IDEC DP² Series Vol. 4 No.9

Development Discussion Policy Paper

Note on Data Cleaning and Panel Data Development of Indonesian Manufacturing Survey Data

Erik Armundito and Shinji Kaneko

December, 2014

Department of Development Policy
Division of Development Science
Graduate School for International
Development and Cooperation (IDEC)
Hiroshima University
1-5-1 Kagamiyama, Higashi-Hiroshima
739-8529 JAPAN

Note on Data Cleaning and Panel Data Development of Indonesian Manufacturing Survey Data

Erik Armundito and Shinji Kaneko

Graduate School for International Development and Cooperation (IDEC)

Hiroshima University

Abstract

When the manufacturing sector plays a significant role in economic development, the demand for statistical data can only increase. Manufacturing data obtained from the annual Indonesian manufacturing survey from 1990 to 2010 cannot be used directly for analysis purposes due to data quality problems. This paper attempts to clean and balance the raw data by proposing several consecutive steps including the statistical modeling and coefficient of variant approaches. The results show that observations for 1,556 firms and 828 firms are obtained each year for the first period panel data and second period panel data, respectively; however, the results should be representative of the entire manufacturing sector in terms of the number of firms and sub-sectors. Furthermore, comparisons with existing data references are needed to verify the results.

Keywords: data cleaning; panel data; Indonesian manufacturing data

JEL Classification Codes: C30, C33, C81

1. Introduction

Data are the basis for all scientific studies and are used by academics, businessmen and practitioners. Collecting good quality data plays a vital role in supplying objective information for identifying problems, improving our analytical understanding of those problems, and thus obtaining appropriate solutions. Using incorrect or inconsistent data can negate the potential benefits of information-driven approaches. Making decisions on the basis of poor quality data is risky and may lead to the distortion of all subsequent analyses and decision making.

Data quality problems, including missing values, duplicate values, misspellings, data inconsistencies, and incorrect data formats, commonly arise in different application contexts and require appropriate treatment to ensure that data and information are reliable. Data cleaning addresses data problems after they have occurred. Error-prevention strategies can reduce but not eliminate many data quality problems. Data cleaning is defined as a three-stage process, involving repeated cycles of screening, diagnosing, and editing of suspected data abnormalities (Van den Broeck, 2005).

A data cleaning approach should satisfy several requirements. First, it should detect and remove all major errors and inconsistencies from individual data sources and when integrating multiple sources. Furthermore, data cleaning should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata. Mapping functions for data cleaning and other data transformations should be specified in a declarative way and be reusable for both other data sources and query processing (Rahm, 2000).

A detailed data analysis is required to detect data errors and inconsistencies that need to be removed. In addition to a manual inspection of the data samples, analysis programs should be used to harvest metadata about data properties and detect data quality problems. After single-source errors have been removed, the cleaned data should replace the dirty data in the original source to provide legacy applications for the data and avoid the need to repeat cleaning work in future data extraction (Rahm, 2000).

The need for data cleaning is centered on improving the quality of data to make them 'fit for use' by users through reducing data errors and improving documentation and presentation. Data errors are common and to be expected (Chapman, 2005); on the other hand, it is important to consider that data collection and observations are often affected by unusual events or disturbances that create spurious effects and result in extraordinary patterns. These unusual values or outliers have adverse effects on understanding the properties of collected data.

Aggarwal (2013) emphasized the definition of an outlier in an available date set. An outlier is described as a particular data point that is significantly different from other data. Outliers are also referred to as deviants, discordant, abnormalities, or anomalies in the statistics and data processing field. Generally, data are formed by one or more creation processes that can represent activity in the system or collected observations. Outliers are generated when the creation process occurs abnormally. Hence, an outlier often encompasses valuable information about the abnormal characteristics of systems and objects that impact data creation processes. The identification of outliers is important in many fields because outliers can contain information that can lead to a process intervention or the prevention of failures and abnormal operating conditions. Thus, there is also a need for effective and efficient methods for outlier detection.

There are various data cleaning approaches available to obtain a reliable data set. Maletic and Marcus (2000) and Basu and Meckesheimer (2007) both suggested that each of the proposed methods has its strengths and weaknesses. Some methods are promising and can be successfully applied to real-world data, while others need improvement.

The Indonesian Statistics Agency (BPS) conducts a manufacturing survey encompassing all manufacturing firms with twenty or more employees on an annual basis for all of the 33 provinces throughout Indonesia. The data set provides comprehensive firm level data for over 22,000 firms. The survey is intended to obtain consistent and accurate manufacturing data to improve national development planning. Because the data are collected from a survey, data quality problems also occur. Data treatment and management are required to build a reliable data set that can be used for any purpose by removing outliers, eliminating missing values, and fixing duplications.

The objective of this paper is to develop a method to clean and balance the raw data from the manufacturing survey, from 1990 to 2000 and from 2001 to 2010, for the purpose of further analysis. Because the data set is in a longitudinal format, data cleaning will result a complete and comprehensive set of panel data consisting of the same firms within periods.

The remainder of the paper is organized as follows. Section 2 presents a brief characterization of Indonesian manufacturing sector data. The methodology is discussed in Section 3. Section 4 presents results and discussion and Section 5 concludes.

2. The Characteristics of Indonesian Manufacturing Survey Data

The BPS's annual manufacturing survey covers medium-size and large-size firms employing 20 or more workers in Indonesia's 33 provinces. Because BPS has branch offices in every province, the survey is conducted simultaneously for all firms in the same period. The data obtained from the survey are expected to have the same characteristics and performance at the particular time of the survey.

The Indonesian manufacturing sector was the engine of growth in the 1980s and for much of the 1990s due to a series of trade reforms following the end of the oil boom. In 1991, the sector's contribution to gross domestic product (GDP) exceeded that of the agricultural sector. Much of the expansion was concentrated in low-skill, labor-intensive, export-oriented industries, and it contributed greatly to a decline in poverty by providing expanded job opportunities. Hence, when the role of the manufacturing sector in economic development increased, the demand for statistical data in the manufacturing sector also increased.

The information obtained from the survey, as illustrated in the survey form, first covers the firms' identity, address, firm status, and location, essentially firm demographics, after which, the following topics are covered:

- *Part I: General Information*, consisting of the main product produced, percentage of capital owned, and number of workers;
- *Part II: Expenses*, consisting of wages of workers, fuel and lubricants, number of generators used, electricity purchased and sold, other expenses, and raw materials;
- *Part III: Production*, consisting of goods produced and percentage of actual production to production capacity;
- *Part IV: Other Income Received*, consisting of manufacturing services received, profit from sale of unprocessed goods, from non-manufacturing services, and from sale of scrap waste;
- *Part V: Fixed Capital*, consisting of estimated value of fixed capital (land, building, machinery and equipment, vehicle, and other), major repair, input costs, output value, and value added.

The numbers of firms reached by the survey varies from year to year. These differences are based on the performance and sustainability of the firm's operation, which are mostly affected by economic conditions at the local, national, and international levels. Table 1 presents the number of firms from 1990 to 2010. There were 16,536 firms surveyed in 1990, which gradually increased to peak at 29,466 firms in 2006, after which the numbers steadily decreased to 22,492 firms in 2010.

Table 1: Number of manufacturing firms from 1990 to 2010

Year	Number of Firms	Year	Number of Firms	Year	Number of Firms
1990	16,536	1997	22,997	2004	20,654
1991	16,494	1998	21,423	2005	20,684
1992	17,648	1999	22,070	2006	29,466
1993	18,163	2000	22,174	2007	27,994
1994	19,017	2001	21,392	2008	25,694
1995	21,551	2002	21,138	2009	24,466
1996	22,385	2003	20,322	2010	22,492

Source: BPS data.

The data set consists of 66 and 23 classifications of manufacturing sub-sectors, based on the 3-digit and 2-digit ISIC (International Standard Industrial Classification) Revision 3, respectively. The 66 3-digit classifications begin with codes 151 to 372, while the 23 2-digit classifications begin with codes 15 to 37. For the purpose of empirical analysis, this paper will only consider 2-digit classification data. Table 2 describes the data from the 2-digit code and classification of manufacturing sub-sectors.

Based on the manufacturing sub-sector classification, the number of furniture and manufacturing firms grew significantly from 1990 to 2000. Approximately 110 new firms were established annually in this sub-sector, followed by the food products and beverages and other non-metallic mineral product sub-sectors, in which approximately 91 and 53 new firms were established annually, respectively. A more detailed description of the number of firms established for each sub-sector from 1990 to 2000 is presented in Appendix I. The number of firms in the food products and beverages sub-sector increased considerably to approximately 71 firms annually during the second period, 2001 to 2010. Followed by the textiles and furniture sub-sectors, these firms added approximately 31 and 20 new firms annually, respectively. Appendix II presents detailed firm numbers from 2001 to 2010.

Table 2: Classification of 2-digit manufacturing sub-sectors

Code	Classifications	Code	Classifications	Code	Classifications
15	Food product and beverages	23	Coal, refined petroleum product and nuclear fuel	31	Electrical machinery and apparatus n.e.c.
16	Tobacco	24	Chemicals and chemical product	32	Radio, television and communication equipment
17	Textiles	25	Rubber and plastics product	33	Medical, precision, optical instruments, and watch
18	Wearing apparel	26	Others non-metallic mineral product	34	Motor vehicle, trailers and semi-trailers
19	Tanning and dressing of leather	27	Basic metals	35	Other transport equipment
20	Wood and product of wood and plaiting	28	Fabricated metal product and equipment	36	Furniture and manufacturing n.e.c.
21	Paper and paper product	29	Machinery and equipment n.e.c.	37	Recycling
22	Publishing, printing and reproduction	30	Office, accounting, and computing machinery		

Source: UNStats.

3. Methodology

3.1. Data

The raw data date from 1990 to 2000 and are obtained from the Annual Indonesian Manufacturing Survey carried out by BPS. The data for the period from 1990 to 2000 consist of an 11-digit identification number, and the data for the period of 2001 to 2010 have 9-digit identification numbers. Because the methods of data collection are slightly different for the two periods, two sets of panel data will be developed.

The data set consists of medium and large sized firms. This grouping is based on the number of workers in each firm, regardless of its amount of capital or output. BPS defines a medium size firm as one with 20 to 100 workers, while large firms are described those with more than 100 workers. Table 3 describes the number of firms and the percentages of medium and large sized firms from 1990 to 2000.

