




Causal Inference Examples

Adam Kelleher
Applied Causal Inference
Fall, 2018



Estimating Population Average Causal Effects in the Presence of Non-Overlap: The Effect of Natural Gas Compressor Station Exposure on Cancer Mortality

Rachel C. Nethery¹, Fabrizia Mealli², and Francesca Dominici¹

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston MA, USA

²Department of Statistics, Informatics, Applications, University of Florence, Florence, Italy



Context

1.1 Natural Gas Compressor Stations and Cancer Mortality

During the last several decades, the United States (US) has witnessed a sharp increase in the incidence of thyroid cancer, which now accounts for 1-1.5% of all newly diagnosed cancer cases (Pellegriti et al., 2013). Increased exposure of the population to radiation and carcinogenic environmental pollutants is blamed, in part, for this increase.



Context

- Natural gas (NG) production has increased over decades
- Most studies focus on exposure to drilling sites.
- This study is about exposure to distribution systems.
- Compressor stations along pipelines can emit chemicals
 - Unintended leaks
 - Intentional “Blowdowns” to reduce pressure
- Pennsylvania and Texas have detected harmful emissions in excess of EPA standards near compressor stations, including benzene (a known carcinogen).



Fine Articulation

health effects of NG distribution systems. Specifically, we aim to provide the first data-driven epidemiological investigation of the causal effects of proximity to NG compressor stations on thyroid cancer and leukemia mortality rates.

communities (Southwest Pennsylvania Environmental Health Project, 2015). In this paper, we exclude from consideration the health impacts of accidents at compressor stations and focus on the potentially harmful exposures to nearby communities resulting from the normal operations of compressor stations. Fugitive emissions, or unintended leaking of chemicals from the compressor



Data Set

- 978 counties from mid-western US
- Data contains county-level ...
 - compressor station exposure,
 - thyroid cancer and leukemia mortality rates, and
 - suspected confounders



Examining Identification Assumptions

- Overlapping support of covariates between exposed and unexposed populations?

relationship. While we would like to apply a classic non-parametric causal inference analysis rooted in the potential outcomes approach (Rubin, 1974), the data exhibit non-overlap, i.e., in some areas of the confounder space, there is little or no variability in the exposure status of the units. Due to this non-overlap, any attempt to adjust for confounding when estimating the population average causal effect must rely upon model-based extrapolation, because we have insufficient data to infer about missing potential outcomes in those regions of the confounder space. Thus non-parametric causal inference methods may yield unreliable results.

Propensity Score Distribution Overlap

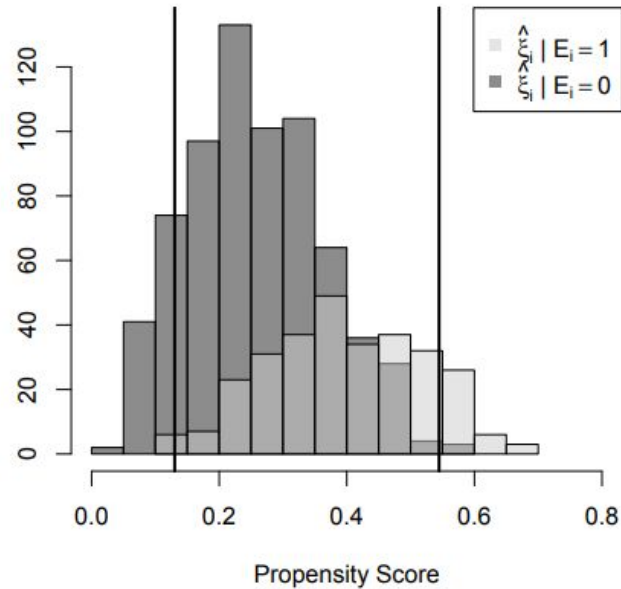


Figure 1: Estimated propensity score histograms stratified by exposure status and overlaid. Bold vertical lines represent the start of non-overlap intervals in both tails of the distribution.



Solution: Novel Contributions!

