# Data Anti-Suppression Methods and an Application of Spatial Concentration Measures to Evaluate Green Energy Value Chain Concentration

A Thesis Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

David Lane Register
May 2012

## Abstract

A number of renewable, or so-called “green energy” products have emerged in response to the environmental, national security, and economic risks posed by the consumption of fossil fuels. The development of the renewable fuels sector, including ethanol, biodiesel, landfill and dairy methane gas, and solar and wind energy have also been considered potential sources of economic growth. But continuously changing market and policy conditions put at risk recent accomplishments in green energy sector development, thus placing a greater emphasis on efficient coordination between producers and suppliers.

Recent advances in firm location theory have produced a variety of analytical tools to determine the effects of geographical proximity on industry upstream and downstream linkages. Locations exhibiting a comparative advantage with respect to attracting green energy business establishments can also be statistically identified with these new approaches. These analytical developments are used to analyze the uneven spatial distribution of businesses and employment, generally described as “concentration”. Findings may supplement more effective policy design and recommendations, and could provide investors with additional information about which locations may be associated with cost savings.

The US County Business Patterns (CBP) database organizes employment and establishment data into a hierarchy by county and has been among the primary data sources for industry concentration analysis. Until recently, studies at finer levels of the geographic and industry hierarchy have been difficult due to suppression of employment data. Several procedures have been developed that use information contained in the CBP databases to impute missing employment records.

This two-paper thesis contributes to the analysis of firm location by: (1) providing a discussion of existing methods of missing employment data imputation and contributing a new imputation procedure, and (2) developing a two-stage method to analyze the geographic distribution of firms and employment, which is applied to ten value chains in the renewable energy sector. Full, accurate datasets and richer analytical methods can be applied to analyze industry concentration patterns and their potential impact on rural economies in future research. The findings and analytical contributions of this thesis may benefit policy makers and investors in green energy, regional scientists and economic geographers.

Table of Contents

List of Figures

List of Tables

# I. Introduction

Renewable, or so-called "green energy" has been considered a potential alternative to provide stable, long-term energy independence (Chichilnisky and Eiseberger 2009). If policies encouraging the development of the renewable energy sector are effective and public support remains optimistic, demand for a new set of skilled labor will grow and businesses will expand or emerge in response to new input and output markets. Newly developed markets for renewable energy technologies will depend on upstream and downstream linkages, or value chains. The relative infancy of these value chains contributes to coordination and logistical uncertainty among energy producers, providers, and consumers. Changing cost structures, overseas competition, and a dynamic policy landscape also contribute to investor uncertainty and perceptions of risk. To meet policy goals and to re-establish the United States as a leader in green energy technology development and production, policy makers require information to justify continued support for green energy industries. The focus of local policy makers will therefore be to determine which characteristics of their communities can be leveraged to attract new investment and retain the jobs and businesses already engaged in renewable energy supply. This two part thesis, first, explores several proposed solutions to employment data disclosure limitations that could hinder the analysis performed in the second part: a two-stage approach to describing the concentration tendencies of renewable energy firms.

Since the late 1980's, the so-called "New Economic Geography" literature has made substantial contributions toward understanding why economic activity is typically unevenly distributed across space, and the role agglomeration economics play in firm location (Fujita and Thisse 2009). Krugman (1991) characterized firm location as a "circularity" of decisions by manufacturers and consumers to the extent that location choices of firms may be, in part, a response to the location decisions of other firms. A broad range of analytical tools describing the location behavior of businesses and jobs have been formulated based on these observations. Empirical analysis of industry concentration has predominantly analyzed the distribution of manufacturing firms in the United States and Europe (Ellison and Glaeser 1997; Maurel and Sédillot 1999; Duranton and Overman 2002; Guimarães, Figueiredo, and Woodward 2007). Most studies conclude that scale economies external to firms and (or) access to local "natural advantages" result in lower variable costs, conferring comparative advantage to certain locations over others.

Geographic concentration of industries participating in the renewable energy sectors can be assessed using this paradigm. Analysis of firms involved in the supply and distribution of green energy products could bridge gaps in knowledge about coordination and logistical problems among value chain members. The location patterns of firms making up the value chains of ten emerging green energy sectors are analyzed in this thesis, including: (1) biodiesel, (2) coal co-firing, (3) wood direct fire, (4) ethanol production from switchgrass, (5) wood ethanol, (6) landfill methane, (7) dairy methane, (8) commercial solar energy production, (9) residential solar panel production, and (10) wind power. Each value chain considered is comprised of a variety of businesses involved in the extraction, production, and distribution or fuel productions, as well as financing operations. There is substantial overlap between these sectors in terms of the types of firms composing them.

To address questions of coordination and location advantage requires complete and accurate data – often a stumbling block for regional economists analyzing firm site selection, employment concentration, and attendant economic impacts. The fact that easily accessible data are often collected by government agencies leads to the problematic issue relating to the depth of employment data available due to disclosure limitations. Data collection agencies are prohibited from disclosing information that could reveal details about the operation of a business and its location. In firm and employment concentration analyses, this shortcoming is most often experienced in the suppression of employment data at relatively finer levels of industry detail or geographic units (for example, counties). Therefore, industry and county-level analyses often resort to alternative data sources that have been imputed. This thesis explores two existing employment data imputation procedures, and proposes a new method to accomplish this task. Enhanced employment data sets are useful for a wide range of applications in economic geography and regional science. Given the increased attention to renewable energy technologies in recent years, such applications may expand the knowledge base needed regarding the wider context of firms comprising this sector and their potential impact on job growth. Overcoming data suppression will also be important for forecasting the expansion of the renewable fuel sector so that investors and policy makers are provided with timely and accurate information about least cost locations.

This research also analyzes the relationship between firm location and the proximity to other similar or related industries. While the cluster analysis is predominantly quantitative, the results are interpreted qualitatively, without making conclusions about the local determinants driving the patterns observed. Information may be useful to state and local policy makers for targeting specific firms, as well as by researchers pursuing more detailed studies in green sector location patterns. The use of this information may also be relevant for investors in renewable energy sectors whose chance of success may be improved by both the support they receive from policy makers and researchers, and by more detailed knowledge about selecting an appropriate location for their businesses.

## References


Chichilnisky, G. and P. Eisenberger. 2009. Energy Security, Economic Development, and Global Warming: Addressing Short and Long Term Challenges. *International Journal of Green Economics* 3(3): 414-446.

Duranton, G. and H.G. Overman. 2002. Testing for Localization Using Micro-Geographic Data. *The Review of Economic Studies* 72(4): 1077-1106.

Ellison, G. and E. L. Glaeser. 1997. Geographic Concentration in US Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy* 105(5): 889-927.

Fujita, M., and J-F. Thisse. 2009. New Economic Geography: An appraisal on the occasion of Paul Krugman's 2008 Nobel Prize in Economics. *Regional Science and Urban Economics* 39:109-119.

Guimarães, P., O. Figueiredo, and D. Woodward. 2007. Measuring the Localization of Economic Activity: A Parametric Approach. *Journal of Regional Science* 47(4):753-774.

Krugman, P. 1991. Geography and Trade. MIT Press, Cambridge, MA, 1991.

Maurel, F., and B. Sédillot. 1999. A Measure of the Geographic Concentration in French Manufacturing Industries. *Regional Science and Urban Economics* 29: 575-604.

## II. Comparison of Methods to Estimate Suppressed Employment Records in the County Business Patterns Database

## Abstract

Regional economists are often forced to limit analyses to relatively high levels of regional or industry aggregation due to extensive data suppression of employment data. Suppression in the US County Business Patterns (CBP) database (one of the most widely used in regional analysis), may be addressed using procedures designed to impute missing data by mining available information. This research compares a goal programming method with an Iteratively Constrained Rebalanced Matrix procedure, which is a standard RAS procedure with constraints. Both procedures extract information about the suppressed employment data from employment flags and establishment size intervals found in the CBP which reduce uncertainty about the true (but unknown) employment records. Imputed employment estimates are constrained to fall within the narrowed intervals, resulting in imputed data sets that aggregate up to the distribution of employment across regions and industries.

Comparison of the optimizations procedures suggests that the interval constraints used in the goal programming approach may lead to indeterminate solutions from record errors in the original CBP data. Similar interval constraints used in the ICRM procedure restrict its flexibility to make estimates that sum to known industry and sector totals. On the other hand, the standard unconstrained ICRM algorithm forces estimates to aggregate up to state and industry totals. The choice between the constrained and unconstrained ICRM procedures may depend on research objectives and data requirements. The ICRM procedures appear competitive compared to the goal programming approach in terms of computational efficiency and programming effort.

## Introduction

Regional economists are frequently challenged by a lack of timely, accurate, or complete data sets. Studies in economic geography and regional science typically rely on firm and employment data to quantify and distinguish differences in the magnitude and types of economic activity. Among the most detailed industry classification systems is the North American Industry Classification System (NAICS), which assigns industry activities (and thus businesses performing these activities) into a hierarchy with unique identifiers. The Bureau of Labor Statistics (BLS) and the Census Bureau use the NAICS system to collect data about the distribution of business establishment and employment. The BLS disseminates employment records through a variety of channels, including the Quarterly Census of Employment and Wages (QCEW), while the Census Bureau develops the annual US County Business Patterns (CBP) databases for public distribution.

A challenge arises with respect to using the County Business Patterns (CBP) database. As most CBP users know, some employment records are subject to non-disclosure rules that limit its usefulness. The US Census Bureau describes the issue with a single line: "In accordance with US Code, Title 13, Section 9, no data are published that would disclose the operations of an individual employer." The Census Bureau uses a confidential algorithm to suppress employment data that could violate this code. Isserman and Westervelt (2006) described these data sets as resembling "a moth-eaten sweater" because of the roughly 1.5 million suppressed data points (about 58% of records). This astonishing lack of detail in the CBP employment data makes clear the case for methods to fill in missing data gaps, from which more accurate results and conclusions about business establishment activity and employment patterns at local levels can be inferred. Clearly, added flexibility from complete, but imputed, datasets could provide for richer econometric analyses with spatially and temporally deeper data.

User-friendly, accurate algorithms may be useful to researchers and the public at large. Algorithms could be applied to impute missing employment records in the CBP datasets for more accurate analysis of industry structure. This chapter compares several three competing methods for attending to data suppression issues in the CBP records.

## Exploring the CBP Databases

The CBP databases are important resources for businesses, universities, research organizations and government agencies. The CBP have been published for public use annually since 1964 by the US Bureau of Census. Data included in the CBP are collected from a variety of sources including the Internal Revenue Service, the Social Security Administration, and the Bureau of Labor Statistics. The data are arranged in a hierarchicy of sectors and counties. Since 1998, sector classifications have used the North American Industrial Classification System (NAICS) which characterizes business establishments according to a two- through six-digit code, with each increasing digit corresponding with a finer level of business establishments and industry structure detail[1]. It is important to define "establishments" in the context of NAICS and the CBP datasets. An "establishment" is a physical location where business is conducted, whereas. A firm or enterprise may own one or more establishments. An example is Wal-Mart. The company is headquartered in Bentonville, Arkansas, but Wal-Mart stores are ubiquitous. According to the CBP, each store represents an establishment in a county.

An array of information is found in the CBP. Data exist for all 50 states, Washington, D.C., Puerto Rico, and Island Areas including American Samoa, Guam, The Commonwealth of the Northern Mariana Islands, and the US Virgin Islands. National, state and county divisions are identified by Federal Information Processing Standard (FIPS) codes. For each location and sector

[1] Prior to 1998, Standard Industrial Classification (SIC) codes were used. While the two systems are different, there are methods whereby the two systems are commensurable.

division, one finds data for total employment, total Quarter-One payroll, annual payroll, the total number of establishments, and the number of establishments broken down by employment size. Table 1 provides a list of the data identifiers and descriptions for a representative CBP dataset.

Despite the abundance of data in the County Business Patterns, its usefulness is limited because employment records may be suppressed. However, some information can be extracted from suppressed records using the employment flag information (Table 2 provides a complete list of employment flags; note that flag "D" is not used) and establishment size intervals. For example, employment for mining (sector 22) in Apache County, Arizona (in 2002) was suppressed with an employment flag of "B" (20-99 employees). There are nine establishments in this sector in Apache County, with seven falling in the 1-4 size interval, one falling in the 5-9 size interval, and one falling in the 10-19 size interval. Therefore, the minimum number of employees is 22, according to the size intervals. However, determining a maximum from these intervals may not correspond with the maximum size of the employment flag, 99. The true maximum number of employees in this sector and county is 56 (the sum product of the number of establishments in each size interval and the upper bound of each interval). Therefore, the interval for employment in this sector for Apache County must be between 22 and 56. These deductions obviously do not arrive at an exact employment count, but they are of value to researchers hoping to mine data from the CBP to more accurately impute employment records.

### Existing Imputation Procedures for Non-disclosed CBP Employment Data

Obtaining exact employment figures is considered impossible for non-federal researchers. Exceptions include federal employees who have been deputized to access the monthly employment survey data. However, there are several imputation methods to obtain *more complete* records. For instance, Isserman and Westervelt (2006) proposed a data mining approach to impute missing employment records. Complete national datasets based on their method can be purchased for $3,000 for each data year. Alternatively, the Minnesota IMPLAN Group's National Data Package is available for purchase and contains imputed employment data derived from the company's private algorithm. Complete national data set at the county level start at $36,330. The BEA provides a web-based service for calculating employment location quotients. However, at present, data requests only accommodate at most three counties per search, making broader regional analysis cumbersome. Clearly, for many researchers and local government agencies, the acquisition of even one data year is costly, let alone multiple years.

There are a handful of other procedures that attempt to overcome County Business Patterns data suppression issues. A commonly mentioned approach for estimating non-disclosed employment data is the midpoint method (Glaeser et al. 1992; Porter 2003), so called because it suggests using the midpoint of the establishment size ranges multiplied by the number of establishments falling within that range[2]. Holmes and Stevens (2002; 2004) employed a similar employment imputation method using the average employment size per establishment rather than the midpoints. This approach may create inaccuracies because estimated employment will not necessarily aggregate up to state, county, national, and industry totals. Furthermore, the midpoint method cannot be applied to the largest employment flag M (100,000+ employees) because there is no upper bound.

The two-step data-mining approach suggested by Isserman and Westervelt (2006) is accomplished by: (1) mining all available data contained within the CBP database through an iterative bound-narrowing procedure; and (2) estimating suppressed employment values based on the narrowed ranges. Their procedure estimates missing employment based on modified establishment size bounds, the NAICS structure, and geographic hierarchies. Bound

[2] This method is recommended by the US Bureau of Census, though as demonstrated, likely not the best estimator.

determination is achieved iteratively, yielding progressively narrower employment ranges. Employment estimates are then made such that they fall within the narrowed ranges and are constrained to add up across industries and counties so that total values (which are observed at state and national levels) are maintained (e.g consistency). The authors used family-oriented titles to keep track of geographical and industrial relationships when adjusting bounds and making estimates. For example, 4-digit NAICS industries are considered *parents* of their successive 5-digit industries (*children*). Additionally, all 3-digit industries are considered *siblings* in the family parented by 2-digit industries. Similar relationships are applied to the geographic hierarchy of the nation, states, and counties. Therefore, states are the geographic parents of its counties (all siblings to each other), while the nation is the parent of all sibling states. Bound narrowing and employment estimation is completed through adjustments, up and down, through these hierarchies. Unfortunately, explanation of the algorithm is somewhat opaque. Detailed inspection of their 2002 dataset also reveals that aggregation consistency with state and national totals is not maintained, and, in some cases, true, known employment numbers change. However, deviations from the CBP-reported, known values of these "re-estimated" employment values appear to be small.

Zhang and Guldmann (2009) developed a goal programming approach to estimate the suppressed CBP employment entries. The linear program minimizes the sum of weighted deviations between the employment estimates and target values, subject to a set of consistency constraints. They use state and national level unsuppressed data, as well as assumptions similar to Isserman and Westervelt's approach to bound narrowing based on employment flags and establishment size intervals, thus creating more precise intervals. However, unlike the Isserman-Westervelt (IW) method, this procedure takes a "top-down" approach, working from the 2-digit to 6-digit NAICS level. The ZG approach can be solved using linear programming solvers, and the system of equations and constraints is accessible.

## Methods and Procedures

The Isserman and Westervelt and Zhang and Guldmann methods are compared to a method proposed here, referred to as an Iteratively-Constrained Rebalanced Matrix (ICRM) procedure. In comparing these three imputation methods, several data years are important. Since Isserman and Westervelt's procedure is documented without specific computational details, comparison is made on the basis of their 2002 CBP imputation, results of which are available for free through their website.

Zhang and Guldmann's (ZG) (2009) article reports a range of results from their imputation of Arizona and Ohio employment records for data year 2000. Therefore, to facilitate comparison of the ZG method with Isserman and Westervelt's approach, their procedure is used for 2000 (for the purpose of ensuring consistency with their reported 2000 results) and 2002 (to compare with the IW results). Finally, 2002 CBP employment data is imputed using the ICRM approach.

### *The ZG Imputation Method*

Parameters of the ZG linear programming problem are defined in Table 3, noting that $X_{ijk}$ is the "decision variable" (i.e. average employment in a missing cell). The objective function minimizes the difference between the absolute deviations of the observed and estimated employment levels

$$\min_{PD,ND,X} \sum_{i,j} w_{ij} * \left(PD_{ij} + ND_{ij}\right) \tag{1}$$

subject to

$$PD_{ij} - ND_{ij} = \sum_k \left[t_{ijk} * X_{ijk} - \left(pmin_{ij} + l_{ij} * \left(pmax_{ij} - pmin_{ij}\right)\right)\right], \quad (2)$$

$$PD_{ij} \geq 0, \qquad ND_{ij} \geq 0, \quad (3)$$

$$\sum_k X_{ijk} = 0 \ \ \forall i,j |\, H_{ij} = 0, \quad (4)$$

$$tmin_k \leq X_{ijk} \leq tmax_k \ \ \forall i,j | H_{ij} = 1, \quad (5)$$

$$fmin_{ij} \leq \sum_k (t_{ijk} * X_{ijk}) \leq fmax_{ij} \ \ \forall i,j | H_{ij} = 1, \quad (6)$$

$$\sum_j \left( \sum_k (t_{ijk} * X_{ijk}) + b_{ij} \right) = c_i \ \ \forall i | \sum_j H_{ij} > 0, \quad (7)$$

$$\sum_i \left( \sum_k (t_{ijk} * X_{ijk}) + b_{ij} \right) = s_j \ \ \forall j | \sum_i H_{ij} > 0, \quad (8)$$

$$X_{ijk} \geq 0, \quad (9)$$

where $l_{ij}$ is an interval location parameter ($\in[0,1]$), $H$ indicates employment entries that have been suppressed, $w_{ij}$ is a weight for the deviation (*i*,*j*), and

$$pmin_{ij} = MAX\left(fmin_{ij}, \sum_k tmin_k * t_{ijk}\right), \quad (10)$$

and

$$pmax_{ij} = MIN\left(fmax_{ij}, \sum_k tmax_k * t_{ijk}\right), \quad (11)$$

where $pmin_{ij}$ and $pmax_{ij}$ represent the endpoints of narrowed employment intervals. Constraints (4) and (9) require that all $X_{ijk}$'s are zero in the cells where employment is suppressed. Constraint (5) requires $X_{ijk}$ to take a value within an employment interval, *k*. Because $t_{ijk}$ is the total number of establishments in county *i*, sector *j*, interval *k*, when it is multiplied by the average establishment size ($X_{ijk}$), and summed across all intervals, it represents total employment for a sector in a county. Constraint (6) ensures that this value falls within the range defined by the employment flags. Constraints (7) and (8) require that the sum of actual employment and estimated employment across all sectors and counties (respectively) equal the total county and sector totals (respectively). Equations (10) and (11) define, based on the establishment and employment flag intervals, the narrowed interval for which each employment estimate ($t_{ijk} * X_{ijk}$) must fall (e.g. they are the minimum and maximum of Isserman and Westervelt's bounds). The interval location parameter $l_{ij}$ identifies the location in the interval where the target will fall and is set equal to 0.5[3] (i.e. the median of $pmin_{ij}$ and $pmax_{ij}$). The weight ($w_{ij}$) for the deviations is set to one in this application to ensure that size intervals are assumed to be of equal importance. Zhang and Guldmann (2009) found this assumption to be competitive with other values. For counties and sectors yielding infeasible solutions, Zhang and Guldmann (2009) recommend re-running the linear program, relaxing the constraints one-by-one to determine where the inconsistency is located as well as how to rectify it. Typically this amounts to manually adjusting the endpoints of the bounds, which can take a considerable amount of time trouble shooting by trial and error. With larger data sets or finer geospatial scales, this trouble shooting procedure would be impractical.

***Iteratively-Constrained Rebalanced Matrix Imputation Approach***

The second approach is a contribution made by this research; an iteratively constrained rebalanced matrix (ICRM) that is solved using a nonlinear program in combination with an interval bound constraint. The objective is to minimize the cross entropy of the row and column

[3] ZG (2009) identify alternative procedures for estimating $l_{ij}$ and $w_{ij}$ to address variability in the location of true, unknown estimates within the bounds, and variability in bound size, respectively. Ultimately, the authors recommend $w_{ij}$ = 1. The location parameter, $l_{ij}$ = 0.5 is used to avoid assumptions about the location of the unknown, true estmates.

scaling factors (Ireland and Kullback 1968), in a process similar to the matrix balancing RAS algorithm (Sinkhorn 1964; Bacharach 1970). The ICRM procedure is iterative, minimizing first a non-linear objective function subject to consistency constraints, followed by a bound-checking step which terminates when a convergence criteria is achieved.

To define the first step, let $c_i$ and $s_j$ be the respective county and sector employment totals for a state; $a_{0(ij)}$ be an employment entry in a cell (suppressed or unsuppressed); and $y_i$ and $z_j$ be county (row) and sector (column) scaling factors. Assume that $\boldsymbol{A}_0$ is the matrix containing the employment levels for a state (Arizona in this research) with rows (columns) corresponding with industry (county) employment totals. Before the algorithm begins, the suppressed employment entries are replaced by the midpoint values corresponding with the adjusted bounds. The algorithm is completed as follows:

1. The scaling factors are determined by minimizing:

$$\min_{y_i, z_j} \sum_i y_i \ln(y_i) + \sum_j z_j \ln(z_j) \tag{12}$$

subject to

$$\sum_j a_{0(ij)}^t \cdot y_i \cdot z_j = c_i \quad \forall\, j \tag{13}$$

$$\sum_i a_{0(ij)}^t \cdot y_i \cdot z_j = s_j \quad \forall\, i \tag{14}$$

$$y_i > 0, z_j > 0 \tag{15}$$

2. The values for $a_{ij}$ are imputed such that

$$a_{ij}^{t*} = \begin{cases} pmin_{ij} \leq a_{0(ij)}^{t-1} \cdot y_i^* \cdot z_j^* \leq pmax_{ij} & if\ a_{0(ij)}\ is\ suppresssed \\ a_{0(ij)} & otherwise \end{cases}. \tag{16}$$

3. Convergence of the algorithm is achieved when the difference in $\boldsymbol{A}^t$ and $\boldsymbol{A}^{t-1}$, after *t* iterations, is a small number.
4. If convergence is not reached, return to step 1.

