Eggers, Andrew C.; Freier, Ronny; Grembi, Veronica; Nannicini, Tommaso

Working Paper

Regression discontinuity designs based on population thresholds: Pitfalls and solutions

DIW Discussion Papers, No. 1503

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Eggers, Andrew C.; Freier, Ronny; Grembi, Veronica; Nannicini, Tommaso (2015) : Regression discontinuity designs based on population thresholds: Pitfalls and solutions, DIW Discussion Papers, No. 1503, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
http://hdl.handle.net/10419/119320

DIW BERLIN

1503

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2015

Regression Discontinuity Designs Based on Population Thresholds

Pitfalls and Solutions

Andrew C. Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions

Andrew C. Eggers – University of Oxford
Ronny Freier – DIW Berlin and FU Berlin
Veronica Grembi – Copenhagen Business School
Tommaso Nannicini – Bocconi University

First Version: October 2013
This version: September 2015

Abstract

In many countries, important features of municipal government (such as the electoral system, mayors' salaries, and the number of councillors) depend on whether the municipality is above or below arbitrary population thresholds. Several papers have used a regression discontinuity design (RDD) to measure the effects of these threshold-based policies on political and economic outcomes. Using evidence from France, Germany, and Italy, we highlight two common pitfalls that arise in exploiting population-based policies (confounded treatment and sorting) and we provide guidance for detecting and addressing these pitfalls. Even when these problems are present, population-threshold RDD may be the best available research design for studying the effects of certain policies and political institutions.

We would like to thank Charles Blankart, Peter Haan, Magnus Johannesson, Juanna Joensen, Henrik Jordahl, Erik Lindqvist, Christian Odendahl, Thorsten Persson, Janne Tukiainen, Tuukka Saarimaa, Viktor Steiner, and David Strömberg. Comments from seminars at Stockholm School of Economics, Potsdam University, Stockholm University, Bayreuth University, Marburg University, University of Bonn (Max-Planck Center), the University of Madgeburg, and the Midwest Political Science Association conference are also gratefully acknowledged. Federico Bruni, Helke Seitz, Sergej Bechtolt, and Moritz Schubert provided excellent research assistance. Some arguments for the exposition on the German case have earlier been circulated under the title: "When can we trust population thresholds in regression discontinuity designs?". Freier gratefully acknowledges financial support from the Fritz Thyssen foundation (Project: 10.12.2.092). The usual disclaimer applies.

I. Introduction

Researchers attempting to estimate the effects of policies face serious endogeneity problems: it is usually impossible to run an experiment in which consequential policies are randomized, and in most observational data it is difficult to locate or construct valid counterfactuals given the various strategic and contextual factors that affect policy choices. In recent years, many researchers have attempted to address these problems by exploiting cases in which policies at the subnational (usually municipal) level depend discontinuously on population thresholds. The use of regression discontinuity designs (RDDs) based on population thresholds was first suggested by Pettersson-Lidbom (2006, 2012), who evaluated the effect of the size of the municipal council on the extent of municipal spending in Sweden and Finland by comparing cities above and below population thresholds that determine council size. Subsequent researchers have used population-threshold RDDs to study the effects of the salary of public officials,[1] gender quotas,[2] electoral rules,[3] direct democracy,[4] fiscal transfers,[5] and (like Pettersson-Lidbom (2006, 2012)) council size.[6] (We survey twenty-eight papers using population-threshold RDDs in Table 4 in the Online Appendix.) Fundamentally, the population-threshold RDD is an attractive research design because at the relevant population threshold we can compare sets of cities that implemented different policies but are comparable in other important respects.

In this paper we highlight two pitfalls that (based on our study of France, Italy, and Germany) complicate the use of population-threshold RDDs. The first pitfall is that the same population threshold is often used to determine multiple policies, which makes it difficult to interpret the results of RDD estimation as measuring the effect of any one particular policy. We show the extent of confounded treatment in the three countries we study, emphasizing that extensive institutional background research is necessary before one can interpret the results of a population-threshold RDD as evidence of the effect of a particular policy. When discussing

[1] See Gagliarducci and Nannicini (2013); Ferraz and Finan (2009); van der Linde et al. (2014); De Benedetto and De Paola (2014).

[2] See Bagues and Campa (2012); Casas-Arce and Saiz (2015).

[3] See Fujiwara (2011); Hopkins (2011); Pellicer and Wegner (2013); Barone and De Blasio (2013); Eggers (2015); Gulino (2014)

[4] See Arnold and Freier (2015); Asatryan et al. (2013); Asatryan, Baskaran and Heinemann (2014).

[5] See Litschig and Morrison (2010, 2013); Brollo et al. (2013); Mukherjee (2011); Baskaran (2012).

[6] See Egger and Koethenbuerger (2010); Koethenbuerger (2012).

potential remedies, we also highlight the "difference-in-discontinuities" design as a possible solution in cases where a treatment of interest changes in tandem with other policies but one can locate a comparable period or setting where these other policies change on their own.

The second pitfall is sorting – the tendency of municipalities to strategically manipulate their official population in order to fall on the desired side of a consequential population threshold. It is well known that the continuity assumption necessary for the RDD may not hold when there is precise manipulation of the running variable (Imbens and Lemieux 2008; McCrary 2008); evidence of manipulation has been produced by Urquiola and Verhoogen (2009) for the case of class size, Barreca et al. (2011) for birth weight, and Caughey and Sekhon (2011) for close elections (though see also Eggers et al. 2015). Our main contribution here is to show conclusive evidence of manipulative sorting in official population numbers in France, Italy, and Germany;[7] we also show that the standard tests for sorting are biased when the running variable is discrete (as in the case of population-threshold RDDs), and we highlight some of the special challenges involved with assessing covariate imbalance in settings where data is pooled from multiple thresholds.

The evidence we present from France, Italy, and Germany shows why carrying out population-threshold RDD in these countries requires care; readers should not conclude, however, that population-threshold RDDs are always problematic or that there are better ways to study the policies that have been addressed with population-threshold RDDs. We suspect that both confounded treatment and manipulative sorting are serious problems in many countries that use population thresholds to assign municipal policies, but even in the countries we study one can identify policies and thresholds such that neither confounded treatment or sorting appears to pose much of a problem. When these problems do arise, there are remedies that we discuss that involve assumptions that will often be weaker than would be necessary for any feasible alternative design. The countries we study are also not representative of all settings where population-thresholds may be carried out; we chose these countries both because many muni-

[7]The paper that is most related to our work in this strand of research is the study by Litschig (2012), which looks at top-down manipulation of population figures in Brazil. In parallel work to ours, two further papers highlight the issue of sorting around population figures in Spain and Belgium (see Foremny, Monseny and Solé Ollé (2015) and De Witte, Geys and Heirman (2015)).

cipal policies depend on population thresholds (and they have done so for a long time[8]) and because we are familiar with these cases from previous work, but we expect that confounded treatment is less severe in countries where population thresholds are less commonly used (e.g. see Hopkins (2011) on the United States) and sorting is less problematic in countries where municipal population counts are linked more closely to national administrative data (e.g. see Pettersson-Lidbom (2012) on Finland and Sweden). Especially given the general challenges we face in studying the effects of policies, it would be a mistake to conclude from our analysis that population-threshold RDDs should be eschewed in favor of other designs – not just because these problems do not afflict every population-threshold, and not just because there are solutions (which we discuss in depth) to these problems, but also because even in the face of these problems a population-threshold RDD may be preferable to the next best design.

II. Confounded treatments

The population-threshold RDD is appealing because it allows the researcher to compare outcomes in a set of cities where one subset is required to implement one policy (say, A) while another identical-in-expectation subset is required to implement another policy (A'). The first common problem we highlight in this paper is that often the population threshold that determines whether policy A or A' is applied will also determine whether other policies (B or B', C or C') are applied; the policy change we hope to study (A vs A') is thus confounded with other policy changes, undermining the appeal of the RDD.

A. Documenting the extent of confounded treatment

Figure 1 summarizes the problem based on our investigation of laws applying to municipalities in France, Italy, and German states. Each dot indicates a population threshold at which at least one policy changes; solid dots indicate that more than one policy changes at the same

[8]The first law on municipal government in revolutionary France (passed 14 December of 1789) includes six provisions dictating features of municipal government as a function of population, including a rule specifying six population thresholds determining the council size. An 1808 law that reformed the constitutional rights of municipalities in Prussia ("Preußsche Städteordnung von 1808") used population cutoffs of 3,500 and 10,000 to assign different rules on council size, voting rights, and budget transparency (among others). In Italy, the *Legge Lanza*, which was drafted in 1865 after the Royal Decree n.3702 from 1859, specified population cutoffs determining council size, executive committees, and voting rights in the former Kingdom of Piedmont and Sardinia.

threshold. In every case, there are some thresholds where just one policy changes, but such thresholds are in the minority.

Table 1 provides detail on the various policies that change at population thresholds up to 50,000 in France; the Online Appendix provides details about population-dependent policies in Italy and Germany (see Tables 5 and 6). As Table 1 indicates, at *every* threshold at which council size increases, the maximum number of deputy mayors also increases, which makes it impossible to disentangle the effect of council size from the effect of additional paid council staff. There is only one threshold (1,000 inhabitants) at which the salary of mayors and deputy mayors increases without the council size also increasing. Many of the most interesting policies change at a single threshold of 3,500 inhabitants at which several other policies (including council size and mayor's wage) also change: the electoral rule used to elect the council, the requirement of gender parity in party electoral lists, and the requirement that the council debate the budget before adopting it.

In the thirteen German states a total of sixty-five different types of municipal policy depend on population thresholds; no state has fewer than fourteen different policies that are determined by population thresholds. (See Appendix Table 6 for details.) The thresholds determining these policies vary across states, ranging from seventy inhabitants to one million. Importantly, of 759 policy-threshold observations across German states (i.e. cases where a policy changes at a given threshold in a given state), only ninety-four do not coincide with another policy change. For mayoral salary, certainly one of the most important of these policies, we find only twelve cases (of 116 in total) in which no other policy changes at the same threshold in the state.[9]

Detecting whether a given treatment is confounded with another treatment can simply be a matter of scouring the legal code for mentions of population thresholds. In some cases enumerating the full set of policies that change at a threshold is more complicated, however, because some policies depend on population thresholds only indirectly. An example of this type of second-order policy is given in Lyytikäinen and Tukiainen (2013): the maximum number of candidates on electoral lists in Finland is a function of the council size, which changes

[9]In Germany and other federal systems the task of locating relevant thresholds is complicated by the fact that higher level authorities may also enact policies based on municipal population thresholds; in Germany, for example, the federal statistical office used a different procedure to implement the 2011 census for municipalities above and below 10,000 inhabitants.

discontinuously at population thresholds.[10] Another example from Baskaran and Lopes da Fonseca (2015) highlights how subtle the interactions among policies can be: in German municipal elections, parties winning less than a certain vote share are denied representation on the council; this constraint is never binding when the municipal council is below a certain size, which implies that there is a population threshold at which the council size increases *and* a vote share cutoff goes into effect (though this would not be clear without detailed knowledge of the electoral system). In short, a researcher should know a setting intimately before concluding that a given policy (and not other policies) changes at a given population threshold.

B. Addressing confounded treatments

Suppose a policy of interest is determined by a population threshold, but other policies change at the same threshold. How can a researcher proceed?

The simplest option is to change the quantity of interest to include the other policies that change in tandem. If two policies move in perfect lockstep, then what initially may seem like an opportunity to learn about the effect of policy A vs A' is at best an opportunity to learn about the effect of policy combination AB vs $A'B'$. In some cases it may be worth studying the effect of this bundle of policies. In France, for example, changes in council size always coincide with changes in the number of deputy mayors; the perfect confounding of these two policies means that it is impossible to separate the effect of the two treatments, but one may be content with estimating the effect of the combination of policies.

If we want to keep the focus on the main policy of interest (A vs A'), then the most promising way to proceed is to look for other settings where the other policy change (B vs B') occurs in isolation; under assumptions we lay out shortly, the difference between the effect of both policies (AB vs $A'B'$) and the effect of the "nuisance" policy (B vs B') gives an unbiased estimate of the effect of A vs A', the quantity of interest. This approach, which combines features of the regression discontinuity design and the difference-in-differences design, is what Grembi, Nannicini and Troiano (2014) call the "difference-in-discontinuity" (diff-in-disc) design. Here we briefly formalize their approach and elaborate on different ways it can

[10] Lyytikäinen and Tukiainen (2013) are also the only paper using population thresholds that tackle the confounded treatment issue.

be applied.

Denote by ab the policy bundle a given municipality receives, where $ab \in AB, A'B, AB', A'B'$. We consider a setting where this bundle is determined by whether the municipality's population Z_i is above or below a threshold value Z_0. Denote by $Y_i(ab)$ the potential outcome when municipality i receives policy bundle ab. Define $Y_{ab}^{+} \equiv \lim_{Z \to Z_0^{+}} E\left[Y_i(ab)|Z_i = Z\right]$ and $Y_{ab}^{-} \equiv \lim_{Z \to Z_0^{-}} E\left[Y_i(ab)|Z_i = Z\right]$. (In words, these are the average potential outcomes for cities at the threshold when they implement policy bundle ab; in the first case the limit is taken from above and in the second it is taken from below.) Similarly, define $Y^{+} \equiv \lim_{Z \to Z_0^{+}} E\left[Y_i|Z_i = Z\right]$ and $Y^{-} \equiv \lim_{Z \to Z_0^{-}} E\left[Y_i|Z_i = Z\right]$ for the observed outcome Y.

Consider a case where the bundle that applies above the threshold is $A'B'$ and the bundle that applies below the threshold is AB. The cross-sectional RDD estimator in that case gives us $Y_{A'B'}^{+} - Y_{AB}^{-}$; by adding and subtracting $Y_{AB'}^{+}$ we get

$$\begin{aligned} \tau_{RDD} \equiv Y^{+} - Y^{-} &= Y_{A'B'}^{+} - Y_{AB'}^{+} + Y_{AB'}^{+} - Y_{AB}^{-} && (1) \\ &= \tau_{ATE|b=B'} + \underbrace{Y_{AB'}^{+} - Y_{AB}^{-}}_{\text{bias}}, && (2) \end{aligned}$$

where $\tau_{ATE|b=B'}$ refers to the average treatment effect of A' vs A for units that received B'.

