# Understanding the Effects of the Neighbourhood Built Environment on Public Health with Open Data

Apinan Hasthanasombat
University of Cambridge
United Kingdom
ah953@cam.ac.uk

Cecilia Mascolo
University of Cambridge
United Kingdom
cm542@cam.ac.uk

## ABSTRACT

The investigation of the effect of the built environment in a neighbourhood and how it impacts residents' health is of value to researchers from public health policy to social science. The traditional methods to assess this impact is through surveys which lead to temporally and spatially coarse grained data and are often not cost effective. Here we propose an approach to link the effects of neighbourhood services over citizen health using a technique that attempts to highlight the cause-effect aspects of these relationships. The method is based on the theory of *propensity score matching with multiple 'doses'* and it leverages existing fine grained open web data. To demonstrate the method, we study the effect of sport venue presence on the prevalence of antidepressant prescriptions in over 600 neighbourhoods in London over a period of three years. We find the distribution of effects is approximately normal, centred on a small negative effect on prescriptions with increases in the availability of sporting facilities, on average. We assess the procedure through some standard quantitative metrics as well as matching on synthetic data generated by modelling the real data. This approach opens the door to fast and inexpensive alternatives to quantify and continuously monitor effects of the neighborhood built environment on population health.

## CCS CONCEPTS

• **Information systems** → *Spatial-temporal systems*; *Data mining*; *Web mining*.

## KEYWORDS

Open data; Population health; Causal inference, Propensity score.

**ACM Reference Format:**
Apinan Hasthanasombat and Cecilia Mascolo. 2019. Understanding the Effects of the Neighbourhood Built Environment on Public Health with Open Data. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313701

## 1 INTRODUCTION

The effects of urban neighbourhood characteristics on health outcomes of its residents is an ongoing research topic [2]. With over half of the population living in urban areas and a projected increase to 68% by 2050 [31], understanding exactly how these neighbourhood aspects impact different disease outcomes is a significant public health concern. Traditional research into the risk factors of many diseases focus mainly on 'direct causes' - those that are close to the biological mechanisms that are responsible for the development of the disease. For instance, high cholesterol and coronary heart disease. However, in order to deliver maximum impact when designing interventions that prevent diseases in the population at large, it has been argued that we should also focus on what puts people at *'risks of risks'* [15, 27] - the factors that lead people to be exposed to the direct causes of disease in the first place. These, which we call 'higher level causes', for instance social-economic status (SES) or the availability of services and amenities in a neighbourhood environment, are useful from a public health point of view, as they can be influenced directly by policy, as opposed to an individual's diet or lifestyle choices.

Previous work on neighbourhood effects on health mainly consists of observational cross-section studies, with the SES of the neighbourhood as the most studied variable, and Body Mass Index (BMI) as the most studied health outcome [2]. We find two main issues. Firstly, a large proportion of these works rely on traditional data sources such as surveys which can be quite expensive at scale and therefore lead to limited geographical and time span coverage. This focus on traditional data sources also means aspects such as the availability of services and amenities in the neighbourhood are neglected, and studies to understand its impact on residents' health are few, even though the built environment was found to have a consistent association with for example, depression [18], amongst neighbourhood structural features. Some previous work have for instance looked at food venue density and BMI [34], and tobacco store density and life expectancy [8]. Secondly, these discussions have mainly revolved around associations, and the daring leap to attempt at a causal conclusion are few. Since these physical aspects of a neighbourhood are controllable and can be planned, this has enormous potential to help inform public policy decisions. The data offered by projects such as Open Street Map (OSM), combined with the growing availability of causal methodology, turns this potential closer to a reality at scale.

In this paper we try to leverage this available open data offering fine grained spatio-temporal granularity and a causality framework which takes us closer to eliciting causal effects. There are advantages to this approach. The use of open data offers a scalable and cost effective solution in addition to the granularity in comparison with survey methodologies. Additionally, physical characteristics are better defined and quantifiable compared to measures such as SES. Specifically, and to illustrate the approach, we investigate the availability of sporting facilities and its potential causal effect on antidepressant prescriptions across over 600 neighbourhoods in London over three years.

In summary, the contributions of this paper are:

*WWW '19, May 13–17, 2019, San Francisco, CA, USA*
2019. ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313701

(1) A general cost effective approach to study the influence of the built environment and neighbourhood characteristics, and the mechanisms that underlie this influence such as access to tobacco, certain types of food, or sports, over population health.
(2) We work on a specific example, exploiting urban open data to show that the effect of sporting facilities on antidepressant prescriptions in London.
(3) Our analysis, using data from Open Street Map (OSM), the UK National Health Service (NHS) prescriptions and the census, shows that the effect of sport venues over prescriptions for antidepressants follows an approximate normal distribution centred at a negative effect for prescriptions with increases in sporting facilities, on average.
(4) We assess our approach by examining the results from the procedure for balance, as well as comparing with results from the null model and of synthetic datasets modelled based on real data, where we have knowledge of the real effect.

This approach opens the door to new methodologies to study urban health and to offer urban planners mechanisms to study these geo-temporal processes at a fine grained scale. We first give a general overview and intuition behind the approach, followed by a more detailed explanation. Section 5 and 6 then describes the results of our analysis on a specific case study and our evaluation. We conclude by discussing the implications of our investigations, its limitations, and potential.

## 2 RELATED WORK

The study of neighbourhood effects on health has grown rapidly over the last two decades [22] [2]. This ranges from grocery store density and BMI [34] to tobacco store density and life expectancy [8]. The neighbourhood variable under study can be split largely into two categories. The tangible aspects of the neighbourhood, such as green spaces, amenities, or venues offering services, and the intangible aspects of the neighbourhood, comprising mostly of social aspects such as social cohesion, social disadvantage, to perceived safety.

The research area dates back to Durkheim's study of suicide in the 19th centuary [5] but the most recent concrete indication of any plausible causal effect is shown by a rare randomised controlled study conducted in the United States which shows that disadvantaged families moving to neighbourhoods with lower poverty rates lead to long term improvements on mental and physical health [17].

Non experimental studies in the literature predominantly feature social characteristics of a neighbourhood [2][22]. However, these intangible variables are often difficult to define, hard to measure and quantify, as opposed to the physical characteristics. Additionally, the traditional methods of surveys lends itself well to the study of social aspects as opposed to the physical, which may explain the lack of studies in the latter area.

Reviews of neighbourhood effects research suggests that apart from a few key studies such as the randomised experiment, evidence is inconclusive regarding many health outcomes [18, 22], and the literature has had its fair share of discussion around the validity of causal estimates from non experimental data [21][12]. Additionally, even though association studies are plentiful, there has been a lack of proposed mechanisms for which high level neighbourhood characteristics can influence health, and of the ones that are proposed, hardly any are related to the tangible aspects of the neighbourhood [4]. This study examines one such potential mechanism, given that the built environment is noted to be consistently associated with depression [18].