Table 3: The number of medium and large sized firms

Year	All firms number	Medium firms		Large firms	
		Number	percentage	Number	Percentage
1990	16,536	12,006	72.6	4,530	27.4
1991	16,494	11,485	69.6	5,009	30.4
1992	17,648	12,147	68.8	5,501	31.2
1993	18,163	12,344	68.0	5,819	32.0
1994	19,017	13,545	71.2	5,472	28.8
1995	21,551	15,110	70.1	6,441	29.9
1996	22,386	15,855	70.8	6,531	29.2
1997	22,997	16,415	71.4	6,582	28.6
1998	21,423	15,056	70.3	6,367	29.7
1999	22,070	15,497	70.2	6,573	29.8
2000	22,174	15,467	69.8	6,707	30.2
2001	21,392	14,734	68.9	6,658	31.1
2002	21,138	14,476	68.5	6,662	31.5
2003	20,323	13,813	68.0	6,510	32.0
2004	20,656	14,117	68.3	6,539	31.7
2005	20,684	14,199	68.6	6,485	31.4
2006	29,465	22,157	75.2	7,308	24.8
2007	27,997	20,921	74.7	7,076	25.3
2008	25,694	18,938	73.7	6,756	26.3
2009	24,466	17,797	72.7	6,669	27.3
2010	22,492	15,976	71.0	6,516	29.0
Total	454,766	322,055		132,711	

Sources: BPS data and author’s calculation

3.2. Variables Construction

The BPS’s raw data contains much information about manufacturing firms, starting from the number of workers to the total value of firm output. Only selected data points are used and will be considered as variables for the purposes of this analysis. The most important step in data cleaning is determining the significant variables, such as main and intensity variables, to minimize additional effort, because the data set consists of more than 22,000 firms. The main variables selected are described below:

- *Capital* (k), measured as the value of total fixed assets (land, building, machinery and equipment, vehicles, and other).
- *Labor wage* (l), measured as the total salary and other incentives of all workers, including production workers and other workers.
- *Raw material* (m), measured as the total materials used to produce a unit of output, both domestic and imported.
- *Value added* (v), measured as the total value generated from the transformation of raw materials into the final product or finished goods, or the difference between the total sales revenue and the total cost of components, materials, and services.
- *Output* (q), measured as the total value generated from the process of manufacturing activity in the form of goods produced, electric power sold, industrial services, trading profits, stock added, semi-finished goods, and other revenue within a year.
- *Energy consumption* (e), measured as the total energy use to operate manufacturing firm within a year in Tons of Oil Equivalent (TOE), including fuel and electricity used;
- *CO_2 emissions,* (CO_2), measured as the common type of gas emitted from the burning of fossil fuels used in manufacturing firms in tons CO_2 equivalent, calculated from fuel combustion used in the manufacturing sector based on the Intergovernmental Panel on Climate Change (IPCC) guidelines (IPCC 2006, 2006).

The intensity variables are defined as follows:

- *Energy intensity* (e_q), the amount of energy used to produce a single unit of manufacturing output, measured as the ratio of total energy used to total output.
- *Labor productivity* (q_l), the total value of output generated per worker, measured as the ratio of total output to total labor wages.
- *Raw material per output* (m_q), the material used to produce a single unit of output, measured as the total materials used to total output.
- *Value added per output* (v_q), the total value added generated per output, measured as the ratio of total value added to total output.

- *Output per capital* (q_k), the total value of output generated per total value of capital, measured as the ratio of total output to total capital.

The information related to monetary units such as capital, labor wages, raw material, value added, and output were originally recorded in thousands of Indonesian rupiah (IDR). To avoid price changes over time, GDP (Gross Domestic Product) deflators are applied to convert these data series to constant prices based on the year 2000. Appendix III shows the Indonesian GDP deflators from 1990-2000. Additionally, to convert the currency from the Indonesian rupiah to the US dollar, the currency rate from the year 2000 is applied (IDR 9,593 = USD 1).

3.3. Data Cleaning

After the variables are constructed and defined, data cleaning is carried out based on the data held by each variable. Several subsequent steps must be performed, including removing missing and zero values, identifying and removing outliers, and smoothing data trends. The result of data cleaning is a data set without data quality problems.

3.3.1. Removing Missing and Zero Values

A number of variables contained missing and zero values, a common data quality problem during data collection. The first step in cleaning data is to remove these values; however, it should be noted that for some variables, zero values can be a real condition that provides important information.

The main variables that contain monetary values such as capital, labor wages, raw material, value added, and output should not have a zero value. Here, the zero values must be removed because it is unreasonable for a capital or output variable to have a value of zero, which implies that there is no production process taking place. The missing values of these main variables must also be removed because precise information cannot be gleaned from a missing value, whether it is actually a zero value or a non-zero value.

The missing values must be removed from fuel and electricity consumption data because this data will be used to determine the main variable of energy consumption. If the unavailability of fuel and electricity consumption data occurs in a certain year, then all observations for that year will be removed. In this case, zero values for fuel or electricity consumption will not be removed, as several firms use only particular energy sources.

3.3.2. Identifying Outliers

The general method for cleaning data involves two major aspects. The first aspect is identifying which observations in a data set are outliers, and the second aspect is addressing the issue of what to do with an observation that has been identified as an outlier. It is important to consider that an outlier may have two interpretations: it can be noise, or it can be an indication of an anomaly that has a specific cause.

Several approaches to identify outliers can be applied based on the assumptions of modeling the outliers and properties of the underlying modeling approaches. Approaches that are commonly used to identify outliers include the model-based approach, the proximity-based approach, and the angle-based approach, each of which can be further elaborated in greater detail. For instance, the model-based approach can be broken down into the statistical modeling approach, the depth-based approach, and the deviation-based approach. Because the data are collected by an annual survey and consist of a large number of observations, a statistical modeling approach is most appropriate here.

3.3.3. Statistical Modeling Approach

The statistical modeling approach can identify outliers in a large data set. The basic idea of this approach is that given a certain kind of statistical distribution, the parameters of a data set can be estimated assuming that all data points have been generated by such a statistical distribution (mean, median, or mode); thus, outliers are the points that have a low probability of being generated by the overall distribution. The basic assumption of the approach is that normal data objects maintain a distribution and occur in a high probability area of this model, whereas outliers deviate significantly.

Kernel density estimation is employed to obtain a clear description of the original distribution of the raw data. Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable, and is also a fundamental data smoothing problem where inferences about the population are made based on a finite data sample. Figure 1 shows a sample raw data distribution for the capital variable from 1990 to 2000, and Figure 2 depicts a sample raw data distribution for the energy intensity variable from 1990 to 2000. Almost all of the variables for both periods showed a Zipf distribution rather than a normal distribution, similar to that shown in Figures 1 and 2.

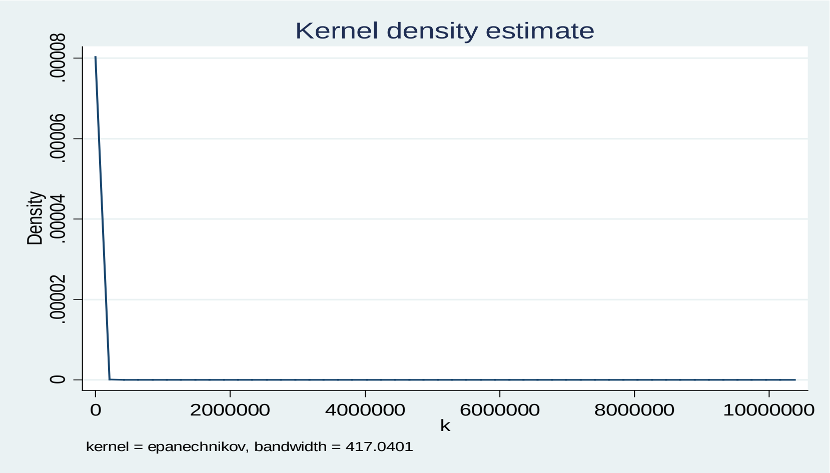


Figure 1: Raw data distribution for the capital variable from 1990 to 2000
Sources: BPS data and author's calculation

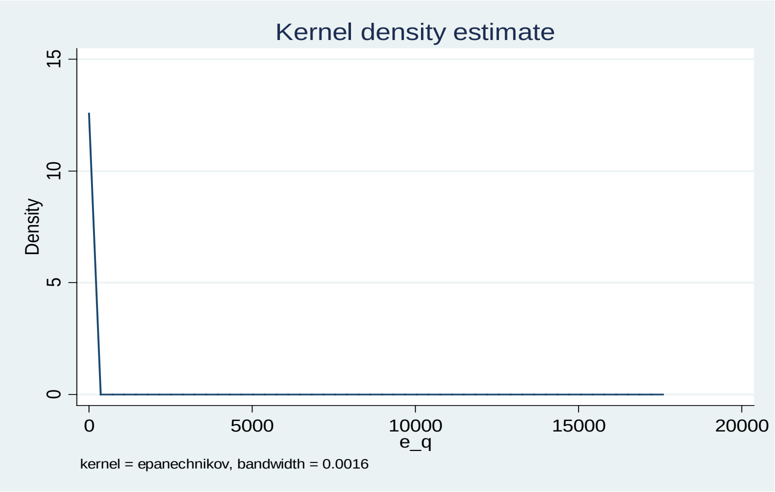


Figure 2: Raw data distribution for the energy intensity variable from 1990 to 2000
Sources: BPS data and author's calculation

Mean, median, and mode, the central tendencies of the statistical distribution of each sub-sector's intensity variables during each period are used and applied to define the range of the intended data set. Multiplying by three certain values, namely 500, 1000, and 1500, the points of the intended data set boundary can be identified. The employment of these three values is a trial and error effort because initially there is no information about the number of outliers in the raw data. Figure 3 illustrates the application of central tendencies multiplied by 500 for the labor productivity variable (q_l) in the food and beverage sub-sector from 1990 to 2000. The mean, median, and mode can be defined in the kernel density distribution graph. Additionally, the points of the central tendencies multiplied by 500 can also be identified.

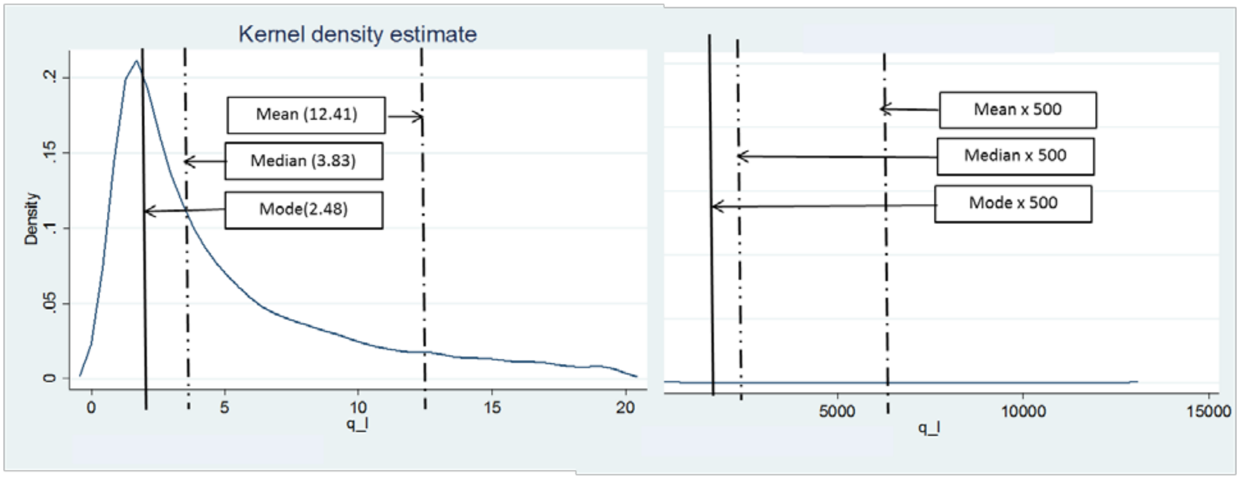


Figure 3: Identification of the boundary points of the labor productivity variable, 1990 to 2000
Sources: BPS data and author's calculation

The purpose of identifying the boundary points in Figure 3 is to obtain the right-hand boundary points of the intended data set. To identify the left-hand boundary points, a similar procedure is carried out using the reverse variable of labor productivity, which is labor per output (l_q). Figure 4 shows the identification of the boundary points of the labor per output variable from 1990-2000.

