Improving Treatment Comparability in an Instrumental Variables Design: Replication and Extension of Hangartner et. al 2018*

Noah Daponte-Smith † Sarah Mohamed ‡ Reed Rasband §

May 8, 2019

Abstract

In "Does Exposure to the Refugee Crisis Make Natives More Hostile?" Hangartner et al. (2018) examine whether exposure to the refugee crisis increases levels of hostility to refugees among natives. They investigate this question using an original survey on the Greek islands in the Aegean Sea conducted during the Syrian refugee crisis and employ instrumental variable techniques to bolster claims of exogeneity. In this paper, we replicate and extend their results. We call into question the validity of their instrument and demonstrate how using matching techniques and distance filtering alters their findings.

*Hangartner, D., Dinas, E., Marbach, M., Matakos, K., Xefteris, D. (n.d.). Does Exposure to the Refugee Crisis Make Natives More Hostile? American Political Science Review, 1-14. doi:10.1017/S0003055418000813

†Ph.D. student, Department of Government, Harvard University, Cambridge, MA, 02138. Email: daponte-smith@g.harvard.edu

‡Ph.D. student, Department of Government, Harvard University, Cambridge, MA, 02138. Email: sarahmohamed@g.harvard.edu

§Ph.D. student, Department of Government, Harvard University, Cambridge, MA, 02138. Email: rasband@g.harvard.edu

List of Tables

1 Placebo Outcomes (Table 1 in original paper) 2

2 Balance Table for Original Data 7

3 Balance Table for Geographically Subsetted Data 11

4 Balance Table for Matched Data at Four Pruning Cut-offs 16

List of Figures

1 Visualization of First-Stage Estimates (Figure 2 in original paper) 2

2 2SLS Estimates (Figure 4 in original paper) 3

3 Islands in Hangartner et al (2018) sample 5

4 First-Stage and 2SLS for Filter at Third Quartile Log Distance 9

5 First-Stage and 2SLS for Filter at Median Log Distance 9

6 First-Stage and 2SLS for Filter at First Quartile Log Distance 10

7 Significance of Results across Matched Datasets 14

1 Introduction

Following the advice and guidelines set forth in "Publication, Publication," this paper first replicates and then seeks to extend the results and analysis in Hangartner et al.'s 2018 paper, "Does Exposure to the Refugee Crisis Make Natives More Hostile?" (King, 2006). The authors examine the long-term attitudinal effects of exposure to the refugee crisis on residents of the Greek Islands. We began by replicating the tables and figures of interest in Hangartner et al.'s paper (Figures 1 and 2 and Table 1), which included the authors' first and second-stage least squares estimations and placebo tests.[1]

[1]Note that our PCA components (which can be seen in the black lines in Figures 2 and 4b-7b) were produced using Stata, using the original authors' code. We then imported the PCA components into R to produce the regression analysis and corresponding figures.

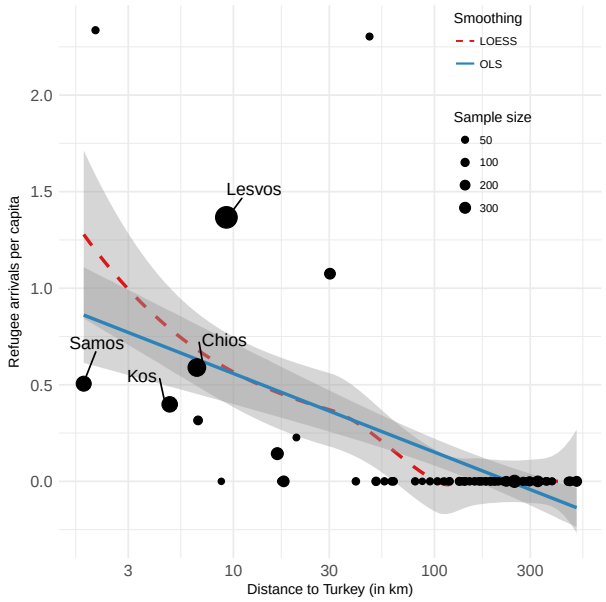


Figure 1: Visualization of First-Stage Estimates (Figure 2 in original paper)

	GD	S/A	ND	PASOK	Female	Age	Education
treatment	0.002	-0.053	0.050	0.001	-0.020	0.007	-0.016
	(0.008)	(0.023)	(0.021)	(0.012)	(0.028)	(0.028)	(0.022)
constant	0.025	0.371	0.191	0.087	0.561	0.477	0.450
	(0.005)	(0.032)	(0.020)	(0.011)	(0.016)	(0.019)	(0.018)
N	2046	2046	2046	2046	2046	2046	2046