- “We introduce a flexible, data-driven definition of sample propensity score overlap and non-overlap regions”
- “We propose a novel approach to estimating population average causal effects in the presence of non-overlap
 - the sample is split into a region of overlap (RO) and a region of non-overlap (RN)
 - distinct models, appropriate for the amount of data support in each region, are developed and applied to estimate the causal effects in the two regions separately”
- A conditioning approach built on bayesian additive regression trees (BART) in the overlap region
 - <https://arxiv.org/abs/0806.3286>
- Extrapolates into the non-overlap region with a bayesian splines method

Results: Simulations

Table 1: Absolute (Abs) bias, 95% credible interval coverage and mean square error (MSE) in estimation of the population average causal effects in simulations from Section 3.1.

Simulation Setting	Method	Abs Bias	Abs Bias (%)	Coverage	MSE
3.1A-i	U-GR	0.12	46.46	0.33	1.25
	U-BART	0.03	10.72	0.99	0.07
	BART+SPL	0.01	5.63	1.00	0.05
3.1A-ii	U-GR	0.17	97.31	0.23	1.69
	U-BART	0.05	31.95	0.89	0.12
	BART+SPL	0.02	12.58	1.00	0.09
3.1A-iii	U-GR	0.23	724.78	0.19	2.34
	U-BART	0.08	427.55	0.71	0.22
	BART+SPL	0.03	100.80	1.00	0.14
3.1B-i	U-GR	0.26	50.42	0.00	0.64
	U-BART	0.13	25.47	0.12	0.33
	BART+SPL	0.11	21.64	0.62	0.27
3.1B-ii	U-GR	0.32	66.58	0.00	0.73
	U-BART	0.18	36.73	0.04	0.48
	BART+SPL	0.15	30.95	0.55	0.36
3.1B-iii	U-GR	0.41	94.79	0.00	0.89
	U-BART	0.24	55.00	0.01	0.68
	BART+SPL	0.21	48.03	0.35	0.49



Results: Simulations (a few results)

- There's lower bias in the new BART+SPL method, compared with the BART method.
- The confidence interval coverage for the new method is conservative

Table 1: Absolute (Abs) bias, 95% credible interval coverage and mean square error (MSE) in estimation of the population average causal effects in simulations from Section 3.1.

Simulation Setting	Method	Abs Bias	Abs Bias (%)	Coverage	MSE
3.1A-i	U-GR	0.12	46.46	0.33	1.25
	U-BART	0.03	10.72	0.99	0.07
	BART+SPL	0.01	5.63	1.00	0.05
3.1A-ii	U-GR	0.17	97.31	0.23	1.69
	U-BART	0.05	31.95	0.89	0.12
	BART+SPL	0.02	12.58	1.00	0.09
3.1A-iii	U-GR	0.23	724.78	0.19	2.34
	U-BART	0.08	427.55	0.71	0.22
	BART+SPL	0.03	100.80	1.00	0.14
3.1B-i	U-GR	0.26	50.42	0.00	0.64
	U-BART	0.13	25.47	0.12	0.33
	BART+SPL	0.11	21.64	0.62	0.27
3.1B-ii	U-GR	0.32	66.58	0.00	0.73
	U-BART	0.18	36.73	0.04	0.48
	BART+SPL	0.15	30.95	0.55	0.36
3.1B-iii	U-GR	0.41	94.79	0.00	0.89
	U-BART	0.24	55.00	0.01	0.68
	BART+SPL	0.21	48.03	0.35	0.49



Results: Real Data

Table 5: Average causal effects of natural gas compressor station presence on 2014 county-level thyroid cancer and leukemia mortality rates and the change in thyroid cancer and leukemia mortality rates from 1980 to 2014.

Outcome	Method	Effect	95% CI
2014 Thyroid Rates	BART+SPL	0.001	-0.017, 0.020
	BART	0.003	-0.007, 0.012
Change in Thyroid Rates 1980-2014	BART+SPL	0.992	-0.308, 2.237
	BART	1.089	0.130, 2.038
2014 Leukemia Rates	BART+SPL	0.006	-0.013, 0.025
	BART	0.005	-0.004, 0.014
Change in Leukemia Rates 1980-2014	BART+SPL	0.913	-0.361, 2.206
	BART	0.988	0.014, 1.958




Results: Real Data

- The new method gives very conservative confidence intervals, which can mean the difference between significant and insignificant effects

Table 5: Average causal effects of natural gas compressor station presence on 2014 county-level thyroid cancer and leukemia mortality rates and the change in thyroid cancer and leukemia mortality rates from 1980 to 2014.

