7. Methods and quality issues in analysing complex and localized services

Most public services are locally delivered, because they fundamentally require personal contact with customers or clients, or need to be provided in places that are spatially accessible by customers. Teachers normally have to be in the same room with a class of children in order to promote education. Health professionals currently cannot do much still to diagnose or treat patients in their absence. And police forces inherently have to deliver the protection of persons and property in the same spaces where people are living and working. Of course the development of second wave of 'digital era governance' is still extending very rapidly the boundaries of what public services are electronically (a-spatially) deliverable (Dunleavy and Margetts, 2010; Margetts and Dunleavy, 2012). Many public services that are currently locally provided may be de-spatialized in future, as has begun to happen with public libraries, given the growth of online information and e-books. But for the moment this is an exceptional case, and local provision still dominates the delivery of public services.

In Western countries the conventional organization of public services is to divide the whole territory of the state into discrete areas, each of which has a local monopoly provider for each service. This pattern always includes municipal councils covering most public services, but it may also involve two tiers of local councils, or higher sub-regional bodies for more specialized or high-cost services. There are often also separate boards or bodies covering hospitals and healthcare on the one hand, and police services on the other – the pattern followed in the UK. In the USA, school boards are also separate, and there are in total 86 000 local bodies, some very local indeed. In federal countries these local providers are brigaded under states or provinces, which originate much of the financial transfers involved, while in more unitary countries (like Japan or England) budgets flows directly from the national-level departments.

In any reasonably large nation, there are usually numerous local providers, most of them doing very similar things to each other under similar laws and budget provisions – thus apparently opening the way for effective

comparisons of productivity rates across providers. With large N datasets, and with locally organized social and administrative data also being extensively available, it should be eminently feasible to undertake multivariate analyses that unpick the role of different causal influences on providers' productivity. So we should expect to see a very large and insightful literature on comparative productivity analyses across organizations and providers, plus the causal factors that influence them. Yet in fact, such a literature still exists only in small pieces. The number of insightful comparative studies has grown recently, but it is not large. Exploring the reasons for this situation structures the main sections of this chapter. The three key factors limiting the insights from existing comparisons have been as follows:

- The importance of quality variations across providers at the local level, given the key characteristics of complex, face-to-face public services, and some of the difficulties of tracking quality levels.
- The different ways in which quality-adjusted outputs are incorporated in parametric and non-parametric analyses.
- The problem that most previous studies have used very restricted sets of independent or explanatory variables. In particular the literature commonly does not cover key variables that bear directly on productivity levels, such as the quality of service management or an organization's level of use of modern ICTs.

7.1 DECENTRALIZED SERVICES AND QUALITY ADJUSTMENTS

Taking account of quality variation pulls productivity analysis towards looking at the *effectiveness* of government services, which is inherently much more difficult to measure. For this reason we argued in Chapter 1 that at central government level it was better to rely on comprehensively measuring agency outputs in all their key dimensions, so as to perfect output weights wherever feasible, rather than to feel impelled towards also bringing in quality weights. While we can get a high level of agreement on how to count basic outcome levels – like the numbers of patients admitted and treated in a hospital, how many children were taught in schools, or the numbers of crimes and arrests in a police area – it is far harder to get agreement on how many of these activities or treatments were effective, or which best indicate the quality of service provided. Ideally we need all stakeholders to agree on what output levels are, which becomes trickier if quality variations are factored in.

In practice, official standards are already extensively defined in healthcare, education or policing, and these often have the effect of 'focusing' public discussions. While stakeholders often disagree about what is important, they are usually ready to accept officially set standards or benchmarks (despite their limitations), and then to focus mainly on comparisons over time (are we getting better or worse?) and across areas (are we doing as well as local areas that are 'need neighbours'?). Statistics useful for assessing overall quality standards are commonly collected, but often only for indicators bearing on particularly crucial activities at the core of that organization's 'mission'. For instance, fire or ambulance services often collect data on the time taken to respond to reach a fire, car accident or medical emergency – even though these urgent responses may comprise only a part of overall service activity. Pass rates in school public exams at standard grades can be compared across schools or school boards, but without controlling for the (probably dominant) influence of non-school institutions (like families, parental support and community values) on children's performance, raw success rates may not say much about the quality of school provision itself. Similarly 'deaths in hospital' rates can be compared, but it is much harder to assess the overall quality of medical care.

However, developments in public management, especially in the later 'new public management' (NPM) era, have tended to ease some of the problems of assessing local public services' quality. The growth of micro-local agencies (such as singly managed schools or hospitals) in quasi-market systems within the public sector has been premised on more comprehensive and regular official surveillance across providers. There are many more published objective indicators (pass rates in schools or morbidity rates in hospitals) useful for citizens in choosing between providers. Periodic standardized audits or evaluations by regulators also focus more directly on local service quality, sometimes producing mainly qualitative reports, but often also some form of summary or 'star' rating. In many ways a prerequisite for more effective quasi-markets is the impartial monitoring and re-regulation of services providers, which enhances comparisons.

Finally, public administration processes often result in the collation (but mostly not the publication) of statistics on 'citizen redress' that have a lot of bearing on quality provision (Dunleavy et al., 2005, 2010a). For instance, there are complaints against municipalities or other local agencies; upwards appeals or complaints from decisions to administrative courts, tribunals, ombudsmen or other appeal bodies; and legal cases and compensation payouts. Depending on your point of view these cases may be just 'the tip of the iceberg' of poor services (as people in the 'redress industry' tend to believe), or may reflect mainly serial complainers and

litigants (as rank and file public officials often believe). Certainly these numbers incorporate some level of 'noise', but variations in citizen redress activity across similar organizations still do give useful indicators of the incidence of poor or unsatisfactory services being provided. We show in Chapter 8 that along with objective indicators they can be applied to try and quality weight service outputs, in this case relating to hospital treatments.

Service quality indicators are particularly indispensable in personalized services, those organized by professional staffs in decentralized delivery chains, especially if they involve elements of 'compulsory consumption' – as with health and social care, statutory education, environmental regulation, policing, social work and law and order services. In general, quality adjustments of outputs and productivity data will always be needed the more complex the service being provided (as in healthcare or policing); and the greater the variations in quality across agencies, localities or time periods being compared.

In three additional circumstances, trying to do without service quality data risks having especially perverse effects on the measurement of outcomes:

1. In many services, unmeasured (perhaps intangible) quality changes may trigger apparent falls in productivity. For instance, if doctors spend more time talking with each patient they see, there is a case for saying here that the quality of service provided improves – yet apparent productivity (if measured in a crude outputs/inputs measure) falls. At the least, the positive quality change means this decline is overstated. At the other extreme, unmeasured quality shading may allow apparent increases in productivity to be recorded, numbers that actually mask a worsening picture. For instance, the compulsory contracting out of hospital cleaning in the UK from the late 1980s onwards inaugurated a 'race to the bottom' in quality standards between cost-cutting contractors. By the mid-noughties this change was being blamed for part of a sharp growth in hospital-acquired infections (HIAs) in the NHS (NAO, 2004, 2009a). As part of the drive to reduce HIAs many hospitals took the cleaning function back in-house, or radically revised their contract specifications to stress quality instead of lowest costs, leading to the removal of more marginal or less reputable firms. Many other implementation changes were also made, and by 2011 the main infection rate (for an infection called MRSA) fell sharply (BBC News, 2011).
2. In many local services, exactly how and when a service is delivered, matters a great deal to what kind of output is being received. Going

into hospital for an acute procedure, but then also getting an infection that greatly extends and complicates your care is a good example of how seriously the nature of a complex service can be changed by poor implementation. In a related way, take the case of a patient who must queue on a waiting list for weeks or months before receiving treatment, during which time their medical condition can worsen dramatically, requiring much more difficult or serious interventions. Equally, getting a police response only long after a crime has finished being committed is a very different service from getting a response that is timely or more effective. *How* or *when* a service is delivered can fundamentally change the nature of *what* is being provided. Comparing across local providers where only outputs numbers are available, without employing useful quality-adjusted data, introduces inaccuracy and unfairness when comparing units with different quality profiles.

3. The strongest perverse effects introduced by a lack of quality information occur when providing a poor service actually directly *increases* apparent output levels, by boosting the demand for a service. For instance, police force X that is poor at detecting, arresting and trying criminals will let far more of them re-offend, and thus have higher crime rates and more emergency call-outs to crime scenes than another force Y, which does more effective preventative and detective work, nips criminal careers in the bud, or uses intelligence to forestall crimes being committed in the first place. Force X here will have the greater numbers of arrests and yet worse crime levels, while Y will have less activity being documented, although its crime level is also less. Similarly we noted in Chapter 2 that a hospital A delivering poorer quality care than a comparator B may have to readmit a lot of patients and redo the treatment, thereby boosting its apparent volume of outputs, whereas hospital B has a longer average hospital stay per patient and far fewer readmissions, so that its volume numbers take a double hit. In these and many other situations then a simple outputs/inputs measure could precisely misidentify the more effective providers.

