Longitudinal Micro Data Study of Regional Selective Assistance in Wales

Report*

Chiara Criscuolo
Ralf Martin

Centre for Economic Performance,
London School of Economics

EXECUTIVE SUMMARY

This report was commissioned by the Welsh Assembly in July 2007 to investigate the feasibility of using linked micro data in the evaluation of the Regional Selective Assistance in Wales. The report discusses in detail the issues related to the matching of administrative records from the Department of Trade and Industry (DTI) Selective Assistance Management Information System (SAMIS) database with the Office for National Statistics (ONS) Annual Respondents Database (ARD), which contains information on output and inputs on a stratified sample of enterprises in the UK. It describes the econometric approaches and methodologies that can be used to evaluate the programme and analyses the feasibility of an evaluation also for size classes and specific industries. The conclusion is that we can conduct a very rich set of analyses, from simple OLS to Instrumental variables also for different size classes and for aggregated industrial sectors. Finally in the appendix we report details on various issues concerning matching with the IDBR.

Table of Contents

Introduction

Introduced in 1972 Regional Selective Assistance (RSA) has become one of the main regional policy instruments in the UK. Its main aim was to safeguard/create employment in "disadvantaged" regions. During the course of the years, different parts of Wales have been eligible to RSA. This report explores the feasibility of an evaluation of RSA combining the SAMIS database, which is the register of all applicants maintained by the Department of Trade and Industry (DTI) with independent performance data from the Office for National Statistics (ONS) Annual Respondents Database (ARD).[1]

Most of the previous evaluation studies of RSA are based on "industrial survey" techniques where senior personnel of a randomly drawn sample of assisted firms are asked to give their subjective assessment of what the counterfactual situation would have been had they not received the grant (see AEP NERA 2003, Cambridge Economics). Few studies have used firm-level econometric techniques to evaluate the direct impact of RSA (Devereux, Griffith and Simpson, 2007; Harris and Robinson, 2003 and Criscuolo, Martin, Overman and Van Reenen, 2007).

Relative to industrial surveys, policy evaluation based on econometric analysis presents three main advantages: firstly, one can compare the firm before and after its exposure to the programme; while industrial surveys can only provide ex-post data on programme participants; i.e. after they have joined the programme. Secondly, and more importantly, we can compare the change in the participating firms' performance to a "control group" of firms who did not – or preferably could not - participate in the programme. Comparing the change in the treatment group's outcomes to the change in the control group's outcomes is the standard "difference in difference" approach that is the mainstay of the evaluation literature.

More conventional evaluation exercises such as performance questions on a programme application form or a survey among programme participants after completion are unlikely to be helpful in determining the causal impact of a program. Thirdly, independent performance data from an independent data source – in our case the ONS Annual Respondents Database - is less likely to be affected by strategic reporting by firms; which might overstate/understate the benefits of the programme.

The implementation of econometric policy evaluation is not free of problems. In our case the main problems derives from the SAMIS database and the ARD coming from different sources and the consequent problems of overlapping and matching.
The two datasets might not overlap because the ARD is a yearly survey of a stratified random sample of businesses, rather than a census, with smaller firms having lower probability of being sampled. This implies, especially for smaller businesses, that there is no guarantee that a recipient of RSA appears in the ARD and if it does, there is no guarantee that we will have sufficient information with which to estimate the impact of the programme. The matching problems arise because ARD and the SAMIS database were not designed to be matched. To combine both we have to rely on a name and postcode matching procedure, which sometimes can be un-reliable in the following ways:[2]

[1] Since 2004 the Welsh Assembly Government also maintains its own database of RSA support (rather than the SAMIS database). This started after the end of the sample period analysed.
[2] The following section contains a more elaborate description of the matching procedure, associated problems and how we are dealing with them.

- Because the ARD samples only a subset of the total population of all businesses in the UK, RSA recipients might not be sampled in the ARD.
- Because of issues such as variations in spelling of names, changes in postcodes or typos in either of the databases we might not be able to identify programme participants in the ARD even though they are sampled.
- A business can be identified in the ARD, but for large businesses that consist of several different enterprises and sites it is not clear which unit has been participating in the programme.

The matching algorithm could associate a business in the administrative dataset with the wrong ARD business. A perusal examination of the datasets reveals that this does not happen very often; i.e. we examined a number of cases in more detail but could not find any obvious mismatches. An obvious mismatch would occur if a programme participant would be associated with several ARD business units that are not owned by the same enterprise or enterprise group.

Matching problems aside we worry, especially in the case of an evaluation at the regional level for Wales, that we might not be able to match a sufficiently large number of RSA recipients to the ARD and that for these matched participants the ARD provides outcome data both before and after they receive support to estimate the causal impact of the program. Having a "large enough" number of matched participants id a key necessary condition for policy evaluation. This is because to run any estimator we need at least as many observations as we have parameters and most of the methods we apply rely on large samples properties to derive the distributions of the estimators (e.g. consistency; asymptotic efficiency); where by large sample we normally mean samples larger than 50.

This report provides our findings on these issues for the **R**egional **S**elective **A**ssistance (RSA) Programme in Wales.

In the remainder of this report we proceed as follows: the next section describes in more detail the institutional background of the Regional Selective Assistance. Section 3 describes each of the three databases underlying the evaluation policy, i.e. the SAMIS database; the IDBR, the interdepartmental business register and finally the ARD. Section 4 reports details on the steps that are required to match ARD and the SAMIS database; discusses the fundamental problems that one encounters in such an exercise and proposes how to deal with them. This is followed by a separate section reporting the results of this matching exercise. Section 6 concludes.

Regional Selective Assistance (RSA)

Institutional Background

Regional Selective Assistance (RSA) was the main regional business support scheme in the UK.[3] From the early 1970s it provided discretionary grants to companies in Assisted Areas. These are disadvantaged regions typically characterised by relatively high levels of unemployment and deprivation.

[3] We discuss our choice of study period below. According to Harris and Robinson (2004), in 1998-9 RSA represented 19% of the UK's industrial policy spending. In April 2004, the RSA scheme was replaced by the Selective Finance for Investment (SFI) scheme. Productivity became an official objective in April 2004, when RSA was replaced by SFI which explicitly requires that projects yield productivity improvements. We discuss the difference between the two schemes below.

It was designed to “create and safeguard employment” in Assisted Areas. There may also be some additional efficiency advantages if there is a spatial element to market failures; e.g. labour market rigidities affect different areas in different ways. Emphasis was given to internationally mobile investments, new products and processes and the manufacturing sector. Assistance could be provided to establish a new business; to expand, modernise or rationalise an existing business; to set up research and development facilities or enable businesses to take the next step from development to production.

Because RSA had the potential to distort competition and trade between European countries it must comply with European Union legislation concerning state aid. In general, this type of assistance has been prohibited by European law except in certain cases. In particular, Article 87 of the Treaty of Amsterdam allows for some state aid in support of the European Union’s regional development policies. The guidelines designate very deprived “Tier 1 Areas” (previously called “Development Areas”) in which higher rates of grant can be offered and slightly less deprived “Tier 2 Areas” (previously called “Intermediate Areas”).[4] The EU regional revised every seven years decide which areas are eligible. Figures 1 to 4 report the maps of the areas eligible for RSA up to 1993; between 1993 and 1999 and after 1999. The Figures show that in all of the maps parts of Wales have been “Assisted Areas”.