Outcome	Method	Effect	95% CI
2014 Thyroid Rates	BART+SPL	0.001	-0.017, 0.020
	BART	0.003	-0.007, 0.012
Change in Thyroid Rates 1980-2014	BART+SPL	0.992	-0.308, 2.237
	BART	1.089	0.130, 2.038
2014 Leukemia Rates	BART+SPL	0.006	-0.013, 0.025
	BART	0.005	-0.004, 0.014
Change in Leukemia Rates 1980-2014	BART+SPL	0.913	-0.361, 2.206
	BART	0.988	0.014, 1.958



Are Donation Badges Appealing? A Case Study of Developer Responses to Eclipse Bug Reports

Keitaro Nakasai*, Hideaki Hata*, Kenichi Matsumoto*

*Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{nakasai.keitaro.nc8, hata, matumoto}@is.naist.jp

Approach: Difference in Differences

- Panel data for donors before and after the donation badge program was introduced
- Extrapolate an appropriate control group to estimate a counterfactual
- **Problem:** Don't want to assume no confounding (a.k.a. ignorability)!
- **Solution:** Choose the control group using propensity score matching!

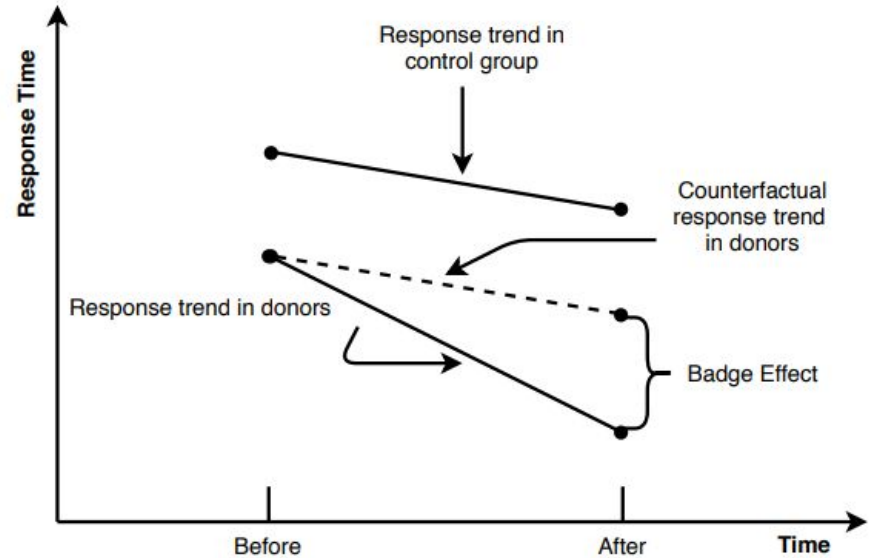


Fig. 1. Example of causal inference framework using a DID model (response time vs. before and after donation badge introduction)



Context

- Few studies have been undertaken on [the effect of] monetary donations to OSS projects
- Donors who contribute 35 USD receive a donation badge
- They'd like to know the benefit to the donor of having a donation badge



Fine Articulation

Based on the framework for causal inference [5], we study how promptly developers respond to bug reports that have donation badges compared with bug reports without donation badges. The analysis revealed that the donation badges decreased response time by about two hours in median. Our findings



Data

Step 1: Bug report collection. From the discussion of proposing donation badges [4], we speculate that November 13, 2014 can be identified as the date of initially implementing donation badges. So we first identify reporters who had



Data

- First, ID reporters who report at least once before and after implementation (It's a CATE!)
- Collect reports from the period two years before and two years after implementation
- Remove reports where...
 - First responses were by the original reporter, or
 - The report was assigned to the original reporter
 - Responses don't come within three days (removes forgotten or postponed reports)
 - (fine articulation, or conditioning?!)
- 60% of reports remain after removal



Data

- Collect reporter metrics
 - Number of months worked at Bugzilla
 - Number of bug report submissions before and after implementation
 - Number of commits worked in git repos
- Collect report metrics
 - Severity level (more severe are fixed faster)
 - OS variables (some OSs are prioritized)
 - Component of application
 - Community size of a component
 - Time (month)
 - Relationship between reporter and responder
- Select a subset of metrics based on the AIC
- Do these satisfy the BDC?