### *Comparison of the IW, ZG, and ICRM Approaches*

To assess the accuracy of the ZG and ICRM methods, each approach is used to enhance the suppressed data at the 2-digit NAICS level for the 2002 Arizona employment data, and then compared to the IW data for Arizona (Table 4). This comparison is based on the assumption that the IW employment estimates are the best available approximation to the US CBP employment numbers. Goodness of fit is assessed for the ZG and ICRM methods with respect to IW data,

$$GOF_p = \sum_{i,j} (Est_{pij} - IW_{ij})^2, \tag{17}$$

where the $Est_{i,j}$ is the estimated employment value from procedure *p* (i.e. ZG, ICRM), and $IW_{ij}$ represents the individual employment values from the IW data set. Smaller values of $GOF_p$ suggest the results of the ICRM or ZG approaches are closer to the IW results.

### *Employment Data Inconsistencies in the CBP*

As noted by ZG (2009), consistency issues may arise with respect to the goal-programming approach. This was realized in their study when they imputed employment data for Ohio, which generated infeasible solutions. To rectify this issue, the authors suggested increasing employment values for sectors and counties identified with the inconsistencies. The obvious implication is that sector and county employment totals will no longer aggregate to reported totals, thus contributing to deviations in state totals. Therefore, if the two-digit level employment totals are used as constraints in more detailed imputation efforts, small deviations at the two-digit level may amplify deviations at lower levels. The usefulness of the ZG method, with respect to the extensive consistency issues within the County Business Patterns database, may therefore be diminished at successively finer NAICS levels. Further complicating this issue is the increased frequency of suppression occurring at finer industry and geographic levels. Relying on similar techniques to that of ZG, the current version of the ICRM method may also suffer from similar complications at finer levels of detail.

Another type of data inconsistency may arise from conflicting employment flags and establishment size intervals. Notice, in Figure 1, the employment flag "A" suggests that suppressed employment is between zero and 19. However, 24 establishments are reported, each with at least one employee. Therefore, the true maximum or true minimum of the suppressed employment value cannot be determined in this case. At the two-digit level, this problem is most common for NAICS sector 99; in the 2002 CBP, there are 61 entries across 22 states for which intervals are indeterminate. One additional inconsistency is noted for NAICS sector 44 in San Juan County, Colorado (FIPS code 08111). With over 16,800 suppressed entries at the two-digit level across the US, the percentage of conflicts may seem negligible. However, the fact that nearly half of the US states are affected suggests that each may suffer from infeasible solutions due to the bound-checking constraints of the ICRM and ZG methods.

In Isserman and Westervelt's approach, inconsistencies of this type may contribute to inconsistencies between industry and geographical parents and children, but appear to be attended to by assuming a deviation tolerance when bounds conflict. For example, the 2002 IW for Arizona indicate that seven imputed values do not fall within the narrowed ranges. Therefore, while the objectives of the ZG and ICRM approaches are based on fulfilling bound constraint requirements (e.g. summability) between industry and geographical siblings, the objective of Isserman and Westervelt's approach is to estimate a complete dataset, where summability appears to accept some margin of error. To illustrate the two procedures here, summability is required. Thus, data sets in which infeasibilities arise with respect to interval constraints or other inconsistencies are not used.

## Results and Discussion

The ZG model is estimated by minimizing (1) subject to (2) through (9) (using both 2000 and 2002 Arizona CBP to calibrate the model). The minimization routine based on the 2002 data returned an optimal solution (Table 5). Compared to the 2002 IW procedure, there appears to be little difference, though the latter underestimates reported employment (by one employee), while the former ensures county and sector totals are met. Note that in all tables, bold numbers indicate estimates that have been imputed.

Imputed sector and county employment data from the ICRM procedure deviate from the actual, reported figures (Table 6). Inspection of the actual CBP data reveals that one bound inconsistency exists for Arizona (again, see Figure 1). In the third step (equation (16)), the ICRM procedure fills in missing employment estimates according to bound constraints similar to those of ZG. The goal-programming approach of ZG presents an advantage in this regard by minimizing the deviation between county and sector total. The *soft* bound constraints in this procedure may be able to overcome bound inconsistencies to create a full data set in which sector and county totals are held constant. While inconsistencies, in the case of Ohio, are not resolved by this approach, it does overcome interval conflicts in the 2002 Arizona CBP. Unfortunately, this flexibility is not yet built into the ICRM procedure.

On the other hand, by omitting the bound checking step in (16), an unconstrained ICRM procedure can be performed, similar to the standard RAS procedure. This procedure creates a data set that aggregates to county and sector totals (Table 7). A similar operation is developed through a modified ZG method by omitting constraints (5), (6), (10), and (11), and reformulating (2) as

$$PD_{ij} - ND_{ij} = \sum_k [t_{ijk} * X_{ijk}]. \tag{18}$$

Results from this procedure are found in Table 8. Note that sector and county totals are maintained. The unconstrained ICRM algorithm and modified ZG approach are also applied to the 2000 data for Ohio, as a response to the inconsistencies noted in Zhang and Guldmann (2009). While the modified ZG approach results in an indeterminate solution, the unconstrained ICRM algorithm reaches optimality, suggesting advantages from the latter in overcoming inconsistent sector or county totals, in addition to interval inconsistencies.

Comparisons of each imputed data set to the IW data using the goodness of fit measure suggest that the constrained and unconstrained ICRM procedures are more similar to the IW data than the ZG method. Table 9 presents the ranked $GOF_p$ results for each computed data set. If the IW data is assumed to represent the closest approximation available of the suppressed CBP employment figures, the ZG procedure without bound constraints is the least desirable. Additionally, eight out of the 83 suppressed values for Arizona do not fall within the narrowed intervals. While the results closest to the IW imputed data are from standard ICRM procedure, they should be considered relative to the fact that step 3 may limit the algorithm's ability to achieve reported state and local totals. Also, note that all imputed values fall within the narrowed bounds, with the exception of the figure imputed for the sector and county that suffers from interval inconsistencies (Figure 1). The unconstrained ICRM algorithm includes a trade-off from the standard ICRM procedure; while the sector and county totals aggregate to the reported values in the CBP, interval constraints will likely not hold. In the absence of bound constraints, 35 imputed values from the unconstrained ICRM procedure fall outside their $pmin$ and $pmax$ ranges.

## Conclusions

Pervasive employment data suppression in the US County Business Patterns databases limits its usefulness. Recent methods to impute suppressed employment data have focused on reducing the uncertainty of employment estimates by using information from employment flags and establishment size intervals reported by the CBP. To date, the most complete national data set available is from WholeData.net, created using a proprietary algorithm about which little is known beyond the general description in Isserman and Westervelt (2006).

In this research, a goal programming procedure developed by Zhang and Guldmann (2009) and a proposed method, an Iteratively Constrained Rebalanced Matrix algorithm, were compared to the IW data, using 2002 Arizona employment data. Both procedures were susceptible bound inconsistencies between employment flags and establishment size intervals. The ICRM algorithm compensates for interval inconsistency at the expense of sector and county totals, while the ZG method permits deviations from any number of constraints. Goodness of fit measures favored the ICRM procedure over that of ZG in terms of how close the imputed data matched the IW data. The unconstrained ICRM algorithm, on the other hand, appears to have merit, using sector and county totals as the only constraints.

Eliminating bound information may appear counter-intuitive to researchers familiar with recent advancements in suppressed employment imputation, but the tradeoffs from their incorporation should be considered. Because flagging methods and suppression algorithms used by the Census Bureau are unknown, bound inconsistencies are impossible to rectify. Therefore, imputation methods designed to use narrowed bounds from the data to create imputed data sets that aggregate to industry and geographic totals necessarily involve a value judgment about the importance of either summability or interval consistency. Because interval consistency appears to be problematic, procedures similar to the ZG or ICRM approaches may produce undesirable results. However, using the unconstrained ICRM algorithm, unsuppressed data serve as *de facto* constraints; complete data sets consistent with reported geographic and industry totals can be created.

Through use of complete and accurate employment data sets, accurately imputed data sets will allow regional economists to produce more reliable results and useful conclusions about changes in employment and industry structure at local levels. The procedures discussed and implemented in this chapter represent simple and cost-effective methods of estimating suppressed CBP employment data. Exact reliability could be assessed using a well-designed Monte Carlo experiment.

## References

Bacharach, M. 1970. Biproportional Matrices and Input-Output Change. Cambridge University Press, London.

Bacharach, M. 1965. Estimating Non-Negative Matrices From Marginal Data. *International Economic Review* 6(3): 294-310.

Glaeser, E.L., H.D. Kallal, J.A. Scheinkman, and A. Shleifer. 1992. Growth in Cities. *Journal of Political Economy* 100(6): 1126-1152.

Holmes, T. J. and J.J. Stevens. 2004. Spatial Distribution of Economic Activities in North America. In: Henderson, J.V., and Thisse, J-F. (eds.) Handbook of Regional and Urban Economics. Elsevier, New York 2797-2844.

Holmes, T. J. and J.J. Stevens. 2002. Geographic Concentration and Establishment Scale. *The Review of Economics and Statistics* 84(4):682-690.

Ireland, C.T. and S. Kullback. 1968. Contingency Tables With Given Marginals. *Biometrika* 55(1): 179-188.

Isserman, A.M, and J. Westervelt. 2006. 1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data. *International Regional Science Review* 29(3): 311-355.

Porter, M.E. 2003. The Economic Performance of Regions. *Regional Studies* 37: 549-578.

Sinkhorn, R. 1964. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics* 35: 876-879.

US Census Bureau. 2006. County Business Patterns (NAICS) Database. Washington, D.C.

Westervelt, J. 2012. Email Message to Author, March 10.

Zhang, S., and J.M. Guldmann. 2009. Estimating Suppressed Data in Regional Economic Databases: A Goal-Programming Approach. *European Journal of Operational Research* 192(2009): 521-537.

## Appendix

Figure 1 - Example of Bound Conflict

| fipstate | fipscty | naics | empflag | emp | qp1 | ap | est | n1_4 |
|---|---|---|---|---|---|---|---|---|
| 04 | 025 | 813990 | | 81 | 459 | 2631 | 18 | 14 |
| 04 | 025 | 95---- | | 83 | 434 | 1560 | 5 | 3 |
| 04 | 025 | 99---- | A | 0 | 0 | 0 | 24 | 24 |

Table 1 - List of CBP Identifiers

| Identifier | Description |
|---|---|
| fipstate | 2-digit state FIPS code |
| fipscty | 3-digit county FIPS code |
| naics | 2- to 6-digit NAICS code |
| empflag | Employment flag |
| emp | Total employment |
| qp1 | Quarter one payroll |
| ap | Annual payroll |
| est | Total establishments |
| n1_4 | Number of establishments with 1 to 4 employees |
| n5_9 | Number of establishments with 5 to 9 employees |
| n10_19 | Number of establishments with 10 to 19 employees |
| n20_49 | Number of establishments with 20 to 49 employees |
| n50_99 | Number of establishments with 50 to 99 employees |
| n100_249 | Number of establishments with 100 to 249 employees |
| n250_499 | Number of establishments with 250 to 499 employees |
| n500_999 | Number of establishments with 500 to 999 employees |
| n1000 | Number of establishments with 1000+ employees |
| n1000_1 | Number of establishments with 1000 to 1499 employees |
| n1000_2 | Number of establishments with 1500 to 2499 employees |
| n1000_3 | Number of establishments with 2500 to 4999 employees |
| n1000_4 | Number of establishments with 5000+ employees |
| censtate | Census state code |
| cencty | Census county code |

Source: US Census Bureau

Table 2 - CBP Employment Flags

| EMPFLAG | Description |
|---|---|
| A | 0-19 employees |
| B | 20-99 employees |
| C | 100-249 employees |
| E | 250-499 employees |
| F | 500-999 employees |
| G | 1000-2499 employees |
| H | 2500-4999 employees |
| I | 5000-9999 employees |
| J | 10,000-24,999 employees |
| K | 25,000-49,999 employees |
| L | 50,000-99,999 employees |
| M | 100,000+ employees |

Source: US Census Bureau

Table 3 – Suppressed Data Estimation – Indices and Parameters

| Identifier | Description |
|---|---|
| $i$ | county index |
| $j$ | sector index (1→21 for NAICS 11→99) |
| $k$ | establishment size interval index (1→12) |
| $t_{ijk}$ | number of establishments in county $i$, sector $j$, and establishment interval $k$ |
| $b_{ij}$ | total employment in sector $j$ of county $i$ ($b_{ij}$ =0 if flagged) |
| $H_{ij}$ | 1 if sector $j$ in county $i$ is flagged, 0 otherwise |
| $s_j$ | total employment for sector $j$□across all sectors |
| $c_i$ | total employment for county $i$ across all sectors |
| $tmax_k$ | employment for upper bound for interval $k$ |
| $tmin_k$ | employment for lower bound for interval k |
| $fmax_{ij}$ | maximum employment of sector $j$ in county $i$, if flagged |
| $fmin_{ij}$ | minimum employment of sector $j$ in county $i$, if flagged |
| $X_{ijk}$ | average size of the establishment in county $i$, sector $j$, and interval $k$ |
| $PD_{ij}$ | positive deviation variable |
| $ND_{ij}$ | negative deviation variable |

Source: Zhang and Guldmann (2009)

Table 4 - IW 2002 Arizona CBP Employment Data

| NAICS | County | | | | | | | | | | | | | | | | Sector Totals | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4001 | 4003 | 4005 | 4007 | 4009 | 4011 | 4012 | 4013 | 4015 | 4017 | 4019 | 4021 | 4023 | 4025 | 4027 | 4999 | Reported | Actual | Difference |
| 11 | **14** | **29** | **31** | **122** | **71** | 0 | **5** | 593 | **11** | **39** | 157 | 142 | 50 | **8** | 563 | 0 | 1835 | 1836 | -1 |
| 21 | **15** | 64 | **69** | **944** | **47** | **2105** | **9** | 1458 | **104** | 871 | 1049 | 260 | **3** | 992 | **121** | 120 | 8231 | 8231 | 0 |
| 22 | **33** | 539 | **389** | 96 | **78** | **36** | **38** | 5816 | 338 | 451 | 1779 | 269 | **89** | 276 | 221 | 0 | 10448 | 10448 | 0 |
| 23 | 169 | 1635 | 2579 | 948 | 236 | **39** | **83** | 117709 | 6012 | 1684 | 23760 | 1562 | 514 | 5753 | 3192 | 0 | 165875 | 165875 | 0 |
| 31 | **165** | 574 | 2618 | **1391** | 259 | **2** | **321** | 118876 | 3174 | 645 | 29755 | 2972 | 640 | 3323 | 2856 | 0 | 167571 | 167570 | 1 |
| 42 | **55** | 493 | 1091 | 362 | 221 | **41** | 102 | 65537 | 832 | 302 | 7634 | 621 | 1980 | 1653 | 3686 | 0 | 84610 | 84610 | 0 |
| 44 | 1237 | 5775 | 7898 | 2219 | 1365 | 135 | 970 | 179339 | 8679 | 4170 | 44045 | 5960 | 2309 | 9171 | 6558 | 0 | 279830 | 279830 | 0 |
| 48 | 81 | 357 | 982 | 236 | 53 | **31** | **37** | 69730 | 894 | 361 | 5188 | 422 | 790 | 697 | 802 | **2** | 80663 | 80663 | 0 |
| 51 | **151** | 551 | 642 | 178 | 112 | **25** | **39** | 42707 | 980 | 332 | 6983 | 317 | 85 | 761 | 567 | 0 | 54430 | 54430 | 0 |
| 52 | 107 | 568 | 797 | 220 | 95 | **16** | 67 | 93454 | 1216 | 425 | 9054 | 727 | 247 | 1269 | 859 | **24** | 109145 | 109145 | 0 |
| 53 | **49** | 625 | 652 | 154 | 86 | **14** | 131 | 32178 | 715 | 288 | 6639 | 654 | 196 | 1488 | 529 | 0 | 44398 | 44398 | 0 |
| 54 | 143 | 2989 | 1202 | 291 | 118 | **3** | **33** | 86677 | 1104 | 452 | 16683 | 935 | 214 | 1810 | 974 | **70** | 113698 | 113699 | -1 |
| 55 | **14** | 98 | 359 | **38** | **13** | **6** | 53 | 37999 | **132** | 217 | 5418 | 140 | **79** | **35** | 187 | 0 | 44788 | 44734 | 54 |
| 56 | 133 | 1504 | 1520 | 321 | 96 | 0 | 147 | 138558 | 1515 | 245 | 27073 | 2347 | 157 | 2310 | 2488 | 14337 | 192751 | 192657 | 94 |
| 61 | 460 | **370** | 519 | 185 | **80** | **4** | 0 | 19935 | 368 | 676 | 4470 | 526 | 221 | 1924 | 662 | **2** | 30402 | 30549 | -147 |
| 62 | 1678 | 4118 | 5685 | 2245 | 966 | **114** | 347 | 132511 | 4857 | 2374 | 42503 | 3575 | 664 | 7417 | 5129 | **2** | 214185 | 214185 | 0 |
| 71 | **39** | 325 | 1141 | 849 | **15** | **15** | **241** | 24338 | 528 | 212 | 7173 | 1056 | 173 | 1347 | 633 | 0 | 38085 | 38085 | 0 |
| 72 | 1064 | 3849 | 8831 | 1547 | 798 | **106** | 1148 | 127187 | 4897 | 3269 | 36520 | 4436 | 1267 | 6650 | 4614 | **2** | 206185 | 206185 | 0 |
| 81 | 195 | 1075 | 1757 | 405 | 556 | **9** | **114** | 52993 | 1818 | 1265 | 14151 | 1501 | 208 | 2272 | 1511 | 0 | 79830 | 79830 | 0 |
| 95 | **9** | 59 | **401** | **42** | **4** | 0 | **2** | 14112 | **12** | **35** | 2942 | 74 | 194 | 83 | **113** | **2** | 18084 | 18085 | -1 |
| 99 | **9** | **6** | **4** | 0 | **2** | 0 | **1** | 222 | 14 | 9 | 137 | 8 | **7** | **5** | **3** | 0 | 427 | 427 | 0 |
| Imputed | 5820 | 25603 | 39167 | 12793 | 5271 | 2701 | 3888 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14561 | 1945471 | | |
| Actual | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | | 1945472 | |
| Difference | 2 | 0 | 0 | 0 | 1 | -2 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | | | -1 |

Note: entries in bold are imputed.

Table 5 - ZG 2002 Arizona CBP Employment Data - With Bounds

| NAICS | County | | | | | | | | | | | | | | | | Sector Totals | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4001 | 4003 | 4005 | 4007 | 4009 | 4011 | 4012 | 4013 | 4015 | 4017 | 4019 | 4021 | 4023 | 4025 | 4027 | 4999 | Reported | Actual | Difference |
| 11 | **19** | **24** | **39** | **100** | **83** | 0 | **12** | 593 | **10** | **35** | 157 | 142 | 50 | **9** | 563 | 0 | 1836 | 1836 | 0 |
| 21 | **11** | 64 | **48** | **975** | **68** | **2069** | **9** | 1458 | **100** | 871 | 1049 | 260 | **4** | 992 | **132** | 120 | 8231 | 8231 | 0 |
| 22 | **22** | 539 | **405** | 96 | **45** | **39** | **53** | 5816 | 338 | 451 | 1779 | 269 | **99** | 276 | 221 | 0 | 10448 | 10448 | 0 |
| 23 | 169 | 1635 | 2579 | 948 | 236 | **26** | **96** | 117709 | 6012 | 1684 | 23760 | 1562 | 514 | 5753 | 3192 | 0 | 165875 | 165875 | 0 |
| 31 | **233** | 574 | 2618 | **1382** | 259 | **4** | **259** | 118876 | 3174 | 645 | 29755 | 2972 | 640 | 3323 | 2856 | 0 | 167570 | 167570 | 0 |
| 42 | **41** | 493 | 1091 | 362 | 221 | **55** | 102 | 65537 | 832 | 302 | 7634 | 621 | 1980 | 1653 | 3686 | 0 | 84610 | 84610 | 0 |
| 44 | 1237 | 5775 | 7898 | 2219 | 1365 | 135 | 970 | 179339 | 8679 | 4170 | 44045 | 5960 | 2309 | 9171 | 6558 | 0 | 279830 | 279830 | 0 |
| 48 | 81 | 357 | 982 | 236 | 53 | **23** | **43** | 69730 | 894 | 361 | 5188 | 422 | 790 | 697 | 802 | **4** | 80663 | 80663 | 0 |
| 51 | **121** | 551 | 642 | 178 | 112 | **35** | **59** | 42707 | 980 | 332 | 6983 | 317 | 85 | 761 | 567 | 0 | 54430 | 54430 | 0 |
| 52 | 107 | 568 | 797 | 220 | 95 | **19** | 67 | 93454 | 1216 | 425 | 9054 | 727 | 247 | 1269 | 859 | **21** | 109145 | 109145 | 0 |
| 53 | **44** | 625 | 652 | 154 | 86 | **19** | 131 | 32178 | 715 | 288 | 6639 | 654 | 196 | 1488 | 529 | 0 | 44398 | 44398 | 0 |
| 54 | 143 | 2989 | 1202 | 291 | 118 | **8** | **35** | 86677 | 1104 | 452 | 16683 | 935 | 214 | 1810 | 974 | **64** | 113699 | 113699 | 0 |
| 55 | **19** | 98 | 359 | **30** | **19** | **9** | 0 | 37999 | **139** | 217 | 5418 | 140 | **70** | **29** | 187 | 0 | 44734 | 44734 | 0 |
| 56 | 133 | 1504 | 1520 | 321 | 96 | 0 | 53 | 138558 | 1515 | 245 | 27073 | 2347 | 157 | 2310 | 2488 | 14337 | 192657 | 192657 | 0 |
| 61 | 460 | **376** | 519 | 185 | **68** | **8** | 147 | 19935 | 368 | 676 | 4470 | 526 | 221 | 1924 | 662 | **4** | 30549 | 30549 | 0 |
| 62 | 1678 | 4118 | 5685 | 2245 | 966 | **112** | 347 | 132511 | 4857 | 2374 | 42503 | 3575 | 664 | 7417 | 5129 | **4** | 214185 | 214185 | 0 |
| 71 | **25** | 325 | 1141 | 849 | **17** | **19** | **249** | 24338 | 528 | 212 | 7173 | 1056 | 173 | 1347 | 633 | 0 | 38085 | 38085 | 0 |
| 72 | 1064 | 3849 | 8831 | 1547 | 798 | **104** | 1148 | 127187 | 4897 | 3269 | 36520 | 4436 | 1267 | 6650 | 4614 | **4** | 206185 | 206185 | 0 |
| 81 | 195 | 1075 | 1757 | 405 | 556 | **19** | **104** | 52993 | 1818 | 1265 | 14151 | 1501 | 208 | 2272 | 1511 | 0 | 79830 | 79830 | 0 |
| 95 | **9** | 59 | **397** | **50** | **8** | 0 | **4** | 14112 | **10** | **39** | 2942 | 74 | 194 | 83 | **100** | **4** | 18085 | 18085 | 0 |
| 99 | **7** | **5** | **5** | 0 | **1** | 0 | **0** | 222 | 14 | 9 | 137 | 8 | **5** | **10** | **5** | 0 | 427 | 427 | 0 |
| Imputed | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | 1945472 | | |
| Actual | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | | 1945472 | |
| Difference | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | 0 |

Note: entries in bold are imputed.