There is a glaring opportunity to leverage non-conventional data sources to help characterise neighbourhoods, and an appropriate methodology that will allow this to be done continuously at scale. Alternative data sources, such as those from mobile phone call detail records, social networks such as Twitter, Foursquare, Google Street View and other web data, have been increasingly used in a wide range of applications from estimating poverty levels [30], measuring social diversity [10], to auditing the neighbourhood environment [29].

There exists studies that investigate how these data sources can be used to understand and improve public health. For instance, the association between human lifestyles and chronic diseases, and its use in predicting diseases evolution in urban areas [34]. Another study used a model based on twitter data to predict disease prevalence, and found associations between the different geographical risk factors (e.g tobacco use, exercise) based on phone interview data, with the model prediction [24]. However these are constrained to associations and predictions. Traditional studies attempting to establish causal links on the other hand, such as that between physical activity, fast food environments, BMI, body fat percentage, and waist circumference, noted lack of data on the food environment to be a limitation [19].

In short, we see three potential improvements to the state of neighbourhood effects research: More focus on physical characteristics of the neighbourhood, a deeper investigation into the potential physical causal mechanisms by which the neighbourhood effects health, and scalability of methodology to continuously monitor effects leveraging non-traditional data sources.

This paper illustrates an alternative approach which addresses these concerns and contributes further to the discussion.

## 3 APPROACH OVERVIEW

We now describe the overall approach. We are interested in studying specific characteristics of neighbourhoods, say the presence of particular services - for instance sporting facilities - over a population health outcome, such as antidepressant prescriptions. The neighbourhood characteristic here is then called the *treatment*. This stems from the fact that we are adopting a causal view; we want to find the effect that applying the treatment would have on the outcome for the average neighbourhood.

*Geographical Separation.* The units under study are then not specific individuals, but rather neighbourhoods. The adopted definition of a neighbourhood in this study is that of a *ward*, an administrative geographical separation, of which, for instance, London has approximately 625. At the beginning of the calendar year, each ward is considered treated with a particular *dose* of treatment, where a dose is the binned value of the absolute value of the treatment variable.

*Randomized Controlled Trials.* In an ideal scenario, the best approach to estimate the causal effect of the neighbourhood treatment to the outcome is to conduct a randomised controlled trial (RCT).

We randomly select half of the wards and strip them of the treatment variable as control, and apply the same amount of treatment to the rest. Why this is infeasible is obvious, but what is the next best alternative?

We run an *observational study*, which aims to achieve the same goal as a RCT, but without data from a randomised experiment. To do this we first look at the intuition behind why RCTs work in the first place. The key point is that randomisation of treatment assignment to each neighbourhood ensures that all of the variables that may also be influencing the outcome are equal in expectation as the number of neighbourhoods increases. This means that the outcomes of the two groups, those exposed to treatment and the control, are comparable because the only difference between the groups is the treatment status. An important point here is that we do not have to know all the variables that also affect the outcome; the randomisation ensures these are balanced even if unobserved.

The observational study conducted here attempts to achieve the same goal as the RCT - construct two groups where the outcome is comparable with the only difference being the treatment status. However, since we no longer have randomisation, two problems arise. Firstly, there could be variables that also affect the outcome, which are only balanced in the groups if the treatment assignment is randomised. Additionally, we also could have variables that also affect the treatment status of the units, as treatment is not randomly assigned. Which means even if we were able to construct two groups with balance on the variables that affected the outcome, the groups may be imbalanced with respect to the variables that determined treatment, and the effect will not be representative of the average effect on the population. The variables that potentially affects both the treatment, outcome, or each other are called *confounding variables* or *confounders*[1]. The game here is then to construct two groups such that on average, the confounding variables are balanced, but the treatment assignment is different.

*Matching.* To achieve this, we use a procedure called *matching*. For each control unit, if we were able to find another unit in the population with the same set of values for all confounding variables but instead has received a different treatment, then the difference between the outcomes of the pair can be calculated and averaged over all pairs in the population. This is called exact matching. However finding an exact match is not always possible, so instead we find, for each unit, a match that would minimise the overall difference between the average value for all confounders between the two groups. This is the basic idea behind the approach.

*Propensity Score.* There are two additional modifications. Firstly, since there could be many confounding variables, high-dimensional matching between the units could be potentially difficult, so we match instead on a single measure called the *propensity score*, which is a function of the confounding variables and is shown to achieve balance, on average, for the confounding variables [28]. Secondly, instead of a binary treatment status, treated or control, there are multiple possible treatment 'doses', where the dose is defined by the bin in which the treatment variable falls into. The two constructed groups now consists of the 'low dose' group and the 'high dose' group, because units can only be matched if they received a different dose of treatment. The balancing nature of the propensity score still applies in this case under some conditions.

*At a Glance.* In summary, we estimate the causal effects by emulating the results of a randomised controlled trial using a matching procedure based on propensity scores to balance the confounding variables between the high dose and low dose treatment group. The difference between the outcomes of these two groups is interpreted as the average causal effect of the treatment on the population.

[1]While in some literatures the word confounder is used to refer to a particular type of confounding variable - the common cause between treatment and outcome - here we use it interchangeably with confounding variable

## 4 METHODOLOGY

In this section we go into more detail on various parts of the methodology before we describe its application to a specific example.

As briefly mentioned previously, the aim is to attempt to emulate a randomised controlled experiment - which means that the confounding variables between the 'treated' and 'control' group in the binary case, or high and low dose group in the multiple treatment case, should be balanced on average.

To achieve this, the wards are matched in such a way that minimises the difference between the confounding variables over all matched pairs, and maximises the difference between the treatment 'dose' e.g the binned value of available sport venues in a ward, in each pair.

In this section we first describe matching in the case of a binary treatment, then introduce the need for the propensity score, followed by how this matching can be achieved with multiple treatment levels.

### 4.1 Matching with Binary Treatment

Finding a match for a treated unit with precisely the same values for all variables in the set of confounders is called exact matching. In practice, especially with more than a handful of confounders, this is infeasible.

Given a confounder $x$, if an exact match cannot be found, the nearest value to the confounder could be used instead. However, where there is more than one confounder, the matching procedure uses the sum of the difference between all of the confounders, and aims to find a matching that minimizes this total sum. This is called nearest neighbour matching. We find a set of pairs of units $M$, given a set of $p$ confounders to minimise the distance

$$\sum_{(i,j)\in M} \sum_{p} |x_p^i - x_p^j|$$

where $x_p^i$ denotes the $p$th confounder of unit $i$. This optimisation problem can be solved as a graph matching problem. Given a graph $G = (V, E)$ a matching is a subset $M \subset E$ such that for all $(u_1, v_1), (u_2, v_2) \in M$, $u_1 \neq v_1 \neq u_2 \neq v_2$ i.e no two edges share a common vertex.

