Testing the Robustness of the Propensity Score for Inferring Causal Effect of the Built Neighborhood on Public Health

ABSTRACT

The inference of causality in the context of public health and the built environment can help inform public policy. Often, correlations are studied, but causal effects are rarely sought. The approach presented in this study proposes a method to infer causal relationships from open data using propensity score matching to simulate randomized controlled trials. Datasets with varying distributions and relationships between features are synthesized to assess the robustness of propensity score matching to different relationships one might find in real-life data. The results are inconclusive, and the numerical values of the inferred causal effects suggest that the method lacks robustness to varying feature values in the data.

CCS CONCEPTS

• **Information systems** → **Spatial-temporal systems**; *Data mining*; • **General and reference** → *Validation.*

KEYWORDS

Propensity matching, maximal graph matching, causal inference, public health, built neighborhood

1 INTRODUCTION

Understanding the effect of the environment on public health can help inform public policy. The natural environment cannot always be influenced. However, it is often possible to exert influence or impose regulations on the built environment. Knowledge about how the presence or absence of certain parts of the built environments, e.g. some type of facility, affects aspects of public health can therefore be extremely important. Yet, the complexity and multidimensionality of problems makes this task very difficult. Reality is often modeled in search of correlations between important factors. The issue with analyzing correlation between built environment factors and public health is that if a correlation is found, there is no guarantee that adjusting the built environment will have the desired effect on public health. Finding causal relationships, however, would serve as a guarantee that the proposed measure will affect the public health in the intended manner.

Finding causal relationships is not straightforward. It often requires making assumptions about the available data. This approach is rarely taken; most studies on the effect of the environment on public health focus on finding correlations. [8] proposes a method for inferring a causal effect using propensity score matching to simulate randomized controlled trials using data of London wards. The method is validated using synthetic data which was generated by randomly sampling distributions mimicking the original data. It is assumed that confounding variables are not correlated and do not influence the built environment or public health variables, from now on called treatment and effect, respectively. This is required for propensity score calculation, as it assumes there are no correlations in the confounding variables. The final results were not conclusive, and the assumptions made in the validations raise questions about how robust the method is to various correlations and relationships encountered in real data.

Individual relationships that could occur in real data are correlations between confounding variables, correlation between a confounding variable and treatment, correlation between a confounding variable and effect. However, these are still simple relationships. One would expect there to be spatial relationships between the different wards as well. For example, the real data from the study [8] shows that older citizens tend to live further out in the suburbs, while younger citizens tend to live more in the city center.

This paper proposes to validate propensity score matching methods more extensively by generating increasingly complex synthetic data and testing the methods on these different datasets. The aim is to produce data that is as realistic as possible and to asses how robust the propensity score matching method is to different relationships found in real-life data that would violate the assumptions necessary for calculating the propensity score. Finally, the study concentrates on producing guidelines for working with propensity score matching for causal inference on the topic of the effect of the built environment on public health.

2 RELATED WORK

Many publications avoid causal language when describing observed relationships, because of the age-old phrase that ’correlation does not prove causation’, instead preferring to hint at causality, thereby avoiding the *faux pas* of contradicting the famous phrase. Demonstrating causality has been approached from different angles, and can be described through three measures: covariation (correlation between cause and effect); temporal precedence (the cause preceding the effect); and confounder control [10][11] . One common issue with inferring a causal relationship is the specific confounder data available, which often comprises of incomplete or biased observations, this can be adjusted for to some extent [1], but the existence of noise into the relationship model is often inevitable. This paper draws on the methods of the research paper [8], and continues work done which looks specifically at covariation and confounder control. Developing methods for identifying causal relationships requires data in which causal relationships exist, many papers look at likely causal relationships such as the effect of a neighbourhood built environment on health [8], which are the subject of many national infrastructure projects [3][12] . To deal with the often unknown confounder relationships methods including propensity score matching have been developed [5][6] , which allow for some mitigation of distortion of results due to unknown relationships. Another such mitigation is to synthesize realistic spatial data with assured levels of causality, which can then be used as ground truths [4]. This allows for complete control of the confounder relationships, and thereby quantification of performance of any causality

Table 1: Confounder ranges

Set of considered confounders & Value range	
Population aged $0 - 15$	$[19.760, 19.611]$
Population aged $16 - 64$	$[69.182, 68.991]$
Population aged 65+	$[11.058, 11.398]$
Greenspace area	$[26, 448, 27.272]$
Homes with deficient access to nature	$[25.727, 25.751]$
Employment & support allowance claims	$[1.026, 1.034]$
Housing benefit claims	$[12.863, 12.871]$
Income support claims	$[3.374, 3.393]$
Incapacity benefit claims	$[2.807, 2.882]$
DWP working age clients	$[14.307, 14.363]$
Part-time employees	$[13.229, 14.688]$
Full-time employees	$[34.089, 39.827]$
Job seekers allowance claims	$[5.569, 5.623]$

identifying method such as the propensity score matching used in this paper.

2.1 Future work

3 METHODS

In this section the methods and the principles behind them will be discussed before the specific experiments will be considered.

3.1 Data structure

The synthetic data will be structured based on ward data from the City of London [2]. This dataset contains shapefile data, as well as ward names and other information. For each experiment, a new dataset will be created using these shapefiles. Confounders will be added with values generated by different principles for each ward. The two variables that will be used to mimic a causal effect are the number of sports venues in a ward as a proxy for the built environment, from now on called the treatment, and the normalized number of prescriptions of antidepressants per ward as a proxy for public health, from now on called the effect. The treatment will either be assigned randomly or based on some other principle, depending on the experiment. The effect will be calculated as a function of the treatment and a factor that describes the strength of the causal effect. The aim is to find the strength of the effect with the proposed methods.

3.2 Propensity score

Ideally, a randomized controlled trial would be performed to assess the relationship between the treatment and effect. This would involve dividing the wards into two groups, removing the treatment from one group and apply a constant value to the second group. Needless to say, that is not possible or humane in a real city. Therefore, a different procedure is needed to calculate the average treatment effect in a meaningful way.

