Frequently Asked Questions & Data Sources

1) What types of jobs are included in the SET data counts? Is it just full time employees or are part-time or seasonal employees accounted for somewhere?

The jobs estimated by EMSI (Economic Modeling Specialists International) and given in the SET data include four classes of workers. These include the following:

- QCEW (Quarterly Census of Employment and Wages): The data available from BLS (Bureau of Labor Statistics) is usually suppressed for rural counties. The QCEW data is unsuppressed by EMSI.
- Non-QCEW: Industry sectors not covered by BLS and the unemployment insurance program, such as the railroad workers.
- Self-Employed: Self-employment estimations prepared by EMSI using ACS (American Community Survey), Non-Employer Statistics, etc.
- Extended Proprietors: These estimates include underreported self-employment, proprietorships, trusts, partnerships and cooperatives. EMSI uses LAPI (Local Area Personal Income) from BEA (Bureau of Economic Analysis) and ACS to develop extended proprietor estimates.

EMSI uses BLS, BEA, ACS 5-Year, CBP (County Business Patterns), Non-Employer statistics from Census and several other sources to estimate the total number of jobs. It applies its own un-suppression algorithms to fill-in the suppressed numbers, especially for rural counties. The jobs are given as annual estimates of full-time and part-time job counts. It does not show full-time-equivalent (FTE) jobs. One full-time and another part-time job are shown as two jobs. Similarly, an individual holding two part-time jobs would be counted as two jobs in the EMSI database. It should be noted that EMSI jobs given in the SET data reports are annualized counts and based on “place of work.”

The EMSI jobs data provided in the SET data snapshot, in general, should exceed the jobs data available from public sources, such as the BLS and BEA. The reasons include Non-QCEW industry sectors, Self-Employed and Extended Proprietors.

EMSI updates its data quarterly. However, only a few underlying datasets from the BLS are available on a quarterly basis. The majority of other federal statistics are published annually. Even though some of the underlying data sources in EMSI are updated quarterly, the SET data are annualized jobs data and should not be used for seasonal employment. For SET regions with higher levels of seasonal employment, QCEW from BLS and QWI (Quarterly Workforce Indicators) from the U.S. Census Bureau can be useful resources. For a rural region, it is likely that the publicly available data would be suppressed especially for more granular industry sectors, such as NAICS (North American Industry Classification System) 6-digit codes. In such cases, we advise SET regions to obtain seasonal jobs estimates from the local sources.

2) Why do the public data show Government as a major employer in my region, whereas, Government is not mentioned within the top 10 industry sectors in the SET data?

The publicly available data on industry sectors might show government as a major employer in the region. However, in most cases those are because of public education, public health and jobs in other kinds of public sector, such as community services. Most of the regions with a major public university or a college would show a large employment in government in the BEA or BLS data. Similarly, in all the regions, public school district jobs are included in the government sector. Some regions might have state or local government run hospitals. For example, there are 1,010 state and local government community hospitals in the USA as per American Hospitals Association annual survey of 2013. The local information on jobs in public universities, colleges and school districts and/or public health should be deducted from the government total jobs to estimate the number of employees in various government departments. Rural regions are unlikely to have a large number of government employees unless the regions have state or national parks, federal lands, federal laboratories, Department of Defense (DoD) facilities, etc.

3) What are a few characteristics of the PCRD (Purdue Center for Regional Development) industry clusters used in the SET data?

The PCRD Industry cluster database was developed during 2005-2007 with a grant support from the U.S. Economic Development Administration Program. The cluster definitions were derived based on extensive literature review, the study of cluster templates available from Harvard University (Prof. Michael Porter's work) and the University of North Carolina Chapel Hill (Prof. Ed Feser's work), and also a benchmark cluster definition derived by using the U.S. input output table (transactions matrix) from the BEA. In a few cases, such as forest and wood products cluster, surveys were conducted via a graduate student's Ph.D. thesis from the Forestry and Natural Resources Department at Purdue University. From the outset, the clusters were defined keeping in mind urban as well as rural regions in the USA. Another distinction was that a county was chosen as the smallest spatial unit with the objective that regions are comprised of a group of adjacent counties. The cluster definitions are based on the NAICS 6-digit industry sectors, the most granular industry classification available. The earliest definitions were based on the NAICS 2002 version, which were updated to NAICS 2007 and 2012 versions respectively. The SET industry cluster data is based on the NAICS 2012 version. Over the course of a decade, we saw several changes in the NAICS definitions such as merger and splitting of various industry sectors. Our earliest effort was to derive cluster job numbers by using publicly available data;

