Quantitative Evaluation Techniques for Regional Policies

(Chapter for the "2018 Handbook on regional growth and development theories" edited by Roberta Capello and Peter Nijkamp)

Augusto Cerqua[a] and Guido Pellegrini[b]

[a] Department of Economics and Quantitative Methods, University of Westminster

[b] Department of Social and Economic Sciences, Sapienza University of Rome

1. Introduction.

In recent years, the vision of what the essential factors for growth are and therefore the role of local policies has drastically changed. The importance of aspects such as human capital, innovation, agglomeration and institutions coupled with the diversified impacts of globalization, have drawn attention to the often-neglected role of space for growth and growth policies (Barca et al. 2012). Moreover, the presence of a wide and persistent inequality in income and joblessness among local areas, regions and countries, exacerbated by the Great Recession, has suggested a more important role for spatially targeted policies. Austin et al. (2018) indicate that place-based policies should be considered in this framework, because "social problems are increasingly linked to a lack of jobs rather than a lack of income" and "subsidizing job creation may be easier at the place level than at the person level". Barca et al. (2012) argue that "space matters and shapes the potential for development not only of territories, but, through externalities, of the individuals who live in them." Therefore, the place-based approach is more appropriate than a space-neutral sectoral approach if the geographical context matters, in terms of social, cultural, and institutional characteristics. These considerations have led to a new spread of place-based policies, often accompanied by skepticism with respect to their results from a significant group of economists and politicians (see, for instance, Glaeser and Gottlieb 2008). Indeed, "a fundamental concern is that spatially targeted policies may simply shift economic activity from one locality to another, with little impact on the aggregate level of output" (Kline and Moretti 2014). It is therefore not surprising that in recent years there has been a particular effort in the development of techniques capable of evaluating the effectiveness of territorial policies.

The multiplication of evaluation studies of place-based policies has mainly been favored by four factors. First, thanks to advances in causal inference literature, and a growing influence of this literature on regional economics, regional policy evaluators can now count on a wide and robust set of evaluation techniques based on the potential outcomes framework proposed by Neyman for randomized experiments and then generalized to observational studies by Rubin (1974). Over the last 15 years, two additional evaluation methodologies have been added to the toolbox of regional policy evaluators: regression discontinuity design (RDD) and the synthetic control method (SCM). Moreover, additional improvements have been proposed for well-established evaluation techniques such as matching and instrumental variable estimators. Second, the increasing availability of detailed data on social and economic phenomena at the spatial level, has allowed researchers to more accurately estimate the impact of regional policies. Third, the use of geographic information

systems (GIS) has increasingly become popular among regional economists. GIS helps regional policy evaluators to identify causal impacts in a more credible way than previously possible, for instance, by detecting the exact geographical boundaries of the areas targeted by the policy (e.g., Neumark and Kolko 2010). Lastly, the development of specific spatial tools developed for popular econometric packages has equipped researchers with powerful tools to analyze the spatial effects of regional policies.

In this survey of place-based policy evaluation techniques, we have chosen to consider only methodologies and studies based on the counterfactual approach. The reason is that we are convinced that to identify the effects of a policy we need a causal model, and the counterfactual approach is the most widely used and convincing approach in this field.[1] The counterfactual approach, typical of program evaluation literature, attempts to compare what actually happened with what would have happened in the absence of the treatment. As each unit can be exposed or not exposed to the treatment (see Holland 1986), the researcher is bound to compare treated units with distinct untreated units. This approach derives from the potential outcomes framework (see Rubin 1974) where pairs of outcomes are defined for the same unit given different levels of exposure to the treatment, with the researcher only observing the potential outcome corresponding to the level of treatment received. Models are developed for the pair of potential outcomes rather than solely for the observed outcome. The potential outcomes framework has a number of advantages over a framework based directly on realized outcomes: i) it allows one to define causal effects before specifying the assignment mechanism, and without making functional form or distributional assumptions; ii) it forces the researcher to think about scenarios under which each outcome could be observed, that is, to consider the kinds of experiments that could reveal the causal effects; iii) it allows formulation of probabilistic assumptions in terms of potentially observable variables, rather than in terms of unobserved components; iv) it separates the modeling of the potential outcomes from that of the assignment mechanism. Of particular importance in Rubin's approach is the relationship between treatment assignment and the potential outcomes (Imbens and Wooldridge 2009). The simplest case for analysis is random assignment of the treatment, which ensures that there are no systematic differences between the treatment and control groups before treatment assignment. This implies that any observed differences in outcomes following the treatment can then be attributed to the treatment itself, rather than to selection bias. Therefore, it is straightforward to obtain estimators for the average effect of the treatment. Randomized experiments have been used in some areas in economics but hardly ever in regional economics. This is why in this survey we will focus on observational studies.

