The Pseudoscience Behind the Racial Threat Theory:

Inconsistencies and Irregularities in Enos n.d. [2012] and Enos (2016)

Social Scientists for Research Integrity

August 2018

1 Introduction

This report originated in research integrity concerns raised over the past years regarding the 2016 article "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior" by Ryan D. Enos, published in the American Journal of Political Science (AJPS). The article, including a 'Supporting Information' appendix, is available at https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12156 and at the author's website, http://ryandenos.com/papers/; [1] the persistent URL for the replication data at the AJPS Dataverse is https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26612. Collective credit is attributed to all scholars who over the years have identified red flags in the work reviewed herein.

In brief, Enos (2016) attempts to probe the "racial threat" theory by examining how turnout and Republican vote in the general elections changed between 2000 and 2004 among White voters who previously resided in close proximity to public housing projects in the Chicago area (primarily inhabited by African Americans) that were demolished starting around 2000. The author purports that the demolition of these projects can be regarded as a "quasi-experiment" (p. 6 [129]), with demolition acting as treatment variable, and "change in white political participation and support for conservative candidates" (ibid.) as the outcome variable, enabling him to recover a "causally identified" effect of "racial threat" (p. 2 [125], p. 17 [139]).

Three hypotheses are formulated on p. 6 [129] of the article in order to test the "racial threat" theory empirically:

> "**H1 (Racial Threat and Turnout)**: After the demolition of the projects, turnout should decline for white voters close to the projects relative to the rest of the city.
>
> **H2 (Proximity and Size)**: [...] the treatment effect should decline as the white voters are farther away from a project and as the population of a project represents a smaller portion of the local outgroup population. [...]
>
> **H3 (Racial Threat and Vote Choice)**: After the demolition of the projects, white voters close to the former projects should experience a decline in racially conservative voting relative to the rest of the city."

Based on empirical analyses performed on both individual voter level and aggregate precinct-level data reported on pp. 8-16 [130-138], the author concludes that he has "presented evidence for racial threat that is relatively free of questions of endogenous confounding" (p. 17 [139]), and both the turnout rates and support

[1] In referencing pages from the article, the first parenthetical number indicates the page in the `enoschicago.pdf` preprint downloadable at the author's website, while numbers within square brackets indicate page numbers in the published version.

for the conservative presidential candidate declined significantly among White voters who had previously resided in close proximity to African-American occupied public housing projects, because, the argument runs, the demolition "caused a significant change in their voting behavior" (ibid.). The author further claims that "the strength of the effect decreased with distance from the project and increased with the size of the outgroup", that "[t]he estimated effect of racial threat, over 10 percentage points, is substantively large in many elections", and that "[t]he white voters in Chicago were threatened by a spatially proximate, yet segregated, outgroup." (ibid.)

Setting aside logically fallacious statements and reasoning errors in Enos (2016), as well as a causal inference discussion on the identification assumptions needed to warrant the above conclusions, this report focuses on the execution of the empirical analyses, in particular the data transformations (at both individual and aggregate levels) that produced the samples on the basis of which results are generated and reported.

2 Noncompliance with the AJPS Replication Policy

The AJPS replication policy is detailed at https://ajps.org/ajps-replication-policy/ and https://ajpsblogging.files.wordpress.com/2016/05/ajps-replic-guidelines-ver-2-1.pdf, henceforth "AJPS Replication Guidelines". Extensive quotes from relevant excerpts are included in the Appendix to this report. In essence, the policy states that[2] (i) authors of accepted manuscripts "must provide replication materials that are sufficient to enable interested researchers to reproduce all of the analytic results that are reported in the text and supporting materials" and "are not permitted to "embargo," or withhold, information that has been used to perform an analysis featured in an AJPS article" (pp. 1-2); (ii) that "[e]very Dataverse Dataset must contain complete information for constructing the analysis dataset(s) from the original data sources (p. 4); (iii) that the author must supply "a software command file for [extracting variables and observations from another, larger, source dataset]" and that "the file always should contain commands for selecting the relevant variables, extracting subsets of observations if necessary, performing any data transformations that are carried out prior to the analysis itself, and assigning missing values" (p. 4); (iv) that in the cases of editor-granted exceptions to posting original data sources, "materials that specify the procedures through which an interested researcher can apply for access to the analysis dataset for replication purposes (including the construction of the analysis dataset from the original source dataset) from the holders of the source data." (p. 10)

A cursory inspection of the replication files associated with the Enos (2016) article immediately reveals that the article is out of compliance with the AJPS replication policy on several counts. Specifically, the following data management steps are not clearly documented in the article or the Supporting Information document, and, most critically, no statistical software code to execute them is included in the replication archive:

1. The process of merging voter file datasets from what the author refers to as the "Illinois voter file", presumably https://www.elections.il.gov/downloads/votinginformation/PDF/NormalizedFormat.pdf, which would have required, according to the link above, "that the three tables (files) be joined to access voting history or to determine political subdivisions. This process requires extensive knowledge of and experience with database software and programming."

2. The process of geocoding of voter addresses, i.e., retrieving their latitude and longitude coordinates, which the author states that he performed (p. 4 of Supporting Information).

[2]Page numbers refer to "AJPS Replication Guidelines" document.

3. The computation of distances from voter residences to the demolition projects. The author claims: "Using GIS software that is integrated with a Structured Query Language database, I measured the distance between every voter and the projects." (pp. 4-5 of Supporting Information). Yet no GIS or SQL code is provided, or even a brief clarification on the metric according to which distance was measured (Euclidean, taxicab, etc.), and how he chose the unique variable value for each voter on the `demo.distance` variable in the `turnout.csv` file (e.g., keeping the minimum of the set of distances with respect to all projects).

4. Obtaining aggregate census tract data, and matching voters to a host of demographic variables by census block. The author states (p. 4 of Supporting Information): "I did this by writing a computer program that sent each unique address in the voter file to the Census Bureau website's address search service." No code for performing this operation is provided, however.

5. Computing probabilities of race conditional on name using Bayes rule by implementing the formula in Equation (1) in a statistical software package like `R` (p. 3 of Supporting Information).

6. The process of merging voter file to homeownership data (p. 5 of Supporting Information).

7. The deletion of observations corresponding to voters residing within 10 meters from the projects.

8. The sources for and process of creating precinct-level aggregate datasets.

In addition to withholding statistical software code used to perform these processes and enable researchers to fully replicate the analyses starting from original data sources, some less consequential forms of non-compliance are the author's failure to provide a codebook for the datasets, and specify the seed used in computational procedures requiring random number generation, such as the bootstrap.

3 Discrepancies With Respect To Official Statistics

The Enos (2016) replication datasets exhibit numerous discrepancies with respect to official statistics released by the Board of Election Commissioners for the City of Chicago (henceforth BECC), https://chicagoelections.com/en/election-results.html, both at individual and aggregate level.

One immediately obvious such discrepancy, which is highly consequential for the author's results, pertains to the turnout figures for 2004. The official number of registered voters in 2004 according to BECC is 1,416,101, and the number of ballots cast is 1,056,830 (see Figure 2), i.e., 359,271 registered voters did not vote according to official statistics. The individual-level turnout dataset Enos starts with before executing the code in the `turnout.R` script (`data.turnout.csv`) has 1,132,646 observations (registered voters). The only data management operations performed prior to this configuration of the dataset were discarding observations for which no race probability could be computed and/or which the author was unable to geocode, and those living within 10 m of the projects—at least according to Enos' own statements on p. 4 of the Supporting Information document, as no `R` code was supplied by the author to ascertain exactly which data transformations were actually applied (in violation of the AJPS replication policy, as described above). Of these 1,132,646 observations (roughly 20% less than official numbers), only 669,115 have variable value 1 on the `vote2004` variable, and the remainder have value 0, i.e., 463,531 registered voters did not vote according to the Enos (2016) replication dataset, which is 104,260 more non-voters than the official figure. While the attrition of 20% of observations can be attributed to the author's stated inability to geocode a number of addresses and/or compute a probability of voter's race given their last name, even assuming all 283,455

registered voters discarded on those grounds had voted, leaving the author with a particularly unlucky draw of registered voters (which is statistically implausible), it would have been *mathematically impossible* for the maximum number of non-voters in a subsample, however peculiar, to exceed the official number of non-voters for the entire sample. Further, there would have been no logical reason for the values on the `vote2004` variable to change from 1 to 0 for at least 104,260 observations in the process of merging datasets and deleting observations that the author was unable to geocode, unless those data transformations had been performed intentionally.

The turnout rate that can be computed on this initial replication dataset (before an additional data transformation consisting of removing voters that had registered after the registration cutoff date of October 10, 2000, set by the author in this particular paper), is approximately 59.08%, a much lower figure compared to the official turnout rate of 74.63% (see code section **R 1** for the calculations).

In the final dataset consisting of 848,061 observations (five less than Enos claims in article) on which article results are reported, turnout is somewhat overestimated for 2000, and severely underestimated for 2004, generating a heavily biased sample: according to the Enos (2016) dataset, turnout decreased from 2000 to 2004 by 13%, from 75.34% to 62.63%, when in reality it increased by 5% from 69.79% to 74.63% (see Figure 1 and Figure 2 for official results, and code section **R 2** for calculations).

The sample sizes and difference-in-differences estimated reported in Figure 1 of Enos (2016) and in the Supplementary Information document on p. 8 can be replicated exactly when running the code provided by the author, so this must be the exact dataset the author used in producing and reporting those results.

Another set of readily detectable irregularities that contradict official statistics surfaces in the aggregate-level data with ward-precint as unit of analysis on which results reported on pages 13-16 [135-138] of Enos (2016) are based. Consider for instance the `data.votechoice.2010.csv` dataset, and the 2004 number of ballots cast and presidential vote variables in this dataset, which are crucial to probing the author's diminished turnout and Republican vote hypotheses in precincts located in close proximity to demolition projects. The author does not provide any explanations as to the procedures for creating this dataset, or statistical software code to reproduce the necessary data management steps, again contrary to AJPS replication policy requirements. The sequence of undocumented transformations that generated this dataset resulted in the deletion of 843 out of the 2,709 total precincts for which BECC has released official 2004 data (i.e., approximately 31% of the Chicago precincts), distributed across 46 distinct wards (out of the 50 Chicago wards), and cumulatively accounting for 315,449 ballots cast in the general election of 2004, and 50,792 votes for the Republican presidential candidate George W. Bush (that is, roughly 27% of the total of 188,056 votes cast for this candidate across the Chicago area). Table 2 provides an overview of the missing precincts in the Enos (2016) data within each affected ward, with within-ward totals for ballots cast in the 2004 election and vote counts for presidential candidate Bush, respectively, in the last two columns (see **R 8** for generating the table).

The data management steps required to assemble such an aggregate dataset from official turnout and presidential vote data sources such as BECC are trivial (see e.g., code section **R 6** for an illustration), and can be expected to be performed in an error-free manner by an entry-level research assistant with minimal quantitative training. It is difficult to conjecture a scenario where accidental errors could have accumulated to result in the deletion of this substantial amount of precincts, and how simple tabulations of totals against official numbers performed as a routine verification step could not have led to the discovery and correction of such accidental errors.