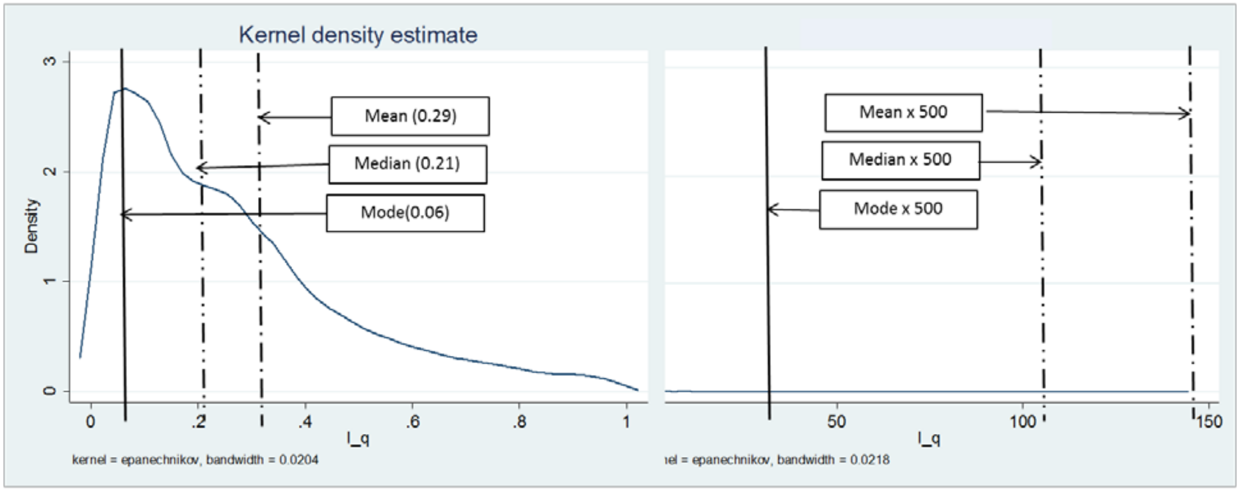


Figure 4: Identification of the boundary points of the labor per output variable from 1990 to2000
Sources: BPS data and author's calculation

The outliers can be identified using these two measures. They are defined as observations beyond the range of the intended data set. Among the three points in the left- and right-hand boundaries, the use of the mode shows that the distribution graph grows stricter. A stricter distribution graph implies that more outliers can be removed, and the more outliers that can be removed, the cleaner the data set is assumed to become. Additionally, among the three certain values, the use of 500 demonstrates that the distribution graph becomes stricter. It must be noted, however, that the more stringent parameters are applied to data cleaning, the less observations are obtained. Consequently, the cleaned data cannot represent the entire manufacturing sector because several sub-sectors will also be removed. For this reason, certain values that are less than 500 are not applied. A similar measure using the mode and the value of 500 is also applied to all intensity variables of all sub-sectors during both periods, from 1990 to 2000 and from 2001 to 2010.

3.3.4. Coefficient of Variant Approach

To further increase the robustness of the cleaned data resulting from the statistical modeling approach, a coefficient of variant (CV) approach is implemented. The coefficient of variant is defined as the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another. CV can also describe the level of the data trend fluctuation.

A single CV value for each sub-sector's intensity variables during each period is used to determine a boundary beyond which observations must be removed. The value of 1 is selected considering that for a Zift or exponential distribution, the standard deviation is equal to its mean. The observations of cleaned data that have a CV value greater than 1 will be removed with the objective of producing a smoother trend line in the cleaned data.

3.4. Panel Data Development

To develop a set of panel data, a firm identification number or firm id is used as a basis from which to synchronize and filter the data set within a period. The firm id is the first character that must be examined for compliance with the standard identification number, and this step will ease the identification of location, sub-sector classification, and particular firms. The firm id takes the form of an 11-digit number for data from 1990 to 2000 and a 9-digit number for data from 2001 to 2010. If a firm id does not comply with the standard identification number, the data must be removed. This firm id examination also seeks to identify whether firms are included in the manufacturing sector category or if they are in other sectors.

Through a balancing process, the data set is stratified by year and compiled with the same firm id. It is expected that some firms with the same firm id exist throughout a period, and firms that do not have the same firm id throughout a period will be removed. By maintaining the same firms across years, the data set is arranged in a longitudinal format. A balanced data set in a longitudinal format is also known as panel data. This process will create two sets of panel data, one for data from 1990 to 2000 and a second for data from 2001 to 2010.

Data cleaning and data balancing produce a data set consisting of the firms that survive. The number of firms for each sub-sector varies based on the number of outliers removed. Further analysis can be carried out if a manufacturing sub-sector comprises more than 15 firms. Otherwise, the results may not represent the entire manufacturing sector.

4. Results and Discussion

Analyzing raw data from a manufacturing survey will return unreliable results, because the raw data commonly has data quality problems. Data cleaning is the first and most important step in any kind of data processing. The purpose of data cleaning is to give researchers access to reliable data so that they can avoid false and misdirected conclusions. After cleaning the data, a data set will be consistent with other similar data sets in a system. Inconsistencies in raw data, including outliers, may have been caused by human error, incorrect formats, or corruption in transmission or storage.

Beginning with removing missing and zero values and identifying outliers, two approaches to cleaning a raw data are applied to obtain a reliable cleaned data set. Then, the cleaned data must be balanced to develop panel data. The results of data cleaning and data balancing are compiled in Table 4. Column A in Table 4 shows the number of annual observations from the BPS survey data from 1990 to 2010, a total of 454,766 observations. Column B shows the result of removing zero values for monetary units of certain variables. In this column, there are zero firm observations for 1996 and 1997 because the capital data were also zero in these years and it is unreasonable for firms to record zero capital values; hence, the observations in those years are removed. The total number of observations after removing zero values is 253,610 or 55.8 percent of the total number of observations in the raw data. Column C exhibits the result of removing missing values for fuel and electricity consumption. There are missing values for the consumption of coal and kerosene in the years 2001 and 2002. The unavailability of coal and kerosene data will have a significant impact on the determination of CO_2 emissions, because coal and kerosene are the main high carbon content fuels. Without coal and kerosene data, CO_2 emissions estimates cannot represent the manufacturing sector's actual emissions. In column B, the number of observations decreased sharply in 2007 because removing zero values eliminated 77 percent of the original observations. To prevent bias from developing in panel data in the future, the 2007 observations are removed. There are 222,062 observations remaining after removing missing values, or 49 percent of the total number of observations in the raw data. Column D is the result of data cleaning using the statistical model and coefficient of variant approaches. In these measures, outliers are identified and removed. Several observations are also removed to reduce the fluctuation in the data trend. After applying these measures, 91,311 observations or 20 percent of the total number of observations in the raw data remain. Column E is the result of data cleaning and data balancing of manufacturing sector data. The firms that do not continuously exist during the study period are removed. The final number of observation is 19,926 or 4.4 percent of the total number of observations in the raw data, which comprises 1,570 firms each year for the first period and 828 firms each year for the second period. Because all observations are removed in the years 1996, 1997, 2001, 2003, and 2007, the developed panel data that can be applied for the purpose of analysis cover from 1990 to 1995 and 1998 to 2000 in the first period, and from 2003 to 2006 and 2008 to 2010 in the second period.

Table 4: The number observation from data cleaning and panel data

Year	A	B	C	D	E
1990	16,536	13,140	13,140	5,429	1,556
1991	16,494	12,926	12,926	6,202	1,556
1992	17,648	14,439	14,439	6,608	1,556
1993	18,163	14,902	14,902	6,764	1,556
1994	19,017	15,488	15,488	7,121	1,556
1995	21,551	17,157	17,157	7,233	1,556
1996	22,386	0	0	0	0
1997	22,997	0	0	0	0
1998	21,423	14,850	14,850	6,312	1,556
1999	22,070	12,845	12,845	6,276	1,556
2000	22,174	13,237	13,237	5,918	1,556
2001	21,392	12,029	0	0	0
2002	21,138	13,059	0	0	0
2003	20,323	13,231	13,231	4,755	828
2004	20,656	13,562	13,562	5,206	828
2005	20,684	13,101	13,101	4,805	828
2006	29,465	11,468	11,468	3,262	828
2007	27,997	6,460	0	0	0
2008	25,694	13,458	13,458	4,961	828
2009	24,466	14,473	14,473	5,543	828
2010	22,492	13,785	13,785	4,916	828
TOTAL	454,766	253,610	222,062	91,311	19,926

Sources: BPS data and author's calculation

Note:
A: raw data
B: result of removing zero values for monetary unit of certain variables
C: result of removing missing values for fuel and electricity consumption
D: result of data cleaning
E: result of data balancing

To acquire other descriptions regarding the number of firms, Table 5 presents the result of data cleaning and data balancing based on sub-sector classifications. The table provides information about the number of observations of raw data, cleaned and balanced data, and the number of observations for the two periods for each sub-sector. In general, the number of observations from 1990 to 2000 decreased more than 90% after being cleaned and balanced. In particular, for several sub-sectors such as coal, refined petroleum products and nuclear fuel; office, accounting, and computing machinery; and recycling, there are zero observation, which implies that no firms were continuously present during the period.

