

**GEOGRAPHIC CONCENTRATION AND ESTABLISHMENT SCALE:  
CAN PANEL DATA TELL US MORE?\***

Salvador BARRIOS<sup>1</sup>, Luisito BERTINELLI<sup>2</sup> and Eric STROBL<sup>3</sup>

May 2003

**Abstract**

In a recent study, Holmes and Stevens (2002) identify for the first time a positive relationship that exists between establishment scale and local industry concentration using a large cross-sectional plant level data set for the US. Using an exhaustive plant level panel data set for Irish manufacturing covering nearly three decades, we are able to extend their analysis in two ways. Firstly, we show that failing to control for fixed effects biases the relationship upward, although the essence of it still remains. Secondly, the link is substantially weaker when plants locate for the first time in an area, but strengthens with age for those that survive in the long run. We link our results to recent contributions on the dynamics of geographic concentration.

Keywords: agglomeration, plant size, Ireland

JEL classification: R12, C23

---

<sup>1</sup>CORE, Université catholique de Louvain, 34 Voie du Roman Pays 1348 Louvain-La-Neuve, Belgium.

<sup>2</sup>CORE, Université catholique de Louvain, 34 Voie du Roman Pays 1348 Louvain-La-Neuve, Belgium.

<sup>3</sup>CORE, Université catholique de Louvain, 34 Voie du Roman Pays 1348 Louvain-La-Neuve, Belgium.

We are grateful for comments by Tom Holmes and Giordano Mion. Any errors are ours alone.

This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors. This research has also benefited from financial support through the RTN research project "Specialization versus diversification: the microeconomics of regional development and the spatial propagation of macroeconomic shocks in Europe" of the European Commission (grant No. HPRN-CT-2000-00072). The second author gratefully acknowledges financial support from the Belgian FNRS. The third author gratefully acknowledges financial support from the European Commission through a Marie Curie Fellowship.

## Section I - Introduction

In a recent paper Holmes and Stevens (2002) [HS] explicitly document, for the first time, what appears to be a simple empirical fact: establishments tend to be bigger in geographical areas where industry is most concentrated. More specifically, using a large cross-sectional database providing information on the location of plants located in the US for the year 1992, they show for all broad sectors of the economy that such a link exists by regressing establishment scale, measured by employment level, on a measure of the geographic concentration of industries. HS's finding thus significantly adds to the existing empirical literature concerning the effects of the distribution of economic activity across space; see, for instance, Henderson (1986, 1994), Kim (1995), and Ciccone and Hall (1996). Moreover, the nature of their analysis complements a number of recent papers by Ellison and Glaeser (1997) and Dumais et al. (2002) on the measurement and decomposition of geographic concentration using plant level data.<sup>4</sup>

In the current paper we are able to extend HS's study by using exhaustive plant level data for Irish manufacturing covering the period 1973-2000. Specifically, exploiting the panel nature of the data allows us to point out two important aspects regarding the link between establishment size and geographic concentration. First, the link is in essence robust to controlling for unobserved industry, region, region/industry and plant level fixed effects that could have potentially been biasing the HS results, although in magnitude somewhat lower. Second, we show that this relationship already exists when a plant first starts up in a particular location. However, at start-up it is weaker compared to that found for incumbent plants, and subsequently only strengthens for long-term survivors over their life cycle. We link these results to the findings concerning the dynamics of geographic concentration by Dumais et al. (2002).

---

<sup>4</sup> A number of studies have now implemented and extended these techniques; see, for example, Maurel and Sédillot (1999), and Duranton and Overman (2002).

The paper is organised as follows. Section II presents the equation tested. Section III provides a description of the data. Section IV presents the results while Section V concludes.

## Section II – Equation tested

HS consider two levels of aggregation in order to study the relationship between scale and concentration: the industry-location level, which they call the *location-level*, and the *plant-level*. For the *location-level*, HS use the simple location quotient

$$Q_{i,l}^x = \frac{x_{i,l}/x_l}{x_i/x} \quad (1)$$

where  $x$ ,  $x_{i,l}$ ,  $x_l$  and  $x_i$  are, respectively, the level of total employment in the manufacturing industry, the level of employment of sector  $i$  in region  $l$ , the total of employment of sector  $i$ , and of region  $l$ . They derive the following equation to be estimated:

$$q_{i,l,t}^s = \beta^s q_{i,l,t}^x + \varepsilon_{i,l,t} \quad (2)$$

where  $\varepsilon_{i,l,t}$  is an *i.i.d.* error term,  $q^x$  is the logarithmic value of (1) and  $q^s$ , a measure of average plant size, is given by the logarithmic value of

$$Q_{i,l}^s = \frac{x_{i,l}/n_{i,l}}{x_i/n_i} \quad (3)$$

where  $n_i$  and  $n_{i,l}$  are the total number of firms in sector  $i$ , and sector  $i$  and region  $l$ . Equation (2) thus examines how the relative industry specialisation of a location affects the average relative size of plants of sectors within that location. Specifically, if  $\beta^s > 0$ , then the more concentrated an industry is within a region the higher the average scale of its plants will be within that location.

The relationship between scale and concentration can similarly be examined at the plant level. For this case one estimates the relationship between the location quotient specific to industry  $i$  and location  $l$  where a particular plant  $e$  is located and the relative size of this establishment compared to the average establishment size in the industry  $i$ . The (nearly) analogous establishment-level quotients at the plant level  $e$  are  $q_e^x$  and  $q_e^s$  specific to each plant:

$$q_e^x = \ln \left( \frac{\frac{x_{i,l} - x_e}{x_l - x_e}}{\frac{x_i - x_e}{x - x_e}} \right) \quad (4)$$

$$q_e^s = \ln \left( \frac{\frac{x_e}{(x_i - x_e)}}{(n_i - 1)} \right) \quad (5)$$

where  $x_i$  is the plant specific level of employment and the other variables are defined as above. One should note that each establishment is assigned the location quotient specific to its industry  $i$  and location  $l$  excluding its own employment level from the calculation. The same applies to  $q_e^s$ . By excluding the current plant employment level, HS index thus avoid potential simultaneity bias given that, for example, the entry of a very large plant in a particular region-industry could increase the location quotient for that location. Changing the notation in (2) provides the equation to be estimated at the plant level:

$$q_{e,i,l,t}^s = \beta_e^s q_{e,i,l,t}^x + \varepsilon_{e,i,l,t} \quad (6)$$

Similarly to above, the coefficient  $\beta_e^s$  obtained will give us the relationship between the actual size of an establishment and the level of concentration of the industry at the location where this plant is located.

### Section III – Data Set

Our data source is the *Forfás Employment Survey* collected by *Forfás*, the policy and advisory board for industrial development in Ireland, since 1973, and we have access to this data up until and including the year 2000. The response rate to this survey is argued by *Forfás* to essentially be nearly 100 per cent, i.e., our data can be seen as including virtually the entire population of manufacturing plants in Ireland at any point in time during the sample period. Information at the plant level includes a unique plant identifier that allows one to link plants across years, their sector of production, the regional location of each plant, the year of start-up, and the level of employment. Sectors are classified according to four digit European NACE Revision 1 codes, which roughly corresponds in the degree of disaggregation to the four digit SIC sectoral classification in the US data used by HS.<sup>5</sup> The location of plants is categorised into the 26 counties of Ireland, the standard Irish administrative regional decomposition. The average size of Irish counties is about 2700 sq. km, while average county size in the US is around 1140 sq. km. Average population densities are however significantly higher for the latter.

Two other features of our data that lie in contrast to that of HS are noteworthy. Firstly, our data provides us with the *actual* level of employment for all plants rather than, as in HS's data, being grouped into size categories, thus avoiding any potential bias resulting from assuming uniform employment distributions within categories. Secondly, our data only covers manufacturing whereas HS had data for the whole economy and thus could examine the issue at hand for all broad sectors. With regard to this latter aspect one should note that HS consistently found a stronger relationship between establishment scale and local industry concentration to be for manufacturing compared to other sectors of the economy.

## Section IV – Empirical Analysis

Our results for the *location-level* regression as described in equation (2) using simple OLS are presented in the first row of Table 1.<sup>6</sup> As can be seen, we find, as HS, that local industry concentration has a positive effect on establishment size. The coefficient  $\beta^s$  is equal to 0.597, which is larger, but still comparable to the one found by the HS for US manufacturing (0.415). We also checked whether this difference is due to the fact that our data spans a long time period by interacting our time dummies with local concentration, but in almost all cases these turned out to be insignificant with very small coefficients, thus suggesting that the relationship has not changed over time.