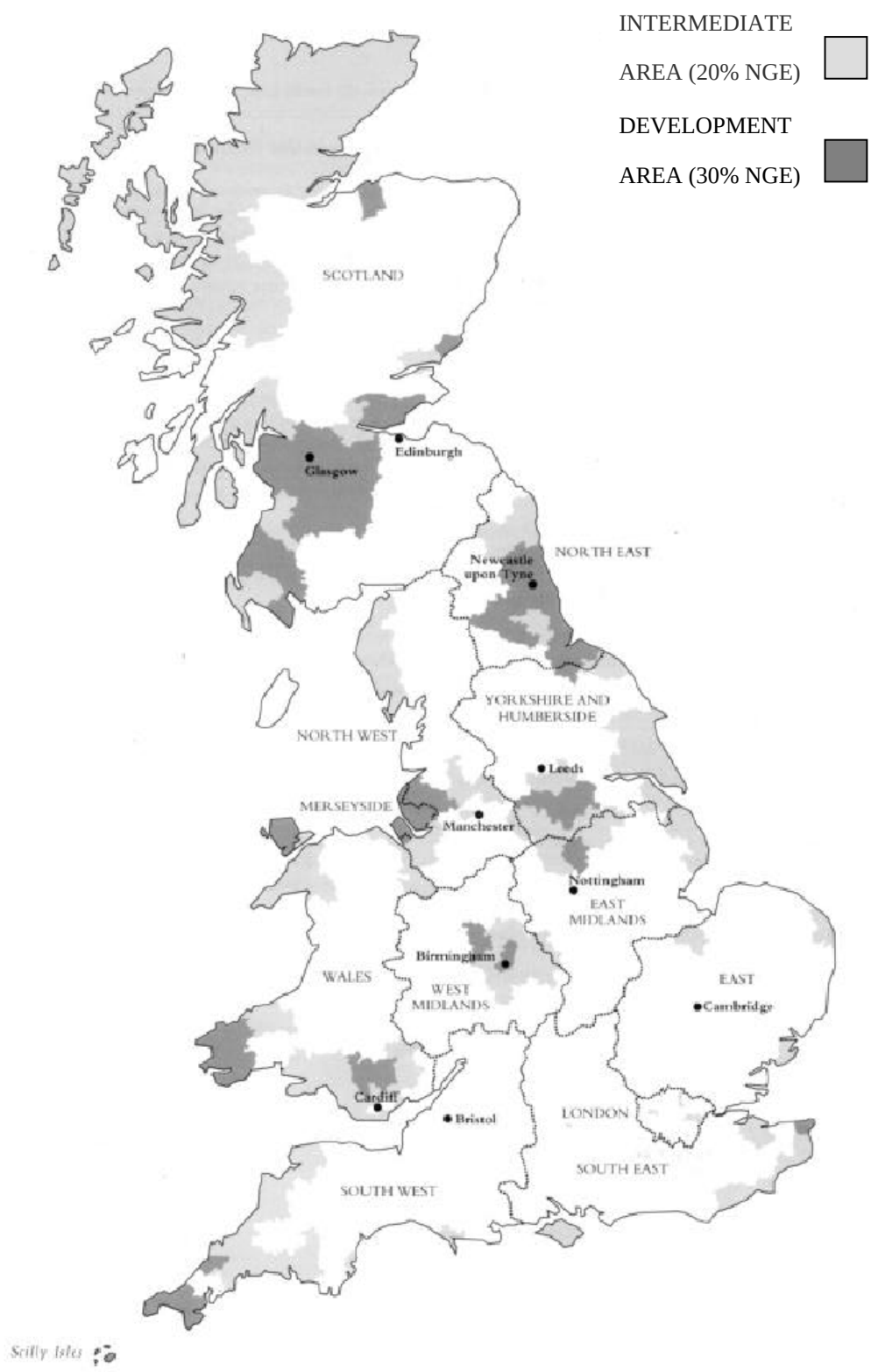
[4] Article 87(3) of the Treaty of Amsterdam defines conditions where State aid may be compatible with EU laws. Article 87(3) (a) allows for “aid to promote the economic development of areas where the standard of living is abnormally low or where there is serious underemployment” [Tier1/Development Areas] and Article 87(3) (c) allows for: “aid to facilitate the development of economic activities or of certain economic areas, where such aid does not adversely affect trading conditions to an extent contrary to the common interest.” [Tier 2 or intermediate Areas] Additional restrictions apply to sectors with over-capacity: motor vehicles, synthetic fibres and yarns, iron and steel, coal, fishery and agricultural products. There is an upper threshold of support that is allowed, which essentially sets a maximum proportion of the firm’s investment that can be subsidised by the member state government. This is referred to as Net Grant Equivalent (NGE) of aid. This is the benefit accruing to the recipient from the grant after payment of taxes on company profits. RSA grants must be entered in the accounts as income and are made subject to tax. Details for calculations of NGEs are available in OJ C74/19 10.03.1998.

Figure 1: Assisted Areas Map prior to August 1st 1993

Notes: The shaded areas are those which are eligible for some Regional Selective Assistance. The dark shaded areas are the very deprived areas eligible for an investment subsidy of up to 30% NGE (Net Grant Equivalent). The light shaded areas are eligible for up to 20% NGE.

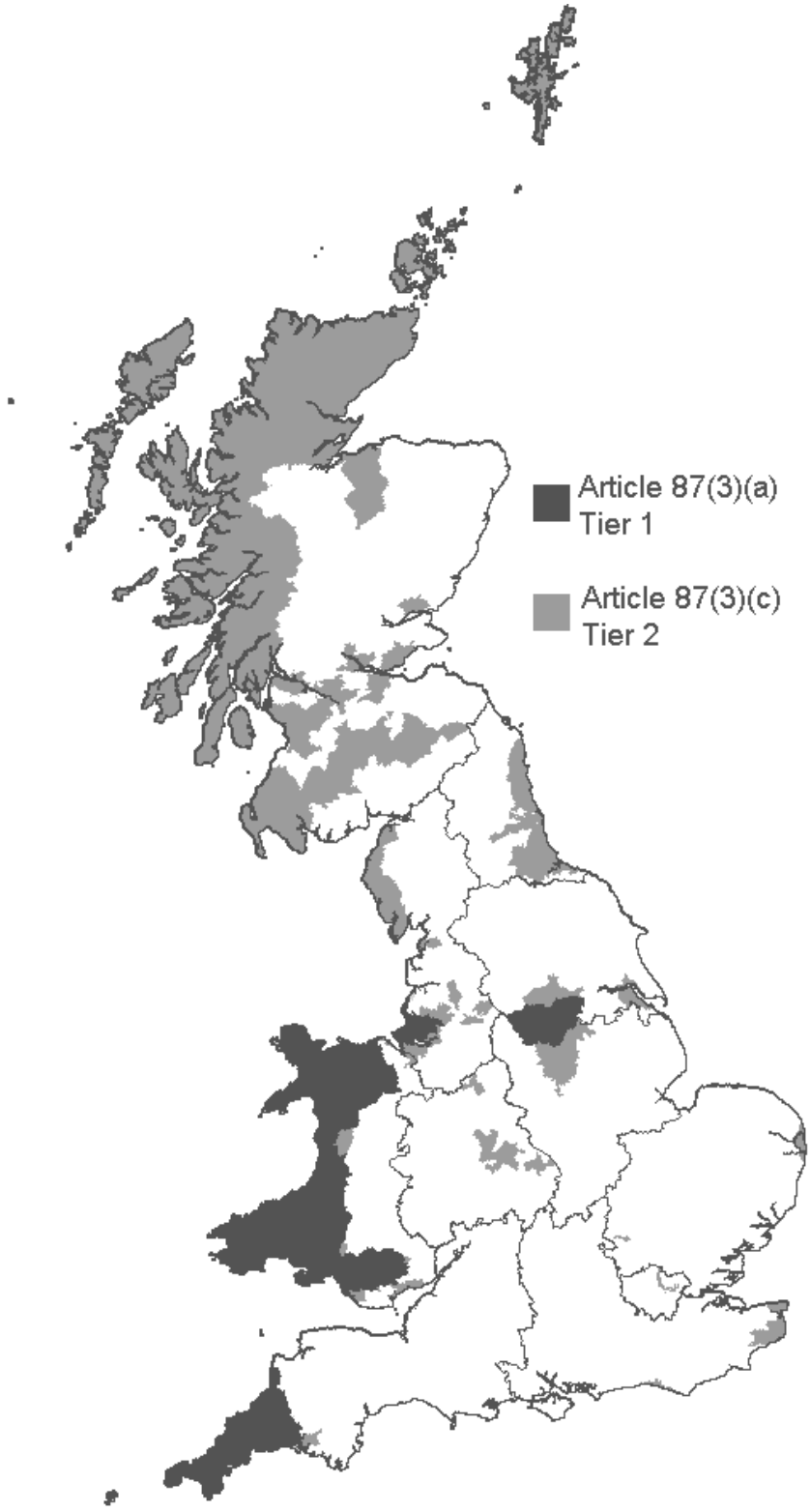
Source: Department of Trade and Industry