however, we found extensive suppression of the data for rural counties, in particular for granular industry sectors. PCRD partnered with EMSI to make use of unsuppressed jobs and wages data for industry clusters. It is to be noted that our industry cluster definitions are not mutually exclusive and an industry sector might appear in the definition of more than one cluster. Similarly, the clusters have broader definitions. For example, the defense and security cluster contains sectors from the defense industries, as well as private and civic security-related businesses. Similarly, health-related sectors and hospitals, are included in the biomedical and biotechnical cluster. The manufacturing supercluster was such a large definition by itself that we subdivided it into six sub-clusters. These sub-clusters are mutually exclusive and an industry sector appears in only one of the six sub-clusters. It is to be noted that cluster definitions are based on export oriented sectors, and the majority of businesses that serve the local economy, such as retail sectors, are excluded from the definitions. The original cluster definition based on NAICS 2002 is available at http://statsamerica.org/innovation/reports/detailed_cluster_definitions.pdf.

4) What are cluster leakages? What does the leakage chart show?

The industry cluster leakage chart is developed by using the "industry supply chain" data available through EMSI's input output estimates for the region. The data focuses on the upstream or backward linkages (purchasing) of the industry sectors contained within the industry cluster definition. Further EMSI develops estimates for the regional purchase coefficients, which shows the proportion of the local demand fulfilled within the region. Hence, the chart shows the total input or regional intermediate demand created by the cluster and the proportions fulfilled within the region versus outside of the region. It is to be noted that these are not MRIO (Multi Regional Input Output) tables, hence we cannot ascertain if the demand was met within the state, other states in the USA or imported from other countries. Please note that an industry sector can simultaneously import and export the same good or service, which is known as cross hauling. Since industry cluster is a group of sectors, cross hauling can exist in specific industry sectors. In case of clusters with higher proportion of demand satisfied from outside of the region, we need to determine if the constituent industry sectors exist or do not exist in the region. According to Deller (2011)[1], if the sectors exist in the region and have the capacity to meet the demand, it is known as "gap" otherwise it is known as the "disconnect." The leakage chart can be helpful in discussing gaps and disconnects and if there are ways to close some of the imports or dollars leaking out of the regional economy.

5) Where are high skilled jobs included? For example, would a geologist be included in mining or in scientific/technical profession or in both?

Skills are associated with occupations and not the industry sectors. A geologist is an occupation and might work both in the mining and scientific and professional services industries. This depends on industry staffing patterns or industry to occupation matrices that are unique for

[1] Deller, Steven. 2012. Targeting Industrial Gaps and Disconnects for Community Economic Development. Choices 27(2).

each of the regions. EMSI estimates industry to occupation matrix at the most granular levels of industries (NAICS 6-digit) and occupations (SOC or Standard Occupation Classification 5-digit) for each of the 3,140 counties in the nation. The SET regions are comprised of counties and hence have regional industry to occupation matrix or staffing patterns. It is to be noted that at the most granular level, a regional economy can have as many as several thousands of sectors, but for occupations, we only have as many as 850 plus occupations.

6) What is the NETS (National Establishments Time Series) database? Why are the job numbers different than the EMSI estimates?

The NETS is an establishment level database and a compilation of the historical D&B (Dun and Bradstreet) records. The data is verified through millions of phone calls every year and primarily it is comprised of NAICS codes, FIPS (Federal Information Processing Standards) or unique county codes, and jobs through the years. Since it is an establishment level data, a business can have multiple establishments in the same county, which is captured by the NETS. Similarly, we have seen a variety of farming establishments, such as green houses, orchards, etc., in the NETS database. We use the temporal data available for FIPS or unique county codes and year-by-year jobs to develop the establishment churn estimating the births, deaths, in-migration and out-migration from the NETS database. We also use the NETS data to estimate sales, establishments and jobs numbers for different stages ranging from Stage 1 to Stage 4 companies. The universe and methodology for NETS and EMSI are entirely different. NETS is mostly self and voluntary reporting of data by companies, whereas EMSI includes publicly available data, un-suppression algorithms and statistical estimates.

As of 2016, we have switched over to YourEconomy.org, a data website based on NETS to estimate the establishments, jobs, and sales. The underlying data remains the NETS and YourEconomy.org has the latest data. This change was necessary as PCRD’s agreement to obtain raw data from NETS through Edward Lowe and Kauffman Foundation has ended. YourEconomy.org provides only the summarized data for NETS.

