CHAPTER 7

SPECIFICATION AND ESTIMATION ISSUES

As mentioned in Chapters 2 and 5, the objectives of this second part of my dissertation are the following. First, I examine the determinants of the allocation of inspections by the municipal government (IMM) and the national government (DCA) among industrial plants in Montevideo. In the presence of three different monitoring institutions, consistency requires the estimation of an inspection equation for SEINCO to explain the inspection strategy developed by this private firm serving the IMM. Second, I empirically test the effectiveness of the inspections and the different enforcement actions of both municipal and state governments in terms of reducing BOD_5 emissions. Third, I also test the effectiveness of these actions in terms of increasing compliance rates. Fourth, since emissions are self-reported, I test for the presence of under-reporting.

In order to fulfill the first objective I estimate an inspection equation for the IMM, another one for the DCA and a third one for SEINCO. In order to fulfill the second objective I estimate two equations, one with BOD_5 concentration level of industrial effluents as a dependent variable and a second with total quantity of BOD_5 emitted. I call this second equation the load equation. The third objective is fulfilled by the estimation of a violation equation. This equation has a specification similar to the BOD_5 equation but its dependent variable is a zero/one dummy variable indicating compliance or violation status.

The idea behind the inspection equations for the IMM and the DCA is, apart from explaining the inspection strategies itself, to estimate probabilities of being inspected that can be used as instruments for actual inspections in the BOD_5, load and violation equations. The inclusion of a probability of being inspected by SEINCO explores the possible impacts that its monitoring activity may have had on BOD_5 levels reported by plants.

7.1 THE MODEL

7.1.1 The Inspection Equations

During the period under study both the municipal (IMM) and national government (DCA) office monitored industrial plants in Montevideo. In fact, all seventy-four industrial plants were inspected at least two times by the IMM. The DCA inspected fifty-eight of these same plants at least once. The remaining sixteen plants were never inspected by the DCA. Parallel monitoring efforts of regulators were not coordinated. As explained in previous chapters, the two offices did not share information on monitoring and enforcement activities on a regular basis. Quite the contrary, information sharing was limited to specific and complicated cases. In fact, the correlation coefficient between the number of inspections of the two offices across time and plants is 0.16. These arguments validate the chosen course of action of estimating separate inspections equations for the two offices. These are presented in the following sections.

7.1.1.1 The IMM Inspection Equation

Equation 7.1 was estimated to fit the probability of being inspected by the IMM:[46]

$$
\begin{aligned}
INSPIMM_{i,t} = {} & \gamma_0 + \gamma_1 INSPIMMCUM_{i,t-1} + \gamma_2 INSPIMMOTHERCUM_{i,t-1} \\
& + \gamma_3 INSPSEINCOCUM_{i,t-1} + \gamma_4 FINEDIMMCUM_{i,t-1} + \gamma_5 VOL_t \\
& + \gamma_6 RF_{i,t} + \gamma_7 PTY_i + \gamma_8 TANNERY_i + \gamma_9 WOOL_i + \gamma_{10} 1997-1998_t \\
& + \gamma_{11} DURINGPLAN_t + \gamma_{12} STREAM_{i,t} + \eta_{i,t} \\
& i = 1,...,74; t = July1997,...October2001
\end{aligned}
\tag{7.1}
$$

$INSPIMM_{i,t}$ is a dummy equal to one if plant *i* was inspected by the IMM in month *t*. In order to specify this inspection equation, I considered that the strategy of the IMM inspectors obeyed five rules. The first one was a “sample without replacement” rule. The IMM classified plants in “Priority 1” and “Priority 2” plants. Priority 1 plants (25 of the 74 plants in my sample) are the heaviest polluters in terms of organic pollution and metals. They account for 80% of this pollution. Interviewed inspectors declared that they try to visit “Priority 1” plants twice and “Priority 2” plants once every six months. But the data do not support this statement. Therefore, in order to capture the sample-without-replacement inspection strategy I included the number of inspections performed in the plant during the last twelve months ($INSPIMMCUM_{i,t-1}$) and the priority group to which the plant belongs (PTY_i , equal to 1 if the plant is a Priority 1 plant) as explanatory variables.[47]

---

[46] Table A.7.1 in the Appendix 7.2 provides a list of all the variables used in this chapter and their definitions.

[47] I also tried the cumulative number of inspections performed in the last six months instead of twelve months, but the model performed better with twelve months in terms of goodness of fit and both the Akaike and Schwarz information criteria.

The second rule mentioned by IMM inspectors was that plants with worse compliance histories and those showing less cooperation with regulators were inspected more often. These plants were those that did not take the promised measures to abate emissions or delayed them. I included $FINEDIMMCUM_{i,t-1}$ to capture the level of cooperation.[48] This variable measures the number of fines imposed by the IMM against this plant in the last twelve months; the more the cumulative number of fines the less the cooperation of the plant in the recent past.[49] This level of cooperation perceived by regulators is not only a function of the recent formal history of the plant. It also depends on non-quantifiable facts on which inspectors based their decisions.[50]

Third, citizens' complaints also triggered inspections but were not included as an explanatory variable because of the unavailability of information about them. Nevertheless, interviewed inspectors declared that most of these complaints were not

[48] The inclusion of the number of detected violations in the last twelve months did not improve the fit of the model. This result is consistent with the policy approach of the period. Effectively, during this period the enforcement efforts were not so much directed at enforcing standards but at decreasing emissions. Towards this end was that the IMM implemented the Industrial Pollution Reduction Plan relaxing emissions standards (See Chapter 2).
Also, the cumulative amount of fines was included instead of the cumulative number. This did not change the results either.

[49] I do not have information on intermediate enforcement actions (e.g., compliance orders, etc.) issued by the municipal government of Montevideo, just those issued by the national government office, DINAMA.

[50] An example is the following: sometimes inspectors are kept waiting at the plant entrance for the length of time needed to make some quick cleanings and other measures (like diluting) to comply with the emissions standards. This is more typical in small plants, with lesser time of effluents retention. Another example is the quickness of response to suggested changes. It is worth noting that this makes the effectiveness of water pollution control very dependent on those specific inspectors with long experience in the job. In other words, a good deal of the compliance history of plants is lost when an inspector retires or is appointed to another office.

originated by “unusual” levels of discharges, but from smells or illegal points of discharge (e.g., streets, brooks) or when the public sanitary system below the streets collapsed.

