Replication analysis of: Rural Roads and Local Economic Development (Asher and Novosad, 2020)

Agustin Fernández Felguera
23 December 2022

1 Summary

The authors study the impact of India's $40 billion national rural road construction program (also known as PMGSY). A program launched in 2000 with the goal of providing all-weather road access to unconnected villages across India. Authors claim they contribute to the literature on two fronts: i) by providing estimates of the impact of investments in transportation infrastructure and ii) by adding an understanding of the barriers to reallocation of labor out of agriculture in developing countries.

The key feature of the paper is its ability to address the endogeneity of road placement by taking advantage of an implementation rule of the program that targeted roads to villages with population exceeding two discrete thresholds (500 and 1,000). This rule causes villages just above the population threshold to be 22 percentage points more likely to receive a road, allowing the researchers to estimate the causal impact of rural roads using a fuzzy regression discontinuity design. The researchers construct a high spatial resolution dataset that combines administrative microdata covering all households and firms in the sample of villages with remote sensing data and village aggregates describing amenities, infrastructure, and demographic information. The analysis sample is restricted to the 11,432 villages that (i) did not have a paved road in 2001; (ii) were matched across all primary datasets; and (iii) had populations within the optimal bandwidth from a treatment threshold.

The primary specification uses local linear regression within a given bandwidth of the treatment threshold, and controls for the running variable (village population) on either side of the threshold. Under the assumption of continuity of all other village characteristics other than road treatment at the treatment threshold, the fuzzy RD estimator calculates the local average treatment effect (LATE) of receiving a new road for a village with population equal to the threshold. They present treatment estimates on five indices of the major categories of outcomes: (i) transportation services; (ii) sectoral allocation of labor; (iii) employment in nonfarm village firms; (iv) agricultural investment and yields; and (v) income, assets, and predicted consumption. On average, they observe outcomes four years after road completion, meaning that they estimate the short to medium-run impact of these roads.

According to the research, the construction of rural roads does not significantly impact agricultural outcomes, income, or assets in the villages where they are built within the first four years after construction. They do find that rural roads lead to a large reallocation of workers out of agriculture. A new road causes a 9 percentage point decrease in the share of workers in agriculture and an equivalent increase in wage labor. These impacts are most pronounced among the groups likely to have the lowest costs and highest potential gains from participation in labor markets: households with small landholdings and working-age men. Their conclusion is that even with better market connections, remote areas may continue to lack economic opportunities and roads alone appear to be insufficient to transform the economic structure of remote villages.

2 Replication

In this section I will replicate the main tables and plots from the paper and I will comment on them. The replication will not focus much on the economic theory or data preprocessing, but on evaluating carefully the empirical design. I will analyze the underlying assumptions in detail and I will test them, providing the necessary robustness checks. As an extension of the paper, I will develop some methodological points that I think are not discussed enough in the paper.

2.1 Data sources

Most of the datasets used in the paper are publicly available in the replication package from (Sam Asher and Paul Novosad, 2020). Note, however, that some variables are constructed from private data or restricted-use data, and so it is not possible to replicate all the figures.

In 2015, the authors scraped the official PMGSY program website to obtain the identities of the 345,797 connected-to-roads villages and their completion dates. As already mentioned, three criteria are used to select from this data those villages appropriate for the analysis. The final sample for the analysis includes 11,432 villages, with 9,656 in the 500 threshold group (between 300 and 700 inhabitants) and 1,776 in the 1,000 threshold group (700 to 1,300 inhabitants).

To get additional information at the village level, the authors combine this dataset with multiple other data sources that go from 2001 to 2014, including household and individual microdata from census and surveys, as well as remote sensing data to measure outcomes that were otherwise unavailable at the village level, such as night lights or crops. They also use the set of asset and income variables in the individual microdata to predict households' consumption.