Now suppose we also have available a second case where the bundle that applies above the threshold is AB' and the bundle that applies below the threshold is AB; we denote potential outcomes in this case as, e.g., $\tilde{Y}_{AB'}^{+}$. The difference-in-discontinuity estimator and estimand are $\tau_{DiDISC} \equiv (Y^{+} - Y^{-}) - (\tilde{Y}^{+} - \tilde{Y}^{-}) = (Y_{A'B'}^{+} - Y_{AB}^{-}) - (\tilde{Y}_{AB'}^{+} - \tilde{Y}_{AB}^{-})$. Consider the following assumption:

Assumption 1:

$$Y_{AB'}^{+} - Y_{AB}^{-} = \tilde{Y}_{AB'}^{+} - \tilde{Y}_{AB}^{-} \qquad \textit{(Local Parallel Trends)}$$

This assumption can be interpreted from two perspectives. Most directly, it states that the effect of B' as opposed to B, holding fixed A, is the same in the first case (in which AB changes to $A'B'$ at the threshold) and the second case (in which AB changes to AB' at the threshold). It is thus analogous to the standard parallel trends assumption in difference-in-differences

estimation, where the two cases being compared are pre- and post-treatment. (Note that Assumption 1 is more local, however, as it must hold only in the neighborhood of the policy threshold.) By rearranging the terms as $Y^+_{AB'} - \tilde{Y}^+_{AB'} = Y^-_{AB} - \tilde{Y}^-_{AB}$, we can see Assumption 1 from a different perspective: it states that the difference in potential outcomes between the two cases should be the same just above and just below the threshold. In this format it is thus analogous to the standard RDD assumption of continuity in potential outcomes across the threshold.

Under Assumption 1, we have that

$$\begin{aligned} \tau_{DiDISC} &= (Y^+_{A'B'} - Y^-_{AB}) - (\tilde{Y}^+_{AB'} - \tilde{Y}^-_{AB}) && (3) \\ &= Y^+_{A'B'} - Y^+_{AB'} = \tau_{ATE|b=B'} && (4) \end{aligned}$$

where the local parallel trend assumption was used to get from the first line to the second line. Thus, under the local parallel trends assumption, the diff-in-disc estimator removes the bias in Equation 2 and yields the effect of A' vs. A conditional on policy B'.

If we make a further assumption,

Assumption 2:

$$Y^+_{A'B'} - Y^+_{AB'} = Y^-_{A'B} - Y^-_{AB} \qquad \textit{(Separability)},$$

which essentially says that the effect of A' vs A in the first case does not depend on whether we are just above the threshold (and thus B' prevails) or below it (and thus B prevails), then we can say that $\tau_{DiDISC} = \tau_{ATE}$, the average treatment effect in the neighborhood of the threshold, which is the standard estimand in RDD.[11]

Given a setting where a policy of interest changes along with a nuisance policy, then, we can use the difference-in-discontinuity (diff-in-disc) design to recover the effect of the policy of interest if we have a second setting in which the nuisance policy changes on its own *and* if we are willing to assume that the effect of changing the nuisance policy (holding fixed the policy of interest) is the same in two settings (Assumption 1). In what situations is this possible?

[11] Assumption 2 thus allows us to generalize somewhat, such that the diff-in-disc estimator gives the effect not just immediately above the threshold but in the entire neighborhood of the threshold.

Grembi, Nannicini and Troiano (2014) illustrate what we might call a "longitudinal diff-in-disc" in order to estimate the effect of fiscal constraints on deficits. Starting in 2001, Italian municipalities below 5,000 were exempted from fiscal constraints that applied to larger cities. A cross-sectional RDD analysis in the post-2001 period using the 5,000 population threshold would thus seem like a good way to study the effect of fiscal constraints vs. no fiscal constraints. The problem is that (as noted in Table 5 of this paper) the salary of the mayor and other executive officers also changes at the 5,000 threshold. Grembi, Nannicini and Troiano (2014) thus implement a diff-in-disc design in which the cross-sectional RDD effect at the 5,000 threshold before 2001 (when fiscal constraints applied to all municipalities) is subtracted from the same effect after 2001 (when fiscal constraints only applied to municipalities below 5,000 in population). This procedure yields a consistent estimate of the effect of fiscal constraints under the local parallel trends assumption that the effect of the other policies that change at this threshold is stable over time and the separability assumption that the effect of fiscal constraints does not depend on these other policies.

Researchers can also consider what we might call a "cross-sectional diff-in-disc" to address the problem of confounded treatment. The key requirement of the cross-sectional diff-in-disc is that the nuisance policies also change at some other threshold or in some other region where the local parallel trends assumption and the separability assumption are plausible, i.e., the effect of the nuisance policies is plausibly the same in the two settings and does not depend on the value of the policy of interest. Arnold and Freier (2015) and Eggers (2015) provide evidence in this spirit by comparing RDD effects measured at different thresholds in the same system in order to "difference out" the effects of nuisance policies. The same approach could of course be used when the effect of the nuisance policies can be measured in an entirely different region or country where the policy of interest does not change; the attractiveness of this design of course depends on the plausibility of the local parallel trends assumption.

III. Sorting

As mentioned above, the appeal of a population threshold-based RDD is partly that the political unit does not choose the policy, which suggests that units just above and below the

threshold should be comparable in all respects other than the policy. As is well known, such an RDD (like any RDD) is less appealing when the units can influence the variable that determines treatment assignment (i.e., population). At an extreme, one could imagine that cities near a population threshold could perfectly control whether they end up above or below the threshold, and thus cities that have policy A differ from cities that have policy A' not just in that policy but also in a whole host of background characteristics that affected their decision to get policy A or policy A'.[12] In such a situation, carrying out an RDD may be no better than a typical observational study in which political units choose their policies.

The problems of strategic sorting in RDD applications are well known (see Imbens and Lemieux (2008); Lee and Lemieux (2010); Urquiola and Verhoogen (2009); Barreca et al. (2011)). Strategic sorting in population figures has been documented by Litschig (2012) for Brazil and it has been briefly mentioned by Gagliarducci and Nannicini (2013) for the Italian case. One of our contributions here is to provide evidence that sorting in RDD studies based on population thresholds is an issue in all three countries that we study. We also demonstrate techniques for diagnosing and explaining manipulation, as well as potential solutions to address this problem.

A. Aggregate graphical evidence

The basic pattern of sorting is documented in Figures 2 (France), 3 (Italy), and 4 (Germany). Because the figures use the same format and reflect the same analysis, we explain the French case in detail and subsequently note only the relevant differences between the French case and the others.

In France, we have population data for eight censuses between 1962 and 2011.[13] For each census, we calculate the difference in population between each city and each major population threshold (i.e., one affecting a policy listed in Table 1) that was in force at the time of the

[12] Alternatively, it may be that only certain cities are able to control whether they end up above or below the threshold, in which case cities that have policy A may differ from cities that have policy A' not only in the factors that affect their policy preferences but also in the factors that affect their ability to manipulate their population figures.

[13] The census years are 1962, 1968, 1975, 1982, 1990, 1999, 2006, and 2011. After 1999, France introduced a new census system that produces annual population estimates for all municipalities; the 2006 census was the first such census.

census; we store all municipality-years in which a city's population was within 250 inhabitants of a threshold. In the left panel of Figure 2 we plot three histograms of these population differences, one for each group of relevant population thresholds (100; 500 or 1,000; and 1,500 and larger). Because there are so many municipality-years, we plot histograms with bin widths of 1. The key evidence of sorting is given by the jumps in each histogram at 0. For example, based on the histogram for the 100-inhabitant threshold, we can see that there were just under 500 cases in which a city was one person short of the 100-inhabitant threshold at which the council size increases, but there were almost 600 cases in which a city cleared that hurdle by one person. The jump is even more striking for the 500- and 1,000-inhabitant thresholds (where the mayor's salary increases).

In the right panel of Figure 2 we depict the McCrary test for all thresholds pooled. This procedure estimates the density of the running variable (i.e., absolute distance in inhabitants to a population threshold) separately on the left and right of the threshold and tests for a jump or drop in the density at the threshold. Not surprisingly (given the histograms in the left panel), the McCrary test indicates a large jump in the estimated density at the threshold.

Figure 3 indicates an even more striking pattern for Italy. Based on the five decennial censuses from 1961 to 2001, we find about 90 cases in which a city cleared the 1,000 or 3,000 population threshold (at which the mayor's wage increases, among other changes) by fewer than 10 inhabitants, but we find only about 20 cases in which a city fell short by fewer than 10 inhabitants; in over 300 cases a city cleared one of the thresholds by fewer than 30 inhabitants, but in fewer than 100 cases did a city fall short by fewer than 30. The pattern of sorting is just as clear (if not as dramatic) at larger thresholds. Again, the McCrary test aggregating all thresholds (right panel) indicates a very large jump in the estimated density at the threshold.

Figure 4 shows the same analysis for Germany. Here we have annual administrative data from 1998-2007 for municipalities from all German states, and our analysis is based on a comparison of each municipality's population to all thresholds in force in that municipality's state. The histograms (left panel) indicate that sorting is nowhere near as severe here, but the McCrary test (right panel) does indicate a significant jump in the density just above the threshold. Note that this analysis includes all thresholds, including many that determine quite

minor policies; in the next section we carry out McCrary analysis for each country for specific types of thresholds.

B. Formal tests at different types of thresholds

We now carry out the McCrary (2008) test for different types of thresholds within countries, still pooling population figures from the various censuses we have collected. Before showing the results, we note that our analysis here and throughout the paper takes account of two biases (previously unrecognized, as far as we know) that arise when applying the standard McCrary test to a discrete running variable. The McCrary test operates by conducting RDD analysis on an under-smoothed histogram of the running variable. The first bias arises because applying the standard algorithm to a discrete running variable tends to result in a histogram with more observations in the first bin to the right of the threshold than in the first bin to the left, even when the density is perfectly flat; fundamentally, this asymmetry arises because with a discrete running variable one can have observations *exactly* at the threshold, and by default these observations are assigned to the first bin to the right. We address this problem by requiring that the bin width of the histogram take an integer value;[14] alternatively, one can simply set the threshold to be -0.5, which eliminates the asymmetry as long as the bin size is not exactly 0.5, 1.5, etc. The second bias arises when a discrete-valued running variable is analyzed using relative deviations from thresholds of different size, e.g. percentage distance from thresholds of 500, 1000, and 10,000 inhabitants; this creates a bias because all thresholds can produce a relative deviation of 0 (which by default goes into the first bin to the right of the threshold) but only very large thresholds can produce a relative deviation of $-\epsilon$. We address this problem by using absolute deviations rather than relative deviations. We explain these biases (both of which tend to increase the likelihood of falsely detecting sorting, especially when data is very plentiful) and our solutions to them in the Online Appendix.

Table 2 reports the results of McCrary analysis (incorporating these adjustments) at different types of thresholds in all three countries. In the top row we assess evidence of sorting in all thresholds, reporting the point estimate (i.e. the effect of crossing the threshold on the log

[14]More specifically, we force the bin size of the McCrary algorithm to the closest integer value to the one chosen by default.

density) and standard error for each test, along with the number of thresholds and observations.[15] Consistent with the previous figures, we find very clear evidence of substantial sorting in France and Italy (with the latter being quite a bit larger) and evidence of small but statistically significant sorting in Germany. In the other rows we assess sorting at particular types of thresholds, e.g., thresholds where the salary of the mayor increases, or thresholds where the council size increases, or thresholds where both increase. In France, we find significant sorting at all types of thresholds, with the smallest effect (and weakest evidence against the null) at thresholds where council size increases (but not mayor's wage) and thresholds at 3,500 inhabitants or higher. In Italy the estimated effects are much larger, with (as in France) smaller effects at larger population thresholds. To give a sense of magnitude, a McCrary effect size of 1.3 (the effect for Italy at all thresholds) implies that the density on the right of the average threshold is almost four times larger than on the left. In Germany the jumps in density are statistically significant for most subsets and smaller but still fairly substantial in magnitude: at thresholds where both the council size and the mayor's salary increases, for example, there are about 15% more cities immediately to the right of the threshold than immediately to the left. The fact that sorting appears to be more severe when we focus on thresholds determining salary and council size is consistent with the idea that local officials strategically manipulate population figures to obtain desirable policies; at these thresholds there is a clear incentive to pass the threshold, while at some others (e.g. thresholds above which cities are subject to more stringent financial oversight) we would if anything expect sorting in the other direction.

Comparing the effects by threshold size shows larger effects for smaller thresholds in Italy and France, suggesting that population size is more easily manipulated in smaller towns. Intriguingly, in Germany the pattern is reversed, with somewhat larger effects at larger thresholds, which may be partly explained by the fact that the salary of mayors in Germany often only increases at larger population thresholds.

At the bottom of Table 2 we conduct the McCrary tests at thresholds at which no policy changes, as far as we are aware. We generated placebo thresholds by taking the midpoint between each actual threshold in each setting (e.g., in France the smallest placebo threshold is

[15] We count only thresholds for which we observe cities within 250 inhabitants of the threshold, which explains why some of the counts differ from the analysis above.

300, which is halfway between 100 and 500) and adding an arbitrary number (117 was picked). In none of the countries do we find discontinuities in the density at these placebo thresholds.

C. How does sorting happen?

The evidence above is consistent with the view that in many municipalities in France, Italy, and Germany, officials and/or citizens respond to population-based policies by manipulating population numbers. We now ask briefly how such manipulation might take place – both because it might indicate how widespread sorting is likely to be beyond these three countries and because it might help us understand the extent to which sorting endangers our ability to learn from RDD in these and other settings.