Table 1: Placebo Outcomes (Table 1 in original paper)

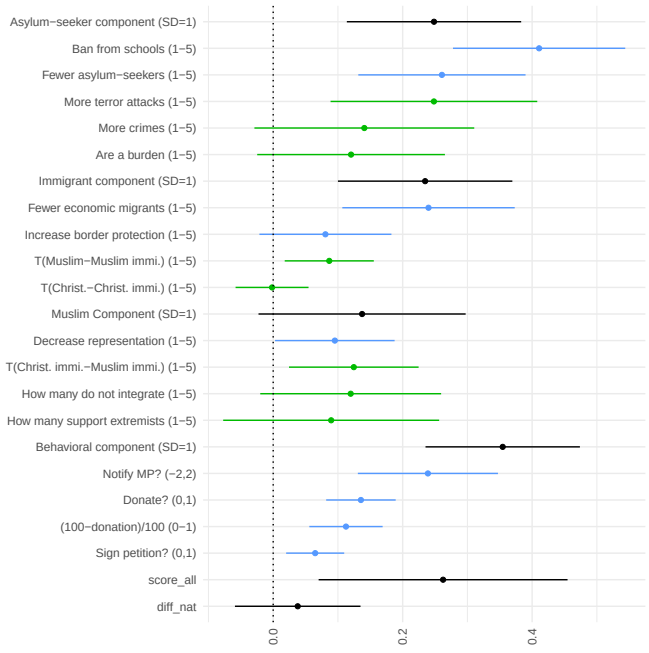


Figure 2: 2SLS Estimates (Figure 4 in original paper)

The authors' original model investigates the effect of refugee arrivals on anti-refugee sentiment. The authors conducted an original survey (N = 2,046) of the Greek islands just after the apogee of the refugee crisis and accordingly use a variety of dependent variables, some generated using principal component analysis. To account for missing data, the authors used multiple-imputation methods. The dependent variables were coded such that higher values indicate higher levels of anti-refugee sentiment, while the independent variable is a binary metric assessing whether an island received refugees. (In robustness checks, the authors present similar results using continuous and logged-continuous independent variables.) The key instrumental specification uses distance from Turkey as an exogenous source of variation in refugee arrivals.

In the following extensions of the original authors' analysis, we first examine the instrumental variable analysis, holding constant the choice of the binary version of the independent variable, and question the validity of the instrument – distance to Turkey (Section 2). Specifically, we are concerned that the instrument of choice may correlate with other unobserved inputs that themselves predict

outgroup hostility. We also posit that the validity of the instrument is highly sensitive to choices about which areas to include in the survey, such that omitting islands which could not have plausibly received treatment attenuates the authors' results. In an attempt to test the robustness of the orignal findings in the face of the these concerns, we subset the sample of islands at the 75th, 50th, and 25th percentiles of distance to Turkey and show first and second-stage least squares estimation of the effect of arrivals on anti-refugee sentiment (Section 3). Recognizing that such cut-offs might be considered arbitrary, we use Mahalanobis-distance matching to see how pruning data at various cutoff points changes our results (Section 4). We also chose to pursue matching given that we found the original dataset to contain considerable imbalance in terms of distance from Turkey, which we may correlate with unobserved determinants of outgroup hostility (see Table 2). With pruning, we find that the estimated effect on hostility remains positive for nearly all outcomes regardless of the number of observations pruned. However, the statistical significance of the results begins to fall as the number of observations pruned increase. Matching does improve balance for logged distance. The paper concludes with a summary of our results and a discussion of the limitations of the methodological approaches used (Section 5).

2 Theoretical Discussion: Instrumental Validity

The analysis in Hangartner et al. is predicated on instrumental variable analysis, intended in this case to correct for possible endogeneity between anti-refugee sentiments and the choice, on the part of the refugee, of the island to which to migrate. The concern, in other words, is that refugees might choose to go to islands already more favorable towards refugees, and that this endogeneity might bias our estimates of the effect of arrivals on anti-refugee sentiments. The instrument the authors employ to resolve the endogeneity problem is the distance of each island from the Turkish coast. On a purely mathematical basis, this instrument appears valid, with an F-statistic, as reported in the original paper, of 133.75, well above the threshold of 10 recommended by Sovey and Green (2011).

