Validating methods for determining the causal effect of the neighborhoood built on public health

Group 1
Julia & John

8th February 2026

1 Summary of selected paper

This proposal report relates to the final project of the Urban Computing module at Leiden University, for the Computer Science Masters course. The presented paper is titled 'Understanding the Effects of the Neighbourhood Built Environment on Public Health with Open Data' by Apninan Hashthanasombat et al. It considers how the built environment of a neighbourhood (as opposed to the natural environment) affects the public health in that same neighbourhood. The paper focuses its examinations on London. The prevalence of sports venues in a neighbourhood is used as a proxy for built environment, and rates of prescribed anti-depressants in the neighbourhood, as a proxy for public health. The paper then looks to ascertain the relationship between sports venues present in a neighbourhood, and rates of prescribed anti-depressants in that same neighbourhood.

1.1 Approach overview

The paper initially gathers data on the two proxies described above, and looks at trends and relationships between the two variables in the data, grouped by neighbourhood. Where a 'neighbourhood' is defined through a 'ward', which is an old British sub-division of London into smaller areas [1]. The paper then looks at the confounding factors, which are meta-factors which can simultaneously affect both variables being looked at, and thereby add noise to any attempt to look at a causal relationship. The paper first identifies a set of relevant variables and confounding variables, and attempts to adjust for their effects. All analysis done in the paper beyond this step makes the assumption that all relevant confounding variables have been adjusted for. This means the set of variables being adjusted must make the effect being studied in the paper identifiable in the causal graph. The difficulty in the approach is actually identifying the confounding factors, which is a process that requires extensive knowledge on the subject-matter. It is furthermore extremely improbable to be able to obtain data on all the confounding factors.

1.2 Data description

The datasets used for this paper focus on London, The paper combines the use of eight datasets which are all publicly available. The **venue data**, in this case the various sports facilities, was obtained from the OpenStreetMap (OSM) project, which is a crowd-sourcing platform to create an editable map of the world with geographical features including the sporting facilities. The National Health Service (NHS) in England, makes available a dataset with **prescription data** broken down at the practice level. This dataset includes which drugs are prescribed, and at what dosage. The **GP and drug data** comes from a monthly updated dataset which allows the prescription data to be interpreted, as it comes with drug name encoding. **Geographical boundaries and demographics data** come from a publicly available dataset with information specific to London, including region borders, population statistics, housing, benefits, crime rates, and access to nature among others. A dataset is maintained by the **Office of national statistics** (ONS) with postcodes, which allows mappings between postcodes and different administrative or electoral areas, or through the use of longitude and latitude specifications.

1.3 Methodology

The paper is trying to identify a causal relationship between the "treatment" which in this case is represented by sporting facilities in a neighbourhood, and a population health outcome, in this case the amount of anti-depressants prescribed by general practitioners in the same neighbourhood. The best way to test for a causal relationship is to performed a randomised controlled experiment, which would involve stripping half the wards of their treatment (sports facilities), and observe the difference between the changed half and the unaltered half. This is not possible or ethical, the paper aims instead to emulate a randomised controlled experiment, by balancing the factors which might be affecting the outcome across wards, and between experimental sub-groups. There are also factors which affect both the treatment level, and the outcome (anti-depressant prescriptions), these variable that cover both ends of the experiment are called "confounding variables". To adjust for these two sets of variables, each ward is assessed in terms of its identified variables and treatment level, and wards are "matched" so that as far as possible the biggest difference between ward pairs are just the treatment level.

1.4 Conclusions

The paper concludes that the slight negative correlation between the treatment (more sporting facilities) and anti-depressant prescriptions in London, is not sufficient to identify a causal relationship. The negatively correlation is not particularly strong, but it can be observed throughout the matched pairs of wards. The paper further concludes that the methodology works, and is generally applicable when considering causal relationships. The paper does suggest that the research question would benefit from having more data, and that the results are not conclusive in terms of neighbourhood built environment's relationship to anti-depressant prescription rate.