Table 6 - ICRM 2002 Arizona CBP Employment Data

| NAICS | County | | | | | | | | | | | | | | | | Sector Totals | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4001 | 4003 | 4005 | 4007 | 4009 | 4011 | 4012 | 4013 | 4015 | 4017 | 4019 | 4021 | 4023 | 4025 | 4027 | 4999 | Reported | Actual | Difference |
| 11 | **19** | **35** | **41** | **116** | **61** | 0 | **8** | 593 | **7** | **35** | 157 | 142 | 50 | **9** | 563 | 0 | 1836 | 1836 | 0 |
| 21 | **12** | 64 | **76** | **907** | **63** | **2088** | **12** | 1458 | **128** | 871 | 1049 | 260 | **4** | 992 | **127** | 120 | 8231 | 8231 | 0 |
| 22 | **52** | 539 | **372** | 96 | **63** | **42** | **53** | 5816 | 338 | 451 | 1779 | 269 | **81** | 276 | 221 | 0 | 10448 | 10448 | 0 |
| 23 | 169 | 1635 | 2579 | 948 | 236 | **57** | **63** | 117709 | 6012 | 1684 | 23760 | 1562 | 514 | 5753 | 3192 | 0 | 165873 | 165875 | -2 |
| 31 | **146** | 574 | 2618 | **1406** | 259 | **3** | **320** | 118876 | 3174 | 645 | 29755 | 2972 | 640 | 3323 | 2856 | 0 | 167567 | 167570 | -3 |
| 42 | **47** | 493 | 1091 | 362 | 221 | **48** | 102 | 65537 | 832 | 302 | 7634 | 621 | 1980 | 1653 | 3686 | 0 | 84609 | 84610 | -1 |
| 44 | 1237 | 5775 | 7898 | 2219 | 1365 | 135 | 970 | 179339 | 8679 | 4170 | 44045 | 5960 | 2309 | 9171 | 6558 | 0 | 279830 | 279830 | 0 |
| 48 | 81 | 357 | 982 | 236 | 53 | **29** | **36** | 69730 | 894 | 361 | 5188 | 422 | 790 | 697 | 802 | **3** | 80662 | 80663 | -1 |
| 51 | **138** | 551 | 642 | 178 | 112 | **27** | **49** | 42707 | 980 | 332 | 6983 | 317 | 85 | 761 | 567 | 0 | 54429 | 54430 | -1 |
| 52 | 107 | 568 | 797 | 220 | 95 | **8** | 67 | 93454 | 1216 | 425 | 9054 | 727 | 247 | 1269 | 859 | **30** | 109143 | 109145 | -2 |
| 53 | **51** | 625 | 652 | 154 | 86 | **11** | 131 | 32178 | 715 | 288 | 6639 | 654 | 196 | 1488 | 529 | 0 | 44397 | 44398 | -1 |
| 54 | 143 | 2989 | 1202 | 291 | 118 | **6** | **43** | 86677 | 1104 | 452 | 16683 | 935 | 214 | 1810 | 974 | **56** | 113697 | 113699 | -2 |
| 55 | **12** | 98 | 359 | **53** | **11** | **8** | 0 | 37999 | **119** | 217 | 5418 | 140 | **83** | **29** | 187 | 0 | 44733 | 44734 | -1 |
| 56 | 133 | 1504 | 1520 | 321 | 96 | 0 | 53 | 138558 | 1515 | 245 | 27073 | 2347 | 157 | 2310 | 2488 | 14337 | 192657 | 192657 | 0 |
| 61 | 460 | **360** | 519 | 185 | **84** | **8** | 147 | 19935 | 368 | 676 | 4470 | 526 | 221 | 1924 | 662 | **4** | 30549 | 30549 | 0 |
| 62 | 1678 | 4118 | 5685 | 2245 | 966 | **109** | 347 | 132511 | 4857 | 2374 | 42503 | 3575 | 664 | 7417 | 5129 | **4** | 214182 | 214185 | -3 |
| 71 | **56** | 325 | 1141 | 849 | **17** | **14** | **222** | 24338 | 528 | 212 | 7173 | 1056 | 173 | 1347 | 633 | 0 | 38084 | 38085 | -1 |
| 72 | 1064 | 3849 | 8831 | 1547 | 798 | **101** | 1148 | 127187 | 4897 | 3269 | 36520 | 4436 | 1267 | 6650 | 4614 | **4** | 206182 | 206185 | -3 |
| 81 | 195 | 1075 | 1757 | 405 | 556 | **8** | **114** | 52993 | 1818 | 1265 | 14151 | 1501 | 208 | 2272 | 1511 | 0 | 79829 | 79830 | -1 |
| 95 | **10** | 59 | **393** | **55** | **8** | 0 | **4** | 14112 | **7** | **40** | 2942 | 74 | 194 | 83 | **100** | **4** | 18085 | 18085 | 0 |
| 99 | **9** | **12** | **13** | 0 | **2** | 0 | **1** | 222 | 14 | 9 | 137 | 8 | **11** | **24** | **11** | 0 | 473 | 427 | 46 |
| Imputed | 5819 | 25604 | 39168 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38202 | 18323 | 293113 | 28504 | 10088 | 49258 | 36269 | 14562 | 1945495 | | |
| Actual | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | | 1945472 | |
| Difference | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 14 | 1 | 0 | | | 23 |

Note: entries in bold are imputed.

Table 7 - Unconstrained ICRM 2002 Arizona CBP Employment Data

| NAICS | County | | | | | | | | | | | | | | | | Sector Totals | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4001 | 4003 | 4005 | 4007 | 4009 | 4011 | 4012 | 4013 | 4015 | 4017 | 4019 | 4021 | 4023 | 4025 | 4027 | 4999 | Reported | Actual | Difference |
| 11 | **6** | **36** | **46** | **138** | **55** | 0 | **6** | 593 | **4** | **34** | 157 | 142 | 50 | **4** | 563 | 0 | 1836 | 1836 | 0 |
| 21 | **9** | 64 | **67** | **903** | **81** | **2070** | **9** | 1458 | **135** | 871 | 1049 | 260 | **13** | 992 | **130** | 120 | 8231 | 8231 | 0 |
| 22 | **51** | 539 | **380** | 96 | **62** | **48** | **48** | 5816 | 338 | 451 | 1779 | 269 | **74** | 276 | 221 | 0 | 10448 | 10448 | 0 |
| 23 | 169 | 1635 | 2579 | 948 | 236 | **61** | **61** | 117709 | 6012 | 1684 | 23760 | 1562 | 514 | 5753 | 3192 | 0 | 165875 | 165875 | 0 |
| 31 | **153** | 574 | 2618 | **1395** | 259 | **6** | **325** | 118876 | 3174 | 645 | 29755 | 2972 | 640 | 3323 | 2856 | 0 | 167570 | 167570 | 0 |
| 42 | **49** | 493 | 1091 | 362 | 221 | **46** | 102 | 65537 | 832 | 302 | 7634 | 621 | 1980 | 1653 | 3686 | 0 | 84610 | 84610 | 0 |
| 44 | 1237 | 5775 | 7898 | 2219 | 1365 | 135 | 970 | 179339 | 8679 | 4170 | 44045 | 5960 | 2309 | 9171 | 6558 | 0 | 279830 | 279830 | 0 |
| 48 | 81 | 357 | 982 | 236 | 53 | **33** | **33** | 69730 | 894 | 361 | 5188 | 422 | 790 | 697 | 802 | **4** | 80663 | 80663 | 0 |
| 51 | **137** | 551 | 642 | 178 | 112 | **39** | **39** | 42707 | 980 | 332 | 6983 | 317 | 85 | 761 | 567 | 0 | 54430 | 54430 | 0 |
| 52 | 107 | 568 | 797 | 220 | 95 | **5** | 67 | 93454 | 1216 | 425 | 9054 | 727 | 247 | 1269 | 859 | **35** | 109145 | 109145 | 0 |
| 53 | **56** | 625 | 652 | 154 | 86 | **7** | 131 | 32178 | 715 | 288 | 6639 | 654 | 196 | 1488 | 529 | 0 | 44398 | 44398 | 0 |
| 54 | 143 | 2989 | 1202 | 291 | 118 | **7** | **54** | 86677 | 1104 | 452 | 16683 | 935 | 214 | 1810 | 974 | **46** | 113699 | 113699 | 0 |
| 55 | **8** | 98 | 359 | **52** | **9** | **7** | 0 | 37999 | **116** | 217 | 5418 | 140 | **83** | **40** | 187 | 0 | 44734 | 44734 | 0 |
| 56 | 133 | 1504 | 1520 | 321 | 96 | 0 | 53 | 138558 | 1515 | 245 | 27073 | 2347 | 157 | 2310 | 2488 | 14337 | 192657 | 192657 | 0 |
| 61 | 460 | **365** | 519 | 185 | **76** | **8** | 147 | 19935 | 368 | 676 | 4470 | 526 | 221 | 1924 | 662 | **7** | 30549 | 30549 | 0 |
| 62 | 1678 | 4118 | 5685 | 2245 | 966 | **112** | 347 | 132511 | 4857 | 2374 | 42503 | 3575 | 664 | 7417 | 5129 | **4** | 214185 | 214185 | 0 |
| 71 | **70** | 325 | 1141 | 849 | **12** | **9** | **219** | 24338 | 528 | 212 | 7173 | 1056 | 173 | 1347 | 633 | 0 | 38085 | 38085 | 0 |
| 72 | 1064 | 3849 | 8831 | 1547 | 798 | **104** | 1148 | 127187 | 4897 | 3269 | 36520 | 4436 | 1267 | 6650 | 4614 | **4** | 206185 | 206185 | 0 |
| 81 | 195 | 1075 | 1757 | 405 | 556 | **5** | **118** | 52993 | 1818 | 1265 | 14151 | 1501 | 208 | 2272 | 1511 | 0 | 79830 | 79830 | 0 |
| 95 | **7** | 59 | **396** | **49** | **9** | 0 | **7** | 14112 | **4** | **40** | 2942 | 74 | 194 | 83 | **104** | **6** | 18085 | 18085 | 0 |
| 99 | **5** | **4** | **5** | 0 | **6** | 0 | **5** | 222 | 14 | 9 | 137 | 8 | **7** | **3** | **3** | 0 | 427 | 427 | 0 |
| Imputed | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | 1945472 | | |
| Actual | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | | 1945472 | |
| Difference | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | 0 |

Note: entries in bold are imputed.

Table 8 - ZG 2002 Arizona CBP Employment Data - Without Bounds

| NAICS | County | | | | | | | | | | | | | | | | Sector Totals | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4001 | 4003 | 4005 | 4007 | 4009 | 4011 | 4012 | 4013 | 4015 | 4017 | 4019 | 4021 | 4023 | 4025 | 4027 | 4999 | Reported | Actual | Difference |
| 11 | **169** | **64** | **663** | **27** | **24** | 2011 | **679** | 593 | **259** | **871** | 157 | 142 | 50 | **48** | 563 | 120 | 1836 | 1836 | 0 |
| 21 | **0** | 539 | **2579** | **96** | **285** | **96** | **122** | 1458 | **338** | 451 | 1049 | 260 | **178** | 992 | **237** | 0 | 8231 | 8231 | 0 |
| 22 | **0** | 1635 | **2618** | 948 | **236** | **0** | **102** | 5816 | 6012 | 1684 | 1779 | 269 | **514** | 276 | 221 | 0 | 10448 | 10448 | 0 |
| 23 | 0 | 574 | 1091 | 1878 | 259 | **0** | **0** | 117709 | 3174 | 645 | 23760 | 1562 | 640 | 5753 | 3192 | 0 | 165875 | 165875 | 0 |
| 31 | **0** | 493 | 0 | **362** | 221 | **0** | **0** | 118876 | 832 | 302 | 29755 | 2972 | 1980 | 3323 | 2856 | 0 | 167570 | 167570 | 0 |
| 42 | **0** | 0 | 0 | 0 | 0 | **0** | 0 | 65537 | 0 | 0 | 7634 | 621 | 0 | 1653 | 3686 | 0 | 84610 | 84610 | 0 |
| 44 | 1237 | 5775 | 7898 | 2219 | 1365 | 135 | 970 | 179339 | 8679 | 4170 | 44045 | 5960 | 2309 | 9171 | 6558 | 43 | 279830 | 279830 | 0 |
| 48 | 81 | 357 | 982 | 236 | 53 | **70** | **67** | 69730 | 894 | 361 | 5188 | 422 | 790 | 697 | 802 | **0** | 80663 | 80663 | 0 |
| 51 | **215** | 551 | 642 | 178 | 112 | **40** | **131** | 42707 | 980 | 332 | 6983 | 317 | 85 | 761 | 567 | 0 | 54430 | 54430 | 0 |
| 52 | 107 | 568 | 797 | 220 | 95 | **63** | 0 | 93454 | 1216 | 425 | 9054 | 727 | 247 | 1269 | 859 | **0** | 109145 | 109145 | 0 |
| 53 | **143** | 625 | 652 | 154 | 86 | **64** | 0 | 32178 | 715 | 288 | 6639 | 654 | 196 | 1488 | 529 | 0 | 44398 | 44398 | 0 |
| 54 | 0 | 2989 | 1202 | 291 | 118 | **0** | **0** | 86677 | 1104 | 452 | 16683 | 935 | 214 | 1810 | 974 | **0** | 113699 | 113699 | 0 |
| 55 | **133** | 98 | 359 | **316** | **96** | **116** | 53 | 37999 | **1515** | 217 | 5418 | 140 | **157** | **2310** | 187 | 14337 | 44734 | 44734 | 0 |
| 56 | 460 | 1504 | 1520 | 321 | 966 | 108 | 147 | 138558 | 368 | 245 | 27073 | 2347 | 221 | 1924 | 2488 | 62 | 192657 | 192657 | 0 |
| 61 | 1678 | **394** | 519 | 185 | **798** | **0** | 347 | 19935 | 4857 | 676 | 4470 | 526 | 664 | 7417 | 662 | **0** | 30549 | 30549 | 0 |
| 62 | 310 | 4118 | 5685 | 2245 | 0 | **0** | 1148 | 132511 | 528 | 2374 | 42503 | 3575 | 173 | 1347 | 5129 | **0** | 214185 | 214185 | 0 |
| 71 | **1064** | 325 | 1141 | 849 | **0** | **0** | **0** | 24338 | 4897 | 212 | 7173 | 1056 | 1267 | 6650 | 633 | 0 | 38085 | 38085 | 0 |
| 72 | 0 | 3849 | 8831 | 1547 | 0 | **0** | 0 | 127187 | 0 | 3269 | 36520 | 4436 | 0 | 0 | 4614 | **0** | 206185 | 206185 | 0 |
| 81 | 195 | 1075 | 1757 | 405 | 556 | **0** | **123** | 52993 | 1818 | 1265 | 14151 | 1501 | 208 | 2272 | 0 | 1511 | 79830 | 79830 | 0 |
| 95 | **26** | 59 | **231** | **316** | **0** | 0 | **0** | 14112 | **14** | **74** | 2942 | 74 | 194 | 83 | **0** | **0** | 18085 | 18085 | 0 |
| 99 | **0** | **11** | **0** | 0 | **0** | 0 | **0** | 222 | 0 | 9 | 137 | 8 | **0** | **0** | **0** | 0 | 427 | 427 | 0 |
| Imputed | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | 1945472 | | |
| Actual | 5818 | 25603 | 39167 | 12793 | 5270 | 2703 | 3889 | 1361929 | 38200 | 18322 | 293113 | 28504 | 10087 | 49244 | 36268 | 14562 | | 1945472 | |
| Difference | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | | 0 |

Note: entries in bold are imputed.

Table 9 – Goodness of fit relative to 2002 IW data

| Rank | Procedure (*p*) | Goodness of Fit |
|---|---|---|
| 1 | ICRM | 41,795 |
| 2 | Unconstrained ICRM | 46,138 |
| 3 | ZG | 51,431 |
| 4 | ZG (without bound constraints) | 751,326,895 |

# III. Green Energy Value Chain Concentration Patterns

**Abstract**

Renewable energy sources are potential responses to the environmental and national security risks posed by dependence on fossil fuels. The development of the renewable fuels sector, including ethanol, biodiesel, landfill and dairy methane gas, and solar and wind energy have also been considered potential sources of economic growth. Market growth in the green energy sector, in terms of input factors and demand expansion, has been notable, but recent increases in foreign competition, rapid technological development, and wavering government support put these achievements at risk. Efficient coordination among renewable energy producers and providers may become critical to the viability of this emerging sector. Coordination between upstream markets and downstream suppliers is more likely to occur in localization economies characterized by the concentration of firms in a particular location.

Geographic concentration of firms that could play a role in the development and sustainability of the renewable energy sector may provide cost savings by solidifying upstream and downstream supply chains, pooling skilled labor, or increasing access to raw materials. A number of analytical tools have been developed, which are capable of quantifying the influence of local factors on firm location decisions and the distribution of employment. These methods are applied to characterize the location patterns of firms making up the so-called "green energy" sector.

This research proposes a two-stage procedure which uses global (industry-level) and local (county-level) measures of industry concentration to analyze the geographic distribution of the businesses and jobs supporting the renewable energy sector. Results suggest differences among value chains in terms of firm location patterns and the availability of local resources. Metropolitan counties tend to exhibit comparative advantage with respect to attracting firms

belonging to the residential solar energy value chain, perhaps due to the cost savings from access to skilled labor. The relatively even distribution of feedstock and labor for firms supporting the biodiesel sector (e.g. production, distribution, financing) suggests an array of locations may be suitable to firms making up this value chain. This energy sector may be of a relatively more footloose nature compared to the residential solar value chain.

## Introduction

Growing concern about the sustainability of energy production from fossil fuels has mobilized significant resources towards the development of technologies broadly defined as renewable, or "green", energy sources. Gillingham and Sweeny (2010) suggested that market failures resulting from society's dependence on fossil fuels have been an important impetus for recent policies promoting the development of alternative energy sources through a windfall of public and private investment dollars. The renewable energy sector is also expected by many to be an important catalyst for rural economic development, building on the 8.5 million jobs and $970 billion in revenue contributed by related businesses in 2006 (American Solar Energy Society 2009). Economic impact analyses have projected substantial gains in regional industry output and employment from renewable energy investments. For instance, Jensen et al. (2010) found that growth in the biodiesel sector could create over 3,000 jobs and increase economic output by over $1.2 billion in the Appalachian Region. Other emerging energy sectors, including solar, ethanol, and dairy methane and landfill gas, were projected to support as many as 28,000 jobs across the region, with increases in regional economic output ranging between $1 million and $12 billion. National impacts from bioenergy production expansion also have the potential to increase farm income by $180 billion by 2025, and $700 billion for the national economy (English et al. 2006). But according to Gillingham and Sweeny (2010), green energy may not be any more effective than other industries at stimulating economic growth (in terms of jobs or revenue). However, the nation's demand for stable and sustainable energy sources positions the renewable energy sector as an ideal candidate for this role.

Despite the potential advantages associated with emerging renewable energy sources, substantial challenges remain amidst uncertain federal support, technological advances, changing industry cost structures, and competition abroad. Favorable policies and low labor costs have contributed to the expansion of green sector industries in China, Brazil, and the United Kingdom, making these countries the primary US competitors. Domestic solar panel producers, for instance, have experienced a sharp drop in prices resulting from oversupply from China and weakened demand in Europe (Spegele 2011; Sweet and Chernova 2011). According to a 2010 report by the Pew Charitable Trusts, in 2010 the United States ranked second (behind China) in terms of investing in "clean energy", and sixth worldwide in terms of the five-year growth rate in green energy investments. The US produced about 4.6% of the world's supply of solar photovoltaic panels, compared to 45% for China.

The effects of waning support for government subsidies and tax credits are also impacting firms engaged in green energy value chains. Tighter state and federal fiscal policy has been a significant factor in downsizing the breadth of renewable energy policy. The impending expiration of tax credits for wind farms, biofuel production, and other energy initiatives paints a less optimistic outlook for industries engaged in the renewable energy sector (Pierobon 2011). Anecdotal also evidence suggests that tax credits have enabled US firms to compete with established energy sources (e.g. coal, oil, and natural gas), but their expiration prior to strengthening value chains may erode what progress has been made (Elliot 2012; Jakab 2012; Tracy 2012). Thus far, the price-cost squeeze faced by renewable energy industries has largely been met through technological innovations. Firms unable to keep pace with innovations are often at a disadvantage. For example, innovation and overseas competition were cited as the primary factors of the recent bankruptcy of Evergreen Solar, Inc. (Gold 2011). Solyndra Corporation, well known for its September 2011 bankruptcy, was unable to maintain a competitive advantage with respect to production practices in light of falling silicon prices.

Federally backed loans of $500 million and an estimated $24 million in tax breaks were not enough for the company to maintain solvency (Milbourn 2011). With the average 2010 price per megawatt hour of retail electricity consumption around $98 (US EIA 2011), renewable energy manufacturers have had difficulty competing with energy produced using established methods (e.g. coal and hydroelectricity). Additionally, though renewable energy production costs appear to be falling, reports suggest that they may be 40% to 50% higher than conventional energy technologies (Denning 2011; Wald 2009). Substitution effects from recent declines in natural gas prices (a result of weakened domestic demand and production surpluses) discourage adoption of alternative renewable technologies (Kahn 2012). Furthermore, agricultural commodity markets impact biofuel markets. For example, low soybean yields in Argentina, due to adverse weather conditions, diminished stocks of soy products, and increased demand from the US and China have contributed to soybean price increases (Romig 2012). As a result, the feasibility of soy-based biodiesel production is diminished.

In addition to these external factors, the economic sustainability of emerging renewable energy sectors will be determined by the ability of firms to coordinate and manage input supply logistics, distribution, financing, and sourcing skilled labor in domestic markets. Coordination among businesses engaged in these emerging energy sectors will be influenced by geographic proximity of selected sites to upstream and downstream input and product markets. Like other supply-oriented firms, renewable energy industries and value chains may also realize benefits from access to transportation and specialized labor, or from knowledge transfers resulting from interactions with nearby firms, academic institutions, or government research facilities. Location-specific characteristics including workforce attributes, local industry specialization, and innovation from competition contribute to the uniqueness of regions (Malecki and Varaiya

1986). These so-called "natural advantages" specific to a location, all else equal, co-determine the site selection decisions of other firms engaged in similar production processes. The role of geography in firm location decisions and industry clustering is observed as variation in the spatial distribution of firms, or more generally recognized as industry concentration. Analysis of firm concentration patterns provides insight into which natural advantages of a location play a role in attracting investment, which in turn contributes to the growth and sustainability of different value chains. Furthermore, quantifying firm location patterns and industry clustering maybe useful for identifying which locations exhibit an optimal mix of natural advantages, or comparative advantage, to the extent that the combination of these factors attract firms belonging to a particular renewable energy value chain.