These problems have been partly addressed in public policy systems that put a lot of emphasis upon targets and key performance indicators (KPIs), such as those in the UK under the Blair and Brown Labour governments (1997–2011). There are obvious incentives for top policy-makers to want to control 'games playing' around targets by local service administrators (Hood, 2006). For instance, on point (3) above the Department of Health introduced a new requirement for local hospital trusts to report

hospital readmissions occurring within six weeks of an initial treatment or operation.

However, some commentators argue that any target-led or KPI-led system of controls *inherently* creates its own distortions in the behaviour of local providers who have to report defined performance. For instance, in the NHS, Hood and Bevan (no date) argue that there are always three different areas of performance: area (i) that is well covered by metrics, and where false positive or negative results do not occur; an intermediate area (ii) where metrics are only partially and inconsistently available, and where false-positive or false-negative readings of service quality abound; and an (always larger) area (iii) where no data is available. In this view, any performance indicator or target system for top-level policy control will necessarily fail because performance in area (i) will not look like (be synchronous with) performance in area (iii). In addition, false positives and negatives will mean that there are no error-free ways of understanding even performance in area (ii), let alone extrapolating to area (iii).

These counsels of caution can easily transfer over into counsels of despair, however. Hood and Bevan consider policy systems in the NHS that were extremely crude and elementary, and which were operated in a rather insensitive, 'command and control' way. Most KPI systems have shown increasing sophistication over time. An alternative 'intelligent centre/devolved delivery' model is conceivable – one where central administrators focus on acquiring a lot of high-quality information covering multiple output and output-quality indicators automatically. For instance, they might use the kind of digital reporting systems commonly used all over large-scale private service companies, like major retailers Tesco and Walmart. Local public service providers in such a system can be assigned more freedom to vary their service strategies and outputs mix, so long as their overall service quality and output levels are maintained within acceptable bounds. Top decision-makers here would have influence with local managers, but not exercise command and control over them on operational matters.

Whatever the balance of these arguments for policy-makers, in studying productivity researchers are usually looking at much more aggregated performance across local agencies. By ensuring that indicators of *all* an organization's main output categories are included in the cost-weighted overall output measure, many of the problems of biases towards 'core' services in indicators can be controlled. For instance, in a fire service it should be feasible to get at least some metrics of non-emergency or fire-prevention work, and in hospitals to control for case-mix effects. Researchers can also usually use administrative statistics on quality levels as part of a wider strategy for quality-weighting outputs. General 'quality

of performance' weights can also help; for instance, looking at local public satisfaction with hospital, police or fire services. There may be problems here with local loyalism (citizens not wanting to run down their area); or alternatively low expectations problems (citizens inured to low-quality services not criticizing what they see as expected or inevitable). But it should not be out of the question to assemble baskets of measures that give a reasonable view of service quality variations across local providers.

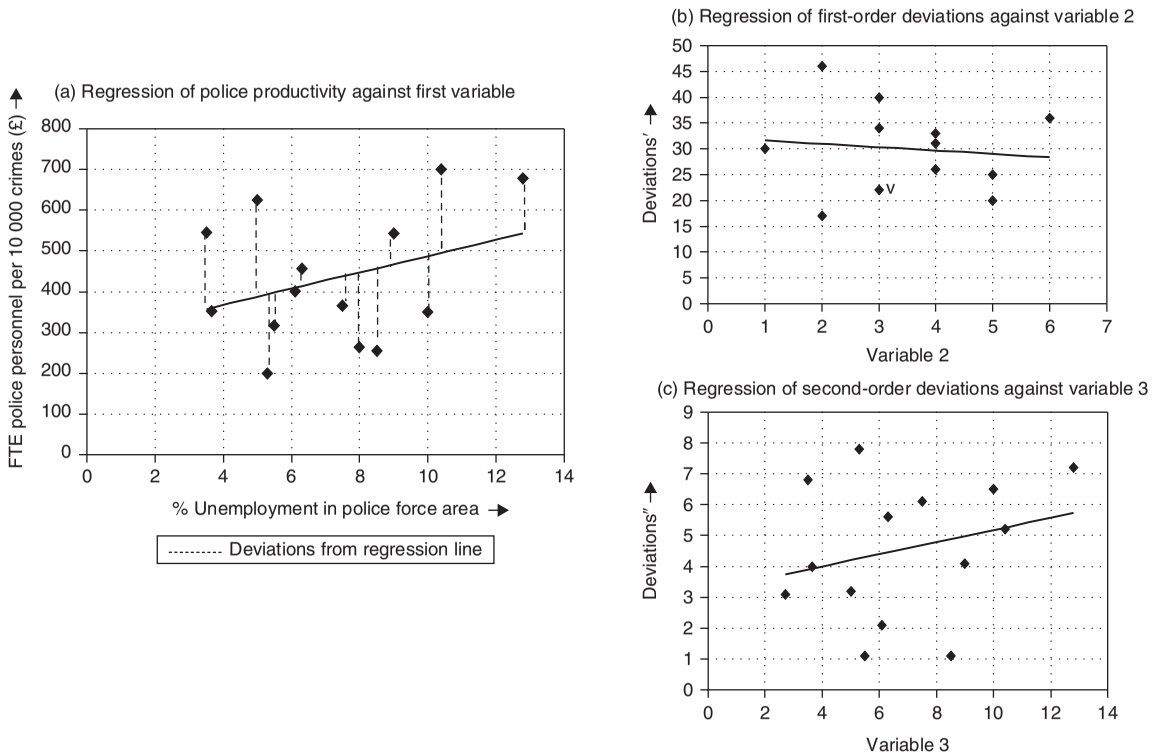
7.2 INCORPORATING QUALITY MEASURES INTO PARAMETRIC AND DEA ANALYSES

With large numbers of comparable public service providers to analyse, it becomes feasible to deploy the two methods discussed in Chapter 1 but not realistically implementable at national government level – namely, parametric approaches using conventional regression techniques; and non-parametric approaches, where we focus on data envelopment analysis (DEA). We look more closely here at how quality-weighting of outputs can be incorporated in both approaches.

Parametric Approaches

Looking across a reasonably large set of local providers, this approach proceeds by seeking to estimate the influence of different explanatory variables (taken one by one) on productivity levels, treating each variable as a parameter for the influence of all other variables. Most commonly implemented via regression analysis, we can briefly give an intuitive and non-technical explanation of a parametric approach. Figure 7.1a shows the observed levels of police productivity (defined as the number of full-time equivalent [FTE] police per 1000 crimes in an area) plotted against a first explanatory variable, chosen here to be the percentage level of local unemployment. (The idea here is that as unemployment increases, so the seriousness of local crimes may increase, necessitating more police people per 1000 crimes.) A 'least squares' regression line is defined, one that minimizes the vertical differences between observed productivity levels for each agency, and the level that would be predicted for that agency given its placing along the horizontal axis showing the values for the local unemployment parameter (or variable). These vertical differences are called first-order deviations and the smaller the sum of deviations the more closely the points fit around the regression line, and more of the variance in productivity can be statistically explained by the X axis variable. Where the fit remains relatively poor (as is the case in Figure 7.1a), the next stage

Figure 7.1a–c An intuitive view of ordinary least squares regression analyses of productivity variations across local providers