A way to do this, is to calculate the propensity score of each ward. The propensity score is a function of the confounding values and models the probability of belonging to a certain group. In this scenario, the groups are assigned by binning the treatment levels into integer values. Propensity score matching can be applied to different cases. The first case is binary matching, which effectively splits the instances into two groups with 0 or 1 treatment. This is not realistic in the case of sports venues and would not yield a realistic estimate of the treatment effect. Another case is multi-level propensity score matching, in which there are different levels of treatment that can be assigned. This applies to the current scenario. However, there is another issue one must keep in mind: there may be different levels of treatment, but there is no knowledge about the difference in effect two treatment levels might have. For example, there is no guarantee that a treatment level of 4 has twice as much effect as a treatment level of 2. Assumed is that a higher treatment level yields a stronger effect, thus an ordinal model must be used. Using an ordinal model allows for not making assumption about the differences between the different treatment levels, except that they are ordered.

The propensity scores, together with the treatment values, will be used to match the wards in pairs that have confounder vectors that are as similar as possible, but with treatment levels that are as much apart as possible. This will be done using a maximal graph matching algorithm. For that purpose, a graph must be constructed.

3.3 Graph matching

The propensity scores of the wards will be used to calculate the weights for a fully-connected, weighted graph with wards at the nodes. The distance between two wards i and j is taken from [9]:

$$d_(i, j) = \frac{|\hat{\beta}^T_{\mathbf{x}_i} - \hat{\beta}^T_{\mathbf{x}_j}| + \epsilon}{|z_i - z_j|} \tag{1}$$

Where $\hat{\beta}^T_{\mathbf{x}_i}$ is the propensity score of ward i, x_i is the vector with confounders of ward i, z_i the treatment value of ward i and ϵ is a small number to deal with division by zero.

Now, the wards are matched using Edmond's Blossom algorithm for maximum graph matching [7]. Maximum graph matching maximizes the total distance between matched pairs. Finally, the average treatment effect (ATE) can be calculated as follows, as done in [8]:

$$ATE = \sum_{(i,j) \in M} \frac{y_i - y_j}{z_i - z_j} \tag{2}$$

With M the set of matched wards, y the outcome, z the treatment value. The obtained ATE can then be compared to the implemented strength of effect in the synthetic data to assess the accuracy of the method.

4 EXPERIMENTS

The model, as described above, will be tested on different synthetic datasets. Each of the four different types of synthetic data with different distributions implemented will be generated with 21 levels of ATE, in $[-10, 10]$. These four categories will be described in the following subsections.

4.1 Random uniform

The first experiment uses synthetic data which is generated by sampling the different confounder values uniformly from the range they occur within the real data. This keeps the confounder value ranges within realistic ranges, to prevent anomalous performance results due to very high or very low value ranges.

4.2 Distributions modelled on data

The synthetic data for this experiment is generated by sampling the values of each confounder from a distribution that is modelled on the normalised distribution of the confounder in the real data. This emulates realistic confounder value distributions, making the performance results more comparable to the real data. The treatment level and outcome are also modelled on the real data. The treatment is approximated by an exponential distribution, based on [8], with the difference being the sign of the term in the exponent:

$$f_z(\mathbf{x}) = P(Z = z|\mathbf{x}_i) = \begin{cases} \frac{1}{f(\mathbf{x}_i)} \cdot e^{\frac{z}{f(\mathbf{x}_i)}} & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{3}$$

where

$$f(\mathbf{x}_i) = \frac{\sum_p x_{ip}}{\alpha} \tag{4}$$

In these equations $\mathbf{x}_i$ represents the confounders of ward i, x_{ip} represents the p^{th} confounder of the same ward and α is a parameter which can be tuned, and that determines the shape of the distribution. The parameter has been tuned as to approach the real distribution as closely as possible. This has been done by eye, since the original data was not available for this study. Note that each ward has its own distribution, from which the treatment level is sampled.

The outcome was modelled using a truncated normal distribution, as introduced in [8]:

$$Y \sim \text{Normal}(\alpha \cdot \sum_p x_{ip}, \beta \cdot \sum_p x_{ip}) + \gamma z_i \tag{5}$$

Where α, β are tunable parameters that determine the shape of the distribution. γ is a factor indicating the strength of the treatment effect. This parameter allows for implementing different causal effects between the treatment and outcome.

The distributions the confounder values are sampled from are shown in Figure 1. An example of what the treatment and outcome distributions look like is shown in Figure 2. Note that these distributions will vary from dataset to dataset based on the confounder values of each ward in a given dataset.

4.3 Spatial correlation: distance to city center

A correlation between age and distance to city center is implemented, with younger people tending to live in the city center and older people living in the suburbs. The relationship is modelled by taking a central ward, and calculating the distance of the every other ward from this central ward. These distance values are normalised to produce a distance coefficient between 0 and 1, where the furthest ward has value 1, and the wards bordering the central ward have value 0. The confounders modelled with a spatial relationship (percentage of population over 65 years, and percentage population under 15 years) have their values chosen through proportional and inversely proportional coefficient relationships respectively, while remaining within the value ranges in the real data.

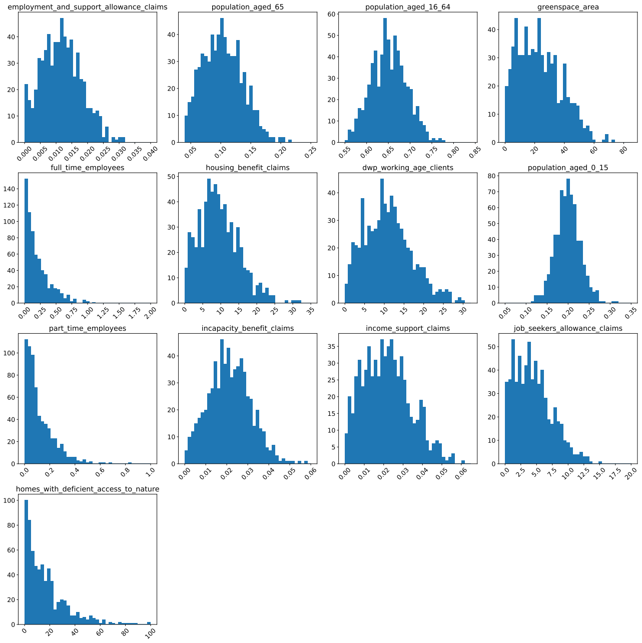

Figure 1: Confounder distributions

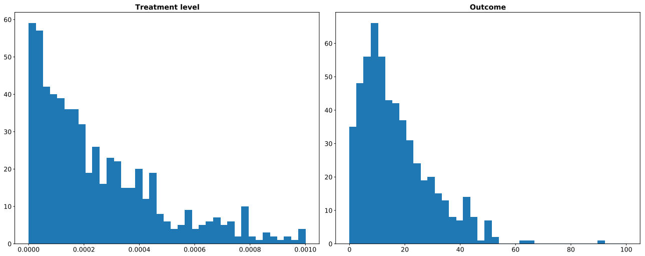

Figure 2: Example of a treatment and outcome distribution of one dataset

4.4 Confounder/outcome correlation

In real-life situations, often many factors contribute to an outcome instead of a single treatment. Therefore, an experiment is conducted in which a direct effect of a confounder is added to the outcome. This is implemented by adding an extra term to Equation 5:

