SERC POLICY PAPER 12

Evaluating Spatial Policies

Steve Gibbons (SERC, What Works Centre and LSE)
Max Nathan (SERC, What Works Centre and NIESR)
Henry G. Overman (SERC, What Works Centre and LSE)

March 2014

This work is part of the research programme of the independent UK Spatial Economics Research Centre funded by the Economic and Social Research Council (ESRC), Department for Business, Innovation and Skills (BIS) and the Welsh Government. The support of the funders is acknowledged. The views expressed are those of the authors and do not represent the views of the funders.

SERC Policy Paper 12

Evaluating Spatial Policies

Steve Gibbons*, Max Nathan and Henry G. Overman***

March 2014

*SERC, What Works Centre and LSE
School of Economics
**SERC, What Works Centre, LSE and NIESR

This note summarises evidence emerging from SERC research (www.spatialeconomics.ac.uk) as well as from the wider urban economics literature. SERC regularly publishes policy papers which are freely available from our website. SERC blog: http://spatial-economics.blogspot.com/. Follow us on twitter: @lse_serc

Abstract

In most countries economic prosperity is very unevenly distributed across space: regions, cities and neighbourhoods seem to be very unequal, whether we look at average earnings, employment, education or almost any other socio-economic outcome. Regional, urban and neighbourhood policies are often based on concerns about these kinds of disparities, and reducing such disparities is a key policy objective in many countries. This paper considers the role of empirical analysis in informing the development of these policies. It is particularly concerned with issues arising in the quantitative evaluation of the impact of policy, the major barriers to more effective evaluation and how these might be addressed in future.

1 / Introduction

In most countries economic prosperity is very unevenly distributed across space. Regions, cities and neighbourhoods seem to be very unequal. This is true if we look at average earnings, employment, education or almost any other socio-economic outcome. Regional, urban and neighbourhood policies are often based on concerns about these kinds of disparities, and reducing such disparities is a key policy objective in many countries. This paper considers the role of empirical analysis in informing the development of these policies. It is particularly concerned with issues arising in the quantitative evaluation of the impact of policy, the major barriers to more effective evaluation and how these might be addressed in future. It draws on a number of academic pieces (particularly Overman, 2012) as well as a recent report for the National Audit Office (Gibbons, McNally and Overman 2013).

Policy makers look to quantitative empirical analysis to do (at least) three things: describe the problem, assess the underlying causes and evaluate the alternative policy responses. Process evaluation, by contrast, seeks to explore and assess the ways public policies are designed and implemented, often using qualitative methods. Process evaluation and qualitative methods are often valuable tools – both in their own right, and as a complement to quantitative approaches. For example, when we are not clear about the factors underlying a result – the 'microfoundations', to use economic language – techniques such as semi-structured interviews or participant observation can help uncover what is going on. However, such approaches are not the focus of this paper: we concentrate on quantitative methodologies.

Analysts working with quantitative methods face a number of barriers in addressing each of the three tasks set out above. First, data availability hampers the provision of basic statistics, as well as constraining further empirical analysis. Second, identifying causal mechanisms, including the impact of particular polices, is difficult. Third, there are a number of practical barriers to putting identification of causal effects at the heart of policy evaluation. As a result, on well-accepted 'standards of evidence' much policy evaluation fails to identify the causal impact of policies (often despite claims to the contrary).

These issues are especially prevalent in spatial analysis. On all three dimensions, and particularly with respect to policy evaluation, research analysing spatial data often falls short of the standards set by current research in other policy areas – for example in clinical treatments and aspects of active labour market policy or international development. This partly reflects the particular difficulties inherent to spatial analysis, but also stems from a failure – by both researchers and policymakers – to adopt methodological developments that might improve both analysis and evaluation.

The rest of this paper is structured as follows. The next section focuses on the availability of data. The two subsequent sections deal with questions of causality and the evaluation of the impact of spatial policies, while a final section briefly concludes. Section 3 focuses on methods and contains material that will be more readily accessible to those with an analytical background. It can be skipped without (much) loss of continuity for those more interested in the broader questions of the role of analysis and evaluation in informing the development of spatial policy.

2 / Data

The most common data availability problem arises from the lack of appropriate data available at the appropriate spatial scale. Unfortunately, for many spatial issues the correct unit of analysis is difficult to define. City or regional boundaries are often administrative creations and for many outcomes, e.g. economic growth, provide a poor substitute for properly defined functional economic areas (see Cheshire and Magrini, 2009, for further discussion). The problem is not simply one of measurement – results may differ substantially as a researcher changes the unit of analysis. This is the so-called Modifiable Unit Area Problem (or MUAP). Sometimes, the difference in results may be valid – for example, a particular policy may have different impacts at the neighbourhood, city and national level. In other situations, however, using the ‘wrong’ unit of analysis – e.g. analysing the impact of local economic development policies using administrative, rather than functional, boundaries – may generate statistical biases that lead to misleading conclusions.

Even when data is available at the appropriate spatial scale, there may be considerable measurement error present in the data. Of course, measurement error is often present in non-spatial data, but the problem can be more pronounced for spatial data because all of the standard problems occur (e.g. is employment correctly defined, measured and recorded) but there is an additional problem of geographical miscoding (e.g. where observations are assigned to neighbouring geographical units because the data on geographical location is inaccurate). These measurement problems become worse as the spatial resolution of the data increases, because any absolute measurement error means greater relative error at smaller scales.

Sampled data creates an additional problem at small scales: For a given sampling rate, smaller spatial scales reduce the average sample size for each spatial unit. Increasing the size of areas and hence the number of observations in each spatial unit reduces problems caused by measurement error (either arising from sampling, or uncertainties about the exact location of boundaries), but can lead to the Modifiable Unit Area Problem if the correct unit of analysis is smaller than the one ultimately used in the analysis.