In terms of the geographic distribution of the missing precincts in the Enos (2016) data, the side-by-

side comparison maps displayed in Figure 4 and Figure 5 reveal a striking pattern of proximity to the demolition projects (superposed as labeled datapoints at the appropriate latitude and longitude coordinates, as detailed in **R 9**): there is a systematic tendency for precincts on a small radius around the demolition projects to be missing (demolition projects almost appear as enclaves surrounded by zero-turnout precincts), a data feature that is particularly propitious to the author's diminished turnout/Republican vote hypotheses. Compounding the effect, some of the missing precincts had, according to official BECC statistics (see right panels in the figures indicated above), very high numbers of ballots cast and/or votes cast for the Republican candidate George W. Bush. Most conspicuously missing is an area in the central region of Chicago around the Northerly Island, Soldier Field, Grant Park, Millennium Park, and vicinity, as well as a shore stripe between Lincoln Park and North Avenue Beach, consisting of precincts with over 1,000 ballots cast and/or over 50% Bush votes (i.e., in the right tail of the distribution of ballots cast/Bush vote percentages). These precincts were in close proximity to demolition projects such as the Harold Ickes or the Cabrini-Green Homes; had they been kept in the dataset it is statistically implausible that results such as the differential effects reported in Figure 5 (p. 14 [136]) of Enos (2016) would have emerged. Along similar lines, precincts with low numbers of ballots cast and/or low percentages of Republican vote in remote locations such as the Far Southeast Side area of Chicago were deleted, which further reinforced the data patterns postulated by the author's hypotheses.

Figure 6 highlights differences between ballots cast and percent Bush votes between the Enos (2016) replication data and the official BECC data, for the precincts that were kept in the former dataset. Most values (grey areas on maps) were identical; in a few precincts however, values appear to diverge from official statistics, in both directions. The observable patterns are consistent with ballot numbers having been increased the most in precincts located at large distances from demolition projects, such as Far North Side or Southwest Side.

Differences between the Enos (2016) and the official BECC data range between -238 and 568 for the number of ballots cast in the 2004 general election within a given precinct, and -22.72% and 26.59% for percentages of votes in favor of Republican candidate Bush (see **R 7**).

Code sections **R 9**, **R 10**, **R 11**, and **R 12** show how the maps were generated. Shapefiles for boundary coordinates of precincts are from https://catalog.data.gov/dataset/precincts-2010/.

4 Other Irregularities in the Datasets

Other anomalies are present in the Enos (2016) datasets, such as the distribution of age at voting registration in the `data.turnout.csv` dataset. Based on the variables in this dataset, over 1800 voters completed the voter registration process before the age of 17, or even prior to their birth (in extreme cases nearly seven decades before they were born). This could have been an error arising when merging datafiles during processes not documented in the `R` code provided by the author, but possibly also voter registration fraud and/or social security number (SSN) fraud (the state of Illinois allows voter registration using SSN for those who lack valid identification). The occurrence of fraud in the 2004 general elections is corroborated by higher than 100% turnout rates in several precincts according to official data, so it is likely that Enos may not be responsible for the existence of this particular irregularity, although mentioning it and removing anomalous observations from the dataset would have been an appropriate course of action. Leaving these observations in the dataset is favorable to the author's hypotheses, since density plots indicate that observations to which voter fraud suspicions are applicable have a very high probability of being Black and very low probability of being White, according to Enos' own race probability computations (see Figure 3, and code section **R 3**).

5 Inconsistencies Across Paper Versions

Early iterations of the analyses reported in Enos (2016) can be traced back to the author's dissertation project at the University of California Los Angeles (UCLA), titled "Spatial Impact: The influence of groups in geographic space on individual political behavior" (submitted in 2010), and, as the author acknowledges on the first page of the article, prior versions of the paper presented at various seminars at Harvard, UCLA, Princeton, and other universities. The dissertation is, as of this writing, no longer available in the ProQuest database; a search returns the message that the thesis is neither available for viewing nor purchase, at the request of the author.

It was possible, however, to retrieve older versions of the paper using The Internet Archive Wayback Machine (henceforth "The Internet Archive"), https://archive.org/about/faqs.php#The_Wayback_Machine, a website launched by a 501(c)(3) non-profit organization. According to the website description at the link above, the organizing principle of the website is centered around "a three dimensional index that allows browsing of web documents over multiple time periods".

The Internet Archive has recorded 83 captures of the author's website, http://ryandenos.com, between June 13, 2009 and August 6, 2018, as shown here: https://web.archive.org/web/*/http://ryandenos.com (see also Figure 8). A search on the specific link to the older version of the paper indicates that it was saved in the Internet Archive 7 times between March 28, 2012 and February 17, 2017: https://web.archive.org/web/*/http://ryandenos.com/papers/chicago_threat.pdf. There are no readily apparent differences across these 7 versions; subsequent analyses in this report are based on the March 28, 2012 capture of the paper (see Figure 9), the earliest version available, which can be retrieved here: https://web.archive.org/web/20120328223557/http://ryandenos.com/papers/chicago_threat.pdf. The old title is "What tearing down public housing projects teaches us about the effect of racial threat on political participation", nearly identical to the 2016 article aside from the synonym substitutions. Although undated, in the sequel this paper will be referenced as Enos (2012).

Except for the fact that there are no aggregate-level analyses in the older version, there is substantial overlap across the two versions, both in the introduction and theoretical framework presented, as well as in the hypotheses and empirical strategy for the individual-level turnout analyses: the same difference-in-differences approach (but without bootstrapped confidence intervals in the former version) is used to report findings in Figure 1 on p. 32 and Figure 2 on p. 33 in Enos (2012) as in Figure 1 on p. 9 [131] of Enos (2016) and Figure A.1 on p. 8 in the Supporting Information document associated with this article. The figures show effect magnitudes and sample sizes for differences in turnout between 2000 and 2004 across various distance ranges from the demolition projects, for White voters only in former, and for both White and Black voters in the latter; each figure panel corresponds to a specific probability of voter race given his/her name. The sample selection procedure, according to the author's own descriptions in Section 6.2 on pp. 26-27 of the older version, and pp. 4-5 of the Supporting Information for the newer version, respectively, is virtually identical. The only stated difference is the registration cutoff date used: January 1, 2000 in the former version, and October 10, 2000, in the latter, respectively.

Consider for example the panels corresponding to probability of White race being 1: panel (a) in Figure 2 on p. 33 in Enos (2012), and panel (h) in Figure A.1 on p. 8 of the Supporting Information for Enos (2016). Peculiarly, the mere change in registration cutoff date tripled the effect magnitude for White voters living within 100 meters of demolition projects, from slightly below -0.05 (decline in turnout by 5%) to around -0.15 (decline in turnout by 15%). Similar increases in effect magnitudes were produced within other categories, as defined by distance. Further, estimates in all categories failed to attain statistical significance at the .05-level

in the Enos (2012) version, as confidence intervals for all distance-defined categories overlapped zero; in the newer version, statistical significance is somehow achieved in 7 out of the 10 distance categories.

Even if the only change performed *had* been shifting the registration cutoff date, failure to state that the effects are extremely sensitive to this choice of cutoff date, resulting in a threefold decline in the effect magnitude in one distance category and loss of statistical significance in all distance categories when the date is changed to January 1, 2000, amounts to an unethical research practice popularly referred to as "p-hacking"— essentially a form of falsification. It can however be demonstrated statistically that additional data transformations aside from cutoff date change must have occurred between the two versions, which are not documented in the latter version, much less theoretically justified. It should also be noted that the author suppressed the discussion regarding the lack of statistical significance included in the 2012 version, and simply states in the 2016 version that his hypothesis received empirical support, capitalizing heavily on the 13.4% estimate in Figure 1 on p. 9 [131], which is patently statistically insignificant, as the corresponding confidence interval overlaps zero. Instead, the author claims, with no justification, that the estimate is "substantively significant", despite the fact that the necessary condition of statistical significance is not met, and the author should have been fully aware of the volatility of this effect and its large fluctuations depending on registration cutoff date.

No replication data were made available by the author for the Enos (2012) version; nor was he under any obligation to do so, as the paper had not been published at the time. However, since the only stated difference is the registration cutoff point, all else is identical in terms of data preprocessing according to the author's descriptions on the relevant pages indicated above, and the data used in Enos (2016) is a *superset* of the data used in Enos (2012), it should be a trivial exercise to recover the 2012 estimates on the basis of the 2016 data by simply modifying the data subsetting statement in the `data.turnout.R` file to change the registration cutoff date from 10-10-2000 to 01-01-2000 (see code section).

The results of this exercise are reported in Figure 7 (see code section **R 4**). Bearing in mind that the left and middle panels are expected to be identical for the rationales delineated above, let us proceed to a side-by-side examination of the mean estimates and sample sizes for voters whose probability of being White was 1 and .99. It can be noticed that both sample sizes and effects are different, implying that *additional*, undocumented data transformations must have been performed between the two paper versions on the subsample of voters who had registered by January 1, 2000. These transformations increased sample sizes in some of the categories, decreased it in others, increased all negative effect magnitudes, changed the direction of previously positive effects to negative, and led to certain previously non-significant estimates becoming statistically significant. These changes, all of which are highly favorable to the author's diminished turnout among Whites hypothesis, occurred *already* on this subset of the data, i.e., *prior* to the change in registration cutoff date, and therefore *cannot* be attributed to a simple change in registration cutoff date. The changes are consistent with *both* the addition and the deletion of observations in several distance-defined categories, which are neither accounted for in the sample selection description nor documented in the replication `R` code provided by the author. For example, in the category defined by the combination of $P(White|name) = 1$ and distance from demolition projects less than 100 meters, the net change of +4 observations (from 16 to 20) produced a threefold increase in the magnitude of negative effect, bringing it in closer alignment with the author's preferred hypothesis. Similarly spectacular increases in effect magnitudes were produced by positive net changes in sample sizes in other categories, whereas deletions of hundreds of observations in the categories corresponding to larger distances reduced the upper bounds of 95% confidence intervals, bringing those effects closer to statistical significance.

It is impossible to determine the exact number of added and deleted observations, but the net change was a reduction in sample size from "about 850 thousand voters", as stated in Enos (2012) p. 27, to 753,909 overall, i.e., at least 100,000 observations were deleted while others were added. The substantial amount of deleted observations is presumably a partial explanation for why turnout is heavily biased downwards with respect to official statistics.

While defensible theoretical rationales for removing observations might be supplied, there is *no plausible source* for the addition of observations corresponding to White voters to the sample. According to the author's statements in the 'Sample Selection' section of Enos (2012) and based on the `R` code, only four (non mutually exclusive) types of observations had been deleted: (i) those he was "unable to geocode" i.e., determine latitude and longitude coordinates of; (ii) those for which he was unable to compute probability of race given last name; (iii) those within 10 m of the projects to preempt potential bias from geocoding error, a subset consisting only of African American voters based on the author's own reporting; (iv) observations for which no registration date was available, as it would have been impossible to determine whether they registered before or after a given cutoff date.