$$Y \sim \text{Normal}(\alpha \cdot \sum_p x_{ip}, \beta \cdot \sum_p x_{ip}) + \gamma z_i + \sum_q \lambda_q \cdot x_{iq} \tag{6}$$

where q is the number of confounders which are to have an additional correlation with the outcome and λ_q the corresponding parameter indicating the strength of the correlation. These correlations are added on top of the spatial correlations. The amount of greenspace area and the number of homes with deficient access to nature are chosen to correlate negatively and positively, respectively, with the outcome. The confounders are chosen arbitrarily, since the aim is to test the propensity score matching method in general.

4.5 Analysis of experimental results

The experiments outlined in the previous sections result in the synthesis of 84 datasets: for every type of experiment, 21 datasets are generated with ATE values in $[-10, 10]$. The treatment level and outcome are both normalised before applying Equation 2. The resulting matched pairs are analysed, yielding estimated ATE values. These can be found in Appendix A. One can note that the estimated values are all either -1, NaN or 1 for the randomly distributed data, and very large or very negative for the other datasets. Intuitively, the last effect might be caused by the very small treatment values. Normalisation did not solve this problem. A possible solution that has been considered, was to remove all matched ward pairs which had treatment differences that lie outside of 3 or 2 standard deviations. Both these methods did not make significant changes to the values of the normalised treatment differences within pairs. Changing the range in which the treatment values are generated could be

5 RESULTS

This section will outline results from the preparations of the experiments and the results of the experiments that have been described in the previous section.

The synthesis of the uniformly at random distributed data is straightforward. The results will be presented further on in this section together with the other experiments.

5.1 Outlier management

Figures 3, 4, 5 and 6 show some results for the 4 distributions: correlated, realistic, spatial, random. The plots show the standardised treatment level on the x-axis, and the standardized effect on the y-axis. The nodes represent the differences in the variable values between matched pairs of wards, this means a high value indicates a big difference. The plots chosen are representative of all the data synthesized for each of the distributions. It is evident from the plots that each of the synthesized datasets for all distributions contains one very noticeable outlier on the treatment variable, which makes visualizing the data difficult as it squashes the other datapoints together. In order to allow for some analysis of the rest of the data, the outlier will be removed entirely.

5.2 Random distribution

The **random** distribution dataset is described in section 4.1, and is demonstrated with three plots showing the matched pair difference of antidepressant prescription, in standard deviations on the y-axis. **Figure 7** shows the standardised values of the matched pair difference percentage of the ward populations that are aged 65 years and above. This is done to give a comparison to the results of the spatially correlated dataset shown in Figure 13. The random distribution shows a significant difference, with a much smaller varaince in matched pair difference in antidepressant prescriptions, the spread across the x-axis is the same because the value-ranges were set, based on the original paper data. Figure 7 shows most of the matched ward pairs having an extremely low difference in effect, and only a small number straying from this. **Figure 8** shows the matched pair difference of treatment level for this distribution, and demonstrated a very clear linear relationship between the treatment and the effect. This is because for this distribution the outcome (antidepressant prescriptions) is based on the Average Treatment Effect multiplied by the treatment value. This results in the linear relationship demonstrated in the graph, which then shows that having a higher matched pair difference in treatment level results in a higher difference in effect. **Figure 9** shows the distributions in the same way as Figure 15, and clearly show that the vast majority of matched pairs shows the same smaller than average difference in antidepressant prescriptions, with an extremely small proportion of matched ward pairs straying from this. In all cases the variance between ward pairs was very small, with one remaining outlier

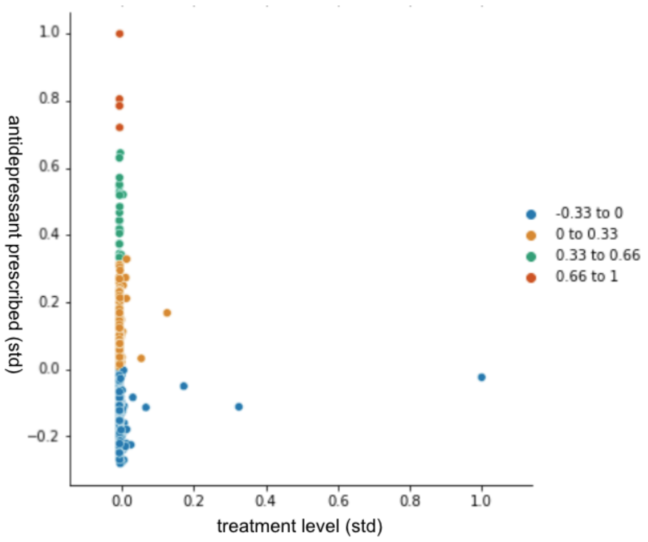

Figure 3: Correlated dataset standardised values including largest outlier, both axes showing standard deviations.

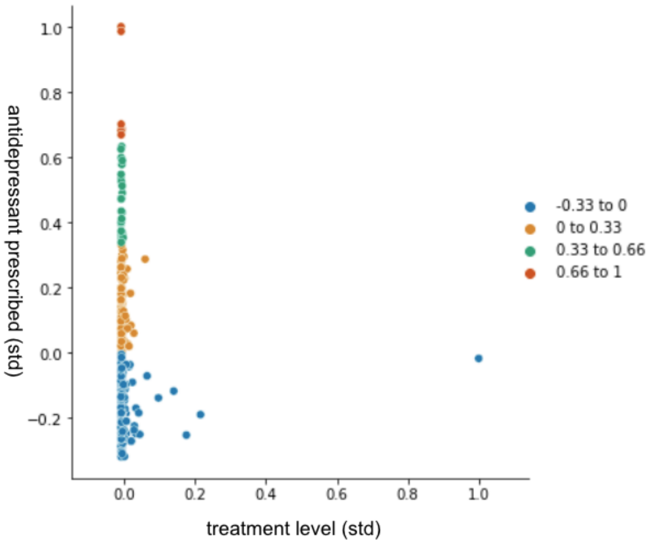

Figure 4: Realistic dataset standardised values including largest outlier, both axes showing standard deviations.