2 Summary of proposed paper

2.1 Problem statement

The results presented in the paper fail to confirm a causal effect. However, we do not know if there exists a causal effect between the studied variables and thus if the method failed to extract the effect or that it accurately predicted that there is none. Therefore, the proposed method lacks validation. It remains a problem in the study of the influence of the environment on public health to find such a causal effect in urban data. Studies focusing on correlation rather than causal effect are more common. However, knowledge about direct causal relationships could greatly help develop public policy to improve the public health. And so the problem of identifying causal factors, and methods to identify them, remains relevant. We do not know if the issue is that the methods used are faulty or because there simply is a small or no causal effect between the studied variables. There is therefore no way to verify the obtained result, and thereby no way to validate the proposed method. Therefore, we propose to validate the proposed method by applying them to data in which a significant causal effect has been determined. Since the authors of the paper concluded that the proposed method would work better for data in which there is a strong causal effect, ideally, we would look for a paper which finds such an effect in a dataset.

2.2 Research question

Does the proposed method for finding causal relationships between the built environment and public health succeed in finding the causal relationship for a dataset in which a strong causal effect has already been confirmed by another study?

2.3 Methodology

To answer the research question, we propose to use the exact same methods as proposed in the paper, but with different data. These methods include simulating Randomized Controlled Trials by running an observational study. This is achieved by dividing the wards into two groups with comparable outcomes but differing treatment status. Then, the wards are matched in pairs such that confounding variables, which consist of ward variables other than the treatment status (number of sports venues) and effect (number of prescriptions), match as well as possible, but the difference in treatment is as large as possible. Here, the problem of dimensionality arises. This problem is addressed by calculating a propensity score for each ward, which is a function of the confounders. The propensity score is used to perform matching with multiple treatment levels, which allows to match wards taking different numbers of sports venues into account. This is the preferred alternative to binary matching, in which the treatment status would be either 0 or 1. Now, with the matching completed, the average treatment effect (the effect of the number of venues on the number of prescriptions) can be calculated. This is done across all wards, per year, resulting in results for the years 2011, 2012, 2013. The obtained effects cannot be compared to a true effect, thus a null model is constructed. This model is constructed by producing synthetic ward data. The synthetic treatment is sampled from a distribution that mimics the observed distribution. The outcome (the number of prescriptions), is sampled from a distribution that is a function of the treatment. This makes it possible to enforce a certain level of average treatment effect, that can be used to validate the method. The data we will be looking for, is by preference a study of the causal effect of the (built) environment on public health in which a strong causal effect has been found. Ideally, this data would also be from London or another large city. We will apply the same methods on the new data and evaluate the results. If time allows, there are several factors of the methods used in the paper of which we could study the alteration. An example would be to alter the calculation of the propensity score.

2.4 Evaluation approach

- **Metrics:** To obtain the best comparison, the same values will be calculated as in the paper. These are the average treatment effect per year, its minimum and maximum value, and the first and third quantile. The second metric is a simple correlation between the normalised amount of treatment (sports venues in the paper) and effect (prescriptions in the paper). The third metric would be joint distributions between dose difference and effect of each matched ward pair. Finally, a statistical analysis needs to be performed of the difference between the distributions presented in the paper and the distributions we will obtain, using a Z-test. Other appropriate statistical methods will be selected depending on our observation of the results.

- **Baselines:** The baseline used will be the results from the paper from which the new data is obtained.

2.5 Data sources and other resources

The data for this project still needs to be obtained. A suitable data source could be [2], a study of the causal effect on older adult's physical activity and well-being. This study contains data of a period of 1 year of several urban wards in Greater Manchester. This means that it would not be possible to compare multiple years. The data is from a different region, but from within the same country as the paper that is studied.

We will contact some researchers to attempt to find other suitable data sources.

The code and data of the original data are available. We are in contact with the authors of the paper and hope to have access to these soon.

References

[1] City of London Parishes Wards. 2015. GENUKI. Accessed on October 17, via https://www.genuki.org.uk/big/eng/LND/parishes

[2] Benton, Jack and Anderson, Jamie and Cotterill, Sarah and Lindley, Sarah and Dennis, Matthew and French, David. 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, volume 18. DOI: 10.1186/s12889-018-5812-z.