Figure 2: Assisted Areas Map after August 1st 1993 and prior to January 1st 2000



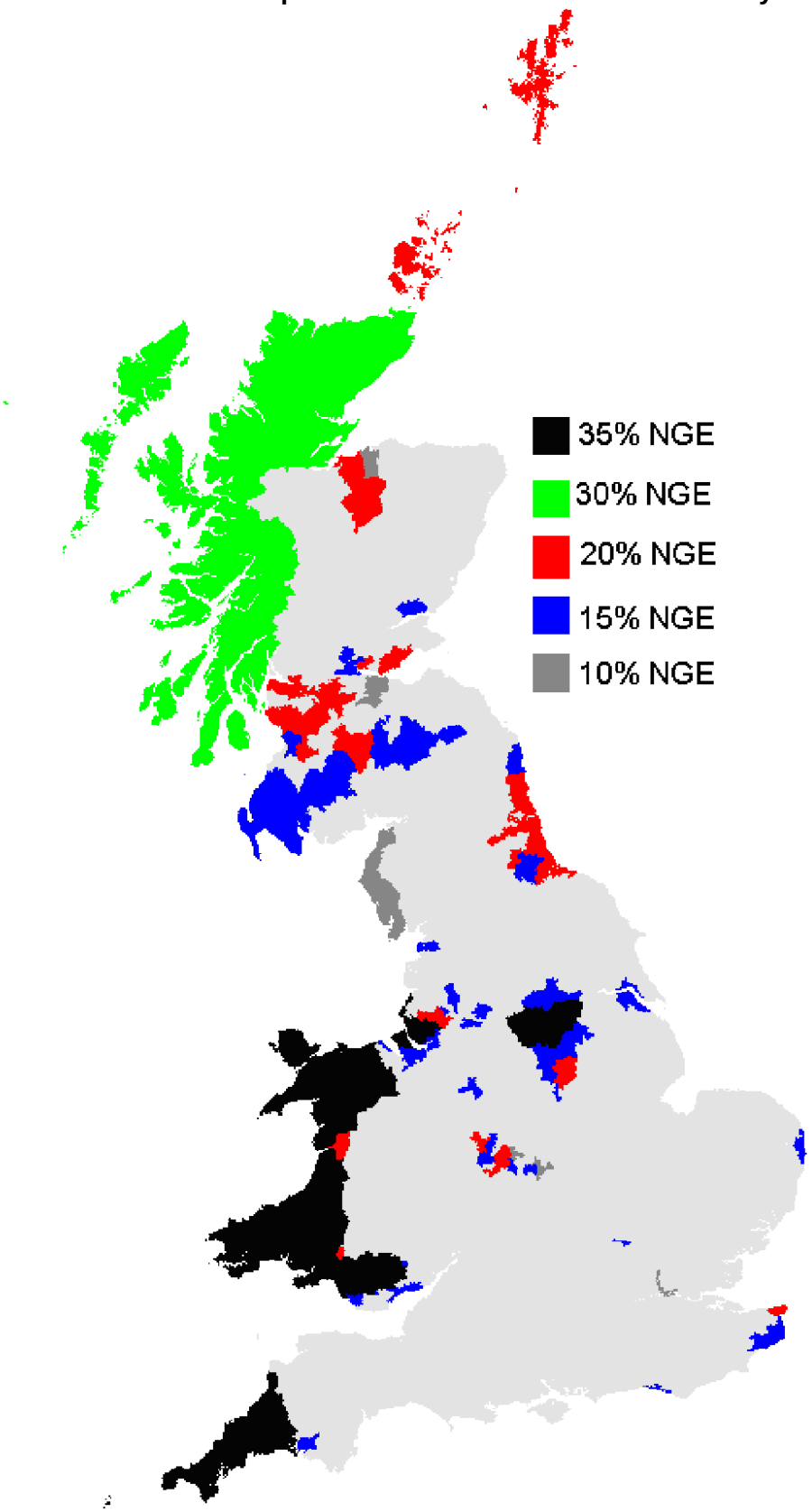
Notes: The shaded areas are those which are eligible for some Regional Selective Assistance. The dark shaded areas are the very deprived areas eligible for an investment subsidy of up to 30% NGE (Net Grant Equivalence). The light shaded areas are eligible for up to 20% NGE. **Source:** Department of Trade and Industry

Figure 3: Assisted Areas Map after January 1st 2000



Notes: The shaded areas are those which are eligible for some Regional Selective Assistance.
Source: Department of Trade and Industry

Figure 4: Assisted Areas Map with detailed NGE rates after January 1st 2000



Source: Department of Trade and Industry
Notes: This shows all the different levels of NGE by area

Formal criteria for receipt of RSA

RSA has been available to manufacturing sector firms and also to services sector firms that served national or foreign markets (i.e. not just the local market). The grants were discretionary and firms could only apply if the supported project satisfied the following criteria. (a) **Location**: The project had to be undertaken within one of the Assisted Areas. (b) **Investment:** The project had to involve capital expenditure on property, plant or machinery; (c) **Jobs:** The project should normally have been expected to lead to the creation of new employment or directly protect some or all of the jobs of existing workers which, without the project, would otherwise have been lost; (d) **Viability:** The project should have had good prospects of viability and should have been expected to help the business become more competitive; (e) **Need:** The applicant had to demonstrate that assistance was necessary to enable the project to proceed as envisaged in terms of nature, scale, timing or location;[5] (f) **Prior Commitments:** As RSA could only be offered where it would make the difference between the project going ahead and not proceeding, there should have been no prior commitment to the project, i.e. the Welsh Assembly Government[6] must have completed its appraisal of the project and issued a formal offer of assistance before the applicant entered into a commitment to proceed with the project; (g) **Other Funding:** The greater part of the funding for the project should have been expected to be met by the applicant or come from other sources in the private sector.[7]

The RSA application process

The process for application is as follows. Interested firms can contact the Welsh Assembly Government (WAG) by post, telephone (and more recently email) with a summary for the project proposal. Within few days (usually three) working days the WAG would let them know whether or not a project appears to meet the general eligibility conditions of the scheme. If it does, a case officer would be allocated to deal with the project who may arrange a meeting or visit to discuss the proposal in more detail. If the project seems likely to be suitable for a full application, the case officer will provide the firm with the necessary forms and advice on the remaining stages of the process. In the application form, firms need to prove additionality - i.e. that they would not go ahead with the project in the Assisted Area - to provide business plans, accounts and reasons for wanting the grant. They then submit this to the Welsh Assembly Government (and before that the National Assembly for Wales and previous to that the Welsh Office). The lag between the day the application is submitted and the decision depends on the amount that the firm is applying for.

Successful applicants will receive an offer letter containing the terms and conditions. The grant is usually paid in instalments in line with project capital expenditure and employment targets. The terms will have been discussed with the firm and will be set out clearly in the “Offer Letter”. Once the firm has formally accepted the offer, WAG will continue monitoring the progress of the project, paying the amounts due in instalments and dealing with any changes that may occur during the life of the

[5] This may be to meet a funding gap, to reduce the risks associated with the project, or to influence the choice of location of a mobile project. It might also be to obtain parent company approval by meeting established investment criteria; or for some other acceptable reason – each case is considered on its own merits.

[6] Previously it was the National Assembly for Wales and before that the Welsh Office.

[7] These may include bank borrowings, hire purchase or lease finance, equity and loan finance from existing or new shareholders and loans from other organisations or institutions. Additional public sector assistance may however, be available towards the project. Any additional assistance must be cumulated with the RSA support and must not breach the European Union State Aid limits.

project. Finally, after payment of the final instalment of the grant, the project enters its Post Completion Monitoring period until all the obligations of the offer are met. Grants may be recovered in full or in part if all the conditions are not met.