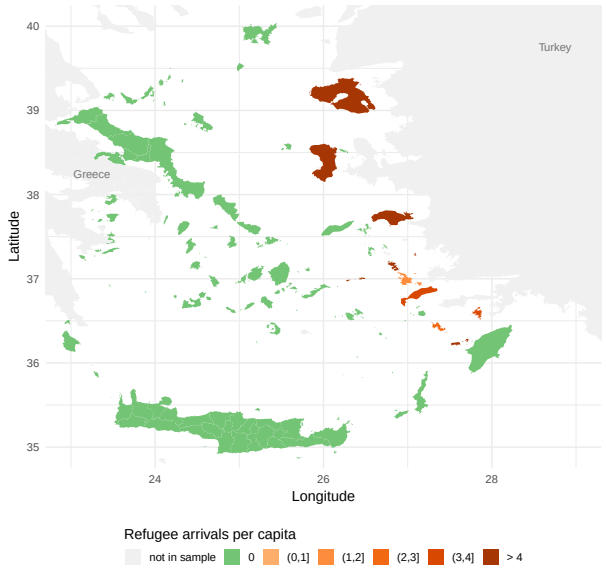


Figure 3: Islands in Hangartner et al (2018) sample

Our primary concerns with the instrument involve geography and the extent of the set of islands that had the possibility of receiving treatment. For elucidation of this argument, consider figure 3, which maps all islands in the sample with a color scale corresponding to the number of refugees per capita. What is apparent from the plot is that, of the 93 islands included in the sample, only 11 received refugees. The 11 islands receiving treatment were all located extremely close to Turkey, often within less than five kilometers. This makes the original structure of the authors' dataset - which includes all Greek islands, regardless of location relative to Turkey or the Greek mainland - somewhat perplexing.

The authors use the following logic to justify both their choice of instrument and their empirical strategy:

> ...distance to the Turkish coast causes dramatic variation in the number of refugee arrivals. This allows us to compare similar residents on neighboring islands, some of which received hundreds of thousands of passing refugees, while others slightly away received none. Second, geographical proximity and the fact that they often belong to the same administrative unit ensures that islands

with and without arrivals are identical across many observable and unobservable characteristics ... (p 2-3).

A deeper analysis of figure 3, though, undermines these claims. The claim that treated and untreated islands are, for all purposes, identical on various covariates is belied by the high distance between the treated islands on the Turkish coast in the east and the dozens of untreated islands on and around the Greek coast in the west, many of which are, in effect, parts of the Greek mainland rather than islands in their own right.[2] Moreover, the very fact that the number of arrivals depends heavily on proximity to the Turkish coast, with a precipitous drop-off in arrivals after the dozen or so closest islands, indicates that the bulk of the islands in the original sample were never plausible candidates for receiving treatment. We are concerned, then, that a substantial portion of the original results may be driven by the inclusion of dozens of islands that did not receive – and never plausibly *could have* received – treatment and that should not be considered identical to treated islands. We posit that the inclusion of survey respondents from islands that should not be considered as plausible treatment candidates may be driving some portion of the authors' results.

Geographic factors are one side of our concerns; the notion that distance may correlate with the outcome variable in a manner independent of refugee arrivals is another. Refugees arrived only in those islands closest to Turkey. The authors argue that distance to Turkey thus might affect anti-refugee sentiments solely through the conduit of refugee arrivals. But one might also argue that residents in islands closest to Turkey – living close to an international border, and indeed one marked by longstanding enmities between the two nationalities – may have different attitudes towards foreigners and towards refugees than do residents of islands less proximate to Turkey. These attitudes would be ingrained well before the onset of the refugee crisis and might provide another channel through which distance to Turkey influences the outcome variable. Futhermore, even after

[2]Some, in fact, lie so close to the Greek mainland as to be linked by bridge!

the onset of the crisis, it may be that areas closer to Turkey feel more threatened by the crisis itself, even if they do not actually host refugees. Thus the potential to be treated, which correlates with distance, may itself influence hostile attitudes.

Moreover, balance in the original dataset presents a third concern. As table 2 shows, the original data is typically balanced on most observable characteristics; however, the distance to Turkey is significantly higher for control units than for treated units. Indeed, the mean logged distance for islands that received refugees is 1.96 (corresponding to about seven kilometers from the Turkish coast), while the mean logged distance for islands that did not receive refugees is 5.24 (corresponding to about 189 kilometers). Since we are concerned that distance may correlate with hostility toward refugees in ways unassociated with actual arrival on an island, this is problematic.