In general, there are specific characteristics of place-based policies that necessarily require a particular specification or even a specific adaptation of counterfactual techniques. First, of course, even though it has not always been obvious in the literature, the assessment must explicitly take into account that these policies act in space. The space is mainly used in two ways. On the one hand, it is represented with a territorial, administrative or economic grid, or through spatial coordinates; this poses various problems, related to the invariance of the results with respect to the chosen grid type and to the optimal level of disaggregation.[2] On the other hand, it can be an instrument to identify the model and the policy effects, as spatial discontinuities determine differences in the admissibility of policies and in the intensity of treatment. Second, probably the greatest difficulty is in the inherent endogeneity of place-based policies: the lower the development of a region, the greater the public

[1] There are also other evaluation approaches, for which we refer, for instance, to Holmes and Sieg (2015).
[2] These aspects are not explicitly dealt with in this survey. For further discussion on the choice of the appropriate spatial units, see Arbia (1989).

intervention. There is, therefore, a need to consider very carefully the presence of endogeneity, especially because in this field, as highlighted by Hanson and Rohlin (2017), "random assignment is nonexistent, treated areas are almost always chosen because of characteristics, and spatial spillovers are likely, all of which make an unbiased evaluation challenging", and then, "in the absence of random assignment, the choice for the evaluators becomes the group of interests in the absence of the program" and which method to use. Third, the presence of interferences between treated subjects, between untreated subjects and between both, leads to the need to adapt the Rubin causal model, as it is based on the Stable Unit Treatment Value Assumption (SUTVA), which postulates that there is no interference between units (see Rubin 1978). Furthermore, the presence of spillovers due to interference creates the need to define various measures of policy impact, considering direct, indirect, and total effects. Indeed, only by taking into account regional policy spillovers is it possible to evaluate the overall effectiveness of a local policy. Finally, the high heterogeneity of these policies is reflected in the presence of considerable heterogeneity in treatment. This requires a more sophisticated analysis from a methodological point of view.

The survey has been organized as follows: Section 2 presents the main types of policy evaluation techniques and the main parameters of interest. In Section 2.1 we examine the principal evaluation techniques to estimate the impact of the policy on all treated or on some subgroups of the treated units. In Section 2.2 we examine the evaluation frameworks developed to estimate the impact of regional policies on neighboring untreated units. In Section 3 we consider the presence of heterogeneity of the regional policy impact and how to deal with it, while Section 4 concludes the survey.

2. Policy evaluation techniques.

As argued earlier, random assignment is rarely feasible in the implementation of regional policies, due to ethical and/or practical concerns and applied researchers must rely on observational studies. When the policy assignment process is not random, treated units are nearly always self-selected or chosen on the basis of certain characteristics and this makes an unbiased impact evaluation a challenging task. Indeed, the estimation of the effect of treatment may be biased by the existence of observed and unobserved confounding factors which makes it difficult to select a suitable control group. This issue is intrinsic in the nature of place-based policies as they mainly target the least developed regions. In addition, as place-base policies are likely to engender spillover effects, it is even more challenging to gauge the overall impact of the policy. Even disregarding spillovers, the non-random treatment assignment processes used by policy makers makes it generally not possible to estimate some of the potential parameters of interest. For instance, without strong parametric assumptions and/or without assuming that treatment impacts across units are homogeneous, it is not possible to retrieve the average treatment effect (ATE) which gives the average impact over all units in the population of interest. Considering the realistic scenario of heterogeneous treatment impacts across units, two parameters might be estimated depending on the data available, the policy assignment rule and the researcher interest: i) the average treatment effect on the treated (ATT) which is the average treatment effect in the subgroup of the population observed to take the treatment; and ii) the local average treatment effect (LATE) which is the average treatment effect over a certain subpopulation of units. The aforementioned parameters are usually the main target of regional policy evaluation studies which adopt econometric approaches focusing on the estimation of the direct effects of policy (i.e., the effect on treated units), disregarding the estimation of

spillover effects.[3] In such studies, this can either mean that the researcher assumes that spatial spillovers are negligible or that even if they are present, the selection of the control group is carried out in a way that allows unbiased estimates of the direct effect of the policy (see, for instance, Busso et al. 2013). Evaluation methods which focus on the estimation of the direct effects of the policy are presented in Section 2.1. In Section 2.2 we will see that when researchers are interested in the estimation of spillover effects, other parameters might be of interest.

2.1. Evaluation methods based on the SUTVA

In this section we review the main quantitative evaluation techniques for regional policies based on the SUTVA. Such assumption postulates that there is no interference between units, that is, that the outcome of one unit is not affected by the treatment status of other units.[4] Therefore, the evaluation strategies based on the SUTVA do not model how units affect each other but assume that even if they interact, the treatment received by one or more of these units does not influence the future outcomes of the other interacting units used as control observations. This implies that spillover effects are ruled out by this assumption. Although many public policies can be credibly evaluated under the SUTVA, this is rarely valid for the evaluation of regional policies as we should expect them to engender spillover effects. This is why regional policy evaluators using the SUTVA should always motivate why they consider it to be a credible assumption in the context under analysis. In most cases researchers justify the adoption of the SUTVA by selecting a control group that is thought to be negligibly affected by the policy, so that the direct effects of the regional policy can be retrieved without bias. Another possibility is to use aggregated geographical areas as units of analysis assuming that they embed policy spillovers. Therefore, evaluation strategies based on the SUTVA implicitly focus on the main direct effects of the regional policies, disregarding or incorporating spillover effects. In the following four subsections, we will describe the basic features, the most recent developments and the most relevant applications in regional economics for the main evaluation strategies based on the SUTVA. In particular, we review the difference-in-differences methodology (Section 2.1.1), matching techniques (Section 2.1.2), regression discontinuity design (Section 2.1.3) and the instrumental variable strategy (Section 2.1.4). For a more technical account of the econometric approaches described here, see Imbens and Wooldridge (2009), Baum-Snow and Ferreira (2015) and Athey and Imbens (2017).