Fourth, the failure to report in subsequent periods also triggered inspections according to the IMM inspectors. As a result, the number of reporting failures in the previous two reporting periods ($RF_{i,t}$) was also included as an explanatory variable.[51]

Finally, unusually high levels of reported pollution sometimes triggered an inspection, although very rarely according to UEI inspectors. One reason is that obviously it could not be optimal for plants to report “peaks” of their emissions. But it is not easy to construct a variable capturing the effect of unusual levels of reported emissions either. There are three reporting periods during the year: March-June, July-October, and November-February, so plants report four months of activity in each report. But they did not have a clear due date for sending their reports. Many plants during the analyzed period sent their reports in the final month of the following reporting period. In other words, regulators were looking at a picture of the plant that was at least four months old. Other plants reported immediately after the end of the reporting period. In short, regulators did not receive the information on emissions at the exact point in time in every period. This complicated the possibility of constructing a variable indicating unusual level of emissions because it was impossible to know at what point in time the regulator

[51] In the first six months of 1997 the UEI implemented a new enforcement strategy. It issued a fax to every plant in its database explaining the new four-month Reporting Form format and communicated to the plants that the municipal government was undertaking a new plan for pollution control. For that reason, in the first reporting period I set the reporting failure history of every plant equal to zero as an indicator that a new enforcement period had begun.

was looking at the information so as to decide on an inspection. For this reason, I opted to include no lagged indicator of reported pollution.[52]

In addition to the variables included to capture these rules, other variables were included to capture other determinants of IMM inspections, for example, $INSPIMMOTHERCUM_{i,\, t\text{-}1}$. This variable measures the cumulative number of inspections performed by the IMM in the rest of the plants. Inspectors knew that the rest of the plants were aware of inspections performed at a specific plant, particularly those in the neighborhood, and that this could have molded their expectations regarding a possible inspection. On the other hand, if the IMM monitoring activities were affected by important budget constraints, as they actually were, the sign of this variable's coefficient would be negative, indicating that the higher the number of inspections performed on other plants in the recent past the smaller was the probability of this plant being inspected given the cost of monitoring campaigns. Therefore, the sign of this variable's coefficient remains an empirical matter.

Another important determinant of IMM inspections during part of the analyzed period was the implementation of the previously Inter American Development Bank-financed "Monitoring Program" in charge of the private consortium SEINCO. The objectives of the program, described in detail in Chapter 2, included establishing a monitoring frequency of industries and water bodies for the IMM. Towards this objective SEINCO conducted regular inspections on industrial plants during 1999-2001. The IMM took advantage of this situation, saving on monitoring resources. $INSPSEINCOCUM_{i,t\text{-}1}$,

[52] In spite of this I ran a model with the average BOD_5 level of the plant in the last six months as an explanatory variable. The resulting coefficient was extremely low (0.00008) and insignificant. The overall fit of the model increased merely 0.000276 as measured by the McFadden R square.

the cumulative number of inspections performed by SEINCO on a plant in the last twelve months, measures this effect.

Also, the Uruguayan industrial sector went through an important contraction during part of the analyzed period. In particular, the industry production volume index dropped 8.6% on average in 1999 and 7.2% in 2001. (During 2000 it increased 2%). The contraction was larger as measured by the industry real GDP: 23% between 1996 and 2001, with an average drop of 4% during the period 1997 – 2001 and 8% during the period 1999 – 2001.[53] Although not recognized by authorities, as a consequence of this contraction, inspectors may have eased or loosened their enforcement pressure on plants, since it was precisely the "difficult economic times" that inspired the Industrial Pollution Reduction Plan. I included the monthly level of the industry production volume index (*VOL*) to capture this possible effect.

Apart from classifying industrial plants according to their priority, the IMM also targeted tanneries and wool washers. The reason for this is that the IMM, in accordance with the IADB, targeted its control efforts at two pollutants, Chromium and BOD_5. These two industries were the most important sources of these pollutants, respectively. For this reason two dummy variables were added to the list of right-hand side variables; *TANNERY* for tanneries, and *WOOL* for wool washers.[54]

[53] The differences in the variation of the volume index (constructed by the National Statistics Institute) and industrial GDP (constructed by the Central Bank) are due to differences in weight of the different sectors in the construction of both indexes. I chose the first one because of monthly availability.

[54] I included sector dummies in place of these two dummies to explore the results. The sector dummies were neither significant nor did they improve the fit of the model in the unconditional regression.

1997-1998 is another dummy variable equal to one in the months of these two years during which the IMM inspectors conducted special monitoring campaigns due to the delay in the implementation of the Monitoring Program by SEINCO. It is interesting to note that IMM inspectors received extra IADB-financed payments for these campaigns.

DURINGPLAN refers to the Industrial Pollution Reduction Plan implemented from March 1997 to December 1999. This variable, a dummy equal to one during these months, was included because the IMM could have changed its monitoring strategy given that its objective during these months was to give more time to plants to incorporate abatement technology.

BOD_5 emissions standard for plants emitting directly into waterways is 60 milligrams per liter (mg/l), while it is 700 mg/l for those emitting into the sewage system. A dummy variable indicating whether the plant was emitting directly into a water body was also included to capture any possible effect of this on the probability of being inspected. This variable is *STREAM*, equal to one if the plant emits directly into a water body.

Finally, $\eta_{i,t}$ is the error term, assumed to be identically and independently distributed with zero mean and to have a logistic distribution. The reason for choosing this distribution, as opposed to a normal distribution, is the following. Although there is a common intercept in the IMM inspection equation, I will test for fixed effects. But probit models do not support a fixed effects specification; only the logit models do. The common panel data structure is a large number of units observed over a relatively short period of time ($N>T$). In this structure, it makes sense to have the desirable asymptotic properties of the maximum likelihood estimators when *N* increases and *T* is fixed. But increasing *N* would mean increasing the number of parameters to estimate (via increasing the fixed effects). This is what is called the "incidental parameter problem". Maximizing the likelihood over all parameters, including the plant-specific effects, will yield inconsistent estimates for large *N* and fixed *T* (Hsiao, 1986). A partial solution to this problem has been developed for the logit model but not for the probit model, which does not support a fixed effect specification because in this case it is not possible to get rid of the plant-specific effects μ_i. The partial solution is the conditional maximum likelihood approach suggested by Chamberlain (1980). Conditioning on the number of inspections actually observed for each plant, one can get rid of the fixed effects and obtain consistent estimates for the rest of the parameters (see Appendix 7.1).