A couple of definitions on graph matching. A *perfect matching* is a matching that includes all the vertices of a graph. This is only possible when the number of vertices are even. A near-perfect matching is a matching when only one vertex is excluded from the matching. This happens when the number of vertices are odd. A matching is called *maximum* if it contains the largest number of edges, which in turn implies it contains the largest number of vertices. A perfect and near-perfect match is always a maximum

matching. The problem of finding maximum matchings can be extended to weighted graphs, where each edge is associated with a distance. We can then ask for a maximum matching that minimises the total sum of all edge weights. Such a matching is called an *optimal matching*. Our problem can be framed as an optimal matching problem with a fully connected graph, where the vertices represent the wards, and the edge weights are the difference between the values of the confounders between each potentially matched pair.

### 4.2 The Propensity Score

Even with nearest neighbour matching, the high dimensionality of the confounders can still be a problem. With many confounders, the optimal match may still yield large overall distances in the matching, and the confounders in the treatment and control group could still remain imbalanced, if the sample size is small relative to the dimension of the confounders. The theory of the propensity score resolves this issue by stating that, instead of matching on all of the confounders, which may possibly be high dimensional, if the matching was performed based on a value called the propensity score, a one dimensional scalar value, then the confounders in the treatment and control group should still be balanced on average, in the limit of a large sample [28]. This was intitially shown to be true in the case of binary treatment. The estimation of the propensity score is discussed next in the context of multiple treatments.

### 4.3 Matching with Multiple Treatment Levels

If we let $z$ denote the level of treatment received, the discussion so far has been constrained to the case where there are only two possible 'doses' of treatment, that is, no treatment at all $z = 0$, and a single value for treatment $z = 1$. For our particular application, we have to consider extensions of this to the case where there are many possible levels of treatment.

The propensity score is unknown, and has to be estimated. Additionally, the balancing nature of the propensity score is only proved in the case where there were two levels of treatment. Joffe et al [11] showed that the balancing property of the propensity score could be extended to the case with varying doses of treatment. This means that matching based on a scalar propensity score will still balance observed confounders when there are different doses of treatment. This is precisely the case when $P(z|x) = P(z|b(x))$, where $x$ are the confounders and $b(x)$ is the propensity score. McCullaugh's ordered logistic model [20] would satisfy this criteria.

In light of this, the propensity score for any unit $i$ corresponds to $\beta^T_{x_i}$ in the model

$$log \frac{P(z_i) \geq d}{P(z_i) < d)} = \theta_d + \beta^T \mathbf{x_i} \tag{1}$$

where $d$ is the different doses of treatment, and can take values $= 2, 3, .., D$. Note that here, the distribution of treatment levels given the confounders depends only on $\beta^T_{\mathbf{x}}$, which means $P(z|x) = P(z|b(x))$, and makes it a balancing score in the case with multiple treatments. An estimate of this value can be obtained using maximum likelihood, and is denoted $\hat{\beta}^T_{\mathbf{x_i}}$.

Another difference when performing matching with doses is that we would like matched pairs to have different levels of treatment. Similar to the case with binary treatment when control units are only allowed to match with treated units and vice versa.

To maximise the difference between the treatment levels of the matched pairs, instead of matching directly with the difference of the estimated propensity score, a modified version of the distance metric that incorporates the treatment dosage, adopted from Lu et al [16] is used. The distance between two units $i$ and $j$, denoted $d_{i,j}$ is then defined as

$$d_{(i,j)} = \frac{|\hat{\beta}^T_{\mathbf{x_i}} - \hat{\beta}^T_{\mathbf{x_j}}| + \epsilon}{|z_i - z_j|} \tag{2}$$

Where $\epsilon$ is an extremely small positive number. The role of $\epsilon$ is to deal with edge cases when either the potential match has the exact same confounders or treatment dosage.

In the case with two treatments, matching can be solved using an optimal bipartite matching algorithm, such as the Hungarian algorithm [13]. This is because the vertices of the graph can be divided into two disjoint sets, those which represent units who received treatment, and those which did not, as units from the same group are not allowed to match, and has no edge in the graph. Here where we have multiple doses of treatment, and each unit can potentially be matched with any other unit, and is called an optimal non-bipartite matching problem. The algorithm used to solve the this problem can be attributed to Edmonds [6]. This algorithm generalises his original blossoming algorithm that constructs maximum matchings, to construct a maximum weighted matching on a graph.

We can now construct a graph, with each vertex representing a unit, and edges connected to other units where the weight is the distance given by equation 2. Using Edmonds algorithm we then can obtain an optimal matching $M$. Given a set of matched pairs $M$, and let $y$ denote the outcome and $z$ the treatment, the average treatment effect (ATE) per dose can then be calculated by

$$ATE = \sum_{(i,j) \in M} \frac{y_i - y_j}{z_i - z_j} \tag{3}$$

## 5 PROOF OF CONCEPT APPLICATION

We demonstrate the feasibility of using the matching procedure described previously to investigate sporting facility availability and its effect on antidepressant prescriptions on the average neighbourhood in London. We first describe the datasets, then the confounding variables, apply the matching procedure and assess the results.

### 5.1 Datasets

To obtain the data required to perform the matching, we combine eight datasets. These include *venue*, *prescription*, *GP*, and *drug* data, *ward boundaries* and *demographics*, the *Greater London area boundary*, and finally *postcodes* in London. All of these datasets are open.

*Venue Data.* The OpenStreetMap (OSM) project is a crowd-sourcing platform that aims to create a freely available and editable map of the world. The map contains data on a wide array of geographical features, such as shops, cycle routes, benches to waterways, which are contributed by users and moderated before being added to the

project. These features can be given a tag, where a list of official tags are maintained [7].

*Prescription Data.* The UK National Health Service (NHS) maintains an open dataset that contains all of the prescriptions, broken down at the practice level, across the whole of England. This dataset specifies which drug, identified using the British National Formulary[2] Code (BNF Code), and the quantity prescribed in GP's across England.

*GP and Drug Data.* Each month a BNF Code to chemical name mapping is published which allows the drug prescribed to be identified. The NHS also maintains a medical practice dataset which allows a practice to be located using its postcode from the practice code.

*Geographical Boundaries and Demographics.* The London data store [1] is a repository for data on different geographical separations in London and contains statistics such as population, housing benefits, crime rates, access to nature, to higher education results. Additionally, the geographical separations at the ward, borough or LSOA level is also available in GeoJson format.

Finally, The office of national statistics (ONS) maintains a postcode dataset that allows mapping between a postcode and different administrative or electoral areas, or latitude and longitude.

## 5.2 Confounding Variables

The confounding variables considered can be broken down into three categories. First of these is the population structure. London, as with many developed cities, is a commercially active area and different parts of the city attract people at different stages of their lives. The data segregates the population into those with age 0-15 (children), 16-64 (working age) and 64+ (retired). Plotting these values as a percentage of the population of each ward on a choropleth clearly shows that central London attracts the majority of the working age group, retirees mostly occupy the outskirts and children are mostly even distributed, albeit with higher concentration in more deprived areas. This is shown in Figure 1.

Additionally, the number of full time and part time employees in an area helps us characterise whether the area is more residential or commercial, which affects both sporting facilities and prescription numbers.