2.1.1. The difference-in-differences estimator and its recent developments

The difference-in-differences (DID) estimator exploits some naturally occurring event that has a certain group of units treated but keeps a similar group of units untreated. This method requires longitudinal data (at least two time periods) and consists in a before and after comparison across these groups of units. In general, the average increase over time in the control group is subtracted from the average increase over time in the treatment group. This double differencing removes biases in second period comparisons between the treatment and control group that could derive from permanent differences between these groups, as well as biases from comparisons over time in the treatment group that could be the result of time trends unrelated to the treatment (Imbens and Wooldridge 2009). The DID estimator delivers unbiased estimates of the ATT if the following

[3] Although it is possible to define other parameters of interest, such as the intention-to-treat (ITT) parameter which gives the average effect of being offered treatment, the ATT and the LATE are by far the most highly valued in the evaluation of regional policies.

[4] It also postulates that there exists only one version of the treatment, or, if there are multiple versions of the treatment, the effects of these different treatment versions are the same.

conditions are satisfied: i) the assignment process does not depend on temporary shocks;[5] ii) without the treatment, the trends of the outcomes relative to the treated group and the control group would have stayed unchanged. The latter condition is called parallel trend assumption.

This estimator is then based on very restrictive hypotheses. Researchers should attempt to mitigate identification concerns by showing that the parallel trend assumption, which cannot be directly tested, was satisfied in the period(s) before the policy implementation. Another possibility is to relax this assumption by checking for potential differences in pre-trends, as in a triple-difference approach using a pre-reform placebo DID.[6] In the evaluation of regional policies, DID is typically used when some cities or regions, experience a treatment, such as a policy change, while others do not. Although the DID estimator has been implemented since the early 1990s, it is still one of the most used approaches for evaluating local policies. Recent examples are Busso et al. (2013) and Mayer et al. (2017) who assess the incidence of urban enterprise zones; Jofre-Monseny et al. (2018) who estimate the impact of large plant closures on the local employment in the affected industry; Mayer and Trevien (2017) who study the causal impact of urban rail transport on firm location, employment and population growth and Ahlfeldt et al. (2015) who estimate agglomeration and dispersion forces in cities.

The synthetic control method (SCM) developed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) builds on the DID but moves away from using a simple average of control units. Indeed, the counterfactual is a weighted average of units, whose observable characteristics are similar to those of the treated unit but have not been exposed to the treatment. In addition, this data-driven procedure reduces discretion in the decision about what to include in the comparison group and allows the effects of unobservable confounders to vary with time. The main limit of the SCM is that it allows one only to look at one treated unit at a time. However, Xu (2017) recently proposed an approach which allows one to unify the SCM with linear fixed effects.[7] This generalizes the SCM to the case of multiple treated units and variable treatment periods, broadening the applicability of SCM also for regional scientists.

Over the last few years, several scholars have adopted the SCM to evaluate the impact of regional policies. For instance, Barone et al. (2016) and Di Cataldo (2017) used the SCM to analyze the economic effects of reducing transfers to lagging regions, while Rickman and Wang (2018) used it to test whether fiscal austerity policies enacted at state level stimulate growth in state economies and reduce their budget deficits. Another example is Castillo et al. (2017) who used the SCM to analyze the long-term impact on employment of a local tourism development policy.

2.1.2. Matching estimators

Matching methods are one of the most used approaches to estimate causal regional policy effects. They ex post mimic an experiment by matching each treated unit to one or more untreated units as similar as possible with respect to a given set of pretreatment variables X. Matching estimators[8] are

[5] For instance, evaluating with the DID a regional policy that targets areas hit by a temporary negative shock in employment and investment, it would deliver upwardly biased ATT estimates as these treated areas' means revert (an "Ashenfelter Dip" problem) (Ashenfelter 1978).

[6] In that case, the identifying assumption would be that there is no contemporaneous change in the differential trend between treated and control units.

[7] See Gobillon and Magnac (2016) for a detailed analysis of the properties of interactive fixed effects, SCM, and DID methods by Monte Carlo experiments.

[8] A large number of different matching estimators have been proposed in the literature. In this survey, we focus on some of the most used approaches and refer to Stuart (2010) for many more details on alternative matching methods such as the inverse probability weighting estimator.

appealing as they are non-parametric and easily interpretable. However, the dimensionality of the space of the matching variables can represent a serious limitation to the implementation of matching. Indeed, if there are a high number of covariates, it may be difficult to identify one or more untreated units to match with every treated unit. A popular alternative is to match on a function of the X: the probability of assignment given the set of characteristics X. This matching method is named propensity score matching (PSM) and was proposed by Rosenbaum and Rubin (1983). The propensity score facilitates the construction of matched sets with similar distributions of the covariates, without requiring close or exact matches on all of the individual variables (Stuart 2010). Matching estimators mainly rely on two crucial assumptions. First, the conditional independence assumption (CIA), i.e., it is assumed that all relevant differences between treated and untreated units are captured by observable attributes (selection on the observables). This implies that conditional on observed confounders, the treatment is as good as randomly assigned. Second, the common support assumption, i.e., every treated unit has at least one counterpart in the control group with the same or very similar observable characteristics. In case such assumptions are satisfied, the ATT is retrieved from the average difference between the treated and the untreated groups with the same values for the confounders.