Inability to geocode an observation (type (i)) would have made impossible the computation of distance from projects. Either this or inability to calculate the probability of race given name, or the combination of types (i) and (ii), would have made it impossible to conclude, e.g., that a given observation has probability of being White equal to 1, and lives within 100 m from the projects. No observation has variable value less than 10 on the distance variable (type (iii)), and had the author chosen to include observations from that pool despite the potential geocoding imprecision argument, the additions would have consisted of African American, not White voters. Finally, all observations included in the analysis above have a non-missing registration date variable value by the sample construction procedure, so the source of type (iv) can also be ruled out.

Consequently, the additions must have occurred either through the modification of existing values on one or more of three variables (distance, race probability, and registration date), or the complete fabrication of new data points not present in original sources. The change in the cutoff date further amplified the magnitude of the negative effects, and induced statistical significance in many previously non-significant categories. In the absence of further information, it is not possible to verify whether additional data transformations might have been performed on the subsample of voters who registered between January 1, 2000, and October 10, 2000.

Implicitly, all subsequent individual-level analyses produced after the dataset creation, including those using matching reported in Figure 2 on p. 10 [132] and Figure 3 on p. 12 [134], as well as the predicted effects in Figure 4 on p. 13 [135], no matter how carefully and honestly executed from that point onwards, are completely unreliable, since they were performed on a sample that had been *heavily distorted*, in the direction of the author's favorite hypotheses, by means of substantial deletion and potentially fabrication of observations.

Notice also that there should exist *no logical connection* between the abnormal decline in turnout produced by undocumeted deletions of observations in the individual-level dataset, and the similarly reduced turnout in the aggregate data caused by the removal of nearly one third of the precincts, explained in a preceding section, as these data come from different sources, and must therefore be processed separately— *unless* the changes leading to these artificially downward biased turnout statistics were *not* performed accidentally and independently from one other.

6 Other Single-Authored Work

The 2016 article is not unique in having attracted research integrity concerns in the academic community. A 2014 Proceedings of the (PNAS) piece by the same author (Enos (2014)), available at http://www.pnas.org/content/pnas/111/10/3699.full.pdf, with Supporting information at http://www.pnas.org/content/pnas/suppl/2014/02/19/1317670111.DCSupporting/pnas.201317670SI.pdf, and reporting similar results, has been the target of criticism in a brief letter by van Hoorn (2014), http://www.pnas.org/content/111/19/E1938:

"The recent field experiment by Enos [(2014)], showing that intergroup contact strengthens exclusionary attitudes, provides a showcase for discussing three main pitfalls. These pitfalls can be labeled as (i) effect sizes, (ii) effect duration, and (iii) subject selection, and have far-reaching implications for the real-life significance of results of attitudinal experiments, sufficient to warrant a different conclusion to the results presented by Enos."

The same work received an indirect characterization as "pseudoscience" in the comments section of a blog post by Columbia Professor Andrew Gelman. The blog post itself is a critique of work by one of Enos' advisees, but the comments refer both to said advisee's and Enos' work: http://andrewgelman.com/2018/05/29/exposure-forking-paths-affects-support-publication/. The comment by user 'Vince S" on May 31, 2018 is worth quoting at length:

> "Ryan:
>
> I'll try to say this politely as I can, but the difference between science and pseudoscience is for the researcher, at least to the extent humanly possible, to be aware of his own biases and how they can color his interpretation of the data, and to properly account for them. There is more to this than just computing p-values.
>
> And both you and Sands [Enos' advisee whose work is criticized in the blog post] make conclusions unwarranted by the data. As you say, Researchers shouldn't claim their effects are generalizable without clear evidence. Evidence, not prior bias (which means getting representative samples). It's up to you whether you want to admit this in the name of doing better science, or dig in your heels and take a defensive posture. [...]
>
> I read your 2014 paper. You do a one-tailed test and don't correct for multiple comparisons in Table 1. If that's what it takes to achieve significance then yes I think most people, including myself, would describe findings as "noisy". (Granted, two-tailed Bonferroni-corrected significance would be barely obtained for the number of immigrants question.) Anyway, the proper conclusion is that results are consistent with a short-term effect of outgroup exposure on exclusionary attitudes, but future research will be necessary to see if the effect generalizes to different populations, and whether the effect lasts longer than a few days (it may well be the case that long-term exposure to outgroups in fact lessens exclusionary attitudes.) The data doesn't "support" or "prove" or "show" that outgroup exposure affects exclusionary attitudes.
>
> I'd also like to point out that the question in "experiment" doesn't match the question in "results". The question in "experiment" is "Would you favor allowing persons that have immigrated to the United States illegally to remain in the country if they are employed and have no criminal history?" whereas it is described later as "Children of undocumented being allowed to stay?" Perhaps this was a typo, but what was the actual question?"
>
> [...] it is still true that both of you used the population in suburban Boston because it was

> convenient. You don't really know whether your results will generalize to, say, the Midwest, or Europe. All you have is the circular argument that your experiment "proves" the phenomenon you claim it does and that, therefore, your results should generalize."

(In his response to the above, Enos dis not clarify what the question actually was and whether a typo had occurred.)

This indicates a pervasive pattern in the author's work, and that the 2016 article was not an isolated instance in using data manipulation techniques intended to divert results in the author's preferred direction, and claiming causal effects where none were warranted.

7 Impact of Report Findings on Other Published Work

Two chapters of the Enos (2017) book, "The Space Between Us: Social Geography and Politics", borrow heavily from the 2016 and 2014 articles discussed above: Chapter 5 ("Boston: Trains, Immigrants, and the Arizona Question") and Chapter 6 ("Chicago: Projects and a Shock to Social Geography"). These chapters include reproducing empirical results in the articles; for instance, Figure 1 on p. 9 [131] of Enos (2016) is reproduced as Figure 6.3 on p. 152 of the book with only cosmetic changes but exact same estimates, produced on the biased sample described in Section 5 of this report. Many of the conclusions made in the book are therefore also implicitly affected.

8 Conclusions

Using straightforward mathematical and statistical arguments, this report has produced evidence to the effect that the datasets used to generate the research results reported in Enos (2016) and available in the corresponding AJPS replication archive do not accurately reflect the electoral landscape surrounding the 2004 general election in the city of Chicago. Both the individual-level voter dataset `data.turnout.csv` and the aggregate precinct-level `data.votechoice2010.csv` are impacted by major distortions in the sample composition that beyond any reasonable doubt skewed results in favor of author's preferred theory and hypotheses.

The effects reported in Enos (2016) are the artifact of data transformations that are neither theoretically explained and motivated on substantive grounds, nor documented in the statistical software code files provided by the author, placing the author out of compliance with the AJPS replication policy. These undocumented transformations, resulting in both the addition and deletion of observations in the individual-level dataset, and the deletion of nearly one third of the precincts and modifications of variable values recording ballots cast and percentage vote for Republican presidential candidate in the aggregate-level dataset, produced samples that are highly at variance with official statistics released by Chicago election authorities. Due to these distortions, every single analysis reported in the article, from simple difference-to-difference estimates, to estimates based on statistical procedures such as regression analysis, predicted turnout probabilities, matching, ecological inference etc., was impacted.

As a consequence, the article does not accurately represent the research processes that the author purports to have performed, and the conclusion that "[t]his causally identified effect supports classic formulations of racial threat" (p. 17 [139]) is unwarranted, as results generated on such heavily distorted datasets cannot be ascribed any measure of credibility.

Statistically, the joint probability of such an accumulation of accidental errors on both individual- and aggregate-level data, occurring independently from one another and inducing severe biases in the direction

of the author's favorite hypotheses, is virtually zero. Of course the errors could still be, in theory, purely accidental. From a legal standpoint, it is Enos who has the burden of proof under the circumstances, as the "honest error" exemption is an affirmative defense (Charrow (2010), p. 70). The first natural step the author needs to take toward meeting this burden of proof is to rectify the noncompliance with the AJPS replication policy by uploading to the dataverse the missing pieces of statistical software code that could demonstrate that the errors were not a deliberate act.

References

Charrow, Robert P. 2010. *Law in the Laboratory: A Guide to the Ethics of Federally Funded Science Research.* Chicago University Press.

Enos, Ryan D. 2012. "What tearing down public housing projects teaches us about the effect of racial threat on political participation." https://web.archive.org/web/20120328223557/http://ryandenos.com/papers/chicago_threat.pdf.

Enos, Ryan D. 2014. "Causal effect of intergroup contact on exclusionary attitudes." *Proceedings of the National Academy of Sciences* 111(10):3699–3704.

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60(1):123–142.

Enos, Ryan D. 2017. *The Space Between Us: Social Geography and Politics.* Cambridge University Press.

van Hoorn, André. 2014. "Significance of attitudinal experiments." *Proceedings of the National Academy of Sciences* 111(19):E1938.

APPENDIX

AJPS Replication Policy: Relevant Excerpts

https://ajps.org/ajps-replication-policy/

> "The corresponding author of a manuscript that is accepted for publication in the American Journal of Political Science must provide replication materials that are sufficient to enable interested researchers to reproduce *all* of the analytic results that are reported in the text and supporting materials. [...] authors also can make their replication files available elsewhere (e.g., their personal website, other data repositories, etc.) as long as *all* of the necessary files are included in the Dataset on the AJPS Dataverse." (emphasis added)

https://ajpsblogging.files.wordpress.com/2016/05/ajps-replic-guidelines-ver-2-1.pdf

pp. 1-2, 'General Principles' section

> "Authors are not permitted to "embargo," or withhold, information that has been used to perform an analysis featured in an AJPS article (except as described below, in the section on "Possible Exceptions to Data Access Requirements"). Instead, authors should provide all information that is required to reproduce and evaluate any analytic result (in quantitative analyses) [...] that is reported in their article."

p. 3- , 'Instructions and Recommendations for Quantitative Analyses'

> "In some rare cases (see the section below on "Possible Exceptions to Data Access Requirements"), the analyses in an AJPS article may be based upon restricted data that cannot be posted in a publicly-accessible location. [...] any such exceptions to the general AJPS replication policy must receive explicit approval from the Editor [...] But, the author still must provide instructions that interested researchers can use to access the data (see the section on ?Information to Reproduce the Analysis Dataset? below), as well as formatting and variable definition information for the data that are analyzed in the AJPS article."

p. 4 'Information to Reconstruct the Analysis Dataset:'

> "Every Dataverse Dataset must contain complete information for constructing the analysis dataset(s) from the original data sources. [...] regardless of the specific details, interested researchers always must be able to follow the author's instructions in order to reproduce the precise data values used for any analyses reported in the AJPS article."