7.1.1.2 The DCA Inspection Equation

As explained in Chapter 2, a previous arrangement between the IMM and the DCA left the main regular monitoring activity in the hands of the UEI during the period analyzed.[55] In fact, as can be seen in Table 2.1 of Chapter 2, out of a total of 760 inspections, 549 were done by the UEI. The arrangement also meant that the DCA would be basically in charge of "initial compliance" by assuring that all plants had and correctly

[55] Nevertheless, a common inspection equation was estimated. This was:

$$\begin{aligned} Insp_{i,t} = {} & \gamma_1 INSPCUM_{i,t} + \gamma_2 DVCUM_{i,t} + \gamma_3 ORDERCUM_{i,t} \\ & + \gamma_4 POSTCUM_{i,t} + \gamma_5 Vol_t + \gamma_6 RF_{i,t} + \gamma_7 Pty_i + \eta_{i,t} \end{aligned}$$

The specification of this equation incorporated both inspection strategies of the IMM and the DCA. The fit of this equation was poorer than the fit of the separate inspection equations for the IMM and the DCA.

operated an effluents treatment plant. However, there is no clear difference between the number of inspections performed by this office on plants that did incorporate abatement technology during the period and those that did not. This means that although the DCA left to the IMM most of the "continuous compliance" control, it did conduct its own regular inspections, although to a much lesser extent.

With all of this in mind, the inspection equation proposed for the DCA was:

$$\begin{aligned} INSPDCA_{i,t} &= \alpha_0 + \alpha_1 INSPDCACUM_{i,t-1} + \alpha_2 INSPDCAOTHERCUM_{i,t-1} \\ &+ \alpha_3 EADCACUM_{i,t-1} + \alpha_4 VOL_t + \alpha_5 TANNERY_i \\ &+ \alpha_6 WOOL_i + \alpha_7 CARRASCO1999_t + \alpha_8 STREAM_{i,t} + \upsilon_{i,t} \\ & i = 1,\ldots,74; t = July1997,\ldots October2001 \end{aligned} \tag{7.2}$$

The first two variables (*INSPDCACUM* and *INSPDCAOTHERCUM*) are defined exactly as *INSPIMMCUM* and *INSPIMMOTHERCUM*, and are included for the same reasons. $EADCACUM_{i,t-1}$ is the cumulative number of compliance orders, fine threats and fines issued by the DCA to the plant up to *t*-1.[56] *VOL*, *TANNERY*, *WOOL* and *STREAM* are the same variables included in the IMM inspection equation. Finally, during 1999 the National Environment Office (Direccion Nacional de Medio Ambiente, DINAMA) performed a special monitoring campaign on those plants in the basin of the Carrasco Stream. This campaign was the result of an agreement between the DINAMA and a non-governmental organization dedicated to fighting pollution of this stream (Asociación Pro-Recuperación del Arroyo Carrasco, APRAC). I included the dummy *CARRASCO1999*

[56] Separating *EADCACUM* into the cumulative number of enforcement orders (*ORDERDCACUM*), the cumulative number of fine threats (*FINETHREATDCACUM*) and the cumulative number of fines (*FINEDDCACUM*) did not improve the results.

for this reason. Finally, I assume a logistic distribution for the errors in this equation for the same reason given for the IMM equation.

7.1.1.3 The SEINCO Inspection Equation

A third inspection equation corresponding to SEINCO was estimated for consistency.

The specification of the SEINCO inspection equation is:

$$\begin{gathered}INSPSEINCO_{i,t} = \beta_0 + \beta_1 INSPSEINCOCUM_{i,t-1} + \beta_2 INSPIMMCUM_{i,t-1} \\ + \beta_3 INSPDCACUM_{i,t-1} + \beta_4 PTY_i + \beta_5 TANNERY \\ {}_i + \beta_6 WOOL_i + \beta_7 STREAM_{i,t} + \delta_{i,t} \\ i = 1,...,74; t = July 1999,..., September 2001\end{gathered} \tag{7.3}$$

INSPSEINCOCUM is the cumulative number of SEINCO inspections. INSPIMMCUM and INSPDCACUM were included to test how SEINCO used the information pertaining to the monitoring activity of the two agencies to develop its own. Finally, according to interviews, SEINCO also inspected “Priority 1” plants more frequently and targeted tanneries and wool washers because of their importance in terms of organic and metal pollution.

7.1.2 The Pollution Equations

7.1.2.1The BOD_5 Equation

Equation (7.4) is a linear pollution equation in the spirit of Magat and Viscusi (1990), Laplante and Rilstone (1996) and Dasgupta, et al. (2001). It assumes a Cobb-Douglas technology.

$$\begin{aligned}\ln(BOD5_{i,t}) = {} & \lambda_1 \ln(P_{q,t}) + \lambda_2 \ln(Labor_{i,t}) + \lambda_3 \ln(Water_{i,t}) \\ & + \lambda_4 \ln(Energy_{i,t}) + \lambda_5 \ln(Flow_{i,t}) + \lambda_6 TECH_{i,t} \\ & + \lambda_7 PINSPIMM_{i,t} + \lambda_8 PINSPDCA_{i,t} + \lambda_9 PINSPSEINCO_{i,t} \\ & + \lambda_{10} INSPIMMCUM_{i,t-1} + \lambda_{11} INSPDCACUM_{i,t-1} + \lambda_{12} FINEDIMMCUM_{i,t-1} \\ & + \lambda_{13} EADCACUM_{i,t-1} + \lambda_{14} DURINGPLAN_t + \mu_i + \nu_{i,t}\end{aligned} \tag{7.4}$$

i = 1, …, 74; t = July 1997, …, October 2001.

Equation (7.4) develops from the idea that the level of concentration of organic pollution in a given month, measured as Biological Oxygen Demand (BOD_5) in mg/l, is a function of two sets of variables, one reflecting the marginal benefits of pollution (i.e., the value of the marginal productivity of pollution) and another reflecting the marginal expected cost of pollution.

Marginal benefits of pollution are represented by the price of the final good (P_q) and the input variables *Labor*, *Water*, *Energy* and *Flow*. Marginal expected costs of pollution are represented by the monitoring and enforcement variables. These are comprised of the probabilities of being inspected by the municipal and national governments ($PINSPIMM_{i,t}$ and $PINSPDCA_{i,t}$) and by the probability of being inspected by SEINCO ($PINSPSEINCO_{i,t}$). These three variables, obtained by fitting the IMM, DCA and SEINCO inspections equations, are included to capture the effect of future possible enforcement actions due to today's pollution decisions.

