CHAPTER 6

DATA SET

6.1 DATABASES

I use three sources of information to construct my dataset. The core information comes from the Municipal Government of Montevideo (Intendencia Municipal de Montevideo, IMM). This is comprised of information regarding production and pollution of industrial plants, plus information regarding monitoring and enforcement activity of the IMM on these plants. The information on production and pollution is obtained from the four-month reports of the plants, described in Chapter 2. The information on inspections is comprised of the number of sampling and non-sampling inspections done per month per plant, and the result of the sample in terms of mg/l of BOD_5. The information on fines levied by the IMM is comprised of the number of fines levied on each industrial plant per month and their amounts. The sample period for all these variables is July 1996 – October 2001, except for fines, which is May 1997 – October 2001.

My second source of information is the Environmental Control Division (DCA) of the Ministry of the Environment. This information includes number and results (in terms of BOD_5 mg/l) of sampling inspections, and number of non-sampling inspections. It also includes the total number of compliance orders issued by the DCA. Types of orders include: an order to present the "Application for the Industrial Discharge Authorization" form (Solicitud de Autorización de Desagüe Industrial, SADI); an order

to present periodic reports of the treatment plant performance; an order to finish the construction of the treatment plant; an order to present the "Start of Operation Report" (Informe de Puesta en Operación, IPO); an order to designate a competent professional responsible for the treatment plant operation; an order to present a "Technical Report" (Informe Técnico, IT), and an order to present modifications to the treatment plant. The DCA sometimes deferred the due dates set in the process of application for the Industrial Discharge Authorization and orders. The data also includes information on these postponements granted by the DCA. Past the due date, the DCA issues a note communicating to the firm that it is potentially subject to a fine due to non-compliance with the previous order. I called this type of action "fine threats". In the case of fines, I have both the number of fines per month per plant and the amount. The sample period for all the DCA variables is June 1996 – October 2001.

Finally, my third source of information is the private partnership MULTISERVICE-SEINCO-TAHAL (SEINCO) that was in charge of the Monitoring Program that the IMM implemented in 1998 as part of the Third Stage of the Urban Sanitary Plan (Plan de Saneamiento Urbano – Tercera Etapa, PSUIII), financed by the IADB. The main objective of the Monitoring Program was described in Chapter 2. This information is comprised of the number and result of the sampling inspections conducted by SEINCO. The period during which SEINCO inspected plants was April 1999 – September 2001.

Table 6.1 gives a summary of all the information just described.

Table 6.1

Data set description

	IMM	DCA	SEINCO
Monitoring And Enforcement Variables	# of sample inspections # of non-sample inspections Result of BOD_5 sampled # of Fines Amount of the fines	# of sample inspections # of non-sample inspections Result of BOD_5 sampled # of compliance orders # of postponements # of Fines Amount of the fines	# of sample inspections Result of BOD_5 sampled
Period	July 1996 – October 2001 except Fines (from May 1997)	July 1996 – October 2001	April 1999 – September 2001

My database includes seventy-four (74) industrial plants located in Montevideo. The selection of these 74 plants is not random. First, they are all privately owned plants. Public industrial plants do not report emissions to the IMM. Second, they were selected from a list of industrial plants that were being sampled by SEINCO during the years 2000 and 2001. Most of these plants were also the ones that were regularly inspected by the UEI. The number of plants in the list did not remain fixed during the consulting period of SEINCO.[40] From a maximum of eighty-seven plants, I excluded twelve (12) plants that reported less than six (6) times during the thirteen (13) reporting periods in my sample although they were active throughout the 13 periods. From the remaining 75 I had to exclude one more because it was not reporting BOD_5 emissions; it reported only metals

[40] For example, it included a maximum of eighty-seven (87) industrial plants in November 2000 – February 2001. In March- June 2001 there were seventy-eight (78) firms in the list. According to SEINCO employees interviewed, the reasons for this change were that some plants closed and others were inactive during some periods. In these cases, "next plants in the list" of the most important polluters of the city (constructed as a result of a previously performed census) were included and inspected.

emissions. Consequently, conclusions from my analysis must be interpreted according to this sample selection bias. It can be said though, that this bias is intrinsic to this type of empirical analysis. It can also be said that, in spite of the latter, plants in the list are responsible for more than 90% of the total industrial organic pollution in the city.