The second is the availability of green spaces. Despite London being one of the greenest capital cities on the planet, there are major discrepancies in the availability of green spaces. The demographic dataset captures this with two values, % area that is green space, and % homes with deficient access to nature. Access to nature not only has been linked with mental health [3] but its availability also determines the type of sporting venue that is available.

Finally, with SES being so commonly linked with a variety of health outcomes [2, 22], we also include several measures of deprivation, ranging from housing benefit claims to job seekers allowance claims. The full set of confounders used are shown in Table 1.

Of course, there are many points of discussion surrounding how we select confounding variables and decide on the causal structure and we defer to Section 7.1 where we touch on the key points. The results of applying the matching procedure is considered next.

[2]A pharmaceutical reference book published by the British Medical Association

| **Population Structure** |
|---|
| Ages 0-15 |
| Ages 16-64 |
| Ages 64+ |
| Full-time Employees |
| Part-time Employees |
| **Green Spaces** |
| % Area Green Space |
| % Homes with Deficient Access to Nature |
| **Deprivation** |
| DWP [3] Working-age client group (Rates) |
| Employment and Support Allowance Claimants |
| Housing Benefit Rates |
| Income Support Claimants |
| Incapacity Benefit Claimants |
| Job Seeker's Allowance Claimant Rates |

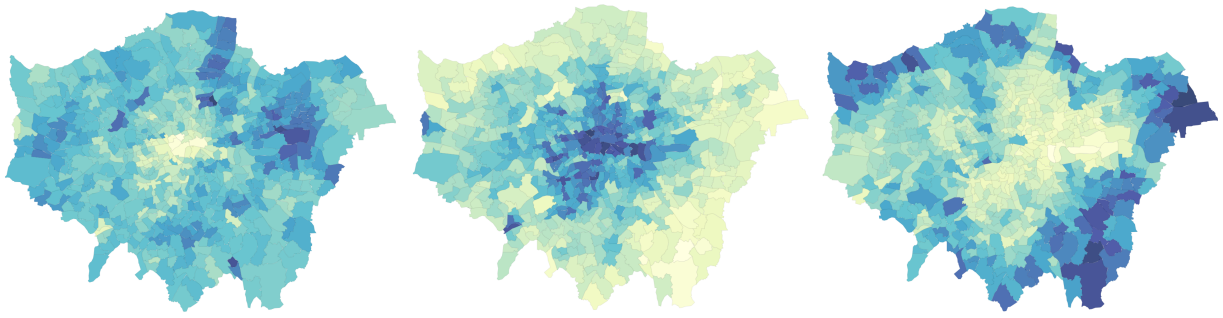
**Table 1: The set of considered confounders.**

## 5.3 The Treatment Effect

In this section we apply the matching procedure as described in Section 4 using the datasets and confounders identified above, to understand the impact of sport venues over antidepressant prescription. Doing this requires the following steps:

(1) Using the OSM data dump (the 'planet file'), the open source Osmium Tool, a boundary file of the Greater London Area and boundaries for London wards, we extracted all venues related to sports between the years 2011-2013 and mapped them to each ward in London using a geometry library. The venues were identified as sports-related through a list of key:value tags, the ones which are included were distilled from the official tag page [23] and are shown in Table 2. The extracted venues and its frequency are represented using the word-cloud shown in figure 2.
(2) To determine prescriptions, the prescription data was filtered by BNF code for antidepressants, the postcode data was filtered to only those in London, the GP data is then used to filter for London GPs and obtain a list of GP IDs with its corresponding prescription numbers for 2011-2013. The GPs were then mapped to each ward similarly.
(3) For each unit (ward) we now have the treatment (sport venues), outcome (antidepressant prescriptions) and associated confounding variables from the demographic dataset. We can now estimate the propensity score, $\beta_{x_i}^T$ in equation 1, calculate the distance using equation 2 and construct a graph to perform optimal non-bipartite matching with Edmonds algorithm.
(4) Given a matching $M$, we could calculate the average treatment effect as in equation 3. This is shown along with the distribution of the effects in the matched pairs in Table 3.

The units here are the change in antidepressant prescription per person, per dosage of sports venue in a ward in London. We can see



(a) Ages 0-15 (b) Ages 16-64 (c) Ages 65+

**Figure 1: A choropleth map revealing systematic bias in age structures in London. Darker colours indicate a higher density.**

| Key | Values Included |
|---|---|
| Sports | All Values |
| Leisure | Pitch<br>Golf Course<br>Swimming Pool<br>Sports Centre<br>Horse Riding<br>Track |
| Club | Sports |

**Table 2: Objects tagged with the key and values listed here were included in the extraction.**

| Year | ATE | Min | 25th % | 75th % | Max |
|---|---|---|---|---|---|
| **2011** | -0.461 | -91.579 | -7.651 | 5.774 | 65.911 |
| **2012** | -1.117 | -66.628 | -7.948 | 7.912 | 55.853 |
| **2013** | -0.529 | -90.034 | -7.573 | 6.454 | 72.320 |
| Null Model | | | | | |
| **2011** | 0.178 | -62.699 | -7.147 | 6.971 | 62.811 |
| **2012** | 0.494 | -68.740 | -7.245 | 8.495 | 62.291 |
| **2013** | -0.085 | -64.700 | -8.180 | 8.161 | 71.384 |

**Table 3: The distribution of the treatment effects in the matched pairs across 3 years, the average distribution on the null model, across ten runs, is shown below.**

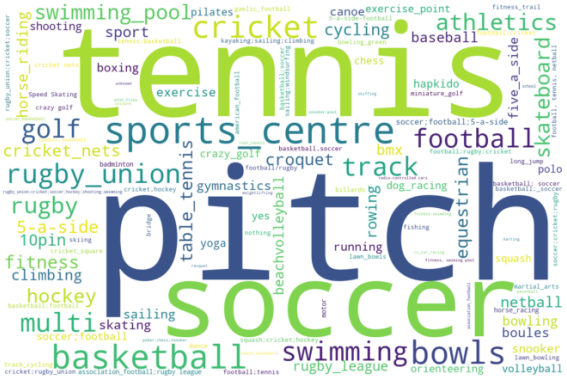



**Figure 2: Tag values from extracted sports venues, size is proportional to frequency**

that the ATE across all three years are negative. This suggests that there is a negative effect. If we compare this with the treatment effect distribution from matching on the null model - where the sport venue counts per wards were randomised - also shown in Figure 3, this also suggests that there is an effect in the negative direction.

Specifically, a unit increase in dosage level is expected, in the long run, to decrease the number of antidepressant prescriptions per person by anywhere between 0.461 to 1.117. With an average prescription level per person value of 23.51 across all wards across three years, this amounts to a decrease of 1.96-4.75% per year. If we looked at the simple correlation between the two variables, we could not discern as much, as shown in Figure 3.