> "The analysis dataset often is created by extracting variables and observations from another, larger, source dataset [...] In such cases, the author must provide a software command file for doing so. [...] the file always should contain commands for selecting the relevant variables, extracting subsets of observations if necessary, *performing any data transformations that are carried out prior to the analysis itself*, and assigning missing values. And, it is important to identify the specific version of the source dataset and the date that it is accessed in order to construct the analysis dataset. The analysis dataset sometimes is created by merging information extracted from several other sources. In such cases, the author must provide the relevant software commands for extracting the data from the separate sources, and for merging the separate subsets

of data into the overall analysis dataset. *Complete reference information* must be provided for all source datasets used to construct the analysis dataset. " (emphasis added)

p. 10 'Possible Exceptions to Data Access Requirements'

"If the AJPS Editor grants permission to withhold some or all of the analysis dataset, the source dataset, or cited data, then the exempted information does not need to be uploaded to the AJPS Dataverse. The author will be required to include a note at the beginning of the published article explicitly acknowledging the limitations on data availability and describing the restrictions that prevent public access to the exempted data. A Dataset still must be created in the AJPS Dataverse, containing materials that specify the procedures through which an interested researcher can apply for access to the analysis dataset for replication purposes (including the construction of the analysis dataset from the original source dataset) from the holders of the source data."

Board of Election Commissioners for the City of Chicago: Official Statistics

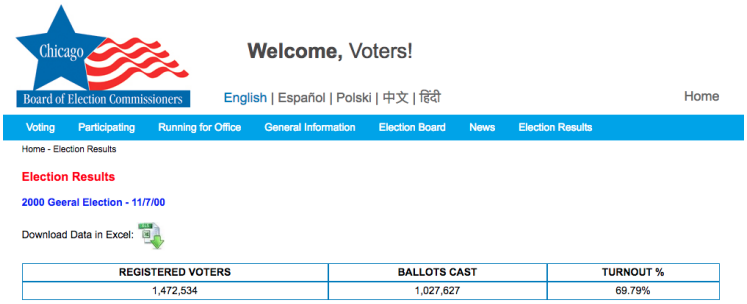

Figure 1: Board of Election Commissioners for the City of Chicago: Official Voter Registration and Turnout Statistics for the General Election of November 2000. See https://chicagoelections.com/en/election-results.html.

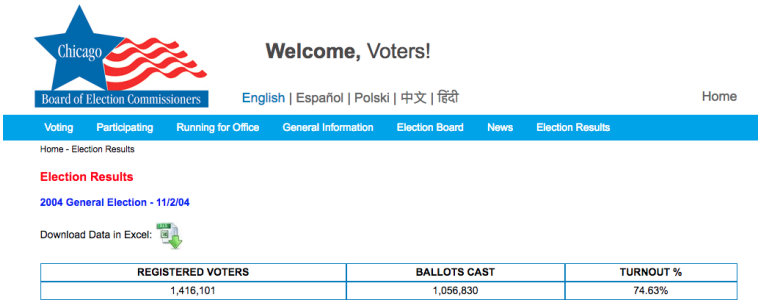

Figure 2: Board of Election Commissioners for the City of Chicago: Official Voter Registration and Turnout Statistics for the General Election of November 2004. See https://chicagoelections.com/en/election-results.html.

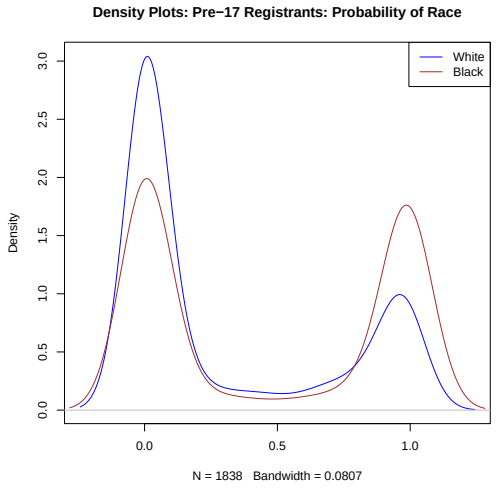

Figure 3: Density plots for voters registered pre-birth or before age 17 in Enos (2016) replication data, `data.turnout.csv` file, by probability of race given name.

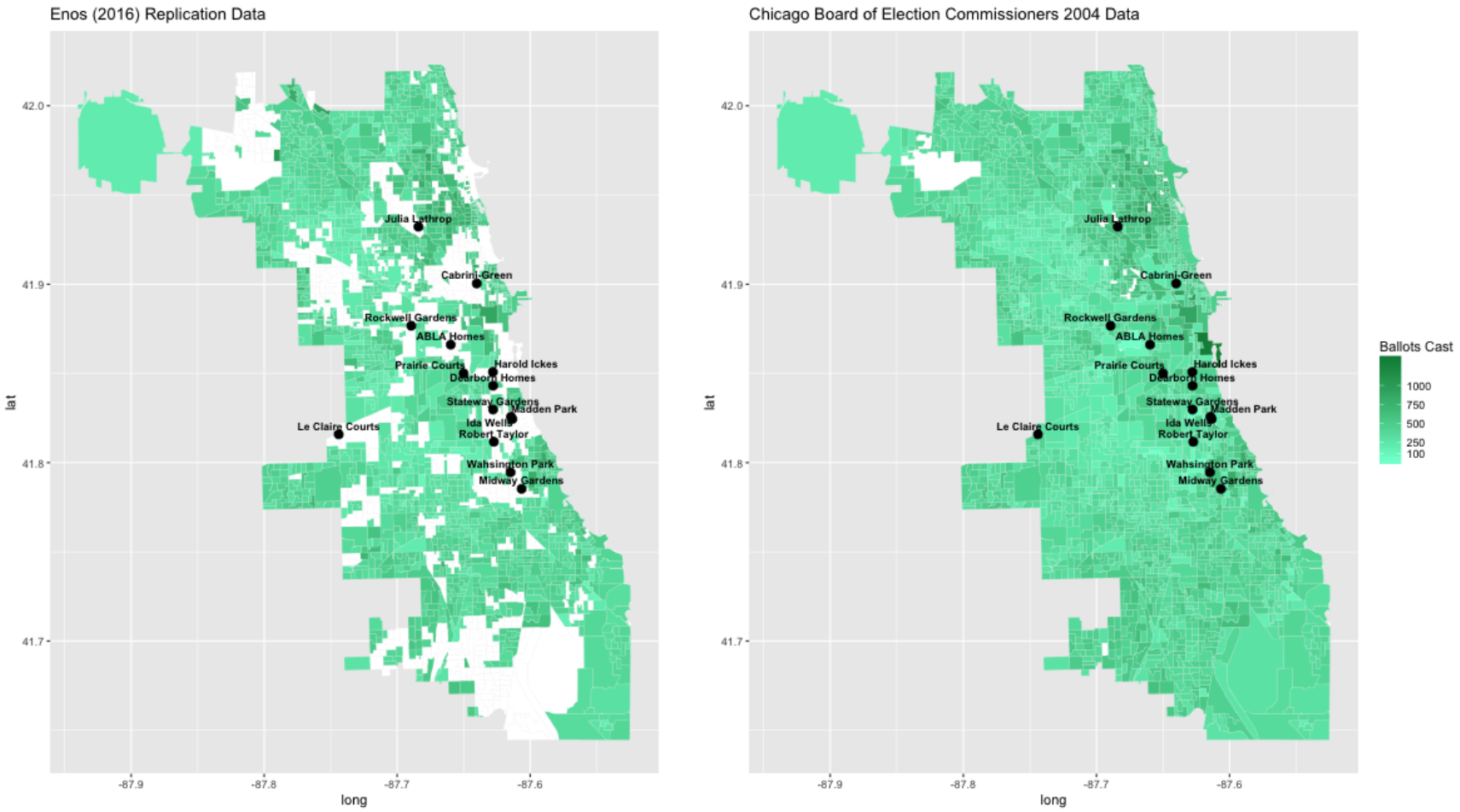

Figure 4: Comparison of Ballots Cast in the General Election of November 2004 in Chicago, by Precinct. Left panel: Geographic distribution based on Enos' `data.votechoice.2010.csv` file (see https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26612). Right panel: Geographic distribution based on official statistics released by the Board of Election Commissioners for the City of Chicago (see https://chicagoelections.com/en/election-results.html). Locations of demolition projects are superposed as labeled dots (see R code for coordinate sources).