For these reasons, poor quality spatial data often create substantial problems in terms of generating simple descriptive statistics. As argued in Overman (2010) the increased use of Geographical Information Systems (GIS) is starting to solve many of these data problems. However, even if data availability becomes less of an issue for analysis, a lack of descriptive statistics for specific administrative units may continue to cause problems in policy development and evaluation. Even though these administrative units may be arbitrary from an analytical perspective they are very important to policy makers. For example, city leaders are often keen to know how their city ‘units’ are performing – both as an input into local decision making, but also because performance is often assessed by comparison to other administrative units (e.g. by central government in making funding decisions). As a result, even when such data might not be a particularly useful indicator of relative performance, or of the impact of a policy, it will still be of great interest to policy makers.

This inability to generate precise descriptive statistics for specific administrative units may leave practitioners feeling that they cannot understand the causes of spatial disparities or of the impact of a specific policy. Specifically, in the absence of descriptive statistics based on representative data a) covering populations or for b) large samples for specific places, it may seem that “no progress is possible”.

Happily, this need not be the case, because sampled data can be informative about the *causes* of spatial disparities even if it cannot ‘accurately’ describe *outcomes for specific places*. For example, a small sample of firms from a Local Authority (LA) may not give a precise estimate of mean productivity in that LA. However, a number of small representative samples of firms from different LAs *can* be used to provide precise estimates of the causal effects of a policy aimed at enhancing firm-level productivity and which is applied across those LAs. Technically, this is because the sampling errors in each LA sample cancel out when averaging across LAs to estimate the impact of the policy.

Of course, this problem of sampled data is not unique to spatial settings. But for some, sampling may appear to raise particular concerns for the evaluation of spatial policies. Suppose you believe that locations are unique in terms of characteristics or their response to a policy (or both). If this belief is strictly true, then there is no role for evidence in informing future policy, no possibility of learning from others; and no value in empirical scientific investigation generally, let alone analysis based on samples. Empirical analysis in all fields proceeds under the assumption that the cases under investigation and the responses to changes in policy or other factors are not completely unique. Cases share similarities in terms of their characteristics and expected response to policy changes. As discussed in the next section, this allows researchers to construct counterfactuals, which are predictions of what would have happened in the absence of policy. Applied work in spatial economics needs to take account of the heterogeneity between places when formulating appropriate research designs but this does not invalidate econometric analysis of the problem at hand.

For example, each city in Britain may appear unique, because it can be characterised in terms of a very large number of different factors including, for example, population, skills, industrial composition and location. While this will prevent researchers from ever finding two exactly comparable cities, we can still get some idea of the likely effect of policy in a specific city by averaging the effects of policy in other cities that share similar combinations of the constituent variables. We discuss these and other techniques in more detail in the next section. In short, while the lack of large samples for administrative spatial units may appear to be a barrier to better-informed policy making at all spatial scales, it need not be.

3 / Causality

Even when appropriate spatial data is available, much spatial econometric and statistical analysis has not paid sufficient consideration to the crucial issue of identification of causal effects. These problems are discussed in detail in Gibbons and Overman (2012). This neglect of causality has profound implications for our ability to understand the drivers of spatial disparities and the subsequent impact of policy designed to tackle these. To see why, we need to consider how empirical analysis might allow us to understand causality in the first place.

Causality is concerned with questions of the type 'if we change x what do we expect to happen to y?' Such questions are central to improved policy making and to the evaluation of particular policies. After all, policy usually seeks to change some policy (or condition) x to achieve some change in outcome y.

The fundamental challenge to answering these questions is that the determinants (x) are not usually randomly assigned. This is certainly the case for most policy interventions, when x (e.g., transport investment) will often be specifically set to (partially) reflect differences in the outcome of interest y (e.g., income). As a result, in real world data we lack the *counterfactual* that tells us what would have happened if x had been set at some different level. This is a problem because it is the comparison of actual outcomes to this counterfactual that identifies the causal impact of determinant x on outcome y. Applied economics has come a long way in its efforts to find credible and creative ways to answer such questions by constructing counterfactuals from observed data (Angrist and Pischke, 2009). Unfortunately, however, such methods have not been widely used in much analysis of spatial data and evaluation of spatial policy – as we shall explain.

In many countries, capacity constraints may limit the extent to which policy is informed by quantitative econometric analysis (which may be poorly communicated, is often technically demanding, and is not always well-covered in degree courses which tend to feed public policy cadres). Coupled with concerns about the underlying data, this can often lead to a focus on research which describes spatial disparities – rather than identifying the underlying causes and the impact of policy. As discussed above, this focus is then reinforced by the understandable political interest in assessing conditions in, and outcomes for, specific administrative ‘places’. As a result of all these factors, the tendency to conflate description with explanation may be particularly pronounced in the area of spatial policy making.

If data availability and the type of analysis undertaken represent significant barriers to informing policy making, a further barrier arises because even when these problems are recognised they can be very hard to address. The fundamental reason for this is that for many spatial economic phenomena and associated policies, suitable identification strategies (i.e. techniques for allowing us to figure out the causal impact of different factors, including specific policies) can be hard to develop.

The question of causality is of crucial importance in assessing the quality of evaluations, and hence their usefulness for informing policy makers. Estimates of the benefits of a policy are of limited use unless those benefits can be attributed, with a reasonable degree of certainty, to the implementation of the policy. Just as with the analysis of causality more generally, solving this problem requires the construction of a valid counterfactual. That is, we need to understand what would have happened to the treated individuals, firms, areas, (i.e. the targets or recipients of the policy) if they had not been treated? This is an outcome that is fundamentally unobservable. The way in which this counterfactual is constructed is the key element of programme evaluation design.