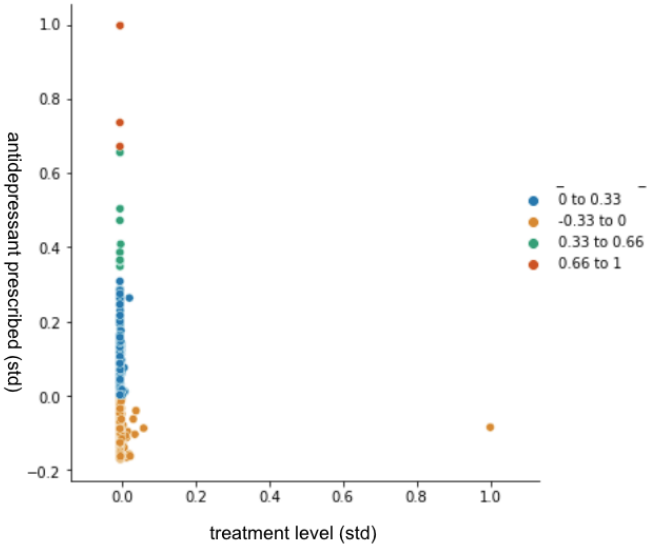

Figure 5: Spatial dataset standardised values including largest outlier, both axes showing standard deviations.

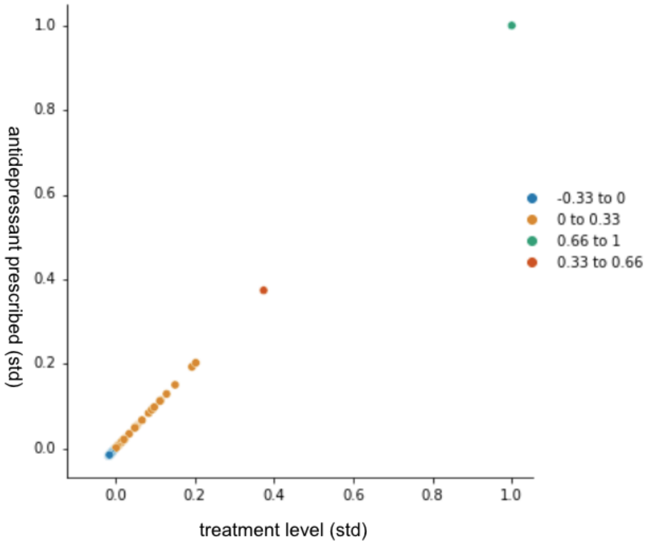

Figure 6: Random dataset standardised values including largest outlier, both axes showing standard deviations.

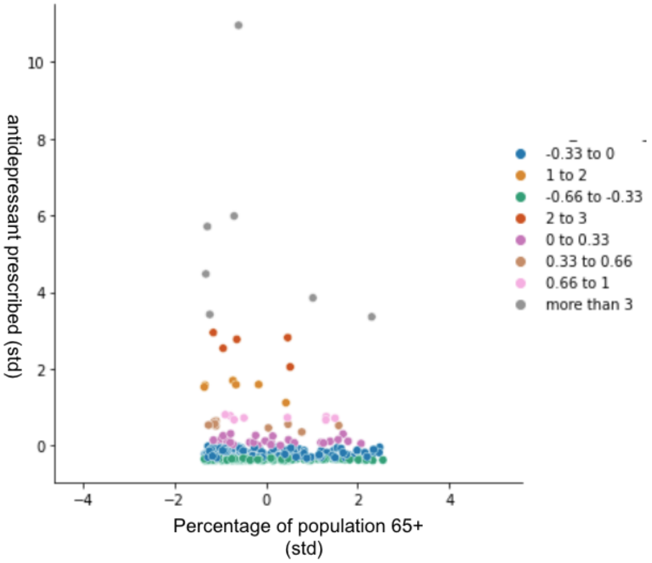

Figure 7: Random dataset standardised values excluding largest outlier, both axes showing standard deviations.

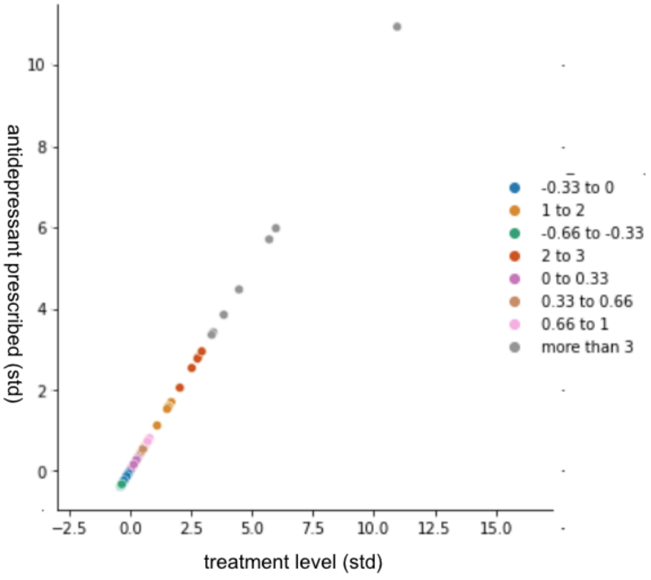

Figure 8: Random dataset standardised values excluding largest outlier, both axes showing standard deviations.

barely visible on the far right of the graph, which skews the results considerably.

5.3 Realistic distribution

The **realistic** distribution dataset is described in section 4.2, and is demonstrated with three plots showing the matched pair difference of antidepressant prescription, in standard deviations on the y-axis. Figure 10, where the matched pair difference in greenspace area is plotted on the x-axis, shows a much wider variety in matched pair differences than Figure 16, with a much more even spread of matched pairs in each level of antidepressant prescription difference. For this distribution it appears that having a larger difference in greenspace area matters less than in the correlated dataset, suggesting that in a real-life situation this confounder matters less than it does in the synthesized correlated data. **Figure 11** shows the matched pair difference of treatment level on the x-axis, and shows a significantly greater spread in matched pair treatment level differences than the correlated dataset. Although again, the vast majority of matched pairs have a smaller than average difference in