In recent years, it has been highlighted that, to a certain extent, matching methods have been misapplied in the literature (see, for instance, Iacus et al. 2012) and a new matching estimator has been proposed: coarsened exact matching (CEM).[9] The idea of CEM is to temporarily coarsen each conditioning variable into substantively meaningful groups, which exactly match these coarsened data, and then retain only the original (uncoarsened) values of the matched data. If different numbers of treated and control units appear in various strata, the econometric model must weigh or adjust for the various stratum sizes. Generally, a weighted regression of the dependent variable on the covariates is adopted at the end of the matching procedure. Iacus et al. (2011) show that CEM dominates commonly used existing matching methods in its ability to reduce imbalance, model dependence, estimation error, bias, variance, mean square error and other criteria. Nonetheless, the inherent trade-off of matching is also reflected in CEM: larger bins (more coarsening) will result in fewer strata; fewer strata will result in more diverse observations within the same strata and, thus, a higher imbalance (Blackwell et al. 2009).

Matching estimators could be combined with the DID. This allows formulation of the main matching hypotheses with respect to the before-after evolution instead of levels. In fact, first-differencing outcomes with respect to a pre-policy period removes selection on the time-invariant unobservables (individual fixed effects and trend effects), while comparing the first-differentiated outcomes for treated units with those of observationally identical non-treated units removes selection on the observables. In other words, the matching DID represents an improvement over both matching and DID because it weakens the identifying assumption for matching by allowing non-observed time-invariant variables to influence performance. However, time-variant unobservables cannot be controlled for and after the matching DID procedure there might still be some residual selection bias.

When applying matching estimators, finding untreated units with characteristics similar to those of the treated units can be challenging. This is the greatest empirical challenge in the use of this method to evaluate regional policies, where policies are essentially endogenous and treatment is chosen according to the economic and social hardship of the units to be treated. A popular approach to overcome such an issue is to select a control group made up of units that applied for the program but were excluded or of eligible units that did not apply to receive the treatment (see Bernini and Pellegrini 2011)

[9] CEM belongs to the "monotonic imbalance bounding" class of matching methods. In this class, the balance between the treated and the control groups is chosen by ex-ante user choice rather than being discovered through the usual laborious process of checking after the fact, tweaking the method, and repeatedly reestimating (Blackwell et al. 2009).

Matching methods have been used extensively by regional economists. Recent applications concerned the study of the impact of capital subsidies (Moffat 2014; Andini and de Blasio 2016), enterprise zones (Reynolds and Rohlin 2015), the opening of a large new plant (Patrick 2016) and the local economy resilience to rare natural disasters (Bondonio and Greenbaum 2018).

2.1.3. Regression discontinuity design

Regression discontinuity design (RDD) is one of the most credible evaluation strategies for the estimation of the causal impact of regional policies. Although RDD has been around since the 1960s, this method was widely used in regional economics and other socio-economic disciplines only in the last decade, after that some influential papers, in particular Hahn et al. (2001) and Lee and Lemieux (2010), have provided a detailed account of the identification assumptions and the implementation of RDD and highlighted the potential for considering it as the non-experimental design closest to experimental design.

The basic idea behind RDD is that the probability of receiving the treatment changes discontinuously with an exogenous continuous variable, referred to as the forcing variable s. The forcing variable is often itself associated with the potential outcomes, but this association is assumed to be smooth. This means that the units close to the threshold but on different sides are otherwise comparable, so any difference in average outcomes between units just to one side or the other is interpreted as evidence of a causal effect of the treatment. In case s fully determines the assignment of the treatment on the basis of a threshold, s*, this approach is called "sharp" RDD. If the magnitude of the jump in probability of receiving the treatment at the threshold value is less than one, the approach is called "fuzzy" RDD.

In regional policy evaluations, the possibility of using RDD often arises from administrative decisions, where public resources are rationed and clear, transparent rules are used for the assignment of the treatment. The main limitation of this design is that it only retrieves the average causal effect in proximity of the threshold and this reduces the external validity of the estimates. In other words, the estimates might concern only a narrow subpopulation that may have different characteristics from the rest of the population of interest to the policy makers.[10]

RDD approaches have been adopted by Becker et al. (2010) and Pellegrini et al. (2013) to evaluate the effectiveness of the EU Cohesion Policy, by Freedman (2015) to examine the labor market impacts of tax incentives, by Cerqua and Pellegrini (2014) to estimate the causal effects of capital subsidies, by Bronzini and Iachini (2014) to estimate the effectiveness of an R&D subsidy program and by Crescenzi et al. (2018) to assess the impact of Smart Specialization Strategy programs.