A standard approach to evaluation design is to create a comparator group of similar individuals not participating in, or for some reason not eligible, for the programme being evaluated. Outcomes can then be compared between the ‘treatment group’ (i.e. those affected by the policy) and the ‘control group’ (i.e. similar individuals not exposed to the policy). Typically, outcomes are then compared by looking at the differences in average post-policy outcomes. The assumption is that the post-policy outcomes in the control group provide an estimate of what would have happened to the treatment group in the absence of the policy. The challenge to effective programme evaluation is to ensure and demonstrate that this counterfactual assumption is plausible, given theoretical reasoning, the institutional context and the evidence in the data.

Box 1: The Scientific Maryland Scale

Level 1: Correlation of outcomes with presence or intensity of treatment, cross-sectional comparisons of treated groups with untreated groups, or other cross-sectional methods in which there is no attempt to establish a counterfactual. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups.

Level 2: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention (‘before and after’ study). No comparison group used to provide a counterfactual, or a comparator group is used but this is not chosen to be similar to the treatment group, nor demonstrated to be similar (e.g. national averages used as comparison for policy intervention in a specific area). No, or inappropriate, control variables used in statistical analysis to adjust for differences between treated and untreated groups.

Level 3: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (e.g. difference in difference). Some justification given to choice of comparator group that is potentially similar to the treatment group. Evidence presented on comparability of treatment and control groups but these groups are poorly balanced on pre-treatment characteristics. Control variables may be used to adjust for difference between treated and untreated groups, but there are likely to be important uncontrolled differences remaining.

Level 4: Comparison of outcomes in treated group after an intervention, with outcomes in the treated group before the intervention, and a comparison group used to provide a counterfactual (i.e. difference in difference). Careful and credible justification provided for choice of a comparator group that is closely matched to the treatment group. Treatment and control groups are balanced on pre-treatment characteristics and extensive evidence presented on this comparability, with only minor or irrelevant differences remaining. Control variables (e.g. OLS or matching) or other statistical techniques (e.g. IV) may be used to adjust for potential differences between treated and untreated groups. Problems of attrition from sample and implications discussed but not necessarily corrected.

Level 5: This category is reserved for research designs that involve randomisation into treatment and control groups. Randomised control trials provide the definitive example, although other ‘natural experiment’ research designs that exploit plausibly random variation in treatment may fall in this category. Extensive evidence provided on comparability of treatment and control groups, showing no significant differences in terms of levels or trends. Control variables may be used to adjust for treatment and control group differences, but this adjustment should not have a large impact on the main results. Attention paid to problems of selective attrition from randomly assigned groups, which is shown to be of negligible importance.

Tools such as the Scientific Maryland which rank pieces of quantitative evidence according to their internal robustness, pay close attention to this key aspect. The scale was developed by Sherman et al (1997) as part of a review of evaluations of policies targeted at crime reduction (summarised in Sherman et al 1998).[1] The SMS is a numerical scale ranging from 1, for studies based on simple cross sectional correlations, to 5 for randomised control trials. The ranking implicitly indicates how effective the research design is in constructing a valid counterfactual for the policy intervention, and hence how reliably the estimated effects can be attributed to the policy in question.

The authors of the original scale did not provide a single precise definition of what type of study falls in each category, but provided a number of indicative descriptions. Box 1 summarises the SMS and gives our interpretation of each level. We discuss the methods in more detail below.

The 'gold standard' of evidence for quantitative analysis is the randomised control trial or RCT, in which cases are randomly assigned to treatment and control groups. Methods exploiting explicit randomisation of this type feature at the top of the most widely-accepted 'hierarchies of evidence', such as the SMS, where it is classed as level 5. Properly implemented, randomisation ensures that treatment and control groups are comparable, allowing the control group to serve as a valid comparison group and thus identifying the causal impact of the policy.

Where randomised control trials are not an option, there are various statistical techniques that can be used to address concerns that the treatment and control groups may not be comparable. This lack of comparability arises in situations in which there is selection into treatment on the basis of observable or unobservable characteristics. In such situations, any difference between treatment and control groups may be a result of these differences in characteristics and nothing to do with the causal effect of policy. A crucial first step in addressing these selection problems is to focus on the comparison of outcomes in treatment and control groups before and after the policy is implemented (difference-in-difference). Availability of panel data, which follows individuals, households, firms or areas through time, is a crucial pre-

[1] These two reports are available at https://www.ncjrs.gov/pdffiles1/Digitization/165366NCJRS.pdf and https://www.ncjrs.gov/pdffiles/171676.PDF.

condition for deploying this technique. The idea behind difference-in-difference is that the change in the control group outcome between the pre and post policy periods estimates the change that would have occurred in the treatment group if the policy had not been introduced (i.e. the counterfactual). Therefore, subtracting this change from the pre to post policy change in the outcome for the treatment group generates an estimate of the policy impact. The crucial assumption is that the outcomes for treatment and control groups would have followed the same trends over time in the absence of the policy. Although this assumption is ultimately untestable, some evidence can be gained by looking for differences in the pre-policy trends. It is not technically difficult or sophisticated to show pre-policy trends across treatment and control groups, if sufficient pre-policy data is collected. There are also variants in the difference-in-difference design that can control for pre-policy differences in trends.

Regression analysis can go some way in achieving this, by statistically ‘controlling’ for differences in characteristics between the treatment and control groups. In its basic OLS form, however, this method only controls for ‘observable’ characteristics on which researchers have data and imposes some potentially restrictive (‘functional form’) assumptions about how these observed characteristics affect outcomes (it assumes that the outcome is determined by a linear combination of the observable characteristics). In many evaluations, one of the most popular solutions to addressing the second of these concerns is matching. Matching (and closely related ‘synthetic control’ methods) involves pairing treatment units with control units or combinations of control units that have similar, or identical, observable characteristics (implying that only treatment and control units which have some overlap in the available characteristics – i.e. ‘common support’ – are used). Matching relaxes the assumption that the outcome is determined by a linear combination of the observable characteristics, and makes more explicit which group the estimates apply to: usually, each treatment unit is paired with a matching control unit, to provide estimates of the ‘average effect of the treatment on the treated’ (i.e. the effect for units which have the characteristics of the treatment group).