Data

- Donor name, date, and donation amount are collected about donations
- These are used to identify badged donors
- When the same name appears with different emails, they're removed from the donors, since their causal state can't be measured (more conditioning!)
- 31 donors were identified



Checking Assumptions

- No assessment of support overlap given!



Propensity Score Matching

- Control state (unbadged reporters) has 957 reporters, but test state (badged reporters) has only 31
- Control units were matched to the treatment units, so 31 control state users were selected
- Between these two sets, there were 1,822 reports
- **Issues:**
 - We're estimating an ATT, not the ATE!
 - It might be better to increase k for kNN! We're dropping almost the entire control group.

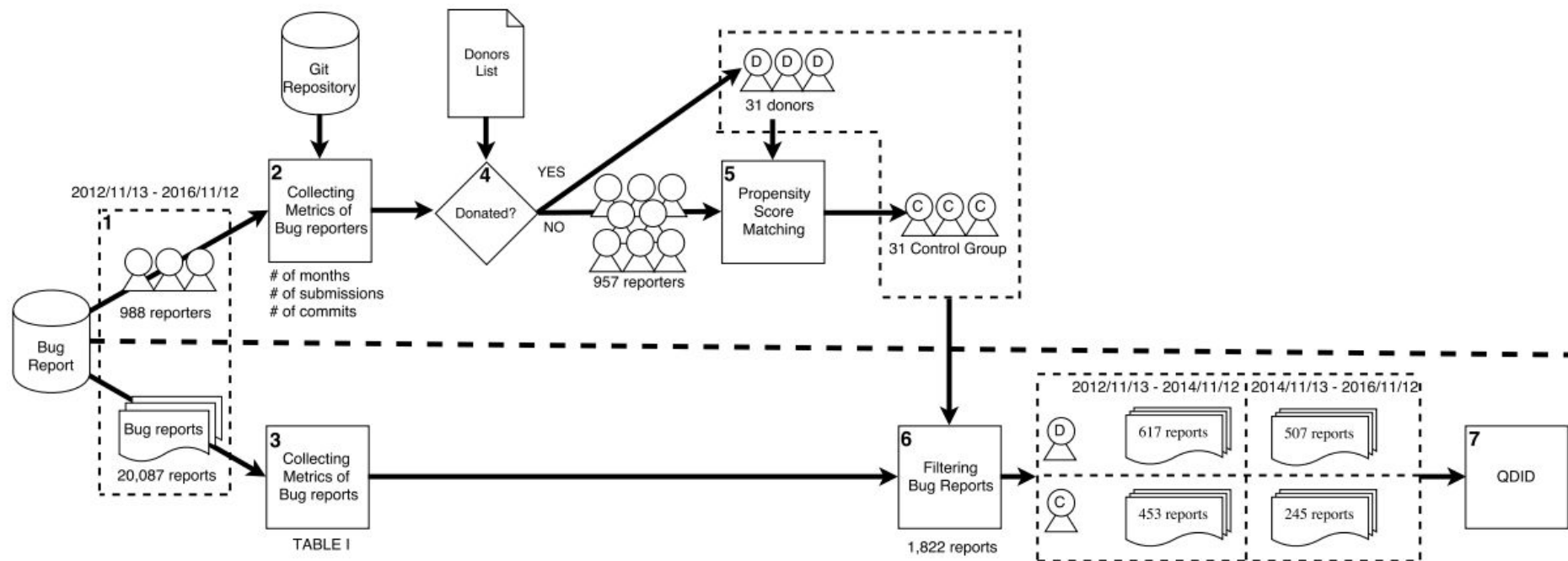


Fig. 2. An overview of the analysis for data collection and causal inference.

Regression Results

$$y = \beta_0 + \beta_1 T + \beta_2 S + \beta_3 (T \cdot S) + \varepsilon$$

Metric	Description	Coeffs (Errors)	t value	Pr (> t)
(Intercept)		2.237 (1.120)	1.998	0.046
Donor	Reporter is a donor or not	2.259 (0.834)	2.708	0.007
Period	Submitted time is before or after the badge introduction	-1.598 (1.227)	-1.302	0.193
Badge	It has a donation badge or not	-2.219 (1.061)	-2.092	0.037
Enhancement	Severity is enhancement or not	0.668 (0.815)	0.820	0.412
Windows	Issue is related to Windows or not	1.119 (0.692)	1.617	0.106
Linux	Issue is related to Linux or not	0.675 (0.949)	0.711	0.477
MacOS	Issue is related to MacOS or not	0.793 (0.907)	0.875	0.382
Component	Response days in median for the belonging components	0.317 (0.198)	1.603	0.109
Community	# of contributors in the belonging components	0.000 (0.001)	0.105	0.917
Time	A numerical order of time in months	0.152 (0.117)	1.303	0.193
Relationship	# of reports in which the reporter and the first responder have worked together	0.005 (0.005)	0.894	0.371

AIC = 4025, Pseudo R^2 = 0.011

It's unclear whether they omitted the T S cross term!