Recent advances in the measurement of the industry concentration provide regional economists with new methods to gauge the influence of agglomeration economies on firm location behavior and business establishment growth. Model-based indexes have linked industry concentration with firm location decision theory, resulting in measures that are able to statistically differentiate systemic versus random patterns of industry concentration (e.g. Ellison and Glaeser 1997; Guimarães, Figueiredo, and Woodward 2007, 2009; Maurel and Sédillot 1999). These approaches identify patterns of spatial dependence associated with external scale economies arising from industry clustering or local natural advantages as evidenced by the geographic distribution of employment. Empirical applications using these methods have largely focused on manufacturing industries in the US and abroad.

This research uses two concentration measures to analyze the influence of geographic proximity on renewable energy industries. A two-step procedure is used to describe firm and employment concentration. In the first step, "global" concentration measures are estimated to describe the degree to which renewable energy activity and firms are regionally concentrated, thus indicating the degree to which the spatial distribution of location determinates (or "natural advantages") influences the site selection decisions of firms belonging to specific renewable energy value chains. The strength of external economies and natural advantages, as measured by the indexes, is tested using a paired bootstrap procedure. Using the global indexes, the local-level dollar value of the cost advantages associated with industry concentration is estimated. The second step uses a "local" concentration index to identify specific locations in a region where firms tend to concentrate (e.g. a county). Local industry concentration indexes are analyzed with respect to a county's typology, including metropolitan, micropolitan, and non-core. This ex-post analysis provides insight into which county types are more likely associated with external economies and harbor natural advantages that may attract renewable energy investment.

The two-stage analysis evaluates firm and employment concentration of ten renewable sectors: (1) biodiesel, (2) coal co-firing, (3) wood direct fire, (4) ethanol production from switchgrass, (5) wood ethanol, (6) landfill methane, (7) dairy methane, (8) commercial solar production, (9) residential solar production, and (10) wind power. The overlap and range of business support structures for these value chains are presented in Figure 2. Concentration indices are estimated for two periods (2002 and 2006), providing a sensitivity analysis with respect to the temporal variability in concentration patterns.

Identification of external economies and natural advantages specific to locations, as indicated by the uneven distribution of firms may be useful to investors and policy makers for identifying low-cost sites. To the extent that attraction and retention of skilled labor, access to upstream and downstream markets, and a well-maintained infrastructure are critical to the sustained growth of renewable energy sectors, state and local policy makers can use this information to forecast the potential impact of local or regional policies designed to attract investment. Results may also be useful to federal policy makers for identifying which sectors might benefit from national policies designed to support or improve local transportation networks, university research capacity, private investment opportunities, or educational funding. Absent knowledge on renewable energy firm location patterns, public and private investment may be formulated such that the cost advantages correlated with geographic concentration are underestimated and underexploited.

### Data Requirements and Considerations

Global and local concentration indexes are estimated using employment and business establishment data. Business establishment data is from the US Census Bureau's County Business Patterns (CBP) database. This data has the advantage of being arranged by county divisions and by six-digit North American Industry Classification System (NAICS) codes. The analysis examines green energy industry concentration for all counties in the contiguous United States ($J$ = 3,078)[4]. However, employment data with similar detail is difficult to obtain. Due to non-disclosure rules, CBP data lacks sufficient detail to facilitate disaggregated employment pattern analyses. Researchers are often compelled to search elsewhere for employment data or conduct analysis at highly aggregated levels. Methods exist whereby missing employment data can be imputed. For example, WholeData.net, the Minnesota IMPLAN Group[5], and other companies use proprietary algorithms to estimate missing employment information, offering enhanced datasets to the public (though generally at significant cost). This study uses employment data for 2002 compiled by WholeData.net. Employment data for 2006 are from the Minnesota IMPLAN Group database. Both data sets allow for highly disaggregated analysis of

[4] In 2002 and 2006, the contiguous United States consisted of 3107 counties and county-equivalents. This analysis excludes Washington, D.C., and combines most independent cities in the state of Virginia with nearby counties.
[5] Note that IMPLAN datasets are arranged by an alternative sector specification to the six-digit NAICS codes. However, a key exists by which the sector codes can be matched to facilitate cross-database analysis.

employment patterns at the county level. Finally, metropolitan, micropolitan, and non-core county classifications from the US Census Bureau's definition files are used to extend location-level analyses to county typology.

Use of employment data from two sources (WholeData.net and IMPLAN) results in some complications in terms of comparison across years. Ideally, in multi-year regional analysis, research would make use of data compiled using the same methodology. However, access to consistent employment data presents further challenges. Employment data from WholeData.net are enhanced versions of CBP datasets, thus representing employment captured at a single point in time. Furthermore, the CBP excludes self-employed individuals, agricultural production, railroad, and government employees, and employees of private households. Employment data from IMPLAN, on the other hand, are annual average estimates created from multiple datasets including the CBP, the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) datasets, and the Bureau of Economic Analysis' Regional Economic Information System (REIS) database. In spite of these differences, data from both sources generally aggregate to observed national and state employment levels. The analysis proceeds by acknowledging some incongruence between the 2002 WholeData.net and 2006 IMPLAN data sets, formulating inter-year conclusions based on these caveats. With respect to the global indexes used to determine industry concentration, analysis of changes should not be too problematic because the estimated indexes represent industry national averages. For the local concentration measures estimated in the second stage, data incongruence is more problematic.

## Industry Concentration and Regional Comparative Advantage

Economic base analysis suggests that industry concentration signals that a specific location is an exporter of some good (Isserman 1977). Firm concentration and industry clustering has also been linked to economic growth (Barkley and Henry 1997; Baldwin and Martin 2004; Brakman et al. 2009). But fostering industry clustering may not necessarily result in economic growth (Feser, Renski, and Goldstein 2008). Nevertheless, policies designed to support business establishment or employment growth through "clustering" are still pursued by some local and regional policy makers. Concentration measures are often used by applied researchers to determine the degree to which a location specializes in the activities of an industry or clusters of industries (Holmes and Stevens 2004). Additionally, local concentration may suggest that a site exhibits comparative advantage through low-cost advantages with respect to attracting, supporting, or retaining firms belonging to a particular industry (Shaffer, Deller, and Marcouiller, 2004).

Factors influencing firm location decisions have been described as "first" or "second" nature (Cronon 1991). First nature factors are associated with advantages or resources assigned by nature (e.g. soil fertility). Second nature factors are linked to society's desire to improve upon the first (Fujita and Thisse 2002; Ottaviano and Thisse 2004; Brakman 2009). Lösch's and Christaller's foundational work on central place theory, von Thünen's early exploration into the role of transportation costs in the distribution of agriculture, Weber's theories on the role of transport costs in economic agglomeration, and Isard's work on industry structure and firm location were among the first to explain the relationship between local "natural advantage" factors and the geographic distribution of economic activity. Built upon these and other foundational works, recent applications of firm location theory suggest that imperfect factor mobility, in the presence of, say, transportation costs or variation in the distribution of skilled labor, endogenize second nature concentration factors. Advantages from second nature factors often manifest as Marshallian external economies of scale: (i) knowledge spillovers, (2) labor-market pooling, and (iii) access to specialized inputs. The relationship of first and second nature factors with geographic concentration of firms and employment is typically explained in terms of increasing returns to scale for firms resulting from technology sharing, labor-job matching, and learning. For example, concentration may facilitate common access to transportation networks, reduce labor-search cost through job market matching, and modify existing technology or production practices from inter-firm knowledge spillovers (Duranton and Puga 2004). Industry structure (e.g. product diversification, degree of vertical integration, average plant size) and workforce characteristics have also been associated with firm-employment concentration and the reinforcement of external economies (Kim, Barkley, and Henry 2000). Industry concentration, therefore, suggests that the presence of local-specific natural advantages gives rise to external economies which in turn reduces variable costs. Localization of firms sends signals to other firms that a location has comparative advantage with respect to a specific industry, and the process continues.

Industry concentration may be more common than expected. Domestic examples include carpet manufacturers (Dalton, Georgia), software development (Seattle), high-tech industries (Silicon Valley, California), dress manufacturers (New York City), and the automobile manufacturing sector in Michigan and Tennessee (Ellison and Glaeser 1997; Holmes and Stevens 2002; Guimarães, Figueiredo, and Woodward 2007, 2009, 2011; Feser and Isserman 2005). Similar findings have been reported for manufacturing sectors elsewhere, especially in the

European Union (Maurel and Sédillot 1999; Lafourcade and Mion 2007; Brülhart and Traeger 2003).

Locations exhibiting comparative advantage with respect to attracting renewable energy firms may be endowed with a wide range of natural advantages (see Figure 3 for a map of solar resource potential and Figure 4 for biomass resources for the biofuel sector). Supply-oriented firms generally locate near areas with abundant resources for energy production (Lambert et al. 2008; Stewart and Lambert 2011). For example, in response to costly, over-land transport of feedstock, biodiesel and ethanol firms tend to locate near feedstock sources – generally oil seed and cellulosic crops and waste products near agricultural locations, or yellow grease (in the case of biodiesel) from urban areas (Jensen, Menard and English, 2007). Similar trends in agricultural regions have been observed for other biofuel sectors using of agriculture and forest raw materials including the cellulosic ethanol, wood direct firing, and animal-waste methane sectors (Schjeldahl 2012). On the other hand, while solar and wind energy value chain firms may find advantages from locating near windy or sun-soaked areas (e.g. the Great Plains, or the southwest), there may also be impetus for some firms in this sector to locate in areas where photovoltaic cells or wind turbine components are produced (which may require specialized labor skill sets), even when access to raw materials (solar radiance or wind) is limited.

Second nature factors that could encourage a site to be selected by firms participating in a renewable energy sector include access to supply or demand markets, skilled labor pools, or transportation networks. Consider the case of Suntech Power's new solar panel manufacturing facility in Goodyear, Maricopa County, Arizona, which opened in 2011. Access to solar radiation was an indirect factor in the company's location decision, as the site was strategically situated 30 miles from the factory's primary downstream market, a utility-scale solar farm

(Cheyney 2011). Another example is the case of REG Seneca Biodiesel in Grundy County, Illinois. The company's biodiesel plant, which opened in 2010, takes advantage of nearby input markets for oilseeds, yellow grease, and animal fats. The plant benefits from immediate proximity to a truck and rail transport, as well as access to skilled labor (REG 2010). Additionally, incentive packages may also weigh heavily on firm site selection decisions. For example, Arizona's Renewable Energy Tax Incentive program has provided a critical advantage in attracting solar energy firms to that state (Bosman 2011). Tennessee's Volunteer State Solar Initiative has fostered research and innovation beneficial to firms engaged in solar value chains in the state through the Tennessee Solar Institute, and has encouraged investment through the program's Solar Innovation Grants (Tennessee 2011). Recently, the West Tennessee Solar Farm began operation, providing electricity to more than 500 residents in Haywood County.

Regional scientists have struggled to develop methods whereby the effects of external economies and natural advantages can be quantified because differentiation among the factors associated with concentration is often difficult and many are unobservable (Brakman et al. 2009). Hoover's (1936) analysis of geographic concentration of the shoe and leather industries used a Coefficient of Location to describe the relationship of employment relative to population. Florence (1939) introduced a location quotient (LQ) to measure industry concentration, comparing relative employment shares between different aggregations of industries across regions. Isserman (1977) presented a theoretical justification for Florence's location quotient, using the LQ to estimate the regional economic impacts from basic (i.e. exporting) firms. Krugman (1991) suggested using Gini coefficients to measure the intensity of firm and employment distributions in the US textile manufacturing industry.

The theoretical models justifying firm and employment concentration measures have advanced since Hoover and Florence, but only recently have these indexes been tied to statistical distributions, which provide testable hypotheses about the strength of the relationship between local characteristics and concentration. For example, Ellison and Glaeser's (1997) "dartboard model" developed an industry concentration index (the EG index) to identify concentration resulting from external economies and natural advantages, rather than random clustering of firms. Maurel and Sédillot (1999) developed an index to measure industry concentration, which was similar to the EG index. Later, Guimarães, Figueiredo, and Woodward (2007) developed alternative derivations of the EG index, measuring the intensity of localization economies using business establishments rather than employment. While these *global indexes* supplement the description concentration at the *industry* level, local measures building on Florence's (1939) location quotient are still commonly used to identify *where* industries concentrate. Derived from a dartboard model similar to Ellison and Glaeser's (1997), location quotients lead to similar conclusions underlying the patterns suggested by the EG index at finer levels of geographic detail (Figueiredo, Guimarães, and Woodward 2007; Guimarães, Figueiredo, and Woodward 2009). The EG index and its modifications reinforce the theoretical role of external economies and natural endowments with respect to firm location decisions, while facilitating comparisons between dissimilar industries across time and spatial units.

## Conceptual Model

### *Global Measures of Industry Concentration*

Carlton's (1983) application of McFadden's (1974) Random Utility Model to industrial location served as the starting point of Ellison and Glaeser's (1997) global concentration index, derived from a model wherein a location's natural advantages encourage firms to invest in that site. Ellison and Glaeser called their model the "dartboard model" because of the index's ability to identify concentration over and above random site selection by firms. They imagined a process whereby firms select sites similar to darts thrown at a dartboard. Some firms concentrate in a location by chance alone, but others may demonstrate a degree of interdependence such that their representative "darts" (decisions) are joined together before being thrown. In other words, firm returns (or costs) have the potential to be tied to a set of similar or related firms to the extent that site selection decisions appear to be made jointly. Firms choose sites to maximize profit (or minimize costs). Firm profits are

$$\log \pi_{ijk} = \log \tilde{\pi}_j + \eta_{jk} + \varepsilon_{ijk}, \tag{1}$$

where $\tilde{\pi}_j$ is the expected profit from choosing location *j* for all *i* firms in industry *k*; $\eta_{jk}$ is a parameter describing the strength of observed and unobserved location characteristics on site selection decisions including external economies and (or) natural advantages specific to location *j* for industry *k*; and $\varepsilon_{ijk}$ is a disturbance that reflects factors idiosyncratic to firm *i*.[6] Ellison and Glaeser introduced the following restrictions such that the expected probability ($p_{jk}$) of locating in region *j* is

$$E(p_{jk}) = \frac{\tilde{\pi}_j}{\sum_j \tilde{\pi}_j} = \frac{e_j}{e} = x_j \tag{2}$$

with variance

$$V(p_{jk}) = \gamma_k x_j (1 - x_j), \tag{3}$$

where $e_j$ is overall employment in area *j*; $e$ is overall (national) employment; and $x_j$ is area *j*'s share of employment in a reference distribution (EG's (1997) study broadly focused on the SIC

[6] EG (1997) derived their index from a model emphasizing natural advantages, but also specified an observationally equivalent alternative emphasizing spillovers. Because the source of excessive concentration cannot typically be identified, the particular motivation chosen for the model is assumed not to affect the results.

20-39 manufacturing sector). This assumption suggests that the expected distribution of industry employment (if firm location is purely random) should equal the distribution of overall firm location activity across all renewable energy industries, which can be described as a binomial random variable. The greater the difference between $x_j$ and $p_j$, the greater the influence localization economies and natural advantages have on firm location decisions. The difference is moderated through $\gamma_k$, which is interpreted as the degree to which industry *k* would locate beyond the level that would be expected from purely random site selection. Finally, Ellison and Glaeser (1997) proposed that the degree to which firms in industry *k* are concentrated is measured as

$$\hat{\gamma}_k = \frac{G_k - \left(1 - \sum_{j=1}^{J} x_j^2\right) H_k}{\left(1 - \sum_{j=1}^{J} x_j^2\right)(1 - H_k)}, \tag{4}$$

where the "raw" industry concentration index is $G_k = \sum_{j=1}^{J} \left(s_{jk} - x_j\right)^2$; $H_k = \sum_{i=1}^{c_k} z_{ik}^2$ is a plant size Herfindahl index; $z_{ik}$ is plant *i*'s share of industry employment; $c_k$ is the number of firms in industry *k*; and the variable of interest, $s_{jk} = e_{jk}/e_k$, is the location's share of industry employment.

It is important to emphasize that $\gamma_k$ characterizes industry concentration arising from the combination of natural advantages and external economies, but because it is an employment-based measure, the parameter also accounts for economies of scale internal to individual firms. While the Herfindahl index accounts for variation in the distribution of firm size, $\gamma_k$ primarily reflects external economies realized by larger plants and discounts those realized by smaller plants. Guimarães, Figueiredo, and Woodward (2007) observed this shortcoming of the EG employment-based index and developed an alternative index based on plant-counts (the GFW index). An advantage of this modified EG index is its ability to adjust for "lumpiness" or variation in establishment sizes that could arise when employment is used to measure establishment concentration. The GFW index, therefore, moderates the influence of scale economies internal to individual firms. Furthermore, the GFW index tends to have a lower variance compared to the EG index in simulation studies (GFW 2007). The reformulated EG index based on plant counts is

$$\hat{\gamma}_{ck} = \frac{c_k G_{ck} - \left(1 - \sum_{j=1}^{j} x_j^2\right)}{(c_k - 1)\left(1 - \sum_{j=1}^{j} x_j^2\right)}, \tag{5}$$

where $G_{ck} = \sum_{j=1}^{j} \left(n_{jk} - x_j\right)^2$; $n_{jk} = c_{jk}/c_k$ is region *j*'s share of plants belonging to renewable energy industry *k*; $c_{jk}$ is renewable energy industry *k*'s establishments in region *j*; and $c_k$ is the renewable energy industry's overall number of establishments in the nation (or entire analysis region). This concentration measure differs from $\gamma_k$ by emphasizing firm counts and replacing the Herfindahl index, $H_k$, with $1/c_k$.[7] Naturally this substitution is based on the rather strong assumption that all plants in a renewable energy value chain contribute equally to concentration.

To illustrate the statistical implication of the concentration coefficients, when $\gamma_{ck} = 0$ ($\gamma_k = 0$), establishment (employment) concentration is purely random. On the other hand, when $\gamma_{ck} = 1$ ($\gamma_k = 1$), it is expected that all establishments (employment) in a renewable energy industry would locate in a single region. In their study, Ellison and Glaeser (1997) adopted the convention that concentration estimates less than 0.02 were "not very concentrated" while those greater than 0.05 were considered "highly concentrated". These classifications are obviously subjective, and the magnitudes of parameters associated with different industries should be interpreted relative to other estimates using similar references distributions. For example, conclusions from Ellison and Glaeser's study about the concentration of the Semiconductor and

[7] The derivation of this index's relationship to the EG index is presented in Appendices A and B of GFW (2007).

Related Device Manufacturing industry, where total manufacturing employment is used as the reference distribution, would not be comparable to an estimate for the same industry using national employment as a reference distribution.

***Spatially-Adjusted Global Industry Concentration Measures***

Despite these recent advances in firm concentration measurement theory, shortcomings remain with respect to capturing a complete picture of the effects of external economies on firm site selection. The first problem relates to the "modifiable areal unit problem" (MAUP), which stems from the imposition of arbitrarily defined spatial boundaries on to data which may or may not have been generated as a function of the selected boundaries (e.g. zip code, city, county, state, or MSA boundaries). The MAUP may bias estimates due to this aggregation problem (Openshaw and Taylor 1979). The second complication, the "checkerboard problem", results from the inability of previous industry concentration measures to account for the proximity of areal units (White 1983; Griffith 1983) [8].

In an attempt to deal with the second issue, Arbia (2001) and Lafourcade and Mion (2007) suggested that spatial autocorrelation of economic activity, as measured by Moran's (1950) *I*, could supplement information from industry concentration indexes by permitting inference about the similitude of concentration patterns between neighboring locations. However, as GFW (2011) demonstrated, measures depending solely on spatial autocorrelation statistics may be unreliable indicators of the geographic concentration of firms because they focus on the similarity between regions and do not account for the economic activity within spatial units. Instead, GFW formulated a spatially-adjusted EG index, which accounts for spillovers, firm size effects, and natural advantages spanning a region. The spatially adjusted EG index establishes a

[8] GFW (2011) (Section 2) present a detailed description of this problem by comparing various concentration measures among permutations of economic activity across a hypothetical region.

link between the Moran's $I$ statistic and the aspatial EG index by including a scaling factor estimated using a neighborhood relational matrix, $\mathbf{\Psi} = \mathbf{I} + \mathbf{W}$, where $\mathbf{I}$ is an identity matrix and $\mathbf{W}$ is a matrix identifying neighboring spatial units. In the absence of neighborhood effects, $\mathbf{\Psi}$ is simply the identity matrix and the index collapses to the standard EG index. The estimator of $\gamma_k^s$ (where the superscript $s$ denotes the spatial reformulation of $\gamma_k$), is

$$\hat{\gamma}_k^s = \frac{G_{sk} - H_E(1 - \mathbf{x}'\mathbf{\Psi}\mathbf{x})}{(1 - H_E)(1 - \mathbf{x}'\mathbf{\Psi}\mathbf{x})} \tag{6}$$

where $G_k^s = (\mathbf{s} - \mathbf{x})'\mathbf{\Psi}(\mathbf{s} - \mathbf{x})$, $\mathbf{x}$ is a $J \times 1$ vector of shares of a reference employment distribution (previously, $x_j$); $\mathbf{s}$ is a $J \times 1$ vector of employment shares (previously, $s_{jk}$); and $H_E = \boldsymbol{z}'\boldsymbol{z}$, where $\boldsymbol{z}$ is a $J \times 1$ vector of average establishment size (previously, $z_{ik}$). Additionally, as GFW indicate (Appendix A; 2011), the spatially adjusted EG index can be extended to their firm-count index, $\gamma_{ck}$ (Eq. 5). No additional attention is given to this index in their study, and its derivation is included here.

The spatially adjusted firm-count index, $\hat{\gamma}_{ck}^s$, takes as its starting point the reparamertization of the spatially weighted EG index to a spatially weighted plant count index (justified on the same set of assumptions proposed by GFW (2007)). As with the conventional $\gamma_{ck}$ index, $H_k$ is replaced by $1/c_k$ and the raw concentration index, $G_{sk}$ is modified as $G_{ck}^s = (\mathbf{n} - \mathbf{x})'\mathbf{\Psi}(\mathbf{n} - \mathbf{x})$, where $\mathbf{n}$ is a $J \times 1$ vector of industry establishment shares (previously, $n_{jk}$). The spatially weighted plant count index follows

$$\hat{\gamma}_{ck}^s = \frac{c_k G_{csk} - (1 - \mathbf{x}'\mathbf{\Psi}\mathbf{x})}{(c_k - 1)(1 - \mathbf{x}'\mathbf{\Psi}\mathbf{x})}. \tag{7}$$

*Optimal Bandwidth Selection for Determining the Geographic Extent of Concentration*

The potential for interregional spillovers, included in the spatially adjusted global indices, is mediated through $\boldsymbol{\Psi}$. Alternative schemes of spatial weighting exist and there is no agreed-upon method to capture all types of spatial interactions (Anselin 1988). Therefore, the neighborhood relational matrix used to formulate $\boldsymbol{\Psi}$ is a data driven approach, instead of an *a priori* specification.