Up until recently, the selection criteria did not include the productivity impact or quality of the jobs. Since 2000, however, projects at the higher end of the productivity distribution, with large R&D investments and better paid jobs have been preferred.

There is great heterogeneity in the size of RSA grants (see also below). The upper limit for the proportion of grant to total capital expenditure on the project varies by Assisted Area. There are two different Assisted Areas (Development Areas and Intermediate Areas and from 2000 Tier 1 and Tier 2) that differ in the degree to which they match with EU Objectives and in which percentage of the projects funded RSA grants cover.

The data

We combine administrative data on support scheme participants[8] with independent business performance data. This involves matching Selective Assistance Management Information System (SAMIS) database of participants, the Interdepartmental Business Register (IDBR) and the Annual Respondents Database (ARD) which we describe in more detail below.

The Selective Assistance Management Information System (SAMIS) database

The Selective Assistance Management Information System (SAMIS)[9] was used to monitor RSA projects. It contains information on 54,322 applications from 1972 to 2003. It includes for all applications information on the name, date and address of the applicant, a project description; the amount applied for, aims and date of application. For successfully completed applications it provides the date in which and the amount of the grant offered and paid (since 1989 additional payment information is available containing date and amount of first and last instalments). For those that were not completed it contains information on why; i.e. whether the project was withdrawn; was accepted but then the firm did not proceed; was not accepted by the firm; or was rejected by Welsh Assembly Government (and before that the National Assembly for Wales and previous to that the Welsh Office) and if so for which reason.

Since the payment information from the SAMIS database is not always accurate, DTI provided us with additional information with more detailed payment information from the Payment RSA database but only for applications after 1988. Therefore in any econometric evaluation we would prefer to restrict our analysis to the period 1989-2003 to minimize measurement error problems deriving from using inaccurate payment information on values and dates. However, in this feasibility study we will report as much as possible on the whole 1972-2003 period. To do so, we will use information on the amounts last offered to the applicants jointly with information on

[8] As described in more detail below we also have information on applicants to the scheme that were rejected for various reasons or had withdrawn their application.

[9] Note that as of 2004 the Welsh Assembly Government also maintains its own database of RSA support (rather than the SAMIS database). However, for the period analysed the SAMIS database covers the RSA grants in Wales.

whether firms proceeded with the project to inform the reader on payment information. Note that the main rationale to do so is that when both sources of information are available; total amount paid to the firm and latest amount offer to the firm are highly correlated (with a correlation coefficient of about 0.8).

In the SAMIS database, there is information on 54,322 program applications from across the United Kingdom, whether the application has been successful or not, from 1972 to 2003. Of these 54,322 observations, 12% are in Wales. We will restrict our analysis to these 6,540 applications.

As mentioned above the SAMIS database contains information on both "successful" and "unsuccessful applications". Therefore, our first task is to distinguish between these two types of applications since in the subsequent econometric analysis we will restrict our attention to "treated" businesses; i.e. "successful applications".

Table 1 reports for each year the total number of applications and the number of applicants which were not offered a grant. For applications that were successful we also report the distribution of latest amount offered (note that the amounts here are in nominal terms). From the Table two main features emerge:

- In each year a small proportion of applicants are offered no grant as shown in column 3; for the whole of the sample period this amounts to a share of 20% (1,340/6,540).
- In each year the distribution of grants offered is very skewed: most of the offers are small with a few very large grants. For example in 1972; the 10^{th} percentile

offer was of £5,000 and the median successful applicant received £26,000 and the 75th percentile is still well below £100,000 (£88,000). However, the 90th percentile is £432,000 and the 99th percentile is more than double at 900,000; i.e. more than 30 times the median value. The skewness of the distribution is also evident when we compare the median offer with the average grant offered which, at 125,670, is almost 5 times larger.

Table 1: Payment information in the RSA database for Wales 1972-2003

				final amount offered (£000s)							
Application year	applications	zero offer	positive offer	p10	p25	p50	p75	p90	p99	mean	Standard deviation
1972	43	17		5	15	26	88	432	900	125.67	222.57
1973	189	86		2	7	22	73	288	797	87.67	168.96
1974	164	63		3	6	21	60	124	420	56.63	94.89
1975	148	53		3	5	18	56	200	1,950	96.91	273.75
1976	117	35		4	7	16	53	150	440	52.82	86.44
1977	164	32		2	6	17	51	100	473	45.73	86.64
1978	212	50		5	11	24	79	170	1,035	86.00	261.14
1979	206	41		5	12	25	66	206	1,400	100.30	232.62
1980	111	29		9	15	35	100	250	5,000	214.75	747.44
1981	177	26		10	16	35	100	400	1,250	129.15	249.92
1982	236	50		10	15	43	100	312	1,270	117.63	212.60
1983	307	41		10	20	50	154	500	2,400	197.72	416.02
1984	349	71		12	25	52	225	876	4,200	313.21	709.20
1985	155	45		15	28	58	180	438	2,200	176.00	375.95
1986	193	33		15	29	75	250	585	3,500	228.82	469.39

1987	218	65	15	30	70	300	750	4,800	326.09	750.71
1988	241	52	14	25	50	230	750	4,800	375.16	1,641.17
1989	283	79	10	25	60	200	600	3,500	320.39	1,528.36
1990	260	59	13	25	60	239	750	3,000	413.07	2,566.95
1991	265	39	13	25	63	300	1,000	3,500	385.96	939.75
1992	228	34	15	40	98	425	1,000	5,000	424.45	887.26
1993	309	96	15	40	90	350	1,000	5,000	453.02	1,402.79
1994	193	18	13	25	70	300	990	3,000	331.03	664.14
1995	204	22	15	25	68	200	900	7,000	389.34	1,124.98
1996	215	26	16	30	90	350	1,300	7,000	800.45	5,112.64
1997	184	16	15	25	80	248	1,000	4,750	404.29	845.62
1998	135	13	24	33	93	250	800	4,000	397.53	876.72
1999	190	28	20	25	75	240	800	5,600	355.29	878.74
2000	158	12	20	48	132	400	1,600	16,000	754.46	2,285.11
2001	282	18	25	52	120	248	880	7,500	462.32	1,210.99
2002	276	32	60	100	180	250	980	10,400	585.79	1,608.24
2003	128	59	78	100	150	250	954	1,750	310.05	389.76
Total	6,540	1,340	10	24	61	200	718	4,163	328.56	1,454.61

Table 2 reports the main reasons for rejections wherever this information is available. As evident by comparing the different rows the lack of two necessary requirements for eligibility i.e. additionality and financial viability are the two main reasons for rejection of applications.

Table 2: Rejection reasons Wales (1972-2003)

Rejection Reason	**Number of applications**
Additionality (Including Timing Problem etc.)	32
Displacement; Disqualified - wrong info provided (e.g. Capital Expenditure/Ineligible activity etc); Efficiency; Employment Link not established	13
Financial Viability	35
Insufficient information provided	10
Other Viability (Technical Feasibility/Management)	11
Total	101

Interdepartmental Business Register (IDBR)

In order to be able to match the administrative information with production data the records from the SAMIS database needed to be matched with the Interdepartmental Business Register (IDBR), which contains both the names of the businesses and the identification numbers used by the Office for National Statistics to conduct the Annual Business Inquiry. The Interdepartmental Business Register (IDBR)10 is essentially a list of all businesses in the UK, their addresses, type of activity and ownership/control structure compiled using a combination of tax records on VAT and PAYE, information lodged at Companies House, Dun and Bradstreet data, and data from other surveys using three aggregation categories: "local units" (plants), "enterprises" and "enterprise groups".[11] A plant or "local unit" is defined as "an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place" and is identified by a unique identifier. A major advantage of the IDBR is that information is available at many disaggregated levels. For our analysis this is particularly useful since we also look at the effect of the policy at the regional level on employment and entry/exit. We therefore need employment and entry/exit information at the local unit level rather than at the enterprise level since enterprises can consist of local units in different regions.[12]

A stratified[13] random sample of enterprises is drawn every year from the IDBR to form the sampling frame for the Annual Business Inquiry (ABI), which provides information on employment, investment, materials, etc. and is described next.