Both standard (OLS) regression analysis and matching rely on the assumption that observable characteristics (those available in the data) are sufficient to account for all differences between treatment and control units that are relevant to the potential outcome of the policy. If this (untestable) ‘selection on observables’ assumption is correct, and there are no differences between treatment and control groups in the mean unobserved characteristics that cause differences in outcomes, then the difference in the outcome variable between the treatment and control units can be attributed to the policy. If the assumption is violated, then there is a standard omitted variables problem – or ‘selection on unobservables’ -, and matching estimates, like standard OLS regression estimates, are biased. To partially address these concerns, good policy evaluation work usually implements ‘balancing’ tests. The aim of these tests is try to demonstrate that there are no statistically significant differences between the treatment and control groups in the mean observable characteristics, in the pre-treatment period (or no correlation between observables and the intensity of treatment). These tests are based on the idea that the degree of selection into treatment on a set of arbitrarily chosen observable factors that determine the outcome, is likely to be a good guide to the degree of selection on unobservable factors that also influence the outcome. Ultimately, however, the extent to which ‘selection on unobservables’ causes OLS or matching estimates to be biased is ultimately untestable on the basis of observable data. Concerns about selection on unobservables have led to the development and implementation of modern programme evaluation techniques which seek to control for unobservable factors for which the researcher has no data. Properly implemented, such techniques are more robust than methods that rely only on selection on observables. It is for this reason that these methods score level 4 on the SMS, while OLS and matching estimators (or other techniques that rely on selection only on observables) only score level three. It is to these more robust techniques that we now turn.

The fundamental idea underlying all of these approaches is that, in the absence of explicit randomisation, ‘quasi-experimental’ sources of randomisation may address selection on unobservables. These sources may occur as a result of institutional rules and processes, changes in these rules and processes or through environmental or other natural phenomena that result in some cases randomly receiving different amounts of treatment than others. All these research designs (for example, instrumental variables, control function approaches and Regression Discontinuity Design) are based on similar principles, using some theoretically informed assumptions about which factors can be considered exogenous, or random, and those that cannot. Justifying these assumptions is never easy, but it is increasingly possible to develop effective strategies on these principles as better data, and better methods of linking data (e.g. GIS) becomes available. Section 4 and Overman (2010) and Gibbons and Overman (2012) provide many concrete examples.

Instrumental variables methods are based on finding, through theoretical reasoning, some specific exogenous variable or variables (instruments) that predict treatment group assignment but not outcomes, so that some part of the assignment process is effectively random. The aim then is to estimate the causal effects of the treatment from the differences in outcomes between units that are assigned to treatment or control groups through variation in this instrument only. An example might be some anomaly in a policy allocation mechanism, which inadvertently results in some individuals being treated while other comparable individuals were not. Since policies are usually designed to avoid this kind of anomaly, finding instruments for policy evaluation can be quite tricky. However, examples might arise during trial or rollout of a programme, or because there are inevitable imperfections in the way policies are targeted to needs. The use of instrumental variables is equivalent to controlling, in a regression of outcomes on treatment, for factors which determine treatment assignment but are uncorrelated with the instrument. This approach is also the basis for 'control function' methods (e.g. the Heckman sample selection correction estimator), which do exactly that, but usually with some additional functional form assumptions imposed from theory.

In academic work, an increasingly popular policy evaluation technique that combines elements of 'matching' and difference-in-difference is a regression discontinuity design (RDD). In RDD, matching is based on some continuous assignment variable. The assignment variable is known to determine assignment of cases – e.g. individuals or areas - into treatment and control groups. For example, area incomes might determine whether or not an area receives assistance under an area targeted policy. Below a known threshold, areas remain directly unaffected by the policy. Above this threshold areas receive assistance under the policy. Comparing the mean outcomes in areas that are 'just above' and 'just below' this income threshold (and/or controlling for incomes in a very flexible way) provides estimates of the causal effect of the policy. The assumption is that the only thing that is different between areas that are closely matched in terms of incomes around the threshold is whether or not they are exposed to the policy. The need for a precisely defined assignment variable, or combination of variables, plus a known threshold, explains why policies that are assigned based on well-determined and documented rules using easily observable characteristics are more amenable to good evaluation. Illustrations of the RDD method are given in the next section.

4 / Policy evaluation in practice

Policy specific *outputs* (e.g., the number of workers trained or firms assisted) are increasingly well monitored by governments. In contrast, many government sponsored evaluations that look at *outcomes* do not use credible strategies to assess the causal impact of policy. By a causal estimate, the evaluation literature means an estimate of the expected difference between the outcome for 'treated' individuals, firms, areas, etc. (i.e. affected by the policy), and the outcome they would have experienced without it. In section 3, we outlined various ways in which we can identify the causal effect of policy. For a variety of reasons, it is often argued that such evaluation strategies cannot be applied to spatial policies. This is simply not the case. Identifying the causal effects of (aspects of) spatial policies using these techniques is feasible and increasingly common in the academic literature. Some concrete examples, mostly drawn from the US and UK, help demonstrate the possibilities.