I conclude this section by presenting the descriptive statistics for the monitoring and input variables in Table 6.2 and the enforcement variables in Table 6.3.[41]

Table 6.2: Descriptive Statistics for Input and Pollution Variables

(Sample July 1997 – October 2001)

Total Potential Observations: 3,848

Variable	Mean	Median	Std. Dev.	Missing Values
BOD_5 (mg/l)	1,031	370	2,334	952
Effluent flow (m3/day)	203	52	453	1,034
Tap water (m3/month)	3,848	784	8,271	638
Underground water(m3/month)	2,793	750	4,873	1,279
Electricity (Kwh/month)	179,409	68,000	278,828	449
Fuel (m3/month)	34	12	50	862
Days worked (per month)	22	23	4.6	594
Number of employees)	122	60	276	342

[41] Descriptive statistics for the levels of production are not presented for space reasons. Also, gas and firewood consumption are not included in the table. The IMM did not ask firms to report gas consumption before 2001, and in 2001 only one plant reported gas consumption in two reporting periods. The problem with firewood is that not all of the industrial plants in the sample use firewood as an input and not all of those who did not use it reported zero consumption. Instead, a value was missing in the respective cell. Thirteen plants did not report firewood consumption for the entire sample period, and 32 plants alternated non-reports of firewood consumption with zero consumption, suggesting that in fact they were not using firewood as an input. Given these, I discarded these two variables from the analysis.

Table 6.3: Descriptive Statistics for Monitoring and Enforcement Variables

IMM and DCA

(Sample July 1996 – October 2001)

Total Observations: 4,736

	Units of Measure	Mean	Std. Dev.	Maximum	Sum	Number of Plants
			IMM			
Sample Inspections	#	0.085	0.286	3	401	74
Result (BOD_5)	(mg/l)	1,582	3,894	49,925		74
Non-sample Inspections	#	0.031	0.212	6	148	74
Total Inspections	#	0.116	0.378	9	549	74
Inspections	Dummy	0.106	0.308	1	502	74
Fines	#	0.003	0.052	1	11	74
Fine (UR)	$	93.6	70	200	1030	74
			DCA			
Sample Inspections	#	0.026	0.158	1	122	74
Result (BOD5)	(mg/l)	1,102	1,720	10,400		74
Non-sample Inspections	#	0.019	0.137	2	89	74
Total Inspections	#	0.045	0.210	2	211	74
Inspections	Dummy	0.044	0.204	1	207	74
Compliance Orders	#	0.024	0.155	2	112	74
Postponements	#	0.013	0.123	2	60	74
Fine threats	#	0.015	0.126	2	72	74
Fines	#	0.001	0.029	1	4	74
Fine (UR)	$	225	50	300	900	74
			SEINCO			
Simple Inspections	Dummy	0.180	0.384	1	663	71
Results (BOD_5)	(mg/l)	1,184	2,545	38,000		71

Notes: (1) Observations for fines levied by the IMM were available from May 1997 (3,996 observations).

(2) Statistics for fines variables are over the non-zero observations

(3) "UR" stands for "Unidad Reajustable", a monetary unit indexed by wages. Its value was approximately US$ 15 in October 2001.

6.2 MISSING VALUES

As evidenced by Table 6.2, I have missing values (MV) in my panel. Observations are missing either because a plant did not report in a given period, in which case I have a missing value for the entire set of variables for that period, or because the report had missing values for one or a subset of variables. I call the first case "unit non-report" and the second case "item non-report".

There were four main reasons for unit non-report. First, the plant went out of business. Second, the plant reported no activity in that period.[42] Third, the plant had not yet started business in that period. Finally, the plant simply failed to submit a report for unknown reasons.