GFW (2011) proposed determining optimal neighborhoods by examining the changes in the index as additional spatial units are introduced in $\boldsymbol{\Psi}$ by the increasing distance between them. Interpreting the results of this procedure as it relates to interregional spillovers provides some insight about the geographic horizon of influence natural advantages may have on firms site selection for a given renewable energy industry. Using county level data, the distance (in miles) at which the ratio of the spatially adjusted index to the conventional index (i.e.$\hat{\gamma}_{ck}^{s}/\hat{\gamma}_{ck}$ or $\hat{\gamma}_{k}^{s}/\hat{\gamma}_{k}$) reaches a maximum is interpreted as the range within which spillovers from external economies or natural advantages influence site selection. In other words, the optimal bandwidth is the average maximum distance at which localization economies influence firm profits. Beyond the threshold, the influence of spillovers on the concentration index diminishes. In this research, the bandwidth procedure is applied to the employment- and establishment-based measures of global concentration to determine the maximum distance at which local factors contribute to concentration. All spatially adjusted global concentration indexes are reported and the relevant hypotheses evaluated at the optimal bandwidth distances, $D_{k}^{*}$ and $D_{ck}^{*}$.

***Local Measures of Industry Concentration***

Significant global indexes of firm concentration warrant further analysis of firm location and employment distribution at a finer spatial resolution. Given significant global industry concentration, a more detailed analysis of these patterns proceeds using location quotients. The standard location quotient (LQ) compares the proportion of employment in a particular industry in a region with the proportion of employment in that industry across all regions (typically, the nation);

$$L_{jk} = s_{jk}/x_j, \tag{8}$$

where $L_{jk}$ is the LQ for renewable energy industry $k$ in location $j$, and $s_{jk}$ and $x_j$ (the reference distribution) are described above. It is generally assumed that when the LQ is greater than one, industry $k$ is concentrated in location $j$.

GFW (2009) extended EG's (1997) dartboard model to formulate a more flexible derivation of the LQ. Under certain assumptions, their derivation is observationally equivalent to (8). They express the LQ as an estimator of the unobservable external economies and (or) natural advantages in location $j$ according to EG's dartboard model. Similar to EG (1997), GFW (2009) assume that the spatial distribution of industry activity replicates the distribution of overall economic activity, using the restrictions in (2) and (3). Location event can be expressed as probabilities, given Eq. (1). The likelihood of observing a particular distribution of firms is then constructed as the product of all location probabilities weighted by a factor of $w_{jk}$, where $w_{jk} = s_{jk} \times\ c_k$, such that

$$l_k = \prod_{j=1}^{J} p_{j|\eta_k}^{w_{jk}} = \prod_{j=1}^{J} \left( \frac{x_j \exp{(\eta_{jk})}}{\sum_{j=1}^{J} x_j \exp{(\eta_{jk})}} \right)^{w_{jk}}. \tag{9}$$

Maximizing $l_k$ with respect to $\eta_{jk}$ and solving the first order conditions[9], it can be shown that $\hat{\eta}_{jk} = \log L_{jk}$, where $L_{jk}$ is the LQ for industry $k$ in region $j$,

[9] GFW (2009) introduce a restriction on $l_k$ requiring that $\sum_{j=1}^{J} x_j \exp{(\eta_{jk})} = x$ to resolve the resulting indeterminate solution, since the $\eta_{jk}$'s are unobservable.

$$L_{jk} = \left(\frac{w_{jk}}{w_k}\right)/x_j \,, \tag{10}$$

and $w_k$ is the sum across regions of all $w_{jk}$'s in the region. As with the global concentration indexes, employment-based location quotients are unable to differentiate localization economies from internal economies of scale. Thus, Figureido et al. (2007) suggest an alternate weighting factor, $w_{jk} = c_{jk}$, which permits the formulation of an establishment-based location quotient derived from the same model.

The advantage of linking the standard LQ with the likelihood function of (9) provides a statistical framework for testing the null hypothesis: LQ > 1. GFW (2009) construct Wald statistics to test for the presence of localization economies influencing firm or employment concentration, basing their hypotheses on the $\eta_{jk}$'s. The first statistic tests if localization economies are present in a region. Following GFW (2009), a Wald test statistic is estimated as

$$W_{jk} = \frac{J[logL_{jk}]^2}{(J-2)w_{jk}^{-1} + \overline{\overline{w_k^{-1}}}}\,, \tag{11}$$

which is asymptotically distributed as a $\chi^2$ variate with one degree of freedom. The null hypothesis is $\eta_{jk}$= 0; location-specific effects are not associated with concentration and thus, the industry is not localized in spatial unit *j*. Rejection of the null hypothesis suggests that the industry exhibits concentration resulting from location-specific advantages rather than random site selection which might exhibit concentration patterns. In addition to this test of regional localization, a test of equality of location-specific factors between two regions, $l$ and $j$, $\eta_{jk} = \eta_{lk}$, with a Wald statistic

$$W = \frac{\left[\log\left(L_{jk}\right) - \log(L_{lk})\right]^2}{w_{jk}^{-1} + w_{lk}^{-1}} \tag{12}$$

distributed as a $\chi^2$ variate with one degree of freedom. This test may be useful in analyses similar to this, in which multiple data years are explored. For example, the statistic can be reformulated in terms of the same location, $j$, at two time periods. The reformulated hypothesis tests the differences between the concentration indexes between periods (e.g. 2002 and 2006): $\eta_{jk2006} = \eta_{jk2002}$. Inferences can be made about changes in location specific affects, demonstrated by increases or decreases in firm or employment concentration. This test is discussed here to bring attention to its potential application, though it is not implemented in this research due to previously discussed differences in data used for 2002 and 2006.

## Methods and Procedures

A two step approach is used to determine: (1) *which* industries tend to globally concentrate across the nation, and of those industries, (2) *where* they are concentrated. In the first step, global concentration of a renewable energy industry is analyzed using the EG and GFW indexes. In step two, industries exhibiting concentration from step one are analyzed using the employment and establishment based location quotients.

Geographic concentration indexes are subject to a wide range of assumptions and different applications. For instance, EG's (1997) index was developed to measure the concentration of industry employment, whereas GFW's (2007) plant-count index was formulated to measure establishment concentration. Furthermore, these measures were developed to express concentration in relative terms. The type concentration measured also depends on the reference distribution chosen by the researcher. To illustrate, recall that EG (1997) applied their index to analyze manufacturing concentration. Geographic concentration of a particular manufacturing industry (defined by the county's share of industry employment, $s_{jk}$) was estimated relative to the county's share of all *manufacturing* employment, $x_j$ (defined by the 2-digit SIC level).

Though not an arbitrary choice, it would be equally admissible to measure the concentration of a particular manufacturing industry using a subset of manufacturing industries (say the 3-digit level NAICS), or all US industries as a reference distribution. It is clear that the selection of a reference distribution is by and large the choice of the researcher, keeping in mind that evidence of concentration is only meaningful relative to the reference distribution selected. In this analysis, the strength of factors associated with the concentration of renewable energy firms is compared across all 60 member industries identified in Figure 2. Estimates indicate the relative strength of the uneven distributions of local factors attracting firms to certain locations. In other words, global indexes larger in magnitude may suggest greater proclivity of firms being attracted to factors found in relatively few locations (e.g. ethanol firms concentrating near agricultural areas specializing in grain production).

***Step 1 – Measuring Global Industry Concentration***

In the first step, employment concentration is estimated using Equations (4) and (6). Establishment concentration is estimated using Equations (5) and (7). Neighborhood relational matrices for (6) and (7) are constructed using county level, population weighted centroids based on the 2000 US Census Bureau definition. Therefore, neighborhood definitions are based on the distance between population centers rather than geographic centroids that lack economic or demographic context.

*Statistical Tests for Global Measures and the Nonparametric Bootstrap Procedure*

EG (1997) and GFW (2011) assess the significance of their indexes using variance estimates based on second moment approximations. However, an alternative variance estimator would be based on a bootstrap resampling procedure. An advantage of this procedure is that it is robust to misspecification that may result from assumptions about the distribution of the data generating process. To illustrate, Greene's (2000) description of this procedure is modified to the spatial plant-count index described in (7). Assume that $\hat{\gamma}_{ck}^{s}$ is an estimate of $\gamma_{ck}^{s}$ based on vectors $\mathbf{n}$ and $\mathbf{x}$ ($\mathbf{s}$ and $\mathbf{x}$ if applying an employment-based index). The bootstrap procedure approximates the standard error for $\gamma_{ck}^{s}$ by resampling all observations, with replacement, from $\mathbf{n}$ and $\mathbf{x}$ and recomputing $\hat{\gamma}_{ck}^{s}$ with each bootstrapped sample. After $B$ times, variance is computed from vector $\hat{\boldsymbol{\gamma}}_{ck}^{s} = [\hat{\gamma}_{ck}^{s}(1), \ldots, \hat{\gamma}_{ck}^{s}(B)]$ as $\hat{\sigma}_{k}^{2} = \frac{1}{B-1}\sum_{b=1}^{B}[\hat{\gamma}_{ck}^{s}(b) - \hat{\gamma}_{0}^{s}]'[\hat{\gamma}_{ck}^{s}(b) - \hat{\gamma}_{0}^{s}]$ and $\hat{\gamma}_{0}^{s}$ is the average of the bootstrapped concentration indexes. In this procedure, $B$ = 1000. Tests at the 1% level are conducted by finding the upper and lower 0.05% quantiles from the vector $\hat{\boldsymbol{\gamma}}_{ck}^{s}$ and constructing 99% confidence intervals. The null-hypothesis of $\gamma_{ck}^{s} = 0$ can be tested, rejecting if the lower bound of the confidence interval is greater than zero, suggesting that observed concentration is nonrandom.

*Uncertainty About Local Cost and Output Differences Associated with Concentration*

Determining the magnitude of concentration describes the impact localization economies could have on costs and eventually firm site selection. However, inference may lack a cost-savings perspective (the cornerstone of firm location theory) if the indexes are not associated with potential cost savings arising from scale economies (EG 1997). Ellison and Glaeser proposed an extension of their EG index to relate the magnitude of firm concentration to the effects of temporal cost shocks on firm location decisions in terms of a "birth-cost" elasticity. In this research, the relationship between changes in costs, defined by local labor income (in dollars) and the number employees in a sector can be deduced from an industry level employment-cost elasticity, $\bar{E}_{k}$, derived from the findings of Jensen et al. (2010) in their analysis of renewable energy employment impacts.[10] Therefore, concentration may indicate the degree to which differences in costs incurred by firms in geographically diffuse locations influence site selection decisions.

Assume that $\bar{E}_k$ is the industry elasticity of local employment ($emp_j$) differences with respect to changes in cost ($cost_j$) between locations. Assume also that differences in employment levels across the nation may reflect each county's propensity to exhibit excess concentration of firm type $k$, such that

$$\bar{E}_k = \frac{\partial emp_j / emp_j}{\partial cost_j / cost_j} = \frac{\sigma(p_j)/E(p_j)}{\sigma(cost_j)/E(cost_j)} = \frac{CV(p_j)}{CV(cost_j)}, \tag{13}$$

where $CV(p_j)$ and $CV(cost_j)$ are the coefficients of variation for the employment and cost distributions (both of which are positive). Thus, the ratio of the coefficients of variation is set to the employment-cost elasticity. Ellison and Glaeser's (1997) restrictions (Equations (2) and (3)) are introduced, redefining the moments of $p_j$ in terms of the observable employment levels, where $\sigma(p_j)$ is the square root of (3). Rearranging (13), it can be shown that

$$CV(cost_j) = \left[\bar{E}_k(x_j)\right]^{-1}\left[\sqrt{\gamma_k x_j (1 - x_j)}\right], \tag{14}$$

where larger values of $CV(cost_j)$ indicate relatively greater uncertainty about costs associated with firms locating in county *j*. It is assumed that concentration will be greatest in locations more likely able to provide cost savings to firms through "natural advantages". Therefore, the relative costs associated with firms locating in areas exhibiting little or no concentration (i.e. no comparative advantage) are more difficult to discern when compared to locations wherein firm concentration is evident.

[10] The cost elasticity is estimated as the projected percent change in direct jobs in the economy divided by the projected percent change in labor income in the economy (in dollars).

It is assumed that least-cost locations are associated with external economies of scale and, thus, permit greater efficiency in production among local firms. Greater efficiency may lead to increases in total industry output (in dollars) for firms of type *k*. The relationship between geographic differences in local firm output and concentration, then, can be expressed similarly to (14), such that

$$CV(out_j) = \left[\overline{EO}_k(x_j)\right]^{-1}\left[\sqrt{\gamma_k x_j(1-x_j)}\right], \qquad (15)$$

where $CV(out_j)$ is the coefficient of variation for local output, and $\overline{EO}_k$ is a job-output elasticity using from Jensen et al. (2010) [11]. Local values of $CV(out_j)$ can be interpreted as the uncertainty in total, county-level output associated with excess concentration of employment in green energy sector *k*.

Based on the relationships in (14) and (15), uncertainty about costs and output are expected to be greater for industries with higher global concentration indexes. If an industry is concentrated, there are relatively fewer locations where investors can expect costs to be lower. In response, they may choose sites where expected cost can be reasonably assumed. Firms in an industry that not exhibiting a tendency to concentrate (i.e. $\gamma_k$= 0), however, are expected to have perfect information about expected costs at any given location; that is, costs and profits are not geographically dependent and agglomeration economies are irrelevant. If only one region is profitable (in other words, higher costs are incurred in all other regions to the extent that firms are deterred, and thus $\gamma_k$= 1), then $CV(cost_j)$ = 0 (costs differences are known with certainty). To summarize the relationship between cost risk, concentration, and the local share of employment in all industries, as $\gamma_k(x_j)$ approaches one, holding $x_j$ ($\gamma_k$) constant, firms in

[11] The output elasticity is estimated as the projected percent change in direct jobs in the economy divided by the projected percent change in total industry output in the economy (in dollars).

industry $k$ are subject to greater (less) uncertainty regarding geographically dependent returns. Similar interpretations hold in the case of output. For example, significant industry concentration suggests that firms experiencing cost savings associated with localization economies may also be more efficient as evidenced by increased value added to the industry.

Because firms may still experience idiosyncratic and market risks (e.g. management decisions, foreign competition, weakened demand, or other shocks not related to a firm's location within the US), these *CV* measures only describe the risk that may be associated with particular locations. As such, interpretations should be considered under *ceteris paribus* conditions. Nevertheless, locations associated with less risk (in terms of lower variable costs) may exhibit comparative advantage with respect to firm site selection decisions. In other words, firm concentration in a region may signal to other investors that potential costs savings are higher (on average) than those found in a region where firms do not concentrate.

***Step 2 – Local Measures of Industry Concentration***

Local measures are constructed by evaluating Equation (10) using the employment and establishment weighting schemes. Thus, the analysis of renewable energy value chains exhibiting global employment or business establishment concentration is extended in the second step. Relevant hypotheses about the strength of external economies are evaluated using equation (11). Specifically, counties where significant instances of localization are identified suggest the presence of external economies (and potentially cost savings) for a particular industry in that location ($\eta_{jk} > 0$). This frames the hypothesis about which counties exhibit comparative advantage with respect to attracting certain firm types of a renewable energy sector. The *p*-values of this test are mapped using ESRI's ArcMap software for the contiguous US to identify counties exhibiting statistically significant concentration.

*An Aggregate Location Quotient*

The relative distribution of firm concentration across the county typology are described using an aggregate location quotient constructed as

$$q_{kt} = \frac{\sum_1^{J_t} \tilde{J}_{jkt}}{\sum_1^{J} \tilde{J}_{jk}} \Big/ \frac{J_t}{J}, \tag{16}$$

where $\tilde{J}_{jkt}$ is equal to one if firms exhibits significant concentration in a county, $J_t$ is the number of counties in typology $t$, and, as before, $J$ is the total number of counties in the US. To illustrate, assume county *j* is classified as a metropolitan county. County level classification as metropolitan, micropolitan, or non-core is assigned based on local population and the degree of access to nearby population centers as defined by commuting ties. Extensive county level concentration of firms in non-core counties relative to metropolitan or micropolitan counties may suggest that cost savings were greatest in these county types, perhaps due to access to cheaper labor or lower land rents. A $q_{kt}$ greater than one suggests that industry *k*'s share of metropolitan counties in which firms concentrate is greater than the national proportion of metropolitan counties; indicating a propensity for firms of the *k*th renewable energy sector to locate in a particular county type. Results from this simple aggregation suggest a relationship between firm location decisions and characteristics unique to these county types.

***Sensitivity Analysis of the Global Indexes Measured in 2002 and 2006***

A sensitivity analysis is conducted to assess differences in global establishment and employment concentration indices between 2002 and 2006. Changes in industry structure over the period may influence the spatial distribution of firms and employment. Differences in the global concentration indexes can be tested for significant change using the paired bootstrap procedure, again using the 99% confidence intervals. Significant differences imply that the tendency toward concentration (positive or negative, relative to the base year) of renewable energy firms has changed between periods.

## Results and Discussion

Summary information about the number of businesses and employees in each renewable energy value chain appears in Table 10. It is important to emphasize that overlap in value chain roles among all industries is reflected in each sector's total. Therefore, the sum of all green energy *sectors'* employment and establishments is greater than the sum of all green energy employees and establishments.

### *Step 1 Results – Global Measures of Firm Concentration*

Tables 11 and 12 (for 2002 and 2006, respectively) present the $\gamma_k$ estimates for both global concentration indexes, along with their respective bandwidths defining neighborhood structure. All of the renewable energy sectors analyzed exhibited geographic concentration in 2002 and 2006. Considerable variation in establishment concentration compared to those of employment are noted (Figure 5 summarizes these differences for both years). The residential solar sector exhibited the greatest magnitude in terms of employment concentration in both 2002 and 2006. By comparison, the biodiesel value chain appears at the opposite end of the employment concentration spectrum.

Investigation into the nature of observed residential solar and biodiesel sector firm location patterns are discussed in detail to motivate the utility of these findings. Residential solar production is composed of two concentrated industries (see Table 13 for industry level global concentration results). Substantially higher concentration levels are observed for the Semiconductors and Related Device Manufacturing industry ($\hat{\gamma}_k^s = 0.032$ in 2002 and 0.031 in 2006). This result is expected, given the oft-noted tendency for industries dependent on skilled labor to concentrate (EG 1997). Therefore, it appears that location decisions for firms engaged in the solar panel manufacturing step of the value chain would likely result in clustering around specialized labor sheds, similar firms, and (or) sunny locations. Panel installation firms in the Commercial Machinery Repair and Maintenance industry are likely more footloose, encouraging site selection across a range of locations where populations demand residential solar panels.

Figure 6 plots the spatial and aspatial global concentration ratios across candidate bandwidths, suggesting that the influence of external economies/natural advantages begins to weaken beyond the distance at which the ratio reaches a maximum. In areas where firms concentrate, optimal bandwidths circumscribe the counties within 20 and 33 miles of neighboring population centers. Several implications arise from this finding. First, note that evidence of firm concentration is expected; if the transport costs of solar panels are relatively low, then location decisions based on, perhaps, access demand markets, may not be too important a source of cost savings. Instead, concentration of downstream suppliers may promote cost savings from increasing returns to scale to assemblers because of improved access to specialized labor. However, while bandwidth estimates may reflect access to labor sheds or potential interaction with other firms in the value chain, questions remain about the extent to which increasing returns from concentration may be offset by changing policy and market realities for solar energy. For example, diminished support for renewable energy subsidies and tax credits, increased foreign competition primarily from heavily subsidized Chinese firms, and expiration of US incentives at the end of 2011 (e.g. the Nonbusiness Energy Property Tax Credit) increase risk to investors and communities. While new tariffs have been imposed on cheaper Chinese solar panels (Gordon 2012), the future of a US role in this sector is unclear. Nevertheless, low transport costs and a demand for skilled labor suggest that location decisions of all firms engaged in the residential solar energy sector may be, on average, primarily driven by access to specialized labor pools.

The observed employment and firm concentration patterns of the residential solar value chain contrast the biodiesel value chain. Costly transport of biodiesel feedstock suggests that a relatively greater dispersion of firms may be expected. Extensive availability of restaurant wastes leads firms engaged in biodiesel production using yellow grease to locate in any number of locations across the US. Firms producing biodiesel using agricultural products (e.g. soybeans or canola) may also be relatively footloose, given the distribution of production potential for these crops. Lower global concentration indexes for biodiesel (compared to those of the residential solar sector) reinforce the expectations that this sector is more evenly distributed across the nation than firms supporting the residential solar value chain.

Figure 6 shows the change in the ratio of the spatial to aspatial global employment concentration estimates for the biodiesel sector in 2002. Biodiesel bandwidth estimates based on the neighborhood relational matrices ranged from 29 to 76 miles, suggesting a relatively large area of influence. The maximum bandwidth was 76 miles. Access to external economies appears to provide increasing returns to biodiesel firms across a wider distance than observed for the residential solar production sector. Note that the maximum bandwidth estimate, 76 miles (the distance at which the ratio of the spatial estimate to the aspatial estimate is greatest), was derived from employment concentration. This may reflect the demand for skilled labor juxtaposed with access to agricultural areas, perhaps implying spillover gains due to larger downstream firms across county borders. Biodiesel from agricultural production may, therefore, benefit from labor sheds that span several counties, while minimizing transport costs by locating in closer proximity to oilseed production areas. Biodiesel production near cities may be more likely to use yellow grease, and potentially, has greater access to skilled labor within influence areas. Concentration of firms in the biodiesel sector, therefore, may be primarily driven by labor availability with feedstock transported from a nearby agricultural periphery (biodiesel from oilseeds), or obtained from within the labor-rich region (biodiesel from yellow grease). This scenario appears to be in line with the disbursed nature of overall firm location activity in the biodiesel sector, perhaps characterized instead by *pockets* of concentration.

Application of the relationships defined in (15) and (16) to the renewable energy sectors demonstrates, the link between cost savings, concentration, and comparative advantage. Each renewable energy sector is ranked in terms of uncertainty in costs and output based on global concentration indexes and estimated elasticities (Table 14). Firms in the residential solar sector were the most concentrated (based on the global employment concentration indexes) and were also subject to the greatest variation in costs and output with respect to differences in the local share of all green energy firms. Firms engaged in the biodiesel value chain, on the other hand, were among the most weakly concentrated and, exhibited comparatively low variation in expected local costs and value-added output. From Figures 7 and 8, it is clear that the expected variation in costs and output, respectively, declines for counties exhibiting increasing shares of all firms engaged in all renewable energy industries ($x_j$). However, it is also clear that for any given industry share, firms supporting the biodiesel sector are subject to less uncertainty in terms of cost savings and output than those engaged in the residential solar energy value chain. Therefore, residential solar energy firms seeking least-cost locations systematically associated with external scale economies and value-added output gains are subject to greater uncertainty about potential profits associated with a location compared to firms engaged in biodiesel value chains, *ceteris paribus*. This finding may also indicate that relatively fewer counties exhibited comparative advantage with respect to attracting residential solar energy sector firms that firms belonging to the biodiesel sector.

