Upscaling a Spatial Survey with Propensity Score Matching: Implications of a Motorbike Ban in Hanoi

Nick Malleson*[1,2], Kristina Bratkova[2], Alexis Comber[1], Phe Hoang Huu[3], Minh Kieu[4], Thanh Bui Quang[5], Hang Nguyen Thi Thuy[6], and Eric Wanjau[2]

[1]School of Geography, University of Leeds, UK
[2]Leeds Institute for Data Analytics, University of Leeds, UK
[3]R&D Consultants, Hanoi City, Vietnam
[4]Faculty of Engineering, University of Auckland, New Zealand
[5]Faculty of Geography, VNU University of Science, Hanoi, Vietnam
[6]VNU Vietnam Japan University, Vietnam National University, Hanoi.

January 17, 2022

Summary

The city of Hanoi, Vietnam, suffers high levels of congestion and air pollution as transport infrastructure has failed to keep pace with a rapidly growing population. This paper presents the use of *propensity score matching* as a means of up-sampling a bespoke travel survey through linkage to census microdata. Among other uses, it is hoped that these new data will begin to shed light on peoples' perceptions of potential transport-related policies, such as a ban on motorbikes from parts of the city centre.

KEYWORDS: Propensity score matching; population synthesis; transport policy; Hanoi

1 Introduction

In many rapidly-developing cities in the Global South, such as Hanoi, Vietnam, transport infrastructure is failing to keep pace with the burgeoning population. This can lead to high levels of congestion, air pollution, and a broad inequity in the accessibility of large parts of the city to residents. This paper presents the preliminary outputs of a new programme of work that is being developed to provide evidence that policy makers can use to better understand and improve transport systems. Specifically, this paper uses the technique of *propensity score matching* to upsample a bespoke travel survey (conduced by the project) by linking it to census microdata. Among other uses, it is hoped that synthetic survey data will begin to shed light on peoples' perceptions of potential transport-related policies, such as a ban on motorbikes from parts of the city centre.

*n.s.malleson@leeds.ac.uk

Specifically, the initial results begin to highlight parts of the city where a ban might be the most disruptive or controversial.

2 Background

In Hanoi, over 90% of the vehicles driven are motorbikes (Ngoc et al. 2017). This compares to approximate 4% in England[1]. Since the introduction of the Doi-Moi policy (Hansen 2017) in the 1980s, the number of motorbikes has increased 10-fold and there are now more than 4 million motorbikes in Hanoi alone (Hansen 2016, 2017). Despite the creation of the a new Hanoi Metro in 2021, on the whole public transport infrastructure has developed slowly which has lead to inevitable increases in personal traffic. This has resulted in chronic pollution at times, with PM2.5 and ozone concentrations regularly exceeding safe levels. In response, the City has developed fast buses, a skytrain system, tightened the standards for new vehicles and imposed petrol quality controls. Some officials have also proposed a radical plan to ban motorbikes in large parts of the city, but this was met with strong public opposition.

3 Data & Methods

Travel Survey In an attempt to support evidence-based policy making for the Hanoi transport system, the *Urban Transport Modelling for Sustainable Well-Being in Hanoi* project[2] created a bespoke household travel survey that asks people for basic demographic information as well as details about their travel behaviour (e.g. common journeys) and preferences (e.g. aspirations for ownership of different types of vehicle). It also asks questions specifically related to a possible ban on motorbikes from the city centre. The COVID-19 pandemic has interrupted the survey on multiple occasions, but at the time of writing $1,500$ households, out of a target of $10,000$, have responded. This is not yet a sufficient number to produce robust results, hence this paper presents preliminary findings.

Upscaling the Survey with Propensity Score Matching (PSM) Even when the full survey has been conducted, there will still be relatively few respondents relative to the full population of Hanoi (8M people) and the survey will inevitably suffer from some socio-demographic and spatial biasses. Fortunately, the project also has access to a sample of micro-data from Vietnam's most recent population and housing census, conducted in 2019 (General Statistics Office 2020). Hence there is an opportunity to use the census as a means of up-scaling the survey, both to artificially increase the number of 'responses' as well as reducing some of the biasses in the survey.

Propensity Score Matching (PSM) is commonly used in medicine and other fields. It attempts to adapt observational studies, in which participants are not randomly sampled, in a way that mimics experimental studies with participants that *are* randomly sampled. In effect, PSM attempts to divide the population in to two groups – those individuals who have received a treatment and those

[1] According to the 2016 National Travel Survey there were 1.1M licenses motorbikes in England compared to approximately 26M licensed cars.