The use of Randomised Control Trials is quite rare in the case of spatial policy, although is more common in other fields like development, education and labour markets (see, e.g, Banerjee and Duflo, 2009; DiNardo and Lee, 2010). Policy experiments are scarce for a number of reasons. One major barrier is ethical. Many academics are comfortable with randomization, or 'experiments', because they are willing to start with the assumption that policy will have no effect. In addition, even if the policy is expected to be beneficial, randomisation is seen as fair in the sense that assignment to treatment or control groups gives everyone the same chance of treatment. This is harder for anyone who starts with the assumption that policy will be beneficial and is uncomfortable with the idea of leaving the decision about treatment to the outcome of a lottery. More generally, randomisation generates ethical concerns that those most in need might not be treated, at least in the short term or that those treated may be inadvertently harmed by the policy. Clinical trials resolve this dilemma by incorporating early stopping boundaries, where a trial is stopped or results reported when evidence first emerges of large significant beneficial (or adverse) effects of treatment. Such procedures may be harder to adopt for many spatial policies given that evaluation will often occur sometime after treatment (either because data is not yet available, or because effects may take months or years to materialise). In some cases, harmful effects of spatial policy experiments might be hard to reverse – which is not usually the case in clinical trials. In addition, large scale field experiments, such as the Moving to Opportunity Programme are costly and still suffer from difficult to avoid design flaws (Katz et al 2001, Kling et al 2005, Sanbonmatsu et al 2012).

On the other hand, small scale experiments suffer from concerns about whether results generalise to other contexts ('external validity'). Concerns about location 'uniqueness' may reinforce perceived problems of external validity for the evaluation of spatial policy. As discussed above (in Section 2), however, it is easy to exaggerate the extent to which location uniqueness presents a particular problem for spatial analysis.

For all these reasons, it is hard to get policy makers to agree to experiments to answer many spatial questions (even when such experiments could be designed). Of course, not all spatial policies are amenable to RCT analyse. For example, it is hard to imagine that randomisation would be a sensible way to allocate much infrastructure spending. Even when randomisation might be possible, other considerations mean that policies will be targeted at specific individuals, areas or firms. In these situations, we need to consider alternative methodologies to assess the causal impact of policy.

4.1 / Policy evaluation: examples

We start with the example of Enterprise Zones (also known as Empowerment Zones in the US and referred to below as EZs). These spatially targeted policies aim to improve economic outcomes in deprived areas. To identify their causal effect we need to figure out what would have happened in these areas in the absence of intervention. One possible identification strategy is to compare these areas to other similar areas that were not targeted by the policy. For some reports, even this simple strategy would improve the quality of evaluations. From an academic perspective, however, such simple comparisons still remain problematic because they require very strong identifying assumptions, so the improvement may not be that great. Specifically, unless we have an exhaustive list of area characteristics that influence local economic outcomes, we should worry that some unobserved characteristic of areas drives both the decision to target the area *and* outcomes in that area. In this case, we would wrongly attribute changes in outcomes to the policy when, in fact, they are driven by unobservable area characteristics.

Much of the recent improvement in the evaluation of programme treatment effects has come from novel ways of addressing exactly this problem, combined with a refined understanding of how to interpret the resulting estimates. One possibility is to compare outcomes for areas that receive funding to areas that applied for, but did not receive, funding. This strategy has been used by Busso, Gregory and Kline (2013) in a recent evaluation of the US Empowerment Zone policy.[2] Such a strategy can be highly effective in removing the influence of unobservables that might bias estimates of policy impact due to non-random self-selection of areas into the application process. Clearly, on its own this does not resolve biases due to non-random selection amongst applicants by the authorities awarding the grants. However, additional information on the rules governing the second stage selection process amongst applicants would allow evaluators to control for these biases too. For example if restrictions to funding limit the number of areas treated, so that selection amongst the applicants is random, or based on some arbitrary criteria (like the design quality of the submission), then selection is less likely to be driven by unobservable characteristics of areas that are relevant to the economic outcomes against which the policy is being evaluated.

In the UK there are now 24 Enterprise Zones located across England, 13 of which are ‘wave two’ areas chosen by competition from 29 potential locations.[3] As with US Empowerment Zones, the 16 sites that lost in the competition may provide a reasonable control group for the 13 that won (depending on how decisions were made as just discussed). Comparing outcomes for the two groups will then tell us whether those that won the competition do better, and we may be willing to attribute this to the policy. Analysis could also compare those that entered the competition to similar areas that did not enter (to see whether those that entered somehow differ from those that do not). For these kind of strategies to achieve identification of the causal effect of the policy requires that, conditional on observable characteristics of areas, treatment is not correlated with any unobservable characteristic that directly influences the outcome of interest.

The timing of policy interventions may provide another possible source of identification. For example, if some EZs are given money or tax breaks before others they should start

[2] Kline and Moretti (forthcoming) also exploits exogenously determined participants / potential participants to identify policy treatment effects over several decades.

[3] 11 ‘Vanguard’ Zones were directly selected by Ministers [DCLG 2011, Enterprise Zones Prospectus]

improving before those given money later. If they do not, that raises questions about whether treatment caused any improvement or instead whether this was caused by some other factor (such as changes in the macro-economy). In practice in the UK, staggered announcements of the recent wave of EZs did not translate in to staggered timing of interventions (business rate discounts applied to all EZs at the same time). However, staggered implementation did happen with US Empowerment Zones, as discussed in Busso, Gregory and Kline (2013) who use a timing strategy to identify possible impacts. In the UK, Single Regeneration Budget expenditure provides an example involving six rounds of funding with timing of implementation varying by rounds. For such timing strategies to work, identification of the causal effect of the policy requires that, conditional on observable characteristics of areas, *timing* of treatment is not correlated with any unobservable characteristic that directly influences outcomes.

Even in situations where we cannot be sure that decisions to fund (or the timing of funding) are uncorrelated with all unobservable characteristics that directly influence outcomes, we may believe that this condition holds for *marginal* decisions. Imagine, for example, that the government makes its funding decisions on the basis of a ranking of projects from best to worst. Such detailed assessment of projects often occurs after a first round process has ruled out the weakest projects (so the sample of projects subject to the more detailed ranking may be those that make it through this first screening process). If the ranking of projects is available then this would allow the comparison of outcomes for otherwise similar areas that were just ‘above the bar’ (and so got treated) to outcomes for areas just ‘below the bar’ (who did not get treated). Sometimes the criteria for treatment will be based on some observable characteristic of areas rather than some ranking based on the quality of bids submitted to the programme under consideration. Then areas that just satisfy the criteria and so get treated can be compared to areas that just fail to satisfy the criteria and so do not get treated.