It may be useful to distinguish among three distinct types of local behavior that could produce the manipulative sorting we observe. First and most simple is *fraud*: officials could simply falsify population numbers, inventing or ignoring residents in order to achieve desired population numbers. Second is what we call *selective precision*: when a municipality is known to be close to a consequential population threshold, officials can selectively order extra checks or expedite/delay procedures that are likely to move the final count in the desired direction.[16] Third is *strategic recruitment*: a municipality could make efforts to attract residents (or repel them) by expediting permits or offering tax incentives or simply encouraging friends to change their official residence.

Do local officials have the means, motive, and opportunity to implement these sorting strategies in the countries we study? The assignment of consequential policies (e.g. the salary of the mayor or the electoral system) based on population thresholds in all three countries provides a clear motive. Local officials in each country are also sufficiently involved in the census and in housing and tax policy to have the means to manipulate. In both France and Italy, mayors are responsible for supervising the census survey at the local level, including hiring and training enumerators; in Germany, municipal registry offices provide reports of births, deaths, and in- and out-flows that state statistical offices use to update census numbers.

[16]Thus selective precision differs from fraud in the sense that the procedures are accurately carried out; the key is that for certain procedures, such as processing new arrivals or checking whether any houses in the municipality should actually be classified as vacation homes, officials can know in advance that implementing the procedure can only increase or decrease the total population count.

In all three countries municipalities are also involved in local development and tax decisions, which suggests that they have the means to recruit residents. Whether local officials have the opportunity to implement these strategies is somewhat more difficult to say. Fraudulently adjusting or fabricating census surveys in order to achieve a desired population number seems risky in systems where central authorities oversee local procedures. For all three mechanisms, the pattern of sorting suggests that local officials must have very precise information about the municipality's ultimate census count at the time when they decide whether or not to engage in manipulation. To see why, note that the most striking feature of Figures 2, 3, and 4 is the deficit of cities narrowly below the relevant thresholds. This indicates that potential manipulators know at least whether the municipality is likely to be very close to the threshold (because cities 1 inhabitant below the threshold appear to be much more likely to manipulate than cities 5 or 10 below) and, in the case of selective precision and recruitment, they know which side of the threshold they are likely to end up on (because cities 1 inhabitant below the threshold appear to be much more likely to manipulate than cities 1 above). This in turn suggests that the manipulation we observe is probably not the result of strategies that would require substantial time to implement, such as issuing permits for new housing; new housing may indeed help a city cross a threshold, but in the time it takes there would likely be stochastic changes in the overall population such that it would not produce sharp sorting right at the threshold. This sharp sorting could, however, be the result of calling for an extra check after initial numbers are tallied (i.e. selective precision) or recruiting a friend from a neighboring municipality to move into a vacant apartment before the census takes place (i.e. strategic recruitment).

The case of France may be instructive in highlighting possible mechanisms for manipulative sorting. The French census is a joint project between the national statistics agency (INSEE) and local municipal authorities: INSEE issues directives; the municipalities hire and train enumerators and submit the results. Municipal authorities are thus involved in interpreting the complex rules that determine how to handle ambiguous cases such as students, members of the military, and people without fixed domiciles. The phenomenon of sorting in the French census was noted as early as 1972 by an INSEE official (Vernet 1972) who suspected that it could be explained by local officials making an extra effort to locate residents when initial tabulations

indicated that they would otherwise narrowly fall below an important threshold;[17] to the extent that these efforts involved locating actual residents (e.g. students who should be enumerated in the municipality), the official's explanation falls under what we call "selective precision". (If "locating" means "inventing", we would call it fraud.) Consistent with this explanation, manipulative sorting in France appears to have diminished over time as central authorities have exercised more oversight over municipalities' data collection procedures. Figure 5 depicts the point estimates and confidence intervals for McCrary tests at three different thresholds over time in France, clearly showing a decline in sorting since the 1980s and a particularly marked drop in the 1999 census. A former census official explained this pattern by noting that for the 1999 census INSEE instituted special measures to strengthen oversight of the census, particularly to ensure that students were only counted once; censuses after 1999 have used a new procedure in which annual population updates for small municipalities are based on local tax files, which may be less prone to manipulation.[18] The variation in sorting over time in France suggests that sorting is less likely to be an issue for population-threshold RDDs in countries like Sweden and Finland where local population figures are collected in a highly centralized way and linked to administrative records.[19]

D. Addressing manipulative sorting

The regression discontinuity design is obviously much less appealing when there is evidence of sorting around the threshold. What can a researcher do in such cases?

One approach is to augment the usual RDD analysis with control variables that capture possible confounding factors. When sorting introduces bias into RDD estimates, it does so because the distribution of covariates differs between the left and right side of the threshold. One way to eliminate this bias, therefore, is to measure these covariates and model their

[17] "The number of municipalities with a population a little below 500 or 1000 inhabitants declines the closer this number gets to 500 or 1000 and increases suddenly for values immediately above these limits. It all takes place as if officials in municipalities where the results obtained by census agents are close to a significant threshold make a maximum effort to enumerate a few individuals who, not having been taken into account in a first tabulation, allow them to cross the desired threshold" (p. 19; authors' translation).

[18] Personal correspondence with Jean-Michel Durr, former Census Director at INSEE.

[19] Consistent with this, Pettersson-Lidbom (2012) and Lyytikäinen and Tukiainen (2013) do not find evidence of sorting in Sweden or Finland.

relationship to the outcome at the threshold. In this approach to sorting, an RDD thus becomes more like a typical observational study, in the sense that one must identify, measure, and control for additional variables. The credibility of the resulting model will depend on what we know about the process of sorting, the extent to which we can measure relevant covariates, and the number of observations near the threshold for model-fitting. It also depends on the extent to which the outcome varies with the unmanipulated running variable. In the best case, such analysis will retain much of the appeal of the ideal RDD; in the worst case, such analysis will be no more attractive (and possibly less attractive) than a pure observational study.

To understand some of the considerations in addressing manipulative sorting through covariate adjustment, consider Figure 6, which captures what we think of as the best-case scenario. Because of sorting, a single binary covariate X is not continuous at the threshold, as shown in the left plot. The right plot shows how this induces bias in the RDD estimate: the expectation of Y conditional on Z (the running variable) and X (the covariate) is completely flat everywhere, but due to the imbalance in X the expectation of Y conditional on Z (but not conditional on X) bends as we approach the threshold, such that the RDD estimate $(Y^+ - Y^-)$ is larger than the effect of the treatment conditional on $X = 1$ or $X = 0$ (given by $Y^+_{x=1} - Y^-_{x=1}$ and $Y^+_{x=0} - Y^-_{x=0}$, respectively). The bias due to imbalance in X can, however, be removed by controlling for X in the RDD analysis. In this very simple case, where $E[Y|X = 1, Z] - E[Y|X = 0, Z]$ is independent of Z, controlling for X is as simple as additively including X in the regression. More generally, one could allow the control function to vary across levels of X or simply estimate the RDD separately across levels of X.

In practice, addressing sorting by controlling for covariates is typically more difficult than in this best-case scenario for several reasons. First, the task of accurately modeling the relationship between the outcome and the covariate (conditional on the running variable) can be difficult; estimates become more dependent on modeling choices and subject to sampling variation. Second, even when we can address the bias due to imbalance in a covariate X at the threshold, we can never completely rule out the concern that our estimates are still biased due to imbalance in other covariates. For both of these reasons, we lose some of the attractive simplicity of the ideal RDD analysis, in which the entire focus is on estimating two conditional

expectations at the threshold.[20]

To make matters worse, it should be remembered that we cannot rule out the possibility of covariate bias even when there is no sign of discontinuity in the density of the running variable (as McCrary (2008) noted). This suggests that every RDD study based on population thresholds should include extensive checks for covariate balance, whether or not there is evidence of sorting in the aggregate – particularly in settings where local officials play a role in producing official figures; when imbalance is evident, the robustness of conclusions to various control strategies should be shown. In the next section (Section IV) we assess the degree of covariate imbalance in the Italian case as an example.

As an alternative to covariate adjustment, researchers can also consider a "donut" RDD analysis that ignores data immediately surrounding the threshold (Barreca et al. 2011). In settings where the sorting appears to be limited to the immediate neighborhood of the threshold, this approach has the advantage that one does not need to measure and control for all potentially unbalanced covariates, nor does one need to worry about measurement error due to misreporting of the running variable. Of course, the very clear disadvantage of the donut approach is that as we drop more data near the threshold, our estimates of the conditional expectation function at the threshold require more extrapolation.

Building on the discussion of the difference-in-discontinuity design in Section III.B, in some circumstances one could take advantage of multiple thresholds to address sorting or at least give an idea of how problematic it is likely to be for one's analysis. For example, one could extend the logic of the diff-in-disc to "partial out" the effect of sorting in the special case where a policy of interest changes discontinuously at a threshold at time t_1 but not at time t_0 and sorting occurs (perhaps due to nuisance treatments) in both time periods. Under the assumption that the bias due to the combination of sorting and the nuisance policies is the same just above the threshold in the two periods (an extension of the local parallel trends assumption), one can use the diff-in-disc to identify the effect of the policy of interest for treated municipalities

[20] Another problem is that manipulative sorting introduces measurement error that induces bias when the conditional expectation depends on the true value of the running variable. That is, cities just above and below the threshold likely differ in their true population, but this variable is not observed and thus cannot be controlled for in a straightforward way. The best way to address this bias would be to obtain good estimates of the true population and include this as a control variable in the analysis.

just above the threshold; under the additional assumption that the effect of the policy of interest is the same for municipalities just above and below the threshold (an extension of the separability assumption), this is equal to the neighborhood average treatment effect. Both of these assumptions are likely to be less attractive than the usual diff-in-disc assumptions: the first assumption will not hold if the policy of interest affects the bias due to sorting, and the second assumption will not hold if the effect of the treatment is different for cities that managed to sort just above the threshold and those that did not.[21]

IV. Sorting and covariate imbalance: the case of Italy

The previous section provided clear evidence of manipulative sorting in France, Italy, and Germany around population thresholds determining municipal policies. This evidence indicates that the key assumption of RDD analysis (the continuity of potential outcomes across the threshold) may be violated in these cases. While we cannot directly test this assumption, we can test for covariate imbalance. In this section we conduct tests for covariate imbalance in Italy. Our goal in this section is to assess the extent of covariate imbalance in Italy and identify covariates that should be controlled for in RDD analysis in that setting. Along the way, we highlight some non-obvious issues that are likely to arise when we test for covariate imbalance using data drawn from multiple different thresholds and/or multiple censuses.

Our main approach to testing for covariate imbalance is to undertake a falsification test in which the covariate is viewed as the outcome in an RDD analysis. Figure 7 shows an example in which the dependent variable is the lagged treatment, meaning an indicator for whether a municipality was above a given population threshold in the *previous* census, given that it was close to that threshold in the current census. The top two plots show this analysis for thresholds at which the salary of the mayor increases. The top left panel shows that the probability of lagged treatment increases with the current running variable (as one might expect) but jumps at the threshold, indicating that cities narrowly above the threshold are

[21] Less formally, and still building on the diff-in-disc idea but in a different way, one could compare RDD estimates at two thresholds where a policy of interest changes but the apparent degree of sorting is much larger at one threshold than the other; if the estimates at these two thresholds are similar, one could conclude that bias due to sorting plays a small role based on the assumption that this bias is increasing in the degree of sorting.

more than 10 percentage points more likely to have been above the threshold in the previous census than cities narrowly below the threshold. The top right panel shows how the estimated jump varies with the (triangular) bandwidth employed for the local linear regression; the black dot shows the bandwidth suggested by the Imbens-Kalyanaraman algorithm (see Imbens and Kalyanaraman 2012) and employed in the figure at left. This clear jump indicates that the RDD analysis for Italy could be biased by the fact that cities above and below the threshold differ systematically in whether they received the treatment in the past. The bottom two plots of Figure 7 use "placebo" thresholds (with no policy changes) and show no evidence of similar persistence; indicating that it is not simply due to stickiness in the population figures that we find the results above.

How should this imbalance be interpreted? The most straightforward interpretation is that officials with influence over population figures prefer to prevent cities from crossing thresholds from one census to the next; for example, if a city has shrunk in population such that it is very close to a population threshold, someone is able to influence the final numbers to keep it above the threshold. Note, however, that a different and more subtle interpretation is also available. Recall that our analysis is based on combining observations near multiple different thresholds across multiple censuses. In such cases, covariate imbalance can emerge simply because the value of the covariate varies across thresholds/censuses *and* the degree of sorting varies across thresholds/censuses.

To see this, suppose we were combining data from a single threshold recorded in just two censuses: one old census, at a time when cities near the threshold were shrinking, and one new census, at a time when cities near the threshold were growing. Suppose sorting was severe in the old census but not the new census. The difference in the severity of sorting means that in the combined data a larger proportion of the observations just above the threshold (compared to just below) will be taken from the old census; because cities were shrinking at the time of the old census, observations from the old census would be more likely to have been above the threshold in the past. Thus even if there were no imbalance in the probability of lagged treatment in either census, we could observe imbalance in this covariate in the combined data.[22]

[22]The same argument could be made when we aggregate data from various thresholds at a single point in time.

The larger point is that it is tempting to interpret imbalance in a particular covariate as the *cause* of the sorting (e.g. the desire of officials not to cross thresholds), but it may simply be an artifact of pooling data from multiple censuses or thresholds in which the degree of sorting varies.

Table 3 addresses this complication by assessing imbalance across several covariates (indicated by rows of the table) while adding controls for the year of the census, the type of the threshold, and other factors. Each of the point estimates in this table is an RDD-based estimate of the effect of crossing population thresholds on the covariate. The estimated effect on lagged treatment (examined graphically in Figure 7) is reported in column 1 of the first row as 0.138 (0.037). Columns 2 to 5 carry out the same analysis but additively include covariates in the RDD analysis: dummies for each year of the census (column 2); dummies for each threshold (column 3); both dummies (column 4); and a set of covariates describing Italian municipalities around the year 2002 (column 5).[23] Columns 6 to 8 show the models from columns 1, 3, and 5 but focusing on "placebo" thresholds where no policy changes. Note that in the absence of sorting we expect no effects in any of these tests; in the presence of policy-induced sorting we expect no effects in the placebo thresholds.