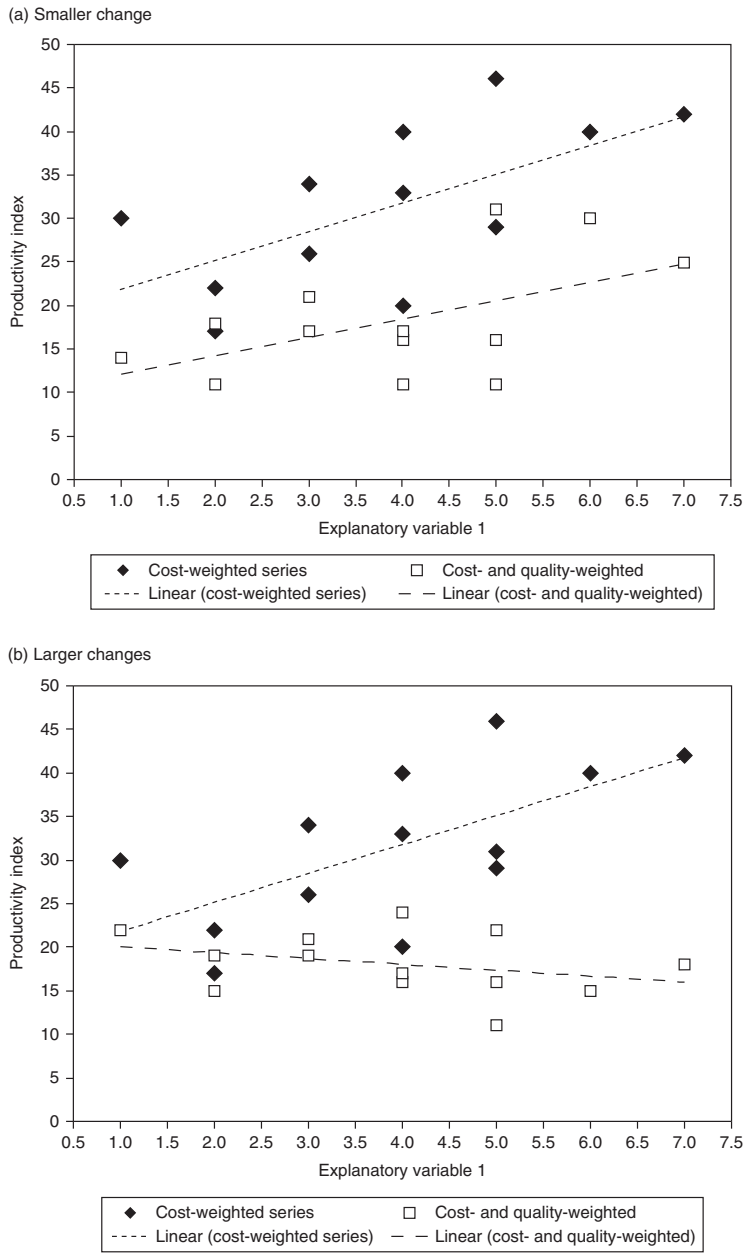


is to redo the analysis, this time matching the first-order deviations in Figure 7.1a against the values for a second parameter B, a step shown in an illustrative way in Figure 7.1b. Again we define a second regression line, measure the second-order deviations from that line and match these deviations in turn against scores for a third explanatory variable, as in Figure 7.1c. At each stage of this process the variance of the points for each local agency will broadly tend to reduce, and it will usually become harder and harder to find additional explanatory variables that make a difference.

In regression approaches, an important influence on the results can often be exerted by the ways in which explanatory variables are defined, how effectively data has been collected and by the order in which different variables enter the equation. Analysts can rely on statistical programmes to enter variables in an automatic way, in a sequence determined by their closeness of fit with the dependent variable (either the original productivity scores or the remaining, still-unexplained deviations calculated at each stage of the analysis). Alternatively, the analyst may define an order in which variables are to be entered, chosen on theoretical grounds to be a logically compelling sequence. Variables are then input in a fixed order, but the statistical package still determines whether any apparent influences found are statistically significant (that is, unlikely to have arisen simply by chance or sample fluctuations). It is common to finish up with quite a large number of alternative final models composed only of significant variables. Models can then be compared in terms of the overall proportion of the variance in the dependent variable that each model explains (the R^2 statistic). So long as the models basically agree in identifying the most theoretically and empirically important variables, we should normally choose the model with the highest R^2 statistic as the best. However, if models with different sets of explanatory variables emerge as almost equally significant, and particularly if the influence assigned to different variables seems to fluctuate from one model to another, then choosing between models becomes more complex.

How does quality adjustment influence the outcome of regression analyses? Essentially each local agency is still situated at the same point on the horizontal scales for the explanatory variables in Figure 7.1. But now the impact of the quality adjustment is to raise or lower the point where each local agency is located on the vertical axis productivity measure, as shown in Figure 7.2. Normally we should expect to see that quality adjustment alters the productivity levels at which scores sit, and incrementally reshapes the new distributions of local providers, without altering them radically, as shown in Figure 7.2a. However, the more that cost- and quality-adjusted productivity or output scores differ from the scores for

Figure 7.2a and b Hypothetical impacts of quality adjustments on data patterns



cost-weighted outputs alone, the more likely it is that the cost- and quality-adjusted plot will have a different shape, including changing the signs of one or more of the explanatory variables from positive to negative or vice versa. Figure 7.2b shows a hypothetical illustrative example of one way this might happen.

To explore more about how parametric approaches work, readers should turn to Chapter 8, which sets out in detail a multivariate regression analysis of how productivity varies across English hospital trusts. The final things to note about this strand of work is that it makes some important assumptions, including (most fundamentally) that the same basic processes operate across the dataset as a whole, and that a 'tournament' of competing regression models is the best way to surface what these processes are. The more inclusive a study's variable set is by including all the factors theoretically linked to shaping productivity levels, the greater the confidence we can have that the analysis is not subject to missing variable bias. Getting stable and consistent scores for the influence of the same parameter on productivity levels across the different regression models tested is also reassuring.

There are a number of considerable advantages to parametric studies. First, they explicitly allow for unknown variables, and for the distribution of productivity levels to be influenced by random shocks – what matters is the location of regression lines and not particularly the locations of individual data points. Second, from regression analyses it is possible to compute 'elasticities', that is, an estimate of the extent to which a change in an explanatory variable can be expected to influence local productivity levels. Knowledge of elasticities is helpful for policy-makers in suggesting where to put resources or effort in trying to boost local productivity levels.

As usual, parametric approaches also have their limits on how far comparative studies can go in understanding which influences within local agencies or local social environments shape strong or weak productivity levels. First, *assembling large N datasets* is crucial. In general, the larger the number of agency data points included (the bigger the N), the more likely it becomes that there will be enough available information (enough degrees of freedom) to sustain an analysis of multiple variables. Now since the number of local providers per country is in fact fixed, this may seem a factor outside the analyst's control – as it certainly is for a single cross-section analysis, a snapshot at one point in time, such as our account in Chapter 8.

However, where it is possible to assemble good-quality data on local agencies' outputs, inputs and productivity scores across a run of years, along with data on explanatory variables for the same time period, then analysts can 'pool' the data for multiple years and use more sophisticated techniques (such as pooled regression analysis) to unpick the influences

discovered. This approach can also check that the pattern of apparent causal influences on productivity remains stable over time, which it may well not do. It can also provide a direct analysis of the parameters most associated with improvements or declines in productivity over time. In practice, there have been fewer but quite important studies that use over-time panel data. In the private sector, Caroli and Van Reenen (2001) and Bloom et al. (2005) have used panel data to estimate the determinants of productivity across firms in different countries. In the public sector, Garicano and Heaton (2010) have used panel data for a set of police departments from 1987 to 2003 to estimate the determinants of productivity, which we discuss later in this chapter. We can expect that as data records and collection improves, it will be possible for scholars to do more comparative productivity studies using panel data.

Second, where *variations in service delivery are compressed* across local providers, this can inhibit the usefulness of regression approaches. The observed variance in local agencies' productivity levels may be extensively constrained by national governments (or state governments in federations) imposing standard laws, regulations and budgetary allocations onto local agencies. In addition, nationalized systems of professional controls may additionally restrict allowable 'good practice' (Dunleavy, 1982). Especially in centralized nations (like the UK and some other European and 'Westminster system' countries), these 'straitjacket' influences may constrain provider agencies into delivering standard services in standard ways, thus inhibiting the scope for local innovations, service variations, or the use of different technologies or business process models. Hence, the range of the dependent variable may be less than we would expect amongst (say) firms in competitive markets. However, in some previously centralized countries (such as the UK), some commentators have seen the creation of micro-local agencies (MLAs, such as individually managed or 'foundation trust' hospitals and local managed schools) as a countervailing tendency, arguing that MLAs have more scope to innovate and deliver services in less preordained, standard ways.