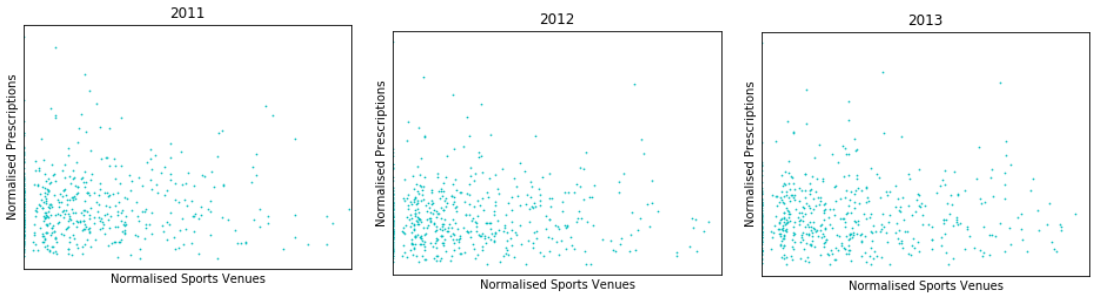
However this result deserves a deeper discussion. With such a relatively small magnitude, a more conservative interpretation would be that of a small negative, or no effect. This is due to the many possible uncertainties, both in the data and assumptions, that will be discussed in Section 7.2. If we look at the entire distribution of effects we can see that it spans a wide range of values, with the distribution skewed towards the negative direction, albeit slightly. This is shown in Figure 4.

In the next section we will explore some more aspects related to these results and assess their validity.

## 6 ASSESSING THE INFERRED EFFECT

How do we assess the reliability of the inferred effects? Since we do not have the 'true' effects to compare with, the effectiveness of the matching procedure in this instance is evaluated using three metrics: the dose difference, the confounder balance, and performance on a synthetic dataset.

Comparing the results to alternative methods used to elicit causal effects such as difference-in-difference or regression discontinuity would also be informative, however the available data does not satisfy the conditions for their use.




**(a) Spearman 0.0286, p-value 0.493** **(b) Spearman 0.00765, p-value 0.855** **(c) Spearman 0.0155, p-value 0.712**

**Figure 3: Simple correlation between the normalised number of sports venues and prescription across wards in London. This paints a very different picture to the obtained effects after matching.**

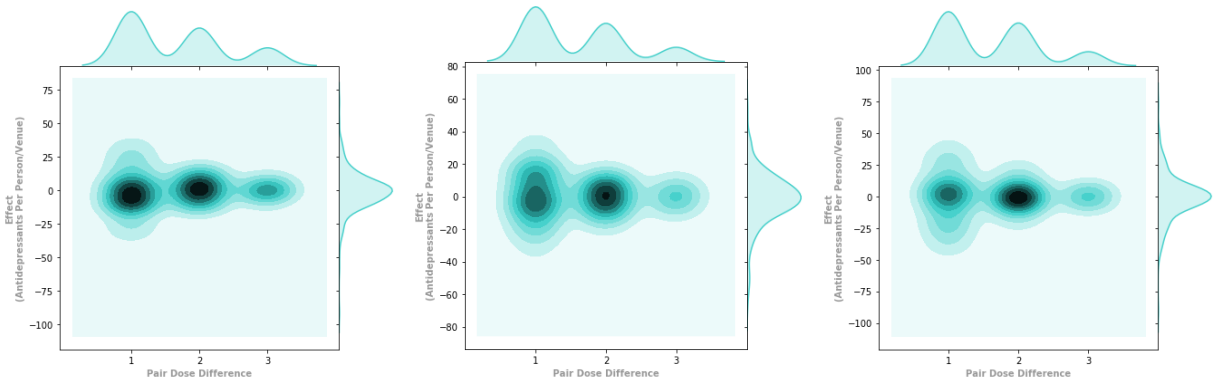
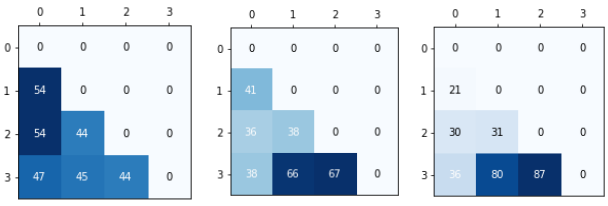



**Figure 4: Plot of the joint distribution between dose difference and effect of each matched ward pair. From left to right: 2011, 2012, 2013.**

## 6.1 Dose Difference

One of the intentions behind the form of the distance metric used, given by Equation 2, is to maximise the difference in treatment doses of each matched pair. This leads to reduced noise when computing the in-pair treatment effect.

Figure 5 shows the distribution of doses in each matched pair across the span of three years. Rows represent the dose of the unit with the higher level of treatment in the pair, and the columns represent the dose of the unit with the lower treatment. Darker colours indicate higher number of pairs with these doses.




**(a) 2011** **(b) 2012** **(c) 2013**

**Figure 5: Distribution of within-matched-pair doses between the high dose unit and low dose unit. Rows represent the high dose unit and column the low dose unit.**

Here one of the aims is to create pairs where the doses are not equal, we can see that this is achieved judging from the diagonals of the heatmap across three years. In fact, the difference in dosage is larger than 1 level in 50.7%, 49.0% and 51.2%, respectively. With approximately 35% of pairs having a difference in dose level of 2 and 15% with a difference of 3 dose levels across three years. The matching procedure has indeed created pairs with different treatment levels.

## 6.2 Confounder Balance

The original aim in any matching procedure is to produce different groups of units that have, on average, similar values for the confounders, such that the treatment effect is representative of the population. Table 4 shows the average value of the confounders between the high dose unit and low dose unit in a pair.

Here we can see that most of the confounders have successfully been balanced across the two groups. The exception is perhaps the normalised full time and part time employee rate, which has a difference of over 5 and 1.4 percent respectively. This is seen clearly by plotting the difference between the two groups as a percentage of the average of the original values, shown in blue in Figure 6.

On further investigation this can be attributed to a couple of extreme values for these two confounders in the ward data, which

| Confounder | High | Low |
|---|---|---|
| Population Aged 0-15 | 19.611 | 19.760 |
| Population Aged 16-64 | 68.991 | 69.182 |
| Population Aged 65+ | 11.398 | 11.058 |
| Greenspace Area | 27.272 | 26.448 |
| Homes with Deficient Access to Nature | 25.751 | 25.727 |
| Employment & Support Allowance Claims | 1.034 | 1.026 |
| Housing Benefit Claims | 12.871 | 12.863 |
| Income Support Claims | 3.393 | 3.374 |
| Incapacity Benefit Claims | 2.882 | 2.807 |
| DWP Working Age Clients | 14.363 | 14.307 |
| Part-Time Employees | 13.229 | 14.688 |
| Full-Time Employees | 34.089 | 39.827 |
| Job Seekers Allowance Claims | 5.623 | 5.569 |

**Table 4: The average normalised values of the confounders in the high dose and low dose groups across all matched pairs in 2011. Other years gave similar results.**

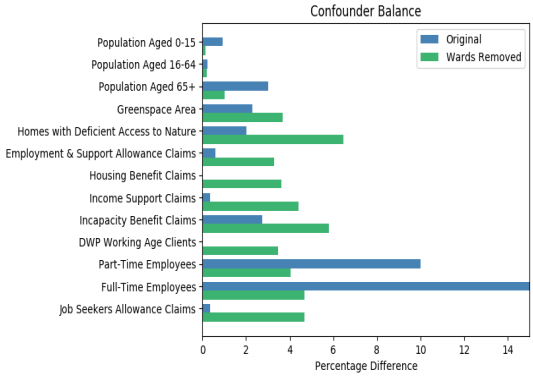



**Figure 6: The percentage difference between the confounders of the high and low dose groups with respect to the average of the original values. The green bars show the balance after the wards of St James, West End, and the City of London was excluded from the matching procedure.**

can be seen clearly in Figure 7, with particular wards having over three times the value of the remaining wards.