We have already seen (in Figure 7) that crossing salary thresholds seems to "affect" the probability of treatment in the previous census. In the top row of Table 3 we see that the imbalance at salary thresholds persists when we control for the year of the census and the threshold being considered. This suggests that the imbalance in lagged treatment is not simply explained by variation in the extent of sorting over time or across thresholds. This imbalance does, however, mostly disappear when we include municipal covariates in column 5, which suggests that some of these municipal characteristics are unbalanced in a similar way, perhaps because they help explain which cities are able to sort. The second row indicates that we do not find a similar effect for the lagged running variable.

In the third and fourth rows of Table 3 we see evidence of imbalance in whether the council size changes at the threshold as well as in the year of the census being considered. This

[23]The covariates are the (log) number of nonprofits per person, the proportion of inhabitants who give blood, the ratio of young to old inhabitants, an indicator for the South, an indicator for whether the municipality is on the seaside, and the proportion of second homes in the municipality. When a given covariate is used as the outcome it is obviously omitted from the list of regressors.

imbalance probably arises for the reason discussed above: sorting is worse at thresholds where both council size and salary change (as shown in Table 2) and in earlier censuses (as shown in Figure 11 in the Online Appendix); in pooled data, therefore, the type of threshold and the vintage of the census is systematically unbalanced, which could cause bias in RDD estimates if the appropriate covariates are not used.

The rest of Table 3 reports similar analysis for a set of covariates we selected because we thought they might explain the aggregate sorting in Italy: two measures of social capital,[24] the proportion of young to old citizens, an indicator for whether the city is in the South of Italy, an indicator for whether the city is located by the sea, and the proportion of second homes. (A large second-home proportion may indicate more opportunities for selective precision.) We find no imbalance in these covariates in the raw RDD. In columns 2-5, we find some imbalance in the proportion of young to old citizens: the analysis indicates that observations just to the right of a threshold correspond to cities that currently have a somewhat older population than observations just to the left of the threshold. Similarly, we find imbalances in the proportion of second homes with borderline significant ($p < .1$), consistent with our speculation.

What can we conclude from this evidence on covariate imbalance at salary thresholds in Italy? The optimistic conclusion is that researchers can productively conduct RDD in Italy using multiple thresholds and/or multiple censuses as long as they include appropriate controls, which based on this analysis would include an indicator for lagged treatment, indicators for the year and the threshold, and controls for the age structure of the population and the proportion of second homes. The pessimistic conclusion is that there are many covariates we have not tested (and many that are unobservable and thus untestable), and thus RDD analysis is likely to be biased even after controlling for the set of covariates we have tested here. The clear implication from this analysis is that studies that pool data across thresholds and years should control for the threshold and year whenever sorting seems like a possibility; whether or not one wants to proceed with a population-threshold RDD setting with evidence of sorting depends, as we discuss in the next section, on what the next best design is.

[24]The measures we use (the number of nonprofit organizations per person and the rate of blood donations) are commonly used in the literature as measures of social capital; see Nannicini et al. (2013).

V. Concluding remarks

We have documented two serious problems with population-threshold RDDs in France, Italy, and Germany. Although important policies depend on population thresholds in each country, these policies often change along with other policies and municipalities seem to strategically manipulate population figures to end up on the desired side of relevant thresholds. We have discussed remedies that researchers might use to address confounded treatment and manipulative sorting; applying these remedies requires additional assumptions, which of course makes the analysis less compelling than a standard RDD. The practical question that remains is whether we should bother to undertake population-threshold RDDs in a setting where these remedies (and associated assumptions) are necessary.

The answer to this question of course depends on what the alternative is – i.e., what the next best research design is for addressing the research question. If the alternative is to carry out another population-threshold RDD in a setting that addresses the research question equally well but does not suffer from confounded treatment and sorting, clearly the alternative would be better. If the alternative is to conduct an observational study in a setting where the municipalities choose their own policies, the answer is less clear.

Ultimately, the choice depends on how much unobservable imbalance remains in the RDD and the observational study after we apply our various corrections, and how much these omitted variables affect the outcome in each setting; this in turn will depend on how well we understand the process by which municipalities choose their policies in the observational setting and how sorting takes place in the RDD, but also how well we can measure and control for the covariates that are unbalanced as a result of these processes. All of these considerations are subjective judgments and cannot be measured in the data. Substantive knowledge is thus necessary; the best answer may be to conduct both sets of analysis. What is most clear to us is that a population-threshold RDD should not be dismissed in favor of alternatives simply because there is evidence of confounded treatment or manipulative sorting. Observational studies have similar problems: policies tend to be correlated with each other in cross-section, and omitted variable bias is always a concern when units choose their own treatments. Given the difficulty of running experiments on consequential policies, it would be unwise to exclusively rely on

purely observational evidence and ignore findings from population-threshold RDDs when these quasi-experiments fall short of the ideal.

More broadly, we also emphasize that, despite the clear challenges of carrying out population-threshold RDDs in the three countries we study, none of our analysis implies that all such designs are problematic. Clearly, researchers should check for confounded treatment, sorting, and covariate imbalance whenever they conduct any RDD; the cases we have shown indicate that these problems can be systematic in some settings. But just as it would be a mistake to discard a specific RDD at the first sign of confounded treatment or manipulative sorting, it would also be a mistake to conclude based on our analysis that all population-threshold RDDs must suffer from the same problems.

References

Arnold, Felix and Ronny Freier. 2015. "Signature requirements and citizen initiatives: Quasi-experimental evidence from Germany." *Public Choice* 162(1):43–56.

Asatryan, Zareh, Thushyanthan Baskaran and Friedrich Heinemann. 2014. "The effect of direct democracy on the level and structure of local taxes.".
URL: *http://www.econstor.eu/bitstream/10419/90143/1/776404695.pdf*

Asatryan, Zareh, Thushyanthan Baskaran, Theocharis Grigoriadis and Friederich Heinemann. 2013. "Direct Democracy and Local Public finances Under Cooperative Federalism." *ZEW Working Paper* 13-038.
URL: *http://www.econstor.eu/bitstream/10419/90143/1/776404695.pdf*

Bagues, Manuel and Pamela Campa. 2012. "Gender Quotas, Female Politicians, and Public Expenditures: Quasi Experimental Evidence." *Econpubblica Working Paper* .
URL: *http://unicreditanduniversities.eu/uploads/assets/UWIN/CAMPA_PAPER.pdf*

Barone, Guglielmo and Guido De Blasio. 2013. "Electoral rules and voter turnout." *International Review of Law and Economics* 36:25–35.

Barreca, Alan I, Melanie Guldi, Jason M Lindo and Glen R Waddell. 2011. "Saving Babies? Revisiting the effect of very low birth weight classification." *The Quarterly Journal of Economics* 126(4):2117–2123.

Baskaran, Thushyanthan. 2012. "The flypaper effect: evidence from a natural experiment with Hessian municipalities." *MPRA Working paper* 37144.
URL: *http://mpra.ub.uni-muenchen.de/37144/1/MPRA_paper_37144.pdf*

Baskaran, Thushyanthan and Mariana Lopes da Fonseca. 2015. "Electoral competition and endogenous political institutions: quasi-experimental evidence from Germany." *Discussion Papers, Center for European Governance and Economic Development Research 237* .
URL: *http://www.econstor.eu/bitstream/10419/109036/1/821986465.pdf*

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti and Guido Tabellini. 2013. "The Political Resource Curse." *American Economic Review* 103(5):1759–1796.

Casas-Arce, Pablo and Albert Saiz. 2015. "Women and power: unpopular, unwilling, or held back?" *Journal of Political Economy, forthcoming* .

Caughey, Devin and Jasjeet S Sekhon. 2011. "Elections and the regression discontinuity design: Lessons from close US house races, 1942–2008." *Political Analysis* 19(4):385–408.

De Benedetto, Marco Alberto and Maria De Paola. 2014. "Candidates' Quality and Electoral Participation: Evidence from Italian Municipal Elections." 8102.
URL: *http://www.econstor.eu/bitstream/10419/96808/1/dp8102.pdf*

De Witte, Kristof, Benny Geys and Joep Heirman. 2015. "Strategic Housing Policy, Migration and Sorting around Population Thresholds." *Mimeo* .

Egger, Peter and Marko Koethenbuerger. 2010. "Government spending and legislative organization: Quasi-experimental evidence from Germany." *American Economic Journal: Applied Economics* pp. 200–212.

Eggers, Andrew. 2015. "Proportionality and Turnout: Evidence from French Municipalities." *Comparative Political Studies* 48(2):135–167.

Eggers, Andrew C, Anthony Fowler, Jens Hainmueller, Andrew B Hall and James M Snyder. 2015. "On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races." *American Journal of Political Science* 59(1):259–274.

Ferraz, Claudio and Frederico Finan. 2009. "Motivating politicians: The impacts of monetary incentives on quality and performance.".
URL: *http://www.econstor.eu/bitstream/10419/35111/1/562100016.pdf*

Foremny, Dirk, Jordi Jofre Monseny and Albert Solé Ollé. 2015. "Hold that ghost: Local government cheating on transfers." *IIPF Conference Papers* .
URL: *https://www.cesifo-group.de/dms/ifodoc/docs/Akad_Conf/CFP_CONF/CFP_CONF_2015/pse15-van-der-Ploeg/Papers/pse15-Olle2.pdf*

Fujiwara, Thomas. 2011. "A Regression Discontinuity Test of Strategic Voting and Duverger's Law." *Quarterly Journal of Political Science* 6(3-4):197–233.

Gagliarducci, Stefano and Tommaso Nannicini. 2013. "Do better paid politicians perform better? Disentangling incentives from selection." *Journal of the European Economic Association* 11(2):369–398.

Grembi, Veronica, Tommaso Nannicini and Ugo Troiano. 2014. "Policy Responses to Fiscal Restraints: A Difference-in-Discontinuities Design." *Harvard Economics Department Working Paper* .
URL: *http://www.econstor.eu/bitstream/10419/68212/1/733920993.pdf*

Gulino, Giorgio. 2014. "Do Electoral Systems Affect the Incumbent Probability of Re-election? Evidence from Italian Municipalities." *Working paper presented at the EEA/ESEM* .
URL: *http://www.eea-esem.com/files/papers/EEA-ESEM/2014/2511/Toulose_ESEM_Giorgio_Gulino.pdf*

Hopkins, Daniel J. 2011. "Translating into Votes: The Electoral Impact of Spanish-Language Ballots." *American Journal of Political Science* 55(4):814–830.

Imbens, Guido and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79(3):933–959.

Imbens, Guido W and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142(2):615–635.

Koethenbuerger, Marko. 2012. "Do Political Parties Curb Pork-Barrel Spending? Municipality-Level Evidence from Germany." *CESifo Discussion Paper* 14-15.
URL: *http://www.cesifo-group.de/dms/ifodoc/docs/Akad_Conf/CFP_CONF/CFP_CONF_2014/Conf-pse14-VanderPloeg/Paper/pse14_Koethenbuerger__19108240_en.pdf*

Lee, David S and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48:281–355.

Litschig, Stephan. 2012. "Are Rules-based Government Programs Shielded from Special-Interest Politics? Evidence from Revenue-Sharing Transfers in Brazil." *Journal of Public Economics* 96:1047–1060.

Litschig, Stephan and Kevin M. Morrison. 2013. "The Impact of Intergovernmental Transfers on Education Outcomes and Poverty Reduction." *American Economic Journal: Applied Economics* 5(4):206–240.

Litschig, Stephan and Kevin Morrison. 2010. "Government spending and re-election: Quasi-experimental evidence from Brazilian municipalities." *UPF Discussion Paper* .
URL: *http://repositori.upf.edu/bitstream/handle/10230/6349/1233.pdf?sequence=1*

Lyytikäinen, Teemu and Janne Tukiainen. 2013. "Voters are rational." *Government Institute for Economic Research Working Papers* (50).
URL: *https://vatt.fi/file/vatt_publication_pdf/wp50.pdf*

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics* 142(2):698–714.

Mukherjee, Mukta. 2011. "Do Better Roads Increase School Enrollment? Evidence from a Unique Road Policy in India." *Syracuse University* .

Nannicini, Tommaso, Andrea Stella, Guido Tabellini and Ugo Troiano. 2013. "Social Capital and Political Accountability." *American Economic Journal: Economic Policy* (5):222–250.

Pellicer, Miquel and Eva Wegner. 2013. "Electoral Rules and Clientelistic Parties: A Regression Discontinuity Approach." *Quarterly Journal of Political Science* 8(4):339–371.