In addition to exploiting thresholds generated by administrative decisions, regional economists can also exploit differences in policies across geographical borders as a source of randomness. In case a geographic boundary splits the areas into treated and control areas, researchers can compare treated units located on the treated side of the geographic border with nearby untreated units located on the opposite side of the geographic border. This identification strategy - referred to as spatial RDD or geographic RDD - is appealing because it controls for confounding unobservables that evolve smoothly over space. Indeed, locations separated by a regional border share the same geography, climate, access to transportation, agglomeration benefits, and access to specialized labor and supplies; the key feature that sets these locations apart is the difference in policies on the two sides of the border (Hagedorn et al. 2015). Location can be specified using two-dimensional RDD in the latitude-longitude space proposed by Dell (2010) or by using the scalar distance to the boundary such as the Euclidean distance or the travel distance by car.

[10] The external validity depends on the homogeneity of the characteristics of treated units throughout the entire population of treated units.

Keele and Titiunik (2015) highlight the fundamental differences between spatial RDD and standard RDD. First, an important assumption of spatial RDD is the so-called compound treatment irrelevance assumption. When studying treatment assignments that change discontinuously at a geographic border, it is common for multiple administrative or political borders to perfectly overlap. This means that potential outcomes might not only depend on the jump in the treatment of interest but also on other region-specific treatments. Therefore, the compound treatment irrelevance assumption states that the potential outcomes are only a function of the treatment of interest. Second, the continuity assumptions needed for identification will hold less often when applied to geography, because when discontinuities are geographic, agents may sort very precisely around the boundaries and undermine the validity of the design.[11]

Following the seminal paper by Holmes (1998), who used spatial RDD to disentangle the effects of state policies from other state specific characteristics, many other regional scholars employed a similar evaluation strategy. For instance, it was used by Giua (2017) to evaluate the effectiveness of the EU Cohesion Policy, by Jofre-Monseny (2014) to analyze the unintended effects on mobility of increasing unemployment protection in neighboring regions and by von Ehrlich and Seidel (2018) to investigate the long-term consequences of temporary regional transfers.

2.1.4. Instrumental variables

The instrumental variables (IV) approach deals directly with selection on the unobservables. The IV method requires the existence of at least one variable exclusive to the assignment rule, known as the instrument. Without any loss of generality, we consider the case of one instrument and one endogenous variable. Such an instrument must be correlated with the endogenous explanatory variable, conditional to the other covariates X[12] and it is supposed to affect only the eligibility to receive the treatment without having a direct impact on the outcomes of interest. The latter assumption is known as the exclusion restriction. It implies that the potential outcomes do not vary with the instrument and any difference in the mean observed outcomes of two groups of units differing only with respect to the instrument can only be due to consequent differences in the eligibility and composition of the treatment group with respect to potential gains from treatment. In general, the critical assumptions underlying IV are substantive and require subtle subject matter knowledge (Imbens 2014).

Depending on the assumptions of the homogeneity or the heterogeneity of the policy effects and on the specific application, the IV estimator allows one to retrieve the ITT, the ATT or the LATE parameters. For instance, with a binary instrument IV estimates recovers the LATE parameter, i.e., the average effect of the treatment for the sub-population whose behavior was influenced by the excluded instrument, conditional to X (Angrist et al. 1996).

IV has been used extensively in regional economics over the past 20 years. Recent applications include Ketterer and Rodriguez-Pose (2018) who assess whether institutions or geography prevail in driving economic growth, Koh et al. (2013) who study whether agglomeration rents are taxable for local governments, Castells-Quintana (2017) who evaluates the causal link between urban concentration and economic growth, Criscuolo et al. (2018) who estimate the causal effects of a business support policy and Filippetti et al. (2018) who assess regional disparities in the effect of training on employment.

[11] Readers can refer to Keele and Titiunik (2015) for a more detailed description of the assumptions and the implementation of the spatial RDD.

[12] If this correlation is strong, then the instrument is said to have a strong first stage. A weak correlation may provide misleading inferences about parameter estimates and standard errors.

Regional economists often use the IV approach to isolate sources of exogenous variation in local labor demand via the Bartik shift-share instrument (Bartik 1991). The idea is to isolate shifts in local labor demand that only come from national shocks in each sector of the economy, thereby purging potentially endogenous local demand shocks driving variation in employment or wages (Baum-Snow and Ferreira 2015). This means that the instrument captures the exogenous changes in local demand because nationwide changes do not reflect local economic conditions. The validity of the instrument requires that certain prerequisites are met. The first is that there is enough variability in the national shock by sector and sufficient variability at the territorial level of the sectoral structure. Another fundamental hypothesis is that the composition of employment is fully exogenous. Third, the instrument is highly dependent on the level of sector categorization used to break down the aggregate national shock.

Moretti (2010) used the Bartik instrument to estimate the local multipliers engendered by the creation of new jobs in manufacturing, while Faggio and Overman (2014) used the same approach to estimate the local multipliers engendered by the creation of new jobs in public employment. Cadena and Kovak (2016) used the Bartik instrument to identify migration responses to local labor demand shocks. Another example is Baum-Snow et al. (2017) who used the migration version of the Bartik instrument (see Card 2001) to investigate how urban railroad and highway configurations have influenced urban form in developing countries.