The three wards with the highest values for both confounders were predictably large commercial areas. These were the St James and West End wards of the Westminster area, containing Mayfair, Soho, large business districts, and the City of London. We may be tempted to remove these wards in the matching procedure, in an attempt to achieve more balance within the confounders. At first glance, doing so may seem a good idea, shown in green in Figure 6.

However, removing the wards, all of which are highly commercial areas of the city, induces a bias, as evidenced by the increased imbalance across the other confounders. The reason behind this may be that removing key business districts means that other business areas are forced to match with less similar areas, perhaps

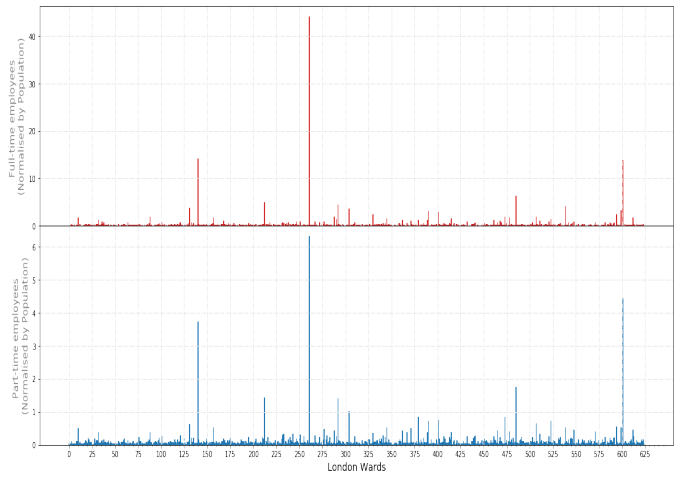



**Figure 7: A plot of full time and part time employees, normalised by population, for the wards in London. Here we clearly see at least three outliers.**

residential ones, generating additional imbalance. Additionally, remove key business districts mean that the average effect is now no longer representative of London as a whole, but rather London without the wards of St James, West End, and the City of London. Therefore for the entirety of London, we have to take into account the way that the balancing errors may have influenced the obtained effect in interpreting said effect.

## 6.3 Generating Synthetic Data

The most common benchmark in the literature for the evaluation of the causal effect estimate is the use of synthetic data [32][14][9][33]. Here, instead of constructing a completely arbitrary artificial dataset, we model the observed dataset and use this model to generate synthetic data. This means that we can generate datasets that mimic the real data, but where we are able to specify the underlying mechanisms - how the confounders affect both the treatment and the outcomes, and how the treatment affects the outcome. As a result, we know the 'true' causal effect of the treatment. We can then assess how well the method recovers this true effect.

More precisely, if we wanted to generate $N$ units $i = 1, ..., N$ (these are artificial 'wards' in our case) each with $p$ confounders (13 here) where we denote the confounders of unit $i$ as $\mathbf{x_i} = x_{i1}, ..., x_{ip}$, we first have to generate values for all confounders for each $N$ units. We then specify two functions, $f_Z(\mathbf{x_i})$ that takes in the confounders and generates a treatment (normalised sporting facilities) for a unit, and $f_Y(z_i, \mathbf{x_i})$ that take in the confounders and the treatment to generate an outcome (normalised antidepressant prescriptions) for that unit.

The distribution of the confounders from the real data is show in Figure 8. We model the confounders using a truncated normal distribution except for part time employees, full time employees, and deficient access to nature, which we model using an exponential distribution. The parameters of the confounder distributions was estimated using maximum likelihood.

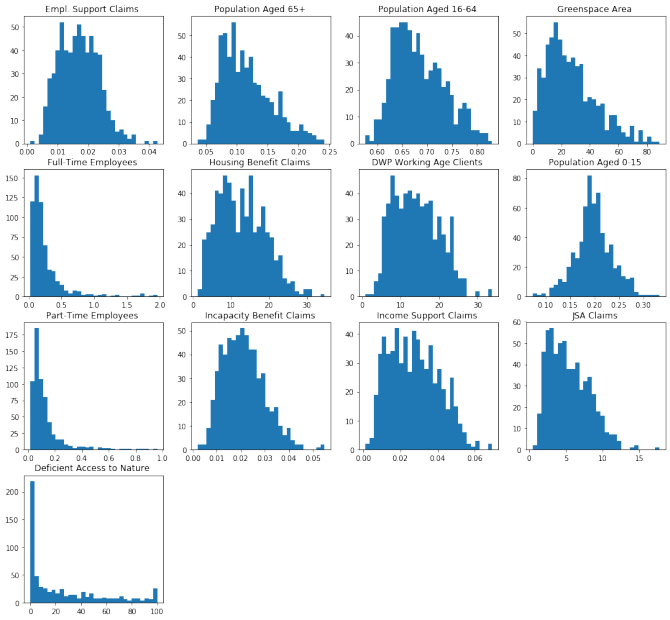



**Figure 8: Distribution of normalised confounders in London wards.**

Similarly looking at the distribution of sporting facilities, we then constructed $f_Z(\mathbf{x})$ as an exponential distribution:

$$f_Z(\mathbf{x}) = P(Z = z|\mathbf{x_i}) = \begin{cases} \frac{1}{f(\mathbf{x_i})} e^{\frac{z}{f(\mathbf{x_i})}} & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

where

$$f(\mathbf{x_i}) = \frac{\sum_p x_{ip}}{\alpha}$$

and $\alpha$ is tuned such that the treatments mimic the observed distribution.

The outcome for each unit, $Y_i$, is determined by $f_Y(z_i, \mathbf{x_i})$ and is constructed as a truncated normal distribution where if $Y > 0$

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

where the values for $\alpha$ and $\beta$ were tuned such that the data generated is within the observed range of outcomes in the real data, and $z_i$ is the treatment for unit $i$. Here, we know that the treatment effect is $\gamma$ per dose by construction.

Figure 9 shows a comparison between the observed distribution of the treatment and outcomes, and the distribution generated by the mechanism just described.

It is important to acknowledge that the functions chosen to generate the data is one of many possible functions that could potentially be the underlying mechanism behind this process. It is virtually impossible to tell which function is more likely than the other, besides the fact that it is able to generate data that closely mirrors the observed data.