Table 5. The result of data cleaning and data balancing for each sub-sector during 1990-2010

No	**Sub-sector**	**Raw data**	**Cleaned data**		**1990-2000**		**2001-2010**	
		Obs.	**Obs.**	**(%)**	**Obs.**	**(%)**	**Obs.**	**(%)**
1	Food product and beverages	99,924	5,946	6.0	4,329	72.8	1617	27.2
2	Tobacco	19,041	544	2.9	180	33.1	364	65.8
3	Textiles	44,980	1,350	3.0	972	72.0	378	28.0
4	Wearing apparel	44,569	1,335	3.0	810	60.7	525	39.3
5	Tanning and dressing of leather	12,160	356	2.9	279	78.4	77	21.6
6	Wood and product of wood and plaiting	32,459	1,004	3.1	801	79.8	203	20.1
7	Paper and paper product	7,926	299	3.8	243	81.3	56	18.5
8	Publishing, printing and reproduction	12,802	832	6.5	657	79.0	175	20.9
9	Coal, refined petroleum product and nuclear fuel	973	0	0	0	0	0	0
10	Chemicals and chemical product	21,325	1,257	5.9	963	76.6	294	23.4
11	Rubber and plastics product	30,429	1,453	4.8	1,026	70.6	427	29.4
12	Others non-metallic mineral product	36,090	1,968	5.5	1,611	81.9	357	18.1
13	Basic metals	4821	152	3.2	117	77.0	35	23.0
14	Fabricated metal product and equipment	17,624	756	4.3	567	75.0	189	25.0
15	Machinery and equipment n.e.c.	8,358	311	3.7	234	75.2	77	24.8
16	Office, accounting, and computing machinery	166	0	0	0	0	0	0
17	Electrical machinery and apparatus n.e.c.	5,784	224	3.9	189	84.4	35	15.6

18	Radio, television and communication equipment	2,815	9	0.3	9	100	0	0
19	Medical, precision, optical instruments, and watch	1,299	118	9.1	90	76.3	28	23.7
20	Motor vehicle, trailers and semi-trailers	5,389	193	3.6	144	74.6	49	25.4
21	Other transport equipment	6,612	231	3.5	189	81.8	42	18.2
22	Furniture and manufacturing n.e.c.	38,157	1,561	4.1	693	44.4	868	55.6
23	Recycling	1,063	0	0	0	0	0	0
	TOTAL	454,766	19,899	4.4	14,103	70.87	5,796	29.1

Sources: BPS data and author's calculation

In Table 5, the cleaned data are also described in two periods. During the first period, the number of firms producing food products and beverages is the largest among the sub-sectors, consisting of 5,946 firms, followed by the others non-metallic mineral product sub-sector (1,968 firms) and the furniture and manufacturing sub-sector (1,562 firms). The food product and beverages sub-sector remains the largest sub-sector in the second period with 4,329 firms after cleaning. Similarly, the others non-metallic mineral product sub-sector remains the second largest sub-sector with 1,611 firms, while the rubber and plastics product sub-sector moves into third place during the second period.

Further appropriate analysis and examinations can be carried out for the selected manufacturing sub-sectors with adequate firm data. Sub-sectors containing less than 5 firms each year will be eliminated, a condition that is based on the level of sub-sector representatives. In Table 5, sub-sectors with less than 45 total observations in the first period and less than 35 observations in the second period are eliminated. The eliminated sub-sectors are: (i) coal, refined petroleum products and nuclear fuel; (ii) office, accounting, and computing machinery; (iii) radio, television and communication equipment; (iv) medical, precision, optical instruments, and watch; and (v) recycling. Of the 23 sub-sectors, only 18 are eligible for further analysis. Tables 6 and 7 summarize the panel data for the periods from 1990 to 2000 and 2001 to 2010.

Table 6: Summary of cleaned and balanced data for the period from 1990 to 2000

Variable	Obs.	Mean	Std. Dev.	Min	Max
Capital (k)	14004	421369.2	1684540	187.6166	4.31E+07
Labor wage (l)	14004	214768.2	793415.9	450.8496	2.24E+07
Raw material (m)	14004	610644.3	2958692	130.6009	1.18E+08
Value added (v)	14004	336897.6	1977243	176.2376	1.40E+08
Output (q)	14004	1082355	5274943	2339.889	2.86E+08
Energy consumption (e)	14004	135.8695	483.5529	0.0123	12341.48
CO_2 emission (co2)	14004	629.0672	2213.923	0.10296	53397.22
Energy intensity (e_q)	14004	2.6E-04	4.82E-04	2.79E-07	8.52E-03
Labor productivity (q_l)	14004	5.2187	7.336084	0.1987	129.1052

Raw material per output (m_q)	14004	0.5076	0.2216	0.0030	0.9766
Value added per output (v_q)	14004	0.3658	0.1778	0.0026	0.9770
Output per capital (q_k)	14004	4.8315	22.2225	0.0291	1601.159

Source: author's calculation

Table 7: Summary of cleaned and balanced data for the period from 2001 to 2010

Variable	Obs.	Mean	Std. Dev.	Min	Max
Capital (k)	5796	130028.9	461210.3	24.90133	9728380
Labor wage (l)	5796	91192.97	244523.6	235.9944	5709646
Raw material (m)	5796	268055.7	1688736	69.7799	5.12E+07
Value added (v)	5796	130350.2	529225.6	357.9417	1.85E+07
Output (q)	5796	437083.1	2107817	1359.84	5.90E+07
Energy consumption (e)	5796	41.32777	177.5197	0.000897	5089.459
CO_2 emission (co2)	5796	242.2456	1156.871	0.00255	25332.13
Energy intensity (e_q)	5796	1.42E-04	2.03E-04	1.82E-07	3.06E-03
Labor productivity (q_l)	5796	4.4722	5.3809	0.0584	66.22058
Raw material per output (m_q)	5796	0.5233	0.2075	0.0011	0.9764
Value added per output (v_q)	5796	0.3898	0.1911	0.0122	0.9666
Output per capital (q_k)	5796	8.3376	41.2793	0.0197	1668.275

Source: author's calculation

The coefficient of variant method is also employed to show the results of data cleaning and data balancing. Each sub-sector's coefficient of variant is compared across the raw and cleaned data to visually present the trend of a particular variable during both periods. Appendix IV depicts the comparison of the raw and cleaned data of the energy intensity variable for all sub-sectors during the first period, and Appendix V describes the comparison of the same variable during the second period. The trends of the cleaned data are smoother than those of the raw data for all sub-sectors, which indicates that the outliers and other data problems have been removed.

The results of data cleaning and data balancing must be verified to attain a consistent and reliable data set. One appropriate way to verify the result of data cleaning is to compare the data with available references, particularly data sets that apply the same treatment and processes, here, to Indonesian manufacturing data. Unfortunately, references describing the results of data cleaning using Indonesia manufacturing data are rare. One considerable exception in this regard is Manning *et al.* (2012), as shown in Figure 5, which presents the relationship between wages and labor productivity in the Indonesian manufacturing sector from 2006 to 2009 and compares their estimation to data from other countries. The Ln (W) in the *y*-axis is the log of average wages calculated as the ratio of wages and salaries paid by employees (converted to current USD) to the number of employees. The Ln (VA/L) in the *x*-axis is the log labor productivity calculated as the ratio of value added (deflated by constant US dollar from the year 2000) to number of employees.

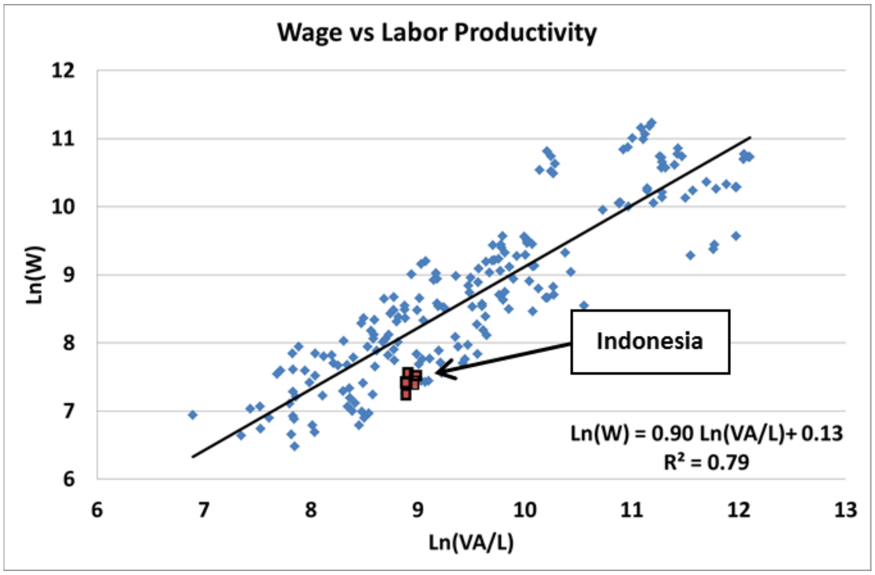


Figure 5. Wage and Labor Productivity of Manufacturing Sectors across Countries, 2006-2009
Sources: INDSTAT in Manning *et.al* (2012), and author's modification.

Figure 6 presents the wage and labor productivity of the Indonesian manufacturing sector based on data cleaning for the years 2006, 2008, and 2009, because the observations in 2007 are removed. Ln (W) is also the log of the average wage calculated as the ratio of wages and salaries paid to employees to the number of employees, and Ln (VA/L) is also the log of labor productivity calculated as the ratio of value added to the number of employees. Applying similar scales to Figure 5, the position of the Indonesian manufacturing sector for the years 2006, 2008, and 2009 can be identified in Figure 6, which is comparable to the reference. In both figures, the Indonesian manufacturing sector is located below the trend line at the coordinates of Ln (W) = 7.5 and of Ln (VA/L) = 8.8.

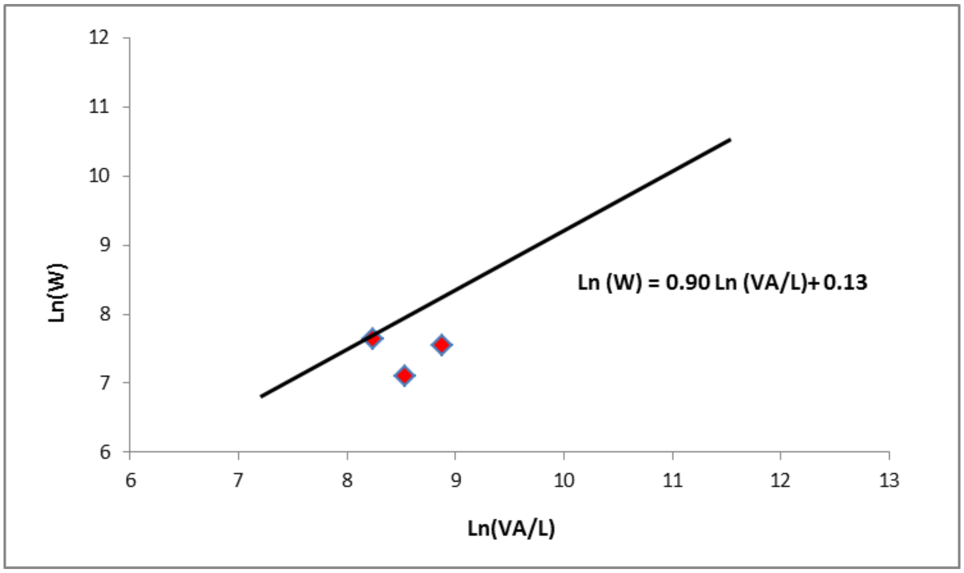


Figure 6. Wage and Labor Productivity of the Indonesian Manufacturing Sector using Cleaned Data for the years 2006, 2008, and 2009.
Source: author's calculation.