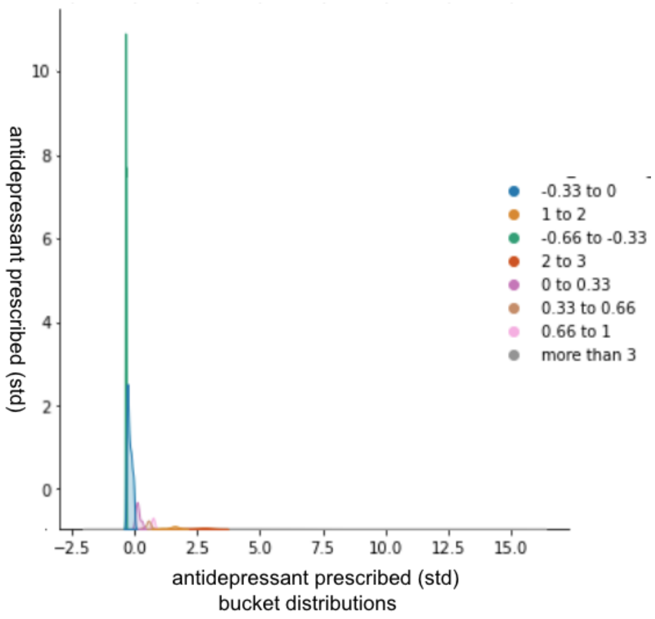

Figure 9: Random dataset standardised distribution excluding largest outlier, axis showing standard deviations.

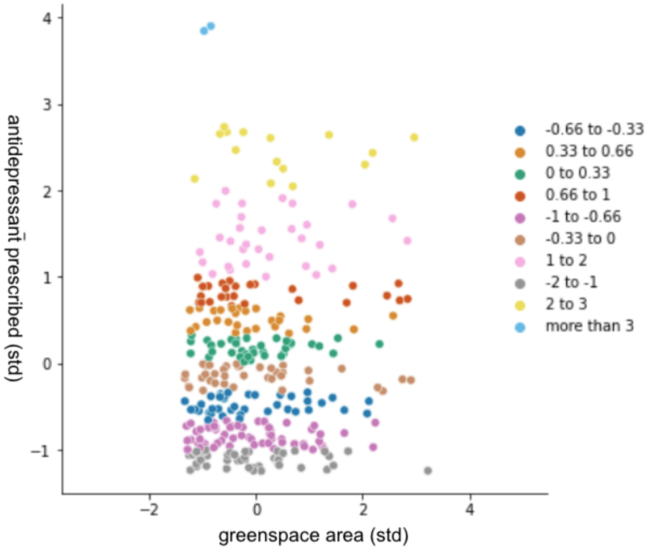

Figure 10: Realistic dataset standardised values excluding largest outlier, both axes showing standard deviations.

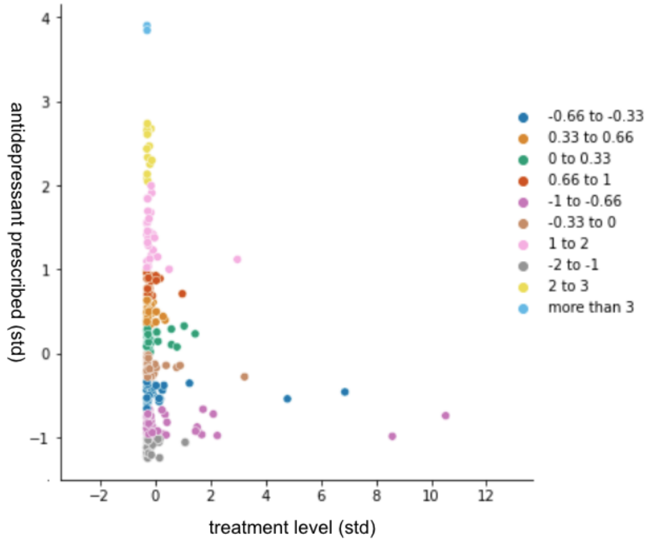

Figure 11: Realistic dataset standardised values excluding largest outlier, both axes showing standard deviations.

treatment level, a small minority of ward pairs display a very much larger difference, and these ward pairs tend to have only a small matched pair difference in antidepressant prescriptions. **Figure 12** shows the matched pair antidepressant prescription difference distribution, where it is evident that the number fo standard deviation buckets used used is much larger than for Figure 18. Smaller matched pair differences in antidepressant prescription levels were more common, but showed much less variance than the distributions of larger matched pair differences.

5.4 Spatial distribution

The **spatial** distribution dataset is described in section 4.3, and is demonstrated with three plots showing the matched pair difference of antidepressant prescription, in standard deviations on the y-axis. **Figure 13** shows the standardised values of the matched pair difference percentage of the ward populations that are aged 65 years and above. This particular confounder, among 2 others, has been modelled to be spatially correlated, where wards further from the centre f the city have a higher population of over 65, and a lower population of 15 years and under. The spread of the matched pairs along the x-axis is similar to Figure 10, but the pairs are squashed together much more densely in terms of antidepressant prescription differences when compared to Figure 10. The majority of wards had a lower than average population of over 65s, which may well be due to the wards which are more central tending to be smaller. Meaning a larger total number of wards would have a lower population of over 65s, and a smaller number of wards would have a very much higher proportion. This conclusion is corroborated by Figure **15**, which shows a similar picture to Figure 12, where most of the wards had a smaller than average population of over 65s, and a smaller number of wards with a higher variance had a larger proportion.

Figure 14 shows the matched pair difference of treatment level for this distribution, and shows a similar sort of spread to Figure 11, suggesting that the spatial data is more realistic (more like the realistic distribution) than the correlated dataset. Again, in general matched ward pairs with a bigger difference in treatment level tend to have a smaller difference in antidepressant prescriptions. Although the majority of ward pairs with a small difference in treatment level also have a small difference in effect.

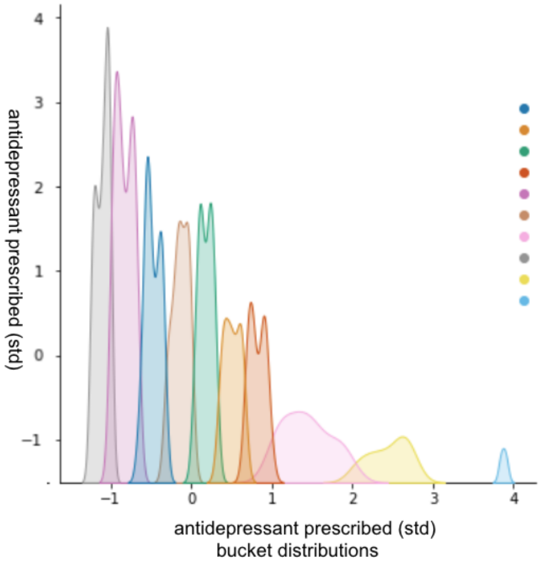

Figure 12: Realistic dataset standardised distribution excluding largest outlier, axis showing standard deviations.

