). *Beyond borders: Global biotechnology report 2007* (EYG No. EJ0003). London, UK: Author. Available on the World Wide Web http://www.ey.com/beyondborders.

Farrell, M.J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A 120*, 253-281.

Gattoufi, S., Oral, M., & Reisman, A. (2004). Data envelopment analysis literature: A bibliography update (1951-2001). *Socio-Economic Planning Sciences*, *38*(2-3), 159-229.

Geiger, D., Verma, T.S., & Pearl, J. (1990). Identifying independencies in Bayesian network. *Networks*, *20*, 507-534.

Goetz, S.J. (1997). *County-level determinants of high-technology firm locations: 1988-94* (TVA Rural Studies Program, Contractor Paper 98-4). Lexington: University of Kentucky, Department of Agricultural Economics.

Goetz, S.J., & Morgan, S. (1995). State-level locational determinants of biotechnology firms. *Economic Development Quarterly*, *9*(2), 174-184.

Goetz, S.J., & Ruspasingha, A. (2002). High-tech industry clustering: Implications for rural areas. *American Journal of Agricultural Economics*, *84*(5), 1229-1236.

Gray, M., & Parker, E. (1998). Industrial change and regional development: The case of the US biotechnology and pharmaceutical industries. *Environment and Planning*, *A30*(10), 1757-1774.

Hall, L., & Bagchi-Sen, S. (2001). An analysis of R&D, innovation, and business performance in the US biotechnology industry. *International Journal of Biotechnology*, *3*(3), 1-10.

Kim, M., & Harris, T.R. (2008, July). *An efficiency analysis of Nevada and Utah counties: Region size leads regional efficiency.* Paper presented at the 2008 American Agricultural Economics Association (AAEA) Annual Meeting, Orlando, Florida.

Meek, C. (1997). *Graphical models: Selecting causal and statistical models*. Unpublished Ph.D. dissertation, Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA.

Munroe, T., Craft, G., & Hutton, D. (2002). *A critical analysis of the local biotechnology industry cluster in Alameda, Contra Costa and Solano counties* (East Bay Bioscience Study) Oakland CA: East Bay EDA. Available on the World Wide Web: http://eastbayeda.org/research_facts_figures/archived_studies.htm.

Narin, F., Hamilton, K.S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, *26*, 317-330.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*, 669-710.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, MA: Cambridge University Press.

Porter, M. (1998). Cluster and the new economics of competition. *Harvard Business Review, 76*(Nov/Dec), 77-90.

Raab, R., & Kotamraju, P. (2006). The efficiency of the high-tech economy: Conventional development indexes versus a performance index. *Journal of Regional Science, 46*(3), 545-562.

Raab, R., & Lichty, R. (1997). An efficiency analysis of Minnesota counties: A data envelopment analysis using 1993 IMPLAN input-output analysis. *Journal of Regional Analysis and Policy, 27*(1), 75-93.

Raab, R., & Lichty, R. (2002). Identifying subareas that comprise a greater metropolitan area: The criterion of county relative efficiency. *Journal of Regional Science, 42*(3), 579-594.

Sambidi, P.R., & Harrison, R.W. (2006). Spatial clustering of the US biotech industry. Paper presented at the American Agricultural Economics Association Annual Meeting, Long Beach, CA.

Sanderson, A., & Dugoni, B. (2002). *Interstate migration patterns of recent science and engineering doctorate recipients* (Info Brief, NSF 02-311). Arlington, VA: Science Resources Statistics (SRS), National Science Foundation.

Shephard, R.W. (1970). *Theory of cost and production functions*. Princeton, NJ: Princeton University Press.

Simar, L., & Wilson, P.W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science, 44*(1), 49-61.

Simar, L., & Wilson, P.W. (2000). A general methodology for bootstrapping in nonparametric frontier models. *Journal of Applied Statistics, 27*(6), 779-802.

Simar, L., & Wilson, P.W. (2002). Non-parametric tests of returns to scale. *European Journal of Operational Research, 139*(1), 115-132.

Simar, L., & Wilson, P.W. (2007). Estimation and Inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics, 136*, 31-64.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge, MA: MIT Press.

US Census Bureau. (2001). *1997 economic census.* Washington, DC: US Department of Commerce, Economics and Statistics Administration.

US Census Bureau. (2005). *2002 economic census.* Washington, DC: US Department of Commerce, Economics and Statistics Administration.

US Patent and Trademark Office. (2008). *Patent counts by country/state and year utility patents, January 1, 1963-December 31, 2007* (Electronic Information Products Division/PTMT MDW 4C18). Alexandria, VA: Author.

Vusovic, S. (2006). *State level location determinants for biotechnology firms*. Unpublished master thesis, Department of Resource Economics, University of Nevada, Reno.

Wilson, P.W. (2006). FEAR: *A package for frontier efficiency analysis with R.* Clemson, SC: Clemson University. Available on the World Wide Web: http://www.clemson.edu/economics/faculty/wilson/Software/FEAR/FEAR-1.12/fear-1.12_user_guide.pdf.

Xia, Y., & Buccola, S. (2005). University life science programs and agricultural biotechnology. *American Journal of Agricultural Economics, 87*(1), 229-243.

Zucker, L.G., Darby, M.R., & Armstrong, J.S. (2002). Commercializing knowledge: University science, knowledge capture, and firm performance in biotechnology. *Management Science, 48*(1), 138-153.

Zucker, L.G., Darby, M.R. & Brewer, M.B. (1998). Intellectual human capital and the birth of US biotechnology enterprises. *American Economic Review, 88*(1), 290-306.