This is an application of the ‘regression discontinuity design’ discussed in the previous section. Examples of applications of this type of analysis to spatial economic policies include Baum-Snow and Marion (2009), and Becker, Egger and von Erlich (2010). As discussed further in Lee and Lemieux (2010), these regression discontinuity designs can only reliably identify the causal effect of a policy on areas that are close to the cut-off in the assignment variable that determines whether an area is treated by the policy or not. There is, therefore, some cost in terms of the extent to which estimated effects generalise to areas that are further away from the policy cut-off. This is sometimes characterised as involving a trade-off between internal and external validity (i.e. the researcher gets good estimates of the causal effect for areas around the threshold but it is not clear whether these would generalise to areas away from the threshold). Some policies, such as the UK's Local Enterprise Growth Initiative (LEGI), may use a combination of cut-off criteria and competition to decide who gets treated (from among those that are eligible). In these cases the probability of being treated under a policy does not change discretely at the cut-off, and 'Fuzzy' discontinuity designs must then be used to assess the efficacy of LEGI (see Elias and Overman, 2013).

Criscuolo, Martin, Overman and van Reenen (2012) use an instrumental variables approach to assess the effects of Regional Selective Assistance (RSA) in the UK, exploiting features of the programme. RSA is designed to raise manufacturing employment, and is targeted at firms in disadvantaged areas with low productivity and high unemployment. Since RSA often targets historically underperforming places and firms, using OLS to evaluate the policy would under-estimate its effects (because poor performers are effectively 'selected in' to the programme).[4] However, RSA eligibility is determined at EU level every seven years, based on state aid rules. These rules clearly predict RSA receipt, but are determined EU-wide and so unrelated to country-specific area trends that can influence firms' performance. The authors are then able to use changes in local area RSA eligibility as instruments for RSA itself. In this example, the effect of RSA is identified by comparing those who receive treatment to those who likely would have received treatment if their area had not become ineligible as a result of the map change.

So far there is nothing specifically *spatial* about these identification strategies (other than that the policy occurs in specific places); indeed, they have been widely used in other applied micro-economic literatures (particularly in the development, education and labour economics fields). However, the fact that the policy intervention occurs in specific places, these places have a geographical location and sometimes a specific boundary beyond which the policy no longer applies. Such boundaries provide a further source of discontinuity, which may be useful in achieving identification. Specifically we can use 'spatial differencing' to compare treated areas to *nearby* non-treated areas. If unobservable characteristics vary smoothly over

[4] Conversely, surveying grant recipients would likely over-estimate its effects, since those who benefitted the most would be most likely to respond to survey requests.

space, but the policy intervention does not, then such a comparison may help control for unobservable characteristics that affect both treatment and outcomes.

As with regular (non-spatial) discontinuity designs, the validity and interpretation of the resulting parameter estimates depend crucially on how the borders of treated areas are determined and what happens to the unobservable characteristics of areas at those borders. If unobservable characteristics vary continuously at the border then spatial differencing may give us the causal effect of the policy even if policy assignment is non-random.. Even if unobservable characteristics do not vary continuously at the border spatial differencing may still help if it eliminates larger spatial trends making it easier to find suitable instruments for the spatially differenced variables. See Duranton, Gobillon and Overman (2011) for further details and an application to the impact of UK business rates on employment. Dachis, Duranton and Turner (2012), Gibbons, McNally and Viarengo (2012) and Einio and Overman (2013) provide further examples.

A further complication arises when using spatial differencing if treatment effects spill-over geographical boundaries to impact non-treated areas. This spill-over might be positive (a ‘multiplier’ effect) or negative (‘displacement’). Regardless of the sign of the effect, if the interest is in the overall aggregate impact of the policy for an area that extends beyond the boundary of the treated zone (as it might be, for example, for EZs) such spill-overs significantly complicate interpretation of estimated coefficients. Specifically, in the presence of positive spill-overs estimates of the effect of policy are biased downwards and vice-versa for negative spill-overs. These issues are discussed further in Neumark and Kolko (2010), but the literature is only just beginning to grapple with the resulting complications.

4.2 / Improving policy evaluation

Official evaluations of government spatial policies (that is, those paid for and sponsored by government) usually make little, use of these kind of strategies to help identify the causal impact of policy. This significantly complicates policy discussions, because reports that are less careful about causality are often willing to make much stronger and broader claims about the impact of policy on the target population (and how that impact was achieved). In contrast, empirical research in the programme treatment effects/quasi-experimental tradition often makes narrower claims about whether the policy has a causal impact for a particular sub-section of the population, where this sub-section of the population depends on the research design used. As a result, policy makers face a difficult trade-off when trying to decide how to evaluate policies. Wide-ranging ‘evaluations’ that are less careful about causality *appear* to provide more information as an input in to the policy making process. Taken at face value, such evaluations allow policy makers to both assess value for money and make changes to policy, while appearing to take in to account evidence about the impact of the policy. Evaluation designs that adopt a focus on causality may appear stronger in terms of internal validity, but more limited in terms of their generalizability. However, it is surely the case that without internal validity – that is reliable causal estimates – there can be no external validity, so careful methods focussing on causal impacts are a necessary, if not sufficient, condition for successful policy evaluation.

As explained above, for researchers Randomised Control Trials represent the gold standard in terms of assessing the causal impacts of policy. To the extent that ethical, political or practical problems prevent the use of pure randomisation, Regression Discontinuity Design may provide a more acceptable mechanism for approximating the ‘pure’ experiment offered by RCT. To do this, the policy experiment must be explicitly designed to assign people to treatment and control according to some assignment variable decided in advance as part of the policy design process. Using RDD in this way at least allows for high confidence in the causal nature of effects identified around the cut-off. Even in cases where this does not prove possible evaluation can still be improved by applying the methods outlined above.