Result

- An estimated -2.21 effect at $p=0.037$ for the effect of badges on response time.
- It looks like bug reports get a roughly 2 hour faster response time when the reporter has a donor badge
- The researchers suspect “signalling” is what is happening: donor reports are prioritized
- What is the CATE we have now?
 - The ATT for reporters who report frequently enough to be included, don't respond quickly to their own reports, report important enough issues to warrant response, and who less likely work at Bugzilla



Data-adaptive doubly robust instrumental variable methods for treatment effect heterogeneity

K. DiazOrdaz, R. Daniel, N. Kreif

Department of Medical Statistics, LSHTM,
Division of Population Medicine, Cardiff University,
Centre for Health Economics, University of York.



Context

- Instrumental variables works with randomized controlled trials (RCTs) with non-adherence
- There was a trial done to evaluate pain-management techniques, but patients didn't always adhere to the treatment



Fine Articulation

- Causal States:
 - Intervention participants were offered 24 sessions introducing them to cognitive behavioural (CB) approaches designed to promote self-management of chronic back pain.
 - The sessions were delivered over three days within the same week with a follow-up session 2 weeks later.
 - At the end of the 3-day course participants received a relaxation CD and self-help booklet.
 - Controls received usual care and the same relaxation CD and self-help booklet.
- Outcome:
 - The primary outcome was pain-related disability at 12 months, using the Chronic Pain Grade (CPG) disability sub-scale.
 - This is a continuous measure on a scale from 0 to 100, with higher scores indicating worse pain-related disability



Data

- The RCT was originally across 27 general practices and community services in the UK
- It consisted of 703 adults with musculoskeletal pain of at least 3 months duration (we're estimating a CATE!)
- 403 patients were randomized into the treatment, and 300 into the control
- Average age of participants was 59.9 years, 81% white, 67% female, 23% employed, 85% with pain for at least 3 years, and 23% on strong opioids (very biased population!)
- This study is a re-analysis, dropping some units with missing data
 - data set consists of 652 participants followed up for 12 months, 374 allocated to active treatment, and 278 in the control (93% of those recruited)
 - Thirty-five individuals (5%) have missing primary outcome data, and a further 4 (<1%) have missing baseline depression score, leaving a sample size of 613



Non-adherence

- In the active treatment, only 179 (45%) attended all 24 sessions, and 322 (86.1%) received at least one session.
- The control arm participants had no access to the active intervention sessions.
- Poor attendance to the sessions was anticipated, and so obtaining causal treatment effect estimates was a pre-defined objective of the study



Identifying Assumptions

- Relevance and unconfoundedness:
 - “We argue that random allocation is a valid IV: the assumptions concerning unconfoundedness and instrument relevance are justified by design”
- Exclusion Restriction:
 - “The exclusion restriction assumption seems plausible with our choices for A, as only those participants receiving at least one training sessions would know how to use the CB coping mechanisms and potentially to improve their disability”
 - “It is unlikely that that random allocation has a direct effect, though since participants were not blinded to their allocation, we cannot completely rule out some psychological effects of knowing one belongs to the control or active group on pain and disability”

Results

Table 1: ATT of the COPERS intervention on CPG, with all-or-nothing binary exposure A , main effect ψ_c and effect modification by depression) ψ_v .