Table 2: Balance Table for Original Data

Variable	mean.Tr	mean.Co	p-value
Age	49.75	49.05	0.29
Education	4.51	4.49	0.67
Female	0.54	0.56	0.42
Household Finance	3.00	2.93	0.11
Log Distance	1.96	5.24	0.00
Golden Dawn Vote	0.03	0.02	0.64

Note: All numbers for age, education, and household finance apart are averages of statistics generated from each of the five missing data imputations. P-values are generated from T-tests of the difference in means. Log distance, female, and Golden Dawn vote had no missing data before imputation therefore their statistics are constant across imputations.

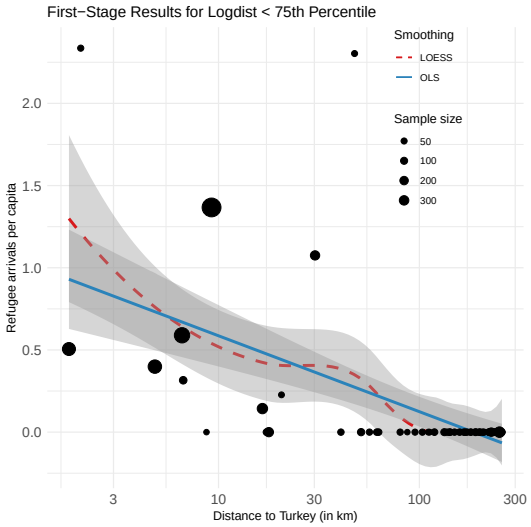
By limiting the sample to respondents living only on those islands relatively close to Turkey, we might be able to better control for pre-existing attitudes towards foreigners than does the existing

instrumental framework. Further, using a more limited sample (of islands closer to Turkey) might reduce differences between treated and control islands on correlates which remain unobserved, therefore reducing potential omitted variable biases.

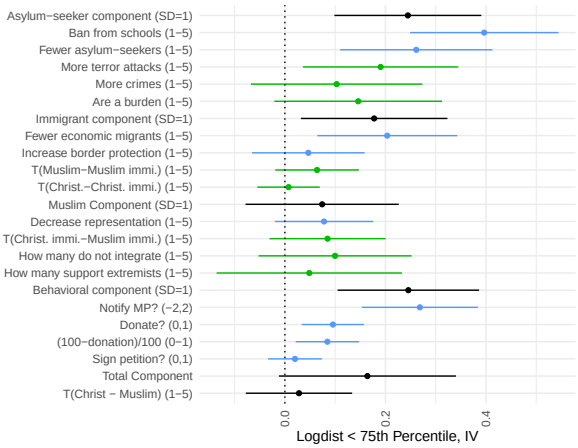
3 First Strategy: Geographic Filtering

To test whether these concerns are valid, we reproduce the authors' two-stage least squares estimation of the effect of arrivals on anti-refugee sentiments by filtering the sample set to include only islands within a certain distance (logged, following the authors' specification) of Turkey. This assumes that the assumptions of instrumental variable regression hold within whichever smaller subset of the data is selected, even if they do not hold in the complete survey. We perform this process three times; the three filtered datasets we produce include islands up to the 75th, 50th, and 25th percentiles of distance to Turkey respectively.[3] The first of these datasets, for instance, eliminates islands extremely proximate to the Greek mainland; the two subsequent datasets, in effect, move the border of exclusion eastward. First-stage and second-stage results are shown in figures 4, 5, and 6 for cutoffs at the 75th, 50th, and 25th percentiles respectively. These figures correspond with figures 1 and 4 in Hangartner et al (2018), and the second-stage plots use as the dependent variable a binary variable indicating whether an island received refugees, in accordance with the original specifications. (Versions using a continuous dependent variable of the per-capita number of arrivals produce very similar results.)

[3] Percentiles are in terms of islands, not individual survey respondents. Because more observations come from islands near Turkey, sample sizes are higher than 75%, 50% and 25% of the original 2,046 observations; see table 3.