With access to only cross-sectional data for one year HS were unable to fully investigate whether the failure to account for time invariant unobservables (to the researcher) possibly correlated with local industry concentration could be driving or biasing their result; see Greene (2000). However, the panel nature and the long time span of our data allow us to explicitly examine this issue. We first proceeded by re-estimating equation (2), by also including regional dummies only to control for unobserved region-specific effects, the results of which are in the second row of Table 1. Accordingly, controlling for regional fixed effects makes no noticeable difference to the size and statistical significance of the relationship. Including instead industry dummies, with and without region dummies, as depicted in the third and fourth row, actually strengthens the link between establishment size and local concentration.

It must be noted, however, that even with industry and region dummies included in equation (2) there may still be time invariant effects specific to each region/industry unit that are biasing the estimate of  $\beta^s$ . For example, a particular region may have certain

---

<sup>5</sup> There are 860 4-digit SIC classes compared to 503 4-digit NACE classes in manufacturing. Ideally we would have liked to convert our sectoral codes to the SIC classification. However, there does not exist a one to one correspondence between these two classification systems at this level of disaggregation.

<sup>6</sup> All regressions presented include time dummies given that we have information across a 28-year period.

(unobservable) natural advantages that are important to particular industries and thus cause greater agglomeration of these. The panel nature of our data also allows us to control for these types of factors if they are time invariant by employing a fixed effects estimator, thus purging all time invariant unobserved factors from (2). The results of our panel estimation detailed in the fourth row show that the coefficient is still highly significant compared to the simple OLS results without industry and regional dummies and about 25 per cent higher.<sup>7</sup> Thus our results indicate that at the *location-level* the relationship between establishment scale and local industry concentration is not an artefact of time invariant unobservables, potentially correlated with local industry concentration.

Proceeding to the estimates of the plant level equation (6), we first used simple OLS without controlling for any fixed effects as given in the fifth row of Table 1. As HS for US manufacturing, we find that the relationship between establishment size and local concentration is statistically significant. Moreover, the size of our coefficient, 0.11 is fairly similar to what the authors find for the US at the county level (0.126). Adding region dummies on their own or in conjunction with industry dummies, the results of which are depicted in rows 6 and 7, respectively, does not alter the significance of the relationship, but does lower its size by a little over 25 per cent.<sup>8</sup> Thus regional and industry specific effects, while not qualitatively changing the relationship, will cause an upward bias if not controlled for. As above, one may also postulate that certain regions have natural advantages that are important for certain industries so that industry/region fixed effects may be driving the observed link rather than local industry concentration. Moreover, the estimates may be further biased because the specification does not include

---

<sup>7</sup> A simple t-test reveals that this difference is statistical significant.

<sup>8</sup> The fact that industry dummies alter the size of the coefficient even though the dependent variable is normalised by average industry size, suggests that the time invariant unobservable industry effects are not very correlated with average industry size.

any other plant level controls.<sup>9</sup> A general fixed effects estimation again allows us to purge all time invariant factors of this kind, which could potentially be correlated with the concentration proxy, from our regression, see Greene (2000).

The result of estimating (6) with a fixed effects estimator are depicted in row 9. Accordingly, this induces only a further 10 per cent decrease in the coefficient, which still remains highly significant. One can thus conclude that at the plant level analysis the relationship between establishment size and local concentration is robust to controlling for all unobserved fixed effects as it was for the *location-level* regression, although for the former the strength of this relationship is somewhat weakened.

Our panel data also allows us to examine how the link between establishment size and local concentration evolves over a plant's life cycle. For example, it may take some time before a plant reaches its efficient scale. Also, one may suspect that there is greater uncertainty about this scale in the earlier part of the life cycle of a plant. Our ability to address the relationship over the life cycle derives from the fact that we can explicitly identify plant births and trace them over their lifecycles as far as these fall within our sample period. Plant births, and hence their subsequent age, are identified by their first year of positive employment and for those plants that existed prior to the first year of collection of the data there is information on the start-up year from which one can readily calculate their age.

We first ran our basic regression, including industry and regional dummies, using observations for the year of start-up of a plant and those for incumbent plants separately. As can be seen from the 10<sup>th</sup> and 11<sup>th</sup> rows of Table 1, the coefficient for plant start-ups

---

<sup>9</sup> We experimented with including the few plant level variables that one can extract from our employment survey, namely, nationality of ownership and plant age (for the plants that started up over our sample period), but these changed little in the estimates. Further details on all non-reported, but discussed, results are available from the authors.

is positive and significant, but about half that for incumbents.<sup>10</sup> A simple t-test verifies that this difference is statistically significant.