But pollution today is also the result of past monitoring and enforcement actions. This is the reason for including the cumulative number of inspections performed during the last twelve months by the municipal government (*INSPIMMCUM*) and the national government (*INSPDCACUM*) and the cumulative number of fines levied by the municipal government (*FINEDIMMCUM*) and the cumulative number of intermediate enforcement actions and fines levied by the national government (*EADCACUM*).[57]

Some cases in the previously cited literature include the contemporaneous number of inspections or a dummy as an explanatory variable to indicate whether the plant was inspected in that month. Those who did not consider under-reporting to be an issue

---

[57] Monetary fines were not the only penalty levied for not complying. Plants could also be temporarily closed. But neither the municipal nor the national government had trustworthy records of these measures. (These types of measures were as uncommon as fines during the period). Another form of penalty implemented was to make professionals in charge of treatment plants legally responsible for sending false reports. During the first phase of the Industrial Pollution Reduction Plan, report forms included at the bottom a copy of the corresponding articles in the legislation that state this legal responsibility. According to the IMM's Industrial Effluents Unit Director, this was done as an explicit enforcement mechanism. The objective was to persuade professionals about the dangers of falsifying information and to act on reluctant plants through them. According to this Director, this type of expected penalty may have had an important impact on emissions levels because plants reluctant to decrease emissions may have encountered increasing difficulties in finding professionals in the market who were willing to cheat at their own personal cost. Apart from its apparent effectiveness, this strategy, which in a sense could be seen as a deviation from the classical theoretical model of enforcement, seems also optimal in terms of institutional compatibility. High fines are rarely feasible to apply in less-developed countries where firms suffer from important cash flow constraints. These alternative penalties are easier to apply because they do not imply a cash payment. At the same time, they do imply significant costs to the firm, either directly (through closing) or indirectly (through the professionals' incentives). Unfortunately, it was impossible to measure their effects.

Finally, *INSPSEINCOCUM* (the cumulative number of past inspections by SEINCO) was originally included in this model but it was dropped due to its correlation of 0.91 with PINSPSEINCO.

included it as another determinant of pollution. These cases estimated pollution and inspections as jointly determined in a system of two equations. Those who did consider the problem of under-reporting, like Shimshack and Ward (2002), did this as an "imperfect" and "weak" test for self-reporting accuracy. My approach was to use fitted values obtained from inspection equations which would serve at the same time as an econometric instrument for actual inspections and as a proxy for probabilities of being inspected. These would capture plants' reactions to the possibility of future monitoring and enforcement actions due to present decisions regarding levels of reported pollution. At the same time I used the information on BOD_5 samples by the IMM, DCA and SEINCO to conduct a difference-of-means test between these and the BOD_5 reported levels to test for the presence of under-reporting and possible changes in the reporting strategy of firms through time.

The reason for including intermediate enforcement actions apart from fines is that with only 11 fines in the whole period (despite frequent violations) it is reasonable to conclude that regulators intended to reduce emissions via these intermediate actions. These may have had their own deterrent effects. This deterrent effect could be explained because fines are not instantaneously applied after a violation is reported or discovered by an inspection. Instead, firms face an increasing probability of being fined. Of course this probability and the amount of the fine is uncertain for the firms. However, firms learn by observing past responses of regulators to violations. This argument is similar to that of Shimshack and Ward (2002).[58]

[58] I ran a version of this equation separating the cumulative number of compliance orders, the cumulative number of fine threats and the cumulative number of fines issued by the DCA. Results did not change.

Eight firms modified their treatment technology during the period, either by constructing nonexistent treatment plants or by significantly modifying existing plants.[59] I included the variable *TECH*, a dummy equal to one in the month that the plant incorporated abatement technology and thereafter, to control the effect of changes in treatment technology on BOD_5 levels.

The last explanatory variable is DURINGPLAN. This variable is the same dummy that was included in the IMM inspection equation. Its value is one during the months of the Industrial Pollution Reduction Plan and zero afterwards. The idea of including this variable in the pollution equation is to test for the presence of different reporting and emitting behavior of plants during the plan. This is possible because during these months emission standards were laxer. With it the IMM intended to give plants time to adopt abatement measures while at the same time complying with pollution regulations. The inclusion of this variable measures the success of the plan.[60]

The parameter μ_i is a plant-specific effect. I chose a fixed-effects model, as opposed to a random-effects model, because I am basing my inference on these 74 specific plants, which were not randomly selected from a large population and are responsible for around 90% of the industrial emissions in the city (Multiervice-Seinco-Tahal, 2001). I did not perform formal tests for the unit effects. To perform these tests under the assumption of non-spherical errors I would have to invert the variance–

---

[59] One more plant incorporated technology the month before the beginning of my study period and two more during 1996.

[60] One caveat to this conclusion is stressed in the next chapter. The after-plan period coincided with one of the most important recessions of the Uruguayan economy in its entire history. As a result, an interpretation of the success of the plan according to a positive sign of the DURINGPLAN dummy could be misleading.

covariance matrix of the errors and, as I explain in detail below, this is not possible when *N* (the number of plants) is larger than *T* (the number of months one observes for each plant), as it is in my case. In spite of this, I performed a Chow test assuming that the errors were spherical. The test strongly suggested rejecting the null hypothesis of common constant terms.[61]

Finally, $v_{i,t}$ is the stochastic disturbance. Following Park, the panel structure of the errors can be:

(1) Panel Heteroskedastic: $E(v_{it}^2) \neq E(v_{jt}^2)$ for $i \neq j$, i and j being different plants, but $E(v_{it}^2) = E(v_{is}^2)$ for $t \neq s$, so $E(v_{it}^2) = \sigma_i^2$. It is assumed that the variance of the errors differs across plants but not across months within each plant.

(2) Contemporaneously Correlated: $E(v_{it}v_{jt}) = E(v_{is}v_{js}) \neq 0$, but $E(v_{it}v_{js}) = 0$, so $E(v_{it}v_{jt}) = \sigma_{ij}$, with all other covariances equal to zero.

Also, the errors can be:

(3) Common Serially Correlated: $v_{it} = \rho v_{it-1} + \varepsilon_{it}$, where the ε_{it} are temporally independent, identically distributed, zero-mean random variables, or

(4) Plant-specific serially correlated: $v_{it} = \rho_i v_{it-1} + \varepsilon_{it}$.

With respect to the latter, Beck and Katz (1995) argue "The assumption of unit-specific serial correlation … seems odd at a theoretical level. Time-series cross-section

---

[61] Under the null of common constant terms, the statistic $\dfrac{\left(R_u^2 - R_p^2\right)/\left(N-1\right)}{\left(1-R_u^2\right)/\left(NT-N-K\right)}$ is distributed as *F[(N-1), (NT-N-K)]*, where the subscript *u* stands for unrestricted model and the subscript *p* stands for the pooled model. The unrestricted model in this case is the FE model and the pooled model is the common-constant OLS model. *K* is the number of explanatory variables. The value of this statistic for this test was 71.69 > *F(73,2705]*.

analysts assume that the “interesting” parameters of the model, $\boldsymbol{\beta}$, do not vary across units; this assumption of pooling is at the heart of Time–Series, Cross- Sections analysis. Why should we expect the “nuisance” ρ show similar pooling behavior? ρ can be interpreted as how long it takes for prior shocks to be removed from the system. Why should this “memory” be the only model parameter that varies from unit to unit?” (Beck and Katz, 1995, p. 638). In spite of the argument presented by this statement, I will test for it.