[2] `https://urban-analytics.github.io/UTM-Hanoi/`

that have not – such that the characteristics of the people in the two groups are nearly identical. Therefore the difference between treatment and control groups can be attributed to the treatment itself, rather than to the differences in the characteristics of the two groups that would have been caused by non-random sampling. Here, the two initial groups, before PSM, are (i) individuals in the survey and (ii) individuals in the census.

The first step in PSM is to calculate a *propensity score*, that is the "probability of treatment assignment conditional on observed baseline characteristics" (Austin 2011). This paper follows Morrissey et al. (2015) and uses the score as a means of identifying individuals in the travel survey who share similar characteristics to individuals in the census, as illustrated in Figure 1. The propensity score is calculated for all individuals (those in the census and those in the survey) and then the nearest neighbour algorithm is used to iterate over all census individuals and find individuals in the survey who have a similar score. In this manner each census individual is matched to a survey individual, so (if successful) the survey is up-sampled and biasses smoothed.

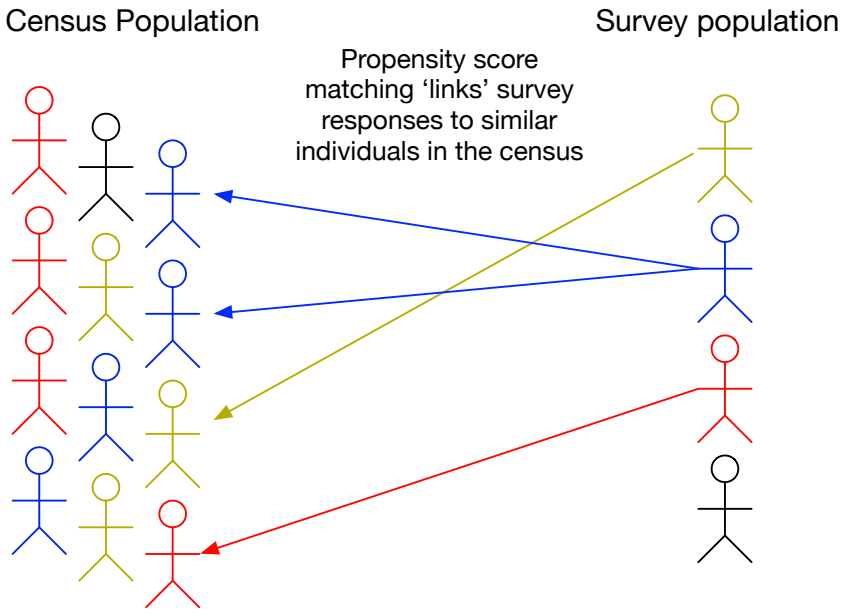

Figure 1: Example of Propensity Score Matching

It is important that a wide array of demographic variables are used to calculate the propensity score so that individuals with different background characteristics are adequately distinguished. Here, currently the score is calculated using sex, age (six groups) and house ownership (owned, rented, other). In future the score will also include information about education levels and spatial location.

The code to conduct the matching is implemented in Python[3]. Following Luvsandorj (2021), the

[3]The code is not currently available as it is stored in a secure environment with the sensitive data, but the research team are committed to open science principles and intend to release the code without the sensitive data when possible.

propensity score is calculated using a logistic classifier in the scikit-learn library and the nearest-neighbours calculation is conducted using the scikit-learn `NearestNeighbors` class.

4 Results

As noted in Section 3, survey collection is ongoing. Currently there are too few surveys for robust results, even after upscaling using propensity score matching. Therefore this section presents preliminary results that are indicative of the kinds of analysis that will be undertaken once a sufficient number of surveys have been returned. As an example therefore, Figure 2 presents the proportion of people in the synthetic (upscaled) survey who are aware of the potential motorbike ban and Figure 3 illustrates their average opinion, where lower numbers are indicative of a lower opinion of the ban. It appears that people in the city centre are more likely to be aware of the ban (Figure 2) which is probably to be expected as they will be the most frequent users of the city centre. Opinion about the ban (Figure 3) appears to be more randomly dispersed, with no clear clusters in favour or against. This neighbourhood variation is worth exploring further. In addition there will be numerous important confounding factors that influence a person's opinion of the ban – for example whether someone owns a motorbike – that will be explored as well.