To obtain a clearer description of the results of data cleaning, the scales of the figure are expanded such that the *x*-axis and *y*-axis both begin at zero, as shown in Figure 7. The new figure compares the wage and labor productivity of Indonesian manufacturing sector for the raw and cleaned data. In this comparison, the position of the Indonesia manufacturing sector based on the raw data during 2006 to 2009 is far from the trend line, which is located at the coordinates of Ln (W) = 1 and of Ln (VA/L) = 9.5. It is can be assumed that data cleaning and data balancing improve the quality of data set, primarily for the data obtained from a survey. Verification of the result of data cleaning and data balancing can be carried out in various ways by comparing the cleaned data with existing references. However, considering the difficulties in obtaining qualified references for this particular data set, Figure 6 provides the best information available.

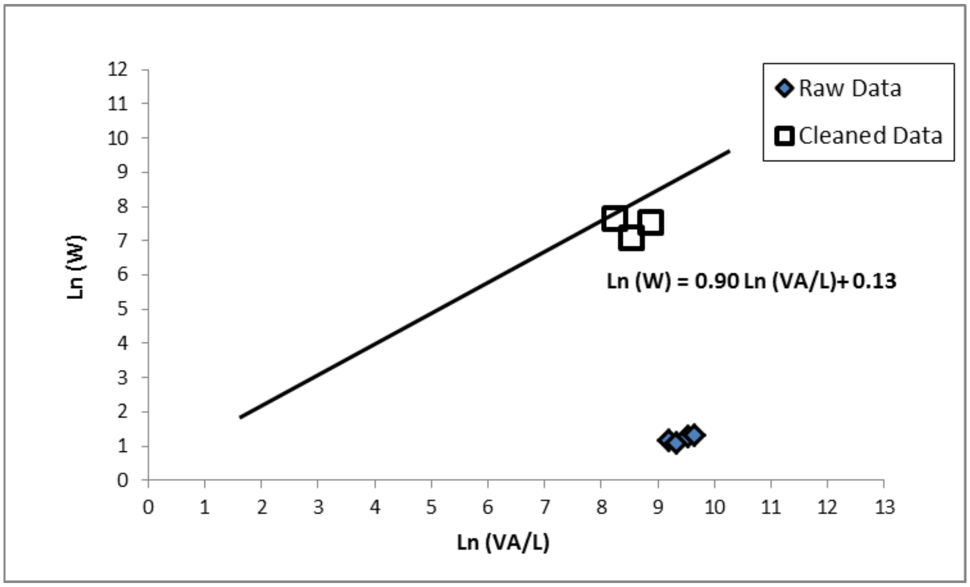


Figure 7. Wage and Labor Productivity of the Indonesian Manufacturing Sector for Raw Data (2006-2009) and Cleaned Data (2006, 2008, and 2009)

Source: author's calculation.

5. Conclusions

This paper carries out data cleaning and data balancing on the annual Indonesian manufacturing survey data from 1990 to 2010 to prepare these data for analysis. The three major findings are summarized below:

- Data quality problems commonly arise from survey data. To cope with these problems, several consecutive steps, including the statistical modeling and coefficient of variant approaches, are applied. It must be noted, however, that the more stringent parameters applied for data cleaning, the fewer observations will be obtained. Consequently, the cleaned data cannot represent the entire manufacturing sector.
- Followed by data balancing, two periods of cleaned panel data for manufacturing are obtained. The first set of panel data consists of cleaned data from 1990-1995 and 1998-2000, and the second set of panel data comprises cleaned data from 2004-2006 and 2008-2010. There are observations for 1,570 firms per year for the first period and for 828 firms per year in the second period.
- Furthermore, to verify the results of data cleaning and data balancing, comparisons with existing references should be made to obtain the same descriptions of a particular variable. Unfortunately, references describe the condition of the Indonesian manufacturing sector are rarely found.

References

Aggarwal, C. C. (2013). Outlier Analysis, Springer;

Basu, S., and Meckesheimer, M. (2007). Automatic Outlier Detection for Time Series: An Application to Sensor Data, *Knowledge and Information Systems*, Volume 11 Issue 2, pp. 137 – 154;

BPS (2008). Large and Medium Industrial Statistics Indicators, *Statistics of Indonesia Catalogue*, No. 6102001;

Chapman, A. D. (2005). Principle and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, version 1.0. *Report for the Global Biodiversity Information Facility*, Copenhagen;

Guyon, I., Matic, N., and Vapnik, V. (1996). Discovering Informative Patterns and Data Cleaning, *AAAI Technical Report* WS-94-03;

IPCC 2006 (2006); IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme, Eggleston H.S., Buendia L., Miwa K., Ngara T. and Tanabe K. (eds). Published: IGES, Japan;

Manning C., Aswicahyono, H., Purnagunawan, M.R. (2012). Labor Demand, Productivity, and Unit Labor Costs in Manufacturing, USAID-SAEDI project, DAI/Nathan Group;

Maletic, J.I., and Marcus, A. (2000). Data Cleansing: Beyond Integrity Checking, *Proceedings of The Conference on Information Quality (IQ2000),* Massachusetts Institute of Technology, October 20-22, pp. 200-209;

Rahm, E., and Do, H. H. (2000). Data cleaning: Problems and Current Approaches, *IEEE Database Engineering Bulletin*, Vol. 23, pp. 3-13;

UNStat, United Nations Statistics Division, http://unstats.un.org;

Van den Broeck, J., Cunningham, S. A., Eeckels, R., and Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities**.** *PLoS Medicine,* 2(10), e267;

Varian, Hal R., (2010); Intermediate Microeconomics: A Modern Approach, 8th Edition, *W. W. Norton & Company*, New York – London;

Appendix I: Number of manufacturing firms based on 2-digit ISIC Revision 3 from 1990 to 2000

Code	Classification	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
15	Food product and beverages	3,655	3,516	3,790	3,943	4,078	4,521	4,670	4,769	4,573	4,666	4,661
16	Tobacco	961	943	902	880	748	815	874	839	785	807	821
17	Textiles	1,828	1,794	1,881	1,953	2,017	2,242	2,173	2,255	2,188	2,055	2,027
18	Wearing apparel	1,766	1,699	1,870	1,798	1,862	2,110	2,159	2,329	1,764	2,214	2,258
19	Tanning and dressing of leather	364	442	481	507	544	606	610	646	600	603	587
20	Wood and product of wood and plaiting	1,357	1,269	1,422	1,491	1,599	1,767	1,692	1,793	1,747	1,779	1,766
21	Paper and paper product	184	217	258	268	305	311	345	359	403	433	431
22	Publishing, printing and reproduction	566	538	548	555	577	645	704	732	535	533	540
23	Coal, refined petroleum product and nuclear fuel	5	9	13	13	12	25	39	37	58	66	57
24	Chemicals and chemical product	864	814	852	892	922	1,008	1,041	1,035	1,055	1,067	1,087
25	Rubber and plastics product	1,190	1,170	1,233	1,249	1,302	1,379	1,481	1,509	1,304	1,371	1,392
26	Others non-metallic mineral product	1,323	1,393	1,461	1,498	1,603	2,027	2,064	2,158	1,948	1,880	1,907
27	Basic metals	161	179	200	216	226	257	283	265	232	225	221
28	Fabricated metal product and equipment	566	584	617	646	722	870	888	969	833	880	892
29	Machinery and equipment n.e.c.	259	271	324	343	365	431	504	485	326	348	347
30	Office, accounting, and computing machinery	7	6	8	9	6	6	6	6	8	8	8
31	Electrical machinery and apparatus n.e.c.	197	216	250	269	319	367	459	393	245	257	259
32	Radio, television and communication equipment	6	9	12	17	19	21	19	22	227	234	227
33	Medical, precision, optical instruments, and watch	51	55	64	59	65	72	62	73	75	63	61
34	Motor vehicle, trailers and semi-trailers	196	204	220	235	241	259	279	279	232	244	246
35	Other transport equipment	242	245	271	279	296	320	322	340	304	320	312
36	Furniture and manufacturing n.e.c.	788	921	971	1,043	1,189	1,492	1,711	1,704	1,909	1,949	1,989
37	Recycling	0	0	0	0	0	0	0	0	72	68	78
	TOTAL	**16,536**	**16,494**	**17,648**	**18,163**	**19,017**	**21,551**	**22,385**	**22,997**	**21,423**	**22,070**	**22,174**

Sources: BPS data

Appendix II: Number of manufacturing firms based on 2-digit ISIC Revision 3 from 2001 to 2010

Code	Classification	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
15	Food product and beverages	4,562	4,543	4,419	4,649	4,718	6,619	6,345	6,078	5,888	5,344
16	Tobacco	811	813	785	807	860	1,282	1,204	1,124	1,044	902
17	Textiles	1,978	1,955	1,916	1,954	1,973	2,968	2,883	2,599	2,454	2,322
18	Wearing apparel	2,055	1,947	1,841	1,860	1,911	3,159	2,952	2,497	2,351	2,177
19	Tanning and dressing of leather	560	540	502	492	475	792	738	688	665	641
20	Wood and product of wood and plaiting	1,739	1,693	1,488	1,437	1,358	1,841	1,704	1,487	1,311	1,159
21	Paper and paper product	345	342	343	357	365	489	469	429	416	401
22	Publishing, printing and reproduction	592	582	568	582	595	947	896	782	733	540
23	Coal, refined petroleum product and nuclear fuel	40	39	42	40	39	61	59	65	62	59
24	Chemicals and chemical product	1,027	1,014	994	1,005	1,007	1,179	1,135	1,075	1,050	1,006
25	Rubber and plastics product	1,493	1,503	1,464	1,491	1,473	1,826	1,760	1,667	1,624	1,550
26	Others non-metallic mineral product	1,657	1,621	1,529	1,513	1,536	2,075	1,952	1,813	1,725	1,613
27	Basic metals	239	237	231	230	226	304	290	273	274	267
28	Fabricated metal product and equipment	906	909	865	861	821	1031	985	902	880	849
29	Machinery and equipment n.e.c.	562	538	508	508	494	565	552	530	505	487
30	Office, accounting, and computing machinery	11	11	10	10	10	14	11	10	10	10
31	Electrical machinery and apparatus n.e.c.	239	243	243	242	238	273	260	247	239	230
32	Radio, television and communication equipment	85	127	138	147	141	193	188	184	178	172
33	Medical, precision, optical instruments, and watch	69	70	63	60	58	62	61	60	58	56
34	Motor vehicle, trailers and semi-trailers	249	280	276	272	270	319	305	296	289	282
35	Other transport equipment	296	306	297	297	285	345	339	317	307	290
36	Furniture and manufacturing n.e.c.	1,876	1,821	1,796	1,830	1,820	3,044	2,821	2,489	2,328	2,099
37	Recycling	1	4	4	10	11	78	85	82	75	36
	TOTAL	**21,392**	**21,138**	**20,322**	**20,654**	**20,684**	**29,466**	**27,994**	**25,694**	**24,466**	**22,492**