Annual Respondents Database (ARD)

The Annual Respondents Database (ARD)[14] is the UK equivalent of the US Longitudinal Respondents Database and is made available by the Office for National Statistics (ONS) based on information from the Annual Business Inquiry (ABI),[15] the mandatory annual survey of UK businesses. The ARD unit of observation is defined by the ONS as an autonomous business unit (also referred to as "reporting units").

---

[10] The IDBR was introduced between 1994 and 1995. Previously, that sampling was on the basis of a Business Register maintained by the Office of National Statistics (the UK Census).

[11] Criscuolo et al (2003) report that in the 1998 IDBR the vast majority of enterprise groups and Reporting Units consist of just one local unit (92%, 149,326 out of 162,477 and 93%, 158,727 out of 171,271 respectively).

[12] Employment information on the IDBR comes from PAYE data if that is the source of the original inclusion and the enterprises operate a PAYE scheme, which in turn if operated at the local unit level, provides independent local unit employment data. Also the IDBR gathers and updates information on employment from the Annual Register Inquiry (see Criscuolo et al., 2003 for details) and the Annual Business Inquiry (ABI). However, employment data is required to construct sampling frames and hence it will be interpolated from turnover data. The IDBR turnover information comes from VAT records if the original source of business information was VAT data; however this information is quite limited as it is only available for single-local unit enterprises that are large enough to pay VAT (the threshold was £52,000 in 2000–01) at both the enterprise and local unit level. For multi-local unit enterprises, no turnover information will be available for local units, since most multi-local unit enterprises do not pay VAT at the local unit level.

13 Stratification is broadly based on industry affiliation; regional location and size. For details see Criscuolo et al. (2003).

[14] More extensive description of the ARD can be found in Criscuolo, Haskel and Martin (2003), Griffith (1999) and Oulton (1997).

[15] Called the Annual Census of Production until 1998.

Some of these business units are spread across several sites but in about eighty percent of all cases a business unit is located entirely at a single mailing address. We call this unit a “firm”.
It is important to note that the ARD does not consist of the complete population of all UK businesses, since the sample is stratified with smaller businesses sampled randomly. It contains the population of larger businesses however (those over than 100 or 250 employees depending on the exact year). Each year the sampled firms account for around 90% of total UK manufacturing employment. The ARD contains a wealth of information, but most importantly for our study it contains information on employment, investment, intermediate inputs and gross output. We are particularly interested in the effect of RSA for employment, investment and productivity.

Matching the SAMIS database with the ARD

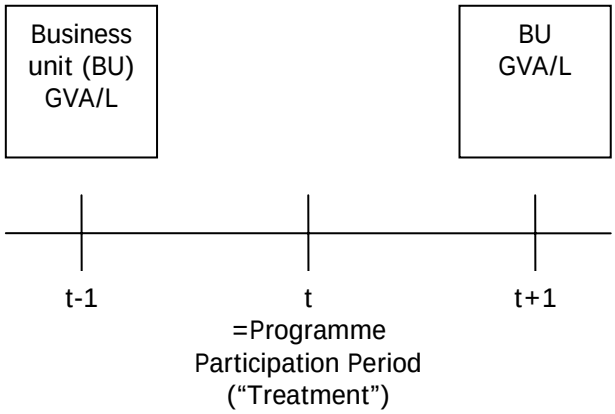
To understand the key issues in matching it is easiest to work backwards; firstly, considering what data we require for matched participants, then considering the problems that we face in identifying those participants.

Data requirements

In order to assess the impact of RSA we require information on the relevant outcome variable for the business unit before and after they receive RSA. Figure 5 represents this graphically. The business unit (BU) receives support at time t.[16] If we observe the information on the relevant outcome variable – e.g. value added per employee – at some point before the participation in the support programme (e.g. t-1) and at some point after (e.g. t+1) then we can assess the impact of the programme on that firm. Of course, to perform an econometric analysis and get results that are representative of all firms we need to be able to do this for a sufficiently large number of programme participants and, for a group of non-participants (i.e. a control group).

Figure 5: Data requirements

[16] For the sake of the example assume that the support programme is completed after one period.



Matching[17] problems

When, as in our case, the performance data comes from some data source un-related to programme participation several problems arise. Firstly, if the database with performance data - as is the case with the ARD – is a survey with stratified random sampling (i.e. not a census) then firms that are programme participants, or are in our preferred control group, simply might not be sampled in years before and after the intervention.

Secondly, there might be discrepancies between the ARD "population", i.e. the sample of reporting units that can be sampled to receive the ABI survey (and thus be in the ARD) and the IDBR population. This might happen because the ABI population is a subset of the IDBR: enterprises on the IDBR are excluded from the ABI population on the grounds of economic activity (e.g. parts of finance, agriculture etc.), legal form, and quality (around 200,000 small units based solely on PAYE registrations are excluded due to the risk of duplication); as are survey specific reporting structures for surveys other than ABI.

Thirdly, the performance database might not hold data on exactly the business unit we require or it might not be possible to determine which of several related business units in the performance database are affected by a particular support scheme. It is these matching problems to which we now turn.

For any dataset it is useful to keep in mind the unit of observation (i.e. the unit that defines a row in the dataset). The ARD unit of observation is referred to as a "reporting unit" and is generally[18] an enterprise; defined as an independent businesses unit. The ARD surveys reporting units (RUs) and provides data on these units on an annual basis. The ARD also provides information on "local units" that are part of that reporting unit. The SAMIS databases on the other hand reports on "incidences" of programme participation. The key difficulty is then to figure out which ARD reporting unit or local unit has potentially been affected by a particular incidence of programme participation.

[17] Note the two senses in which we are using the term "matching". First is the mundane sense of merging the DTI administrative data with the ARD. Second is the econometric sense of matching which involves the construction of a valid control group matched to the treatment group.
[18] There are some exceptions to this rule.

The steps involved for that purpose are as follows:

- **Matching SAMIS with IDBR**: DTI uses name and postcode from its administrative data to match a list of participants (and possibly applicants) to the Inter-Departmental Business Register (IDBR). This matching may occur at the local unit, reporting unit, enterprise, PAYE unit and sometimes Company's House Reference Numbers (CRN) levels.

- **Matching SAMIS-IDBR with ARD:** We exploit the match at each level to get a pointer from the matched SAMIS-IDBR administrative record to the ARD RUs that potentially could have been affected by RSA. There are three possible outcomes: (i) we cannot map the support to any reporting unit in the ARD; (ii) we map the support uniquely to one reporting unit in the ARD (iii) We map the support to multiple ARD RUs. Figure 6 illustrates this graphically.

- **Data pre- and post- participation:** For each ARD reporting unit that is matched to a record of RSA support, we then examine if we have sufficient pre- and post- programme participation performance data to do econometric evaluation analysis.

This raises a number of issues. Consider a binary treatment (i.e. simply a one for getting an incident or a zero for not getting it)[19]. Firstly, as Figure 6 illustrates, even if we have a unique match between an observation in the administrative dataset and an ARD reporting unit that unit might report for several local units and only a subset of them may actually be affected by the programme. Depending on whether RSA is used to create or safeguard employment, the fact that it only affects a subset of local units may lead us to overstate or understate the programme impact if we do not control for this in the evaluation. Similarly, for RSA records matched with multiple ARD RUs we might have the problem that not all ARD RUs are affected by the programme.