The fact that the relationship between start-up size and local concentration is weaker than for the entire data set suggests that the link may become stronger as plants age. Of course, the age a plant acquires, and thus possibly its size, at any point in time is intrinsically linked to its ability to survive. Many plants that enter a (local) market may not be able to survive over the long-term. If survivability also determines the way a plant's size evolves according to local concentration, then it is clearly important to take account of differences between plants according to their ability to survive in the long run. To do this in a simple way we classified plants into survivors and non-survivors. The choice of when a plant should be considered a survivor is, however, not clear-cut. Due to data considerations in terms of having a reasonable and representative sample size of each group, we assume that plants that survive more than ten years are 'survivors'. Choosing different thresholds would have resulted in more unbalanced groups and often unfeasible sample sizes.<sup>11</sup> One must note, however, that we did also experiment with using a five years cut-off point and that this produced qualitatively similar results.

To proceed we included a set of ten age dummies and their interaction with the local concentration measure in our base plant level specification, thus allowing us to estimate the impact of age on the link between size and local concentration. This specification was estimated for observations on plants no more than ten years old, for the survivor and non-survivor groups separately. The results of this exercise are reported in Table 2. We also provide plots of the coefficients of the total effect by age, obtained

---

<sup>10</sup> Given that there is only one observation per plant, we are not able to run a fixed effects estimation for the start-up observations. We thus also used OLS, but included region and industry dummies, for the incumbent plant regression. The coefficient using fixed effects for incumbents, which comprise an overwhelmingly majority of all observations, was, unsurprisingly, nearly identical to that of the overall sample.

<sup>11</sup> Choosing ten years as the cut-off points means excluding all plant births after 1990. We thus also limited our total sample to only observation from prior to this year for this exercise. Excluding the last ten years from our data and estimating our overall plant level regression produced almost identical results.

by adding the coefficients for the base year and the appropriate interaction terms, in Graphs 1 and 2 for survivors and non-survivors, respectively.<sup>12</sup> For survivors, all the interaction terms proved to be significant, and it is clear from the graph that in general the strength of the relationship increases as the plants age. In contrast, for non-survivors not only are few of the interaction terms statistically significant, but any positive relationship is clearly absent. As a matter of fact, for those that survived more than six but less than 11 years, the relationship is actually significantly weaker than that at birth.<sup>13</sup> As a matter of fact, for those short-term survivors that last at least 8 years the relationship becomes in aggregate negative – although this latter result must be viewed with considerable caution as it is sensitive to our choice of cut-off point for long-term survival and due to a very small proportion of the sample of non-survivors.

Our results with regard to the life cycle of plants can be put in the context of those found by Dumais et al. (2002) in their investigation of the components driving geographic concentration for the US. Specifically, they discover that births of plants act to decrease, while plant closures tend to increase geographic concentration, and this result is confirmed for Irish manufacturing by Barrios et al. (2003) using the same data as here. Relating this to our findings, one can think of births decreasing geographic concentration in two, not necessarily mutually exclusive, ways. For one, new plants may be more likely to locate in less concentrated regions. However, it may also be the case that new plants that locate in less concentrated areas tend to be larger than those that locate in more concentrated areas. The fact that we find that the positive relationship between size and geographic concentration also holds for plant births, although it is weaker, suggests that the latter is unlikely to serve as an explanation for the decreasing

---

<sup>12</sup> The local concentration variable on its own was significant for both groups and its value can be read as the intercept of the line.

<sup>13</sup> One must note, however, that the number of observations for each age group decreases substantially the higher the age.

effect of births on geographic concentration. In fact, for this explanation to hold we would have had to find a negative coefficient for plant births and this was not the case.