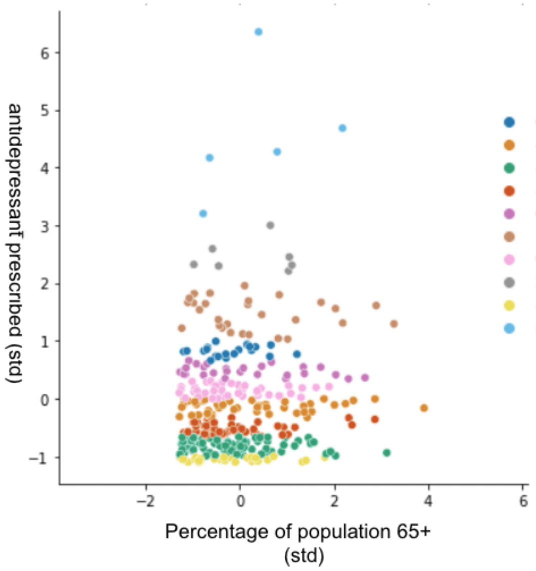

Figure 13: Spatial dataset standardised values excluding largest outlier, both axes showing standard deviations.

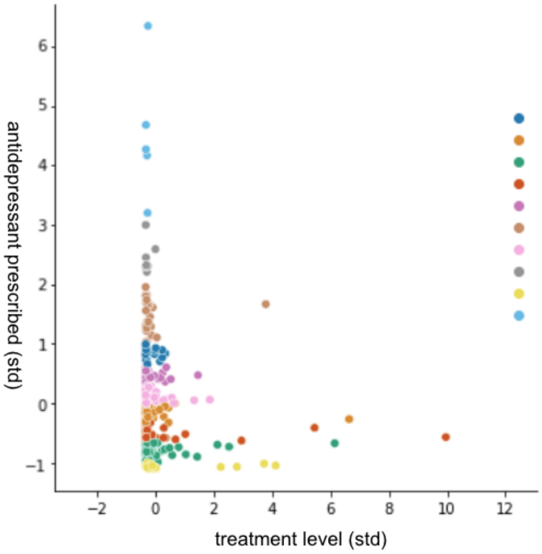

Figure 14: Spatial dataset standardised values excluding largest outlier, both axes showing standard deviations.

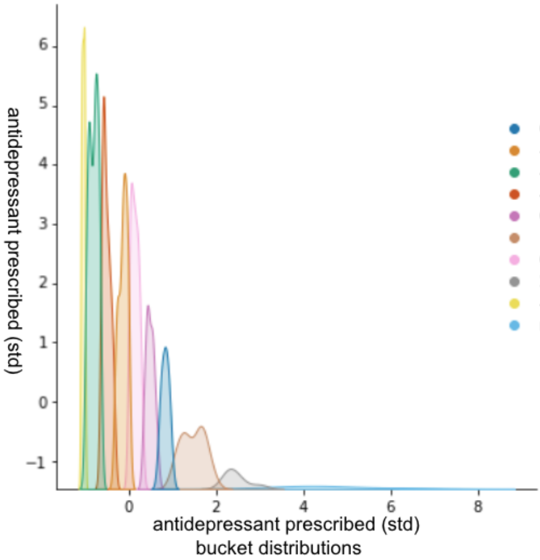

Figure 15: Spatial dataset standardised distribution excluding largest outlier, axis showing standard deviations.

5.5 Correlated distribution

The **correlated** distribution dataset is described in section 4.4, and is demonstrated with three plots showing the matched pair difference of antidepressant prescription, in standard deviations on the y-axis. The effect of confounders on the dataset is shown in Figure 16, where the matched pair difference in greenspace area is plotted on the x-axis. Having a larger difference in greenspace area between the pairs does not necessarily mean that the difference in antidepressant prescriptions is higher, but matched pairs that do have a high difference in greenspace area do generally also have higher differences in antidepressant prescriptions. **Figure 17** shows the matched pair difference of treatment level on the x-axis, and seems to show that although the differences in treatment level are extremely small (due to a second outlier which was not removed). Interestingly, the greatest variation in treatment level difference occurred in pairs with a smaller difference in antidepressant prescriptions. **Figure 18** shows the matched pair antidepressant prescription difference distribution, and shows that the majority of matched pairs had a very small difference in effect, and a very small minority had an extremely large difference in effect. This second category makes the graphical analysis difficult without removing more of the outliers.

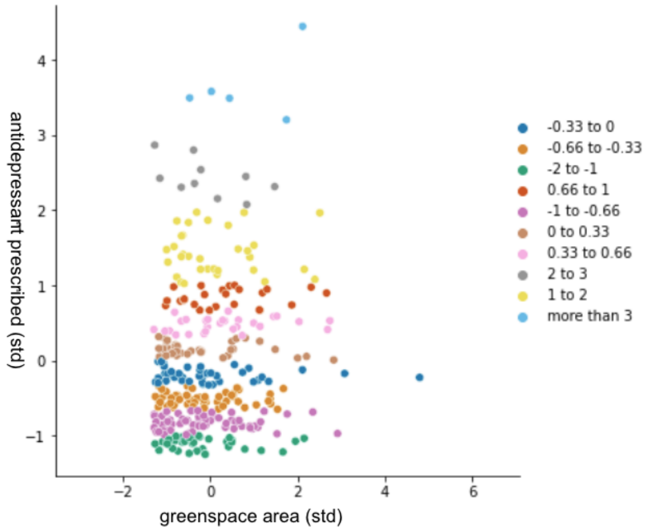

Figure 16: Correlated dataset standardised values excluding largest outlier, both axes showing standard deviations.