Pettersson-Lidbom, Per. 2006. "Does the Size of the Legislature Affect the Size of Government? Evidence from Two Natural Experiments." *Stockholm University Working Paper* .
URL: *http://www.gsb.stanford.edu/sites/default/files/documents/12.6.05%20Lidbom.pdf*

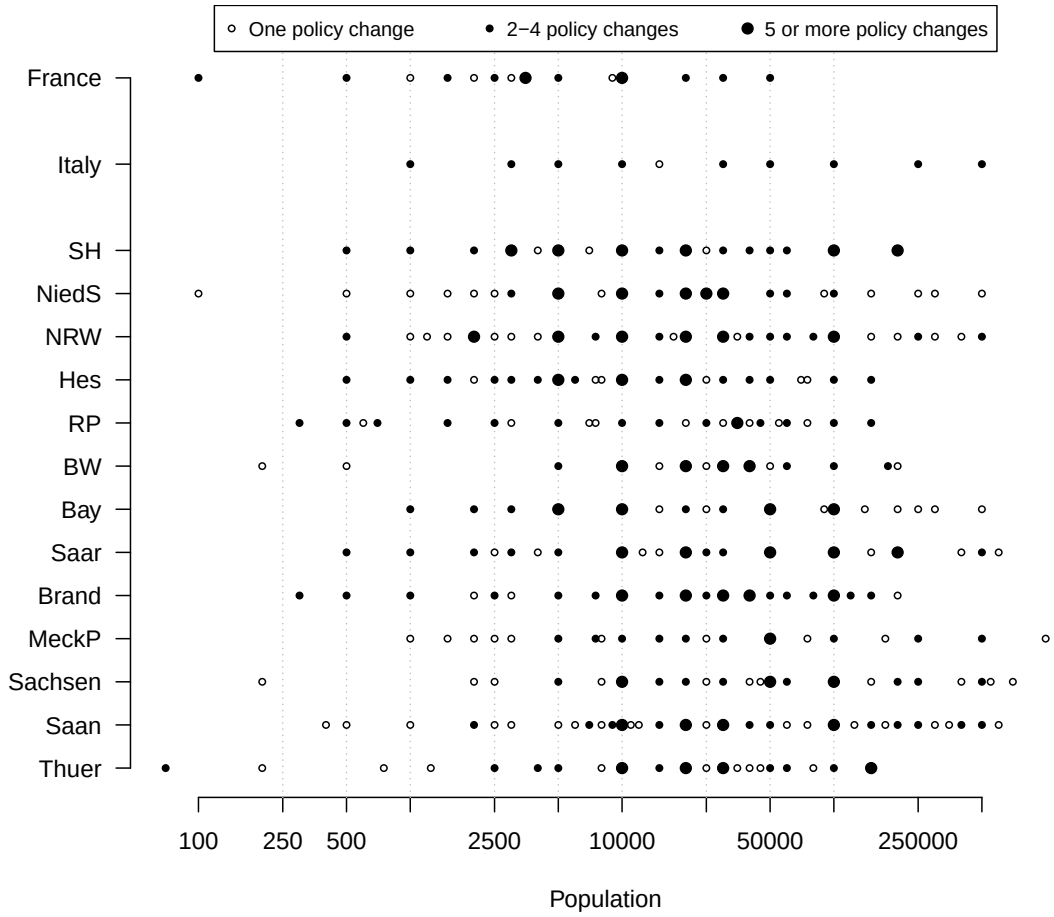
Pettersson-Lidbom, Per. 2012. "Does the Size of the Legislature Affect the Size of Government: Evidence from Two Natural Experiments." *Journal of Public Economics* 98(3–4):269–278.

Urquiola, Miguel and Eric Verhoogen. 2009. "Class-size caps, sorting, and the regression-discontinuity design." *The American Economic Review* 99(1):179–215.

van der Linde, Daan, Swantje Falcke, Ian Koetsier and Brigitte Unger. 2014. "Do Wages Affect Politicians' Performance? A regression discontinuity approach for Dutch municipalities." *Utrecht University School of Economics Discussion Paper* 14-15.

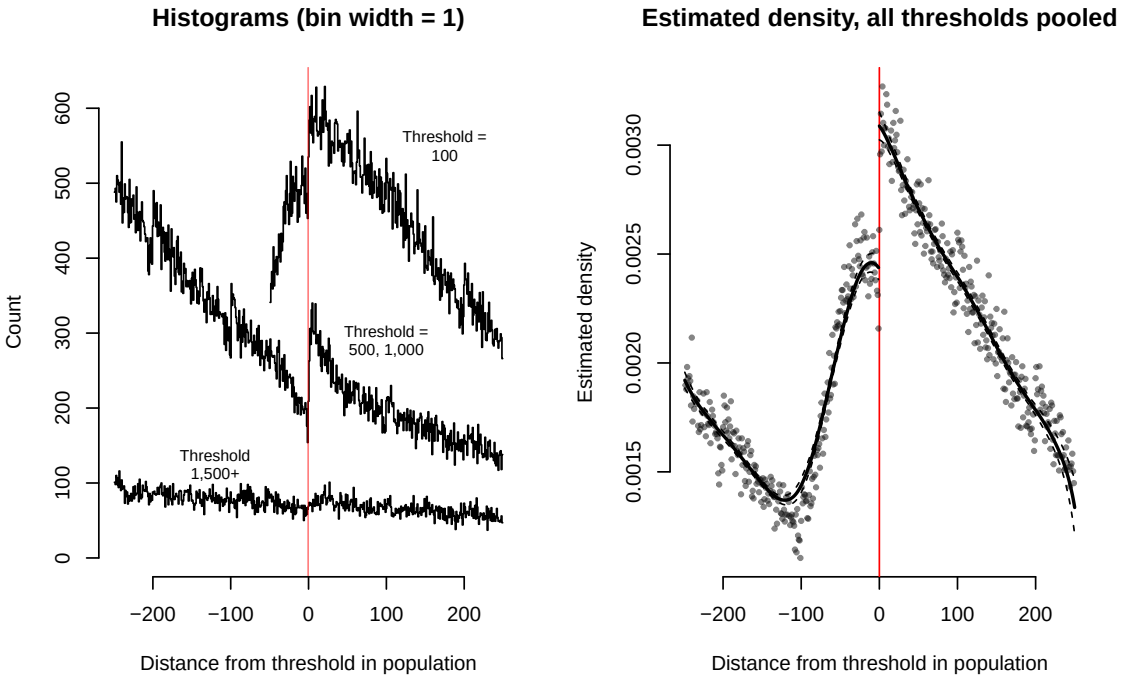
Vernet, Maurice. 1972. “Population de la France : le nombre et la loi.” *Economie et statistique* 36(1):3–19.

Figure 1: Population thresholds at which municipal policies change: France, Italy, and German states



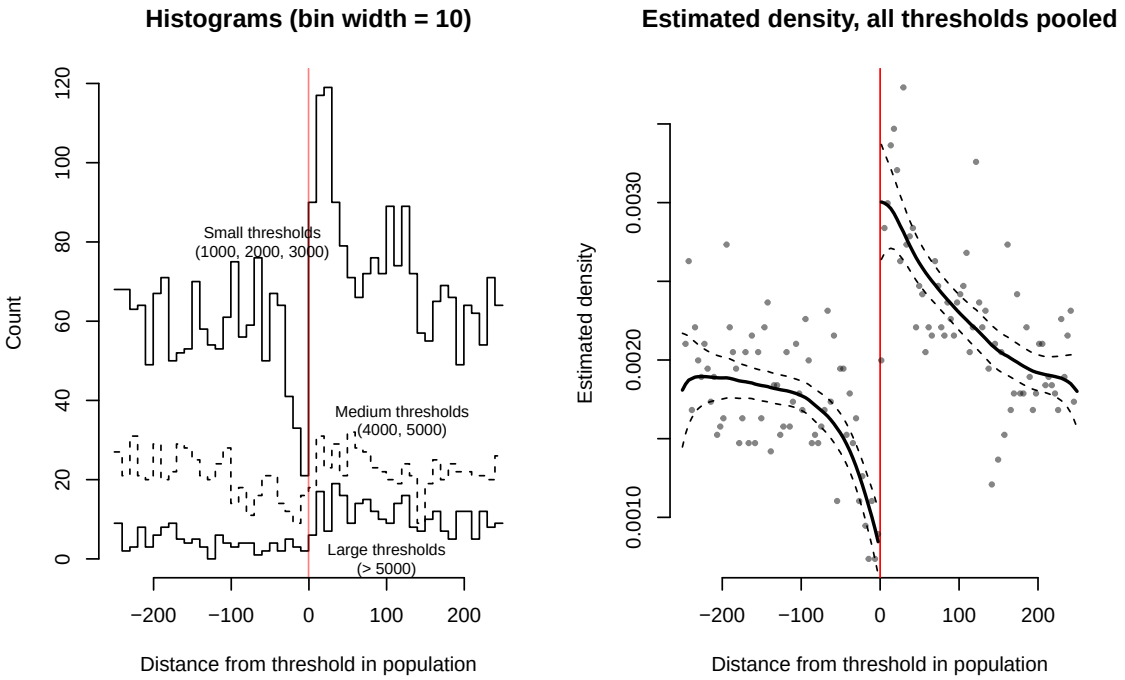
NOTE: Each dot indicates a population threshold at which a policy changes; solid dots indicate more than one policy changing at the same threshold.

Figure 2: Sorting in municipal population in France, 1962-2011 pooled



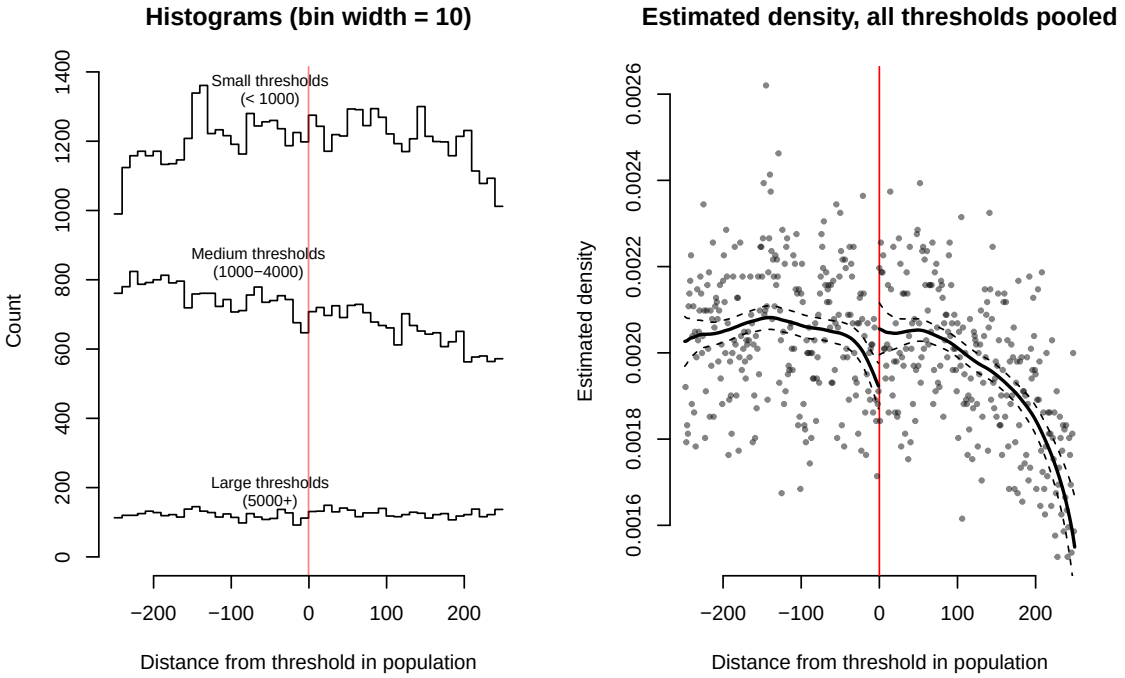
NOTE: The left plot depicts three histograms, one for each group of thresholds (100; 500 & 1,000; 1,500 and larger). In each case the bin width is 1, meaning that the top of the line indicates the number of data points (municipality-years) with population that is exactly a given amount (e.g. 50 inhabitants) from the threshold. The right plot depicts the McCrary analysis for all cases pooled.

Figure 3: Sorting in municipal population in Italy, 1961-2001 pooled



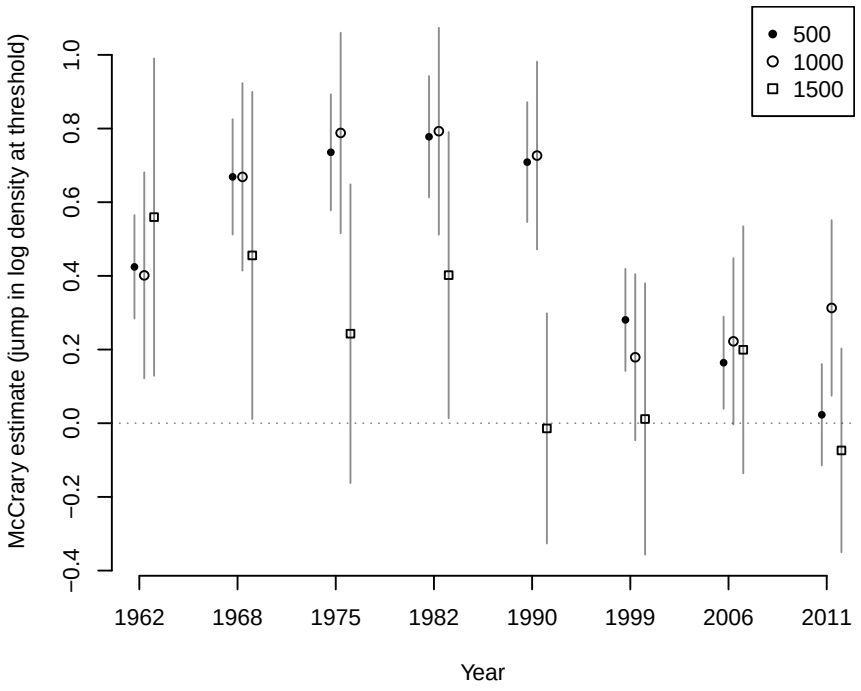
NOTE: In the left plot the bin width is 10, meaning that the top of the line indicates the number of data points (municipality-years) with population that is in a given interval (e.g. 40-49 inhabitants) from the threshold. Otherwise see notes to Figure 2.