One can similarly interpret our findings with regard to the result on plant closures by Dumais et al. (2002). The increasing effect of closures on geographical concentration could be due to plants being more likely to close in less concentrated areas. However, it may also be the result of the possibility that plants that close in less concentrated areas are larger. Given that in general the probability of closure is much higher in the early years of a plant's life cycle, see, for instance, Geroski (1995), our finding of a positive relationship between size and geographic concentration for at least the earlier years of non-survivors, would tend to rule out the latter explanation. To further check this we re-ran (6) for all observations of plants exiting Irish manufacturing in the year immediately before closure including industry and region specific dummies and, as can be seen in the last row of Table 1, we found that the coefficient of interest to still be positive and significant, although slightly lower than for the overall sample.<sup>14</sup>

#### **Section IV – Conclusion**

In this paper we test the relationship between establishment scale and the regional concentration of industries using exhaustive employment panel data on Irish manufacturing plants for the period 1973-2000. Our analysis extends the recent study by Holmes and Stevens (2002) for the US. Besides confirming their evidence of positive link between these two variables, our results bring two important additional conclusions: first, we find that this relationship is qualitatively robust to controlling for fixed effects although we discover that not doing so may bias the result upward at the plant level; second, the link is substantially weaker for start-up plants but strengthens as plants get

---

<sup>14</sup> As with plant births, given that there is at most one observation per plant we were not able to control for plant specific fixed effects.

older and survive in the long run; this latter result provides further insight into the link between plants' life cycle and the dynamics of agglomeration.

## References

- Barrios, S., L. Bertinelli and E. Strobl (2003). "The Dynamics of Geographic Concentration in Ireland", mimeo.
- Ciccone, A. and R. E. Hall, 1996. "Productivity and the Density of Economic Activity", *American Economic Review* 86(1), 54-70.
- Dumais, G., G. Ellison and E.L. Glaeser (2002). "Geographic Concentration as a Dynamic Process", *Review of Economic and Statistics*, 84 (2), 193-204.
- Duranton, G. and H. Overman, 2002. "Testing for Localization Using Micro-Geographic Data", CEPR DP 3379
- Ellison, G. and E. Glaeser, 1997. "Geographic concentration in U.S. manufacturing industries: a dartboard approach". *Journal of Political Economy* 105 (5), 889-927.
- Geroski, P.A. (1995) "What do we know about entry?", *International Journal of Industrial Organization*, 13, 421-440.
- Greene, W.H., (2000). Econometric Analysis, 4<sup>th</sup> Edition, Prentice Hall International, Inc.
- Holmes, T. J. and J.J. Stevens, (2002). "Geographic Concentration and Establishment Scale", *Review of Economics and Statistics* 84(4), 682-690.
- Henderson, J.V. (1986). "Efficiency of Resource Use and City Size", *Journal of Urban Economics* 19(1), 47-70.
- Henderson, J.V., (1994). "Where Does an Industry Locate?", *Journal of Urban Economics* 35(1), 83-104.
- Kim, S., (1995). "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U.S. Regional Manufacturing Structure, 1860-1987", *Quarterly Journal of Economics*, 110(4), 881-908.
- Maurel, F. and B. Sédillot, (1999). "A Measure of the Geographic Concentration in French Manufacturing Industries". *Regional Science and Urban Economics* 29(5), 575-604.

**Table 1 – General Estimations**

<b>ROW</b>	<b>Sample</b>	<b>Regional FE</b>	<b>Industry FE</b>	<b>General FE</b>	<b>Obs.</b>	<b><math>\beta</math></b>	<b>p-value</b>
<b>(1)</b>	Location Level	No	No	No		<i>0.597</i>	0.00
<b>(2)</b>	Location Level	Yes	No	No		<i>0.614</i>	0.00
<b>(3)</b>	Location Level	Yes	Yes	No		<i>0.691</i>	0.00
<b>(4)</b>	Location Level	Yes	Yes	Yes		<i>0.749</i>	0.00
<b>(5)</b>	Plant Level – All obs.	No	No	No		<i>0.111</i>	0.00
<b>(6)</b>	Plant Level – All obs.	Yes	No	No		<i>0.115</i>	0.00
<b>(7)</b>	Plant Level – All obs.	Yes	Yes	No		<i>0.077</i>	0.00
<b>(8)</b>	Plant Level – All obs.	No	Yes	No		<i>0.080</i>	0.00
<b>(9)</b>	Plant Level – All obs.	Yes	Yes	Yes		<i>0.070</i>	0.00
<b>(10)</b>	Plant Start-Ups	Yes	Yes	No		<i>0.042</i>	0.00
<b>(11)</b>	Incumbent Plants	Yes	Yes	No		<i>0.081</i>	0.00
<b>(12)</b>	Closures	Yes	Yes	No		<i>0.055</i>	0.00