Table 6.3 shows that there were a total of sixty-two (62) non-reports over a potential 962 observations (74 plants times 13 reporting periods). Six of these correspond to four plants that ceased production (for different reasons). Twelve correspond to "no-activity" periods of three different plants. Sixteen (16) correspond to three plants that started business in periods four, five and nine, respectively. The remaining twenty-eight (28) correspond to "random" non-reports.

[42] I treated these as missing values because in some cases the firms indicated (usually in a letter to the Director of the Industrial Effluents Unit of the IMM) that they were producing "very low" quantities and therefore it was not worth reporting emissions. Even more, in one case the letter was followed by three non-reports in the following periods without any clear information regarding the exact point in time in which production re-started.

Table 6.4: Distribution of Reporting Failures by Reason

Reason	Number of Non-Reports	Number of Plants
"Ceased Production"	6	4
"No Activity"	12	3
"Not in Business Yet"	16	3
"Random"	28	13
Total	62	23

Tables 6.4 and 6.5 break down the distribution of these non-reports even further. Table 6.4 shows the frequency distribution of the number of reporting failures by the number of industrial plants. Table 6.5 shows the frequency distribution of the number of reporting failures by reporting period.

Table 6.5: Distribution of Reporting Failures by Number of Industrial Plant

									Total
Number of Non-Reports	8	6	5	4	3	2	1	0	62
Number of Plants	1	2	2	2	4	2	8	53	74

Table 6.6: Distribution of Reporting Failures by Period

Period	Number of Non – Reports	Period	Number of Non – Reports
1	10	8	3
2	6	9	2
3	6	10	4
4	5	11	6
5	2	12	6
6	2	13	8
7	2	Total	62

There are several reasons for item non-reports. One is that some firms never report a specific variable. Others report a specific variable unsystematically. For example, in the case of underground water consumption some firms report zero consumption in some periods and do not report in others. Finally, other values appear to be randomly missing.

Taking into consideration item and unit non-reports there were a total of 5,557 observations missing for the inputs and pollution variables described in Table 6.2 plus the production variables reported by the industrial plants, out of a total of 40,924 possible observations. In other words, 13.6% of the data set was missing.

6.3 DEALING WITH MISSING OBSERVATIONS

The problem with MV is that estimation based only on the complete observations (those having no missing values) may bias parameter estimates.

Several methods are used in the applied literature and others are proposed in a more recent theoretical literature to deal with missing values. The issue when selecting a method to deal with missing values is that some of them (for example, imputing means) may reduce the efficiency of the final estimators. Nevertheless, it is not the purpose of this section to review these methods, but to inform the reader about how I deal with my missing observations. A review of these methods, along with a discussion of their properties, can be found in Little and Rubin (1987) and Little (1992). For the case of

panel data, a review of the literature of incomplete panels and selection bias can be found in Verbeek and Nijman (1992b).

6.3.1 "Missing at Random" and "Ignorability"

First, one should distinguish between the concepts of "missing completely at random" (MCAR) and "ignorability"(Little and Rubin, 1987). Call Z the complete data set. Z is an $n*(k+1)$ matrix, where n is the number of observations and k is the number of independent variables, excluding the intercept. Now, $Z = Z_{obs} + Z_{mis}$, where Z_{obs} and Z_{mis} are the subsets of observed and missing values, respectively. Define a "response indicator" matrix R, such that $r_{i,j} = 1$ if $z_{i,j}$ is observed and zero otherwise. Then Z_{mis} is MCAR if $f(R/Z,\theta) = f(R,\theta)$ for all Z, where θ is a scalar or vector that indexes the density function f. That is, data is MCAR if the "missing-ness" is independent of the particular realization of the data at hand. In other words, the probability distribution of the missing observations does not depend on the particular sample at hand. Similarly, Z_{mis} is "missing at random" (MAR) if $f(R/Z,\theta) = f(R/Z_{obs},\theta)$ for all Z_{mis}, which means that data is MAR if observations for one or more variables are missing when certain values are realized for other observed variables. Finally, data are not MAR if the missing observations depend on the values of the unobserved variables for those cases; i.e., one does not observe a certain variable or the whole set of variables when the value of some variable is larger or smaller than a specific amount.