Third, the *widespread contracting out* of public services has produced in many countries (especially the UK) the development of large, oligopolistic firms that deliver the same services in many districts. Whether it is maintaining traffic lights, cooking school meals, collecting refuse, or providing hospital ancillary services, these companies use the same standard operating procedures nationwide. So the *apparent* diversity of local purchasers may actually be somewhat illusory in terms of the real number of providing organizations at work. A wide range of municipalities or hospital boards are signing contracts with an underlying and much smaller number of main corporate contractors, who implement services

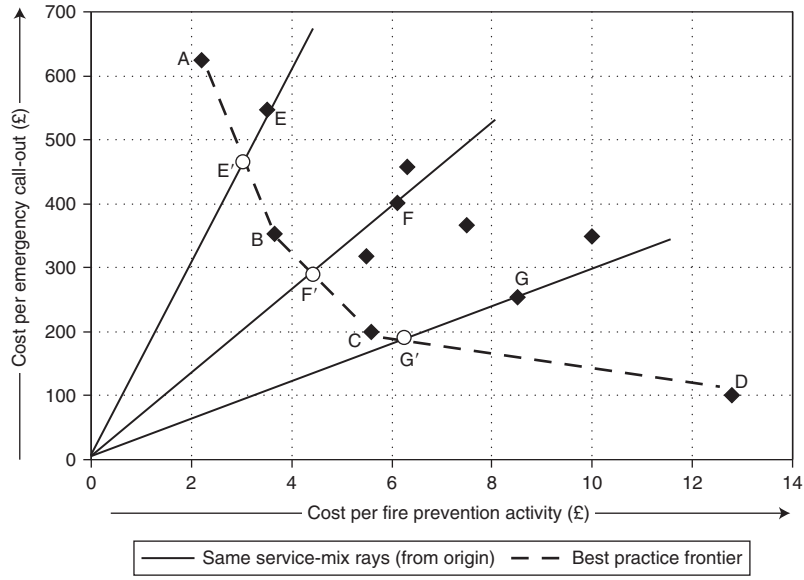
production. Comparing across localities in such a situation may say something about the choices made by the decision-makers within the provider bodies (the councils or hospital boards), and especially illuminate their commissioning or contract-negotiating skills, or perhaps just the dates when the contracts were signed. But the real organizational units undertaking actual production (and hence in an operational position to be able to improve productivity directly) are the smaller number of major firms in operational control of services. There are analytic approaches that might address such situations (such as hierarchical regression models), but so far they have not been applied in this area. The contractual data needed on who is actually delivering implementation where have also not generally been available.

Data Envelopment Analysis

One of the fundamental assumptions of parametric approaches is that there exists a production function that conforms to a particular probability distribution, which can be more successfully approximated by improving the available data, exploring more potentially important explanatory variables, and so on. By contrast, there are a set of analytic approaches that do not assume a given probability distribution. We focus here on one particular non-parametric approach, data envelopment analysis (hereafter DEA), which originated in operational research in the late 1970s, and was explicitly intended to help illuminate the assessment of performance of 'decision-making units' (DMUs) in non-market contexts (Charnes et al., 1978; Cooper et al., 2011). In particular, DEA seeks to cope with situations where production functions are not known, and to undertake comparisons based much more clearly on the best known data about what is possible for DMUs to achieve. We have already introduced (in Chapter 2) a simple (single-output) example of how DEA does this. It may be helpful to consider here a slightly more complex (two-dimensional) illustration of how the approach works.

Suppose that we have data for how a set of local fire services perform on two different kinds of outputs – the number of call-outs to fires or accidents that they have attended in a year, an indicator of their emergency services load; and the number of fire-prevention activities they have undertaken, such as inspecting premises for fire safety issues, or advising businesses and households on fitting fire alarms and taking other prevention measures. We also have data for the total costs of each fire service (or for the number of FTE personnel each employs, if we are examining only labour productivity). Even if we do not have accurate cost weights for outputs data, the DEA approach can be useful here in locating services

Figure 7.3 A data envelopment analysis of local fire services' productivity (hypothetical example)



against each other, and in estimating how far performance might be improved – either by saving costs and personnel, or by expanding the demand for services (essentially applicable only for preventative work in the case of fire services). We compute separate ratios for each local agency of the costs per fire prevention activity, shown as the horizontal dimension in Figure 7.3, and for the costs per emergency call-out, shown on the vertical axis. We then graph the combined performances for a set of agencies on both dimensions. (To keep the diagram clear we show only a few data points, but a real analysis might have between 100 and 500 data points.)

The DEA approach argues that because the service mix between prevention and emergency responses varies widely, the best way to assess the relative efficiency of agencies is to compare their performance along same service-mix lines (shown as rays out from the origin). The best-performing agencies are those that in terms of their particular ray are as close to the origin (i.e., have the lowest combined costs of provision) as possible. A 'best practice' frontier is defined by joining up all those agencies closest to the origin as shown, with the requirement that the frontier must 'envelope' all other agencies' data points – in this case it is defined by four agencies, A, B, C and D. All the remaining agencies are higher cost

and less productive. To see how much they are adrift from best practice, and what they could achieve if they could match best practice organizations, we look along the service-mix ray running through them, and extrapolate to where that particular ray cuts the best practice frontier. For instance, for agency E that point is E′, and for agencies F and G it is F′ and G′ respectively.

The strength of DEA is that it draws on the available data to determine what is feasible. It does not try to identify the production possibility frontier (as parametric approaches do). The comparisons that DEA draws derive from known data about decision-making units, and its data demands are much more modest than for parametric studies. The approach can be extended to more than two dimensions of performance using linear programming statistical techniques to compare across multiple output/input ratios. These more complex models cannot be conveyed visually, but they are achievable using relatively simple software packages.

Critics of DEA argue that the approach is very vulnerable to the correct identification of the DMUs that are on the frontier. If the data for these particular agencies are not accurate, or if the frontier cases are very distinctive or unusual DMUs that are systematically unlike the main mass of agencies, then basing the whole frontier analysis on comparisons with them will not produce useful insights.

Crudely done DEA studies might even misinform policy-makers about what improvements are feasible. For instance, a study of Australian hospitals in the state of Victoria found that small, rural hospitals (using local generalist family doctors as part-time medical staff) were much cheaper than hospitals with a full-time staff of specialist doctors in towns and major cities. But that did not mean in any way that the bigger hospitals could or should try to match the lowest cost-base units, since their mission and case mixes were completely different (Steering Committee for the Review of Commonwealth/State Service Provision, 1997, pp. 51–67).

Yet DEA techniques have also been extended in recent years so that multivariate analyses can be conducted. In principle, different best-practice frontiers can be defined for different classes of agencies. Similarly the scope for feasible improvements does not necessarily have to be defined against the best-practice frontier for all agencies (which may often reflect special factors unique to different areas or organizations). Instead, just as with regression analyses, policy-makers can consider much more feasible kinds of comparative insight. For example, government executives, ministers or audit bodies can estimate how much input costs might be saved if DMUs in the lowest-performing three quartiles could improve their performance to match those agencies whose performance defines the 75 per cent frontier (where the second quartile starts). Alternatively it is

possible to estimate how much output levels or services quality might be improved with the same change.

DEA approaches have become increasingly sophisticated in recent years. In fact, they have been extensively applied in private sector contexts for improving efficiency and productivity. Some useful applications to measuring local agencies' productivity have also been undertaken. The low data requirements for this approach mean that it can be deployed effectively even in circumstances where estimating cost-weighted outputs is difficult because the same staff or inputs are used to produce multiple activities or output streams.

The DEA techniques are especially helpful in conditions where the quality aspects of service provision are recognized as important, but where quality measures are partial or may not even be very tangible (as with services) – as in healthcare (Solà and Prior, 2001; Clement et al., 2008). This approach can help when it is not feasible to estimate the costs of providing better-quality services (or the savings that might be made from reducing service quality), given available information. Combining available quality indicators with data on multiple different outputs is a feasible application of DEA packages.

7.3 CAPTURING INDEPENDENT VARIABLES THAT ACTUALLY SHAPE PRODUCTIVITY

One of the most disappointing aspects of the literature on local service providers' productivity is that only a very restricted range of explanatory variables is considered, especially in analyses undertaken by economists and operational researchers. A large number of economic studies have focused on issues around economies of scale and scope, trying to produce definitive answers on what is the 'optimum' size of hospital or police force for maximizing outputs/inputs productivity (or less commonly for achieving maximum effectiveness in service delivery). Theoretically issues here have absorbed quantitative analysts. And if the cumulated studies could lead to especially clear or agreed conclusions then they might have high relevance for policy-makers – especially when governments periodically consider how to reorganize hospital care, or how to restructure the area structure of police forces.

However, the overwhelming conclusions of most studies in this vein has been that reliable evidence of scale economies peters out fairly quickly in the transition from small to medium-size facilities or local authorities. In most public services, scale economies are not readily apparent in the transition from medium-size to large facilities or municipalities – because

big hospitals, big police forces and big city local authorities often confront higher-intensity problems, which are inherently more expensive to handle. In addition, most decision-makers facing operational choices are working with a fairly given set of facilities and production processes, and they are not remaking long-run service architectures. So their practical ability to change the existing scale and scope of services being provided is often small.