Figure 4: Sorting in municipal population in German states, 1998-2007 pooled



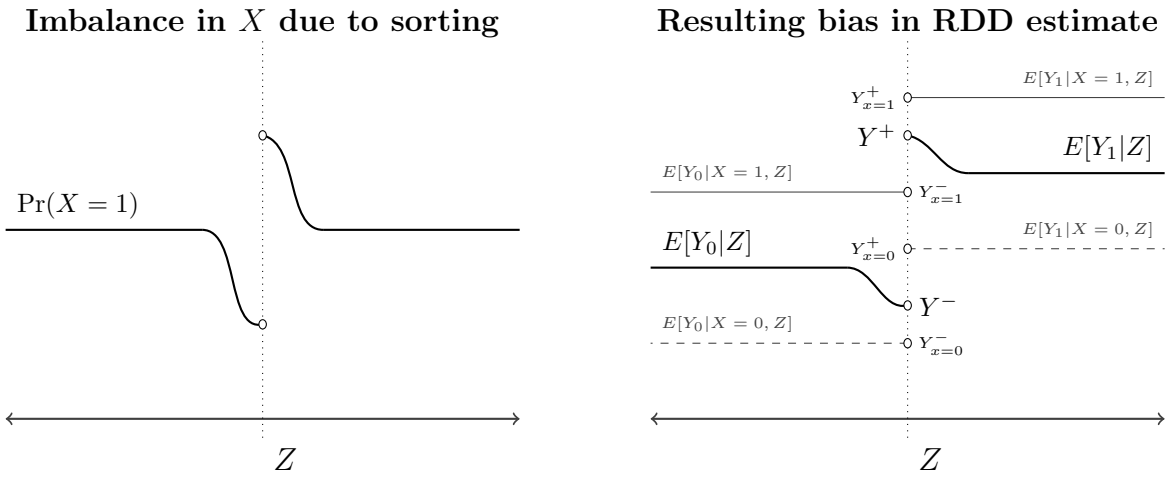
NOTE: In the left plot the bin width is 10, meaning that the top of the line indicates the number of data points (municipality-years) with population that is in a given interval (e.g. 40-49 inhabitants) from the threshold. Otherwise see notes to Figure 2.

Figure 5: Sorting over time in France at the 500, 1,000, and 1,500 population thresholds



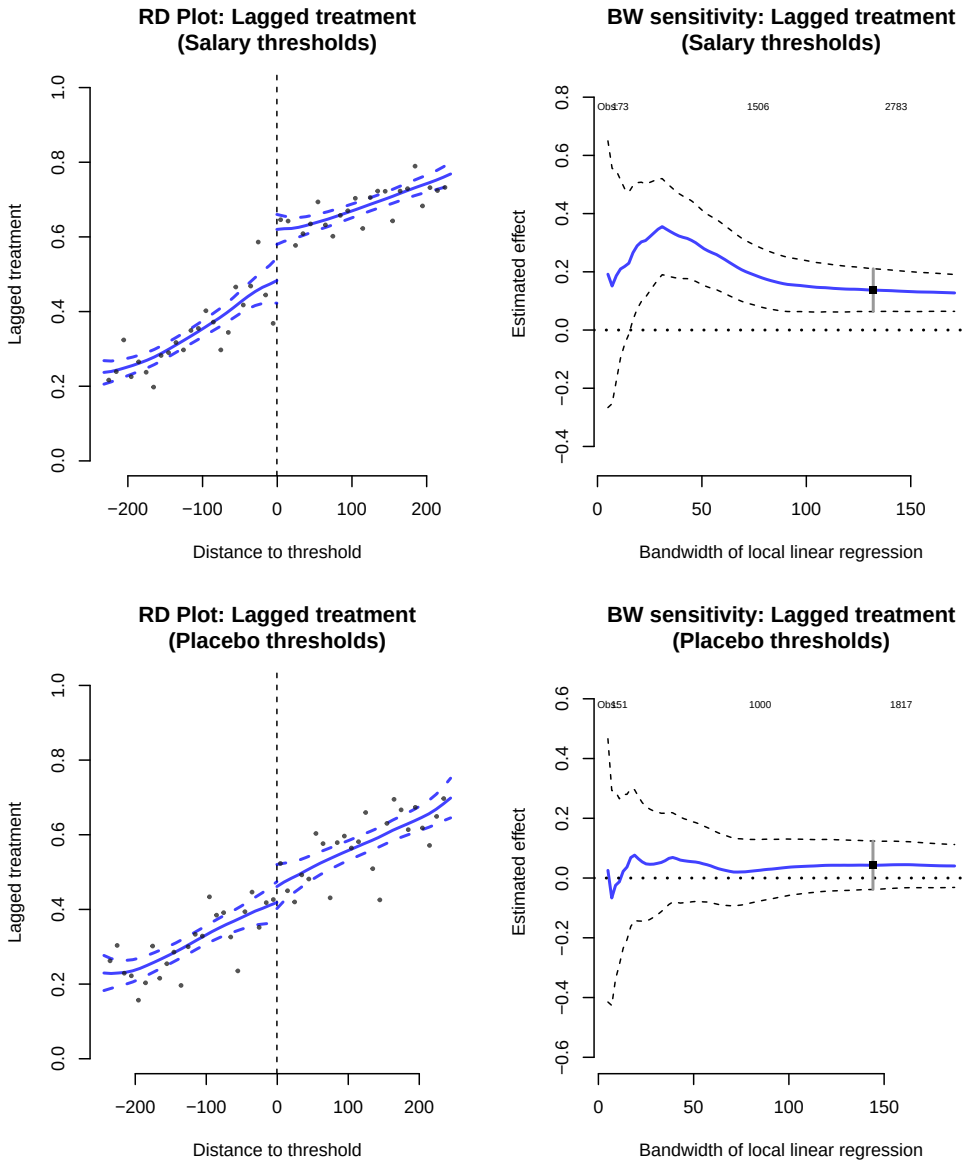
NOTE: Each point corresponds to the McCrary test statistic (the estimated jump in the log density of the running variable at the relevant threshold) for a given threshold in a given census in France. Lines show 95% confidence intervals.

Figure 6: Best case scenario for addressing manipulative sorting with covariate adjustment



NOTE: Suppose covariate X is not continuous at the threshold due to sorting, as shown in the left plot. If X is also related to the outcome, as shown in the right plot, then the usual RDD estimate $(Y^+ - Y^-)$ will be biased. The bias due to imbalance in X is removed if the RDD is estimated conditional on X, e.g. as $Y^+_{x=1} - Y^-_{x=1}$.

Figure 7: Imbalance in lagged treatment in Italy



NOTE: The dependent variable in the RDD analysis above is "lagged treatment" – an indicator for whether a municipality was above a given threshold in the *previous* census, given that it was close to that threshold in the current census. The left panel in each pair shows the dependent variable in binned means of the running variable (gray dots) and the local linear regression estimate at the Imbens-Kalyanaraman optimal triangular bandwidth; the right panel shows the sensitivity of the estimated effect to the bandwidth, where the optimal is shown with a black dot.

Table 1: Population thresholds in French municipalities

	Policy changes at k inhabitants (in tsd)													
	0.1	0.5	1	1.5	2	2.5	3	3.5	5	9	10	20	30	50
Council size	x	x		x		x		x	x		x	x	x	x
Salary of mayor and deputy mayors		x	x					x			x	x		x
Max. number of deputy mayors	x	x		x		x		x	x		x	x	x	x
Max. number of non-resident councilors	x	x												
Must have a cemetery					x									
Prohibition on commercial water supply							x							
Campaign leaflets subsidized						x								
Council must approve property sales						x								
Electoral system – PR or plurality								x						
Gender parity								x						
Outsourcing scrutiny								x						
Council must debate budget prior to vote								x						
Committees follow PR principle								x						
Amount of paid leave for council work								x			x		x	
Commission on accessibility									x					
Max. electoral expenditure										x				
Outsourcing commission											x			
Max. municipal tax on salaries											x			x
Debt limit												x		

NOTE: The table identifies population thresholds (in thousands) at which given policies change. This is a partial list of policies, chosen to highlight the variety of policies that depend on population thresholds and the extent to which the same threshold often determines multiple policies. *Source*: French legal code.

Table 2: Summary of McCrary sorting tests

Sample	France		Italy		Germany	
	# of thresholds (# of close obs)	McCrary Test statistic	# of thresholds (# of close obs)	McCrary Test statistic	# of thresholds (# of close obs)	McCrary Test statistic
Total						
All years, all thresholds	21	0.238***	7	1.328***	195	0.068***
	(311,392)	(0.014)	(4,756)	(0.136)	(101,520)	(0.025)
Specific thresholds						
Salary increase	14	0.497***	6	1.331***	78	0.135***
	(140,421)	(0.026)	(4,730)	(0.134)	(11,579)	(0.061)
Salary increase (no council)	7	0.533***	3	0.840***	21	0.001
	(35,329)	(0.049)	(2,125)	(0.211)	(447)	(0.321)
Council increase	15	0.215***	3	1.909***	120	0.071***
	(267,558)	(0.015)	(2,605)	(0.197)	(81,669)	(0.026)
Council increase (no salary)	12	0.139***	0	n.a.	63	0.063***
	(162,466)	(0.018)	(0)	n.a.	(70,537)	(0.029)
Council and/or salary increase	21	0.240***	6	1.331***	141	0.072***
	(302,887)	(0.014)	(4,730)	(0.134)	(82,116)	(0.027)
Council and salary increase	7	0.475***	3	1.909***	57	0.149***
	(105,092)	(0.029)	(2,605)	(0.197)	(11,132)	(0.063)
Threshold size						
Small thresholds (<3500)	7	0.237***	2	1.644***	61	0.054***
	(306,520)	(0.014)	(3,295)	(0.178)	(93,873)	(0.026)
Big thresholds (>=3500)	14	0.239*	5	0.700***	134	0.216***
	(4,872)	(0.122)	(1,461)	(0.247)	(7,647)	(0.077)
Placebo thresholds	20	0.008	9	-0.044	186	-0.009
	(215,986)	(0.018)	(2,800)	(0.133)	(85,326)	(0.025)

Notes For each test, we report four numbers: the number of unique population thresholds (e.g. 14 in the first test for France) at which we observe municipalities with populations within 250 inhabitants of the threshold; the number of observations within 250 inhabitants of these thresholds (e.g. 273,274); the estimated difference log frequency above vs. below the threshold (e.g. 0.256) and the standard error of that estimate (0.015).

Table 3: "Effects" of crossing threshold on covariates (Italy)

	Obs [BW]	Thresholds where salary changes (1)	(2)	(3)	(4)	(5)	"Thresholds" w. no changes (6)	(7)	(8)
Lagged treatment	2592	.138***	.113***	.128***	.108***	.041	.024	.028	-.055
	[132]	(.037)	(.037)	(.037)	(.037)	(.036)	(.041)	(.041)	(.040)
Lagged running variable	2522	80.115	55.194	84.583	63.567	27.862	95.758	124.452	1.985
	[128]	(63.630)	(63.642)	(58.183)	(57.660)	(53.600)	(97.363)	(79.532)	(72.847)
Both salary and council size change	1590	.268***	.207***						
	[81.7]	(.049)	(.046)						
Year of census	3056	-7.091***		-3.172***			-2.871*	-1.988	
	[158.8]	(1.045)		(.889)			(1.620)	(1.386)	
Population at threshold (in 1000s)	1678	.147	-.091				.346		
	[86.4]	(.401)	(.396)				(.532)		
Log nonprofits/person	1647	.001	.011	.026	.023	-.000	.004	.002	-.052
	[84.5]	(.057)	(.058)	(.058)	(.058)	(.058)	(.045)	(.045)	(.045)
Proportion donating blood	1570	-.003	-.003	-.003	-.003	-.004**	-.001	-.001	-.001
	[80.5]	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)	(.002)
Log young/old ratio	1815	.045	-.089**	-.045	-.089**	-.086**	.027	-.007	-.035
	[93.8]	(.056)	(.040)	(.047)	(.040)	(.039)	(.053)	(.045)	(.034)
South	1777	.037	.037	.036	.033	-.023	-.012	-.010	-.031
	[91.3]	(.048)	(.048)	(.048)	(.048)	(.040)	(.050)	(.050)	(.041)
Seaside	2077	-.026	-.033	-.031	-.032	-.033	-.012	-.017	-.044**
	[106.9]	(.023)	(.023)	(.023)	(.023)	(.024)	(.020)	(.020)	(.020)
Proportion vacation homes	1897	.025	.043*	.033	.037*	.037*	.032	.040*	.021
	[97.8]	(.022)	(.022)	(.022)	(.022)	(.022)	(.022)	(.021)	(.020)
Year dummies:			✓		✓	✓			✓
Threshold dummies:				✓	✓	✓		✓	✓
Other municipal characteristics:						✓			✓

Notes: Each point estimate comes from a different RDD analysis in which the row variable is the dependent variable and we pool data from multiple censuses and population thresholds. Model (1) includes no extra control variables; Model (2) includes a dummy for threshold type (e.g. mayor's salary, council size, both); Model (3) includes a dummy for each threshold (e.g. 1000, 2000); Model (4) includes controls for municipality characteristics (e.g. South, blood donations) other than the dependent variable; Model (5) includes threshold dummies and municipal characteristics.

Supplementary information

A. Issues in detecting sorting

The McCrary test based on the work by McCrary (2008) has become a standard method to test for sorting in RDD settings. The test checks for manipulation of the RDD running variable by closely examining the distribution of this variable around the threshold. Importantly, the test has been designed for continuous variables around a single threshold. In this appendix, we illustrate two difficulties that researchers need to consider when applying this test to a setting where the underlying RDD running variable is discrete and/or the researcher pools different thresholds.

Issue 1: Bin size selection with discrete variables

The first issue relates to the selection of one of the key parameters in the test, i.e. bin size. The McCrary test proceeds in two steps. The bin size is important in the first step of the McCrary test which produces an undersmoothed histogram from the data. Given this bin size selection, the histogram is in the second step smoothed using a local linear regression (involving a selection of the bandwidth h).