Practical estimation procedures use the concept of "ignorability" instead of the concept of MCAR. Ignorability is a weaker concept than MCAR. A missing data mechanism is said to be ignorable for both sampling-based and likelihood-based

inferences when the data is MCAR. But it is also ignorable (only for likelihood-based inferences) when data is MAR, although not MCAR. Finally, it is non-ignorable when the data is not MAR (Little and Rubin, 1987). Therefore a missing data mechanism can be ignorable for inferences purposes even if missing values are not MCAR.

Verbeek and Nijman (1992a) proposed a formal test for ignorability in linear regression models of panel data. The test is worth performing because of the complexities involved in estimating a panel incorporating the selection rule. The advantages of the test are its simplicity and the fact that it takes into account both wave (unit) and item non-response (although the authors refer to the latter only when information on the dependent variable is missing).

I cannot perform the test proposed by Verbeek and Nijman because I have zero observations for my balanced sub-panel. (I have no month in which all the 74 plants reported). Consequently, I proceed with my unbalanced panel.

This option is justified by three reasons. First, and most obvious, I have no choice, other than to perform no estimation at all. Second, that it is fairly simple to conclude that there exists selection bias in my data set due to non-reporting. I have twelve (12) observations missing as a consequence that the plants informed “no activity” or “very low” activity. Missing-ness is then clearly related to the level of production in those cases. In other words, the selection rule is not independent, among other possible things, of the overall economic situation of firms. These twelve cases make my selection rule not ignorable. Third, I do not think this source of non-ignorability of the selection rule is important in terms of bias because in most cases plants were actually not working and not emitting, as proved by inspections performed in those cases. If this is true, and if I

assume that item non-responses are missing at completely at random, which I do, then the missing observations do not hide any unknown information

6.3.2 Imputing item non-responses

In spite of the fact that I proceed with an unbalanced panel, I impute for the item non-responses before estimating my parameters of interest. The reason is that item non-responses account for 55.4% of the total 5,009 observations missing for the Input and Pollution variables.

According to the literature on missing values, there are basically two criteria to follow when imputing values for item non-reports: conditional mean imputation and multiple imputation (Little, 1992).

Conditional mean imputation methods are based on Buck (1960), Dagenais (1973) and Beale and Little (1975). The basic idea is to use the information on the observed Xs or on the observed Xs and Ys to fill in missing values, correcting for the variances and covariances. Least squares on the filled-in data produce consistent estimates assuming MCAR, which I assumed for my item non-responses.

Multiple imputation is proposed as a way to handle the problem that whatever the conditional mean imputation procedure, "estimated standard errors of the regression coefficients from ordinary least squares or weighted least squares in the filled-in data would tend to be too small, because imputation error is not taken into account." (Little, 1992, p. 1232). By multiple imputation, basically, one imputes $m \geq 2$ values for each missing observation to obtain m different data sets. With each data set one obtains the desired estimates and "averages" them to obtain a final parameter estimate and variance

estimate that "correct" for the underestimation of variances produced by filling in missing observations. (Rubin, 1987).

Both conditional mean and multiple imputation methods were developed and applied for cases of cross-section data and therefore share a problem when applied to panel data: it makes little sense to fill in item non-responses of one plant conditioning on information observed for the rest of the plants, with different technologies, management and output. In other words, it makes little sense to "average" across plants.

I solved this problem by performing the imputations within plants. This way I not only preserve between-plant variability, minimizing bias and variance problems for the final estimates, but I also use plant-specific information about the missing values.[43] Within-plant imputation leaves aside multiple imputation because this would produce *m* data sets for each different plant, and there is no clear way to handle all this information to obtain the final panel estimates. Consequently, I use an iterated Buck procedure within plant to impute for item non-reports, in the spirit of the suggestion made by Beale and Little. I present this iteration briefly below.

Assume for each firm that there is a data set consisting of *N* observations and *k* variables, but one or more of the *k* variables are not observed in some of the *N* observations. Define the following variables:

$\tilde{x}_j = \sum_{i \in C} x_{ij}$; where $C<N$ is the subset of complete observations. Then $\tilde{x}_j$ is the average of the variable x_j over the set of complete observations.