7.1.2.2 The Load Equation

The reason for estimating a BOD_5 pollution equation is that emission standards are defined in terms of concentration of organic matter (as measured by BOD_5). In addition, it is interesting to test whether the monitoring and enforcement strategy of regulators during the period had an effect on the total organic load discharged by plants and to compare it with the results obtained with the BOD_5 equation. An interesting issue that may arise with this comparison is whether regulators’ effectiveness is masqueraded by the dilution of effluents in clean water, for example. The estimated load equation is specified exactly as the BOD_5 equation except for the obvious fact that it cannot include *FLOW* as an explanatory variable because *LOAD* is defined as BOD_5 times *FLOW*. *LOAD* is then measured in kg/day.

7.1.2.3 The Violation Equation

In order to test the effectiveness of regulators regarding the compliance status of plants I estimate a conditional fixed-effects logistic model with a dummy variable equal to one if the plant reported a violation as a dependent variable. Violations were defined with respect to the laxer standards during the Pollution Reduction Plan.

The violation equation has the same explanatory variables as the BOD_5 equation, but it has fewer observations. Five plants were dropped from the sample because they release effluents into the soil and there are no standards set for BOD_5 in this case. Also, fourteen additional plants that complied or did not comply in every month of the period and therefore did not add any likelihood to the conditional model were also dropped from the sample (see Greene (1997), p. 899).

7.2 ESTIMATION ISSUES

7.2.1 Inspection Equations

Although there is a common intercept in the inspection equations (7.1), (7.2) and (7.3) I test for the presence of fixed-effects in each of them using a Hausman test that compares the conditional (fixed-effects) and the unconditional logit estimates.

Under the null hypothesis (H_0)$\alpha_i = \alpha$, both the conditional logit estimator (CLE) and the unconditional logit estimator (LE) are consistent, but the CLE is inefficient because it fails to use the information in H_0. (i.e.: any plant-specific, time-invariant characteristic captured by the fixed effect). Under the alternative of plant heterogeneity, the LE is inconsistent whereas Chamberlain’s CLE is consistent and efficient. The Hausman chi-squared statistic is the following:

$$\chi^2 = \left(\hat{\beta}_{CLE} - \hat{\beta}_{LE}\right)' \left(var[CLE] - var[LE]\right)^{-1} \left(\hat{\beta}_{CLE} - \hat{\beta}_{LE}\right)$$

For the unconditional logit estimator the row and column corresponding to the constant term are dropped from the estimated covariance matrix *var[LE]*. The results of the tests for the three inspection equations are presented in Table 7.1.

Table 7.1: Hausman Tests for Fixed Effects on Inspections Equations

	Degrees of Freedom	Chi Square	Prob>Chi Square
IMM	9	98.7	0.000
DCA	6	91.48	0.000
SEINCO	4	238.6	0.000

In the three cases it is very clear that I should reject the null of no fixed effects in favor of the alternative of fixed effects.

Because fixed effects are never actually estimated in the conditional logit (again, see Greene (1997) pg. 899), the results of these tests suggest the following trade-off. On the one hand, without estimates for the fixed effects I cannot obtain predictions for the probabilities of inspections.[62] On the other hand, if I specify an unconditional logit to be able to predict probabilities I do not recognize plant specific effects and I obtain inconsistent estimates of my parameters.

The chosen solution was to estimate an unconditional logit to predict the probabilities and a conditional logit to interpret the estimated coefficients. The two models cannot be specified identically because the conditional (fixed-effect) logistic

[62] STATA has two options for predicting the probabilities. The first one is conditioning on a single inspection within plant. This is not what I was looking for. The second option is conditioning on the fixed effects being zero. This one gave me predicted probabilities greater than 0.8 in every case. This could not be used either.

regression eliminates any variable without within-plant variability. This implies that *TANNERY, WOOL*, and *PTY* are omitted from the IMM and SEINCO inspection equations and *TANNERY* and *WOOL* are omitted from the DCA inspection equation.

The resulting equations are:

$$
\begin{gathered}
INSPIMM_{i,t} = \gamma_i + \gamma_1 INSPIMMCUM_{i,t-1} + \gamma_2 INSPIMMOTHERCUM_{i,t-1} \\
+ \gamma_3 INSPSEINCOCUM_{i,t-1} + \gamma_4 FINEDIMMCUM_{i,t-1} + \gamma_5 VOL_t \\
+ \gamma_6 RF_{i,t} + \gamma_7 1997-1998_t \\
+ \gamma_8 DURINGPLAN_t + \gamma_9 STREAM_{i,t} + \eta_{i,t} \\
i = 1,...,74; t = July1997,...October2001
\end{gathered}
\tag{7.5}
$$

$$
\begin{gathered}
INSPDCA_{i,t} = \alpha_i + \alpha_1 INSPDCACUM_{i,t-1} + \alpha_2 INSPDCAOTHERCUM_{i,t-1} \\
+ \alpha_3 EADCACUM_{i,t-1} + \alpha_4 VOL_t + \alpha_5 CARRASCO1999_t \\
+ \alpha_6 STREAM_{i,t} + \upsilon_{i,t} \\
i = 1,...,74; t = July1997,...October2001
\end{gathered}
\tag{7.6}
$$

$$
\begin{gathered}
INSPSEINCO_{i,t} = \beta_0 + \beta_1 INSPSEINCOCUM_{i,t-1} + \beta_2 INSPIMMCUM_{i,t-1} \\
+ \beta_3 INSPDCACUM_{i,t-1} + \beta_4 PTY_i + \beta_5 TANNERY_i \\
+ \beta_6 WOOL_i + \beta_7 STREAM_{i,t} + \delta_{i,t} \\
i = 1,...,74; t = July1999,...,September2001
\end{gathered}
\tag{7.7}
$$

where all the variables are as defined in Equations (7.1), (7.2) and (7.3), and γ_i, α_i and β_i represent the plant-specific fixed effect in the three equations, respectively.