2.2 Empirical strategy

The primary specification uses local linear regression within a given bandwidth of the treatment threshold as recommended by (Gelman and G. Imbens, 2019), and controls for the running variable (village population) on either side of the threshold. The thresholds may not perfectly determine road construction for a given village, but they create a discontinuity in the probability of receiving a road. We can use such discontinuities to produce instrumental variable estimators of the effect of the treatment (close to the discontinuity). The researchers use a two-stage instrumental variables specification in which equation 1 is the relationship of interest and equation 2 is the first-stage:

$$\begin{aligned} Y_{v,j} = \beta_0 + \beta_1 \text{Road}_{v,j} + \beta_2 \left(\text{pop}_{v,j} - T\right) \\ + \beta_3 \left(\text{pop}_{v,j} - T\right) \times \mathbf{1}\left\{\text{pop}_{v,j} \geq T\right\} + \zeta \boldsymbol{X}_{v,j} + \eta_j + \epsilon_{v,j} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{Road}_{v,j} = \gamma_0 + \gamma_1 \mathbf{1}\left\{\text{pop}_{v,j} \geq T\right\} + \gamma_2 \left(\text{pop}_{v,j} - T\right) \\ + \gamma_3 \left(\text{pop}_{v,j} - T\right) \times \mathbf{1}\left\{\text{pop}_{v,j} \geq T\right\} + \nu \boldsymbol{X}_{v,j} + \mu_j + \upsilon_{v,j} \end{aligned} \tag{2}$$

Here, $Y_{v,j}$ is the outcome of interest in village v and district-threshold group j, T is the population threshold, pop_{vj} is baseline village population, $\boldsymbol{X}_{v,j}$ is a vector of village controls measured at baseline, and η_j and μ_j are district-threshold fixed effects. There are eleven controls that include socio-economic variables at the village level, agricultural output, or dummies for the presence of village amenities. District-threshold fixed effects are district fixed effects interacted with an indicator variable for whether the village is in the 1,000-person threshold group.

The variable Road is an indicator that takes the value 1 if the village received a new road before the year in which Y is measured, which is 2011, 2012, or 2013[1]. The coefficient of interest is β_1 as it captures the effect of a new road on an outcome variable.

To get an optimal bandwidth for this specification the researchers use the method proposed by (G. Imbens and Kalyanaraman, 2012) and obtain the value 84. The choice of this parameter affects significantly the number of observations used in the estimation: No observations are dropped for the bandwidth of 84 while for a value of 60, 3,140 villages are dropped (27% of the sample). However, as shown in table 1, the estimates of the first-stage equation are robust to changes in this parameter[2].

On a side note, it is important to note that the running variable population is discrete, and conventional RD methods deal with continuous variables, as they ensure that some observations will have scores arbitrarily close to the cutoff in large samples. However, it seems that mass points in the data do not change the results significatively and the population variable has 167 different values in the final sample.

Lastly, there are several other estimating methods for RD designs different from the one just presented. I use the package *rdrobust* (Calonico et al., 2017) to correct for the bias from the standard estimation used in the paper and the package *RDHonest* which finds the best estimator in terms of the minimax MSE. *Rdrobust* allows me to use controls and it does not need to estimate any additional parameter, since for the bandwidth choice I use the same value as in the paper. For *RDHonest* things are harder, first, because it does not allow the use of controls. Second, the smoothness constants $M1$ and $M2$ should not be data-driven: a researcher should choose them a priori. Since this is hard to do, I use the automatic estimation of the parameters proposed in (Armstrong and Kolesár, 2020).

2.3 Identification of Fuzzy Regression Discontinuity (FRD)

In the Sharp RD case, in general we can identify treatment effects with only the continuity assumption, that is, $E[Y_1|Pop, W]$ and $E[Y_0|Pop, W]$ are continuous in the population variable around the threshold. In the FRD design, the interpretation of the causal effect is more complicated because we need to add IV assumptions. As discussed in (Cattaneo and Titiunik, 2022), we require the monotonicity restriction (no-defiers) which rules out villages when their compliance decisions (road construction) are the opposite of their assignments (program priority).

We can state that under the assumption of continuity of all other village characteristics other than road treatment at the treatment threshold, and under the monotonicity assumption, the fuzzy RD estimator can be interpreted as the local average treatment effect (LATE) of receiving a new road for

[1] The authors mention these are not particularly unusual years in the Indian economy, in terms of economic growth and the weather.

[2] I also checked the robustness of the rest of the results and the estimates do not change significantly with changes in the bandwidth parameter.

complier villages with population equal to the threshold.