In the private sector the analysis of economies of scope has focused on how firms and organizations acquire additional capabilities from handling many different but related activities. For instance, with flexible manufacturing systems, modern firms have adapted their machinery, staff training and production processes so as to produce relatively short runs of many different products. Within the government sector there has not been any equivalent progress on economies of scope. Before the 1980s, in countries with highly modernized local governments and health authorities (like the UK), the issue of scale enlargement was chiefly discussed in terms of achieving a scale that would allow specialist facilities to be run by a single council or provider. But since then there has been a lot of progress in collaborative contracting and in constructing variably sized *coalitions* of public authorities to run expensive or specialist facilities. There are also now much better ICT and organizational networking arrangements for decentralized agencies involved in partnerships. Accordingly, groups of small or medium-sized local authorities or health bodies can often collaborate efficiently to provide specialist facilities that are collectively used and funded. The development of micro-local agencies in countries with previously centralized local authority provision (such as the UK and Sweden) has also undermined many previous claims for economies of scope, which proved not to be evidence-based.

In healthcare most comparisons have not been focused on overall productivity in hospitals or district heathcare providers, but on much more specific dependent variables, such as morbidity in hospitals for different kinds of treatments or operations. A large number of studies in a clinical audit vein have focused on just one kind of treatment, where the dependent variable is very consistently defined across hospitals. In the UK there have been some comparisons also of how hospitals or district health bodies have performed in terms of meeting specific or high-priority targets set by the central government (Cooper et al., 2010). But at any one time many things might be going right or going wrong in hospitals, only some of which relate specifically or solely to performance on a particular area of treatment. In any normal hospital, some medical teams will be performing well, while others are marking time, and still others are in a period of some decline in their activities or competences. So the performance of individual

teams, in combating particular diseases or carrying out particular treatments will often look very different from an analysis of overall hospital productivity.

In particular, more segmented analyses do not usually capture well the importance of 'whole organization' elements arising from the general management of hospitals. For instance, a medical team that is performing well in terms of achieving high success rates with operations of type X at low cost may still be severely constrained in output or productivity terms if their hospital confronts a budgetary crisis that requires all units to cut back or to limit the scale of their operations. Similarly, where general patient care in a hospital is poorly managed in terms of prevention of infections, then large-scale outbreaks of hospital-acquired infections occur – as in the UK at the Maidstone and Tonbridge Wells Hospital Trust from April 2004 to September 2006 (Healthcare Commission, 2007). Here the otherwise good performance of many different medical teams across different wards may well take a general hit, because their patients become infected and have to stay in hospital longer with complications.

Perhaps surprisingly, there are only a smallish number of local productivity analyses that address issues of hospital-wide performance, and allow for administrative factors to shape performance. In particular, very few studies include variables covering overall organization factors bearing directly on productivity, such as the use of IT, or the quality of management and extent of modern management practices (but see West et al., 2006; Borzekowski, 2009). Far fewer studies cover such aspects because it is hard for analysts to specify and acquire data on these kinds of explanatory variables. In assessing management factors in particular, the problem is that we need independently specified indicators of the quality of managers or the modernity of management practices (separately measured from the organization's overall performance).

Of course, 'hybrid' data can be obtained on hospitals' or local agencies' overall performance that incorporate reference to their management, chiefly indices from external evaluators of how well a hospital or a local agency is being run. For instance, in the UK, hospital trusts were awarded 'star' ratings by NHS evaluators. Local authorities were given an assessment for managerial quality by a national body called the Audit Commission. From 2002 to 2008 this focused just on local councils' own services and activities, and was modestly called the Comprehensive Performance Assessment (Audit Commission, 2011). From 2008 to 2010 the evaluations also looked at how municipalities worked cooperatively with other local service providers (like the police force and health service); in this form it was called the Comprehensive Area Assessment. Locally

managed schools in England have also been graded by an inspectorate, Ofsted.

The trouble with using any of these assessments, however, is that they were compound measures. Normally the evaluators made visits to each body assessed, and looked at a wide range of factors in making their judgements – such as the visible quality of service indicators, evidence of competence and good morale amongst staff, financial performance, the governance of the institutions and how effectively organizational leaders tackled problems. Nonetheless, at root the evaluations were also heavily based on units' objective performance. So instead of being genuinely independent measures of managerial quality, the evaluators' assessments largely incorporated and rested on the dependent variables that we are interested in analysing, namely the productivity, cost-effectiveness and output achievements of the unit being studied. As a result the correlations between such overall organizational management assessments and productivity performance may be so close that the overall assessments cannot be used as explanatory variables.

For information and communication technologies the problem has been quite different, namely that useful information on how far and in what ways ICTs are used in service delivery has been hard to find. In earlier periods, when IT was first being adopted and early automation processes were underway within local service providers, data on ICT costs and on the diffusion of particular styles of ICT use across local providers were more useful indicators than they are now. However, as we noted for private sector firms in Chapter 1, data on ICT spending levels no longer capture the actual use of modern ICTs very well. In the digital era a high relative spend on ICT does not necessarily signal an organization that is using a lot of up-to-the-minute technology, since web-based approaches are not as expensive as older ones. It may indeed indicate the reverse, an organization struggling along with older legacy systems and making little effective use of online transactions and internet-based services.

In the remainder of this section, we look at two recent and important studies that have innovatively addressed these problems. The first is an analysis by Van Reenen, Bloom and others that tries to unpick the influence of management practices on the performance of English NHS hospital trusts. The second is a large-scale, over-time analysis by Garicano and Heaton of how far the adoption of ICTs and related management practices by US police forces has helped to improve their performance across a long period. We also describe briefly an alternative approach that uses web-based research methods to assess management practices and ICT use, methods that we go on to apply in Chapter 8 to analyse productivity in English hospital trusts.

Analysis of Management Practices in Shaping Performance

A sophisticated interview-based approach to assess the quality of hospital managements in shaping performance was deployed by Nicholas Bloom, John Van Reenen and colleagues. The Bloom and Van Reenen (2010) method was developed for analyses of productivity and performance in private sector manufacturing firms, and it was later generalized to apply also to services firms and then to public sector organizations (see Bloom et al., 2005 and 2009a). It essentially involves the analysts drawing up a structured set of dimensions of management good practice (covering maybe 15 to 25 dimensions). These are derived from the previous literature and they are closely contextualized to fit the nature of the industry being covered. For each dimension the aim is to be able to score firms or agencies into one of three overall categories – not implementing that aspect of good practice, or partially implementing it, or fully implementing it.

To uncover this information the research team phoned an appropriate manager or member of professional staffs in 100 NHS hospital trusts and 21 comparator private hospitals, contacting either one or two persons per hospital (Bloom et al., 2009a). In a strongly evidence-based way, they sought to discuss with each interviewee 18 dimensions of hospital management, surfacing many different nuggets of information without using a fixed questionnaire. Instead, most questions were open-ended ones, and a more dialogic, 'elite interview' style of enquiry was undertaken on each dimension, until the interviewer could classify the organization's rating with some confidence. Interviewers were trained graduate students, 'double-blinded' by not being aware of a hospital's performance in any way. Each interviewer conducted around 46 interviews, so that they became experienced evaluators of responses. A key feature of this approach was that interviewees were also not told that they were being scored in any way.

Figure 7.4 shows examples of how questions were asked and ratings defined. In the hospitals study, the research team contacted 161 people, always including a senior hospital trust administrator in each case, and also reaching a senior doctor as well in some cases. At the end of the process, hospitals were assigned a composite score for their ratings for their management quality across all the dimensions.