The inspection equation above implicitly assumes that every plant has information on the number of inspections, orders, fine threats, and fines that both DINAMA and the IMM perform on every plant at any moment, and they use this information to form their probabilities of being inspected. In the real world this is impossible without a government policy of information disclosure. This is not the case in Montevideo. Therefore, the estimated probabilities of inspections will surely differ from the ones formed by plant-managers. One possibility to move closer to reality is to estimate a separate inspection equation for every plant. This was not possible because first, some plants did not have enough observations to assure the "asymptotic" properties of maximum likelihood estimators and second, other plants had zero observations for *FINEDIMMCUM* or *EADCACUM* or *RF* during the whole period.

7.2.2 The Pollution Equation

Given that no current endogenous variable ($BOD_{5\,i,t}$) appears on the right hand side of the inspection equation, my statistical system of equations is recursive. Estimates obtained are consistent. Simultaneous-equations-estimation procedures are not necessary.

Provided that there is no contemporaneous correlation between the error term in the pollution equation ($v_{i,t}$) and the error term in the inspection equations ($\eta_{i,t}, \nu_{i,t}, \delta_{i,t}$), the fitted values obtained from the inspection equations (which are a proxy for the probabilities of inspections) will be uncorrelated with $v_{i,t}$, and a least squares estimator will yield consistent estimates of the parameters of the pollution equation.[63]

The natural estimation approach would have been to use feasible generalized least squares (FGLS), either A) a seemingly unrelated (SUR) weighted least squares correcting for (1) plant heteroskedasticity and (2) contemporaneous correlation of errors across

[63] Since the reduced form for the inspection equation is identical to the (structural) inspection equation, my estimation procedure is the same as 2SLS but in a system that is not simultaneous.

plants, or B) a Parks estimator correcting for (1) plant heteroskedasticity, (2) contemporaneous correlation of errors across plants and (3) serial correlation. The problem with this approach is that I have a number of cross-sections (N=74) that is larger than the number of time periods (T = 52). Therefore, I cannot estimate my model using FGLS because the error covariance matrix is not invertible.[64]

The easiest way to circumvent this problem would have been to discard the appropriate number of plants. This procedure would have left me with a panel of 52 plants and 52 time periods (*N = T,* the necessary condition for $\mathbf{\Sigma}$ to be invertible). However, the panel would have been unbalanced since the number of plants with zero

---

[64] To simplify notation, assume that $T_i = T$ for all i (i.e, all plants have the same number of observations; i.e., the panel is balanced) and there is no serial correlation among the errors. Then, the form of $\mathbf{\Omega}$, the (52*74)*(52*74) covariance matrix is as follows:

$$\mathbf{\Omega} = E[\nu\nu'] = \begin{pmatrix} \sigma_{1,1}\mathbf{I}_{52} & \sigma_{1,2}\mathbf{I}_{52} & \cdots & \sigma_{1,74}\mathbf{I}_{52} \\ \sigma_{2,1}\mathbf{I}_{52} & \sigma_{2,2}\mathbf{I}_{52} & \cdots & \sigma_{2,74}\mathbf{I}_{52} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{74,1}\mathbf{I}_{52} & \sigma_{74,2}\mathbf{I}_{52} & \cdots & \sigma_{74,74}\mathbf{I}_{52} \end{pmatrix} = \mathbf{\Sigma} \otimes \mathbf{I}_T$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,74} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,74} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{74,1} & \sigma_{74,2} & \cdots & \sigma_{74,74} \end{pmatrix}$$

and $\sigma_{ij} = E(v_{it}v_{jt})$ as defined above. Call $\hat{\mathbf{\Sigma}}$ the estimated $\mathbf{\Sigma}$ using $\hat{\sigma}_{ii} = \sum_t (\tilde{y}_{it} - \tilde{x}_{it}b_{FE})^2 \big/ T$ and $\hat{\sigma}_{ij} = \sum_t (\tilde{y}_{it} - \tilde{x}_{it}b_{FE})(\tilde{y}_{jt} - \tilde{x}_{jt}b_{FE}) \big/ T$, where $\tilde{y}$ and $\tilde{x}$ are the "within" transformation of the dependent and independent variables and b_{FE} is the corresponding fixed-effects estimates. Then, $\hat{\mathbf{\Sigma}}$ is a *(74*74)* matrix, but its rank is, at most, 52 (the lesser of *T=52* and *N=74*). $\hat{\mathbf{\Sigma}}$ is not of full rank and therefore is not invertible. If $\hat{\mathbf{\Sigma}}$ is not invertible, neither is $\hat{\mathbf{\Omega}}$.

non-reports is 43. Now, there is no selection procedure for these 52 plants that would have not introduced selection bias.[65] The method chosen to avoid the singularity of $\hat{\mathbf{\Sigma}}$ and at the same time to use the information of the 22 (74 – 52) plants that I have "in excess" of *T* was to obtain consistent point estimates of my parameters and then calculate robust standard errors for these estimates.[66] This method not only circumvents the impossibility of applying FGLS to produce consistent estimates of the parameters but it also allows me to draw correct inferences about the coefficient estimates.[67]

---

[65] The selection bias in question could have been explored in several ways. First, run two different regressions, one with the balanced panel comprised of the 43 plants with zero missing values and one with the unbalanced panel comprised of the 31 plants with at least one month not-reported, and then compare. Second, run C_{52}^{74} regressions with unbalanced panels and calculate the average, the standard deviation and the number of times the estimates were statistically significant. The obvious drawback of this second alternative is the amount of work involved. Of course, there is also the possibility of running 74 different regressions (one for each plant), reporting 74 different estimates for each parameter and their pooled descriptive statistics, but this would go against the original idea of the poolability of the data.

[66] I am grateful to Manuel Arellano for suggesting this to me via e-mail communication.

[67] A considered but discarded course of action was Panel Corrected Standard Errors (Beck and Katz, 1995 and Beck *et al.*, 1993). These authors suggest to obtain the point estimates by inefficient but consistent OLS and then to obtain correct estimates of their standard errors by $Cov(b_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$. "Any serial correlation of the errors must be eliminated before the panel-corrected standard errors are calculated"(p. 638). In such a case $\hat{\boldsymbol{\Omega}}$ would be exactly as in footnote 23, but with $\hat{\sigma}_{ii} = \sum_t (y_{it} - x_{it}b_{OLS})^2 \big/ T_i$ and $\hat{\sigma}_{ij} = \sum_t (y_{it} - x_{it}b_{OLS})(y_{jt} - x_{jt}b_{OLS}) \big/ T$. I did not use Panel Corrected Standard Errors mainly for two reasons. First, the motivation of Beck and Katz (1995) for suggesting them was the overconfidence produced by Park's (FGLS) standard errors, a point already made by Freedman and Peters (1984). My motivation here is somewhat different since I cannot use FGLS in the first place due to the fact that *N*>*T*. Panel Corrected Standard Errors were developed for panels with *T*>*N*. The second reason is an empirical one. I have two plants (#52 and #72) that did not have contemporaneous (common) observations. Furthermore, the unbalanced nature of the