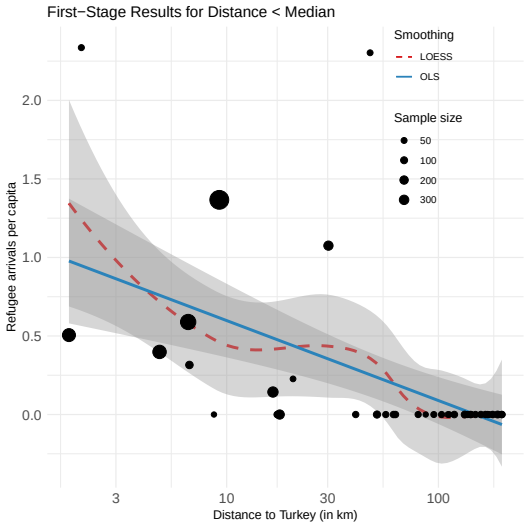


((a)) First-Stage 1

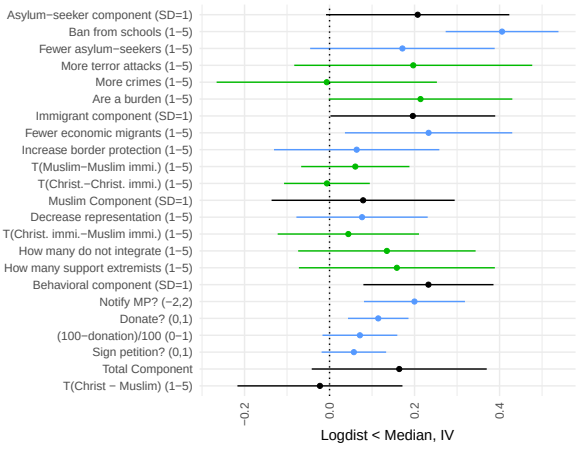


((b)) 2SLS

Figure 4: First-Stage and 2SLS for Filter at Third Quartile Log Distance

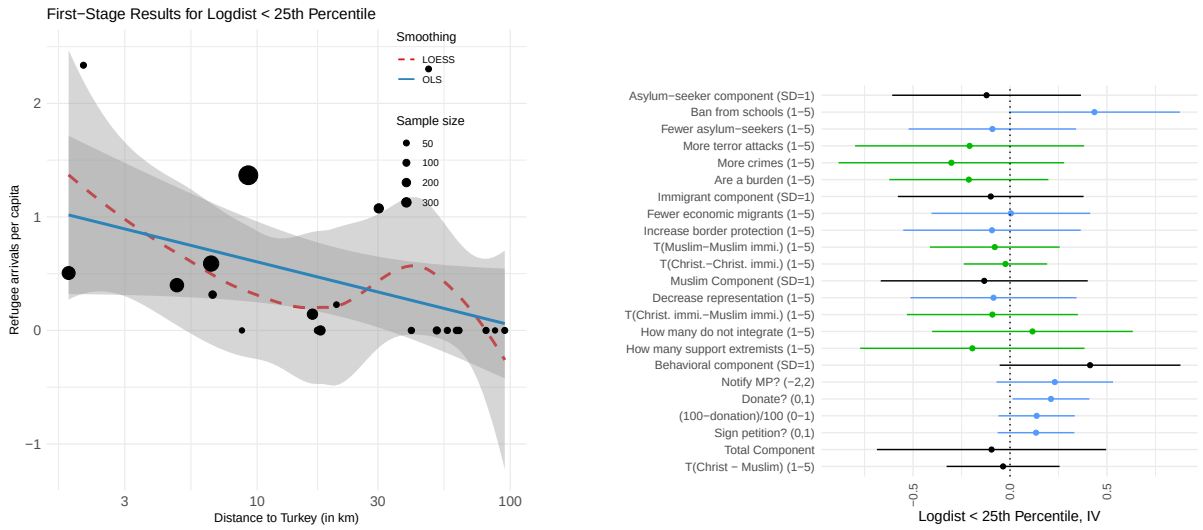


((a)) First-Stage



((b)) 2SLS

Figure 5: First-Stage and 2SLS for Filter at Median Log Distance



((a)) First-Stage

((b)) 2SLS

Figure 6: First-Stage and 2SLS for Filter at First Quartile Log Distance

To test how geographic subsetting affects the balance between treated and control groups, we present balance tests in table 3. As would be expected, the balance in terms of logged distance improves as subsetting progresses. In particular, when looking only at observations from the 25% of islands closest to Turkey, the mean distance for the control group shrinks to 35.5 kilometers, while the mean distance for the treated group remains at seven kilometers. While geographic subsetting does not eliminate the difference in balance between treated and control units, it does give us higher confidence that we are comparing groups that are similar in both observed and unobserved characteristics.

Table 3: Balance Table for Geographically Subsetted Data

Variable	Distance Percentile	Obs.	mean.Tr	mean.Co	p-value
Age	25th	1168	49.75	48.02	0.18
	50th	1395	49.75	48.74	0.25
	75th	1711	49.75	48.99	0.30
Education	25th	1168	4.51	4.37	0.22
	50th	1395	4.51	4.46	0.54
	75th	1711	4.51	4.51	0.94
Female	25th	1168	0.54	0.48	0.18
	50th	1395	0.54	0.57	0.38
	75th	1711	0.54	0.57	0.20
Household Finance	25th	1168	3.00	3.12	0.13
	50th	1395	3.00	2.97	0.65
	75th	1711	3.00	2.96	0.44
Log Distance	25th	1168	1.96	3.57	0.00
	50th	1395	1.96	4.45	0.00
	75th	1711	1.96	4.91	0.00
Golden Dawn Vote	25th	1168	0.03	0.03	0.61
	50th	1395	0.03	0.03	0.68
	75th	1711	0.03	0.03	0.74

Note: All numbers for age, education, and household finance apart are averages of statistics generated from each of the five missing data imputations. P-values are generated from T-tests of the difference in means. Log distance, female, and Golden Dawn vote had no missing data before imputation therefore their statistics are constant across imputations.