Sources: BPS data

Appendix III: Indonesian GDP deflator

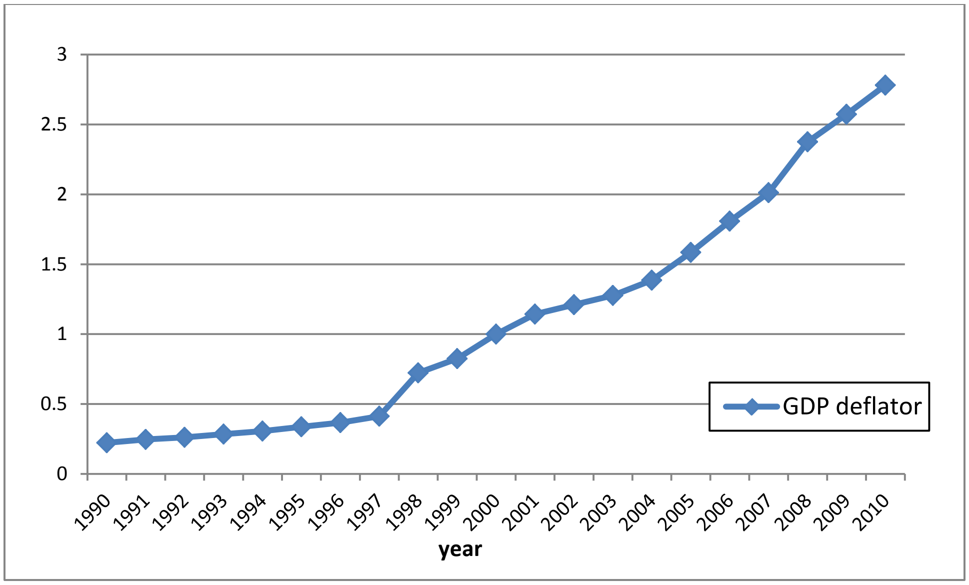
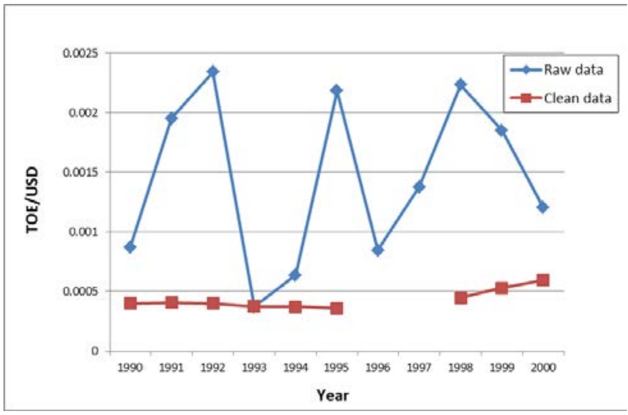


Figure III.1. Indonesian GDP deflator, 1990-2000
Source: http://www.econstats.com/

Appendix IV: Comparison of raw data and cleaned data of energy intensity variable for all sub-sectors, 1990-2000

(Sources: BPS data and author's calculation).



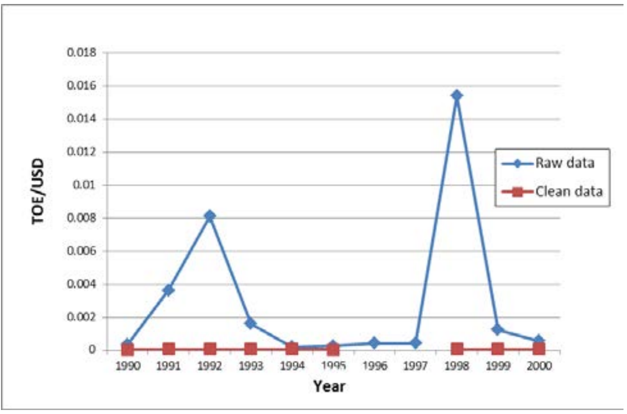
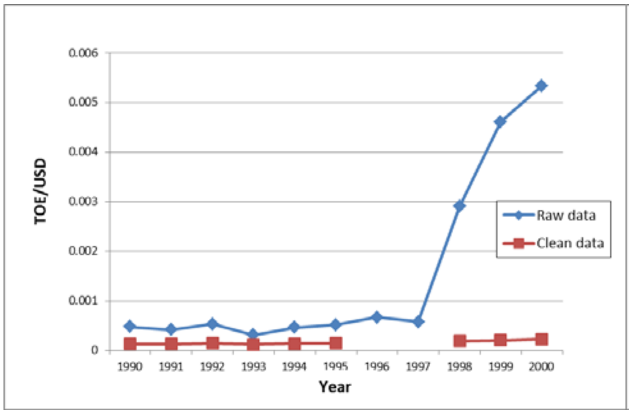


Figure IV.1 Food product and beverage

Figure IV.2. Tobacco



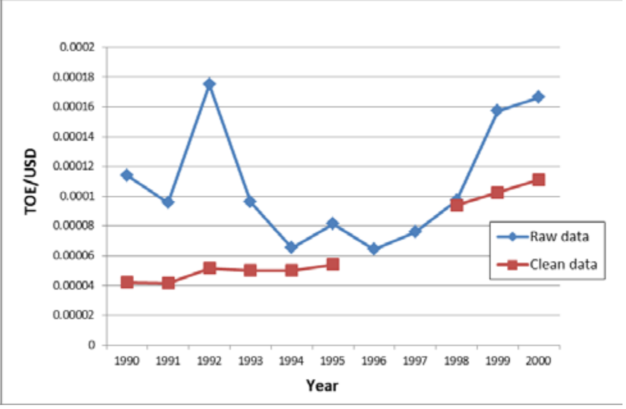
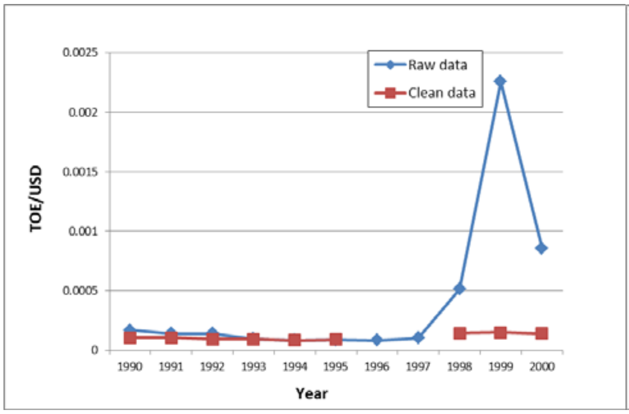