### *Step 2 Results – Local Measures of Firm Concentration*

Local measures supplement the global concentration indices by identifying where renewable energy concentration is observed. Figures 9, 10, 11, and 12 identify local patterns (observed in 2002) in the establishment and employment location quotients for biodiesel and residential solar energy value chains. Note that establishment concentration patterns (Figures 10 and 12) suggest more counties exhibit firm concentration in these two sectors than employment concentration. These patterns appear to be the result of comparing establishment distributions with an employment based reference distribution ($x_j$).

The patterns of local employment concentration of residential solar energy firms corroborate the implications of the global concentration indexes regarding firm site selection near locations exhibiting external economies (e.g. Silicon Valley or the Southwest US). The magnitude of employment concentration observed in the Appalachian Region is also remarkable – particularly in Virginia, West Virginia, and Southeastern Pennsylvania. When compared with the distribution of solar radiation (Figure 3; NREL 2008), it appears that concentration in the Appalachian Region is likely not because of sunlight. The prevalence of employment concentration across the Sunbelt is also notable. Therefore, it remains unclear if firms that could participate in residential solar value chains may be more attracted to areas endowed with relatively more sunlight hours and radiation or if firms making up this sector perceive some advantage from locating near other factors (as in the case of Appalachian Region concentration). Certainly, in the case of Suntech Power’s facility in Maricopa County, Arizona, proximity to solar radiation was an important site selection factor. With an employment based location quotient of 2.88 (significant at the 1%

level), cost savings appear to attract solar energy firms; perhaps because of access to sunlight, upstream or downstream markets, or specialized labor sheds. Inspection of the underlying local market structure reveals that Maricopa County is home to number of solar panel manufacturers and related industries, including First Solar, another dominant global competitor. Arizona's Renewable Energy Tax Incentive program has also served as a factor in attracting solar companies to the region (e.g. Saint-Gobain Solar in 2011). Therefore, it may be assumed that Maricopa County exhibits comparative advantage with respect to attracting solar energy value chain firms, driven simultaneously by the availability of sunlight, a developing infrastructure of supporting businesses, and a history of favorable policy incentives to producers. In the case of the biodiesel sector, extensive employment concentration (Figure 11) supports the implication from the first step of the analysis that site selection of firms belonging to the biodiesel sector may be more geographically disbursed than firms oriented toward the residential solar sector (Figure 10). Additionally, note the general lack of establishment (Figure 12) or employment concentration across much of the Midwest. Employment and establishments both appear more likely to concentrate near densely populated areas (e.g. The Eastern Seaboard, Atlanta, Florida, Southeast Texas, Los Angeles, and Seattle, among others). As would be expected, county-level total biomass resources appear to be most abundant near the population centers of the East and West coasts, and across Midwest, likely due to the prevalence of agricultural resources distributed across the region (Figure 4; NREL 2009). The lack of biodiesel value chain concentration in these areas (given the relatively high transport costs of soybeans and other oilseeds) suggests that many of the counties with a comparative advantage with respect to feedstock availability may lack other factors critical to the site selection decisions of biodiesel value chain firms. An employment location quotient of 1.22 (significant at the 1% level) indicates that Grundy County, Illinois may possess a comparative advantage with respect to attracting firms belonging to the biodiesel sector. REG Seneca, therefore, may enjoy reduced feedstock and final product transportation costs due to access to an extensive local transportation hub (a shared resource arising associated with external scale economies). Costly transportation of biodiesel inputs and products suggests that counties with access to an efficient transportation network (e.g. Grundy County) may possess comparative advantage with respect to attracting firms engaged in the biodiesel value chain.

From the aggregate location quotients (Tables 15 and 16), it is clear that metropolitan counties exhibited employment concentration with greater frequency than did non-metropolitan counties with respect to all value chains. Metropolitan counties appear to be, on average, associated with greater cost savings to renewable energy firms. For example, Maricopa County, Arizona and Grundy County, Illinois, both metropolitan counties, may have contributed to less uncertainty with respect to potential costs or output for Suntech Power and REG Seneca.

***Sensitivity Analysis***

Differences between observed patterns of establishment concentration from the spatially-adjusted GFW indexes for any sector were not evident at the 1% level of significance. However, estimates from six sectors appear to come from significantly different distributions over the time period. Specifically, firms in the biodiesel, wood direct fire, ethanol from switchgrass, wood ethanol, and wind energy sectors may have experienced declining tendencies toward concentration over the period. The firms belonging to the residential solar sector, on the other hand, may have increased their tendency to locate near other similar firms. However, observed differences in confidence intervals may, instead, be an artifact of the data collection practices involved in the data sets. Regardless, differences in the estimates over the period are subtle, indicating that any changes, whether statistically significant or not, are modest.

## Conclusions

Renewable energy technologies have been the focus of many policy makers and investors for their role in promoting economic development and transitioning away from dependence on fossil fuels. Yet the patterns of industry concentration that could result in cost-savings to firms, and eventually to the flow of investments to counties has not been investigated. Data from 2002 to 2006 was analyzed to describe the geographic landscape related to ten renewable energy sectors. A two-step procedure was developed to analyze firm concentration using recent advances in industry concentration analysis. The utility of this approach was motivated by focusing on the biodiesel and residential solar energy sectors. Metropolitan counties appear to have a comparative advantage with respect to attracting larger firms in both value chains. It appears that site selection decisions of firms engaged in the solar energy value chain are made independent of access to solar resources, largely gravitating toward population centers, suggesting that access to factors associated with population centers (e.g. skilled labor or knowledge spillovers) may be of greater importance than proximity to natural solar resources. When compared to the distribution of agricultural feedstock resources, observed concentration patterns for firms engaged in value chains for biodiesel produced from oilseeds implies slightly more serious implications than in the solar case. Because oilseed production is typically performed away from metropolitan areas associated with skilled labor availability, access to a well-developed transportation network (often found in metropolitan areas) may reduce costs associated with feedstock transport. Firms engaged in the production of biodiesel using yellow grease, however, may be able to capitalize on the availability of skilled labor *and* feedstock nearby, as yellow grease typically obtained from restaurant wastes in areas with comparatively denser population (metropolitan locations).

These concentration patterns suggest several implications for a range of stakeholders. If the objective is to determine where firms engaged in biodiesel value chains may experience relatively greater returns, policy makers and investors may observe that metropolitan areas generally possess the necessary structures to support various levels of value chain activity. On the other hand, localization economies appear to be relatively weak in rural counties that may produce a range of agricultural feedstock used in biodiesel and ethanol production or for electricity generation from direct firing of wood products. Investors are often forced to weigh the relative importance of access to raw materials versus access to localization economies. Non-metropolitan policy makers may need to develop policy tools aimed at improving comparative advantage (perhaps in terms of transport infrastructure or labor pool quality), whereas metropolitan policy makers' focus may be on attracting specific renewable energy support firms.

The relative riskiness associated with investments in locations exhibiting a given local share of green energy firms can be interpreted with respect to stakeholders' risk tolerance. For example, risk-averse investors may opt to locate their firms in areas where concentration is higher, and thus expected costs are known with greater certainty. Policy makers with the objective of spurring economic growth may design policies to attract firms in industries exhibiting the least output uncertainty for their particular region. Clearly, identification of each location's comparative advantage with respect to attracting firms in each of the renewable energy sectors is important to reduce risk exposure to tax payers (through more efficient policy initiatives) and investors. It is important for investors and policy makers to keep in mind that risk is only limited to the influence of geography on firm profits and does not relate to the risk associated with market and policy conditions, or even managerial decisions.

Additional limitations of this research should be considered, as well. First, certain renewable energy industries may give rise to regional leakages. For example, Suntech Power's Maricopa County plant acquires many of their inputs from other states and countries (Cheyney 2011). If specialized labor is unavailable in the immediate area, workers hired from across local borders may spend their wages spent outside a particular county. Second, crowding out of existing local firms may also occur following new investments, potentially resulting in little or no net gains in economic output. Third, location quotients, as an indicator of comparative advantage, must also be carefully considered by investors with respect to the limitations of the measures used. While location quotients may suggest evidence of factors supporting concentration, careful inspection of local economic and industry structure is critical. Finally, temporary policy incentives used to attract renewable energy may also affect firm location decisions to the extent that concentration may arise in the absence of any other localized sources of comparative advantage. Therefore, upon policy expiration, advantages associated with certain locations may be lost.

Taken together, these implications suggest that comparative advantage with respect to attracting renewable energy activity may not be as straightforward as many would hope. Substantial heterogeneity across the ten value chains suggests that deeper investigation of the individual value chains will be necessary in the future. The local measures described in step two may be useful for regression analysis to identify which county level factors (e.g. average education levels, resource availability, or federal policy funds allocated) are associated with firm concentration and increasing returns to external scale economies.

## References

Anselin, L. 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, The Netherlands.

Arbia, G. 2001. The Role of Spatial Effects in the Empirical Analysis of Regional Concentration. *Journal of Geographical Systems* 3(3): 271-281.

Baldwin, R.E., and P. Martin. 2004. Agglomeration and Regional Growth. In: Henderson, J.V., and Thisse, J-F. (eds.) Handbook of Regional and Urban Economics. Elsevier, New York 2671-2712.

Bosman, T.H. 2011. Clean Energy Seeks Customer Access. SiteSelection.com. January. http://www.siteselection.com/issues/2011/jan/alternative-energy-hot-spots.cfm.

Bradsher, K. 2011. Solar Panel Maker Moves Work to China. The New York Times, 14 January.

Brakman, S., H. Garretsen, and C. van Marrewijk. 2009. The New Introduction to Geographical Economics. Cambridge University Press, Cambridge, UK.

Brülhart, M., and R. Traeger. 2003. An Accounty of Geographic Concentration Patterns in Europe. HWWA Discussion Paper 226.

Carlton, D.W. 1983. The Location and Employment Choices of New Firms: An Econometric Model with Discrete and Continuous Endogenous Variables. *The Review of Economics and Statistics* 65: 440-449.

Cheyney, T. 2011. Raising Arizona Solar: Suntech bets on US market with Goodyear module manufacturing plant. PV-Tech.org. 18 May. http://www.pv-tech.org/chip_shots_blog/raising_ arizonasolar_suntech_bets_on_us_market_with_goodyear_module_manufa.

Cronon, W. 1991. Nature's Metropolis: Chicago and the Great West. W. W. Norton & Company, Inc., New York.

Duranton, G. and D. Puga. 2004. Micro-foundations of Urban Agglomeration Economies. In: Henderson, J.V., and Thisse, J-F. (eds.) Handbook of Regional and Urban Economics. Elsevier, New York 2063-2118.

Elliot, D. 2012. Wind Energy Tax Break Fails in US Senate. CBS News, 14 March, http://www.cbsnews.com/8301-505245_162-57397173/wind-energy-tax-break-fails-in-us-senate/.

Ellison, G. and E.L. Glaeser. 1999. The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration? *The American Economic Review* 89(2): 311-316.

Ellison, G. and E. L. Glaeser. 1997. Geographic Concentration in US Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy* 105(5): 889-927.

English, B.C., D.G. De La Torre Ugarte, K.L. Jensen, C. Hellwinckel, R.J. Menard, B. Wilson, R.K. Roberts, and M. Walsh. 2006. 25% Renewable Energy for the United States By 2025: Agricultural and Economic Impacts. The University of Tennessee Bio-based Energy Analysis Group.

Feser, E., H. Rensky , and H. Goldstein. 2008. Clusters and Economic Development Outcomes: An Analysis of the Link Between Clustering and Industry Growth. *Economic Development Quarterly* 22(4): 324-344.

Figueiredo, O., P. Guimarães, and D. Woodward. 2007. Localization Economies and Establishment Scale: A Dartboard Approach. FEP Working Papers 247.

Florence, P.S. 1939. Report on the Location of Industry. Political and Economic Planning, London, U.K.

Fujita, M., and J-F. Thisse. 2009. New Economic Geography: An appraisal on the occasion of Paul Krugman's 2008 Nobel Prize in Economics. *Regional Science and Urban Economics* 39:109-119.

Fujita, M., and J-F. Thisse. 2002. Economics of Agglomeration: Cities, Industrial Location, and Regional Growth. Cambridge University Press, Cambridge, UK.

Gillingham, K., and J. Sweeney. 2010. Market Failure and the Structure of Externalities. In: Moselle, B., Padilla, A.J., and Schmalensee, R. (eds.) Harnessing Renewable Energy in Electric Power Systems – Theory, Practice, Policy. RFF Press, Washington, D.C.

Gold, R. 2011. Overrun by Chinese Rivals, US Solar Company Falters. *Wall Street Journal*, 17 August.

Gordon, M. 2012. US Govt Sets New Tariffs on China Solar Panels. *Associated Press*, 20 March.

Greene, W.H. 2000. Econometric Analysis, 4th Edition. Prentice Hall. Upper Saddle River, NJ.

Guimarães, P., O. Figueiredo, and D. Woodward. 2011. Accounting for Neighboring Effects In Measures of Spatial Concentration. *Journal of Regional Science* 51(4):678-693.

Guimarães, P., O. Figueiredo, and D. Woodward. 2009. Dartboard Tests for the Location Quotient. *Regional Science and Urban Economics* 39:360-364.

Guimarães, P., O. Figueiredo, and D. Woodward. 2007. Measuring the Localization of Economic Activity: A Parametric Approach. *Journal of Regional Science* 47(4):753-774.

Guimarães, P., O. Figueiredo, and D. Woodward. 2006. Geographic Concentration and Establishment Scale: An Extension Using Panel Data. *Journal of Regional Science* 46(4): 733-746.

Guimarães, P., O. Figueiredo, and D. Woodward. 2004.Industrial Location Modeling: Extending the Random Utility Framework. *Journal of Regional Science* 44(1):1-20.

Holmes, T. J. and J.J. Stevens. 2004. Spatial Distribution of Economic Activities in North America. In: Henderson, J.V., and Thisse, J-F. (eds.) Handbook of Regional and Urban Economics. Elsevier, New York 2797-2844.

Holmes, T. J. and J.J. Stevens. 2002. Geographic Concentration and Establishment Scale. *The Review of Economics and Statistics* 84(4):682-690.

Hoover, E.M. 1937. Location Theory and the Shoe and Leather Industries. Harvard University Press, Cambridge, MA.

Isserman, A.M. 1977. The Location Quotient Approach to Estimating Regional Economic Impacts. *Journal of the American Institute of Planners* 43(1): 33-41.

Jakab, S. 2012. Solar Turnaround Eclipsed by Subsidy Cuts. *Wall Street Journal*, 23 February.

Jensen, K.L., D.M. Lambert, R.J. Menard, and B.C. English, and W. Xu. 2010. Projected Impacts of Green Jobs Development in the Appalachian Region. Bio-Based Energy Analysis Group, The University of Tennessee, Institute of Agriculture.

Jensen, K., R.J. Menard, and B.C. English. 2007. US and Tennessee Biodiesel Production – 2007 Industry Update. Report prepared for USDA Rural Development.

Kahn, C. 2012. Natural gas prices stay at 10-year low. Associated Press. 12 April.

Kim, Y., D.L. Barkley, and M.S. Henry. 2000. Industry Characteristics Linked to Establishment Concentration in Nonmetropolitan Areas. *Journal of Regional Science* 40(2): 231-259.

Krugman, P. 1990. Increasing Returns and Economic Geography. NBER Working Papers 3275, National Bureau of Economic Research, Inc.

Krugman, P. 1991. Geography and Trade. MIT Press, Cambridge, MA, 1991.

Lafourcade, M., and G. Mion. 2007. Concentration, Agglomeration, and the Size of Plants. *Regional Science and Urban Economics* 37: 46-68.

Lambert, D.M., K.T. McNamara, and M.I. Beeler. 2007. Location Determinants of Food Manufacturing Investment: Are Non-metropolitan Counties Competitive? Selected paper presented at the American Agricultural Economics Association Annual Meeting, Portland, OR.

Lambert, D.M., K.T. McNamara, and M.I. Garrett. 2006. An Application of Spatial Poisson Models to Manufacturing Investment Location Analysis. *Journal of Agricultural and Applied Economics* 38(1): 105-121.

Lambert, D.M., M. Wilcox, A. English, L. Stewart. 2008. Ethanol Plant Location Determinants and County Comparative Advantage. *Journal of Agricultural and Applied Economics* 40(1):117-135.

Lichtenberg, R.M. 1960. One-Tenth of a Nation: National Forces in the Economic Growth of the New York Region. Harvard University Press, Cambridge, MA, 1960.

Malecki, E.J. and P. Varaiya. 1986. Innovation and Changes in Regional Structure. In: Nijkamp, P. (ed.) Handbook of Regional and Urban Economics. Elsevier. New York 629 – 645.

Marshall, A. 1890. Principles of Economics. Macmillan, London, England.

Maurel, F., and B. Sédillot. 1999. A Measure of the Geographic Concentration in French Manufacturing Industries. *Regional Science and Urban Economics* 29: 575-604.

McFadden, D. 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, ed. P. Zarambka. Academic Press, New York, NY, 1974.

Milbourn, M.A. 2011. $24 million in state tax breaks not enough. The Orange County Register, 6 September.

Minnesota IMPLAN Group. 2006. IMPLAN. Stillwater, Minnesota.

Ottaviano, G., and J-F. Thisse. 2004. Agglomeration and Economic Geography. In: Henderson, J.V., and Thisse, J-F. (eds.) Handbook of Regional and Urban Economics. Elsevier, New York 2563-2608.

Pierobon, J. 2011. Tall Hurdles for Renewable Energy Policy in 2012. RenewablesBiz, 7 December, http://www.renewablesbiz.com/article/11/12/tall-hurdles-renewable-energy-policy-2012.

Puga, D. 2010. The Magnitude and Causes of Agglomeration Economies. *Journal of Regional Science* 50(1): 203-219.

REG, LLC. 2010. Renewable Energy Group welcomes Seneca (Illinois) to its biodiesel network. 9 April. http://www.soypower.com/pdfs/Seneca%20plants%20joins%20REG%20network, %20040910.pdf.

Ricardo, D. 1817. On the Principles of Political Economy and Taxation. London.

Romig, S. 2012. Argentina soy prices advance again. Wall Street Journal Market Watch. 12 April. http://www.marketwatch.com/story/argentina-soy-prices-advance-again-2012-04-12.

Sarmiento, C. and W.W. Wilson. 2006. Spatial Competition and Ethanol Plant Location Decisions. Selected paper presented at the American Agricultural Economics Association Annual Meeting, Orlando, FL.

Schjeldahl, D. 2012. Advanced Energy Markets. SiteSelection.com. January. http://www.site selection.com/issues/2012/jan/sas-alternative-energy.cfm.

Spegele, B. 2011. Chinese Company to Tap US Wind. *Wall Street Journal*, September 20, 2011.

Stewart, L., and D.M. Lambert. 2011. Spatial Heterogeneity of Factors Determining Ethanol Production Site Selection in the US, 2000-2007. *Biomass and Bioenergy* 35(3): 1273-1285.

Stewart, L., D.M. Lambert, M.D. Wilcox, and B.C. English. 2009. Tennessee Agriculture and Forestry Industry Clusters and Economic Performance, 2001-2006. *Journal of Agricultural and Applied Economics* 34(1): 172-195.

Sweet, C., and Y. Chernova. 2011. First Solar Revamps Amid Weak Market. *Wall Street Journal*, 15 December.

Tennessee Department of Economic and Community Development. 2011. State Energy Program. http://www.tn.gov/ecd/recovery/sep.html

The Pew Charitable Trusts. 2010. Who's Winning the Clean Energy Race? Washington D.C. Available at: http://www.pewtrusts.org/uploadedFiles/wwwpewtrustsorg/Reports/ Global_warming/G-20%20Report.pdf.

Tracy, R. 2012. Wind-Power Firms on Edge. *Wall Street Journal*, 2 February.

US Census Bureau. 2006. County Business Patterns (NAICS) Database. Washington, D.C.

US Department of Energy. 2010. *Renewable Energy Consumption and Electricity Preliminary Statistics 2009*. Washington DC: Energy Information Administration, August.

US Energy Information Administration. 2011. Average Retail Price of Electricity to Ultimate Customers by End-Use Sector, 1999 through 2010. http://www.eia.gov/electricity/ annual/pdf/table7.4.pdf.

Wald, M. 2009. Cost Works Against Alternative and Renewable Energy Sources in Time of Recession. *New York Times*, 28 March, http://www.nytimes.com/2009/03/29/business/ energy-environment/29renew.html?_r=2.