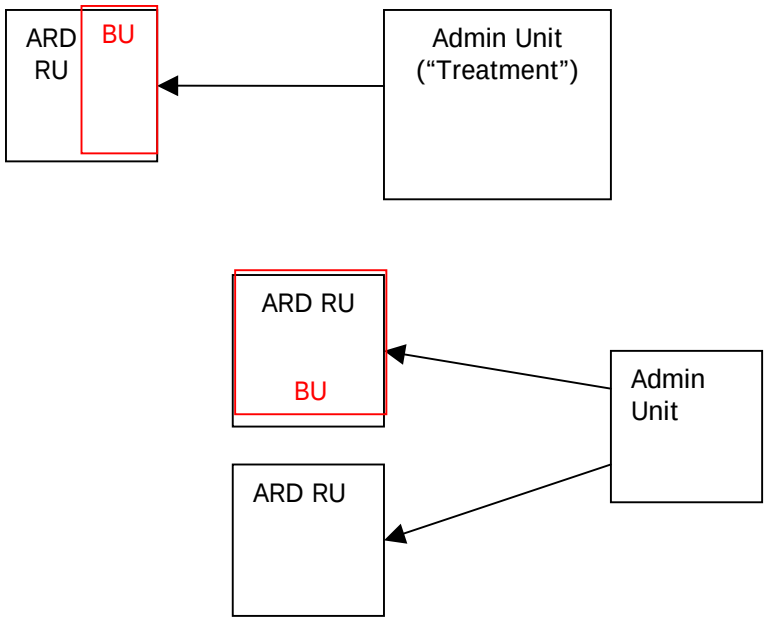
The fundamental problem with these issues is that we generally do not know which case applies, which makes it hard to control for it in an econometric model. RSA by definition should only apply to local units at a specific location within Assisted Areas. However, this might not be clear-cut. For example if the head office applies for the RSA on behalf of one of its local units then the administrative dataset would hold the postcode of the head office and/or the postcode of the the postcode of the local units where the project takes palce. As a consequence we would associate the administrative record with the head office local unit instead of (or in addition to) the programme participating local unit.

The SAMIS database holds some basic information on the characteristics of the business unit participating in the programme. With this information we might be able to make some progress on these issues. For example, if we know the employment of the participating business unit we can use this as an additional matching variable to distinguish between several suggested ARD RUs; secondly we know for each of the units matched to the IDBR whether this are single plant that are not part of a larger group. If this is the case no ambiguity arises.

[19] These problems are less severe if the treatment is continuous and measured in cash terms. For example, if a firm received an RSA grant of £100,000 then this is equivalent to two grants of £50,000. A binary treatment is not so obvious to aggregate.

Other than this possibility, the key strategy to handle these problems is to be aware of them and examine the robustness of any subsequent econometric findings to different ways of controlling for them. This can be done for example by looking at differences in evaluation results when using only a sample of RUs with unique match to a treatment unit or by aggregating over several RUs in the case of multiple matches.

Figure 6: Matching unique and multiple matches



Matching results

Matching with IDBR

As shown in Table 3, using name, postcode and CRN numbers, DTI has managed to map the information in the SAMIS database to the IDBR for 4,434 RSA applications; i.e. 66% of the cases.

Table 3: the RSA database and the success rate of IDBR matching

Year	Applications	Matching with IDBR %
1972	43	30.2
1973	189	37.0
1974	164	42.7
1975	148	39.2
1976	117	40.2
1977	164	40.2
1978	212	50.9
1979	206	52.4
1980	111	43.2
1981	177	40.1
1982	236	44.1
1983	307	45.6
1984	349	55.3
1985	155	63.9
1986	193	60.6
1987	218	69.7
1988	241	69.3
1989	283	65.4
1990	260	70.8
1991	265	77.4
1992	228	82.9
1993	309	84.5
1994	193	88.6
1995	204	89.2
1996	215	90.2
1997	184	91.8
1998	135	94.1
1999	190	88.9
2000	158	88.0
2001	282	89.7
2002	276	93.5
2003	128	91.4
Total	**6,540**	**66**

The table shows the number of applications and the percentage of successful matches with the IDBR for each year from 1972 to 2003. The earlier period, between 1972 and 1986 has lower rates of successful matches with the IDBR, although increasing steadily over the years. For the period 1987-2002, the rate of successful matches increases from 65% to an average level of more than 80% after 1993. Given this, it might be preferable to focus only on the latter period in the second stage of the

analysis. The better matching rate in the second period reflects the fact that the IDBR was introduced in 1994 and is not likely to contain information for "units" that have closed down before 1993. As the first step of matching to the ARD is a match to the IDBR this will bias us towards matching grants pre 1993 to firms that have survived at least until 1994.

When matching to the IDBR, there are two main issues:

- One "IDBR unit" might have applied for and received several grants.
- Name and address identifiers from the SAMIS database for one project are sometimes attributed to several IDBR units;

Our strategy to deal with this in the econometric analysis is as follows: we create a basic sample for analysis where we count all reporting units in SAMIS that have received at least one RSA grant as treated units. If a RU has several RSA treatments we will aggregate over these to get total subsidy amounts. To examine robustness of results to these issues we will also report results for only the sample of RUs that had unique matches at this stage. For that purpose it is important that this restricted sample is large enough. We therefore examine the relevance of each issue in turn.

Table 4 shows the number of applications by firms that have an IDBR match and how many of these applications have been successful. The table shows that most reporting units (RUs) in the data have only applied to (67%) and received (53.2%) RSA once.

Table 4: Number of multiple applications and awards

	Number of applications by the same RU	Number of grants awarded to the same RU
0		813
1	2,209	1,757
2	586	440
3	260	156
4	120	73
5 or more	129	65
Total	3,304	3,304

Notes: e.g. there are 2209 reporting units that have applied only once, 586 have applied 2 times. From column 2 we see that 1757 reporting have received 1 grant, 440 have received 2.

Table 5 reports on multiple matches of the same SAMIS record to different RUs. Almost 67% of applications are matched to only one IDBR record; another 20% are matched to two different IDBR units and the other 13% to three or more IDBR units.

Table 5: Number of Multiple reporting units per treatment

Match to IDBR	Number	Proportion
1	3,682	66.79
2	1,078	19.55
3	417	7.56
4	168	3.05
5 and more	168	3.05
Total	5,513	100

Matching with ARD Sample

To undertake analysis of outcomes such as employment, investment or productivity we have to rely on the ARD sample. The tables below present summary results of the matching to the ARD sample. The aim is to show whether there is sufficient information in the matched RSA-IDBR -ARD sample to undertake further evaluation.

Table 6: Matches of SAMIS-IDBR with ARD

	All applications		Only successful applications	
	1	2	3	4
Year	match with IDBR	match with ARD	match with IDBR	match with ARD
1972	13	10	<10	<10
1973	70	41	34	25
1974	70	41	28	19
1975	58	39	28	19
1976	47	28	25	16
1977	66	53	45	37
1978	108	76	60	42
1979	108	74	63	45
1980	48	30	22	13
1981	71	38	44	24
1982	104	64	64	41
1983	140	90	96	64
1984	193	141	106	80
1985	99	74	51	39
1986	117	80	86	58
1987	152	99	96	69
1988	167	102	107	68
1989	185	100	104	62
1990	184	110	120	79
1991	205	125	144	101
1992	189	123	146	100
1993	261	182	155	120
1994	171	106	135	87
1995	182	110	138	90
1996	194	131	150	108
1997	169	120	135	99
1998	127	80	89	54
1999	169	110	123	85
2000	139	86	119	77
2001	253	138	218	119
2002	258	150	216	130
2003	117	66	65	41

Table 6 reports the number of successful matches with the IDBR (column 1) and the ARD (column 2) for both recipients of RSA and unsuccessful applicants. In column 3 and 4 we report the number of matches with the IDBR and ARD but only for successful and accepted applications that receive RSA, respectively.

Table 7 examines to what extent there is sufficient information in the ARD sample to undertake various kinds of econometric evaluation exercises. In particular we want to find out if there are sufficiently many firms for which we have valid data before and after an RSA support spell. The first row, labelled "ALL" refers to all ARD sample RUs that could be matched with at least one RSA record, irrespective of whether this record is linked to multiple or only one ARD RU. We examine to what extent there is sample information before and after RSA (i.e. the treatment in this case, as illustrated in Figure 5). When a RU has received several RSA grants we only consider the earliest of these, which is a useful way to assess if the minimal requirements for econometric evaluation are met. Clearly it might be interesting to find out if the impact of several RSA is different from only one. However, this would imply more stringent requirements on the data as we would now require at least one valid observation before and after *each* treatment.