2.2. Evaluation methods not based on the SUTVA

The average difference in outcomes between treatment and control groups can be given a causal interpretation using the methods presented in Section 2.1 only in the case of the absence of interference. However, the SUTVA appears completely unrealistic in many evaluations of territorial policies, which often have the purpose of generating spillovers between treated and untreated units to engender local development. There are several reasons for interference among treated and non-treated units, such as neighborhood effects, network effects, and agglomeration effects. Angelucci and Di Maro (2016) identify the four main types of spillover effects: (1) externalities, (2) social interactions, (3) context equilibrium effects and (4) general equilibrium effects. The first two effects operate from the treated units to the untreated population: the interaction may depend on an indirect effect occurring in the physical or economic space that treated and untreated units share or depend on the social interaction between the treated and untreated units. The second two effects act indirectly through effects on the context factors, in particular rules and behavior, or through changes in the price system, generated by changes in supply and demand, which are important if local markets are not very open to the outside world.

The possibility of interference is higher when the target population is a subset of the local economy, loosely defined as the geographic unit or local institution within which the target population lives and operates (Angelucci and Di Maro 2016). For example, subsidized firms can use subsidies to crowd-out non-subsidized firms from the local market, or to employ the most qualified local workers coherently with the economic rationale of the regional policy (De Castris and Pellegrini 2012). Moreover, subsidies may also indirectly influence the local nontarget population, for example by affecting the interactions within the local network or by spurring imitation effects. Similar effects come with respect to interventions on human capital, that can crowd-out workers of neighboring areas. Clearly, the extent of spillovers depends on their characteristics and the structure of the markets.

The presence of spillover effects violates the SUTVA in all evaluation contexts, so even in randomized trials this creates two types of issues for policy evaluators. First, estimators commonly

used to retrieve the direct policy effects are likely biased, with the bias depending on the level of interference but also on the degree of association between individual and neighborhood treatments (Forastiere et al. 2018). Second, direct treatment effects are not sufficient to summarize the policy impact. Indeed, it is also crucial to estimate the indirect policy effects in order to assess the overall policy effects (direct policy effects + indirect policy effects).[13] Sobel (2006) shows that ignoring interference can lead to entirely wrong conclusions about the effectiveness of the treatment, leading to inappropriate policy recommendations and incorrect understanding of data-generating models.

To design an evaluation strategy that accounts for the presence of spillover effects, one needs to understand and identify which untreated units are in fact subject to spillovers. This theoretical analysis is fundamental, because it pinpoints the subset of untreated units which is most likely to be indirectly affected by a particular treatment, and therefore it allows identification of the possible restrictions that determine the identification of a causal model which takes into account spillovers. The most common evaluation strategy in this strand of literature exploits this theoretical information to split the untreated units in those indirectly affected by policy and those not affected by the policy. For instance, when individuals can be partitioned into groups, it is often plausible to assume that interference occurs within groups but not across groups. This assumption is referred to by Sobel (2006) as partial interference. Using such an assumption, in randomized trials it is possible to retrieve unbiased estimates of direct, indirect, total and overall policy effects. Halloran and Hudgens (2016) provide a clear exposition of this. The authors consider two clusters (e.g., groups or regions) of individuals. In the first cluster, a certain portion of individuals in the cluster is treated and the rest remains untreated. In the second group, no one in the cluster is treated. The direct effect of treatment in the first group is defined by comparing the average outcome when an individual is treated with the average outcome when an individual is not treated. The indirect effect is defined as a contrast between the average outcome when an individual is not treated in the first group compared with the second group. The total effect is defined by comparing the average outcome when an individual is treated in the first group to the average outcome when an individual is not treated in the second group. The overall effect is defined by the contrast in the average outcome in the entire first cluster compared to the average outcome of the entire second cluster. For causal inference when units are connected along more complicated network structures, see Eckles et al. 2017.

Although interesting, partial interference is often an unrealistic assumption in regional policy evaluations. Indeed, units do not necessarily aggregate in specific and identifiable groups wherein they interact only within those groups but not with units in other groups. Another possible approach is to use the proportion of treated units within a group as a measure of interaction between units. For example, Arpino and Mattei (2016) use such approach to evaluate the effects of state aids to companies in the presence of interference. For a more general framework which considers how to separate individual treatment effects from spillover, interaction, and general equilibrium effects, see Huber and Steinmayr (2017).

If the presence of interference is assumed to depend directly on the distance, geographical but also economical or reliant on a network, it is possible to use more complex types of restrictions, which constrain the effects of spillover to follow a certain spatial pattern. This approach is at the basis of spatial econometric models (see, among others, Anselin 2006; Arbia 2014), which use a spatial

[13] In addition, moving from randomized experiments to observational studies implies a further complication, i.e., the typical assumptions of the methods presented in Section 2.1 must be extended - say, to include the treatment of neighbors, and individual and neighborhood covariates - to guarantee identification and valid inference.

weight matrix to model the interactions between units.[14] Spatial econometrics extends traditional econometrics by considering the potential effects associated with the locational aspect of data. In such models, spatial dependence expands the information set to include information from neighboring units. Spatial econometricians generally are not particularly concerned about how these spatial effects are engendered, but rather they focus on how they can be captured. In addition, they do not primarily aim to estimate causal models but are more interested in presenting a spatial model which accurately describes the data generating processes and then to estimate the parameters of the model by nonlinear methods. Therefore, in this strand of literature questions of identification have been addressed by asking which spatial processes best fit the data and the notion of causality has been at most marginal. The limited consideration given to the estimation of credible causal policy parameters and some relevant identification issues have exposed the field of spatial econometrics to harsh critiques (see, for instance, Gibbons and Overman 2012) which have pushed spatial econometricians to a more consistent consideration of causal interpretation in recent years (see Vega and Elhorst 2015).