It is not a standard practice to include controls in the RD design, as the identification argument does not require covariates since is based on continuity. The authors only mention very briefly that "village controls and fixed effects are not necessary for identification but improve the efficiency of the estimation". As shown in (Frölich and Huber, 2019), even when the assumption of balancing is met (covariates are identically distributed on both sides of the threshold), such that the estimate is identified with and without controlling for any X, controlling for covariates can still be useful to reduce the variance of the estimator, as long as the covariates have predictive power either for the outcome or for the treatment.

When using a FRD design it is interesting to show the RD plot of both the outcome discontinuity and treatment discontinuity. For the latter, we want to see how strong is the instrument and we can use a Sharp RD design to do so. Figure 1 shows the share of villages that received new roads before 2012 in each population band relative to the threshold; there is a substantial discontinuous increase in the probability of receiving a road in the threshold. Table 1 presents first-stage estimates using the main estimating equation at various bandwidths. As suggested by the figure, crossing the treatment threshold raises the probability of treatment by approximately 22 percentage points.

	60	70	80	90	100	110
Road Priority	0.224	0.221	0.217	0.214	0.213	0.215
SE	0.019	0.018	0.017	0.016	0.015	0.014
Observations	8339	9720	11099	12457	13871	15238

Table 1: First Stage: Effect of Road Prioritization on Road Treatment

Note: This table presents first-stage estimates of the effect of being above the treatment threshold on a village's probability of treatment. The dependent variable is an indicator variable that takes on the value 1 if a village has received a PMGSY road before 2012. The first column presents results for villages with populations within 60 of the population threshold (440–560 for the low threshold and 940–1,060 for the high threshold). The second through sixth columns expand the sample to include villages within 70, 80, 90, 100, and 110 of the population thresholds. The specification includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates.

Table 2 presents the mean values for various village baseline characteristics and for the whole sample and for the two population bands, including the set of controls that are used in all regressions. In column 4 we see that there are several variables that do not pass the t-test, in part because many village characteristics are correlated with size and we are considering the whole sample (not using a bandwidth). However, more importantly, column 7 shows, using the main specification with controls[3], that none of the control variables have a significant coefficient for the discontinuity[4]. This failure to reject the null hypothesis of covariate balance can be seen as evidence of the comparability of control and treatment groups near the cutoff.

Authors first observe that there is a set of states that did not follow guidelines regarding the population eligibility threshold and more importantly, some village officials manipulated the official population data used in the PMGSY program to be above the threshold required for having priority in the road construction. Fortunately, the population data used to assign priority in the program is from the 2001

[3]In principle the optimal bandwidth for testing discontinuities in covariates may not be the same as the optimal bandwidth for the treatment. So I follow the practice of testing robustness to variations in bandwidth.

[4]The closest one is Medical Center because there was a policy (Total Sanitation Campaign) about hospitals based on the size of the village.

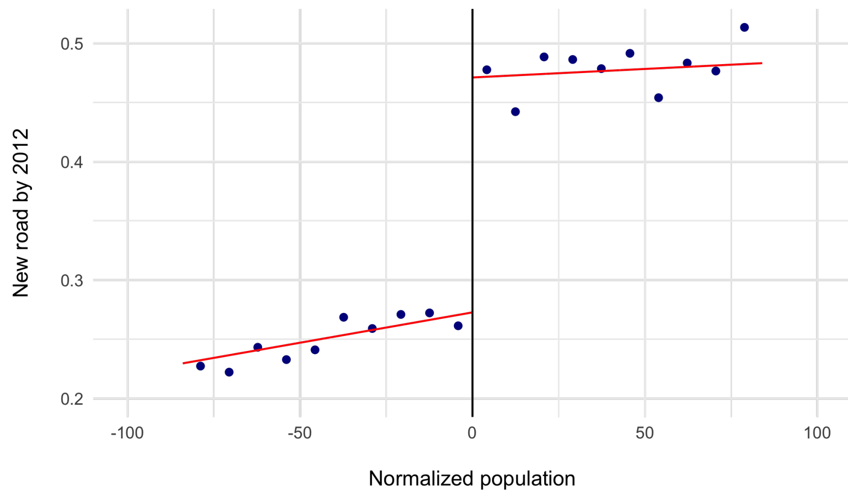

Figure 1: First Stage: Effect of Road Prioritization on Probability of New Road by 2012
Note: The figure plots the probability of getting a new road under PMGSY by 2012 against village population in the 2001 Population Census. The sample consists of villages that did not have a paved road at baseline, with baseline population within an optimal bandwidth (84) of the population thresholds. Populations are normalized by subtracting the threshold population.

census, which was collected before the start of the PMGSY program. Additionally, the researchers get rid of those villages in which there may be manipulation of the population variable. The distribution of this variable is shown in figure 2.