The data for the management quality variable were then added into a dataset that included a large number of other possibly important independent variables, and were then regressed against a very useful and inclusive range of dependent variables. These included some specific indicators of hospitals' medical performance (covering the 28-day mortality rate for emergency admissions for acute myocardial infarction and the rate for

Figure 7.4 Two examples of how the Bloom et al. study of English hospitals assessed management quality on 18 different dimensions

(3) Continous improvement			
Tests process for and attitudes to continuous improvement and whether things learned are captured/documented a) How do problems typically get exposed and fixed? b) Talk me through the process for a recent problem that you faced c) How do the different staff groups get involved in this process? Can you give examples?			
	Score 1	Score 3	Score 5
Scoring grid:	No, process improvements are made when problems occur, or only involve one staff group	Improvements are made in regular meetings involving all staff groups, to improve performance in their area of work (e.g., ward or theatre)	Exposing problems in a structured way is integral to individuals' responsibilities and resolution involves all staff groups, along the entire patient pathway as a part of regular business processes rather than by extraordinary effort/teams

(4) Performance tracking			
Tests whether performance is tracked using meaningful metrics and with appropriate regularity a) What kind of performance indicators would you use for performance tracking? b) How frequently are these measured? Who gets to see data? c) If I were to walk through your hospital wards and theatres, could I tell how you were doing against your performance goals?			
	Score 1	Score 3	Score 5
Scoring grid:	Measures tracked do not indicate directly if overall objectives are being met, e.g., only government targets tracked. Tracking is an ad-hoc process (certain processes aren't tracked at all).	Most important performance indicators are tracked formally; tracking is overseen by senior staff.	Performance is continuously tracked and communicated against most critical measures, both formally and informally, to all staff using a range of visual management tools

Note: The dimensions shown here are those for 'continuous improvement' and 'performance tracking'.

Source: Bloom et al. (2009a, p. 28).

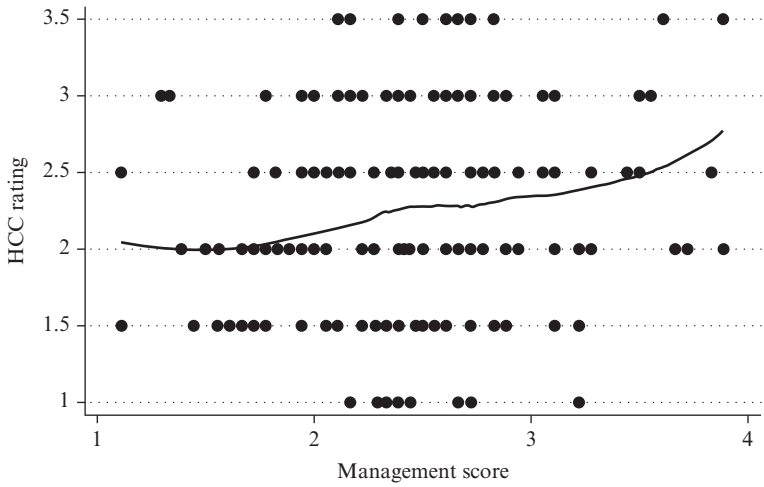
non-elective surgery), and two more general performance indicators (the size of the waiting list for all operations and the rates of MRSA infection, a particular 'superbug' on which a lot of public and NHS attention focused at the time of the study, in 2006). The study also looked at composite performance judgements made by a Department of Health body, the Healthcare Commission (HCC), which evaluated hospitals along two dimensions (on a scale from 1 to 4):

> The efficiency of resource use is measured [by HCC] by the number of spells per medical employee, bed occupancy rate and the average length of stay. Service quality is measured by clinical outcomes (readmission risk and infection rates), waiting times and a measure of patient satisfaction as well as job satisfaction of the staff. (Bloom et al., 2009a, pp. 6–7)

The basic distribution of the HCC rankings against the research team's scores for management quality is shown in Figure 7.5. This is a classic example of the kind of 'bait ball' pattern familiar across policy systems where decentralized agencies deliver services that are paid for from central government grants and are in turn subject to relatively intensive levels of central supervision. In 2006 the NHS in England was especially in the grip of a 'targetology' wave (described further in section 8.1, Chapter 8). Just as in nature, sardines or tuna form a 'bait ball' to maximize the chances of individual escape from predator attacks (such as sharks or dolphins), so this kind of pattern is functional for local agencies. So long as local managers can perform within the middle mass they effectively become invisible to national regulators and budget controllers. Hence few hospitals appear as excellent on even one dimension (in the top left or bottom right quadrants of the figure) because the effort to achieve excellence in one area could imperil performance and attract criticism if other areas lag or go wrong in the process. Only a few hospitals did well on both dimensions, at the top right. Similarly, there are few hospitals that are poor on both measures (in the bottom left quadrant) – because if their organization gets into this sub-target zone, hospital managers have strong incentives to focus attention on and fix the areas of conspicuously lagging performance. Managers hence rationally assign extra resources not to try to be excellent, but just to get themselves back safely hidden within the main mass of the 'bait ball'. It is clear in Figure 7.5 that there is no close fit between management quality and overall performance, although the average HCC scores did rise from 2.29 in hospitals scored lowest on management, to 2.58 for those with middling scores, and to 2.81 for those scored highest on management quality (all on a possible numerical range running from 1 to 4).

This initial basis for analysis looks unpromising. But a great advantage

Figure 7.5 The basic pattern of English hospital trusts' performance, 2006, charted against their scores for management quality, as found by Bloom et al.



Note: Each point represents a survey response. The vertical axis shows the average hospital care score (on a range from 1 to 4) in 2005–06. The horizontal axis shows the average management score assigned by Bloom et al. (2009a) across their 18 questions. The line is the local linear regression line.

of multiple regression techniques is that the picture can change greatly when we take account of controls for other aspects of hospital trusts' situations. After controlling for hospitals' differing case mixes, and their size, regional location, and whether they were speciality hospitals, Bloom et al. found that better managed hospitals (on their scores) were significantly associated with somewhat lower morbidity for emergency admissions, lower waiting lists and with fewer hospital staff planning to leave. They also found highly significant associations between hospitals being scored better on their management measures and achieving high HCC performance ratings. The size of effects here was large, with management quality accounting for around one-seventh of the variance in hospitals' average HCC rankings.

The research team also found that the patterning of the management quality scores showed that scores were higher in foundation trusts (a category of larger, better managed trusts to which in 2006 the central ministry assigned more independence to set their own policies), and in hospitals with more clinically qualified managers, and of larger size. Management scores also seemed to be stronger in those parts of the country where

multiple hospitals were competing for patients (as in big cities) than in hospitals that were local monopolies.

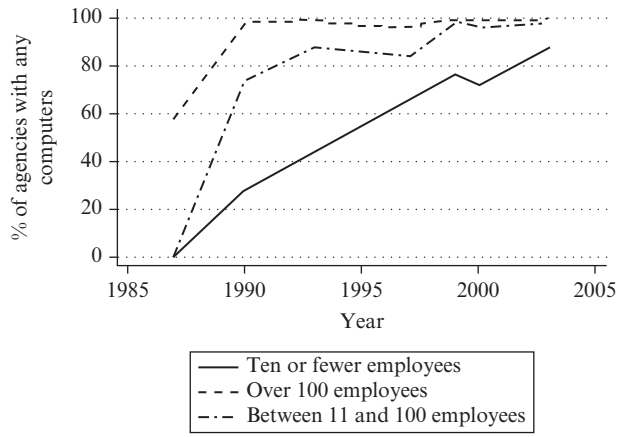
Overall, Bloom et al. (2009a) concluded: 'Our measure of management quality was robustly associated with better hospital outcomes across mortality rates and other indicators of hospital performance' (p. 15). This is clearly an innovative approach and an impressive study. Its conclusions seem to offer rarely available and strong evidence to support the widely held 'managerialist' conviction that the quality of local management and organizational leadership clearly or obviously condition the performance of complex organizations like hospitals.

A key area of vulnerability, however, is the narrow informational basis for deriving the management scores themselves. Across most of the organizations characterized, the information basis is limited to the (admittedly detailed) responses from just one person, and in no case were more than two people interviewed. This seems a fragile foundation for grounding such a key explanatory variable. And although the study incorporates 'noise' controls relating to the interviewers for each case, it does not seem to include any variables that characterize the interviewees or could control in any way for their almost certainly variable perspectives. The famous Graeme Allison (1971) dictum that 'where you sit determines where you stand' applies in any organization. And without any real cross-checking of the management quality information from one respondent with other respondents, it is hard to know what reliance to place on them. The same criticism applies to this research approach in other contexts – for instance, to studying differing management practices across US, European and British manufacturing firms, where the same research team typically talked to one or two people only and drew similar, widely noticed conclusions about the importance of management practices (Bloom et al., 2005). Of course, the research team could legitimately respond to this criticism by asking: 'Well if not our approach, what alternative would you suggest for surfacing genuinely independent information on management practices?' This is an issue to which we return below in discussing web-based research methods, and where we implement an alternative approach in Chapter 8.