When combining counterfactual analysis with a spatial econometric model, the need for different measures of policy impact is evident, as the mechanisms leading from direct effects to indirect effects through the spatial weight matrix are explicit. In addition, in contrast with traditional analyses, the impact of policy intervention can be different in different locations. The main summary impact measures that can be calculated for each independent policy variable included in the model are the average direct effect, the average indirect effect and the average total effect (see LeSage and Pace 2009). In the most common spatial econometric models, spillovers arise as a result of impacts passing through neighboring observations and back to the observation itself[15] and the aforementioned summary impact measures take this feedback effect into account.

2.2.1. Some recent findings of this literature

The literature concerning the empirical applications of the models that relax the SUTVA and analyze the presence of spillovers is not vast. Some work has concerned analysis of enterprise-zone programs. For instance, Neumark and Kolko (2010) develop a method of precisely identifying enterprise-zone boundaries over time and find no evidence of employment spillovers looking at control areas located at an increasing distance from the subsidized enterprise zones. Ham et al. (2011) compute a triple difference estimate and find positive but statistically insignificant spillover effects on neighboring areas in terms of unemployment and poverty rate, but Hanson and Rohlin (2013) find negative spillover effects on neighboring areas in the number of establishments and employment. An analysis that takes into account how spillovers are modified according to geographic distance is that of Einiö and Overman (2016), who find negative spillover effects that quickly diminish in space. Concerning business incentive programs, Cerqua and Pellegrini (2017) consider the presence of interference among firms by assuming that spillovers are possible only within the same economic sector and within a certain geographic distance. They show that treated firms benefit from a large increase in employment, but such increase is partially determined to the detriment of the untreated firms.

[14] Although the proposal to place more weight on closer observations is widely accepted, the true spatial matrix is generally unknown (Vega and Elhorst 2015). Moreover, in some applications even relatively small perturbations in the spatial weights matrix will have salient consequences in the empirical results (Ward and Skrede Gleditsch 2008).

[15] The magnitude of this type of feedback depends on: (1) the position of the region in space (or in general in the connectivity structure), (2) the degree of connectivity among regions governed by the spatial weight matrix used in the model, (3) the parameter measuring the strength of spatial dependence, and (4) the magnitude of the coefficient estimates (LeSage and Pace 2009).

The literature concerning the analysis of public policies in a counterfactual context using spatial econometric models is even more limited. Dubé et al. (2014) use the spatial DID estimator, which extends the DID estimator allowing for spatially correlated treatments and local spatial interactions in treatment responses, to evaluate the effects of the development of a commuter rail transit system on housing prices. They find that even if the mean effect is positive and significant, some houses may experience lower property values compared to the houses experiencing no changes in environmental amenities. Chagas et al. (2012) use a spatial version of the PSM to identify the effect of sugarcane production on respiratory diseases. The effects on the surrounding areas are relevant, and almost as large as the effects on the producing areas. De Castris and Pellegrini (2012) implement an evaluation strategy based on a spatial econometric model which allows evaluation of the net spatial effects of capital subsidies. Under certain assumptions, such an approach disentangles spillovers engendered by the policy from those that cannot be attributed to the intervention. They find a modest spatial crowding out whereby subsidized regions attract employment and investment from neighboring areas.

3. Heterogeneity of regional policy impact

Most evaluations of regional policies aim to estimate an average causal impact of the policy on the treated or on a specific subset of the population of interest. In the absence of randomized experiments this is a challenging task and the methods presented in Section 2 have been developed to retrieve point estimates which are as accurate as possible. Nevertheless, policy makers are not only interested in the average impact of a policy but also on how the impact changes with respect to other parameters. In other words, policy makers would like to know the heterogeneity of the impact with respect to other covariates, intensity of treatment and even geographic features. In this section we present the most recent approaches to study the heterogeneity of the impact starting from the heterogeneity with respect to one or more variables.

The extension of the RDD proposed by Becker et al. (2013), called heterogeneous LATE, allows assessment of the treatment effect heterogeneity in the neighborhood of the threshold. This method allows one to estimate the LATE for different values of one or more covariates z different from the forcing variable s. The main assumption underlying the validity of this approach is that the variables z are uncorrelated with the error term in the outcome equation, conditional on s. Becker et al. (2013) applied the heterogeneous LATE to evaluate how the causal impact of the EU Cohesion Policy changes with respect to human capital and the quality of institutions. Percoco (2017) applied the same methodology to estimate how the causal impact of the EU Cohesion Policy changes with respect to the local economic structure.