Nevertheless, we still want to test that in the sample used, units cannot precisely manipulate the assignment variable to influence whether they receive the treatment or not. The authors use the test proposed by (McCrary, 2008) for a discontinuity in 0 and they show a plot with a nonparametric regression fitted to each half of the density distribution. The point estimate for the discontinuity is 0.014 (Log difference in heights), with a standard error of 0.05.

However, I noticed that the test is quite sensitive to the choice of the bin parameter and so I use a tuning parameter-free method proposed by (Cattaneo, Jansson, and Ma, 2020), which does not require boundary-specific data transformations (such as prebinning). The heuristic idea of the test is that whereas other nonparametric density estimators are constructed by smoothing out a histogram-type estimator of the density, this estimator is constructed by smoothing out the empirical distribution function using local polynomial techniques. The calculations are done using the R package *rddensity*, which is maintained by the same authors of the paper; and the graphical representation of the test is shown in figure 3. The confidence intervals displayed are corrected for the RD bias and the plot has an additional histogram in the background. The statistic of the test is close to zero and the p-value is high (close to 0.9), so we can conclude that the density of population is continuous at the threshold.

2.4 Main results

The most relevant table in the paper presents treatment estimates on five indices of the major families of outcomes: (i) transportation services; (ii) sectoral allocation of labor; (iii) employment in nonfarm village firms, (iv) agricultural investment and yields; (v) income, assets, and predicted consumption.

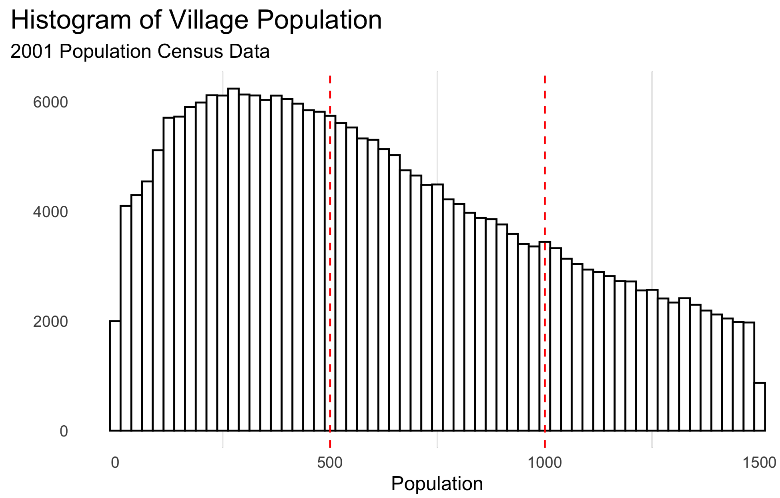

Figure 2: Distribution of running variable

Note: The figure shows the histogram of village population as recorded in the 2001 Population Census. The vertical lines show the program eligibility thresholds used in this paper, at 500 and 1,000.