Figure IV.3 Textile

Figure IV.4. Wearing apparel



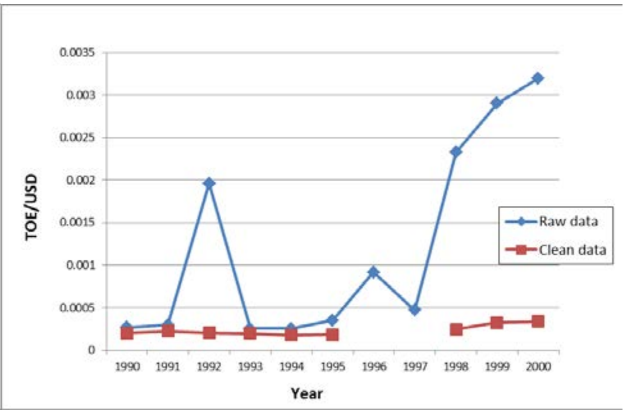


Figure IV.5 Tanning and dressing of leather

Figure IV.6. Wood and product of wood and plaiting

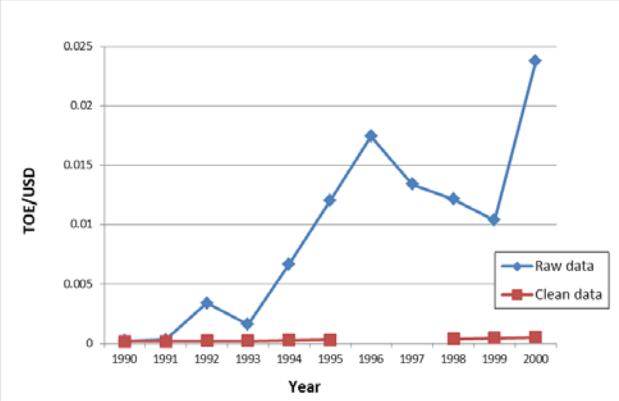


Figure IV.7. Paper and paper product

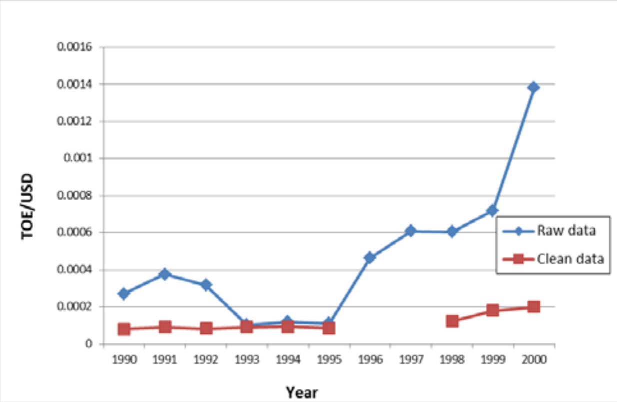


Figure IV.8. Publishing, printing and reproduction

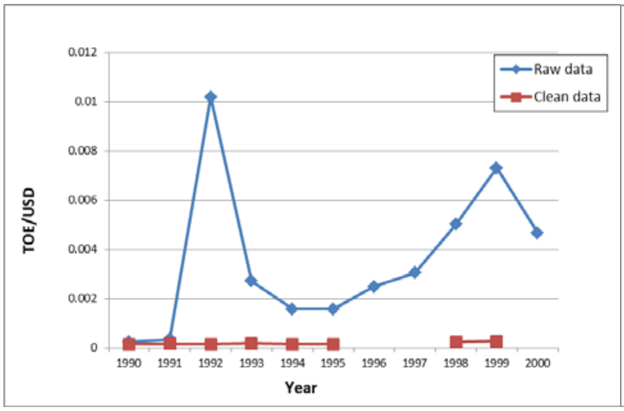


Figure IV.9. Chemicals and chemical product

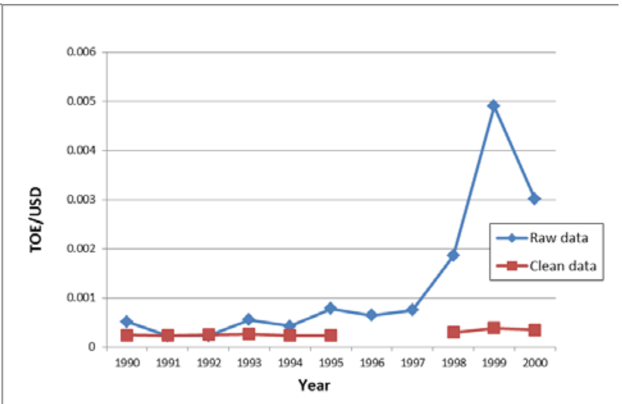


Figure IV.10. Rubber and plastics product

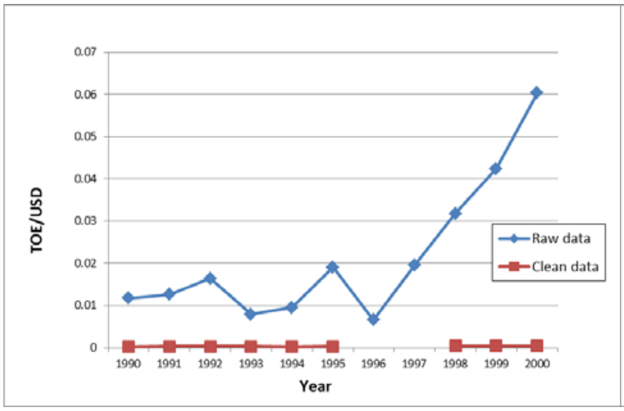


Figure IV.11. Others non-metallic mineral product

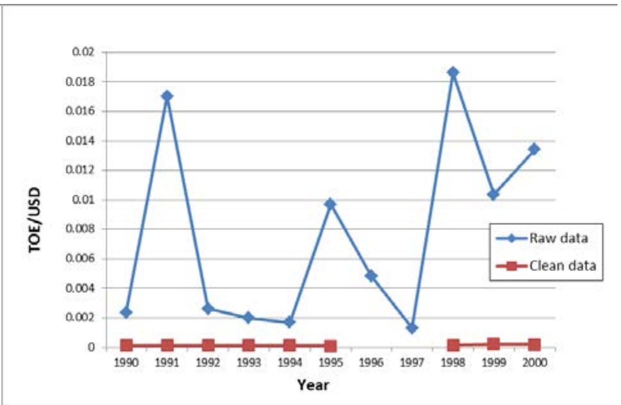


Figure IV.12. Basic metals

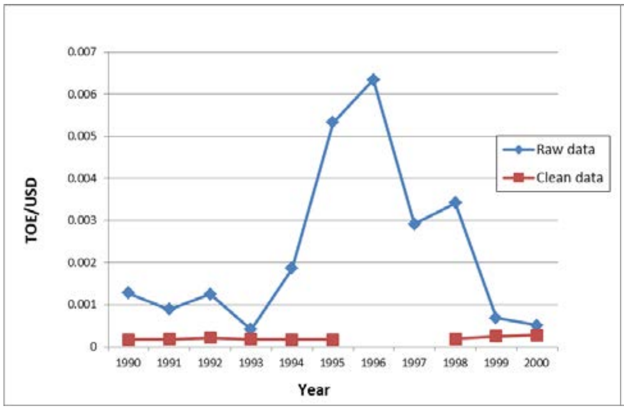


Figure IV.13. Fabricated metal product and equipment

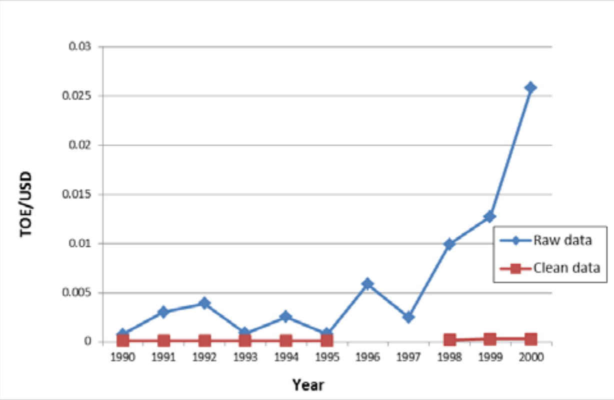


Figure IV.14. Machinery and equipment

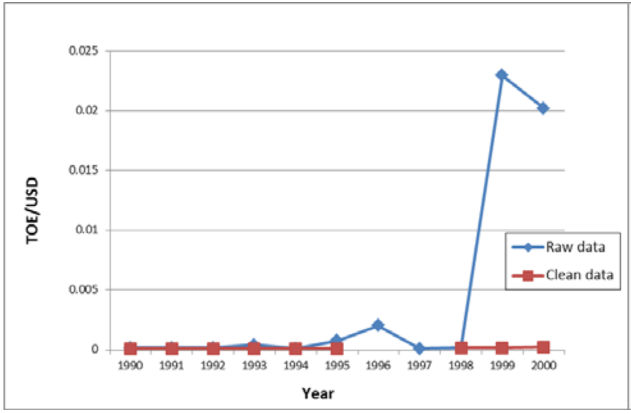


Figure IV.15. Electrical machinery and apparatus

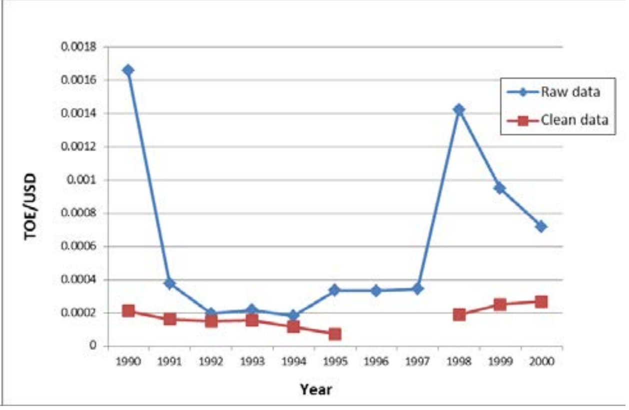


Figure IV.16. Motor vehicle, trailers and semi-trailers

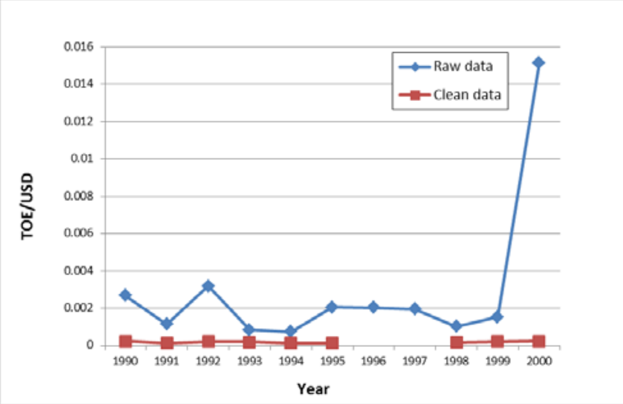


Figure IV.17. Other transport equipment

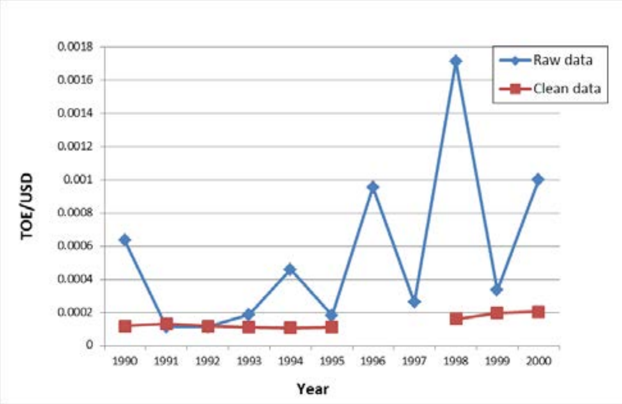
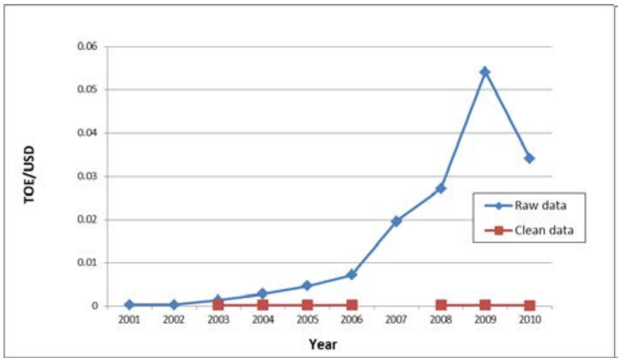


Figure IV.18. Furniture and manufacturing

Appendix V: Comparison of raw data and cleaned data of energy intensity variable for all sub-sectors, 2001-2010

(Sources: BPS data and author's calculation).



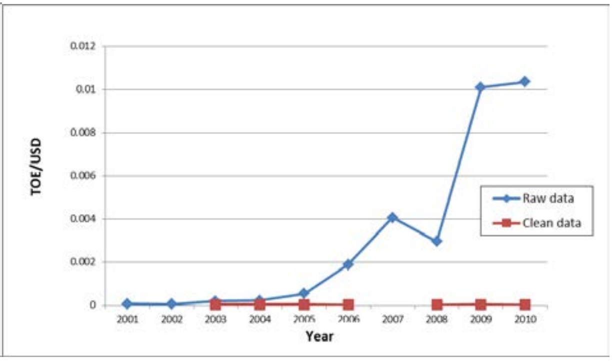
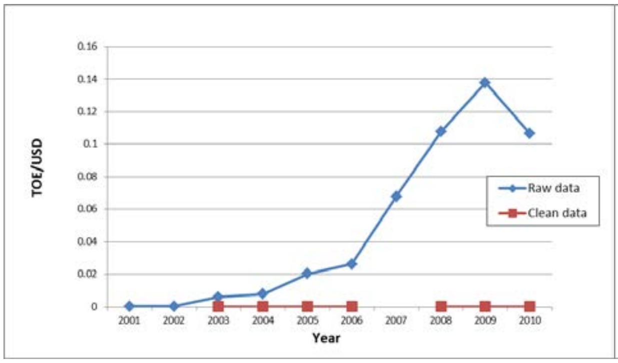


Figure IV.1 Food product and beverage

Figure IV.2. Tobacco



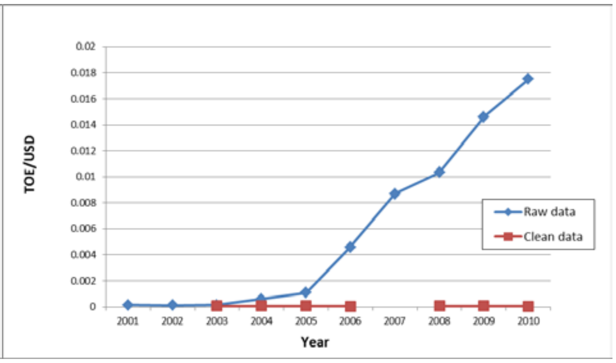
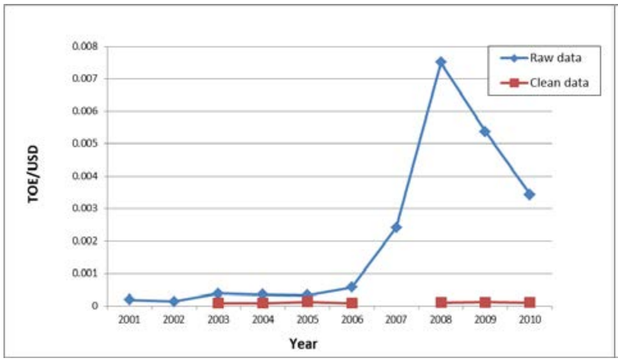