Therefore, I first ran a least squares dummy variables (LSDV) model to obtain residuals to transform the data as in Cochrane-Orcutt. With the transformed data I run a second LSDV to estimate the parameters of the BOD_5 equation and the LOAD equation. Because my *T* is “large” (i.e., 52) this allows me to get consistent estimates of the “fixed effects”. With the residuals of the second LSDV, I calculate Arellano’s (1987) robust standard errors:

$$\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{N}\left(\tilde{\boldsymbol{X}}_i'\hat{\tilde{\varepsilon}}_i\hat{\tilde{\varepsilon}}_i'\tilde{\boldsymbol{X}}_i\right)\right)\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}$$

where $\tilde{\mathbf{X}}$ denotes the matrix of explanatory variables after the within-mean transformation, $\tilde{\mathbf{X}}_i$ denotes the same matrix but for plant *i*, and $\hat{\tilde{\varepsilon}}$ are the estimated residuals obtained from the second LSDV.

Arellano’s robust standard errors assume no contemporaneous correlation and are robust to panel heteroskedasticity and serial correlation. The reason for not calculating Arellano’s robust standard errors with the original data and instead transforming the model to eliminate autocorrelation of the errors first is that this technique assumes that *N* is large and *T* is small and the asymptotic results are derived as $N \to \infty$. In my panel, although it is true that *N*>*T*, it is also true that *T*=52 cannot be considered small. Therefore, by transforming the data to eliminate the serial correlation of the errors first I am taking into consideration’s Arellano’s (2003) cautionary note that when *T* is not small the robustness of this technique to serial correlation may decrease.[68]

---

panel greatly diminishes the number of observations to calculate the covariances $\hat{\sigma}_{ij}$. In other words, I cannot calculate all $\hat{\sigma}_{ij}$ to form $\hat{\Omega}$.

68 I thank Gabriela Sanromán for pointing this out to me.

7.3 TESTING THE ERROR STRUCTURE

7.3.1 Testing Contemporaneous Correlation of the Errors

As explained in Footnote 15, I have two plants (#52 and #72) that did not have contemporaneous (common) observations. In other words, I cannot calculate all $\hat{\sigma}_{ij}$ to perform the test for contemporaneous correlation of the errors, and therefore I cannot test the validity of the assumption of no contemporaneous correlation of the errors that underlies the application of Arellano's robust standard errors. Nevertheless, this assumption is justified by the fact that the unbalanced nature of the panel greatly diminishes the number of observations to calculate the covariances $\hat{\sigma}_{ij}$. Given that I have no observations that are common to all of the cross-sections, the estimated residual covariance matrix would be formed by temporally mismatched sources. While this procedure is consistent (as the number of observations within cross-sections approaches infinity), it is not likely to be a good estimator in this setting.

7.3.2 Testing for common vs. plant specific serial correlation of the errors

The Durbin-Watson statistic of the LSDV regression was 1.2812. This value suggested rejecting the null hypothesis of non-autocorrelation of the errors in favor of the alternative of first-order autocorrelation.

Having rejected the null of no autocorrelation, it is convenient to test whether this autocorrelation is common to all plants or is plant-specific. I perform this test as a classical Chow test extended for the case of *N* linear regressions, one for each plant. The restricted model is the pooled model $v_{it} = \rho v_{i,t-1} + \varepsilon_{it}$. The unrestricted model is:

$v_{it} = \rho_i v_{i,t-1} + \varepsilon_{it}$, so the null hypothesis can be written as H_0 $\rho_i = \rho$ for all *i*. To perform

this test I calculated ρ as $\rho = \frac{\sum_i \sum_t v_{it} v_{i,t-1}}{\sum_i \sum_t v_{i,t-1} v_{i,t-1}}$ and ρ_i as $\rho_i = \frac{\sum_{t_i} v_{it} v_{i,t-1}}{\sum_{t_i} v_{i,t-1} v_{i,t-1}}$, *i=1,..., 74,*

that is, as the OLS estimates of the pooled and unpooled models, with v_{it} being the residuals of the LSDV model estimation.

The test statistic is:

$$F = \frac{\left(\hat{v}'\hat{v} - \hat{v}_1'\hat{v}_1 - \hat{v}_2'\hat{v}_2 - \ldots - \hat{v}_N'\hat{v}_N\right)/(N-1)*(K+1)}{\left(\hat{v}_1'\hat{v}_1 + \hat{v}_2'\hat{v}_2 + \ldots + \hat{v}_N'\hat{v}_N\right)/N*\left(T-(K+1)\right)}$$

where , as usual, *N* is the number of plants, *T* is the number of months and *K* is the number of parameters (equal to one in this case). Under H_0, *F* is distributed $F((N-1)*(K+1), N*(T-(K+1)))$. The value of *F* obtained was 1.4509. The critical value for *F*((74-1)*(2),(74*(41-2))) = *F*(146, 3034) tends to one. Therefore, the test suggests that the null hypothesis of common autocorrelation be rejected in favor of the alternative hypothesis of plant-specific autocorrelation.

7.3.3 Testing for panel heteroskedasticity

Finally, I test for the presence of panel heteroskedasticity with three different tests: Bartlett, Levene and Brown-Forsythe. The results of these tests are presented in Table 7.2. All the tests in this table suggest rejecting the null hypothesis of panel homoskedasticity in favor of the alternative that not all of the plant-specific errors' variances are the same.

Table 7.2. Test for the Equality of Variances Between Residuals

Sample: 1997:07 2001:10
Included observations: 52

Method	Df	Value
Bartlett	73	1194.6
Levene	(73, 2812)	8.3630
Brown-Forsythe	(73, 2812)	6.9412

APPENDIX 7.1

Fixed Effects Logit

(Taken from Greene (1997), P. 899)

In a fixed effects logit model:

$$\Pr ob(y_{it} = 1) = \frac{e^{\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it}}}{1 + e^{\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it}}}$$

In this non-linear model it is not possible to sweep out the heterogeneity by taking differences from the mean, as in the linear case. In addition, even if it is possible to estimate the parameters, any desirable properties of the estimated individual effects, α_i, will depend on increasing T, which will not make sense in the typical panel. This will be particularly problematic for maximum likelihood estimators , whose only desirable properties are asymptotic. The solution, as Chamberlain suggests, is to remove the heterogeneity by some other means, and thereby finesse the problem of estimating the α_i s.