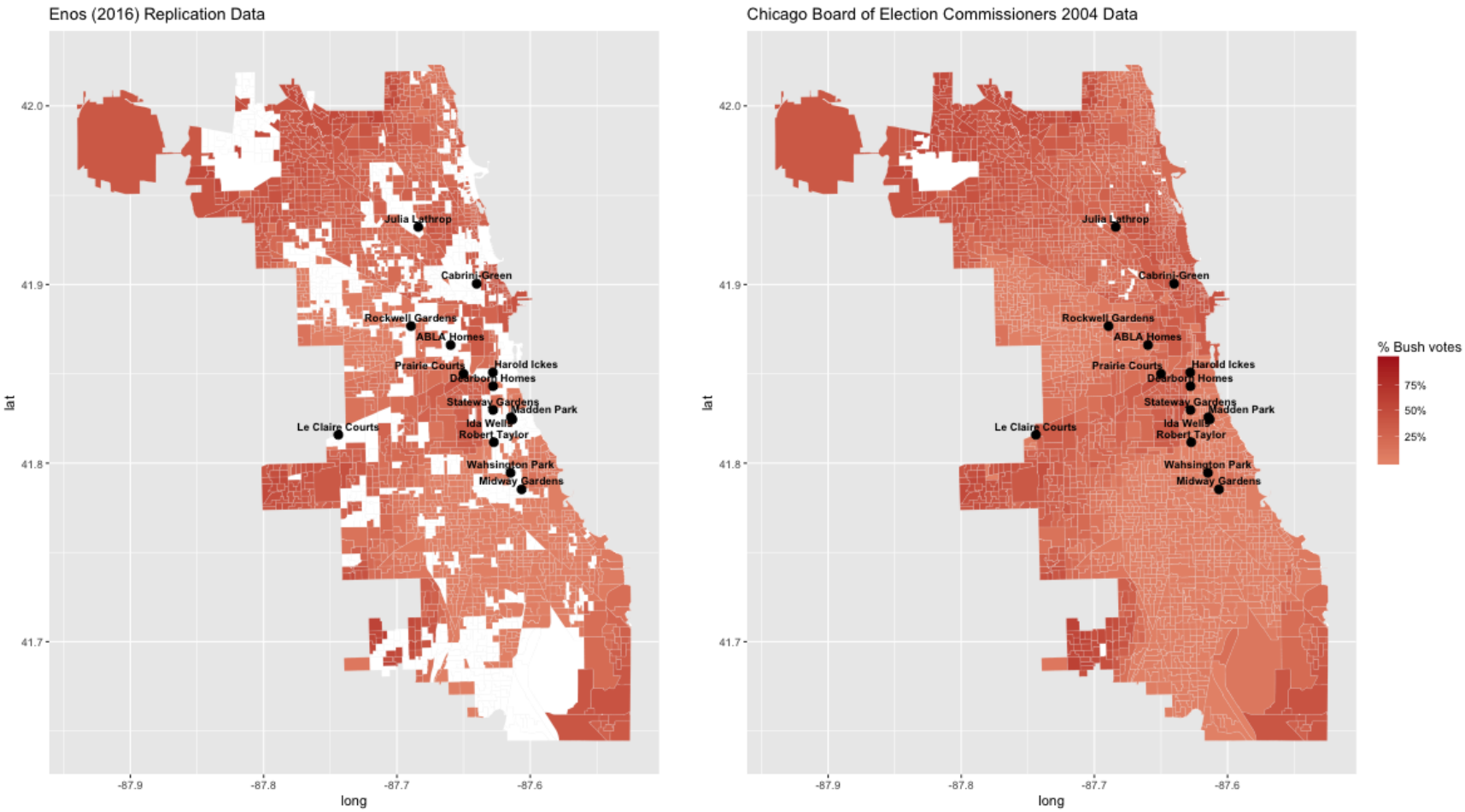

Figure 5: Percent Votes for Presidential Candidate G. W. Bush in the General Election of November 2004 in Chicago, by Precinct. Left panel: Geographic distribution based on Enos' `data.votechoice.2010.csv` file (see https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26612). Right panel: Geographic distribution based on official statistics released by the Board of Election Commissioners for the City of Chicago (see https://chicagoelections.com/en/election-results.html). Locations of demolition projects are superposed as labeled dots (see R code for coordinate sources).

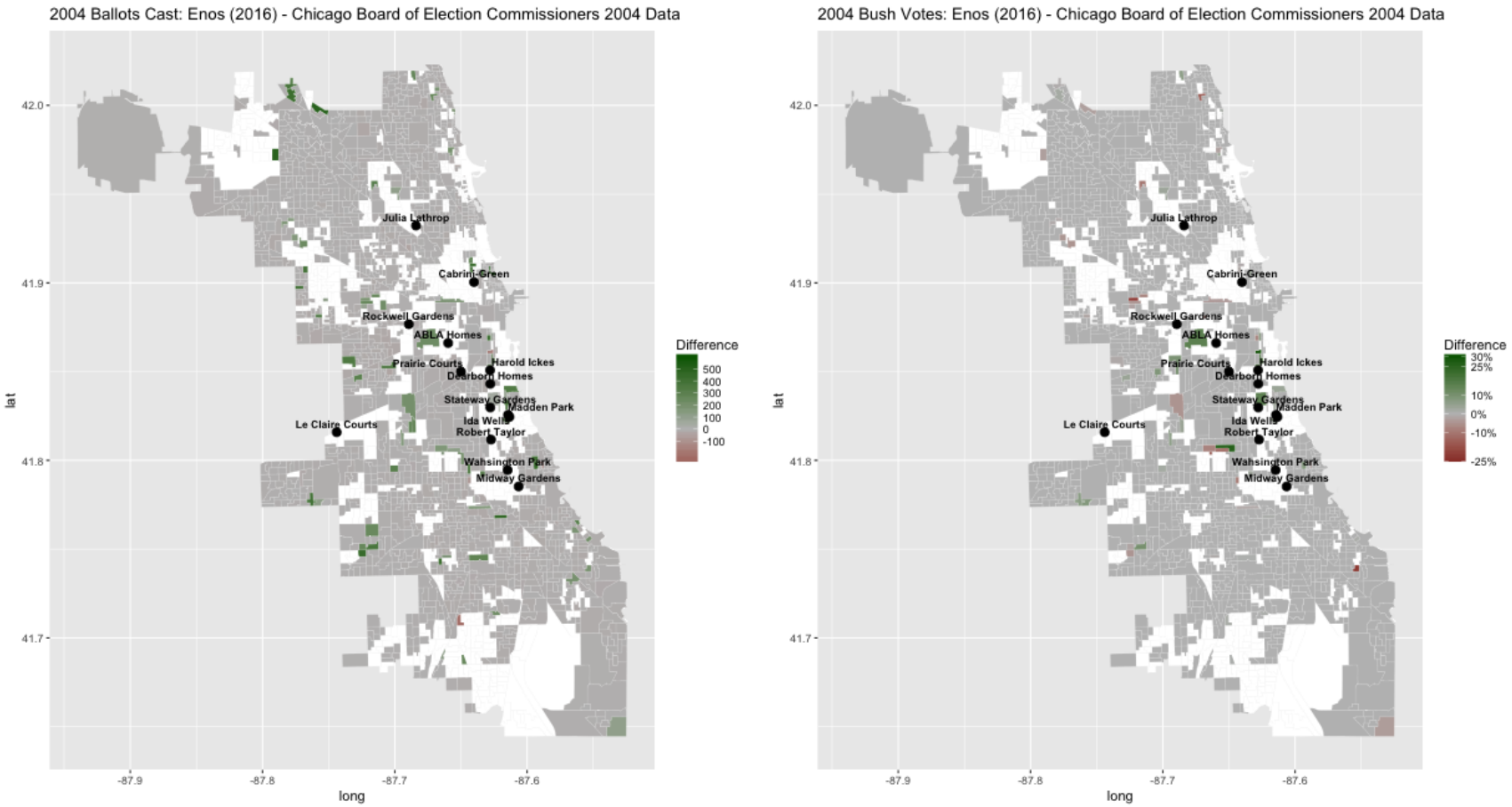

Figure 6: Differences between Enos' `data.votechoice.2010.csv` data (see https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26612) and official statistics released by the Board of Election Commissioners for the City of Chicago (see https://chicagoelections.com/en/election-results.html). Left panel: Ballots Cast in the General Election of November 2004 in Chicago. Right panel: Percent Votes for Presidential Candidate G. W. Bush in the General Election of November 2004 in Chicago. Locations of demolition projects are superposed as labeled dots (see R code for coordinate sources).

Ward	Missing Precincts	Ballots Cast	Bush Votes
1	1,2,3,4,5,7,8,10,11,14,17,19,21,22,23,25,26,27,28,33,36,37,38,39,40	14205	2706
2	3,4,5,9,10,13,14,15,16,18,19,20,24,25,26,27,29,30,31,32,34,36,38,39,41,44,45,49,52,55,56,57,58,59	13107	1582
3	1,2,4,6,8,9,11,14,15,16,18,21,23,24,27,29,30,31,32,33,38,40,42,43,44,46,47,51,52,53,54,55,56,57,58,59,60,61,62,63	11464	593
4	4,5,6,7,14,18,19,20,26,33,35,37,39,40,42,49,53,54,55,56,57	7990	280
5	33,55	623	18
6	8,9,35,45,50,65,66	2603	65
7	11,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65	6339	203
8	67,68,69,70	1665	47
9	2,3,4,6,7,8,11,13,14,15,16,17,18,19,21,24,25,26,27,28,30,33,35,36,37,39,41,42,44,45,47,49,50,52,53,54,55,56,57,58	15842	526
10	28,40,47,48,49	1620	166
11	5,21,24,31,34,46	1978	602
12	3,6,7,11,12,18,19,25,26,27,28,29	2733	532
13	5,6,9,10,11,14,15,16,23,25,28,30,31,33,34,36,40,44,46,47,48,49,50,51,52,53,54	9180	2547
14	8,12,13,22,23,25,27,35,36,37	2559	605
15	49,50,51,52	1155	59
16	1,4,8,10,17,18,19,29,30,39,42,43,44,45,46,47	4480	313
17	56,57,58,59,60,61,62,63,64	3338	61
18	19,32,51,52,63,64	2228	125
19	4,6,7,11,14,15,29,30,32,33,34,35,36,37,38,40,41,42,43,44,48,50,51,59,64,65,66,67	12791	4250
20	1,2,3,6,7,11,12,13,14,16,17,18,19,23,24,32,33,37,39,42,47,50,51,52,53,54,55	9532	229
21	3,4,7,17,23,32,43,46,65,67,71,72,73,74,75,76	6361	179
22	3,27,30	829	109
23	7,21,31,37,55,56	2259	499
24	4,54,55,56,57,58,59	2278	84
25	12,14,18,24,25,26,27,28,29,30,31	4905	1100
26	38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63	5725	910
27	2,6,9,11,12,13,14,16,20,21,27,32,35,37,38,39,41,42,43,44,46,48,49,50,51,53,54,56,57,58,59,60,61,62,63,64,65,66	12870	1576
28	3,5,9,11,13,16,21,25,27,31,33,35,37,39,40,41,43,44,45,46,47,51,52,56,57,58,61,62,63,64,65,66,67,68,69,70,71,72,73	10717	330
29	6,7,10,11,16,18,19,25,30,36,39,49,50,51,52,53,54	6163	515
30	3,6,8,10,16,17,18,21,22,25,27,34,36,42,43	4190	1039
31	19,46,47,48,49,50,51	1599	394
32	46,49,50,51,52	2497	825
33	1,2,3,4,5,6,7,9,11,12,15,17,19,20,22,24,26,27,28,29,30,31,32,33,34	10036	2169
34	1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25,27,28,30,31,32,34,36,37,38,39,41,42,43,44,45,46,48,49,50,54,55,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73	23972	648
35	11,13,31,33,35,36	2069	436
36	43	335	102
37	1,2,3,4,5,6,7,8,12,13,14,15,17,18,19,20,21,22,23,24,25,26,27,28,29,31,32,33,34,35,36,37,38,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58	17876	886
40	13,16,48,49	1561	364
41	1,2,3,4,8,9,11,12,13,15,16,17,18,20,22,24,25,28,29,30,31,32,34,35,36,37,38,40,41,42,43,44,45,48,49,50,52,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71	21204	9726
42	3,12,14,17,22,41,53,58,59,67,68,69	6680	2427
43	2,3,4,7,8,9,10,11,17,20,21,22,23,24,25,29,31,32,41,43,49,52,55,56,57,58,60,61,62,63,64,65,66,67	14768	4958
44	9,11	1787	575
46	3,5,7,8,9,11,12,13,14,16,17,19,20,24,25,32,33,34,35,37,38,40,41,42,43	15035	2803
47	3,25,44,46,47,48,49,50,51	4205	1092
48	3,9,11,15,17,20,29,32,34,47,49,50,51,52,53,54,55,56	7147	1195
49	38,39,40,41,42,43,44	2949	342
	Total: 843 precincts in 46 wards	**315449**	**50792**

Table 2: Missing precincts in Enos (2016) aggregate data for the 2004 general election (`data.votechoice.2010.csv`), by ward. Last two columns indicate total ballots cast and total vote counts for presidential candidate G. W. Bush, respectively, across all missing precincts for each ward.