McCrary (2008) suggests the following bin size selection procedure:

$$\widehat{b} = 2\widehat{\sigma}n^{-1/2} \tag{A.1}$$

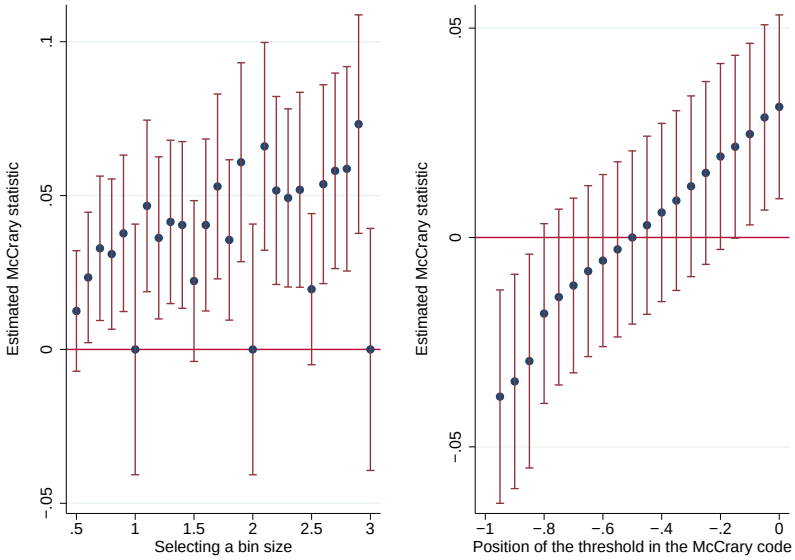
where $\widehat{\sigma}$ is the sample standard deviation of the forcing variable. Naturally, $\widehat{b}$ is not confined to the values of the discrete variable distribution.

Let us first consider how the bin size works in discrete distributions. Assume that we have a discrete distribution which takes the values -10 to 9 in increments of 1. The threshold is at 0 (which is also counted into treatment). Further assume that we observe 100 observations for each discrete value. Note that bin size, x, in the McCrary is defined in the following way: $\{..., [-2x, -x), [-x, 0), [0, x), [x, 2x), ...\}$.

Now consider that we use a cut-off point exactly at 0 and choose a bin size of 1. That would give us 20 equal sized bins with 100 observations each. If we now vary the bin size to 0.5, we observe a crucial imbalance. To the right of the threshold (the treatment side) we first have a bin [0,0.5) with 100 observations, followed by an empty bin [0.5,1). On the left of the threshold, we observe the reverse. Here, the first bin [-0.5,0) is empty, and the second bin [-1,0.5) holds 100 observations. A bin size smaller than the discrete steps in the distribution, thus, creates empty bins which are distributed asymmetrically to the right and left of the threshold.

Even for non-integer bin sizes larger than one you will tend to have more possible population values per bin to the right than to the left. Consider a bin size of 1.5 in the above example. The first bin to the right [0,1.5) now includes 200 observations while the second bin [1.5,3) holds 100 observations only. Again, the reverse is true for the left side. Here, the first bin [-1.5,0) includes only 100 obs, and the second bin [-3,1.5) includes 200 obs. When doing a McCrary on a discrete running variable, if you use a non-integer bin size there will be a bias toward finding a jump in population. That is because for any bin size x and any $k > 0$ the number of integer values in [0, kx) is weakly larger than the number of integer values in [$-kx$, 0).

Figure 8: Simulating different bin sizes and cut-off values



NOTE: In this figure, we manipulate the bin size (left panel) or the exact position of the cutoff (right panel) in the McCrary routine. The patterns signify that the choice of those parameters can be crucial in a discrete settinng. *Source:* Own calculations.

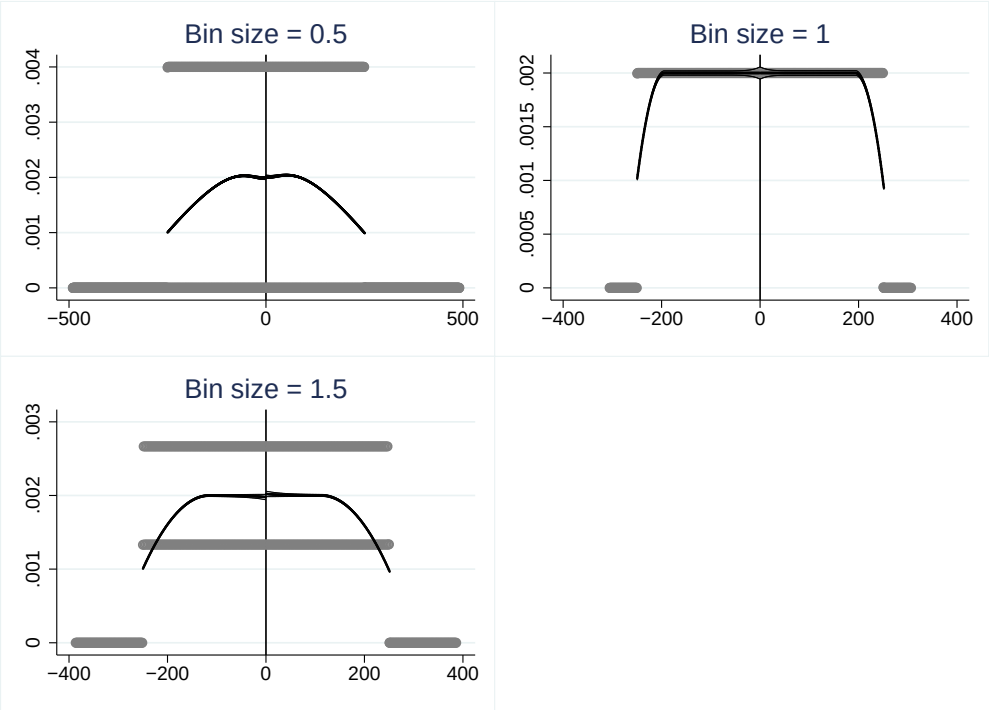
To illustrate the problems in the estimation of the McCrary statistic, we simulated a data set consisting of 200,000 observations (perfectly) uniformly distributed to 500 discrete values [-250,249] in increments of one. We set the threshold to 0 (zero is included in the treatment). Given this setup, the data are constructed such that no sorting of any magnitude should be found. Implementing the McCrary Stata routine (using the cut-off point at 0 and leaving the bin size and bandwidth to be calculated by the algorithm), we obtain an estimate of 0.031 (0.011), a bin size of 0.65 and a bandwidth of 191.3 indicating a positive sorting result.

In the following we manipulated the bin size. In the left panel of figure 8, we vary the bin size between 0.5 – 3. The figure shows that depending on the exact bin size the estimated McCrary statistic varies significantly and often signals a false positive sorting result.

In panel 2 of figure 8, we illustrate a similar problem of discrete bins when the cut-off value in the McCrary test is manipulated. Note, that due to the discrete values there is no true cut-off point. Any value between (-1,0] could be said to be the cut-off point. In the graph we highlight that the estimated McCrary statistic can vary from significant negative to significant positive depending on the exact positioning of the cut-off value.

In figure 9, we again highlight the problem that certain bin sizes create artifically jumps in the density of the running variable (by construction). In the left upper panel, we use a bin size of 0.5 which creates empty bins. With a bin size of 1 (upper right panel) all bins are of similar size. Increasing the bin size to 1.5 (lower left panel) again highlights that the choice of bin size creates distinct sets of bins with more or less observations in it.

Figure 9: Issues with bin size and discrete values in McCrary tests



Notes: The figure shows the graphical output of the McCrary routine. The focus is on the distinct sets of observations when there is a bin size of 0.5 (upper left panel) or a bin size of 1.5 (lower left panel). Only when there is a bin size of 1 do we see that all the simulated observations are correctly bined. *Source:* Own calculations.

It is important to understand how the bin size interacts with the second important parameter in the test statistic, the bandwidth h. For the case of continuous variables, McCrary (2008) finds that the choice of bin size is unimportant as long as h is large compared to b (he suggests $h/b > 10$). Even for discrete variables, any asymmetric grouping in the bins will become less important with a larger bandwidth. However, in our simulations we found that while an increased bandwidth had a mitigating effect on the bias from choosing a particular bin size, a false positive test result could be obtained even with $h/b > 100$ or more.

We see two solutions to this problem. First, one can restrict the bin size in the McCrary test to the set of integers. Guidance as to which multiple of the increment to chose can be obtained from the original McCrary test. For a bin size of below 1 in the original McCrary, the minimum bin size of 1 should be set. For larger bin sizes, the closest multiple of an increment value of the discrete running variable distribution should be used.

A second solution works as follows. Redefine the breakpoint, c, in the McCrary away from zero to half the distance between 0 and the next integer in the discrete distribution (in our above example this would be -0.5). Now, define the bins to the right and left of this new breakpoint to be $\{..., [c-2x, c-x), [c-x, c), (c, c+x], (c+x, c+2x], ...\}$. This solves the asymmetry problem, does not throw out any data, and allows the McCrary process to pick the bin size and bandwidth with the original algorithm.

A third, but inferior solution is to drop the observations at 0 from the analysis. This, however, comes at the

cost of losing (potentially crucial) information just around the threshold.

Issue: Pooling different thresholds

One of the particular issues that is linked to RDDs based on population thresholds is that researchers often try to pool the data from different thresholds. This is due to the fact that the type of policies that we evaluate around population thresholds often allow for such pooling. Council size, the salary of mayors or transfers increase discontinuously at one threshold and then again increase further at a larger threshold. By pooling the data, researchers increase the power of the test (often crucial for RDDs which typically require a lot of data). While pooling may those be a good idea, it also creates a number of issues especially for detecting potential sorting effects.

Generally, pooling will give more weight to thresholds with many observations. For the application with population thresholds and given that we choose an absolute scale, the problem is further aggravated as a small town has a larger chance to be close to the small threshold compared to large towns in proximity of the larger threshold. This implies that even more weight is given to smaller towns and smaller thresholds.

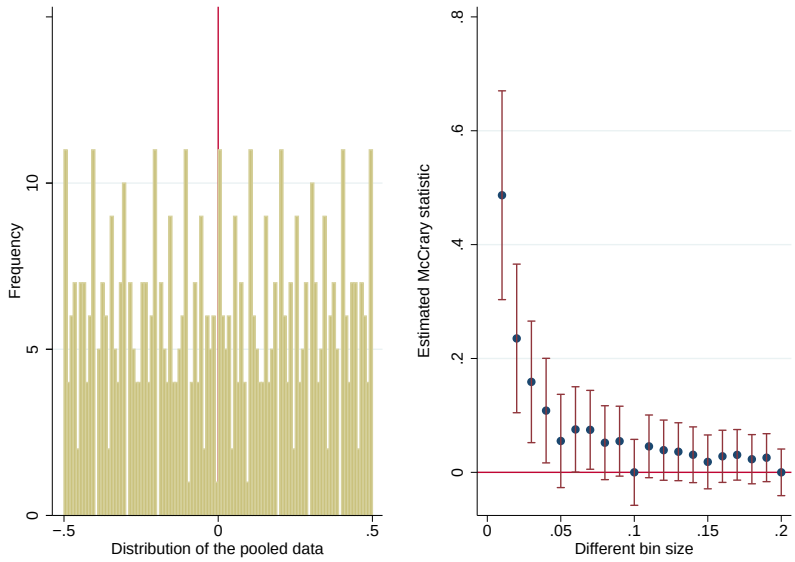
Pooling thresholds can cloud the issue of sorting when the sorting is positive at some and negative at other thresholds. While a larger council size may be a desirable features when the council sizes are still small (leading to positive sorting), we may consider the chance that larger council size eventually looses its attractiveness and may even become a negative aspect for a town (leading to negative sorting). When examining the overall pooled sorting effects, the negative and positive sorting may cancel and hide the extent of the sorting. This is related to the more general point made by McCrary (2008) that sorting around one threshold could go in two directions at the same time (e.g., in our application some municipalities want to be treated and sort above and other may choose to avoid the treatment and want to stay below the cutoff). This implies that researchers cannot fully exclude the possibility of sorting even when the sorting test is negative. Note that a similar concern applies when multiple treatments change at one threshold. While some municipalities sort above the threshold to benefit from some policies (e.g., larger transfers, higher salary for the mayor), other towns may try to stay below the cutoff to avoid another policy (e.g., tighter financial oversight).

Finally, pooling over thresholds requires a choice of the joint measure of deviation. While there are many different ways to define a scale, two alternatives are natural candidates: absolute and relative deviation from the threshold. The absolute deviation (e.g., number of inhabitants from the threshold) is the easiest and most harmless alternative. The approach of the relative deviation (percent deviation from the threshold) also has an intuitive appeal mainly as we can imagine larger political units to be prone to larger shifts in population. We therefore like the aspect that the relative deviation includes more observations in a close band in the case of large thresholds. The problems with pooling on a relative scale start when the underlying score variable is discrete (as population).

Consider the case of the following population thresholds. Say, a policy changes at 10 different thresholds: 1,000, 2,000, ..., 10,000. Also assume that we have a universe of towns in the size of 500 to 10500 with exactly one town at each particular value (1 town of size 500, 1 of size 501,... 1 of size 10499, 1 of size 10500).

Now, consider a pooling strategy with absolute deviations. For each town we measure the shortest absolute distance to the next threshold and stack the samples on top of each other. We end up with a distribution

Figure 10: Pooling different thresholds



NOTE: This graph highlights the uneven distribution of an initially uniformly distributed universe of towns when pooling is done on a relative scale (left panel). Also, we show how the McCrary routine estimates in this flawed pooling exercise which turn out to give false positive test results when the bin size is very small (right panel). *Source:* Own calculations.

(similar to the case above) in which we have values between -500 to 500 and exactly 10 observations for each integer (including the zero). We are back to the problems described above, however, no additional issue arises.

We now turn to the case of relative scaling. Say, we measure the distance to the next threshold in the form $\frac{(z-c)}{c}$, where z is the size of the town and c is the closest threshold (a log transformation will give a similar representation). Using this relative transformation results in a particular type of discrete distribution. Each town that has a size of exactly a threshold value will get the relative distance of 0. There are 10 such towns. Now, the town that has 1001 will be at 1/1000, so will be the towns at 2002, 3003, 4004,..., 10010. Thus, there are also 10 observations at this discrete value in the distribution. However, the town at 2001 will be at the 1/2000 point in the distribution. Here, there are only towns at 4002, 6003, 8004 and 10005. Hence, there are only 5 towns in this discrete bin. In the extreme, consider the town at 10001 which stands alone at the 1/10000 bin. Similarly, towns at 9001, 8001, 7001, and 6001 stand alone at their particular distribution value. The same pattern occurs on the other side of the thresholds (here, the observations at 9999, 8999, 7999, 6999, 5999 stand by themselves).