Analysing a Larger Over-time Dataset

Analysing a much larger, over-time dataset is the central innovation that allows a 2010 study by Luis Garicano and Paul Heaton to explore the role of IT investments in shaping the performance of US police departments. Using survey data completed every three years across the period 1965 to 2006 by many different US police departments, the authors built

Figure 7.6 The spread of ICT use across US police departments, 1965 to 2005



Source: Garicano and Heaton (2010, p. 174).

up a dataset of police performance with 19 400 cases for many variables, around 13 000 to 16 000 for almost all others, and never less than 5000 cases. Their primary focus was on whether the huge take up of computers and other new IT by police forces across the USA produced any noticeable improvements in their performance, assessed not using productivity but measured in terms of arrest rates (an activity index) and crime rates (a measure of the problem environment). Figure 7.6 shows that larger police departments with over 100 employees were early users of ICTs, followed swiftly by medium-size departments, and at a much more gradual rate by the smallest (mainly rural) departments with less than ten employees. The authors' chosen IT use indicator was a compound measure of whether departments in a given year used computers for crime analysis, investigation and dispatch, and used data records for arrests, service calls, criminal histories, stolen property, traffic citations and warrants. The transition from mainframes to PCs to mobile computing was also covered.

To control for other variables the dataset also covered the size and complexity of departments (in terms of employee numbers, number of special units, number of organizational levels and total written directives); the make up of departments across uniformed officers, field operations staff and technical staff; the educational and training requirements for officers; and the demographic make up of the unit's local population (in terms of its size, the local racial make up, local poverty rates, mean

household incomes and the education levels). The dependent variables for performance considered included the offence rates and clear up rates for overall crime (and those for sub-sections such as violent crime and property crime), and information about assaults on officers or deaths of police officers.

The authors' first and rather dismaying results were that they could not find any statistically significant effects demonstrating that the large-scale take up of IT by police departments was associated with any improvement in how they performed on the dependent variables. Rather than lowering clear up rates, increased IT use actually led to a growth in crimes being recorded, especially in relation to property crime. Detailed analysis showed that this negative association with crime amelioration was solely an information effect: 'Offense reports increase by 10% once computers are available for record keeping' (Garicano and Heaton, 2010, p. 184). Other than that, neither overall ICT use measures nor individual measures showed distinctive impacts on outputs or outcomes. The exceptions were that police forces making more use of computers and IT also subsequently upped their recruitment requirements to demand college education for all employees, and increased the amount of training they undertook. Careful checks were carried out for contaminating effects in the data, including the possibility that IT use may have only longer-lagged effects; that departments that were stressed, recently failing or mismanaged might be more likely to adopt IT as part of 'turnaround' reorganizations; that IT use responded chiefly to the level of contextual IT use in the local communities served; or that IT use responded only to the financial strength of the departments. After all these checks failed to shake the central finding, the authors conclude: 'It is surprising that IT appears to exert little effect on policing outcomes given the widespread use of IT in modern police departments' (Garicano and Heaton, 2010, p. 180). The authors explicitly link this central finding to the early wave of studies in private firms that found no boost from ICTs to firms' productivity or profitability.

They turn next, however, to exploring whether IT changes might have had effects but only *in combination* with shifts in management policies and in the organization of business processes to align them with the new technologies. Garicano and Heaton argue that a particular style of policing associated with the COMPSTAT approach introduced by New York Police Commissioner Bill Bratton was the most relevant change (see Bratton and Malinowski, 2008). They especially focused in their dataset on five aspects – the use of information technology for crime data collection and analysis; the adoption of a problem-solving paradigm to reduce crime rates; the use of feedback for priority-setting and evaluation; police

forces using a geographic-based structure for deploying personnel to particular local areas; and stronger internal accountability. (Their data did not cover another key element of COMPSTAT, the empowerment of middle managers, and it only partly illuminated a final element, encouraging internal accountability for results.) The authors classed police departments that showed simultaneous elements of at least five of the seven COMPSTAT practices above in half the years covered as showing good management practices and then looked to see if this made a difference to police performance:

> The results are striking. Whereas the estimates on each of the individual management practices are of negligible magnitude and generally statistically indistinguishable from zero, the combination of practices into a COMPSTAT system yields positive and significant effects on clearance rates. The coefficient estimates of around 2% imply a roughly 10% gain relative to the average clearance rate of 22%. For violent crimes, which in many cases are an area of particular investigative emphasis, the point gains are even greater. (Garicano and Heaton, 2010, p. 193)

The authors further conclude that:

> One additional reason for the weak aggregate relationship between general IT and policing outcomes may be that while many agencies utilize some type of IT, relatively few have yet implemented all of the complementary management practices that allow IT to impact police effectiveness. (Ibid., p. 195)

Garicano and Heaton's overall conclusion is thus that:

> [W]hile the effects of general information technology on crime fighting and deterrence are statistically insignificant (in spite of our large samples), this effect becomes relatively large when IT adoption is undertaken as part of a whole package of organizational changes. That is, our results are a clear endorsement of what we have called here the complementarity hypothesis, and suggest that police departments, like firms, are likely to only enjoy the benefits of computerization when they identify the specific ways the new information and data availabilities interact with existing organizational practices and make adjustments accordingly. (Ibid., p. 196)

Overall, this is an impressive study that reaches carefully based conclusions, and takes issue with previous work (e.g., Levitt, 2004). There are two main issues. First, as is to be expected in a long-standing survey-based dataset, the information on ICT use in police departments is rather old-fashioned. It tracks well the onset of automation and the spread of initial computer use to new activity streams within police departments, especially in small departments. But the information seems of declining relevance

for the assessment of practices in medium to large police forces, especially for the modern period with the bureaucratic development of websites, the primacy of online information, the use of online transactions and the adoption of social media in community-building.

Second, unlike Bloom et al., Garicano and Heaton do not have any directly accessed or purpose-designed information on management practices. By looking at survey questions included in their dataset that can be plausibly related to the COMPSTAT paradigm, they are able to partly overcome this barrier and to uncover a significant composite 'IT plus management change' effect – one that seems consistent with a great deal of research reaching similar conclusions for private sector firms. We carry over these key lessons to our own work in Chapter 8, which looks at hospital trusts in England. We seek to capture well-based information on trusts' use of modern ICT approaches, and their adoption of good practice management strategies.

Using New, Web-based Research Methods

Using new web-based research methods is the key step that allows us to avoid Bloom et al.'s dependence on responses from just one or two individuals per hospital trust, and to track more modern ICT use than that assessed by Garicano and Heaton. Across the social sciences, e-methods that exploit the availability of a great deal of new and different kinds of information in digital formats are only just beginning to develop. For instance, one can still scan most social science methods textbooks in vain for any guidance on these issues. Yet digital and web-based information sources are accessible, efficient, and at times offer major advances on previously available information. Websites provide new landscapes and sources for research. Especially given the transparency and accountability requirements that apply within the public sector, the websites of local government agencies and central government departments offer a uniquely objective picture of organization strategies and activity. Web-logs, list-servs, e-mails, usage statistics and Twitter followers shed interesting light on where an organization's internal activities and external interactions are concentrated. Empirical sociologists recognize that both in business and in government the accumulation of massive amounts of transactional and administrative data has more or less made obsolescent the social sciences' previous primary reliance on survey data (Savage and Burrows, 2007 and 2009). In particular, the volume, comprehensive inclusiveness, objectivity and non-reactive qualities of transactional datasets have many advantages for analysts, decisively shifting the focus of research away from survey-

Table 7.1 The four different ways that an organization's underlying pattern of activities and their online presence can be related

Organization Presents Itself Online as Doing	Organization is Actually Doing	
	A lot	Not much
A lot	1 Web census analysis correctly identifies high activity situation	2 Facade activity
Not much	4 Organizations with 'stealth' activities	3 Web census analysis correctly identifies low activity situation

based, small sample datasets to massive datasets that are transactional censuses, not sample studies (Dunleavy, 2010a).

The argument for this way of researching starts from the premises that in advanced industrial countries every salient public sector organization is now on the internet in some form, and so their websites, aspects of their transactional systems, and use of social media, along with forums, blogs and other elements are largely open for inspection. Their websites at least (but not intranets) can also be systematically crawled for information, although this needs to be done slowly because anti-virus software will otherwise repel attempts to fast-crawl a site (see Escher et al., 2006; Petricek et al., 2006). Sophisticated network techniques can then be used to analyse the 'graph structure' in the web data. Essentially, how the organization communicates with citizens, customers, businesses, civil society or other government or political bodies is fully open for analysis by political scientists.

Of course, what the organization *says online that it is doing*, and what the organization *actually is doing* may vary. Because of this problem, many social scientists have prematurely dismissed websites as useful sources of information, portraying them as simply public relations 'fronts' for organizations. There is a serious potential difficulty here, but in fact it is relatively easily managed in researching government sector organizations, as Table 7.1 shows. For online methods to work it is important that the vast bulk (say 95 per cent) of all organizational situations will be covered by cells 1 and 3, where an organization is either doing a lot or a little, and its web presence (when critically assessed using online research approaches) accurately reveals that situation.