It is also interesting to note that in many circumstances government could get such analysis, particularly of RCT or RDD type designs, at little direct cost because this kind of evaluation has the potential to be published in top academic journals (c.f. a number of the references provided above). This publication potential acts as an incentive to academics to undertake the evaluations without seeking funding from the government departments concerned. Such ‘open evaluation’ will not work for all policies (because the degree of academic interest will usually depend on the extent to which the policy 'design' allows causal effects to be identified) but it could work for a proportion of them. In short, when appropriate, policy evaluation of this kind does not need to be big, expensive and centralised. Instead, it can be outsourced by using open evaluation in the academic (and wider non-governmental) community.

A major barrier to such an open approach to evaluation is, once again, the availability of data. But now the issue concerns the availability of information from official sources on the government policy to be evaluated. So a first step in moving to a more open evaluation model would be to require good information to be recorded at all stages of the policy making process. For example, when policy is implemented via bids for funding for spatially targeted projects, information is needed on whether selection of projects is competitive, how decisions are made, what is the location and timing of intended and actual expenditure, what types of expenditure (buildings, capital grants, training) are funded, and so on. As is clear from the earlier discussion, to deploy these robust assessment techniques, this information on bids needs to be available whether the bids are successful or not.

In practice, nearly all of this information will be available and processed when appraising the bids before a decision is made. The only additional costs involved arise from doing this in a consistent, well documented, manner and in making this data available. Recording all of this detail would involve a small amount of expenditure, but does take time at a point when officials are sometimes under pressure to make decisions and start spending money.

Unfortunately, it is arguable that costs of acquiring and recording the information are not the major barrier in terms of data availability for open evaluation. Assuming all this information (on the policy process and outcomes) is available, there is one remaining major barrier. Specifically, effective policy evaluation needs the government to make all this information available to researchers. For all kinds of reasons, governments remain reluctant to do this.

Of course, a genuine reason for resistance to transparency is that some of the information may be confidential (more so when it relates to individuals or firms than areas). Fortunately, governments and statistical agencies appear to be increasingly willing to find mechanisms for solving this specific problem. In the UK, for example, they do this by making data 'publicly' available in a secure data environment with controlled access and detailed disclosure rules (e.g. the Secure Lab at the ESRC funded UK Data Service). Again, there will be some cost to maintaining this data and providing access to it.

The final challenge in achieving more careful policy evaluation is officials' and Ministers' (understandable) desire for rapid results. To perform the kind of analysis discussed in this section requires data on the policy and for a range of outcome variables (e.g. firm performance or, employment) for an appropriate number of geographical areas. That outcome data is usually only available with a time lag of several years – which complicates the interaction between evaluation and policy formulation, because policy makers are often working on shorter time scales. But once the data becomes available, if the policy design is such as to interest academics, researchers will then spend many (unpaid) hours figuring out whether the policy in question had any causal impact on outcomes.

In short, with a little patience and transparency, open evaluation has the scope to significantly increase our understanding of the causal impact of government urban policy at very little (direct) financial cost. In addition, such evaluation can also increase our understanding of how the spatial economy functions. For example, evaluation of place-specific policies can tell us the extent to which other 'amenities' are likely to get capitalised into land values. Policy evaluation of transport projects can tell us whether or not market access (through the transport network) affects productivity. Looking at the impact of training policies can help increase our understanding of local labour markets. The literature is only just beginning to explore these issues, but experience from other fields suggest that we might learn a lot more from such an approach.

5 / Conclusions

In the researcher's ideal setting, public policies would be designed with evaluation in mind and would contain an element of randomisation in the way individuals were assigned treatment. As discussed above, however, there are significant political difficulties, as well as a number of practical and statistical considerations, that may prevent the use of randomization – especially in spatial policy contexts. In the absence of successful randomization, evaluations always face limitations on how certain they can be that estimates are causal, because it can be difficult to construct a credible control group if evaluation has not been considered at the time the policy was devised. For example, if a policy is implemented at a national level, it will often be difficult to construct a valid counterfactual. Similarly, if policy targets all 'poorly performing' individuals or areas, there may be no appropriate control group that is not subject to the policy. Another very difficult situation arises where policy is targeted on the basis of criteria that are known to the decision maker, but unknown (or only partially known) to those undertaking the evaluation. In this case, it is difficult to be sure that any control group is truly similar to the treatment group before the policy is introduced.

To some extent, this paper has been concerned with the ‘process’ by which econometric analysis and evaluation influence policies. It has considered a number of barriers that limit this influence – specifically in terms of the availability of data, the limitations of analysis and of the evaluation of government policy. But the paper also indirectly highlights a number of constraints facing policy makers at local and national level (both in the UK and in many other countries). Policy makers are accountable for the performance of particular places. This means that they need to be interested in data for administrative units even if they understand that these might not adequately capture how spatial disparities are developing and what, if any, impact policy is having. Where there is a lack of analytical capacity, this often exacerbates the problems caused by any disconnect between the data used for analysis and that used to assess the performance of different administrative units. In terms of the analysis, policy makers may perceive ethical or political problems with decision making processes, such as randomisation or competitive bidding, that many researchers advocate as ‘ideal’ for evaluation purposes. More careful evaluations call for upfront costs in terms of systematic data collection but only deliver results longer term once outcome data becomes available and analysis has been undertaken. Political imperatives, e.g. an incentive to show short term results, can easily over-ride the desire of officials to take a longer term view of the impact of the policies for which they are responsible.