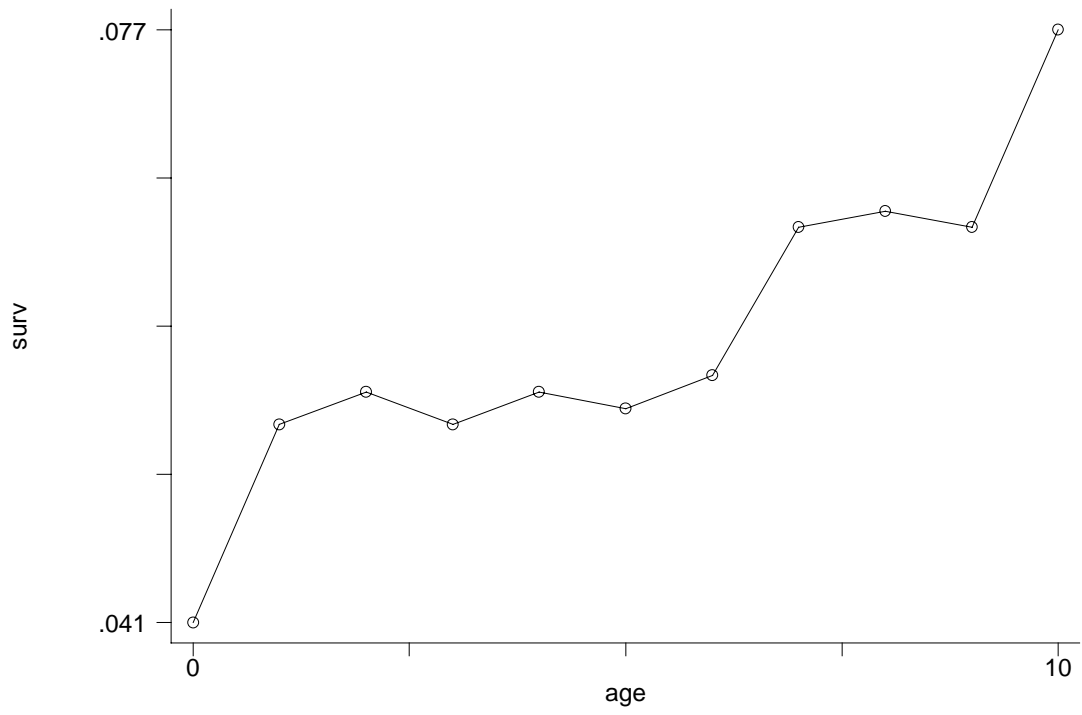
Notes: (1) p values in parantheses. (2) All regressions included time dummies. (3) General FE refers to a fixed effects estimation where all time invariant effects are purged from the equation; see, Greene (2000).

**Table 2 – Estimations for Survivors and Non-Survivors**

	<b>Survivors</b>	<b>Non-Survivors</b>
<b>q<sup>x</sup></b>	0.041*** (0.000)	0.033*** (0.001)
<b>q<sup>x</sup> * AGE1</b>	0.012* (0.062)	0.007 (0.311)
<b>q<sup>x</sup> * AGE2</b>	0.014** (0.030)	0.006 (0.403)
<b>q<sup>x</sup> * AGE3</b>	0.012* (0.075)	-0.004 (0.635)
<b>q<sup>x</sup> * AGE4</b>	0.014** (0.041)	0.002 (0.850)
<b>q<sup>x</sup> * AGE5</b>	0.013** (0.049)	-0.009 (0.365)
<b>q<sup>x</sup> * AGE6</b>	0.015** (0.027)	-0.013 (0.245)
<b>q<sup>x</sup> * AGE7</b>	0.024*** (0.001)	0.005 (0.684)
<b>q<sup>x</sup> * AGE8</b>	0.025*** (0.001)	-0.027* (0.077)
<b>q<sup>x</sup> * AGE9</b>	0.024*** (0.002)	-0.053*** (0.005)
<b>q<sup>x</sup> * AGE10</b>	0.036*** (0.000)	-0.083*** (0.002)
<b># Obs.</b>	53639	27842
<b># Plants</b>	5760	6094
<b>R-squared</b>	0.23	0.07

Notes: (1) \*\*\*, \*\*, and \* represent one, five and ten per cent significance levels, respectively. (2) standard errors in parantheses. (3) A constant term and age and time dummies on their own are also included but not their coefficients not reported.

Graph 1:  $\beta$ -coefficient for survivor plants



Graph 2:  $\beta$ -coefficient for non-survivor plants