The three sets of figures displayed demonstrate that the results become progressively less robust, and the instrument progressively less valid, as the line of exclusion moves west and consequently excises more untreated observations. At the third-quartile cutoff, many of the original results become non-significant, though still with positive signs; nearly all are positive and non-significant at the median cutoff. At the first-quartile cutoff, no variables have significant coefficients. When we run

the authors' original analyses while selecting only data which we view as lying plausibly within the potential treatment group, it is far less clear that refugee arrivals do, in fact, heighten anti-refugee sentiments among natives.

4 Another Strategy: Matching

A fair critique that might be raised of our analysis in the section above is that the cutoffs chosen for subsetting the data were arbitrary. At a certain level, this is true: There is nothing special about distance quartiles that makes them a natural cutoff point. To address these concerns, we use a matching procedure to rerun the procedure that generated the original quantities of interest.

Our primary motivation for using matching is three-fold: first, as demonstrated in table 2, the original dataset contains considerable imbalance on covariates of interest; second, using a matching procedure in which we progressively raise our matching standards to prune more and more data allows us to see how results change as we iteratively subset; third, pre-processing data using matching has been shown to reduce dependence on modeling assumptions (Ho et al., 2007), which in this case may somewhat alleviate dependence on the assumptions of two-stage least squares regression.

Rather than attempting to resolve the imbalance in distance between islands that did and did not receive migrants by using geographic subsets of the data, we match observations using Mahalanobis distance matching. The covariates used for matching include age, education, gender, household finance, logged distance to Turkey, and voting for Golden Dawn (an extreme-right Greek nationalist party). Noting the balance-sample size trade off that exists in matching analysis (King, Lucas and Nielsen, 2017), we do not provide a single revised matched estimate. Instead, we use the algorithm proposed by King, Lucas and Nielsen (2017) to rerun the instrumental variables analysis using eight different post-matching subsets of the original data, successively pruning 150 observations at a time in an attempt to improve balance.

Our procedure is the following: For each hostility outcome variable, we generate a matched set

of data for each pruning cutoff between 150 and 1,200 observations. For non-principal component analysis outcomes, this yields a dataset ranging from 846 to 1,896 observations. Since there are missing values in the principle component variables generated by the original analysis, the pruned datasets for those outcomes range from 397 observations (for the smallest matched set for the behavioral component) to 1,746 observations (for the largest matched set for the immigration component). After generating the matched data sets, we use each to run the original instrumental variables model preferred by Hangartner et al. (2018). Like the authors, we use logged distance to Turkey to predict refugee arrivals, using the binary measure of arrivals as the independent variable. We then regress the hostility outcome on the arrival prediction, controlling for gender, age, and education level and clustering the standard errors by municipality, in accordance with the original analyses.

Figure 7 shows the valence – positive or negative – and the statistical significance of the beta coefficients for refugee arrivals generated from this procedure. For comparison, results are also shown for the original, unmatched data, with zero observations pruned. The estimated effect on hostility remains positive for nearly all outcomes regardless of the number of observations pruned. However, the statistical significance of the results begins to fall as the number of observations pruned increases. This would be expected given the loss of sample size, but we would stress that even pruning 1,050 observations still leaves a dataset of 996 observations. Nevertheless, the loss of statistical confidence shown in figure 7 is driven primarily by widening confidence intervals and not attenuation of the effect sizes themselves; within each outcome the β coefficients for treatment – not displayed here – seldom decrease by more than about 0.1 standardized units relative to the original results.