Some of these issues stem from fundamental differences in goals between researchers and policy makers. Others are more easily addressed, and we have made a number of suggestions in this paper. First, collecting data for functional spatial units – such as metropolitan areas – can better align the spatial scales used by policy makers and analysts. Second, designing policies with clear, well-documented rules and easily observable recipient characteristics helps researchers use programme features to evaluate those policies. Competitive application processes, clear cut-off criteria, and carefully designed variations in the location or timing of interventions, for example, can all provide helpful bases for evaluation. Third, using such institutional features of policies to help improve both policy evaluation and the understanding of the causes of spatial disparities increases the relevance of academic research to the policy making community. Fourth, secure data services allow governments to share data in a way that maintains some control over exactly how that data is used. In turn, opening up public datasets allows for ‘open evaluation’: where the academic community can provide longer-term assessments of the impact of policy, even if policy makers attention remains focused on the short term.

One significant problem remains, however – the econometric strategies outlined here can often be hard to explain to those with little or no economics or statistical background. Even if policy analysts understand and wish to apply these techniques in decision-making, for example, it is not always easy to persuade elected leaders of their value, since the latter often lack specialist knowledge and may be more concerned with the political concerns outlined above. This suggests a fifth recommendation: better methods training and capacity-building for civil servants and local policymakers, both analysts and those involved in programme design.

Simply having evidence to hand is not sufficient to make good policy: policymakers need objectives, and principles to guide these. Conversely, purely belief or principle-driven policy making can be completely misguided and such approaches still win out over evidence-based policy making in many situations. Strong evidence is a necessary condition for effective policy making, and addressing the barriers above will help move us towards this. Addressing these problems also makes for good research and evaluation – regardless of any influence on policy.

References

Angrist, J. and J.S. Pischke (2009). Mostly Harmless Econometrics. Princeton, NJ: Princeton University Press.

Banerjee, A. and E. Duflo (2009). The Experimental approach to development economics. Annual Review of Economics 1: 151-178.

Baum-Snow, N. and J. Marion (2009). The effects of low income housing tax credit developments on neighbourhoods. Journal of Public Economics 93: 654-666.

Becker, S., P. Egger and M. von Erlich (2010). The effect of EU Structural Funds on regional performance. Journal of Public Economics 94 (9-10) 578-590.

Busso, M., J. Gregory, and P. Kline (2013). Assessing the Incidence and Efficiency of a Prominent Place Based Policy. American Economic Review, 103(2): 897-947.

Cheshire, P. and S. Magrini (2009). Urban growth drivers in a Europe of sticky people and implicit boundaries. Journal of Economic Geography 9: 85-115.

Criscuolo, C., R. Martin, H. G. Overman and J. van Reenen (2012). The causal effects of an industrial policy. Spatial Economics Research Centre Discussion Paper SERCDP0098. London, SERC.

Dachis, B, G. Duranton and M. Turner (2012). The effects of land transfer taxes on real estate markets: Evidence from a natural experiment in Toronto. Journal of Economic Geography 12 (2): 327-354.

DiNardo, J., and D. S. Lee (2010). Program evaluation and research designs in Ashenfelter, O. and D. Card (eds) Handbook of Labor Economics Vol 4. Elsevier.

Duranton, G., L. Gobillon and H. G. Overman (2011). Assessing the effects of local taxation using microgeographic data. Economic Journal 121: 1017–1046.

Einio, E. and H. G. Overman (2013). The Effects of Spatially Targeted Enterprise Initiatives: Evidence from UK LEGI. Processed, LSE.

Gibbons, S., S. McNally and H. G. Overman (2013). Review of government evaluations: A report for the UK National Audit Office. http://www.nao.org.uk/wp-content/uploads/2013/12/LSE-Review-of-selection-of-evaluations-with-appendices1.pdf.

Gibbons, S., S. McNally and M. Viarengo (2012). Does additional spending help urban schools? An evaluation using boundary discontinuities. IZA Discussion Paper 6281, IZA Bonn.

Gibbons, S. and H. G. Overman (2012). Mostly pointless spatial econometrics. Journal of Regional Science, 52, 2, pages 172–191.

Katz, L. F., J. R. Kling and J. Liebmann (2001). Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment. The Quarterly Journal of Economics 116(2): 607-654.

Kling, J. R., J. Ludwig and L. Katz (2005). Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment. The Quarterly Journal of Economics 120(1): 87-130.

Kline, P. and E. Moretti (Forthcoming). Local Economic Development, Agglomeration Economies and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority. The Quarterly Journal of Economics.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. Journal of Economic Literature 48: 281-355.

Neumark, D. and J. Kolko (2010). Do enterprise zones create jobs? Evidence from California's enterprise zone program. Journal of Urban Economics 68: 1-19.

Overman, H. G. (2010). GIS a job: What use Geographical Information Systems in Spatial Economics. Journal of Regional Science 50: 165-180.

Overman, H. G. (2012). Geographical Economics and Policy. In Fischer M M,Nijkamp P (Eds.) Handbook of Regional Science. Springer Heidelberg New York Dordrecht London

Sanbonmatsu L, Marvakov J, Potter N, Yang F, Adam E, Congdon WJ, Duncan G, Gennetian L, Katz LF, Kling JR, Kessler R, Lindau ST, Ludwig J, McDade T (2012). "The Long-Term Effects of Moving to Opportunity on Adult Health and Economic Self-Sufficiency." Cityscape 14(2): 109-136.

BIS Department for Business Innovation & Skills

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Llywodraeth Cynulliad Cymru
Welsh Assembly Government

Spatial Economics Research Centre (SERC)
London School of Economics
Houghton Street
London WC2A 2AE

Tel: 020 7852 3565
Fax: 020 7955 6848
Web: www.spatialeconomics.ac.uk

SERC is an independent research centre funded by the Economic and Social Research Council (ESRC), Department for Business Innovation and Skills (BIS) and the Welsh Government