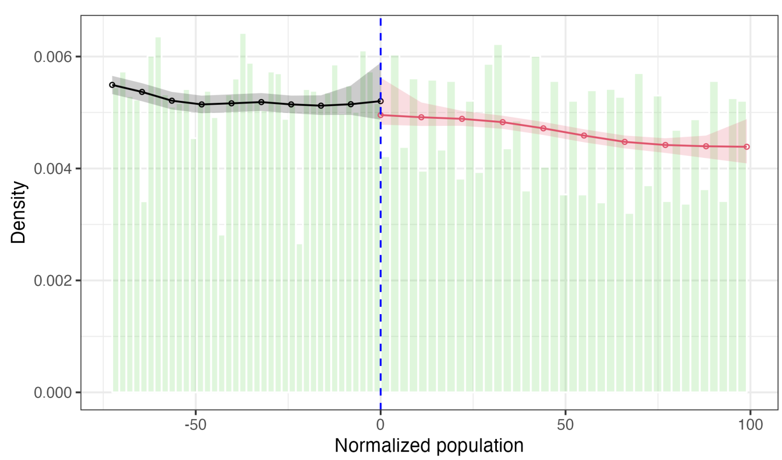

Figure 3: Cattaneo, Jansson, Ma (2017) manipulation test of the running variable

Note: The figure uses the normalized village population (reported population minus the threshold, either 500 or 1,000). It plots a nonparametric regression to each half of the distribution following Cattaneo, Jansson, Ma (2017), and a histogram in the background. testing for a discontinuity at zero.

	Full	Below	Over	Difference	p-value on diff.	RD estimate	p-value on RD
Primary school	0.956	0.951	0.961	0.010	0.007	-0.012	0.720
Medical center	0.163	0.153	0.175	0.022	0.001	-0.092	0.151
Electrified	0.425	0.408	0.443	0.035	0.000	-0.001	0.992
Distance from nearest town (km)	26.782	26.811	26.749	-0.062	0.882	-4.818	0.172
Land irrigated (share)	0.281	0.275	0.288	0.014	0.014	-0.020	0.647
ln land area	5.160	5.107	5.220	0.113	0.000	-0.093	0.360
Literate (share)	0.456	0.452	0.460	0.008	0.006	-0.007	0.762
Scheduled caste (share)	0.142	0.141	0.144	0.004	0.276	-0.021	0.492
Land ownership (share)	0.736	0.737	0.734	-0.003	0.475	0.004	0.904
Subsistence ag (share)	0.440	0.443	0.436	-0.007	0.181	0.024	0.545
HH income > INR 250 (share)	0.757	0.755	0.759	0.004	0.431	-0.029	0.532

Table 2: Summary Statistics and Balance

Note: The table presents mean values for village characteristics, measured in the baseline period. The first eight variables come from the 2001 Population Census, while the final three come from the 2002 BPL Census. Columns 1–3 show the unconditional means for all villages, villages below the treatment threshold, and villages above the treatment threshold, respectively. Column 4 shows the difference of means across columns 2 and 3, and column 5 shows the p-value for the difference of means (t-test). Column 6 shows the regression discontinuity estimate, following the main estimating equation, of the effect of being above the treatment threshold on the baseline variable (with the outcome variable omitted from the set of controls), and column 7 is the p-value for this estimate, using heteroskedasticity robust standard errors. An optimal bandwidth of 84 around the population thresholds has been used to define the sample of villages.

These indices aggregate information from numerous variables and are generated to have a mean of 0 and a standard deviation of 1. They are constructed to address concerns about multiple hypothesis testing within families of outcomes.

Table 3 shows these estimates for four different methods: (Controls-Conventional) Uses the main estimating equation that includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects. These are the estimates shown in the paper. (Controls-Bias corrected) Uses the package *rdrobust* to correct for the bias in the estimates. (No controls - Conventional) Same equation as row 1 but removing all the control variables. (No controls-Minimax) Uses the package *RDHonest* (minimax approach) with no controls. Heteroskedasticity robust standard errors are reported below point estimates in all cases.

Overall, the first column shows a large positive effect on the availability of transportation services, and the second shows that roads cause a significant reallocation of labor out of agriculture. However, for the last two columns, we observe very small and insignificant positive effects. These results broadly summarize the findings of this paper: rural roads lead to increases in transportation services and reallocation of labor out of agriculture, but not to major changes to village firms, agricultural production, or predicted consumption.