7) How is the data gathered? Who submits the data? (i.e. businesses? Individuals? Government agency?) Why is this time frame chosen (for any chart/table)? What is the source of the data?

Please refer to the enclosed PowerPoint about data sources used and website addresses for the regional data snapshot and the cluster reports. Most of the data are publicly available federal data sources from the U.S. Census Bureau, Bureau of Labor Statistics, and selected data programs run by these agencies. There are two data sources that are proprietary: EMSI and NETS/Youreconomy.org. EMSI un-suppress the publicly available jobs data for rural counties at the most granular level of industry sectors. It also develops estimates for self-employed and proprietors that are included in the total jobs. The objective is to provide an idea about the total number of jobs available at the regional level except seasonal jobs. EMSI collects data from roughly 80 different federal agencies and provides jobs estimates by industries as well as occupations. The majority of the long-term data starts from 2003, which is the post-recession period after the dot com bust of 2001-2002. Similarly, the majority of the near-term data starts from 2009, which is the post-recession period after the housing bubble. Most of the economy related data from EMSI (clusters, industries, and occupations) are from 2009-2014. We have used 2014 as the end-year because BEA and QCEW, BLS have published finalized job numbers for 2014, which are the underlying data sources for EMSI. Once 2015 is published, we will switch over to 2015 as the end-year. The industry cluster leakage data from EMSI is only available for 2012 and 2013 as those are based on input-output (IO) tables.

8) How are clusters identified? What do they include?

Please refer to FAQ number 3 on details about PCRD's industry clusters. The industry clusters are defined based on industry sectors. In our case, we used NAICS 6-digit industry sector codes to develop the cluster definition. Usually, clusters are defined based on NAICS 3-digit and 4-digit codes, however, in our research we found significant heterogeneity with a NAICS 3-digit or 4-digit code. The clusters are defined at the most granular level.

9) Which data are actual reported numbers vs. algorithm or some other type of estimation? How are estimations determined? On what are they based?

The employment numbers of industries, occupations, and industry clusters are estimates. As explained in FAQ number 1, the jobs are comprised of four parts. QCEW, the first part is obtained from the BLS. Algorithms are used to un-suppress jobs by industry sectors in those counties where the public data has been suppressed to comply with the federal disclosure regulations. Please note that un-suppression algorithms have been discovered and published in academia. Although EMSI has its own proprietary algorithm, we do know that control totals (total jobs at NAICS 2-digit level, county level, state level) are used to ensure that estimated jobs are not off the mark.

Following reference can be used to understand the un-suppressing processes:
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: 311-335.

The 2nd part includes jobs in the non-QCEW activities that are not counted under the QCEW program by the BLS. This includes railroad workers and armed forces. QCEW also does not count self-employed and proprietors. The details of QCEW coverage are included at (http://www.bls.gov/cew/cewover.htm).

The 3rd and 4th parts include estimation for self-employed and proprietors that are explained in detail in FAQ number 1.

10) Are there alternate ways to depict Arts/Entertainment (tourism) cluster that includes the retail sector?

Arts, Entertainment, Recreation and Visitor Industry cluster broadly defines various aspects of tourism including sporting, gaming, historical, amenities, and natural aspects. The U.S. EDA funded industry cluster research was tasked to develop a set of definitions that are applicable nationally and at the county level. Tourism cluster definitions can be specific to the location dependent on unique endowments. Whereas pilgrimage-based tourism or popular gaming locations can attract visitors year long, in many places tourism is seasonal. The challenge with seasonal tourism is that data can only be gleaned at the local level. Similarly, if we want to include retail, we may need to estimate the proportion of retail dependent on or serving the tourism industry. A portion of retail would be serving the local populations. It will be better to define the tourism cluster by specific location with input from the local residents and businesses.

11) Clarify when/where K-12 and public college appears in the data, if at all. What is included in the "Education" cluster?

K-12 and public colleges are actually part of the government sectors. Earlier even EMSI could not differentiate the job numbers for public colleges, public hospitals from other kinds of government jobs. It is only recently that government sector has been classified in much more detail at the NAICS 6-digit level so that we can know local and state level public education and public health jobs. The "Education" cluster definition do not include jobs for K-12 and college. Currently, the cluster definition only includes private education and training related industries. The primary reason was lack of granular data. We use to advise the regions that public education and public health job numbers could be available locally and can be added to the private sector job numbers. K-12 education does not bring outside money to the region but higher public education does, which can be added to the cluster definition if the SET committee agrees.