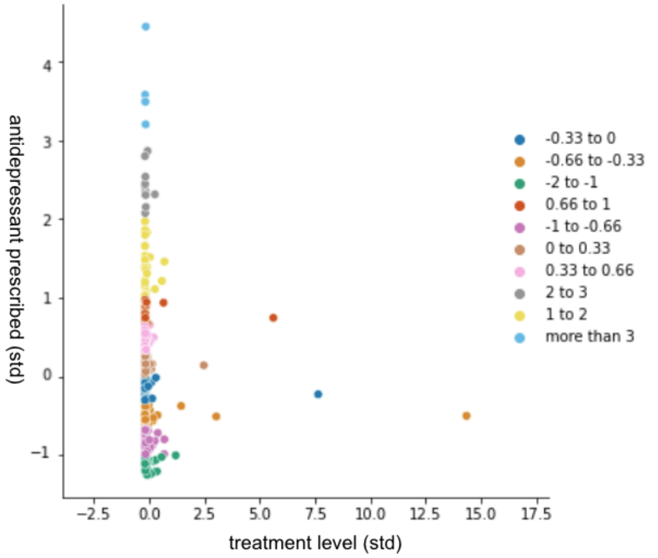

Figure 17: Correlated dataset standardised values excluding largest outlier, both axes showing standard deviations.

6 CONCLUSION & DISCUSSION

The results are highly inconclusive. In all the datasets a very far-out outliers appears, the reason for which is unclear. Moreover, the estimated ATEs of the different datasets are highly unrealistic.

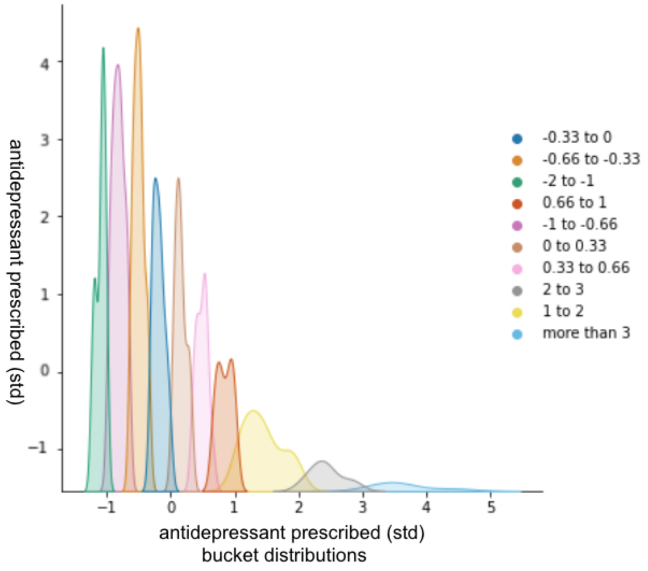

Figure 18: Correlated dataset standardised distribution excluding largest outlier, axis showing standard deviations.

One can conclude from the issues with calculating the ATE, that this method is very sensitive to the parameter ranges of the different values involved. This is something that has to be carefully considered from the start. The values need to be normalised and similarly scaled. It is important to consider how this should be done while preserving the original information in the data as much as possible. The interplay between the treatment and outcome complicates the process of data synthesis. The distributions of the treatment and outcome vary based on the specific confounder values of each ward in a dataset, thus these distributions would ideally be tuned for each dataset, if the goal is to emulate real data. Manual tuning was not sensible for this research, given that 84 datasets were considered.

All in all, it can concluded that this method of causal inference, which already makes many assumptions, in any case is very sensitive to the scaling of the input data. There may be more, undiscovered factors, which have contributed to the problem. These observations do not speak in the favour of the robustness of propensity scoring used for causal inference from these types of datasets.

It is important to note that the choice of topic, i.e. public health and the built environment, is an arbitrary topic to test the proposed methods. It is also a topic which could greatly benefit from tools to find causal relationships. However, to seriously consider propensity score matching as a tool for finding causal effects between the built environment and public health, many other factors must be considered. For example; the size of a ward might affect where people go to sports venues or visit the doctor. If a ward is very small, people might go to neighbouring wards. Even if a ward is not particularly small, people living close to the border of a ward, might visit doctors and sports venues just across the border. Another factor to be taken into account is connections between business wards and residential wards: a person might have a gym at their work place, while they visit a doctor that's in the ward they live in. Furthermore, one could argue that wards that are better connected in terms of public transport would be more likely to attract more people than wards that are less well connected. There are also some issues with the sports venues: factors such as the size, or for example the age groups that can join the venue, are left out. A suitable adjustment for future research on this topic would be to adjust the treatment level based on the size of a sports venue, or how many people are a member there. There are also socioeconomic factors. More expensive sports venues could have a different effect in poorer wards than they would have in more wealthy wards. These are all non-trivial complicating factors, it needs to be seriously considered how these issues could best be addressed for the purpose of causal inference.

To elaborate further on the issue of applying causal inference on the specific problem of public health and the built neighborhood, as opposed to using it as a casus for testing the methods, the choice of treatment and effect need to be carefully considered. Intuitively, one could expect there to be some sort of effect between the number of sports venues and the number of antidepressant prescriptions. However, is this really a suitable proxy for the built environment and public health? When applying the proposed causal inference models to real-life situations for informing public policy, the choice of treatment and outcome must be carefully considered.

6.1 Future work

The results of this paper are inconclusive, as were the results of the original paper, suggesting the method for propensity scoring analysis used in these two papers may be sub-optimal. Avenues for future work include thorough research into understanding what caused the outliers found in the results. Another possible research direction would be to consider the calculation of the estimated average treatment effect and how the normalisation or scaling of the different confounders, treatment and outcome affects this. Alternatively, other methods for causal inference might be applied to the studied problem, in order to compare the methods.

REFERENCES

[1] Cande V. Ananth and Enrique F. Schisterman. 2017. Confounding, Causality and Confusion: The Role of Intermediate Variables in Interpreting Observational Studies in Obstetrics. *HHS Public Access* (2017).

[2] Greater London Authority. [n.d.]. Statistical GIS Boundary Files for London. https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london

[3] Anderson J. Cotterill S. et al. Benton, J.S. 2018. Evaluating the impact of improvements in urban green space on older adults' physical activity and wellbeing: protocol for a natural experimental study. *BMC Public Health* (2018).

[4] M Birkin and M Clarke. 1988. Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environment and Planning A: Economy and Space* 20, 12 (1988), 1645–1671. https://doi.org/10.1068/a201645 arXiv:https://doi.org/10.1068/a201645

[5] Marco Caliendo and Sabine Kopeinig. 2008. SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. *Journal of Economic Surveys* 22, 1 (2008), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6419.2007.00527.x

[6] Rajeev H. Dehejia and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics* 84, 1 (2002), 151–161.

[7] Jack Edmonds. 1965. Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards B* 69 (1965), 125–130.