Comparing methods, the major differences can be seen when we correct the main specification (row 1) from bias (row 3). The value of the coefficients changes completely and can lead to different conclusions, especially regarding the impact on firms. Additionally, while we do not observe important differences between rows 3 and 4, which use no controls but different algorithms, we do see that there is a substantial difference between the estimates with controls and without in rows 1 and 3, especially for the last two columns. While the reduction in the standard errors is expected, the difference in the coefficients is less clear. This fact is even more puzzling if we test the robustness of the results to the choice of controls. If we remove the control dummy presence of a medical center for the main specification, the estimates change significatively: For the transportation index, the estimate is 0.633 instead of 0.410, while for Ag Occupation is -0.277. In the specification used in the paper, many of

Estimates	Transportation (i)	Ag occupation (ii)	Firms (iii)	Ag production (iv)	Consumption (v)
Controls	0.410**	-0.341**	0.269*	0.082	0.033
Conventional	(0.189)	(0.162)	(0.158)	(0.125)	(0.138)
Controls	0.333*	-0.206	0.523***	0.181	0.118
Bias corrected	(0.184)	(0.159)	(0.155)	(0.123)	(0.135)
No controls	0.469**	-0.425**	0.282	0.242	0.105
Conventional	(0.219)	(0.207)	(0.200)	(0.191)	(0.204)
No controls	0.474**	-0.426**	0.283	0.235	0.105
Minimax	(0.221)	(0.209)	(0.201)	(0.195)	(0.205)
Observations	11,432	11,432	10,678	11,432	11,432
R^2	0.178	0.282	0.296	0.534	0.496

Table 3: Impact of New Road on Indices of Major Outcomes

Note: This table presents regression discontinuity estimates of the effect of a new road on indices of the major outcomes in each of the five families of outcomes: transportation, occupation, firms, agriculture, and consumption. It shows four different procedures: (Controls-Conventional) Uses the main estimating equation that includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects. (Controls-Bias corrected) Uses the package rdrobust to correct for the bias in the estimates. (No controls - Conventional) Same equation as row 1 but removing all the control variables. (No controls-Minimax) Uses the R package RDHonest (minimax approach) with no controls. Heteroskedasticity robust standard errors are reported below point estimates in all cases. The number of observations and the R^2 correspond to the main specification.

the controls have highly significative coefficients.

In the paper, the authors continue the analysis by examining the components of each of these indices to explain the impacts of roads in more detail, and present results on heterogeneity. This analysis is similar to the one already done for this last table 3 (the methodology is the same) so this part is omitted here. Moreover, some variables use private data so it is not possible to replicate all the figures.

2.5 How the impact of a new road depends on the size of the village?

The authors use the common practice of normalizing all of the cutoffs to zero and estimate only one effect[5] but in reality, there are two thresholds (500 and 1,000 population). The final sample has villages from 6 states, two of which use the two thresholds to assign priority to road construction, and the four others use the 500 pop. threshold[6].

As the authors discuss in the Conceptual Framework, there are a large number of factors that influence how the arrival of a road in a village can improve its economic outcomes. It would not be unreasonable to think that the population of the village can be connected with many of these factors, for example,

[5] Researchers often prefer one takeaway summary effect that can be more precisely estimated by using all the data.

[6] The pogram used an additional threshold for 250 inhabitants but the sample was too small. As suggested in 2 the number of villages with less than 250 pop. is small, especially for the latest periods analyzed.

Estimates	Transportation (i)	Ag occupation (ii)	Firms (iii)	Ag production (iv)	Consumption (v)
500 pop. threshold	0.740*** (0.117)	-0.269*** (0.098)	0.524*** (0.099)	0.027 (0.077)	0.040 (0.083)
1000 pop. threshold	0.227 (0.184)	-0.278** (0.132)	0.174 (0.146)	0.069 (0.109)	-0.155 (0.121)

Table 4: Impact of New Road on Indices of Major Outcomes by threshold

Note: This table presents regression discontinuity estimates of the effect of a new road on indices of the major outcomes in each of the five families of outcomes: transportation, occupation, firms, agriculture, and consumption. The data is divided into two subgroups: Villages with around 500 population and villages with around 1,000. The estimates are obtained by using the main estimating equation that includes baseline village-level controls for amenities and economic indicators, as well as district-cutoff fixed effects. Heteroskedasticity robust standard errors are reported below point estimates in all cases.

you can think of economies of agglomeration. In this case, the treatment effect should be different for the two thresholds and that would affect substantially the evaluation of the policy.

There is a lack of theoretical investigation on how to combine observations from all cutoffs to estimate economically relevant average effects. That is to say, how to properly analyze Multi-Cutoff RD Designs. There has been studied recently in the JMP (Bertanha, n.d.). Moreover, closely related to this topic, there is some literature about the study of treatment effect extrapolation (how to get further away from the discontinuity), e.g. (Bertanha and G. W. Imbens, 2014), (Angrist and Rokkanen, 2015), among others.