Table 7: Matches RSA-IDBR with ARD pre- and post-treatment

	before treatment	after	in treat year	1 year after	2 years after	3 years after	4 years after
All	268	1012	235	224	216	192	180
Growth	159	735	140	134	129	115	110

Returning to Table 7, in row 1 we give the number of RUs where we have data on productivity levels (value added per employee) in particular years relative to the year in which the RSA treatment took place. Column 1 shows that there are 268 ARD RUs that are associated with at least one RSA grant and have at least one valid observation on productivity levels in a time window of 5 years before the firm receiving the grant. Column 2 shows that there are 1012 ARD RUs with at least one RSA grant and that report at least one productivity measurement in any year after the grant. For econometric evaluation we need *both*: i.e. an observation before and after the grant. We consider this in the remaining columns. Column 3 shows that there are 235 cases with at least one observation before and at least one observation after the evaluation. In terms of assessing if the minimal requirements for econometric evaluations are met, column 3 is the most interesting one as it reports the largest sample of participants we can get under the most favourable assumption regarding the speed of impact (i.e. that it begins to occur in the same year as the RSA grant). The remaining columns address the concern that any impact of RSA might take time to materialise by giving the number of RUs that have one (or more) measures of productivity at least one year after the benchmark (column 4); at least two years after the benchmark (column 5) etc. By definition the numbers of RUs decline as we move forward through the columns, but the numbers appear sufficiently large to consider some form of econometric evaluation. The next row examines the same statistics for productivity growth. We now need data in at least two years before the treatment and two years after (to calculate changes in productivity). Clearly the numbers of useable firms fall and, indeed, they almost halve compared to row 1. However the samples still seem large enough for an evaluation analysis.

Table 8: Matches RSA-IDBR with ARD pre- and post-treatment for different size-classes

Employment	Before treatment	after	in treat year	1 year after	2 years after	3 years after	4 years after
less than 50	51	340	34	29	27	26	24
50 to 99	53	204	45	40	39	37	32
100 to 249	70	243	69	68	66	54	52
250 and more	94	225	87	87	84	75	72

Table 8 repeats the same exercise for various sub samples of RUs according to their size. We have divided the sample in 4 size classes. The first row reports matching rates for small firms, i.e. with less than 50 employees; the second row for firms with less than 100 employees; the third for firms with less than 250 employees and the last for large firms with 250 employees or more. From comparing columns 3 to 7 in the table across rows we can see that the number of successful matches increases if we look at larger firms. This reflects the fat that the ARD is a random sample of firms stratified by size with only larger firms being sampled every year.

Table 9: Matches RSA-IDBR with ARD pre- and post-treatment for different sector

Sector	Before treatment	after	in treat year	1 year After	2 years After	3 years after	4 years after
Extraction of crude petroleum and natural gas	<10	<10	<10	<10	<10	<10	<10
Other Mining and Quarrying	<10	<10	<10	<10	<10	<10	<10
Manufacture of food products and beverages	33	83	30	29	26	24	24
Manufacture of tobacco products	<10	<10	<10	<10	<10	<10	<10
Manufacture of textiles	<10	25	<10	<10	<10	<10	<10
Manufacture of wearing apparel	<10	16	<10	<10	<10	<10	<10
Manufacturing, tanning and dressing of leather	<10	<10	<10	<10	<10	<10	<10
Manufacture of wood and cork	<10	27	<10	<10	<10	<10	<10
Manufacture of pulp, paper and paper products	10	40	<10	<10	<10	<10	<10
Publishing and printing	15	45	12	12	12	12	10
Manufacture of coke, petroleum and nuclear fuel	<10	<10	<10	<10	<10	<10	<10
Manufacture of chemicals and chemical products	27	79	25	24	24	20	19
Manufacture of rubber and plastic products	26	104	25	22	21	17	17
Manufacture of other non-metallic mineral products	10	35	<10	<10	<10	<10	<10
Manufacture of basic metals	11	45	11	11	11	<10	<10
Manufacture of fabricated metal products	24	121	21	20	19	16	16
Manufacture of machinery and equipment n.e.c.	25	95	21	20	19	17	17
Manufacture of office machinery and computers	<10	<10	<10	<10	<10	<10	<10
Manufacture of electrical machinery n.e.c	<10	50	<10	<10	<10	<10	<10
Manufacture of radio, television	10	40	<10	<10	<10	<10	<10
Manufacture of precision instruments	12	33	10	10	10	<10	<10
Manufacture of Motor Vehicles, Trailers	13	50	13	13	13	12	11
Manufacture of Other Transport Equipment	<10	17	<10	<10	<10	<10	<10
Manufacture of furniture, manufacturing n.e.c.	17	73	13	11	10	<10	<10
Recycling	<10	<10	<10	<10	<10	<10	<10
Electricity, gas, steam and hot water supply	<10	<10	<10	<10	<10	<10	<10
Construction	<10	<10	<10	<10	<10	<10	<10
Sale, maintenance of motor vehicles	<10	<10	<10	<10	<10	<10	<10
Wholesale trade	<10	11	<10	<10	<10	<10	<10
Retail trade	<10	<10	<10	<10	<10	<10	<10
Hotels and restaurants	<10	<10	<10	<10	<10	<10	<10
Supporting transport activities; travel agents	<10	<10	<10	<10	<10	<10	<10
Real Estate activities	<10	<10	<10	<10	<10	<10	<10
Renting of machinery and equipment	<10	<10	<10	<10	<10	<10	<10
Computer and related activities	<10	<10	<10	<10	<10	<10	<10
Research and Development	<10	<10	<10	<10	<10	<10	<10
Other Business Activities	<10	<10	<10	<10	<10	<10	<10

Finally, Table 9 report successful matches for different sectors. The table shows that looking at 2-digit sectors in most industries we do not have enough observations to conduct econometric analysis. Here the solution therefore would be to aggregate up to 1-digit sectors or intermediate level of aggregation (e.g. high-tech; medium tech and low-tech)

Conclusions and Strategies for Evaluation of RSA

This report discusses issues related to the matching of administrative records from the SAMIS database to the ARD and describes the econometric approaches and methodologies that can be used to evaluate the impact of RSA in Wales.

The main aim of this study was to investigate the feasibility of an evaluation of the Regional Selective Assistance in Wales and the feasibility of an evaluation also for size classes and specific industries. From our analysis we conclude that we can conduct a very rich set of analyses, from simple OLS to instrumental variables also for different size classes and for quite aggregated manufacturing sectors.

Our conclusion is therefore that using linked micro data in the evaluation of the impact of the Regional Selective Assistance in Wales is feasible in a possible second stage of the analysis

As we mention in the Appendix in this second stage we would take particular care in overcoming selection bias when estimating treatment effects, by using a suitable instrumental variable that is correlated with participation in the programme but - conditional on programme participation - not with productivity; employment and investment. One possible instrument in the case of RSA is the geographical variation in the extent to which projects can be funded and the changes over time of the map of the Assisted Areas map, as shown in Figures 1 to 4 of this report.

Appendix: Econometric evaluation methods

The objective of econometric evaluation techniques is to estimate the causal impact of a particular programme on an outcome variable such as productivity, employment or/and investment; i.e. we try to answer the question: "what would productivity be like if the programme had not taken place?" We propose to answer this question by examining whether there is an impact at the individual businesses level (micro impact).

Following the literature we define programme participation very broadly as "treatment". Ideally, to evaluate the effect of treatment we would want to observe what would happen to the "treated" firm had it not participated in the programme. However, because this is not observable, we need to find an alternative approach that allows us to evaluate "the treatment effect on the treated", i.e. the effect of the programme on participating firms. In what follows we will describe the assumptions under which different methodologies consistently estimate the "treatment on the treated".