# Appendix

Figure 2 - Renewable Energy Industry Structure and NAICS - IMPLAN Bridge

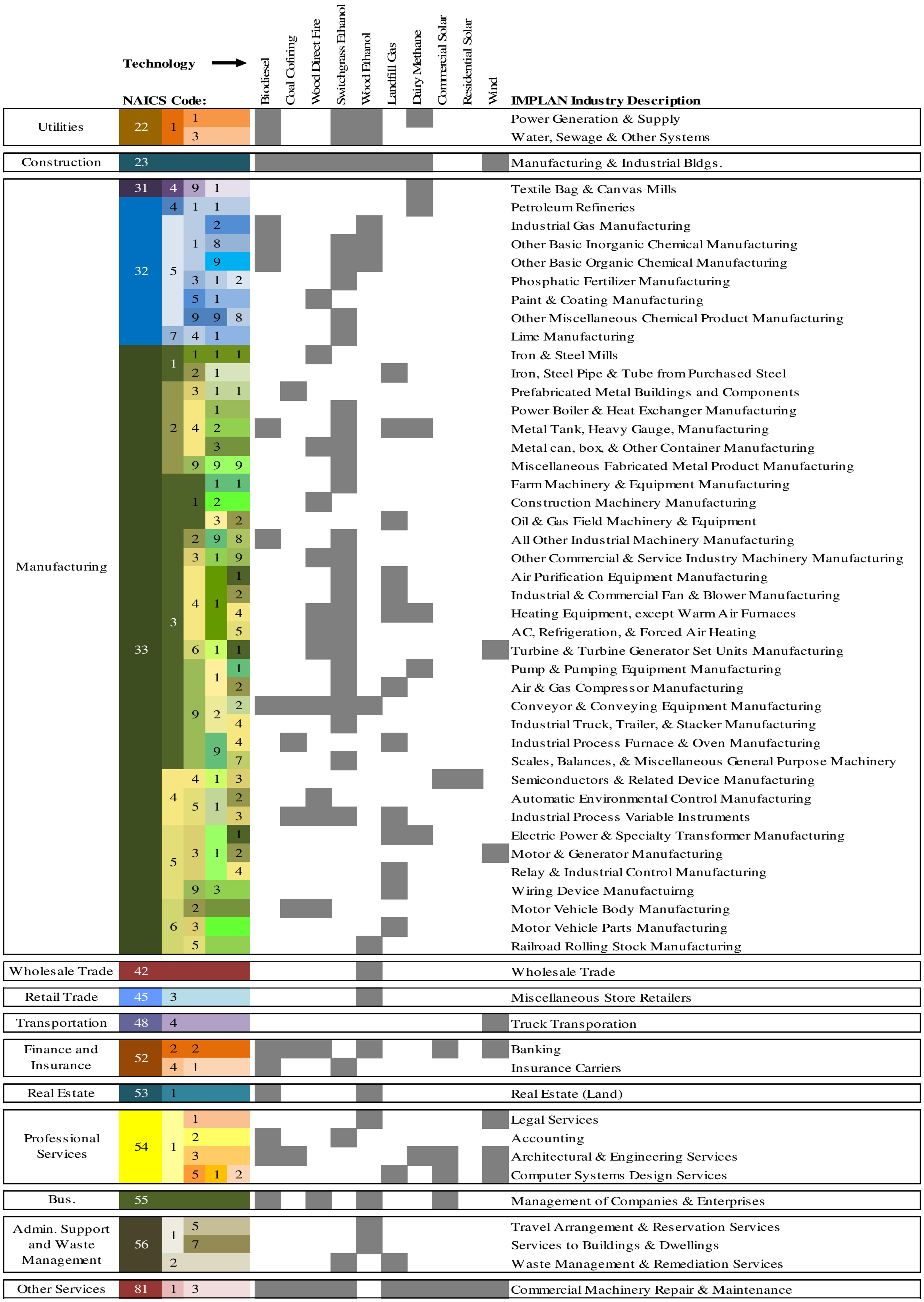


Figure 3 - NREL Solar Resource Potential

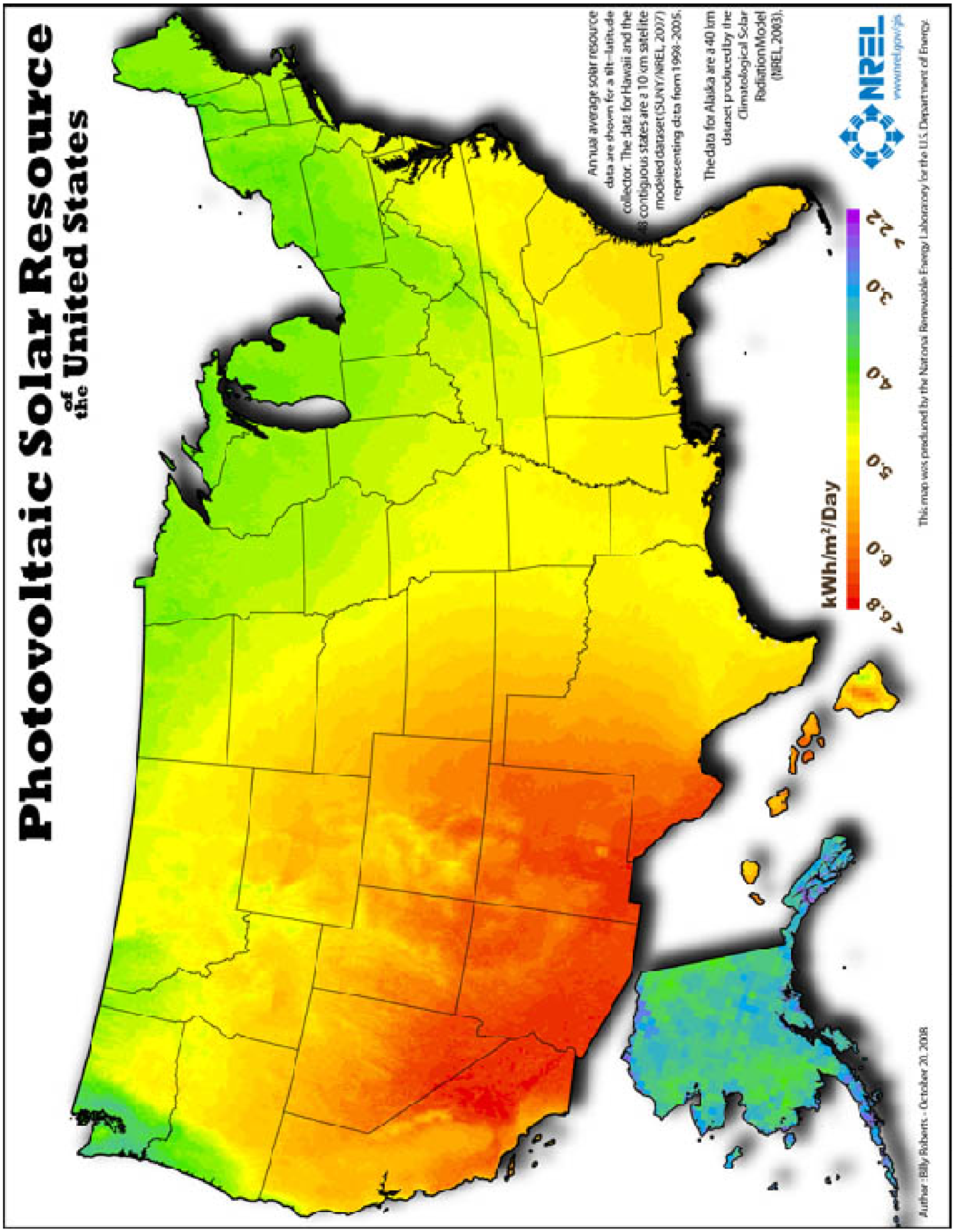


Figure 4 - NREL Biomass Resource Potential

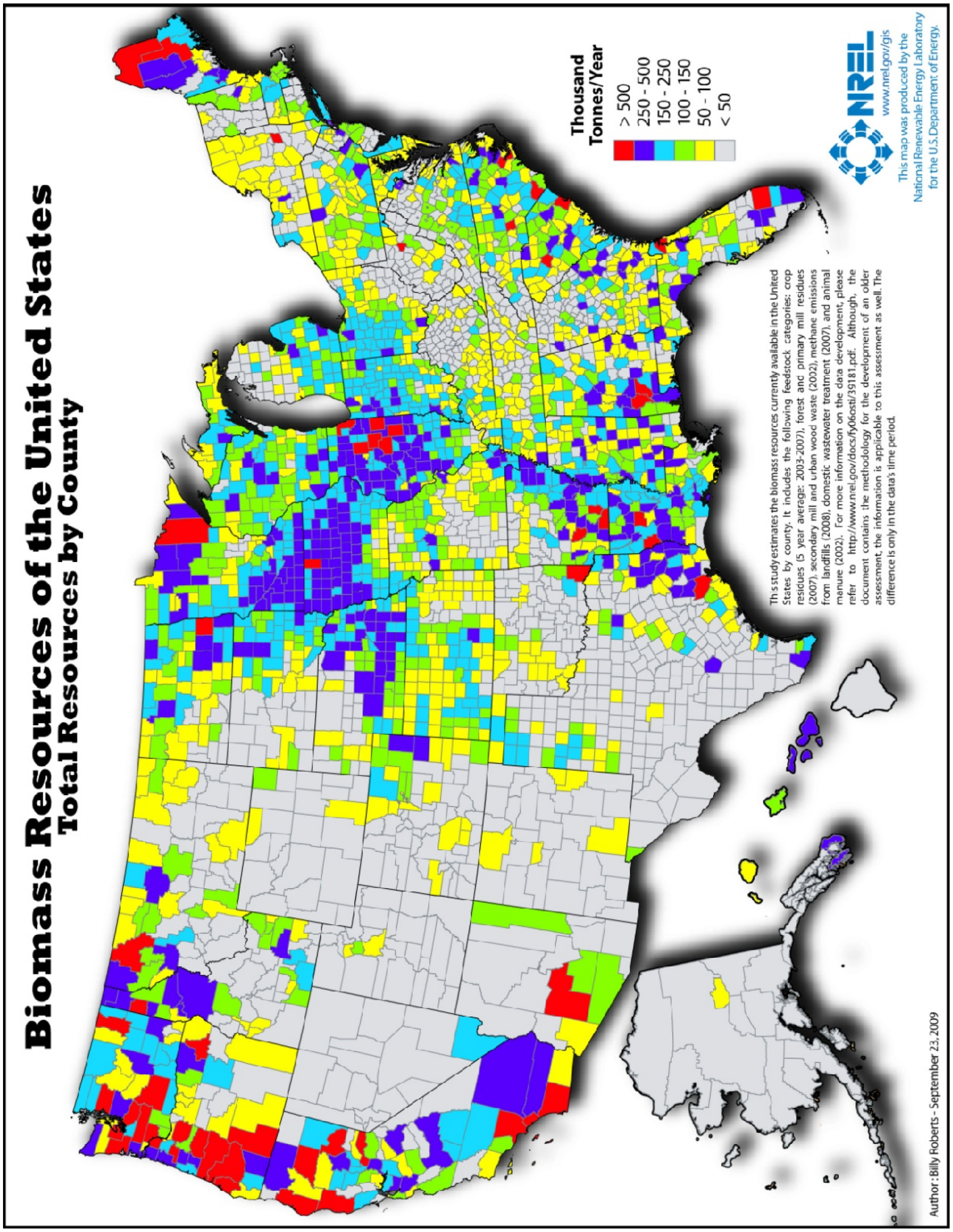
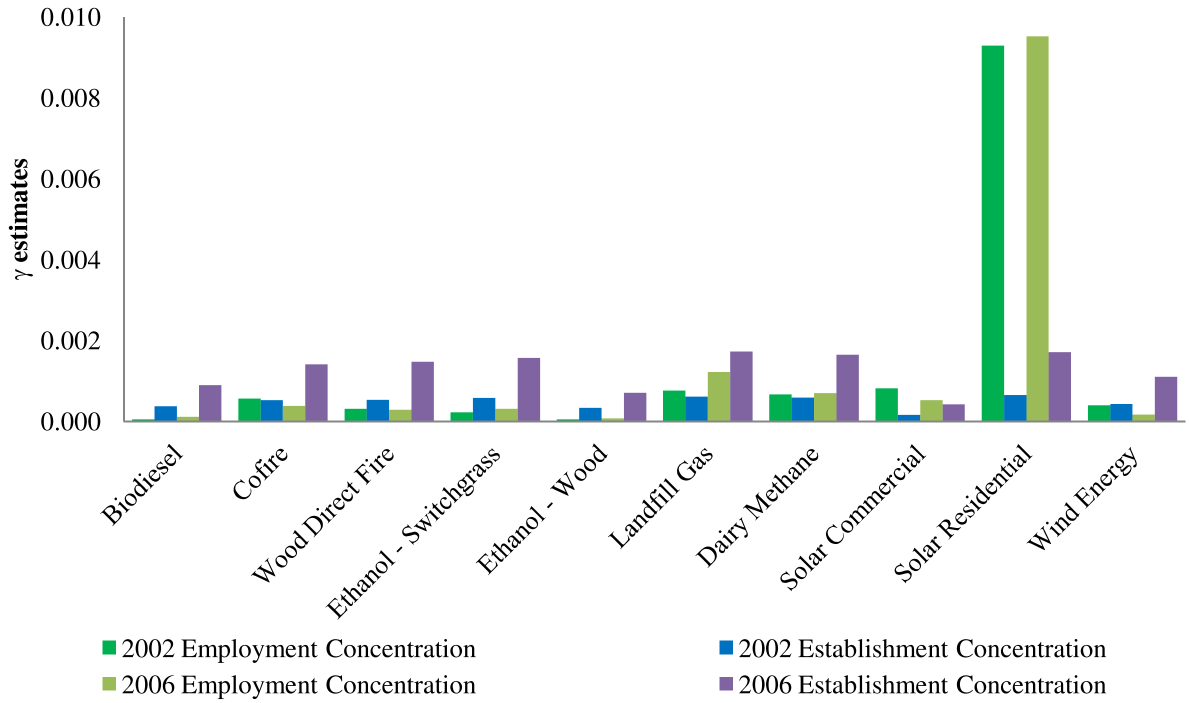
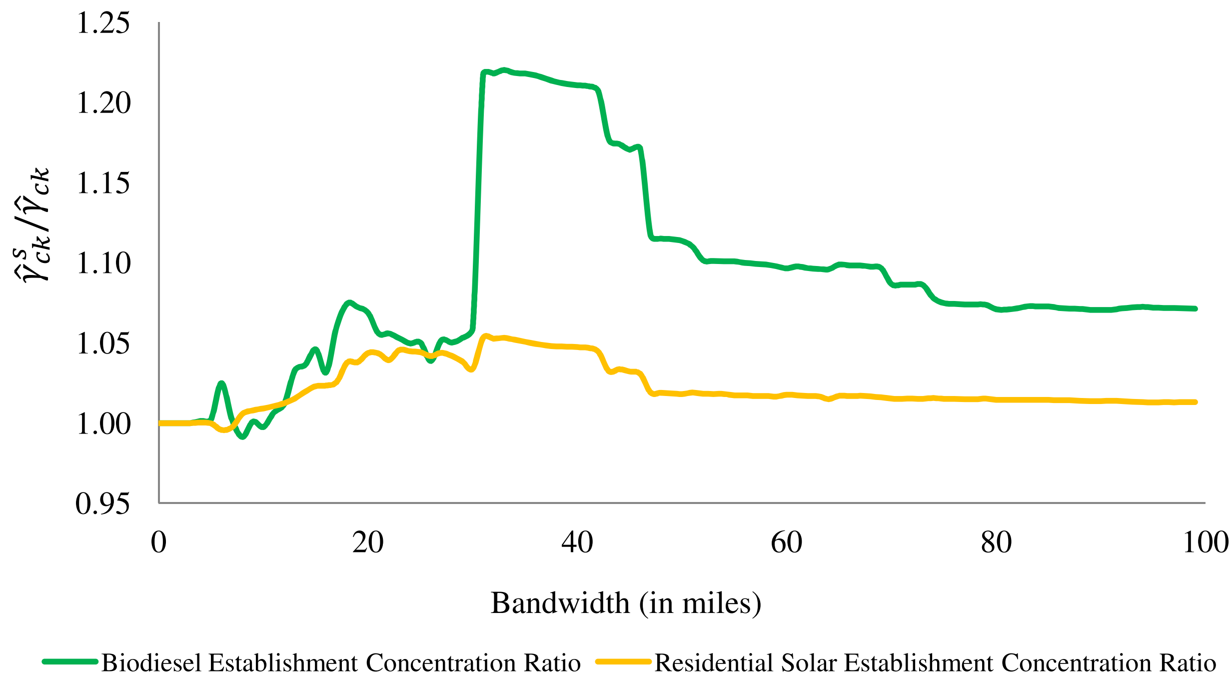


Figure 5 - Global Concentration Index Comparison (2002 & 2006)




Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN

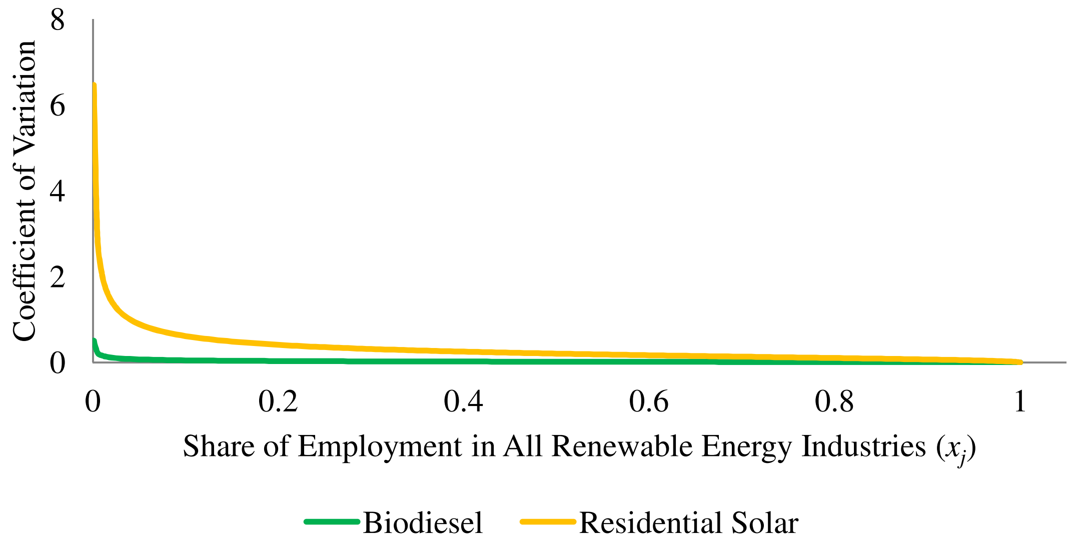
Figure 6 - Spatial-Aspatial Index Ratio and Optimal Neighborhood Bandwidths




Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN

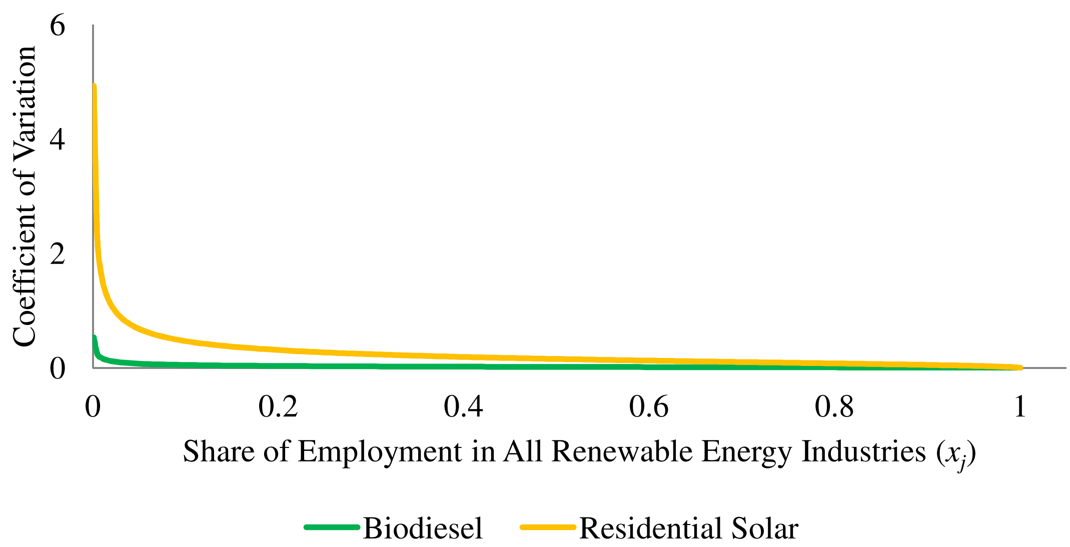
Note: In this example, the optimal bandwidth for the biodiesel sector was 33 miles and for the residential solar sector was 31 miles.