[8] Apinan Hasthanasombat and Cecilia Mascolo. 2019. Understanding the Effects of the Neighbourhood Built Environment on Public Health with Open Data. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 648–658. https://doi.org/10.1145/3308558.3313701

[9] Bo Lu, Elaine Zanutto, Robert Hornik, and Paul R Rosenbaum. 2001. Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse. *J. Amer. Statist. Assoc.* 96, 456 (2001), 1245–1253. https://doi.org/10.1198/016214501753381896 arXiv:https://doi.org/10.1198/016214501753381896

[10] Harmen Oppewal. 2010. Concept of Causality and Conditions for Causality. *Wiley Sons, Ltd* (2010).

[11] Julia M. Rohrer. 2018. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018), 27–42.

[12] ZonMw. 2020. Project catalog Healthy Neighborhood. (2020).

A ESTIMATED ATE VALUES

Table 2: Uniformly at random distributed data

ATE	Est. ATE	Min	25th %	75th %	Max
−10	−1.0	−1.0	−1.0	−1.0	−1.0
−9	−1.0	−1.0	−1.0	−1.0	−1.0
−8	−1.0	−1.0	−1.0	−1.0	−1.0
−7	−1.0	−1.0	−1.0	−1.0	−1.0
−6	−1.0	−1.0	−1.0	−1.0	−1.0
−5	−1.0	−1.0	−1.0	−1.0	−1.0
−4	−1.0	−1.0	−1.0	−1.0	−1.0
−3	−1.0	−1.0	−1.0	−1.0	−1.0
−2	−1.0	−1.0	−1.0	−1.0	−1.0
−1	−1.0	−1.0	−1.0	−1.0	−1.0
0	*nan*	*nan*	*nan*	*nan*	*nan*
1	1.0	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0	1.0
8	1.0	1.0	1.0	1.0	1.0
9	1.0	1.0	1.0	1.0	1.0
10	1.0	1.0	1.0	1.0	1.0

Table 3: Realistically distributed data

ATE	Est. ATE	Min	25th %	75th %	Max
−10	−150.836	−228587.109	−499.003	448.288	122829.517
−9	26597.272	−385866.424	−600.426	576.252	9072827.62
−8	905.163	−103956.9	−463.986	424.235	362591.684
−7	556.453	−146240.968	−325.962	479.484	206254.903
−6	−2427.006	−1162632.525	−521.067	572.925	316138.405
−5	266.879	−393589.886	−460.658	386.284	488693.656
−4	980.165	−93500.62	−285.54	669.194	225567.81
−3	1207.704	−86594.399	−481.439	894.029	83833.161
−2	−740.236	−45604.158	−972.737	253.717	62525.846
−1	−1991.72	−395585.624	−594.947	457.934	118724.698
0	9738.667	−42843.467	−401.617	730.675	1112341.838
1	356.911	−44098.977	−592.777	409.361	202699.477
2	−192.549	−85721.335	−508.08	320.839	49949.455
3	34054.437	−2357892.413	−432.319	564.651	13963399.255
4	265.081	−73309.625	−397.332	610.84	90833.324
5	−163.623	−122421.027	−506.793	481.354	217973.704
6	18.967	−101986.736	−709.57	259.2	266815.7
7	−580.25	−779801.334	−426.871	574.058	513404.902
8	46124.429	−2102351.897	−576.61	622.556	16993648.356
9	−16255.701	−4634992.227	−455.549	508.605	292685.121
10	−62.691	−120794.901	−545.639	580.354	73187.941

Table 4: Spatial confounder distribution

ATE	Est. ATE	Min	25th %	75th %	Max
−10	−5174.72	−738394.727	−545.069	511.519	106509.78
−9	83.787	−50.06	−0.534	0.81	25787.114
−8	−2972.751	−684408.055	−645.864	572.29	273945.573
−7	−485.256	−452414.383	−442.893	358.439	444041.162
−6	−170.211	−71422.321	−373.593	450.19	56754.121
−5	−65.348	−20907.934	−0.247	0.267	14208.041
−4	1598.472	−582529.239	−469.298	552.629	307322.023
−3	−2843.47	−645375.067	−420.961	614.682	173527.489
−2	−39197.895	−12497096.771	−448.99	540.56	365584.591
−1	−2922.188	−913576.904	−0.219	0.256	373.117
0	−1719.338	−498500.925	−881.874	490.906	37090.252
1	1842.316	−243992.328	−362.286	365.013	491533.279
2	122.263	−2162.997	−1.158	1.388	40857.065
3	234.766	−64971.77	−542.023	712.261	96458.44
4	266.858	−121257.158	−679.843	367.507	167353.281
5	613.857	−54509.842	−428.832	656.367	111731.59
6	4283.015	−209642.302	−496.273	427.79	1557944.041
7	91.022	−96.277	−1.254	2.539	29114.132
8	970.433	−339.991	−0.459	0.356	318742.587
9	−107.592	−38949.737	−1.664	1.671	3842.761
10	−118.115	−292456.255	−439.707	602.442	134146.518

Table 5: Confounder & outcome correlation

ATE	Est. ATE	Min	25th %	75th %	Max
−10	−234.238	−123577.108	−461.535	452.435	106777.673
−9	−2985.732	−500466.072	−559.131	519.354	59519.819
−8	6673.978	−445183.452	−444.604	341.407	2351391.413
−7	−8968.083	−2593239.303	−611.451	362.873	125724.986
−6	−1326.931	−562901.309	−521.991	364.617	333086.64
−5	179.624	−633.056	−0.366	0.612	59432.228
−4	3078.283	−105564.666	−208.258	734.551	253635.248
−3	1935.447	−201057.525	−566.773	495.141	334497.177
−2	−7109.411	−1905600.522	−562.494	368.581	462213.683
−1	−2206.267	−534005.556	−357.788	400.032	213600.294
0	−202.812	−220870.814	−480.06	400.056	223358.435
1	997.156	−167665.861	−559.023	471.106	479518.43
2	1358.524	−145589.923	−578.141	553.034	745905.56
3	−148.696	−195760.414	−442.732	748.427	128915.16
4	−328.482	−702980.553	−488.195	560.802	659005.661
5	−3779.451	−1238627.558	−496.079	318.334	148259.361
6	2417.18	−36394.135	−586.362	745.081	533700.339
7	384.314	−118203.937	−384.839	843.521	78998.322
8	166.801	−176.885	−2.087	1.682	41894.518
9	10843.546	−110823.745	−637.775	450.27	3658450.118
10	−1463.738	−513653.563	−343.592	598.55	63081.542