The first step is always to estimate the treatment effects in each threshold. Table 4 shows the difference in the results between the two thresholds. The coefficients for columns (i) and (iii) are very different and lead to different conclusions. The standard error for the estimates in the 500 pop. subset are lower but this may be due to the larger number of villages in this threshold. But after that, as stated by (Bertanha, n.d.), the ability to combine different local effects to estimate an average effect depends on how comparable the researcher believes these effects are, it depends crucially on heterogeneity assumptions connecting local treatment effects.

To continue the analysis I should combine these two estimates with an appropriate weighting scheme in a way that is adequate with the assumptions I make. Unfortunately, it was not possible to do this due to time constraints.

3 Final remarks

In this replication study, I analyze in detail the empirical strategy and I examine several potential improvements to the methodology of the original study. Among these, I give particular attention to the use of multiple cutoff points, as this approach has significant relevance for interpreting the results and was not addressed in the original paper.

However, I also look at how the results of the study can change depending on the estimation method used. When instead of using the most standard method we use one that corrects for the bias in the estimation, the results differ drastically. The authors do not analyze how robust are their results to different estimation approaches.

A related problem is the use of covariates. The authors only mention in the paper that village controls and fixed effects are not necessary for identification but improve the efficiency of the estimation. However, the estimates change notably when we remove the controls or change some of them. As adding controls in a FRD design is not the standard practice and it is not required for identification, it could be reasonable to complement the results shown with a specification with no controls.

In empirical research, there are often a large number of decisions that a researcher must make that can influence the results of the study. These decisions can include choices about the research design, sample selection, data collection, and so on. While some of these decisions may not significantly change the results of the study, it is important for researchers to carefully consider all of these decisions.

References

Angrist, Joshua D. and Miikka Rokkanen (2015). "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff". In: *Journal of the American Statistical Association* 110.512, pp. 1331–1344.

Armstrong, Timothy B. and Michal Kolesár (2020). "Simple and honest confidence intervals in nonparametric regression". en. In: *Quantitative Economics* 11.1, pp. 1–39.

Asher, Sam and Paul Novosad (Mar. 2020). "Rural Roads and Local Economic Development". en. In: *American Economic Review* 110.3, pp. 797–823.

Bertanha, Marinho (n.d.). "Regression Discontinuity Design with Many Thresholds". en. In: ().

Bertanha, Marinho and Guido W Imbens (Dec. 2014). *External Validity in Fuzzy Regression Discontinuity Designs*. Working Paper 20773. National Bureau of Economic Research.

Calonico, Sebastian et al. (2017). "Rdrobust: Software for Regression-discontinuity Designs". In: *The Stata Journal* 17.2, pp. 372–404.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma (July 2020). "Simple Local Polynomial Density Estimators". en. In: *Journal of the American Statistical Association* 115.531, pp. 1449–1455.

Cattaneo, Matias D. and Rocıo Titiunik (2022). "Regression Discontinuity Designs". In: *Annual Review of Economics* 14.1, pp. 821–851.

Frölich, Markus and Martin Huber (Oct. 2019). "Including Covariates in the Regression Discontinuity Design". en. In: *Journal of Business & Economic Statistics* 37.4, pp. 736–748.

Gelman, Andrew and Guido Imbens (July 2019). "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs". In: *Journal of Business & Economic Statistics* 37.3. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.2017.1366909, pp. 447–456.

Imbens, Guido and Karthik Kalyanaraman (July 2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". In: *The Review of Economic Studies* 79.3, pp. 933–959.

McCrary, Justin (Feb. 2008). "Manipulation of the running variable in the regression discontinuity design: A density test". en. In: *Journal of Econometrics*. The regression discontinuity design: Theory and applications 142.2, pp. 698–714.

Sam Asher, World Bank and Johns Hopkins SAIS and Dartmouth College Paul Novosad (June 2020). "Replication data and code for: Rural Roads and Local Economic Development". en. In: Publisher: Inter-university Consortium for Political and Social Research (ICPSR) Type: dataset.