Most regional policies are evaluated using one of the methods presented in Section 2, which apply to binary treatments. However, regional policies usually allocate different resources to treated units, i.e., they use continuous treatment instead of binary treatment. Since the early 2000s, researchers have tried to generalize evaluation methods to the case of continuous treatment and the main example of this is generalized propensity score (GPS) matching (Hirano and Imbens 2004; Imai and Van Dijk 2004). GPS matching is a nonparametric method to estimate treatment effects conditional on observable determinants of treatment intensity. Conditioning on the GPS, which represents the conditional density of the actual treatment given the observed covariates, and the actual treatment levels it is possible to estimate the average value of the dependent variable at different levels of the treatment and the GPS. This then allows the researcher to estimate the dose-response function after averaging such estimates of the average value of the dependent variable over the covariates X, while keeping the treatment level t fixed. The main assumption behind GPS is that once all relevant pre-treatment covariates are controlled for, the assignment of different intensities of treatment with the same GPS can be considered as good as random. This means that the adoption of a GPS approach presumes the availability of all the covariates which determined the assignment process of

different intensities of treatment. This assumption is stronger than the CIA assumption for the matching estimators discussed in Section 2.1.2 as it requires that the selection on the observables assumption holds for all possible values of a continuous treatment rather than only for the binary treatment. GPS matching has been adopted by regional policy evaluators in several circumstances. For instance, Bia and Mattei (2012) used it to estimate the dose-response function of capital subsidies, while Becker et al. (2012) adopted GPS matching to estimate the continuous relationship between European funds and economic growth. Recently, Cerqua and Pellegrini (2018) proposed an alternative approach to estimate the heterogeneity of the impact with respect to treatment intensity. The authors proposed to extend the RDD framework to the case of continuous treatment by exploiting as a source of local randomness the presence of sharp or fuzzy discontinuities in the assignment of continuous treatment. Therefore, this approach allows estimation of the average effect among units treated at different levels around the discontinuity. The main assumption behind this approach is that, after conditioning on the observable variables affecting treatment assignment and the intensity of treatment, treatment assignment is as-if randomized for those units near the policy assignment threshold.

Another possible extension to the regional policy evaluation methods comes from the spatial RDD estimator, which is usually adopted to retrieve the causal impact of a policy in proximity of a geographic boundary as seen in Section 2.1.3. However, the spatial RDD could also be used to assess the spatial variation in treatment effects with respect to specific locations as suggested by Keele and Titiunik (2015). Since the boundary is an infinite collection of points, it is possible to select some of these points along the boundary and estimate the treatment effect in each one of them. This process will produce a collection of treatment effects that can vary along the boundary, leading to a treatment-effect curve, where each effect can be mapped in its specific location. To the best of our knowledge, no regional policy has been evaluated assessing the heterogeneity of the impact along the geographic border. However, such an approach could be used to evaluate the heterogeneity of the impact with respect to: i) one or more covariates in case units have different characteristics along the border; ii) treatment intensity in case units receive different levels of treatment along the border; iii) specific geographic features such as the presence of rivers, mountains or woods which might hinder the linkage between neighboring areas. In addition, the investigation of the latter heterogeneity might allow us to gain some insights of the presence/absence of policy spillovers across the boundary.

4. Conclusions.

In recent years, the use of counterfactual techniques for the evaluation of regional policies has greatly expanded, mainly due to the availability of ever better data from the point of view of geographical location, the characteristics of the subjects involved, the temporal and spatial coherence. This process will continue rapidly with the increasing accessibility to large datasets with a wealth of previously unthinkable information.

However, the availability of larger and better data and the development of more specific research questions has posed new challenges to the econometric techniques developed to estimate regional policy impacts. The result is a methodological development of techniques dedicated principally to the evaluation of these policies. In the first place, the observation that these policies act in their own space, often defined by natural or administrative boundaries, has spurred the development of techniques capable of incorporating this feature in the definition and identification of evaluation models. Secondly, the presence of spillovers between subjects and regions, often a desired outcome of these interventions, required a relaxation of the SUTVA, which pushed scholars to propose different identification hypotheses, which require validation also on a theoretical level, in order to allow the estimation of the policy impact in a wider context.

This survey has given an account of these advances, and of some recent or still evolving methodological developments. Future challenges also go in other directions, based on new methodologies and access to increasingly broad and timely data sources. Mainly, we expect that the ex-post evaluation of regional policies will be increasingly influenced in the near future by empirical approaches coming from other disciplines, such as psychology, political science and computer science. For instance, regional policy evaluators might try to combine the policy evaluation techniques toolbox with mediation analysis. Mediation analysis aims to identify and specify intermediate paths and variables, through which the intervention produces the causal effect of interest (see, for instance, MacKinnon et al. 2007). Therefore, it explicitly adds a modeling component and, at the same time, it prepares the tools to falsify or corroborate the assumptions on which the model is based. Using mediation analysis would allow regional scientists to investigate and test for potential mechanisms behind the causal estimates. Another potential future venue for regional policy evaluators is to consider machine learning techniques. Although machine learning is concerned primarily with prediction (see, among others, Andini et al. 2017) by training complex models (e.g., random forest, neural nets) to maximize predictive accuracy, it is possible to envisage the use of machine learning in the ex-post evaluation of regional policies as well. Indeed, the increasing availability of big data will probably spur a use of such techniques to improve on the creation of counterfactual scenarios (see Varian 2014).