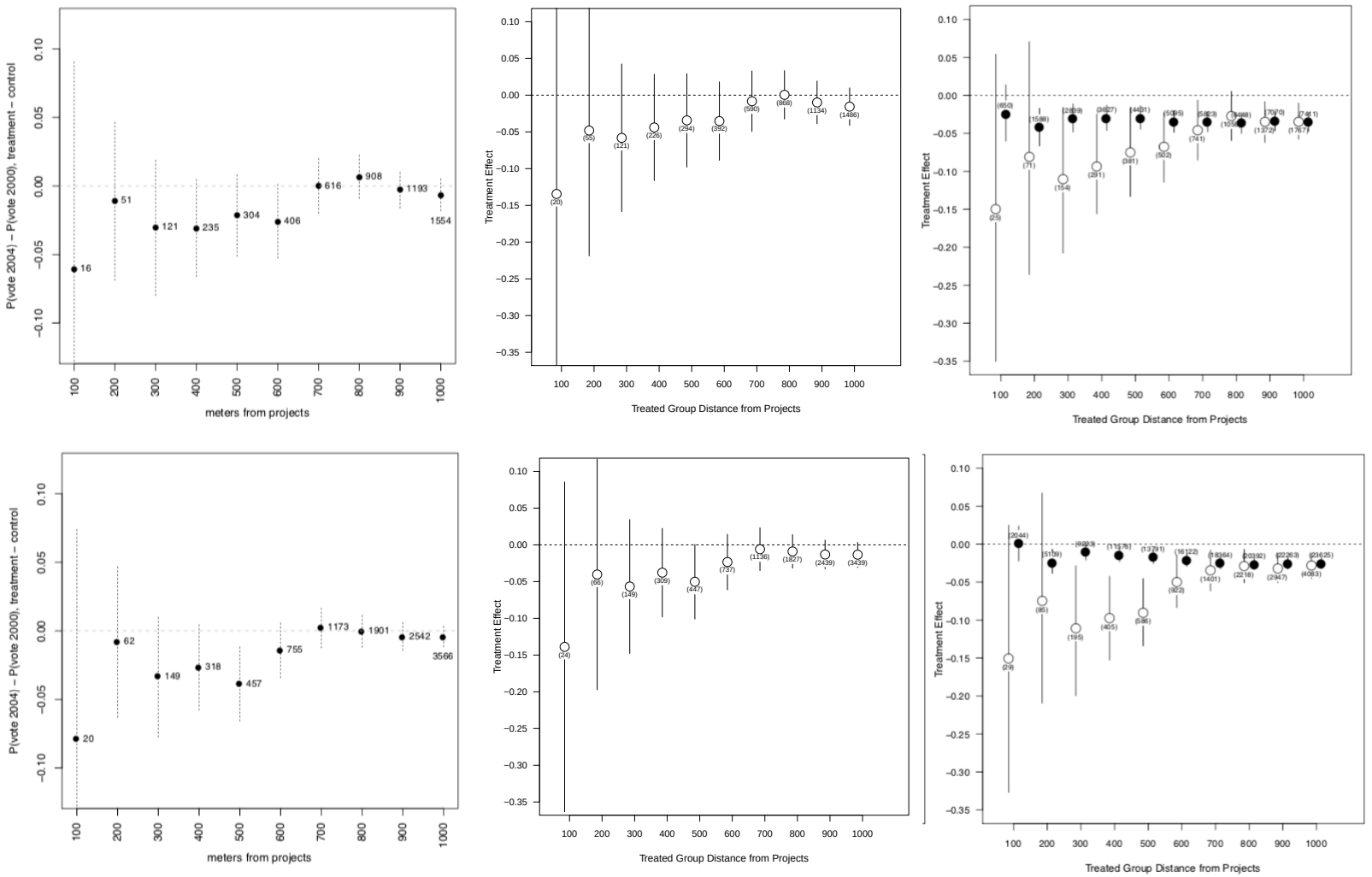

Figure 7: Comparison of Enos' results based on `data.turnout.csv` file (see https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26612) across paper versions. Left: Original estimates reported in Enos (2012), Figure 1, panel (a), p. 33 (registration cutoff date January 1, 2000). Middle: Estimates on 2012 subset (registration cutoff date January 1, 2000) using Enos (2016) replication data. Right: Estimates reported in Supporting Information for Enos (2016) article (registration cutoff date October 10, 2000). Top row: $P(White|name) = 1$. Bottom row: $P(White|name) = .99$

R Code to Reproduce Findings in This Report

Note: Separate .R code file also available.

R 1: 2004 Turnout: Official statistics vs. Enos (2016) replication data

```
##  \cite{Enos2016} 'data.turnout.csv' dataset: 2004 turnout
dim(data)[1] # 1132646 registered voters vs. 1416101 officially
dim(data[which(data$vote2004==1),])[1] # 669115 ballots cast vs. 1056830
    officially
dim(data[which(data$vote2004==1),])[1]/dim(data)[1] #59.08% turnout vs.
    official 74.63% turnout
#(see https://chicagoelections.com/en/election-results.html)

1416101-1056830 # BECC: registered voters - ballots cast
# according to BECC 359,271 registered voters did not vote
dim(data)[1]-dim(data[which(data$vote2004==1),])[1]
dim(data[which(data$vote2004==0),])[1]
#according to \cite{Enos2016} 463,531 registered voters did not vote
1416101-1056830-(dim(data)[1]-dim(data[which(data$vote2004==1),])[1] )
#104,260 values of 1 on vote2004 variable changed to 0
```

R 2: 2004 Turnout: Official statistics vs. Enos (2016) replication data

```
#-----------------------------------------------------------------
use.data = data[data$reg<"2000-10-10"&is.na(data$reg)==F,]
#-----------------------------------------------------------------

### Turnout rates
dim(use.data[which(use.data$vote2000==1),])[1]/dim(use.data)[1] #75.34%
dim(use.data[which(use.data$vote2004==1),])[1]/dim(use.data)[1] #62.63%
```

R 3: Possible voter registration/SSN fraud

```

### Voters who registered before the age of 17 or pre-birth
use.data$yr_reg=as.numeric(substring(use.data$reg,1,4)) #create
    registration year variable
summary(use.data$yr_reg) #get summary stats on registration year
use.data$age_reg=use.data$yr_reg-(2004-use.data$age) #create age @
    registration variable
summary(use.data$age_reg) # min is -65.11, i.e., voter registered 65 yrs
    prior to birth
pre17=use.data[which(use.data$age_reg<17),] #get pre-17 registrants subset
```

```
# pre-17 registrants (possible SSN and/or voter registration fraud)
pdf("pre17reg.pdf") #figure showing density by race
plot(density(pre17$whitename),col="blue",
     main="Density Plots: Pre-17 Registrants: Probability of Race")
lines(density(pre17$blackname),col="brown")
legend("topright", lty=c(1,1), col=c("blue","brown"), c("White","Black"))
dev.off()
```

R 4: Replicate results in Enos n.d. [2012]

```
### Replicate results in Enos n.d. [2012]

### These lines of code from Enos' turnout.r file
# create vectors of distances to demolition projects & race|name probabilities
# 9x10 storage matrix for results
# results displayed in Fig. 1 p. 9 of \cite{Enos2016} (enoschicago.pdf document)
# and Supplemental Information Fig. A1 p. 8 (enoschicagosi.pdf document)
#-------------------------------------------------------------------
##distances used repeatedly in estimation below
dists = seq(from = 100, to = 1000, by = 100)
##basic diff in diffs in paper, estimated across multiple definitions of white and distances
cat('begin basic difference-in-differences estimation \n')
namepcts = c(seq(from = .91, to = .96, by = .01),.975,.99,1)
##matrices for stroing results
res.mat = matrix(nrow=length(namepcts),ncol=length(dists))
#-------------------------------------------------------------------

### These lines of code from Enos' turnout.r file
#only modification: change cutoff date to Jan 01, 2000, instead of Oct 10, 2000,
# to obtain the subset of observations used in Enos n.d. [2012] (chicago_threat.pdf document)
#"I discarded all voters that had moved or newly registered after January 1, 2000" (p. 27)
#-------------------------------------------------------------------
###registration is Illionis is cutoff 27 days prior to election day, limit to these individuals
#use.data = data[data$reg<"2000-10-10"&is.na(data$reg)==F,]
use.data = data[data$reg<="2000-01-01"&is.na(data$reg)==F,]
```

```
#-------------------------------------------------------------------------

### Get no. of observations
dim(use.data)[1] #753909 vs. "about 850 thousand voters" (Enos n.d.[2012] p. 27)
# approx. 100,000 voters discrepancy

### Difference-in-difference estimates for Whites
# with conventional confidence intervals as in Enos n.d. [2012] (chicago_threat.pdf)
# replicate Figure 2 on p. 33 estimates
### The following lines adapted from \cite{Enos2016} turnout.R file
#only difference: they produce conventional rather than bootstrap CIs; much faster
# mean estimates reported in \cite{Enos2016} are NOT bootstrapped; only CIs; not much difference
# applying this code to \cite{Enos2016} sample yields same mean estimates as in article & Suppl Info
## Create storage matrices for means & ub & lb of CI
white.treat.effect.mean=res.mat
white.treat.effect.conf.lower=res.mat
white.treat.effect.conf.upper=res.mat
white.treat.N=res.mat
## Loop thru indices populate storage matrices w/ estimates
for(j in 1:length(namepcts)){
useW = use.data[use.data$whitename>=namepcts[j],]
for(h in 1:length(dists)){
Wtreat = useW[useW$demo.distance<=dists[h],]
Wcont = useW[useW$demo.distance>dists[h],]
white.treat.N[j,h] = nrow(Wtreat)
ttest=t.test(Wtreat$vote2004-Wtreat$vote2000,Wcont$vote2004-Wcont$vote2000)
white.treat.effect.mean[j,h]=mean(Wtreat$vote2004-Wtreat$vote2000)-mean(Wcont$vote2004-Wcont$vote2000)
white.treat.effect.conf.lower[j,h] = ttest$conf.int[1]
white.treat.effect.conf.upper[j,h]=ttest$conf.int[2]
}
}
```