Chamberlain (1980) suggested estimating this model for panels with large *n* and small *T*. (It is not actually limited to small *T*, but the computation of the probabilities becomes "unwieldy" as *T* grows). His suggestion is that we consider the set of T_i observations for unit *i* as a group and maximize the conditional likelihood function

$$L^c = \prod_{i=1}^{n} \text{Prob}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iT_i} = y_{iT\,i} \Big| \sum_{t=1}^{T_i} y_{it}) \qquad (1)$$

that is, the likelihood function for each set of T_i observations conditioned on the number of 1's in the set, instead of maximizing the usual unconditional likelihood function,

$$L = \prod_{i} \prod_{t} \text{Prob}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iT_i} = y_{iT\,i}) \qquad (2)$$

Example: Assume a large number of cross-sectional units (n), each observed in two periods, 1 and 2.

The unconditional likelihood is:

$$L = \prod_{i=1}^{n} \text{Prob}(Y_{i1} = y_{i1}) \text{Prob}(Y_{i2} = y_{i2}) \qquad (3)$$

For each pair of observations we have the following possibilities:

(1) $y_{i1} = 0$ and $y_{i2} = 0$, $\text{Prob}(0{,}0 | \text{sum} = 0) = 1$

(2) $y_{i1} = 1$ and $y_{i2} = 1$, $\text{Prob}(1{,}1 | \text{sum} = 2) = 1$

The *ith* term in equation (1) above for either of these two possibilities is 1. So when we take logs these terms disappear. The units (plants) with all observations equal to 1 or 0 are dropped out. They contribute nothing to the conditional likelihood function.

The remaining possibilities are:

(3) $y_{i1} = 0$ and $y_{i2} = 1$, $\text{Prob}(0{,}1 | \text{sum} = 1) =$

$$\frac{\text{Prob}(0{,}1 \text{ and sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \frac{\text{Prob}(0{,}1)}{\text{Prob}(0{,}1) + \text{Prob}(1{,}0)}$$

(4) $y_{i1} = 1$ and $y_{i2} = 0$, $\text{Prob}(1{,}0 | \text{sum} = 1) =$

$$\frac{\text{Prob}(1{,}0 \text{ and sum} = 1)}{\text{Prob}(\text{sum} = 1)} = \frac{\text{Prob}(1{,}0)}{\text{Prob}(0{,}1) + \text{Prob}(1{,}0)}$$

For this pair of observations the conditional probability is

$$\frac{\dfrac{1}{1+e^{\alpha_i+\beta' x_{i1}}}\dfrac{e^{\alpha_i+\beta' x_{i2}}}{1+e^{\alpha_i+\beta' x_{i2}}}}{\dfrac{1}{1+e^{\alpha_i+\beta' x_{i1}}}\dfrac{e^{\alpha_i+\beta' x_{i2}}}{1+e^{\alpha_i+\beta' x_{i2}}}+\dfrac{e^{\alpha_i+\beta' x_{i1}}}{1+e^{\alpha_i+\beta' x_{i1}}}\dfrac{1}{1+e^{\alpha_i+\beta' x_{i2}}}}$$

$$=\frac{e^{\beta' x_{i2}}}{e^{\beta' x_{i1}}+e^{\beta' x_{i2}}}$$

By conditioning on the sum of the two observations the fixed effects are removed. The product of the terms such as the preceding, for those observations sets for which the sum is not zero or T_i, constitutes the conditional likelihood function.

APPENDIX 7.2

DEFINITIONS OF VARIABLES

Table A.7.1

Name	Definition
1997-1998	Dummy equal to one in months of 1997 and 1998 during which the IMM conducted special, IADB-financed monitoring campaigns
$BOD5_{i,t}$ =	Biological Oxygen Demand concentration of discharges, in mg/l
$CARRASCO1999_{i,t}$=	Dummy equal to one in the months of 199 during which the DCA conducted a special monitoring campaign in the Carrasco stream
$DURINGPLAN_{t}$=	Dummy equal to one during the Industrial Pollution Reduction Plan months
$ENERGY_{i,t}$ =	Total energy consumption in mega joules (MJ)
$EADCACUM_{i,t}$=	Number of enforcement actions (orders and fines) imposed by the DCA against the plant in the last twelve months
$FLOW_{i,t}$ =	Daily average effluent flow (m3/day)
$INSPDCA_{i,t}$ =	Dummy equal to one if plant *i* was inspected by the DCA in month *t*
$INSPDCACUM_{i,t-1}$=	Number of inspections performed by the DCA in the plant during the last twelve months
$INSPDCAOTHERCUM_{i,t-1}$=	Number of inspections performed by the DCA in the rest of the plants during the last twelve months
$INSPIMM_{i,t}$ =	Dummy equal to one if plant *i* was inspected by the IMM in month *t*
$INSPIMMCUM_{i,t-1}$=	Number of inspections performed by the IMM in the plant during the last twelve months
$INSPIMMOTHERCUM_{i,t-1}$=	Number of inspections performed by the IMM in the rest of the plants during the last twelve months
$INSPSEINCO_{i,t}$=	Dummy equal to one if plant *i* was inspected by SEINCO in month *t*
$INSPSEINCOCUM_{i,t-1}$=	Number of inspections performed by SEINCO in the plant during the last twelve months
$FINEDIMMCUM_{i,t-1}$=	Number of fines imposed by the IMM against the plant in the last twelve months
$LABOR_{i,t}$ =	Total days-employee worked
$LOAD_{i,t}$ =	*(BOD5*FLOW)* = Total organic pollution discharged in (mg/day)
$PINSPIMM_{i,t}$ =	Probability of being inspected by the IMM
$PINSPDCA_{i,t}$ =	Probability of being inspected by the DCA
$PINSPSEINCO_{i,t}$ =	Probability of being inspected by SEINCO
$P_{q,t}$=	Price of the good produced
PTY_{i}=	Dummy equal to 1 if the plant is a Priority 1 plant
$RF_{i,t}$ =	The number of reporting failures in the previous two reporting periods
$STREAM_{i,t}$=	Dummy equal to one if the plant emits directly into a water body
$TANNERY_{i}$=	Dummy equal to one if the plant is a tannery

$TECH_{i,t}$ =	Dummy equal to one after plant modified their treatment plants
$VIOL_{i,t}$ =	Dummy equal to one if the plant reported a violation
VOL_t =	Monthly level of the industry production volume index
$WATER_{i,t}$ =	Total water consumption in m3/month
$WOOL_i$=	Dummy equal to one if the plant is a wool washer