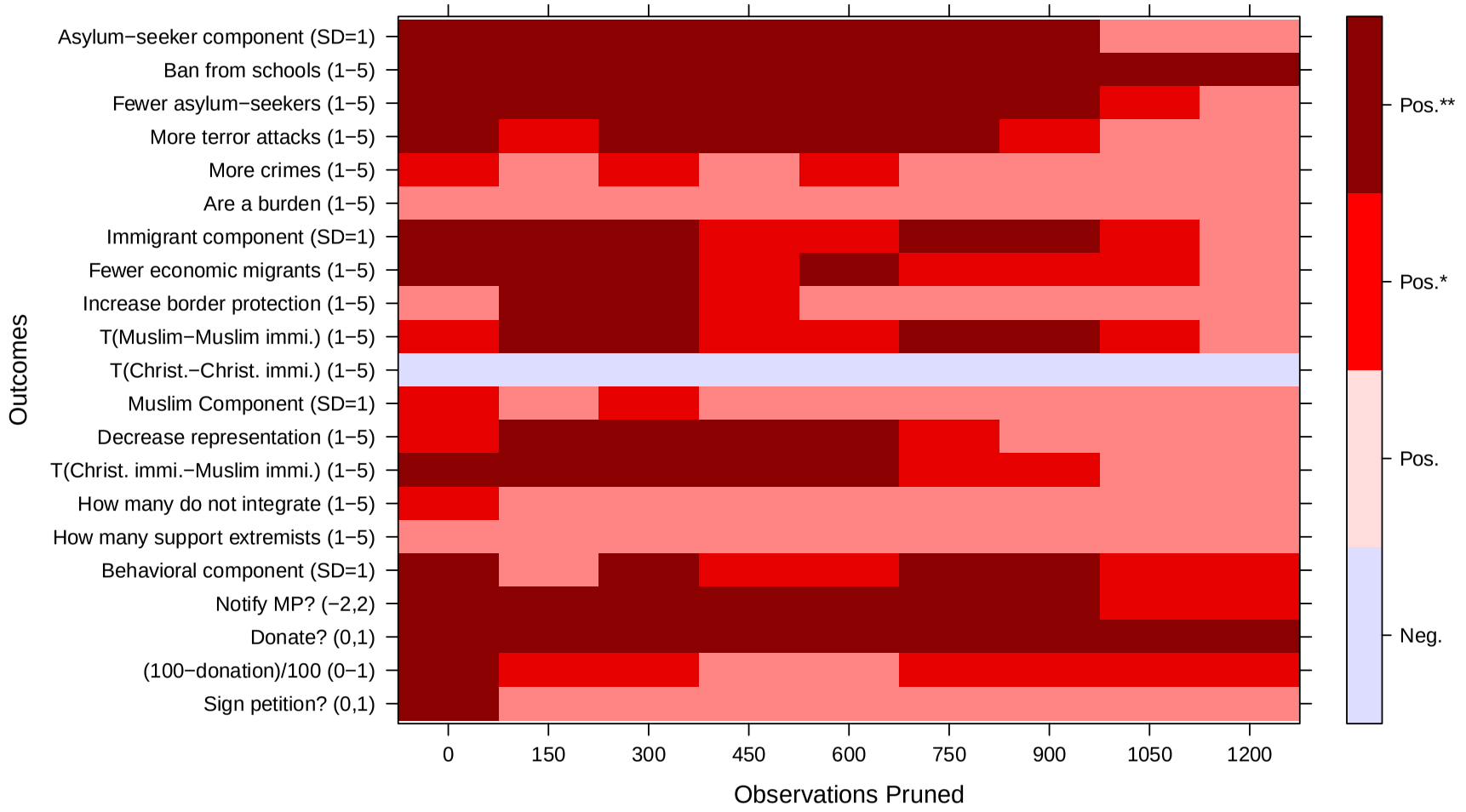


Figure 7: Significance of Results across Matched Datasets

Note: Cells are colored based on the value of the 2SLS β coefficient for refugee arrivals in each matching specification.

** $p < 0.01$, * $p < 0.05$

Of course, we are interested in whether and to what degree the matching procedure improves balance. Table 4 shows balance tests conducted at four different points along the matching frontier for non-PCA outcomes. Covariates remain balanced as matching progresses; while some p-values drop below 0.05 in some specifications (e.g. for age and education), the absolute difference between treated and control group observations is small. Our principal concern in this paper is with imbalance in terms of logged distance, which is what we believe may correlate with unobserved determinants of distance. Matching does improve balance for logged distance; when 1,200 observations are dropped, the mean logged distance for individuals living on islands with arrivals is 2.26 (9.6 km), while for islands without refugee arrivals the mean is 4.12 (51.6 km). While this difference is an improvement over the original data set, it is still substantial and statistically significant.