12) Please explain why retail is excluded.

PCRD cluster definitions were based on industry sectors that were export oriented and bring outside money to the region. Retail usually serves the local population and some part of large retail (big boxes) do attract outside residents. It was difficult for us to determine in general what part of retail was export oriented and hence retail cluster definition was not developed. Prof. Michael Porter, Harvard Univ., has developed a definition for retail industries which they term as the local (non-basic) cluster. The non-basic or those industries that generally serve the local population. PCRD cluster definitions were focused on basic or export-oriented industry sectors.

Data Sources

The following is information about the sources of data used by the PCRD to prepare the regional data snapshot and cluster reports.

- **Census Bureau:**
  Socio-demographic variables are mainly obtained from the Census Bureau through American Fact Finder. (http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t) Here are the list of data tables utilized in this report.

  - Total Population: P001 (2000), P1 (2010), and PEPANNRES (2014);
  - Race/Ethnicity: QPT6 (2000) and PEPSR6H(2014);
  - Population Pyramid: QTP1 (2000) and PEPAGESEX (2014);
  - Educational Attainment: S1501 (2014)

- **General Patent Statistics Reports:**
  The Patent Technology Monitoring Team periodically issues general statistics reports that profile patenting activity at the U.S. Patent and Trademark Office (USPTO). (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm#by_geog)

- **EMSI (Economic Modeling Specialists Intl.):**
  Some regional economic variables are estimated by EMSI. We apply the latest version of data selecting QCEW, non-QCEW, self-employed and extended proprietors in class of workers (http://www.economicmodeling.com/).

- **LAUS (Local Area Unemployment Statistics):**
  LAUS is a U.S. Bureau of Labor Statistics (BLS) program that provides monthly and annual labor force, employment and unemployment data by place of residence at various geographic levels. LAUS utilizes statistical models to estimate data values based on household surveys and employer reports. These estimates are updated annually and used for the SET data reports. Annual county-level LAUS estimates do not include seasonal adjustments. (http://data.bls.gov/cgi-bin/dsrv?la)

- **NETS (National Establishment Time Series):**
  NETS is an establishment-level database, not a company-level database. This means that each entry is a different physical location, and company-level information must be created by adding the separate establishment components. Please refer to the enclosed Frequently Asked Questions (FAQ) for details.

- **Youreconomy.org:**
  The youreconomy.org is an online information tool available from the Edward Lowe Foundation that provides the NETS database. Please refer to the enclosed Frequently Asked Questions (FAQ) for details. Please note that Youreconomy.org would be shifting to Infogroup Business Data from May 1st, 2016. The universe covered by Infogroup data is different than the NETS, which is based on the Dun & Bradstreet historical records.

- **SAIPE (Small Area Income and Poverty Estimates):**
  SAIPE is a U.S. Census Bureau program that provides annual data estimates of income and poverty statistics at various geographic levels. The estimates are used in the administration of federal and state assistance programs. SAIPE utilizes statistical models to estimate data from sample surveys, census enumerations, and administrative records. (https://www.census.gov/did/www/saipe/data/interactive/saipe.html)

- **LEHD (Longitudinal Employer-Household Dynamics):**
  LEHD is a partnership between U.S. Census Bureau and State Department of Workforce Development (DWD) to provide labor market and journey to work data at various geographic levels. LEHD uses Unemployment Insurance Program and Quarterly Census of Employment and Wages (QCEW) data from DWDs and census administrative records related to individuals and businesses. (http://lehd.ces.census.gov/).

- **OTM (On the Map):**
  OTM, a product of the LEHD program, is used in the regional data snapshot report to develop commuting patterns for a geography from two perspectives: place of residence and place of work. At the highly detailed level of census blocks, some of the data are developed synthetically to maintain confidentiality of the worker. However, for larger regions mapped at the county level, the commuteshed and laborshed data are fairly reasonable.

  OTM includes jobs for a worker employed in the reference as well as previous quarter. Hence, job counts are based on two consecutive quarters (six months) measured at the "beginning of a quarter." OTM data can differ from commuting patterns developed from state annual income tax returns, which asks a question about "county of residence" and "county of work" on January 1st of the tax-year. OTM can also differ from the American Community Survey data, which is based on a sample survey of the resident population. (http://onthemap.ces.census.gov/)