Estimating the sorting in this case is problematic when the zeros count into either of the sides. Assume treatment starts at zero (to the right side). We now have a bin at zero which has (by construction) observations from all thresholds. To the left of this bin (on the non-treatment side) there are 5 bins which relate to only one threshold respectively and in our example hold only one observation. For small bin sizes, the relative pooling creates an asymmetry exactly at the threshold.

We illustrate the issue by simulating the above example. The left panel of figure 10 illustrates the resulting histogram. Noticeable, the histogram is mirrored to the right and left of the zero. Depending on whether the zero is counted to the right or to the left, we can see that there exists a sorting issue right at the threshold.

In the next step, we estimated the McCrary test statistic on a sample where we duplicated every observation above 20 times (this brings us to a comparable sample size as above). To estimate the McCrary and avoid the first issue (see above) we set the breakpoint to -0.005. In panel 2 of figure 10, we vary the bin sizes between 0.01 and 0.2.[25] We find that for small bin sizes the McCrary test signals a false positive and significant sorting result which is entirely driven by the particular issue related to the zeros.

While the bin size issue (see above) can be remedied without loss of information, the second problem concerning the zeros in pooled data, when the distance is measured in relative terms, cannot be solved. The only technical solution to obtain correct inference on the scope of sorting in this case is to drop the observations at zero altogether. However, in doing this, we lose (potentially critical) information.

[25]Here, we also manipulated the bandwidth to be exactly 10 times the bin size. If we let the algorithm choose a bandwidth, we do find positive point estimates, however, they are not statistically different from zero. The reason is that the algorithm chooses a bandwidth which is more than 200 times larger than the bin size and thus smooths away any difference at the threshold.

Table 4: RDD studies using population thresholds

Authors	Publication status	Research focus	Country Time	Thresholds
	Panel 1: Studies using a simple RDD			
Egger & Koethenbuerger	2010, AEJ:App	Council size	Germany 1984-2004	1'000, 2'000, 3'000 5'000, 10'000, 20'000 30'000, 50'000, 100'000 200'000
Pettersson-Lidbom	2011, JPubE	Council size	Finland 1977-2002	2'000, 4'000, 8'000 15'000, 30'000, 60'000 120'000, 250'000, 400'000
			Sweden 1977-2002	12'000, 24'000, 36'000
Fujiwara	2011, QJPS	Single round vs. runoff	Brazil 1996-2008	200'000
Hopkins	2011, AJPS	Ballots	US 2005-2006	5% of citizen, 10,000
Litschig	2012, JPubE	Sorting	Brazil 1991	17 brackets
Barone & de Blasio	2013, IRLE	Single round vs. runoff	Italy 1993-2000	15'000
Brollo, Nannicini, Perotti & Tabellini	2013, AER	Transfers	Brazil 2001-2008	10'189, 13'585, 16'981 23'773, 30'564, 37'356 44'148
Gagliarducci & Nannicini	2013, JEEA	Wage of mayors	Italy 1993-2001	5'000
Litschig & Morrison	2013, AEJ:App	Transfers	Brazil 1982-1988	10'189, 13'585, 16'981
Pellicer & Wegner	2013, QJPS	Election rules	Morocco	25,000

continued on next page ...

... continued from previous page

Authors	Publication status	Research focus	Country Time	Thresholds
			2003, 2009	
Hinnerich-Tyrefors & Pettersson-Lidbom	2014, Econometrica	Direct democracy	Sweden 1918-1938	1'500
Arnold & Freier	2015, PuCh	Signature requirements	Germany 1995-2009	10'000, 20'000, 30'000, 50'000, 100'000
Eggers	2015, CPS	Electoral rule	France 2001-2008	3'500
Bordignon, Nannicini & Tabellini	R&R, AER	Single round vs. runoff	Italy 1993-2007	15'000
Ferraz & Finan	2012, WP	Wage of mayor	Brazil 2004-2008	10'000, 50'000, 100'000 300'000, 500'000
Mukherjee	2011, WP	Infrastructure	India 2001-2009	500
Baskaran	2012, WP	Transfers	Germany 2001-2010	5'000, 7'500, 10'000 15'000, 20'000, 30'000 50'000
Hirota & Yunoue	2012, WP	Council size	Japan	2'000, 5'000, 10'000 20'000, 50'000, 100,000 200'000, 300'000, 500'000 900'000
Egger & Koethenbuerger	2013, WP	Council size	Germany 1984-2004	1'000, 2'000, 3'000 5'000, 10'000, 20'000 30'000, 50'000, 100'000 200'000
Lyytikäinen & Tukiainen	2013, WP	Council size	Finland 1996-2008	2'000, 4'000, 8'000 15'000, 30'000
De Benedetto, &	2014, WP	Wage of mayors	Italy	1'000, 5'000, 50'000

continued on next page ...

. . . continued from previous page

Authors	Publication status	Research focus	Country Time	Thresholds
De Paola			1991-2001	
van der Linde et al.	2014, WP	Wage of politicians	Netherlands 2005-2012	8'000, 14'000, 24'000 40'000, 60'000, 100'000 150'000, 375'000
	Panel 2: Studies using a Difference-in-Discontinuity			
Casas-Arce, Saiz	2015, JPE	Gender Quota	Spain 2003-2007	5'000
Baques & Campa	2011, WP	Gender Quota	Spain 2003-2007	5'000
Grembi, Nannicini & Troiano	2012, WP	Fiscal rules	Italy 1999-2004	5'000
Asatryan, Baskaran, Grigoriadis, Heinemann	2013, WP	Citizen referenda and spending	Germany 1980-2011	10'000, 20'000, 30'000, 50'000, 100'000
Asatryan, Baskaran, Heinemann	2013, WP	Citizen referenda and taxes	Germany 1980-2011	10'000, 20'000, 30'000, 50'000, 100'000
Gulino	2014, WP	Electoral rules	Italy 1985-2000	5'000

Notes: All papers are also cited in our references.

Table 5: Population thresholds in Italian municipalities

	Policy changes at k inhabitants (in tsd)											
	1	3	5	10	15	25	30	50	60	100	250	500
Size of the city council		x		x			x			x	x	x
Wage of the mayor	x	x	x	x			x	x		x	x	x
Wage of the executive officers	x		x	x				x				
Attendance fee for city councilors				x			x					
Maximum number of executive officers				x						x	x	x
Electoral Rule (plurality/runoff)					x							
Neighborhood councils							x			x		
Hospitals						x						
Health district									x			
Balanced-budget rule			x									

NOTE: The table identifies population thresholds (in thousands) at which given policies change. This is a partial list of policies, chosen to highlight the variety of policies that depend on population thresholds and the extent to which the same threshold often determines multiple policies. *Source*: Italian Law on the Local Finance.

Table 6: Population Thresholds in Germany (by rule and state)

#	Institution/ Rule	Population thresholds in the different German states													Σ
		SH	NiedS	NRW	Hes	RP	BW	Bay	Saar	MVP	BB	Saan	Sax	Th	
	Councils														
1	Council size	(12)	(30)	(10)	(9)	[15]	(10)	(10)	[6]	(15)	(11)	(12)	(14)	[11]	165
2	Full-time council members / Deputies	(3)	[2]			[4]	[1]	[1]	[4]	[1]	[1]	[1]	(6)	(5)	29
3	City districts		[1]			[1]	[1]	[1]	1						5
4	City district council				1	[1]			(3)		[1]	[1]		[3]	10
5	Administrative units		(2)							(2)	[1]	[1]		[1]	7
6	Council of the administration units									[4]	(5)				9
	Wages														
7	Wage of mayors	[7]	[9]	[9]	(11)	[8]	(11)	(9)	(7)	(7)	(7)	(10)	(12)	[9]	116
8	Additional compensation of the mayor	(5)	[5]	[5]		[7]				[7]		[7]			36
9	Wage of head of admin. units	[3]								[1]		[5]	[3]	(8)	20
10	Wage of deputies								[5]				[8	[7]	20
11	Wage of mayors in recreational cities	[1]			[1]		[1]			[1]					4
	Fiscal rules														
12	Status of a larger city	[1]	[2]	(6)	[1]	[1]	[1]	[1]	[1]		[1]		1		16
13	Status of a county free city	[1]						[1]				1			3
14	Fiscal equalization		[6]	19*	[7]		[1]	(7)	(7)		(6)	(3)	(6)	[7]	69
15	Special transfers for mergers									[1]			[1]	[1]	3
	Citizen involvement														
16	Citizen request	[6]	[3]	[2]		(6)				[1]				1	19
17	Petition for referenda	[6]		[7]	[2]	(4)	[4]	[6]	[5]	[1]	(2)	[3]		[1]	41
18	Quota requirements for referenda	[6]		[2]				[2]						[2]	12
	Elections														
19	Signatures for party lists		[2]	[2]		[14]	[5]		[1]		[5]		(7)		36
20	Signatures for mayoral candidacy		[1]	[1]			[4]				[1]				7
21	Election districts	[4]	1		1	[3]				1	(6)	1	1	[1]	19
22	Ballot districts			1		[2]	1			[2]	[1]	1			8
23	Reevaluation of an election				[1]		[2]						[1]		3
	Committees/commisioners														
24	Equal opportunity commissioner	[1]		[1]					[1]	[1]	1	[1]	[1]	[1]	8
25	Integration council			(3)	1	[1]									5
26	Accounting agency	[1]			[1]				[1]			[1]	[1]		5
27	Oversight regulation	[1]			[1]		1								3
28	Council for top-secret issues						[3]						[2]		5
29	Open council						(4)	[2]							6
	Mayor														
30	Mayor status	(2)			1		[2]	[2]					[2]	[2]	11
31	Mayor title	[1]			[1]				[1]			[1]			4
32	Deselection of mayors			[2]									[1]		3
33	Qualification of mayor								[1]	[1]		[1]			3

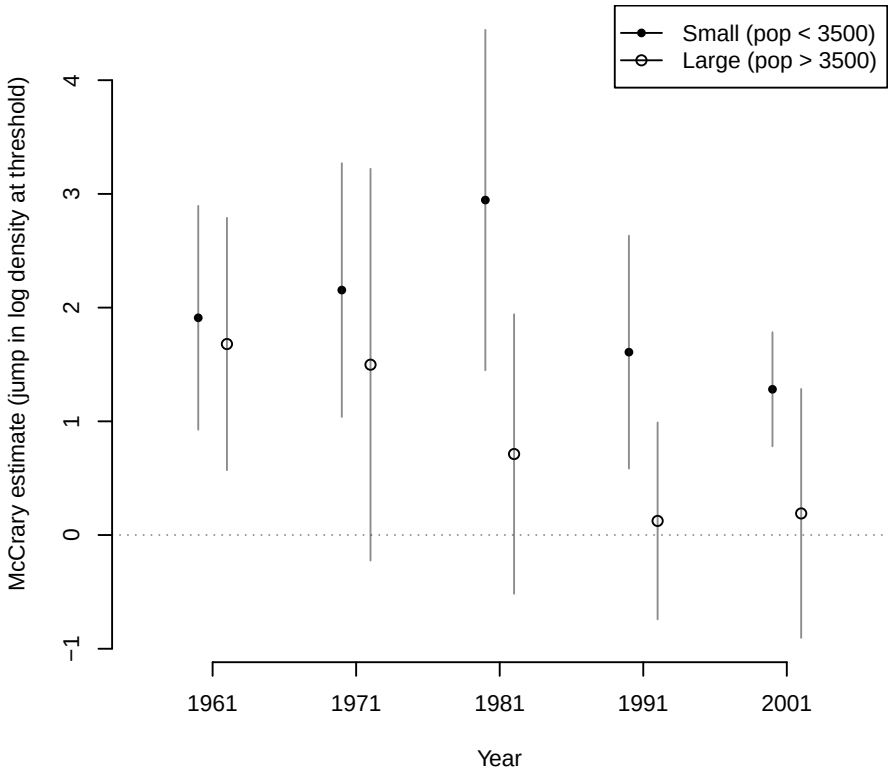
continued on next page . . .

. . . continued from previous page

#	Institution/ Rule	Population thresholds in the different German states												
		SH	NiedS	NRW	Hes	RP	BW	Bay	Saar	MVP	BB	Saan	Sax	Th
	Unique rules													
34	Direct democracy	[1]												
35	Consolidated accounts	[1]												
36	Youth welfare office		[1]											
37	Construction oversight		[2]											
38	Mayor deputy recall				[1]									
39	County financing				[1]									
40	Additional transfers				[1]									
41	Key transfers				[3]									
42	Deputies in administrative units					[6]								
43	Pension funds					[1]								
44	Municipal treasurer						[1]							
45	Office hours on election day						[1]							
46	Election statistics						[1]							
47	Economic status						(8)							
48	Accounting committee							[1]						
49	Water management							4						
50	Vehicle Tax							[1]						
51	Deputy mayor status								[1]					
52	Naming of city districts								[1]					
53	Cooperation councils								[1]					
54	Council after mergers									[1]				
55	Outside administrative units									[2]				
56	Forced administrative reconsideration									[1]				
57	Administrator qualification requirement									[1]				
58	Format of the ballots										[2]			
59	Add compensation of head of admin											[4]		
60	City district elections												[2]	
61	Infrastructure subsidies												[4]	
62	Country-side municipalities													[1]
63	General committee													[1]
64	Annual audits													[1]
65	Election proposals													[1]
	Σ	60	67	51	44	73	63	47	46	52	51	54	72	64
	# of different policy issues	19	14	14	18	15	20	15	17	19	15	17	15	19
	# of different population thresholds	21	34	22	20	22	20	17	14	22	23	22	27	17
	# of thresholds with unique change	9	18	10	10	4	7	8	6	6	9	12	12	3

Notes: *Source*: Own research.

Figure 11: Sorting over time in Italy



NOTE: Each point corresponds to the McCrary test statistic (the estimated jump in the log density of the running variable at the relevant threshold) for a given set of thresholds in a given census in Italy. Lines show 95% confidence intervals.