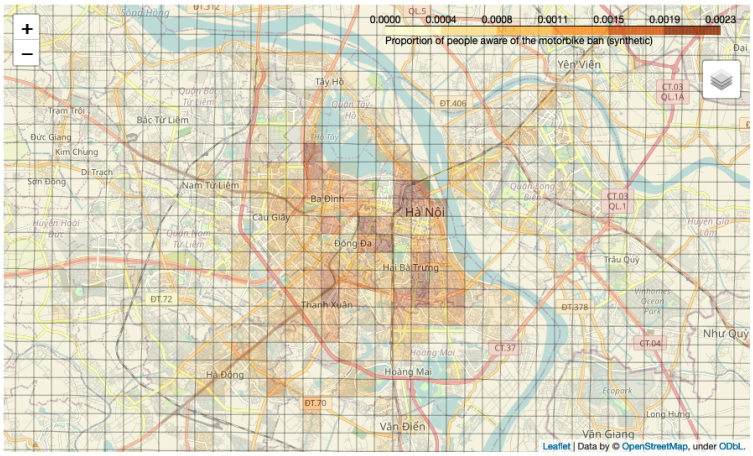

Figure 2: Proportion of people in the synthetic survey who are aware of a potential motorbike ban

5 Conclusion

This paper has presented some preliminary work from the *Urban Transport Modelling for Sustainable Well-Being in Hanoi* project. In particular, it has upscaled a bespoke household travel survey and begun to explore resident's opinions on a potential motorbike ban. Although the pandemic has limited the number of surveys that have been returned, and there are currently too few for rigorous analysis, we anticipate increased survey collection as Hanoi begins to leave strict pandemic-related restrictions. Future work will involve a more thorough analysis of the synthetic data as well as an exploration into many of the confounding factors.

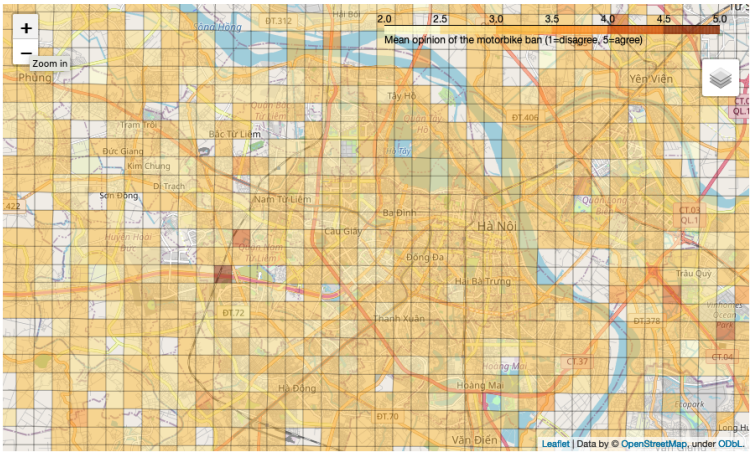

Figure 3: Mean opinion on the motorbike ban (lower numbers mean less favourable opinion)

6 Acknowledgements

This work has received funding from the British Academy under the Urban Infrastructures of Well-Being programme [grant number UWB190190].

References

Austin, P. C. (2011), 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies', *Multivariate Behavioral Research* **46**(3), 399–424.

General Statistics Office (2020), *Completed Results of the 2019 Viet Nam Population and Housing Census*, Statistical Publishing House, Vietnam.

Hansen, A. (2016), 'Hanoi on wheels: Emerging automobility in the land of the motorbike', *Mobilities* pp. 1–18.

Hansen, A. (2017), 'Transport in transition: Doi moi and the consumption of cars and motorbikes in Hanoi', *Journal of Consumer Culture* **17**(2), 378–396.

Luvsandorj, Z. (2021), 'Propensity Score Matching: Beginner's guide to causal inference from observational data'.

Morrissey, K., Clarke, G., Williamson, P., Daly, A. & O'Donoghue, C. (2015), 'Mental Illness in Ireland: Simulating its Geographical Prevalence and the Role of Access to Services', *Environment and Planning B: Planning and Design* **42**(2), 338–353.

Ngoc, A., Hung, K. & Tuan, V. (2017), 'Towards the Development of Quality Standards for Public Transport Service in Developing Countries: Analysis of Public Transport Users' Behavior', *Transportation Research Procedia* **25**, 4560–4579.