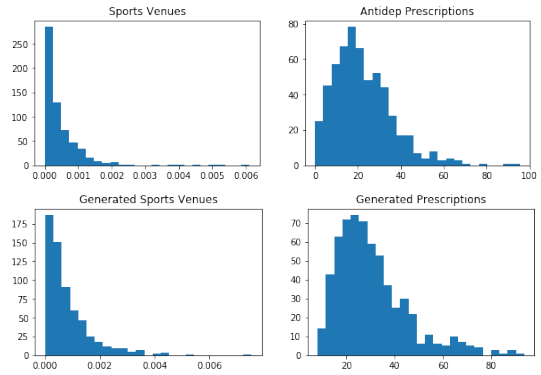



**Figure 9: Comparison between normalised sports venues and antidepressant prescriptions in the data and the generated dataset.**

## 6.4 Matching on Synthetic Data

Now that we could generate units along with its confounders, treatment, and outcome, we can run the same matching procedure used on the real dataset on the synthetic dataset. Since we know the true ATE of the synthetic dataset by construction, we can evaluate how the matching procedure performs.

The synthetic process was run to generate a total of 6250 units, each batch of 625 having a different set of confounders. This is done four times for four different 'true' ATE values, two for each direction, one with a small effect and one large effect. The results of the treatment effects calculated from the matchings are shown in Table 5.

| ATE | Est. ATE | Min | 25th % | 75th % | Max |
|---|---|---|---|---|---|
| -1 | -0.292 | -75.2 | -7.20 | 6.17 | 63.2 |
| -10 | -9.30 | -73.7 | -15.9 | -2.75 | 87.5 |
| 1 | 1.45 | -70.5 | -5.30 | 8.55 | 70.6 |
| 10 | 10.4 | -69.9 | 3.48 | 17.3 | 68.5 |

**Table 5: Treatment effects for the synthetic dataset. Each row was obtained from 6250 generated units, with each group of 625 containing a different set of confounders.**

Here we can see a similar outcome to the treatment effects obtained from the real dataset. The effects span a wide range, however the distribution is skewed towards the direction of the true effect. In the case where the magnitude of the 'true' effect is large, over 75% of all estimated effects lie in the correct direction.

The estimated effects also lie in the correct direction. We can see that in larger true effects, the matching procedure is able to give a better estimate. Where the effect is small, the inferred effects are less accurate. This gives some indication to how we should interpret the results obtained from real data, and since the obtained effects are small, we should err more on the side of a conservative interpretation.

## 7 DISCUSSION

We now turn our attention to discuss the potential sources of error in this approach. First we address the issue of confounding variable selection which was deferred from section 5.2 and then address the sources of error more generally.

### 7.1 Confounding Variable Selection

In the current causal inference literature there does not exist a method of testing which variables are relevant to a causal model for a particular problem from a set of possible variables. The closest that we have come across in this respect are methods in causal discovery [26], which at the current best can be used to find the equivalence class of directed acyclic graphs that could represent the class of models that potentially produced the observed distribution of variables, under some restricting assumptions, and given a pre-specified set of variables.

Therefore the current best option in determining confounding variables related to a causal question is to use existing subject matter knowledge, or literature relevant to the problem at hand. This is a fundamental problem of causal inference, where we cannot be certain whether we have identified all variables relevant to the cause and effect relationship at hand. It may be the case that in the future it is discovered, hypothetically, that people who smoke are severely more likely to develop depression, in which case the proportion of smokers in an area, or some variable that closes correlates with it, should additionally be considered a confounder.

An additional point to note is that of selecting key variables and edges that would have enough of an effect to be worth considering. Precisely how large is 'enough' still remains an open issue. People who have pets may indeed be slightly less prone to develop severe depression, but is the effect noticeable enough? Does the distribution of people who own pets consistently different in various areas? Systematic ways to answer these types of questions still remain unknown.

### 7.2 Sources of Error and Open Problems

The challenge with inferring cause is the amount of care that must be taken to eliminate all sources of errors throughout the entire process that could impact the results. In this project we by no means claim that this has been achieved, and discuss the potential sources of errors that we have conceived of here. The errors can be classified into three major categories; those that emerge from the assumptions about the causal structure, from the data itself, and from simplifying assumptions.

The most glaring source is that of the causal structure. Here the inferred effects is only a good estimate if we have included all of the confounders that are indeed relevant to the problem. This means that there are no unobserved factors that could influence the treatment and outcome. This is otherwise known as the strong ignorability assumption in Rubin's causal model, or called identifiability in Pearl's framework [25], and we have discussed above.

The second source of error are those from the data. Does the dataset of venues include all existing sports venues in London? If not, are they missing at random, or does it depends on some factor, such as the area? If it is the latter case, how do we assess this bias in exclusion from the dataset, and how do we incorporate this into an uncertainty of the estimated effect? The same issue applies with the prescription data. A potential direction forward is to assume the observed data is some fraction of reality and find the thresholds, under some additional assumptions, where the results significantly deviate from those obtained here.

Additionally, can the numbers in the dataset be taken at face value? For instance, the statistics on the demographics of a ward may have itself been generated by a statistical model. Even if we had the uncertainty attached to any particular value, how do we incorporate this uncertainty when inferring the effect?

The final major source of error are the simplifying assumptions. For instance, does the number of antidepressant prescriptions represent the rate of depression in an area well? Or does the exclusion of private prescriptions in the data create enough of a bias? The same question about incorporating uncertainties, if these were known, applies here. In general, issue of propagating uncertainties through some causal inference procedure is an interesting and open issue.

## 8 CONCLUSIONS

In this paper we have illustrated an approach to investigate the causal effects of the built neighbourhood environment on population health using open data at scale. This is important for several reasons. Firstly, it provides the potential to investigate the physical underlying mechanism of how a particular aspect of the neighbourhood influences health. Many higher level neighbourhood characteristics such as income or SES have been associated with multiple health outcomes, but how exactly does this affect health remains a question. Is it because people with higher SES eat healthily, exercise more often, or smoke less? Looking at the available services in a neighbourhood allows us to tackle these types of potential mechanisms. Secondly, the physical environment has the advantage of being better defined and quantifiable as opposed to more abstract measures such as general deprivation. Additionally, it provides a low cost and fast alternative to traditional methods of data collection, often with extra granularity. Lastly, these features of a neighbourhood are controllable from a policy point of view, as opposed to individual lifestyle choices, and is perhaps a more effective choice from a public health intervention standpoint.

Specifically, we have investigated the extent to which the availability of sporting venues have a causal role to play on the prescription of antidepressants in the neighbourhoods of London. Is the strength of evidence obtained and the magnitude of the indicated effect large enough to warrant a mass campaign encouraging sporting facilities? Probably not. However, it does give us insight into some important issues. With the preliminary knowledge of the possible effect size being rather modest, decision makers can now decide whether pursuing this question further, to obtain a better estimate of the effect, for example by collecting additional data or investigating to what extent the assumptions hold, is worth pursuing. It may be the case that the resources available should be put into other issues that are now known to be more cost effective, given the evidence obtained here. Or similarly, this method can be adopted to investigate related questions on venues and health.