R 5: Chicago Board of Election Commissioners: General Election 2004: Turnout Data by Precinct

```
##### Official BECC 2004 turnout data
#install.packages("readxl") #first install pkg if not already installed
require(readxl)
#Data ranges for wards in Board of Election Commissioners 2004 turnout
    Excel file
rngs=c("A10:D50","A54:D113","A117:D180","A184:D241","A245:D300",
"A304:D370","A374:D439","A443:D513","A517:D575","A579:D628",
"A632:D682","A686:D715","A719:D773","A777:D814","A818:D870",
"A874:D921", "A925:D989", "A993:D1057", "A1061:D1128", "A1132:D1187",
"A1191:D1267", "A1271:D1301", "A1305:D1361", "A1365:D1424", "A1428:D1459",
"A1463:D1526", "A1530:D1596", "A1600:D1673", "A1677:D1731", "A1735:D1778",
"A1782:D1833", "A1837:D1889", "A1893:D1927", "A1931:D2004", "A2008:D2044",
"A2048:D2103", "A2107:D2165", "A2169:D2222", "A2226:D2273", "A2277:D2326",
"A2330:D2401","A2405:D2474", "A2478:D2545", "A2549:D2599", "A2603:D2656",
"A2660:D2703", "A2707:D2758", "A2762:D2818", "A2822:D2866", "A2870:D2915")
# Export data for each ward to a "ward#.csv" file
for(ward in 1:50){
  write.csv(assign(paste("ward", ward, sep = ""),
                   cbind(ward,as.data.frame(read_excel("dataexport_2004
                       turnout.xlsx",range = rngs[ward])))),
            paste("ward",ward, ".csv" ,sep=""),row.names=FALSE    )
}
# Stack ward datasets vertically
allframes=lapply(1:50,function(x)
  read.csv(paste("ward",x,".csv",sep=""), header=TRUE))
BECC2004turnout = as.data.frame(do.call(rbind,allframes))
colnames(BECC2004turnout)[2] <- "precinct"
colnames(BECC2004turnout)[5] <- "Turnout"
#Order dataframe by ward then precinct
BECC2004turnout =BECC2004turnout[
  order( BECC2004turnout[,1], BECC2004turnout[,2] ),
  ]
#
# Verify totals
sum(BECC2004turnout$Ballots.Cast) #1,056,830 ballots cast
sum(BECC2004turnout$Ballots.Cast)/sum(BECC2004turnout$Register.Voters) #
    74.63% turnout
#
```

R 6: Chicago Board of Election Commissioners: General Election 2004: Presidential Vote Data by Precinct

```
##### Official BECC 2004 presidential vote data
#Data ranges for wards in Board of Election Commissioners 2004
    presidential vote Excel file
rngs=c("A11:H51","A55:H114","A118:H181","A185:H242","A246:H301",
"A305:H371","A375:H440","A444:H514","A518:H576","A580:H629",
"A633:H683","A687:H716","A720:H774","A778:H815","A819:H871",
"A875:H922", "A926:H990", "A994:H1058", "A1062:H1129", "A1133:H1188",
 "A1192:H1268", "A1272:H1302", "A1306:H1362", "A1366:H1425", "A1429:H1460"
    ,
"A1464:H1527", "A1531:H1597", "A1601:H1674", "A1678:H1732", "A1736:H1779",
 "A1783:H1834", "A1838:H1890", "A1894:H1928", "A1932:H2005", "A2009:H2045"
    ,
"A2049:H2104", "A2108:H2166", "A2170:H2223", "A2227:H2274", "A2278:H2327",
"A2331:H2402","A2406:H2475", "A2479:H2546", "A2550:H2600", "A2604:H2657",
"A2661:H2704", "A2708:H2759", "A2763:H2819", "A2823:H2867", "A2871:H2916")
# Export data for each ward to a "ward#.csv" file
for(ward in 1:50){
  write.csv(assign(paste("ward", ward, sep = ""),
                   cbind(ward,as.data.frame(read_excel("dataexport_2004
                       president.xlsx",range = rngs[ward])))),
            paste("ward",ward, ".csv" ,sep=""),row.names=FALSE    )
}
# Stack ward datasets vertically
allframes=lapply(1:50,function(x)
  read.csv(paste("ward",x,".csv",sep=""), header=TRUE))
BECC2004president = as.data.frame(do.call(rbind,allframes))
colnames(BECC2004president)[2] <- "precinct"
colnames(BECC2004president)[4] <- "Kerry_count"
colnames(BECC2004president)[5] <- "Kerry_pct"
colnames(BECC2004president)[6] <- "Bush_count"
colnames(BECC2004president)[7] <- "Bush_pct"
colnames(BECC2004president)[8] <- "Badnarik_count"
colnames(BECC2004president)[9] <- "Badnarik_pct"
#Order dataframe by ward then precinct
BECC2004president =BECC2004president[
  order( BECC2004president[,1], BECC2004president[,2] ),
  ]
#
# Verify totals
```

```r
sum(BECC2004president$Bush_count)
sum(BECC2004president$Kerry_count)
#
```

R 7: Differences BECC 2004 vs. Enos (2016) aggregate data

```r

## \cite{Enos2016} replication data: 'data.votechoice.2010.csv' file
setwd('..')
setwd("./Enos_Chicago_Rep")
data.2010 = read.csv('data.votechoice.2010.csv')
# extract ward & precinct number from ward_pre variable
data.2010$ward=gsub(" .*$", "", data.2010$ward_pre)
data.2010$precinct=gsub(".+\\s+", "", data.2010$ward_pre)

## Merge BECC datasets to Enos data by ward & precinct
library(plyr)
setwd('..')
BECCdata=join(BECC2004turnout,BECC2004president, by=c("ward","precinct"))
combined=join(BECCdata, data.2010, by=c("ward","precinct"))

### Differences BECC 2004 vs. \cite{Enos2016} aggregate data
combined$diff.ballots=combined$votes_cast_2004_president-combined$Ballots.
    Cast #RE-BECC ballots
summary(combined$diff.ballots) # min -238, max. 568
combined$diff.Bushv=combined$bush2004_pct-combined$Bush_pct #RE-BECC Bush
    vote
summary(combined$diff.Bushv) # min -22.72%, max. 26.59%
```

R 8: Missing precincts in Enos (2016)

```r
## Table: missing precincts in \cite{Enos2016}
remiss=combined[which(is.na(combined$ward_pre)==1),c(1:2,4,9)]
dim(remiss) #843 missing precincts
# Idenify missing precinct #s by ward & store in matrix
precs=matrix(NA, nr=length(unique(remiss$ward))+1,nc=4)
colnames(precs)=c("Ward","Missing Precincts","Ballots Cast","Bush Votes")
vect=unique(remiss$ward)
for(i in vect){
  nw=which(vect==i)
   precs[nw,1]=i
   precs[nw,2]=paste(remiss[which(remiss$ward==i),2],collapse=",")
   precs[nw,3]=sum(remiss[which(remiss$ward==i),3])
```

```
  precs[nw,4]=sum(remiss[which(remiss$ward==i),4])
}
precs[length(unique(remiss$ward))+1,3]=sum(remiss$Ballots.Cast)
precs[length(unique(remiss$ward))+1,4]=sum(remiss$Bush_count)
# Export matrix to Latex as table
library(xtable)
table=xtable(precs)
print(table, include.rownames=FALSE)
```

R 9: Adding demolition projects as points on maps

```
###Demolition projects coordinates to plot on maps
#p. 1 in Supplemental Information for \cite{Enos2016}
#https://en.wikipedia.org/wiki/Cabrini%E2%80%93Green_Homes
#geo:41.900417,-87.640139
#https://www.topozone.com/illinois/cook-il/city/ida-b-wells-housing-
    project/
#41.824478, -87.6133836
#https://wikivividly.com/wiki/Rockwell_Gardens
#https://tools.wmflabs.org/geohack/geohack.php?pagename=Rockwell_Gardens&
    params=41.876713_N_87.6894_W_type:landmark_region:US-IL
#41.876713, -87.6894
#https://illinois.hometownlocator.com/maps/feature-map,ftc,3,fid,1772628,n
    ,stateway%20gardens.cfm
#41.8297556, -87.6278288
#http://www.lat-long.com/Latitude-Longitude-1764331-illinois-Robert_Taylor
    _Homes__historical_.html
#41.8117003,-87.6272727
#https://www.latlong.net/place/washington-park-chicago-il-usa-19936.html
#41.794552,-87.614731
#https://livingnewdeal.org/projects/julia-c-lathrop-homes-chicago-il/
#41.932234,  -87.683995
#https://tools.wmflabs.org/geohack/geohack.php?pagename=ABLA_Homes&params
    =41_51_58_N_87_39_35_W_
#41.866111, -87.659722
# https://www.travelmath.com/cities/LeClaire+Courts,+IL
# 41.8158657, -87.7439436
#http://www.lat-long.com/Latitude-Longitude-1737787-illinois-Midway_
    Gardens__historical_.html
# 41.7853118 , -87.6064379
#https://www.tripmondo.com/united-states/illinois/cook-county/chicago/
    prairie-courts/
```

```
# 41.850,  -87.650
#http://www.lat-long.com/Latitude-Longitude-1737470-illinois-Dearborn_Homes.html
#41.8430889, -87.6278291
# https://wikivividly.com/wiki/Harold_L._Ickes_Homes
#41.8508, -87.6280
#http://digital.chipublib.org/digital/collection/ChicagoParks/id/4851/
#41.825670, -87.614416

d <- data.frame(
lat=c(41.900417,41.824478,41.876713,41.8297556,41.8117003,41.794552,
                        41.932234,41.866111, 41.8158657, 41.7853118, 41.850,
                             41.8430889,
                        41.8508,41.825670),
lon=c(-87.640139,-87.6133836,-87.6894,-87.6278288,-87.6272727,-87.614731,
                        -87.683995,-87.659722, -87.7439436,-87.6064379,
                            -87.650, -87.6278291,
                        -87.6280,-87.614416))
projects=c("Cabrini-Green", "Ida Wells", "Rockwell Gardens", "Stateway Gardens",
           "Robert Taylor", "Wahsington Park", "Julia Lathrop", "ABLA Homes",
           "Le Claire Courts", "Midway Gardens", "Prairie Courts", "Dearborn Homes",
           "Harold Ickes", "Madden Park")
d$xa=c(0,-0.017,rep(0,8),-0.025,0,0.025,0.025)
d$ya=c(0.005,-0.002,rep(0.005,12))
pointLabels<-annotate("text",x=d$xa+d$lon,y=d$ya+d$lat,size=3,fontface="bold",label=projects)
#####
```

R 10: 2004 Ballots Cast: Enos (2016) replication data vs. official BECC statistics

```
# Maps: necessary libraries
library(rgdal)
library(ggplot2)
library(ggmap)
require(RColorBrewer)
### Create maps
#Download and unzip 'Precincts - 2010.zip' file from data.gov inside 'Additional Data' subfolder
#https://catalog.data.gov/dataset/precincts-2010/resource/18ed6619-7fee-4
```

```
    d5f-bfb4-ff38c09600d8?inner_span=True
setwd("./Additional Data/Precincts - 2010")
# Convert shapefile into dataframe, merging w/ BECC precinct-level data
shape <- readOGR("geo_export_6f371c5e-e7fe-42d0-b1da-a2ca78a90da1.shp")
shape@data$id <- rownames(shape@data)
shape@data   <- join(shape@data, combined)
shape.df     <- fortify(shape)
shape.df     <- join(shape.df,shape@data, by="id")

### Map: BECC data: Ballots cast 2004
png("BECC2004Ballots.png",width=650,height=650)
map=ggplot(data = shape.df) +
  geom_polygon(aes(x = long, y = lat, fill = Ballots.Cast, group = group),
               color = "white", size=.05) +
  coord_quickmap() +
  scale_fill_gradient(low = "aquamarine", high = "springgreen4", na.value="white",
                      breaks=c(100,250,500,750,1000),name="Ballots Cast",
                      labels=c("100","250","500","750","1000"),
                      limits=c(0,1350))
map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+ggtitle("Chicago Board of Election Commissioners 2004 Data")
print(map2)
dev.off()

### Map: \cite{Enos2016} data: Ballots cast 2004
png("REdata2004Ballots.png",width=650,height=650)
map=ggplot(data = shape.df) +
  geom_polygon(aes(x = long, y = lat, fill = votes_cast_2004_president, group = group),
               color = "white", size=.05) +
  coord_quickmap() +
  scale_fill_gradient(low = "aquamarine", high = "springgreen4", na.value="white",
                      breaks=c(100,250,500,750,1000),name="Ballots Cast",
                      labels=c("100","250","500","750","1000"),
                      limits=c(0,1350))
map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+ggtitle("\cite{Enos2016} Replication Data")
print(map2)
dev.off()
```