Table 4: Balance Table for Matched Data at Four Pruning Cut-offs

Variable	Obs. Pruned*	mean.Tr	mean.Co	p-value
Age	450	49.54	49.03	0.49
	900	49.78	49.40	0.66
	1050	50.00	48.75	0.19
	1200	50.22	48.05	0.04
Education	450	4.48	4.34	0.04
	900	4.44	4.26	0.02
	1050	4.44	4.28	0.07
	1200	4.40	4.27	0.17
Female	450	0.55	0.59	0.14
	900	0.57	0.60	0.32
	1050	0.58	0.60	0.55
	1200	0.58	0.59	0.86
Household Finance	450	3.04	3.04	0.97
	900	3.07	3.15	0.18
	1050	3.08	3.16	0.14
	1200	3.09	3.15	0.33
Log Distance	450	2.06	4.98	0.00
	900	2.20	4.61	0.00
	1050	2.22	4.37	0.00
	1200	2.26	4.12	0.00
Golden Dawn Vote	450	0.01	0.01	0.86
	900	0.00	0.00	0.74
	1050	0.00	0.00	0.63
	1200	0.00	0.01	0.54

Note: All numbers for age, education, and household finance apart are averages of statistics generated from each of the five missing data imputations. P-values are generated from T-tests of the difference in means. Log distance, female, and Golden Dawn vote had no missing data before imputation therefore their statistics are constant across imputations. *The original data had 2,046 observations.

5 Discussion and Conclusion

In this paper we have provided further analysis of an important recent finding by Hangartner et al. (2018): that significant increases in outgroup hostility may emerge within communities that host refugees, supported by survey data in Greek islands affected by the ongoing Syrian refugee crisis. While the authors make a laudable attempt to circumvent problems of endogeneity – namely, that hostility may rise in response to refugee arrivals, that levels of hostility may also determine refugee arrivals – using an instrumental approach, we are concerned that the instrument of choice, distance to Turkey, does not satisfy the exclusion restriction. In particular, we are concerned that individuals living closer to international borders may have different attitudes towards foreigners than those living farther away, and that even close areas that do not host refugees themselves may feel a greater sense of threat when on the front lines of a refugee crisis. To rephrase this theoretical point in methodological terms, we are concerned that the treatment and control groups differ in their unobserved characteristics – xenophobia and threat perception, among other factors – and that the imbalance on these characteristics may bias results. We are also concerned, in a similar vein, that the large scope of the original sample size biases results towards significance and positive signs by introducing into the sample a bevy of islands which could never have plausibly received treatment.

Rather than simply discarding Hangartner et al.'s findings, we instead pre-process their survey data using twelve different specifications that have the potential to resolve the concerns we raise. First, we attempted to limit analysis to three subsets of the data that are closest to Turkey, which to yield valid results assumes that the exclusion restriction holds within those subsets, but not for the data as a whole. This results in wider confidence intervals and the loss of statistical significance for some outcomes); under the strictest specification, the coefficient for most outcomes to become negative. Second, we used Mahalanobis distance matching to generate nine different subsets of data that have improved balance in terms of logged distance to Turkey. This procedure also reduces the

degree of confidence we have that the findings are not happenstance, although the effect on hostility uniformly remains positive. None of these twelve methods are a panacea for achieving comparability between treatment and control group; even the most restrictive cut-offs do not achieve balance in terms of distance.

The methods employed in this paper do not allow us to resolve in a definitive manner two crucial questions: how close islands receiving and not receiving treatment must be in order to be seen as comparable; and how much balance on covariates must exist between treatment and control groups for critical assumptions of causal inference to hold. For this reason, we have presented twelve different specifications that restrict the data analyzed to varying degrees, allowing the reader to determine which specification they prefer. Instrumental variables assumptions are notoriously difficult to satisfy, and it may be that none of these methods suffice to retain confidence in the original findings. On the other hand, it may be that some restrictions presented here go too far. Where the Goldilocks zone lies is, ultimately, a matter of taste – a subjective decision on the part of the researcher and the reader. Based on our analysis, we think it is likely that refugee arrivals increase hosts' hostility to some degree, although we cannot be highly confident in this relationship. We simply submit that the findings do not hold muster under the strictest geographic and matching restrictions.

References

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos and Dimitrios Xefteris. 2018. "Does Exposure to the Refugee Crisis Make Natives More Hostile?" *American Political Science Review* pp. 1–14.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political Analysis* 15(3):199–236.

King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics* 39:119–125. **URL:** *http://gking.harvard.edu/papers*

King, Gary, Christopher Lucas and Richard A Nielsen. 2017. "The balance-sample size frontier in matching methods for causal inference." *American Journal of Political Science* 61(2):473–489.

Sovey, Allison J and Donald P Green. 2011. "Instrumental variables estimation in political science: A reader's guide." *American Journal of Political Science* 55(1):188–200.