[43] An example of the latter is to use monthly volumes of effluents discharged divided by days worked in the month to impute the monthly average effluent flow.

$\hat{x}_{ij}$ is the filled-in data where $\hat{x}_{ij} = x_{ij}$ (the observed value) if the variable *j* is observed in the observation *i* or $\hat{x}_{ij} = \tilde{x}_j + \sum_{l \in p} b_{jl}(x_{il} - \tilde{x}_l)$, where b_{jl} is the partial regression coefficient of x_j on x_l over the complete observations, and $p<k$ is the set of observed variables in observation *i*. In other words, $\hat{x}_{ij}$ is the fitted value of a linear regression on the *p* observed variables for observation *i* using the complete observations.

$\bar{x}_j = \sum_{i=1}^{N} \hat{x}_{ij} / N$; the mean of variable *j* over the filled-in data.

$a_{jk} = \sum_i (\hat{x}_{ij} - \bar{x}_j)(\hat{x}_{ik} - \bar{x}_k) + c_{ijk}$; the jk^{th} element of the corrected matrix of sums of squares and products, where c_{ijk} is the corrected term and equals the residual variance computed from the regression of x_j on the observed variables in that observation *i* over the complete cases, if only x_j is missing in observation *i*, or the residual covariance computed from the regression of x_j and x_k on the rest of the observed variables in that observation if both x_j and x_k are missing in that observation, always regressing over the complete cases. In mathematical notation, call v_{jk} the covariance of $(x_j - \sum_p b_{jp} x_p)$ and $(x_k - \sum_p b_{kp} x_p)$ where *p* is the subset of observed variables in the observation in question. Then, $c_{ijk} = v_{jk}$ if x_j and x_k are both unknown $(j \neq k)$ or if only x_j is unknown $(j = k)$, or 0 otherwise.

The steps of the version of the iterated Buck's procedure proposed by Beale and Little are:

1. Fit all the missing items as suggested by Buck and compute a_{jk}.

2. Calculate $\bar{x}_j$ and substitute it for $\tilde{x}_j$ in $\hat{x}_{ij} = \tilde{x}_j + \sum_{l \in p} b_{jl}(x_{il} - \tilde{x}_l)$

3. Repeat until $\bar{x}_j$ and a_{jk} have no further significant changes.

To perform this procedure I construct the following variables for each plant: (1) *WATER = TAP + UW*: Total water consumption (in m^3/month) equals the sum of tap (*TAP*) water and underground water (*UW*) consumed; (2) *ENERGY = EL*3.6 + FUEL*43,752.06*: Total energy consumption in mega joules (MJ), where *EL* is the electric energy consumed in Kwh/month and *FUEL* is the quantity of fuels consumed per month in m^3; (3) *LABOR = WD*EMPLOY*: Total employee-days worked, where *WD* is the total number of days worked in the month and *EMPLOY* is the total number of employees in that month; (4) *POLLUTION = FLOW*BOD*$_5$**1000*: Total organic pollution discharged in (mg/day), where *BOD*$_5$ was already defined and *FLOW* is the average flow level of discharges, in m^3/day; (5) *PRODUCTION* = Quantity produced by month.[44] The original variables were fitted using these constructed variables. I estimated the auxiliary linear regressions with the variables in natural logarithm forms. These did not necessarily provide better fits than auxiliary regressions with variables in their original form, but they are closer to "the spirit" of a Cobb-Douglas type of production function.[45]

Finally, I do not use the monitoring and enforcement variables in this imputation for two reasons: first, I conserve degrees of freedom in the auxiliary regressions within

[44] In twenty-five cases this variable involved standardizing units of measure to be able to add different products.

[45] A document describing the distribution of missing values per variable by industrial plant, the processes followed to impute for item non-responses in each plant, and the corresponding iteration procedures is available from the author upon request.

firms, and second, it would be like cheating to use these variables to impute for the MV and then use the resulting data to test for their effect on pollution.