It is hoped that this contributes to the conversation about urbanisation, and more generally, data-driven approaches in research tackling social and policy questions.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Damon Wischik for helpful discussions, and also thank the anonymous reviewers for their valuable and helpful comments.

A. Hasthanasombat's work is supported through the Cambridge Trust and King's College Cambridge, and he would like to also thank the Department of Computer Science and Technology for their additional support.

## REFERENCES

[1] 2018. London Datastore. https://data.london.gov.uk/dataset/lsoa-atlas

[2] Mariana C. Arcaya, Reginald D. Tucker-Seeley, Rockli Kim, Alina Schnake-Mahl, Marvin So, and S. V. Subramanian. 2016. Research on neighborhood effects on health in the United States: A systematic review of study characteristics. *Social Science & Medicine* 168 (Nov. 2016), 16–29. https://doi.org/10.1016/j.socscimed.2016.08.047

[3] Jo Barton, Rachel Bragg, Carly Wood, and Jules Pretty. 2016. *Green exercise: Linking nature, health and well-being*. Routledge.

[4] Alexandra Blair, Nancy A. Ross, GeneviÃĺve Gariepy, and Norbert Schmitz. 2014. How do neighborhoods affect depression outcomes? A realist review and a call for the examination of causal pathways. *Social Psychiatry and Psychiatric Epidemiology* 49, 6 (June 2014), 873–887. https://doi.org/10.1007/s00127-013-0810-z

[5] Emile Durkheim. 2005. *Suicide: A study in sociology*. Routledge.

[6] 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.

[7] Open Street Map Foundation. 2018. Map Features - OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/Map_Features#Route

[8] Panagis Galiatsatos, Cynthia Kineza, Seungyoun Hwang, Juliana Pietri, Emily Brigham, Nirupama Putcha, Cynthia S. Rand, Meredith McCormack, and Nadia N. Hansel. 2018. Neighbourhood characteristics and health outcomes: evaluating the association between socioeconomic status, tobacco store density and health outcomes in Baltimore City. *Tobacco Control* 27, e1 (July 2018), e19–e24. https://doi.org/10.1136/tobaccocontrol-2017-053945

[9] Xing Sam Gu and Paul R. Rosenbaum. 1993. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics* 2, 4 (1993), 405–420. https://doi.org/10.2307/1390693

[10] Desislava Hristova, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo. 2016. Measuring Urban Social Diversity Using Interconnected Geo-Social Networks. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. ACM Press, Montréal, Québec, Canada, 21–30. https://doi.org/10.1145/2872427.2883065

[11] Marshall M. Joffe and Paul R. Rosenbaum. 1999. Invited Commentary: Propensity Scores. *American Journal of Epidemiology* 150, 4 (Aug. 1999), 327–333. https://doi.org/10.1093/oxfordjournals.aje.a010011

[12] Markus Jokela. 2014. Are Neighborhood Health Associations Causal? A 10-Year Prospective Cohort Study With Repeated Measurements. *American Journal of Epidemiology* 180, 8 (Oct. 2014), 776–784. https://doi.org/10.1093/aje/kwu233

[13] Harold W Kuhn. 1956. Variants of the Hungarian method for assignment problems. *Naval Research Logistics Quarterly* 3, 4 (1956), 253–258.

[14] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns.. In *IJCAI*. 3768–3774.

[15] Bruce G. Link and Jo Phelan. 1995. Social Conditions As Fundamental Causes of Disease. *Journal of Health and Social Behavior* (1995), 80–94. https://doi.org/10.2307/2626958

[16] 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.

[17] Jens Ludwig, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. 2012. Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science* 337, 6101 (Sept. 2012), 1505–1510. https://doi.org/10.1126/science.1224648

[18] C. Mair, A. V. Diez Roux, and S. Galea. 2008. Are neighbourhood characteristics associated with depressive symptoms? A review of evidence. *Journal of Epidemiology & Community Health* 62, 11 (Nov. 2008), 940–946. https://doi.org/10.1136/jech.2007.066605

[19] Kate E. Mason, Neil Pearce, and Steven Cummins. 2018. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health* 3, 1 (Jan. 2018), e24–e33. https://doi.org/10.1016/S2468-2667(17)30212-8

[20] Peter McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)* 42, 2 (1980), 109–142. http://www.jstor.org/stable/2984952

[21] J. Michael Oakes. 2004. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. *Social Science & Medicine* 58, 10 (May 2004), 1929–1952. https://doi.org/10.1016/j.socscimed.2003.08.004

[22] J. Michael Oakes, Kate E. Andrade, Ifrah M. Biyoow, and Logan T. Cowan. 2015. Twenty Years of Neighborhood Effect Research: An Assessment. *Current Epidemiology Reports* 2, 1 (March 2015), 80–87. https://doi.org/10.1007/s40471-015-0035-7

[23] OSM. 2018. Key:sport - OpenStreetMap Wiki. https://wiki.openstreetmap.org/wiki/Key:sport

[24] Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. (2011), 8.

[25] Judea Pearl. 2009. *Causality*. Cambridge University Press. Google-Books-ID: f4nuexsNVZIC.

[26] Jonas Peters, Dominik Janzing, and Bernhard SchÃűlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press. Google-Books-ID: uie9AQAACAAJ.

[27] Jo C. Phelan, Bruce G. Link, Ana Diez-Roux, Ichiro Kawachi, and Bruce Levin. 2004. "Fundamental causes" of social inequalities in mortality: a test of the theory. *Journal of Health and Social Behavior* 45, 3 (Sept. 2004), 265–285. https://doi.org/10.1177/002214650404500303

[28] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (April 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[29] Andrew G. Rundle, Michael D. M. Bader, Catherine A. Richards, Kathryn M. Neckerman, and Julien O. Teitler. 2011. Using Google Street View to Audit Neighborhood Environments. *American Journal of Preventive Medicine* 40, 1 (Jan. 2011), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034

[30] Chris Smith-Clarke and Licia Capra. 2016. Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 425–434. https://doi.org/10.1145/2872427.2883076

[31] United Nations Social and Economic Affairs Division. 2018. World Urbanization Prospects.

[32] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. 2015. Causal Inference via Sparse Additive Models with Application to Online Advertising. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, Austin, Texas, 297–303. http://dl.acm.org/citation.cfm?id=2887007.2887049

[33] Fani Tsapeli, Mirco Musolesi, and Peter Tino. 2017. Non-parametric causality detection: An application to social media and financial data. *Physica A: Statistical Mechanics and its Applications* 483 (Oct. 2017), 139–155. https://doi.org/10.1016/j.physa.2017.04.101

[34] May C. Wang, Soowon Kim, Alma A. Gonzalez, Kara E. MacLeod, and Marilyn A. Winkleby. 2007. Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. *Journal of Epidemiology and Community Health* 61, 6 (June 2007), 491–498. https://doi.org/10.1136/jech.2006.051680