R 11: 2004 Bush Votes: Enos (2016) replication data vs. official BECC statistics

```
### Map: BECC data: Bush vote 2004
png("BECC2004Bushvote.png",width=650,height=650)
map=ggplot(data = shape.df) +
  geom_polygon(aes(x = long, y = lat, fill = Bush_pct, group = group),
               color = "white", size=.05) +
  coord_quickmap() +
  scale_fill_gradient(low = "darksalmon", high = "firebrick", na.value="
      white",
                      breaks=c(.25,.50,.75),name="% Bush votes",
                      labels=c("25%","50%","75%"),
                      limits=c(0,1))
map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+ggtitle
    ("Chicago Board of Election Commissioners 2004 Data")
print(map2)
dev.off()

### Map: \cite{Enos2016} data: Bush vote 2004
png("REdata2004Bushvote.png",width=650,height=650)
map=ggplot(data = shape.df) +
  geom_polygon(aes(x = long, y = lat, fill = bush2004_pct, group = group),
               color = "white", size=.05) +
  coord_quickmap() +
  scale_fill_gradient(low = "darksalmon", high = "firebrick", na.value="
      white",
                      breaks=c(.25,.50,.75),name="% Bush votes",
                      labels=c("25%","50%","75%"),
                      limits=c(0,1))
map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+ggtitle
    ("\cite{Enos2016} Replication Data")
print(map2)
dev.off()
```

R 12: Differences BECC 2004 vs. Enos (2016) aggregate data

```
### Map: Differences: \cite{Enos2016} data - BECC 2004 Data
png("RE_BECC_DiffBallots.png",width=650,height=650)
map=ggplot(data = shape.df) +
  geom_polygon(aes(x = long, y = lat, fill = diff.ballots, group = group),
               color = "white", size=.05) +
  coord_quickmap() +
  scale_fill_gradient2(mid="gray", low = "firebrick4", high = "darkgreen",
```

```
        na.value="white",
8                           breaks=c(-100,0,100,200,300,400,500),name="
                                Difference",
9                           labels=c("-100","0","100", "200","300","400","500")
                                ,
10                          limits=c(-250,600))
11 map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+
12   ggtitle("2004 Ballots Cast: \cite{Enos2016} - Chicago Board of Election
         Commissioners 2004 Data")
13 print(map2)
14 dev.off()
15
16 ### Map: Differences: \cite{Enos2016} data - BECC 2004 Data
17 png("RE_BECC_DiffBushv.png",width=650,height=650)
18 map=ggplot(data = shape.df) +
19   geom_polygon(aes(x = long, y = lat, fill = diff.Bushv, group = group),
20                color = "white", size=.05) +
21   coord_quickmap() +
22   scale_fill_gradient2(mid="gray", low = "firebrick4", high = "darkgreen",
         na.value="white",
23                          breaks=c(-.25,-.10,0,.10,.25,.30),name="Difference"
                                ,
24                          labels=c("-25%","-10%","0%","10%","25%", "30%"),
25                          limits=c(-.25,.30))
26 map2=map+ geom_point(data=d, aes(x=lon, y=lat),size=3)+pointLabels+
27   ggtitle("2004 Bush Votes: \cite{Enos2016} - Chicago Board of Election
         Commissioners 2004 Data")
28 print(map2)
29 dev.off()
```

Internet Web Archive: Captures of Enos' personal website

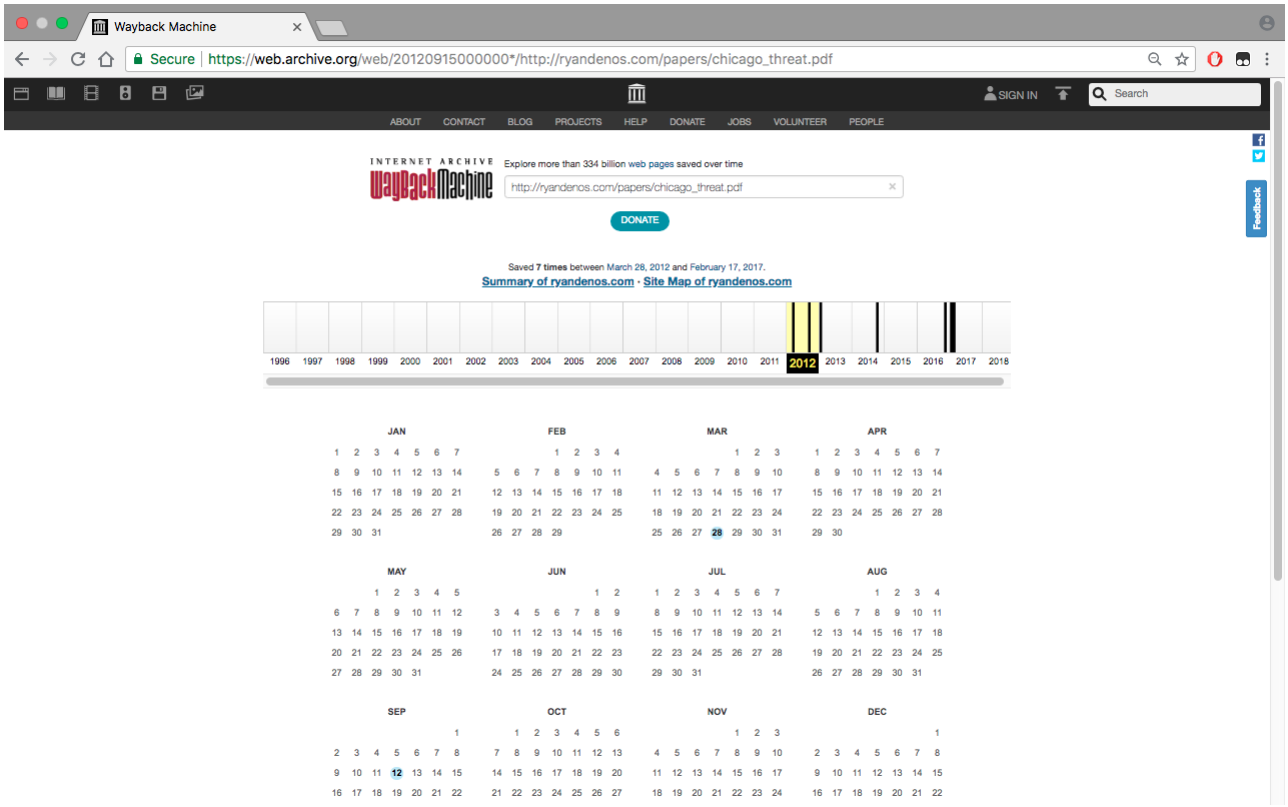

Figure 8: Internet Archive captures of Enos' website.

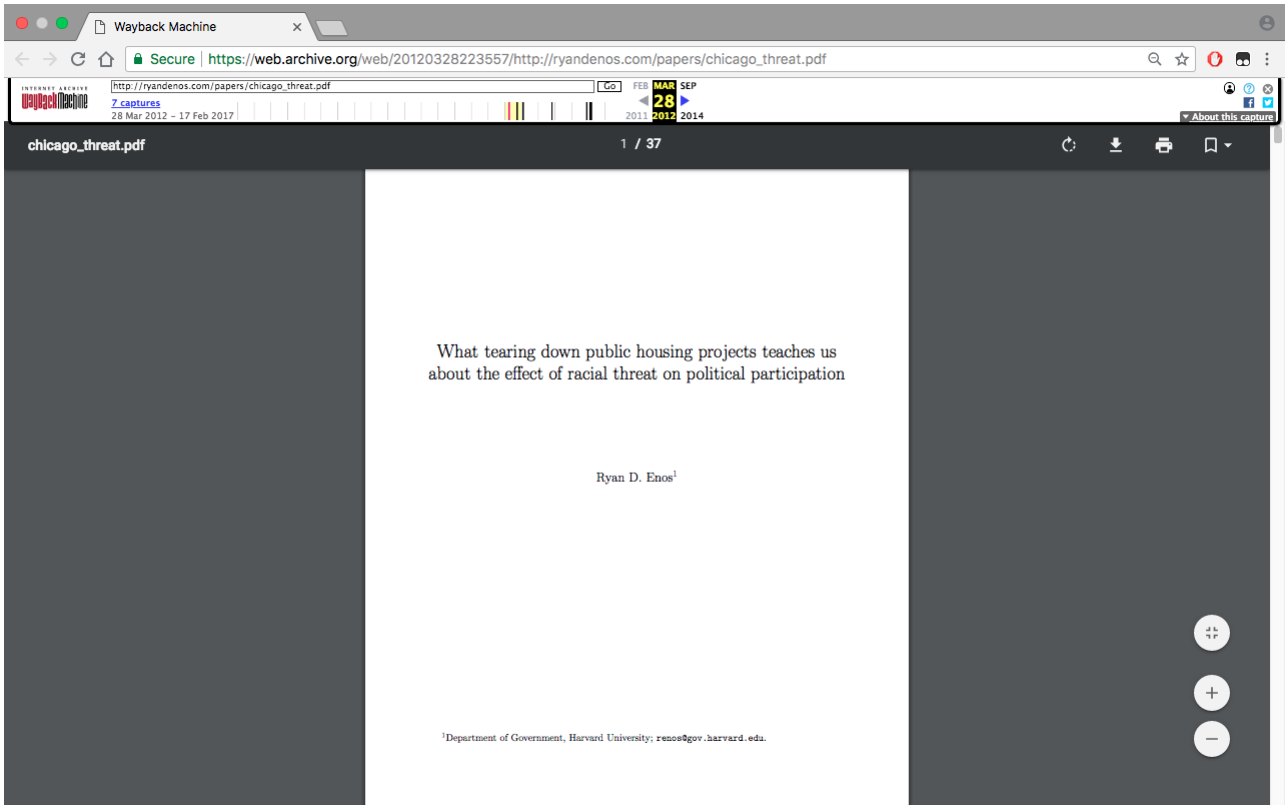

Figure 9: Internet Archive capture of March 28, 2012 version of the paper.