	ψ_c	SE	ψ_v	SE
TSLS	2.94	4.67	-0.58	0.57
IV-g	2.78	4.66	-0.53	0.54
IV-g SL	2.10	4.75	-0.45	0.54
IV-TMLE	3.16	4.74	-0.64	0.56
IV-TMLE SL	2.22	4.88	-0.51	0.58

We perform each of the methods in turn, TSLS, IV-g and IV-TMLE to estimate $ATT(v)$. As Table 1 summarises, the use of DR methods, even after using Super Learner does not result in a material change in the point estimates or SEs, compared to standard TSLS. All 5 estimators give the same inference, namely, there is no evidence of an average treatment effect in the treated, and there is no evidence of effect modification by baseline depression. The lack of statistical significance may

Results



- In this case, using a new and improved method (doubly robust, efficient) didn't substantially change the effect estimate or the error bars compared with two-stage least squares



Ethnic Diversity Increases Scientific Impact

Bedoor K AlShebli^{*}, Talal Rahwan^{*}, Wei Lee Woon^{*}

Department of Computer Science, Masdar Institute, Khalifa University of Science and Technology,
Abu Dhabi, United Arab Emirates

Email: {bedoor.alshebli, talal.rahwan, wei.woon}@ku.ac.ae



Context

- Lots of studies (see citations) have shown significant benefits of diversity
 - Economic vibrance
 - Innovation
 - health
- They study how diversity relates with scientific impact, where “diversity” is each of
 - ethnicity
 - Discipline
 - Gender
 - Affiliation
 - Academic age



Fine Specification

- Causal States:
 - For each diversity type, we distinguish between group diversity, where the unit of analysis is the paper's set of authors, and individual diversity, where the unit of analysis is a scientist's entire set of collaborators. In both cases, we use Gini Impurity to quantify diversity
- Outcomes:
 - As a proxy for scientific impact, we consider the number of citations received within five years of publication
 - Average is compared with other papers in that subfield.



Naive Result

- “Remarkably, from both the group and individual perspectives, we find that a subfield’s ethnic diversity is the most strongly correlated with impact (r is 0.77 and 0.55 for [group and individual ethnic diversity], respectively)”



Data

- Microsoft Academic Graph dataset
 - 1,045,401 multi-authored papers
 - 1,529,279 scientists
 - 8 main fields and 24 subfields of science
- for each such scientist with at least 10 collaborators, we analyze his/her entire set of collaborators
 - 5,103,877 collaborators over 9,472,439 papers
- Five aspects of collaborations (mentioned before) define diversity and are provided.



Identification Assumptions

- Approach: exact matching
- Group Diversity confounders:
 - year of publication
 - number of authors
 - field of study
 - affiliation ranking
 - authors' impact prior to publication. T
- individual diversity confounders
 - academic age;
 - number of collaborators
 - Discipline
 - affiliation ranking



Procedure:

- Compute percentiles of diversity to define treatment and control groups
- Do exact matching, and compute differences in impact between the groups.

the effect of these factors on the phenomena under investigation. In our case, when studying *group* ethnic diversity, the treatment set consists of papers for which $d_{eth}^G > P_{100-i}(d_{eth}^G)$, and the control set of papers for which $d_{eth}^G \leq P_i(d_{eth}^G)$, where $P_i(d_{eth}^G)$ denotes the i^{th} percentile of d_{eth}^G . This process is repeated using $i = 10, 20, 30, 40, 50$, corresponding to progressively larger gaps in ethnic diversity between the two populations. Thus, if ethnic diversity does indeed increase scientific impact, we would expect to find a significant difference in impact between the two populations, and that it increases in tandem with the

Table 1: Results of coarsened exact matching on group ethnic diversity. T and C are the treatment and control populations respectively; T' and C' are the populations of matched treatment and matched control papers respectively; \mathcal{L}_1 is the multivariate imbalance statistic [7]; δ is the relative impact gain of T' over C' , i.e., $\delta = 100 \times (\langle c_5^G \rangle_{T'} - \langle c_5^G \rangle_{C'}) / \langle c_5^G \rangle_{C'}$. A t-test shows that δ is statistically significant; see the resulting p -values. For more details, see Section S3.