The two other possibilities here would represent problems, if they are widespread and cannot be detected. 'Facade' activity (cell 2) occurs when a lively online presence masks low levels of underlying ('real')

activity (or activity of a different kind) by the organization. This situation sounds plausible, yet it is actually much harder for an organization to sustain than might appear at first sight. Virtually all significant government organizations' web operations are now salient, complex and interlocked with their fundamental transactions systems and ways of working and doing business. All the chapters in Part I noted that essential business processes in government now operate via the internet – the time is long past when websites just held press releases aimed at creating a public relations gloss. Websites are too expensive to maintain properly simply for propaganda purposes. And facade content is anyway clearly visible for critical researchers using modern methods (see Table 7.2). Indeed, any intelligent citizen browsing a facade website can quickly detect it. The whole concept of 'digital era governance' stresses that, increasingly, government bureaucracies are *becoming* their websites, so that the organizational socio-technical system is increasingly manifest on the web. Indeed, it has to be completely manifest or else modern, pared-down systems of risk-adjusted administration will collapse (Dunleavy et al., 2008).

The second problem in Table 7.1 concerns covering organizations with large-scale 'stealth' activities (cell 4). They are delivering public services, or doing things politically, but they are not telling citizens or talking about what they do on the web. Again this sounds possible, yet who exactly are these bodies? Certainly this is irrational behaviour for any public service bureaucracy in an advanced industrial country that is citizen-facing or business-facing. It is also counterproductive for most interest groups, civil society organizations and parties, to be implementing activities yet masking this from the public and society. Cell 4 is far more typical of some kinds of companies, especially those providing commercially confidential services. Only a few special purpose government agencies (such as intelligence services and defence agencies), and their opponents in terrorist organizations, may actually have critically important classes of activity shielded from web revelation. Even police services and foreign affairs ministries must now increasingly operate on the internet, or risk being marginalized from society's key information networks (Escher et al., 2006; Hood and Margetts, 2007). Similarly, even many modern terrorist movements rely extensively on online sites to raise funds, maintain communications, distribute ideological memes and provide for decentred patterns of cell organization (Burke, 2007, p. 39).

Not only does web-based research use digital information as a new source of data, but it also entails adopting new methods for assembling and critically analysing information. There is a paradigm shift in research approaches, a switch away from 'reactive' and 'obtrusive' methods (such

as sending out surveys where people are asked explicitly to give all the information collected, or to react to questions or propositions devised solely by the researchers). Instead, analysis moves towards 'non-reactive' and 'unobtrusive' measures, where information is collected without the interposition or even awareness of research subjects (the people or organizations being studied), and without the biases that can occur from researchers posing questions in ways that skew responses. The new *non-reactive measures* minimize the effect of question biases, peer influences, subjects giving 'improving' responses that conform to public or cultural norms, or delivering what they consider that higher tiers of government or the researchers themselves want to hear. Table 7.2 shows how the two broad approaches compare in several ways.

Non-reactive approaches relating to websites are also cheap to implement and facilitate comprehensive coverage of all agencies in a given category, rather than having to rely on sub-sets of data that cover only parts of the populations involved. Why sample when you can conduct a comprehensive census? Why worry about many aspects of conventional statistical significance if you can include the whole population in your datasets from the outset (also avoiding all missing case problems)? Why base analysis on a handful of cases (left largely unsituated in the wider field of all similar organizations) when you can cover them all, in detail? This basic shift of approach can be easily varied and extended in numerous ways – for example, using external or internal search engines and specialist media tracker sites to track the foci of memes in macro-content through their incidence in discourses; crawling websites for in-links and out-links to other websites and organizations (Escher et al., 2006); and using modern networking analysis to track influences (Cho and Fowler, 2007; Christakis and Fowler, 2008, 2010).

Our key approach in Chapter 8 relies on trawling systematically through the websites of organizations, and recording in great detail a large series of objectively recordable pieces of information that the website reveals about the organization. It is very important that the coding of items should be as objective as possible, and be as little open to misinterpretation as possible, otherwise researcher biases could creep in. This would vitiate some of the key advantages of using non-reactive methods, which include avoiding the need for personal judgements by respondents or researchers, and generating results that are completely replicable by other observers. Hence web-census methods need to focus on recording the answers to unambiguous questions that either require 'Yes/No' or 'Present/Not present' dichotomous codes, without the researcher needing to make complex, qualitative judgements. Examples of the single, unambiguous questions needed are:

Table 7.2 Comparing reactive surveys of organizations and non-reactive, web-census methods (WCM)

	Surveys of Organizations	Web-census Methods (WCM)
Coverage	A statistically representative sample, covering a fraction of the whole population of organizations	The whole population of organizations
Instrument defined by	Researchers define a strictly limited number of questions. Question wording effects extensively condition subjects' responses. Any incorrect or inappropriate single question wordings can contaminate significant sections of analysis and results	Researchers identify a large number (several dozen to hundreds) of discrete items to be coded as present or not. Items are structured and weighted to tap theoretically relevant dimensions. Any single incorrect item has a tiny impact on overall indices
Type of methods approach	*Reactive methods* (surveys, interviews) – those contacted may report erroneously, edit their responses or misrepresent situations	*Non-reactive methods* – items are coded as objectively present/absent in the organizations' websites, using simple dichotomies
Researcher–subject interactions	Obtrusive – respondents know the study is underway and the precise content of its questions	Unobtrusive – organizations need not be alerted that a study of them is underway
Costs	Substantial	Low
Key 'meaning' problems	Responses may be artefacts of the questions asked. Responses are a poor guide to these people's actual behaviour. The effects of interviewer and coder judgements may be hard to spot or control for	Organizational behaviours are established, but the salience and meaning of items coded may be disputed (at both an individual and an aggregate level).
Key problems with interpreting the information gathered	Who exactly in the organization completes and returns surveys varies a lot, and may not be known. The 'authority' status of the actual respondents is typically unclear, along with how far they consulted others	1 Controlling for 'facade' activity – which shows up clearly in well-designed coding frames 2 'Stealth' activities that are not detectable on organizations' websites (not likely to be a problem with *normal* government or civil society organizations?)
Key technical problems with datasets	Small sample sizes. Extensive non-response. Extensive missing data problems in achieved responses. Mistakes cannot be post-corrected by researchers without going back to respondents	Complete returns are always achieved, without missing data or non-response problems. Mistakes and miscodings can be post-corrected by researchers

- Is a copy of the agency's Annual Report easily findable on the site?
- Does the agency website include a link to the national ombudsman?
- Does the agency website include a link to a next-tier-up appeals body?
- Does the agency provide information for prospective staff thinking of moving into the local area?
- Does the agency provide users with any online guidance on how to make a complaint?
- Does the agency website use pictures/videos/social media?
- Does the agency have a Twitter page? A Facebook page?

(Some terms here may need explicit standardization, for example, in the first question here, 'easily findable' means that an experienced web researcher having spent an hour on the agency website, could not find the Annual Report online.)

A necessary feature of a web-census approach is that questions are reductionist – they focus on small, precisely codable characteristics of the organization's online presence. Taken on their own, in themselves, no single one of these indicators ever means very much. However, the method is holistic. It works by covering dozens of these small characteristics, whose cumulative presence or absence clearly does build up a picture and have meaning. Hence researchers must choose multiple small indices that are well adapted to assessing deeper-lying agency characteristics, and which can be simply aggregated into overall scores for each agency. This is the approach we deploy in Chapter 8 for characterizing the management practices and extent of ICT use across English hospital trusts in a manner that is fully independent from considerations of their overall performance, or of how they are ranked by health service inspectors or regulators.

Conclusions

Because multiple local providers play key roles in delivering most major public services in welfare states, we can compare across providers and also deploy more sophisticated parametric and non-parametric methods. Yet so far the insights derived from doing so have not been particularly illuminating, partly because of the importance of quality adjustments in complex services and the greater difficulties of getting large-scale data on intangible quality variations. Too many analyses have also focused on easily quantifiable but relatively remote influences on the productivity performance of local providers (such as the sizes of agencies or area populations), while not covering factors that in theoretical terms are likely to have the most immediate consequences for productivity change. Modern,

web-based and non-reactive methods can prove helpful in expanding the range of explanatory variables to include factors that seem likely to really matter for productivity – such as management practices and overall management quality, and how far local agencies make full use of modern ICT in delivering services