Figure IV.3 Textile

Figure IV.4. Wearing apparel



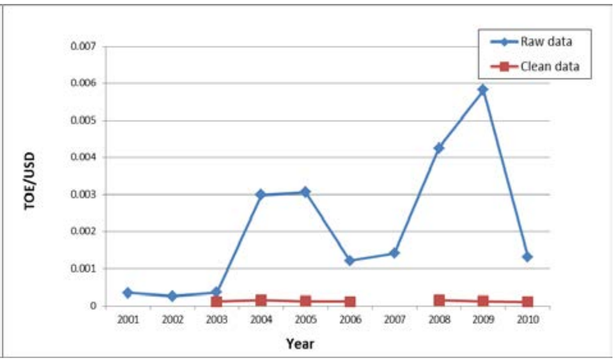


Figure IV.5 Tanning and dressing of leather

Figure IV.6. Wood and product of wood and plaiting

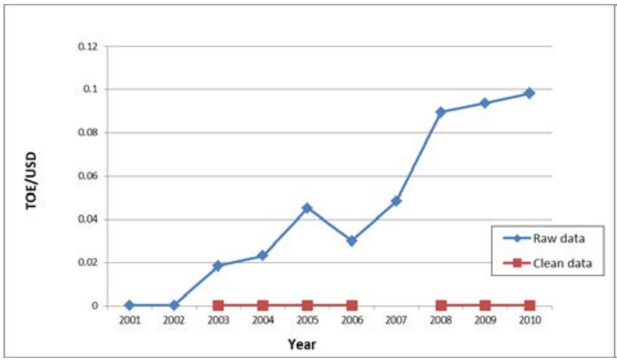


Figure IV.7. Paper and paper product

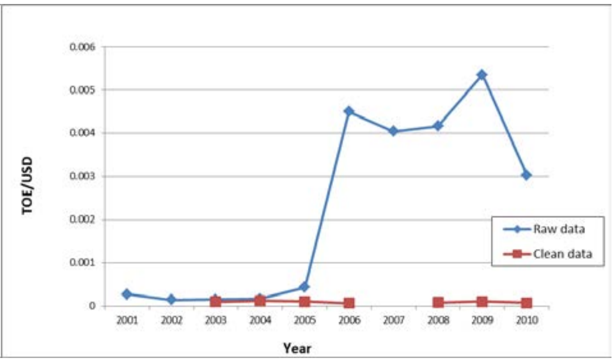


Figure IV.8. Publishing, printing and reproduction

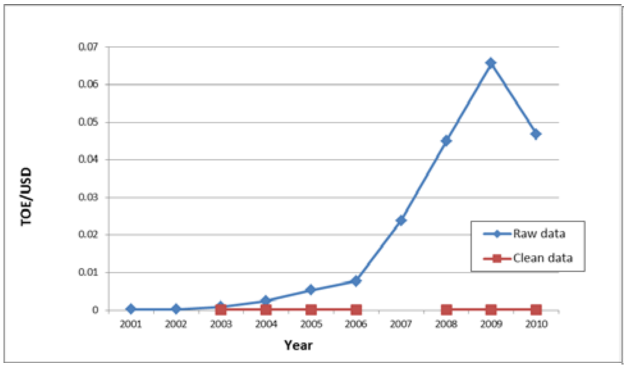


Figure IV.9. Chemicals and chemical product

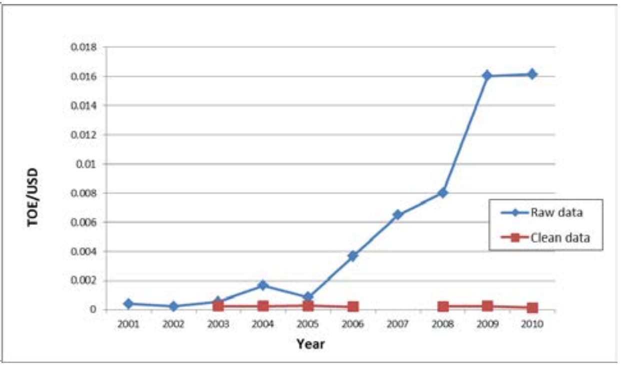


Figure IV.10. Rubber and plastics product

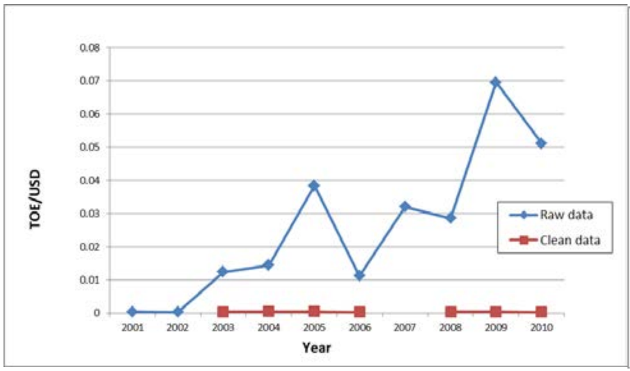


Figure IV.11. Others non-metallic mineral product

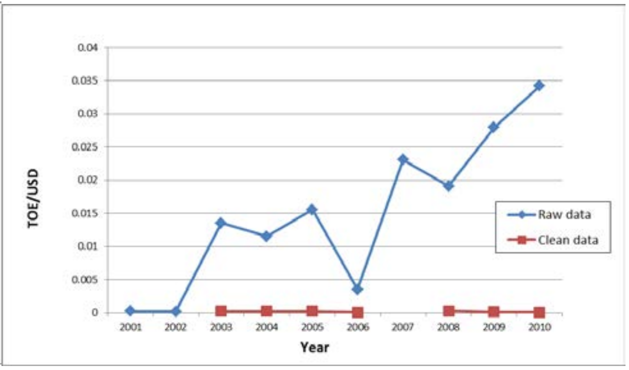
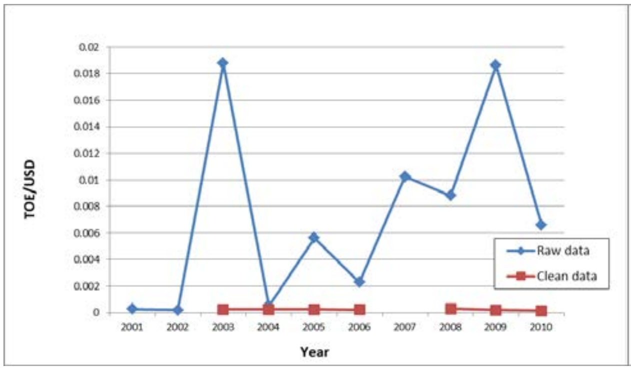


Figure IV.12. Basic metals



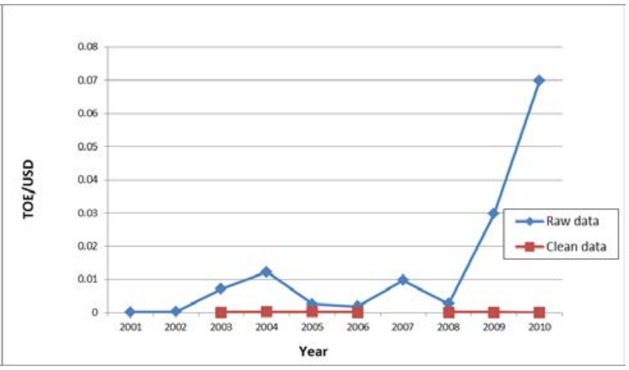
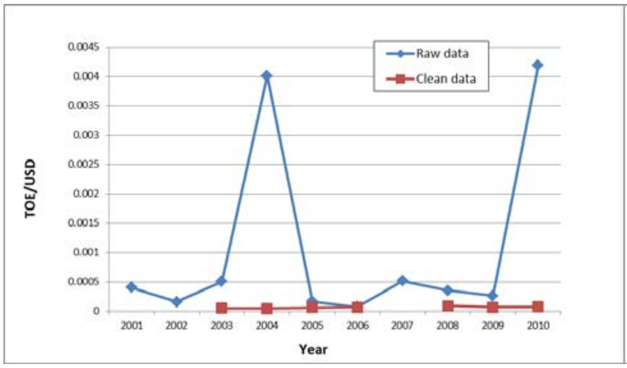


Figure IV.13. Fabricated metal product and equipment

Figure IV.14. Machinery and equipment



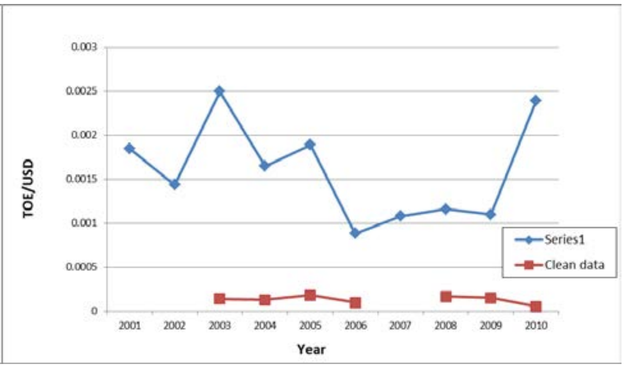
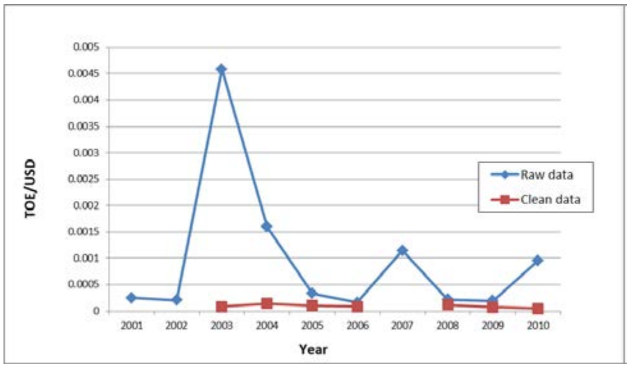


Figure IV.15. Electrical machinery and apparatus

Figure IV.16. Motor vehicle, trailers and semi-trailers



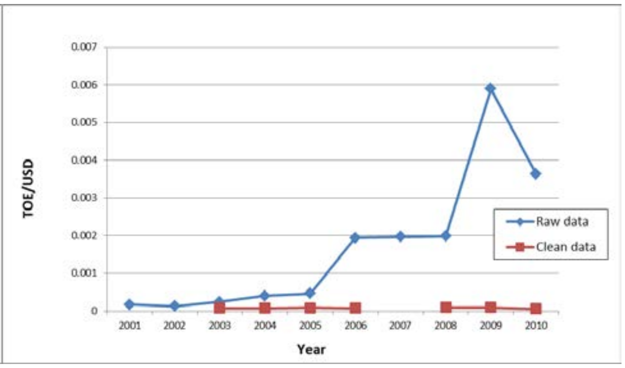


Figure IV.17. Other transport equipment

Figure IV.18. Furniture and manufacturing