	$ T $	$ C $	$ T' $	$ C' $	\mathcal{L}_1	δ	p
$T : d_{eth}^G > P_{90}(d_{eth}^G)$ $C : d_{eth}^G \leq P_{10}(d_{eth}^G)$	45,710	17,802	16,477	16,322	0.37	11.64	0.001
$T : d_{eth}^G > P_{80}(d_{eth}^G)$ $C : d_{eth}^G \leq P_{20}(d_{eth}^G)$	45,710	24,827	16,546	22,855	0.37	12.97	6.85e-06
$T : d_{eth}^G > P_{70}(d_{eth}^G)$ $C : d_{eth}^G \leq P_{30}(d_{eth}^G)$	58,889	56,662	39,934	55,250	0.22	7.43	6.14e-05
$T : d_{eth}^G > P_{60}(d_{eth}^G)$ $C : d_{eth}^G \leq P_{40}(d_{eth}^G)$	78,340	63,129	59,370	61,834	0.27	7.42	1.28e-05
$T : d_{eth}^G > P_{50}(d_{eth}^G)$ $C : d_{eth}^G \leq P_{50}(d_{eth}^G)$	127,629	63,129	72,376	62,121	0.25	5.03	0.003

Table 2: **Results of coarsened exact matching on individual ethnic diversity.** The notation is as per Table 1.


	$ T $	$ C $	$ T' $	$ C' $	\mathcal{L}_1	δ	p
$T : d_{eth}^I > P_{90}(d_{eth}^I)$ $C : d_{eth}^I \leq P_{10}(d_{eth}^I)$	139,822	84,270	31,500	48,801	0.61	55.46	7.38e-18
$T : d_{eth}^I > P_{80}(d_{eth}^I)$ $C : d_{eth}^I \leq P_{20}(d_{eth}^I)$	168,575	168,475	67,686	152,379	0.40	43.26	8.87e-145
$T : d_{eth}^I > P_{70}(d_{eth}^I)$ $C : d_{eth}^I \leq P_{30}(d_{eth}^I)$	252,801	251,423	174,457	237,525	0.36	30.51	1.11e-165
$T : d_{eth}^I > P_{60}(d_{eth}^I)$ $C : d_{eth}^I \leq P_{40}(d_{eth}^I)$	346,137	336,570	231,951	320,815	0.34	26.80	3.70e-224
$T : d_{eth}^I > P_{50}(d_{eth}^I)$ $C : d_{eth}^I \leq P_{50}(d_{eth}^I)$	437,600	404,782	322,820	391,162	0.27	18.63	2.39e-194



Exercise

- 1) Simulate some panel data. We'll do some analysis similar to the Eclipse Bug Report paper. Specifically,
 - a) Simulate N units over T discrete time steps.
 - b) Add a time-varying causal state, where a binary treatment turns on and stays on for each test unit at a different time, but never turns on for control units
 - c) Add confounding between treated and untreated units in any way you like!
- 2) Create a propensity score model for D^* (as we defined it in class)
- 3) Match test and control units.
- 4) Center each test time series defining the start time $t=0$ as the time treatment turns on, and the same time is the start time for the matched control unit. $t < 0$ before the start, $t > 0$ after.
- 5) Aggregate the test time series and control time series to find the time series for average outcomes in each group
- 6) Take the difference between these time series to estimate the causal effect time series.





Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs*

Matias D. Cattaneo[†] Luke Keele[‡] Rocío Titiunik[§]
Gonzalo Vazquez-Bare[¶]

August 15, 2018



INFLUENCE ESTIMATION ON SOCIAL MEDIA NETWORKS USING CAUSAL INFERENCE

Steven T. Smith, Edward K. Kao, Danelle C. Shah, Olga Simek, and*

*Donald B. Rubin**

MIT Lincoln Laboratory
Lexington MA 02421 USA

{ stsmith, edward.kao, danelle.shah, osimek } @ll.mit.edu

Harvard University
Cambridge MA 02138 USA
rubin@stat.harvard.edu



Distinguishing between Personal Preferences and Social Influence in Online Activity Feeds

Amit Sharma

Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
asharma@cs.cornell.edu


Dan Cosley

Information Science
Cornell University
Ithaca, NY 14853 USA
drc44@cornell.edu



DISCOVERING EFFECT MODIFICATION AND RANDOMIZATION INFERENCE IN AIR POLLUTION STUDIES

KWONSANG LEE¹, DYLAN S. SMALL², AND FRANCESCA DOMINICI¹



Bayesian Propensity Scores for High-Dimensional Causal Inference: A Comparison of Drug-Eluting to Bare-Metal Coronary Stents

Jacob V Spertus¹ and Sharon-Lise T Normand^{1,2}

1: Department of Health Care Policy, Harvard Medical School

2: Department of Biostatistics, Harvard TH Chan School of Public Health