Let us start from the case in which programme participation is completely random. A random allocation of the treatment creates directly comparable treatment and control groups and allows researchers to estimate the treatment effect simply as the difference between the means of the outcome variable in the treatment and control groups.

However, it is very unlikely that programme participation is random. Instead, participation is likely correlated with the expected benefits from the treatment. Since participation is not random, if firms who participate are simply compared with those who did not, the estimates will suffer from selection bias.

In this case quasi-experimental methods need to be used to construct suitable control groups. The idea is to use observed data together with some appropriate identifying assumptions to "construct" the missing counterfactual using control groups: firms to whom the intervention is applied (the treatment group) are matched with an "equivalent" group from which the intervention is withheld and the average value of the outcome indicator for the treatment group is compared with the average of that for the constructed control group.

We will group the various estimators in two broad categories:

The first category includes all the methods that assume that participants and non-participants only differ in terms of observable characteristics which we can control for (what econometricians call "ignorability assumptions" or "selection on the observables"). In this category we include OLS and other regression methods, methods based on propensity score and other matching methods.
The second category includes all the estimators based on the existence of an instrumental variable that helps explain participation to the program but has no direct effect on the outcome (in our case productivity).

Let us start by assuming that it is possible to control for all possible reasons why outcomes might differ between participants and non-participants and that there is a single homogenous effect of the programme on participants. In this case one might try estimating the treatment effect using OLS. Multivariate regression analysis is used to control for observable characteristics that distinguish participants and non-participants. The treatment effect is estimated as the differences in the mean

outcomes of the two groups, participants and non-participants, conditional on the set of variables that cause outcome and participation.

There are several potential sources of bias. First is there may be differences in the effect of the program across different firms. At best we can try and estimate the mean effect rather than the effect on each firm. Secondly, there may be omitted variables. Thirdly, there may be a lack of common support between the treatment and control group as signalled by large differences in the empirical distributions of observables between the two groups of firms. Matching estimators try to solve these sources of bias incurred by OLS, but still rely on the assumptions that there is only selection on observables (so do not deal with the second problem).

The idea of matching estimators is to identify among the non-treated firms a "control" group of firms with similar observable characteristics (common support) to the treatment group. In this case, the difference in the average outcome between the treated and the matched non-treated firms (the control group) consistently estimates the effect of the treatment on the treatment group.

However, we face further problems if both observable and *unobservable* firm characteristics drive participation outcomes. In this case, matching estimates, although more flexible than OLS estimates, are still affected by bias due to these unobservables.

One possible solution to this bias is provided by the "difference-in-difference" estimator. This method compares a treatment and a control group (first difference) before and after the intervention (second difference). Once the mean difference between the "after" and "before" values of the outcome variable for each of the treatment and control groups is calculated, the difference between these two mean differences is calculated. This difference in difference is the estimate of the impact of the program.

The main drawback of this approach is the need to identify the control group: a group that is unaffected by the program but who would have responded identically to changes in the environment as would the treatment group. There are many possible strategies for constructing the control group such as using geography – some areas are eligible for the program (e.g. pilots) and some are not and a strategy is to compare the firms in pilot areas with those in non-pilot areas. A variant of this is to use firms on either side of a boundary. Another possibility is size – firms just below a size threshold are eligible whereas those just above are not eligible. Having multiple possible dimensions enables use to test for the identification assumptions underlying the difference in difference approach.

Formally, this method estimates the effect of the treatment consistently if we assume that the unobservables that affect participation decision and outcome are separable in to (i) an individual-specific effect (constant over time); (ii) a common macroeconomic effect, which is the same across all firms (common trends assumption) and (iii) an idiosyncratic shock that is not correlated with participation and the outcome of interest. Note that matching can be combined with difference in differences. We might select a sub-sample of the non-pilot areas for example that are more closely matched on the observables with the treatment group[20].

[20] See Blundell, Costa-Dias, Meghir, and Van Reenen (2004) "Evaluating the employment effects of a mandatory job search assistance programme" *Journal of the European Economic Association*, for an extensive discussion of this in the context of evaluating labour market programs.

If we are not willing to accept the assumption of selection on observables nor assumptions (i)-(iii) on unobservables, we must turn to our second category of estimators which relies on the use of Instrumental Variables (IV). Instrumental variables are chosen so that they determine program participation, but do not affect outcomes given participation. This identifies the exogenous variation in outcomes attributable to the program, acknowledging that the treatment may be non-random. The "instrumental variables" are first used to predict program participation; then one sees how the outcome indicator varies with the predicted values. Instruments might be constructed on the basis of variations in programme availability in different geographical areas or over time. However, in general, finding a suitable instrument is not that easy since it must satisfy the criteria of being correlated with the participation choice while being correctly excluded from the productivity equation.[21]

In the results sections of this report we are discussing for each of the programmes which instruments might be available or if only the more basic evaluation techniques can be applied in a possible stage 2 of the project.

[21] The control function approach is similar in nature to the IV estimator. We do not discuss it here for brevity.

References

Angrist, Joshua D. (2004) “Treatment effect heterogeneity in theory and practice” Economic Journal, vol. 114(494), pp. C52-C83.

Armstrong, Harvey W. (2001) “Regional Selective Assistance: Is the Spend Enough and Is It Targeting the Right Places?” Regional Studies, vol. 35(3) May, pp. 247-57.

Blundell, R.; Monica Costa Dias; Costas Meghir and John Van Reenen (2004) “Evaluating the Employment Impact of a Mandatory Job Search Program” Journal of the European Economic Association, vol. 2(4), pp. 569-606.

Criscuolo C., Haskel J. and Martin R., (2003): “Building the evidence base for productivity policy using business data linking”, *Economic Trends* 600 November 2003, pp.39-51. www.statistics.gov.uk/articles/economic_trends/ETNov03Haskel.pdf.

Department of Trade and Industry (2001) “DEPARTMENT OF TRADE AND INDUSTRY (2001). The government’s expenditure plans, 1999–2000, the regions, p.41

Griffith Rachel S. (1999) “Using the ARD establishment level data to look at foreign ownership and productivity in the UK.” Economic Journal vol. 109, pp F416-F442.

Heckman, James J.; Robert, J. LaLonde and Jeffrey A. Smith (1999) “The Economics and Econometrics of Active Labor Market Programs” in Handbook of Labor Economics, Volume 3 O. Ashenfelter and D. Card eds., pp. 1865-2097.

Jaffe, A. B. (2002) “Building programme evaluation into the design of public research programmes”, Oxford Review of Economic Policy, Vol.18, 1, Spring, 22-34.

Jones, Jonathan and Colin Wren (2004) “Do Inward Investors Achieve their Job Targets?” Oxford Bulletin of Economics and Statistics, vol. 66(4), pp. 483-514.

Jones, Jonathan and Colin Wren (2004) “Inward Foreign Direct Investment and Employment: A Project-Based Analysis in North-East England” Journal of Economic Geography, vol. 5(4), pp. 517-44.

Oulton Nicholas (1997) The ABI Respondents Database: a new resource for industrial economics research. Economic Trends vol. 528, pp. 46–57.

Perry, J. and Vale, S. (2003) Dissemination of Information on Business Registers in the UK, paper prepared for the Joint UNECE/Eurostat Seminar on Business Registers, Luxembourg June 2003.

Roome, Marjorie (2005) “Linking DTI Business Support Schemes data to the IDBR and ARD with implications for the future use of the administrative data in scheme assessment”, DTI report

Wren, Colin (2005) “Regional Grants: Are They Worth It?” Fiscal Studies, vol 26(2), pp. 245-75.

Wren, Colin (2003) “Investment Scale as a Signal in Industrial Assistance Schemes with Employment Objectives” Economica, vol. 70, pp. 331-52.

Wren, Colin and David Storey (2002). “Evaluating the Effect of ‘Soft’ Business Support upon Small Firm Performance” Oxford Economic Papers, vol. 54, pp. 334-65.

Wren, Colin and J. Taylor. (1999) “Industrial Restructuring and Regional Policy” Oxford Economic Papers, vol. 51, pp. 487-516.