Figure 7 – Variation in Expected Local Costs




Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN

Figure 8 - Variation in Expected Local Value-Added Output




Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN

Figure 9 - Residential Solar Value Chain Employment Concentration

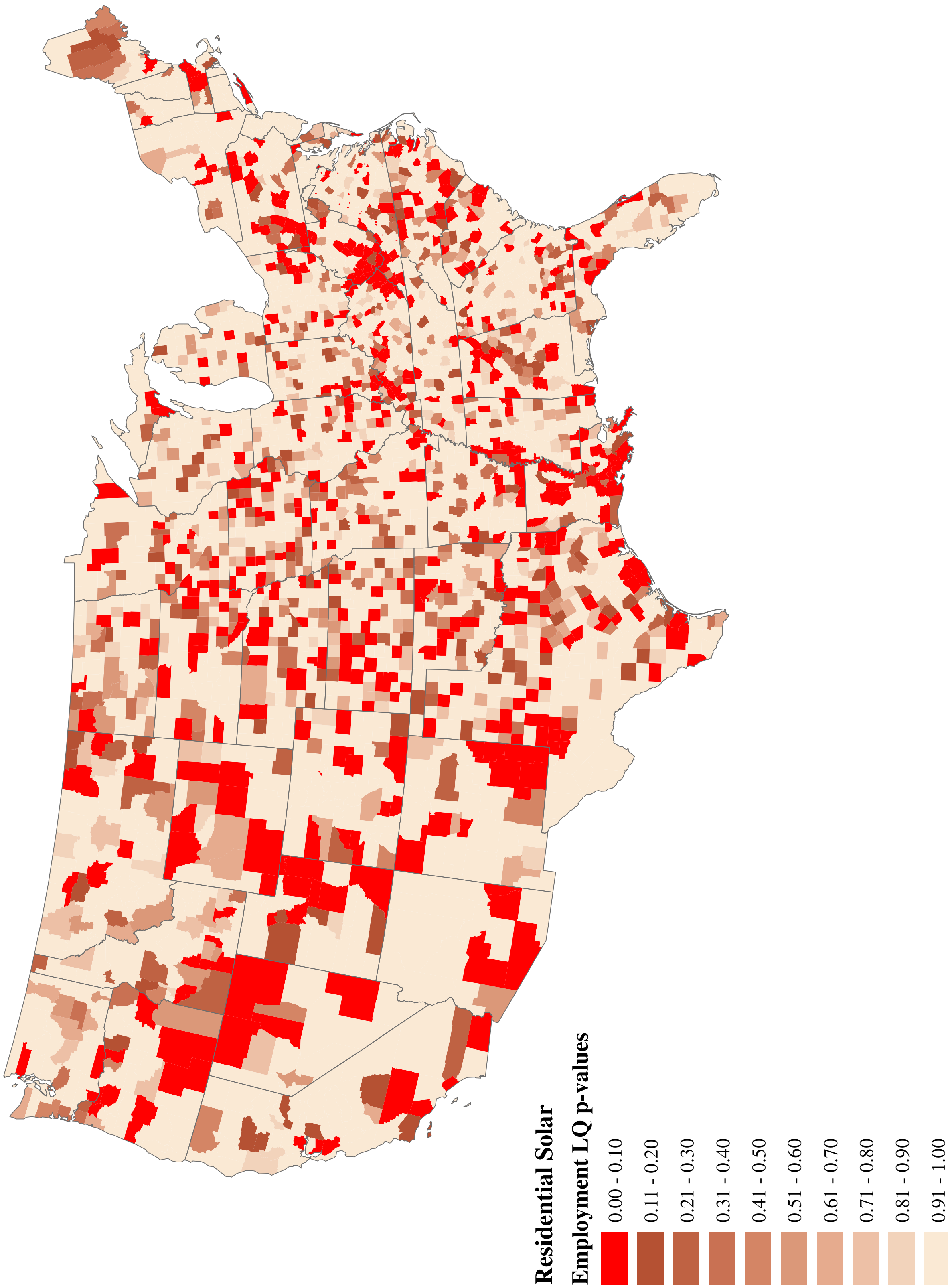


Figure 10 - Residential Solar Value Chain Establishment Concentration

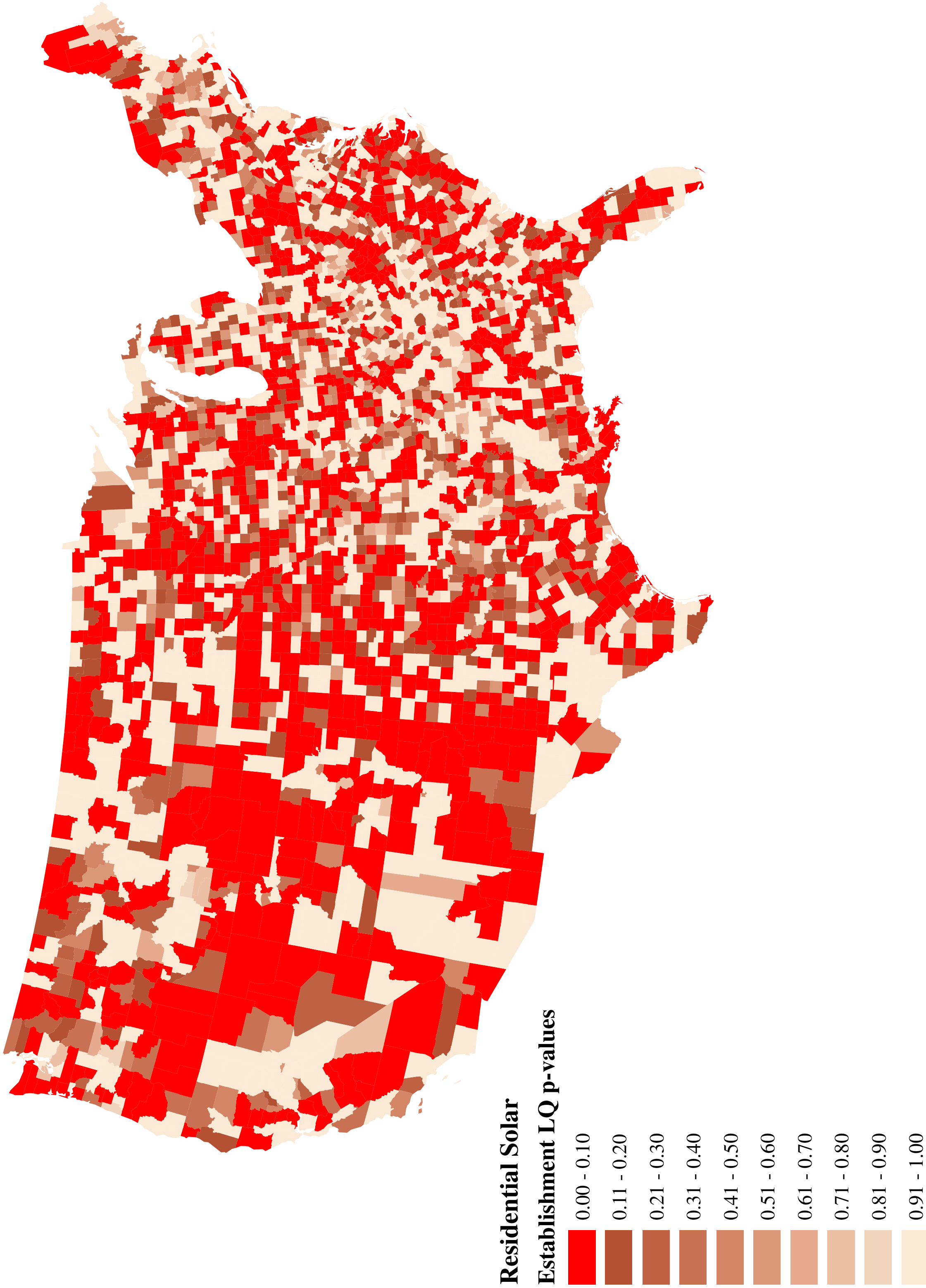


Figure 11 - Biodiesel Value Chain Employment Concentration

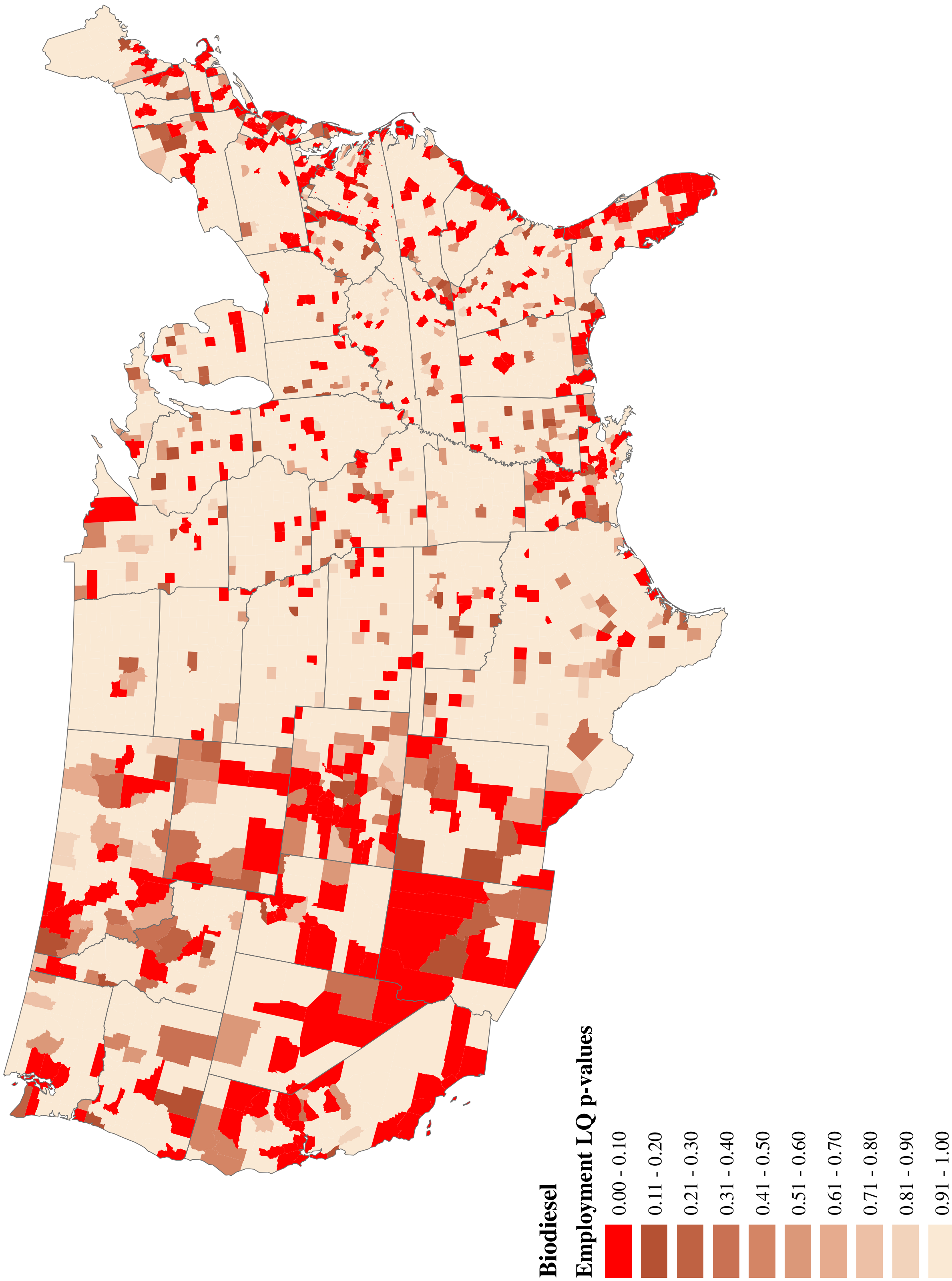


Figure 12 - Biodiesel Value Chain Establishment Concentration

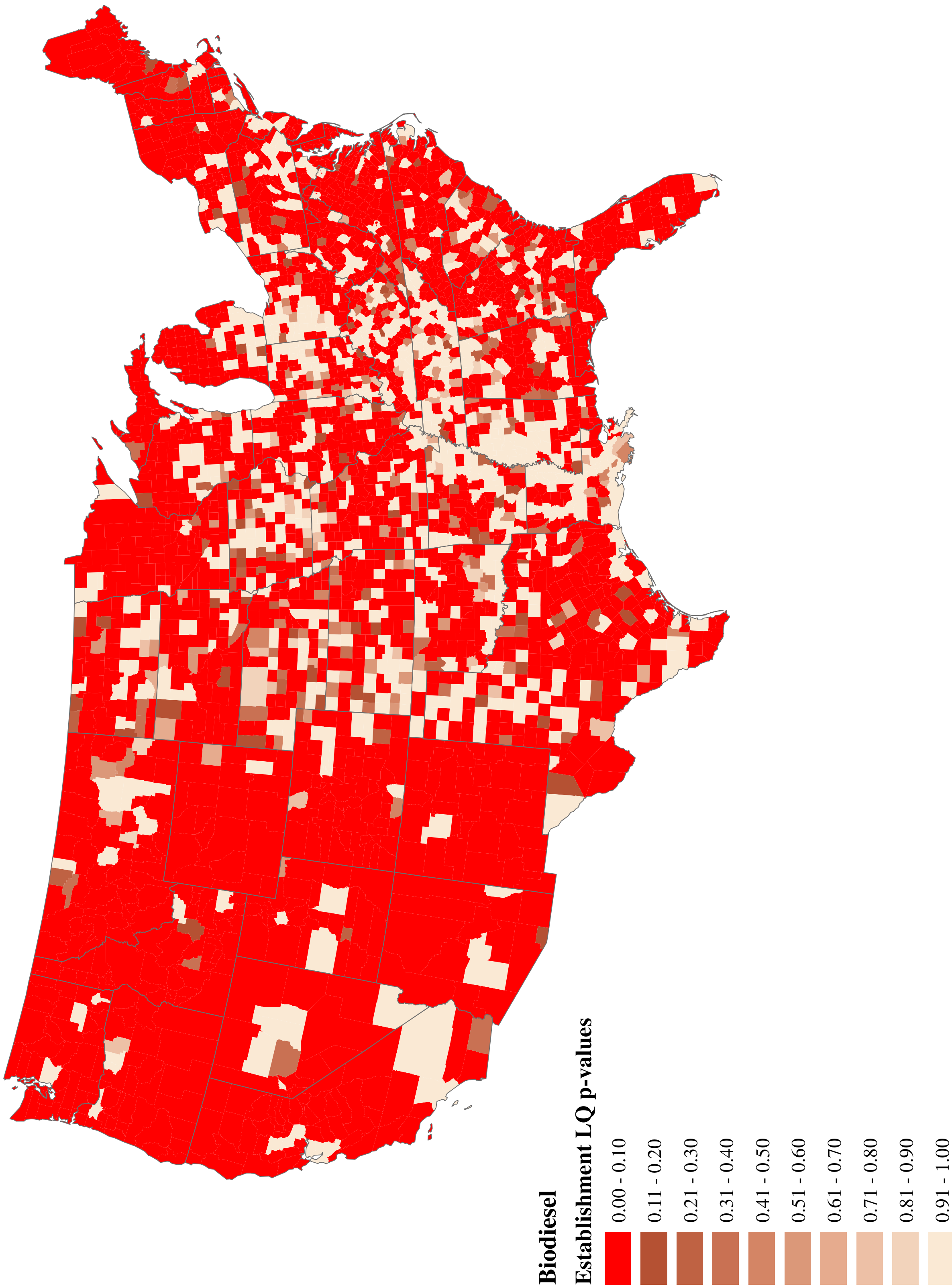


Table 10 – Renewable energy sector employment and establishments - 2002 and 2006

| Value Chain | Establishments | | Employment | |
|---|---|---|---|---|
| | 2002 | 2006 | 2002 | 2006 |
| Biodiesel | 2,061,911 | 2,366,005 | 12,944,391 | 17,275,533 |
| Cofire | 782,710 | 1,042,672 | 9,015,782 | 5,417,058 |
| Wood Direct Fire | 777,306 | 981,447 | 9,206,047 | 5,586,651 |
| Ethanol - Switchgrass | 1,122,755 | 1,028,278 | 10,552,367 | 6,674,636 |
| Ethanol - Wood | 1,904,160 | 2,225,963 | 20,513,715 | 24,810,396 |
| Landfill Gas | 763,581 | 878,564 | 9,714,170 | 3,702,717 |
| Dairy Methane | 757,450 | 947,920 | 7,558,072 | 4,648,078 |
| Commercial Solar | 146,060 | 336,135 | 5,142,390 | 6,888,514 |
| Residential Solar | 1,098 | 25,528 | 42,911 | 549,916 |
| Wind Energy | 1,045,790 | 1,396,584 | 13,778,125 | 10,234,347 |
| Total Renewable Energy Sector | 2,498,469 | 2,723,175 | 29,959,974 | 34,605,146 |

Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN

Table 11 – 2002 global industry concentration estimates and optimal bandwidths

| | Establishment Concentration | | Employment Concentration | |
|---|---|---|---|---|
| Value Chain | $\hat{\gamma}_{ck}^{s}$ | $D_{ck}^{*}$ | $\hat{\gamma}_{k}^{s}$ | $D_{k}^{*}$ |
| Biodiesel | 0.00074*** | 33 | 0.00005*** | 29 |
| Coal Cofire | 0.00126*** | 32 | 0.00056*** | 20 |
| Wood Direct Fire | 0.00132*** | 32 | 0.00031*** | 7 |
| Ethanol - Switchgrass | 0.00136*** | 33 | 0.00023*** | 25 |
| Ethanol - Wood | 0.00054*** | 33 | 0.00005*** | 27 |
| Landfill Gas | 0.00153*** | 32 | 0.00076*** | 34 |
| Dairy Methane | 0.00144*** | 32 | 0.00066*** | 31 |
| Commercial Solar | 0.00031*** | 29 | 0.00082*** | 20 |
| Residential Solar | 0.00137*** | 31 | 0.00937*** | 31 |
| Wind Energy | 0.00096*** | 33 | 0.00040*** | 33 |

Source: 2002 US CBP, 2002 WholeData.Net. *** denotes significance at 1% level

Table 12 – 2006 global industry concentration estimates and optimal bandwidths

| | Establishment Concentration | | Employment Concentration | |
|---|---|---|---|---|
| Value Chain | $\hat{\gamma}_{ck}^{s}$ | $D_{ck}^{*}$ | $\hat{\gamma}_{k}^{s}$ | $D_{k}^{*}$ |
| Biodiesel | 0.00089*** | 33 | 0.00011*** | 76 |
| Coal Cofire | 0.00141*** | 31 | 0.00038*** | 21 |
| Wood Direct Fire | 0.00147*** | 31 | 0.00028*** | 14 |
| Ethanol - Switchgrass | 0.00157*** | 31 | 0.00031*** | 31 |
| Ethanol - Wood | 0.00070*** | 33 | 0.00007*** | 33 |
| Landfill Gas | 0.00173*** | 32 | 0.00122*** | 33 |
| Dairy Methane | 0.00165*** | 31 | 0.00070*** | 37 |
| Commercial Solar | 0.00042*** | 16 | 0.00052*** | 20 |
| Residential Solar | 0.00171*** | 33 | 0.00951*** | 20 |
| Wind Energy | 0.00110*** | 33 | 0.00017*** | 17 |

Source: 2006 US CBP, 2006 IMPLAN *** denotes significance at 1% level

Table 13 – Global industry concentration estimates for residential solar industries

| Year | Industry Description | IMPLAN Code | NAICS Code | $\hat{\gamma}_{ck}^{s}$ | $\hat{\gamma}_{sk}$ |
|---|---|---|---|---|---|
| 2006 | Semiconductors & Related Device Manuf. *** | 311 | 334413 | 0.03142 | 0.04667 |
| | Commercial Machinery Repair & Maintenance *** | 485 | 8113// | 0.00150 | 0.00148 |
| 2002 | Semiconductors & Related Device Manuf. *** | 311 | 334413 | 0.03279 | 0.03869 |
| | Commercial Machinery Repair & Maintenance *** | 485 | 8113// | 0.00121 | 0.00175 |

Source: 2002, 2006 US CBP, 2002 WholeData.Net, 2006 IMPLAN *** denotes significance at 1% level

Table 14 – Relative green energy sector uncertainty rank

| Value Chain | Cost Variation | Output Variation |
|---|---|---|
| Biodiesel | 9 | 7 |
| Coal Cofire | 5 | 6 |
| Wood Direct Fire | 7 | 8 |
| Ethanol - Switchgrass | 6 | 5 |
| Ethanol - Wood | 10 | 9 |
| Landfill Gas | 2 | 2 |
| Dairy Methane | 3 | 3 |
| Solar Commercial | 4 | 4 |
| Solar Residential | 1 | 1 |
| Wind Energy | 8 | 10 |

Table 15 – 2002 aggregate location quotients

| | Metro | | Micro | | Non-Core | |
|---|---|---|---|---|---|---|
| Value Chain | EST | EMP | EST | EMP | EST | EMP |
| Biodiesel | 0.937 | 1.885 | 1.003 | 0.773 | 1.050 | 0.397 |
| Cofire | 0.956 | 1.534 | 1.022 | 0.931 | 1.025 | 0.601 |
| Wood Direct Fire | 0.944 | 1.708 | 1.022 | 0.841 | 1.036 | 0.506 |
| Ethanol - Switchgrass | 0.908 | 1.360 | 1.013 | 1.069 | 1.068 | 0.672 |
| Ethanol - Wood | 0.859 | 1.950 | 0.997 | 0.839 | 1.116 | 0.310 |
| Landfill Gas | 0.948 | 1.371 | 1.000 | 1.079 | 1.042 | 0.657 |
| Dairy Methane | 0.944 | 1.360 | 0.987 | 0.963 | 1.052 | 0.727 |
| Commercial Solar | 1.246 | 2.310 | 1.130 | 0.533 | 0.733 | 0.174 |
| Residential Solar | 0.794 | 1.366 | 1.036 | 0.970 | 1.149 | 0.718 |
| Wind Energy | 0.884 | 1.357 | 1.013 | 0.960 | 1.088 | 0.730 |

Source: 2002 US CBP, 2002 WholeData.Net.

Table 16 – 2006 aggregate location quotients

| | Metro | | Micro | | Non-Core | |
|---|---|---|---|---|---|---|
| Value Chain | EST | EMP | EST | EMP | EST | EMP |
| Biodiesel | 1.049 | 1.748 | 1.047 | 0.671 | 0.936 | 0.561 |
| Cofire | 1.045 | 1.501 | 1.056 | 0.899 | 0.935 | 0.644 |
| Wood Direct Fire | 1.043 | 1.351 | 1.049 | 1.164 | 0.940 | 0.630 |
| Ethanol - Switchgrass | 0.986 | 1.158 | 1.020 | 1.101 | 1.001 | 0.820 |
| Ethanol - Wood | 0.957 | 1.695 | 1.009 | 0.828 | 1.031 | 0.523 |
| Landfill Gas | 1.036 | 1.227 | 1.009 | 1.139 | 0.966 | 0.744 |
| Dairy Methane | 1.023 | 1.262 | 1.015 | 0.935 | 0.974 | 0.820 |
| Solar Commercial | 1.492 | 2.265 | 1.098 | 0.575 | 0.550 | 0.189 |
| Solar Residential | 0.870 | 1.011 | 1.070 | 1.003 | 1.070 | 0.990 |
| Wind Energy | 0.961 | 1.546 | 1.004 | 0.920 | 1.029 | 0.597 |

Source: 2006 US CBP, 2006 IMPLAN.

# IV. Summary and Conclusions

This thesis focused on exploring methods to enhance incomplete employment data sets and contributing a two-stage approach to analyzing geographic concentration patterns of firms and employment belonging to the green energy sector. Comparison of methods attending to the problem of employment data suppression suggest that the procedure contributed here, the Iteratively Constrained Rebalanced Matrix (ICRM) approach, and the unconstrained IRCRM algorithm may be more flexible than an existing goal programming (ZG) approach in imputing missing employment data in the US County Business Patterns (CBP) databases.

More robust conclusions about the accuracy of the imputation procedures will be possible through comparison across randomly created, suppressed data sets in a Monte Carlo simulation procedure. Full datasets will be useful for a wide range of applications, including more detailed analysis using the two-stage approach described geographic concentration. In the absence of a well defined approach to imputing suppressed employment records at detail levels of industry and geographic aggregation, the second chapter proceeds by making use of commercially available enhanced data sets.

Findings from the first step of the two-step approach to describing renewable energy sector firm concentration patterns suggest varying degrees of significant concentration for firms engaged in all ten value chains. The utility of this approach was developed through a detailed discussion of findings and implications for the biodiesel and residential solar energy sectors. Global indexes suggest that firms engaged in residential solar energy value chains may experience cost savings in relatively fewer locations than those of the biodiesel sector. The greater variation in the geographic distribution of residential solar business establishments and employment indicates that firms in this sector may be subject to greater spatial variation in the distribution of industry-specific, local “natural advantages” (and their attendant cost savings).

Industries exhibiting significant global concentration were analyzed using local concentration indexes in the second step. Location quotients were used to determine which counties exhibited comparative advantage with respect to attracting renewable energy firms, observed by concentration of renewable energy firms or employment. Based on explorations of the market and policy conditions, and economic landscape in counties exhibiting firm concentration, inferences were made about which local factors may influence firm site selection decisions. Residential solar energy firms appear to be attracted to areas endowed with greater sun exposure and (or) to locations that facilitate access to skilled labor. Similarly, firms supporting the biodiesel sector tend to concentrate near population centers, presumably to access skilled labor, transportation hubs, and feedstock (for production using yellow grease or animal fats). Concentration near soybean-producing locations is also evident, perhaps driven by firms producing biodiesel using oilseeds.

Implications of these results may be useful to policy makers and investors in determining which industries may benefit from characteristics found in relatively few placed (as evidenced by global measures), and which locations are associated with lower risk, in terms of value-added output or variable costs, to firms belonging to a particular industry (from local measures). Potential green energy investors may find that locations exhibiting comparative advantage with respect to attracting firms similar to their own are, perhaps, associated with relatively lower transportation or labor search costs. Therefore, policies designed to expand renewable energy markets or promote local economic growth may be effective when aimed at improving local workforce quality or expanding inter-regional transportation infrastructure. Improvements in logistical efficiency for renewable energy firms or in coordination between upstream and downstream markets may simultaneously reduce the national dependence on fossil fuels and foster economic development.

The two-step approach to analyzing firm concentration has potential for a range of applications beyond that described in this thesis, including more detailed analyses of individual renewable energy value chains, or numerous other industries. The primary limitation that may arise from this approach will likely be a lack of employment data at sufficient levels of disaggregation. The proposed data imputation procedures may be useful for overcoming this limitation, however. As the accuracy of data imputation methods is improved by future research, the utility of the two-step industry concentration analysis procedure will improve as well. Analysis at finer industry and geographic levels will strengthen conclusions about the nature of firm concentration and the spatial distribution of economic activity across the nation.

## Vita

David Lane Register was born in Ocala, Florida to David and Sandy Register. In 2010, he completed a B.S. degree in Food and Resource Economics at the University of Florida in Gainesville, Florida. In May 2012, he obtained an M.S. degree in Agricultural Economics from the University of Tennessee, in Knoxville, Tennessee. Upon completion of his M.S. degree, he returned to the University of Florida to complete a Master of Business Administration degree. He is grateful to his professors, fellow students, friends, and family for their help and support in completing this thesis.