Up-scaling a Spatial Survey with Propensity Score Matching –

Towards larger-scale analysis

Nick Malleson, Eric Wanjau, Alexis Comber, Kristina Bratkova, Hang Nguyen Thi Thuy, Thanh Bui Quang, Phe Hoang Huu and Minh Kieu

Context

- We have our survey (~30,000 people)
- But that is only ~0.6% of the Hanoi population
- Issues with bias and sparse geography
- **Would like to estimate how the survey results might vary across the city**

Aims

- Upscale the survey to make it more representative (larger sample and less bias)
- Combine the census microdata and the survey to create a rich, synthetic population
- Better understand the possible implications of a motorbike ban
- (Work in progress!)

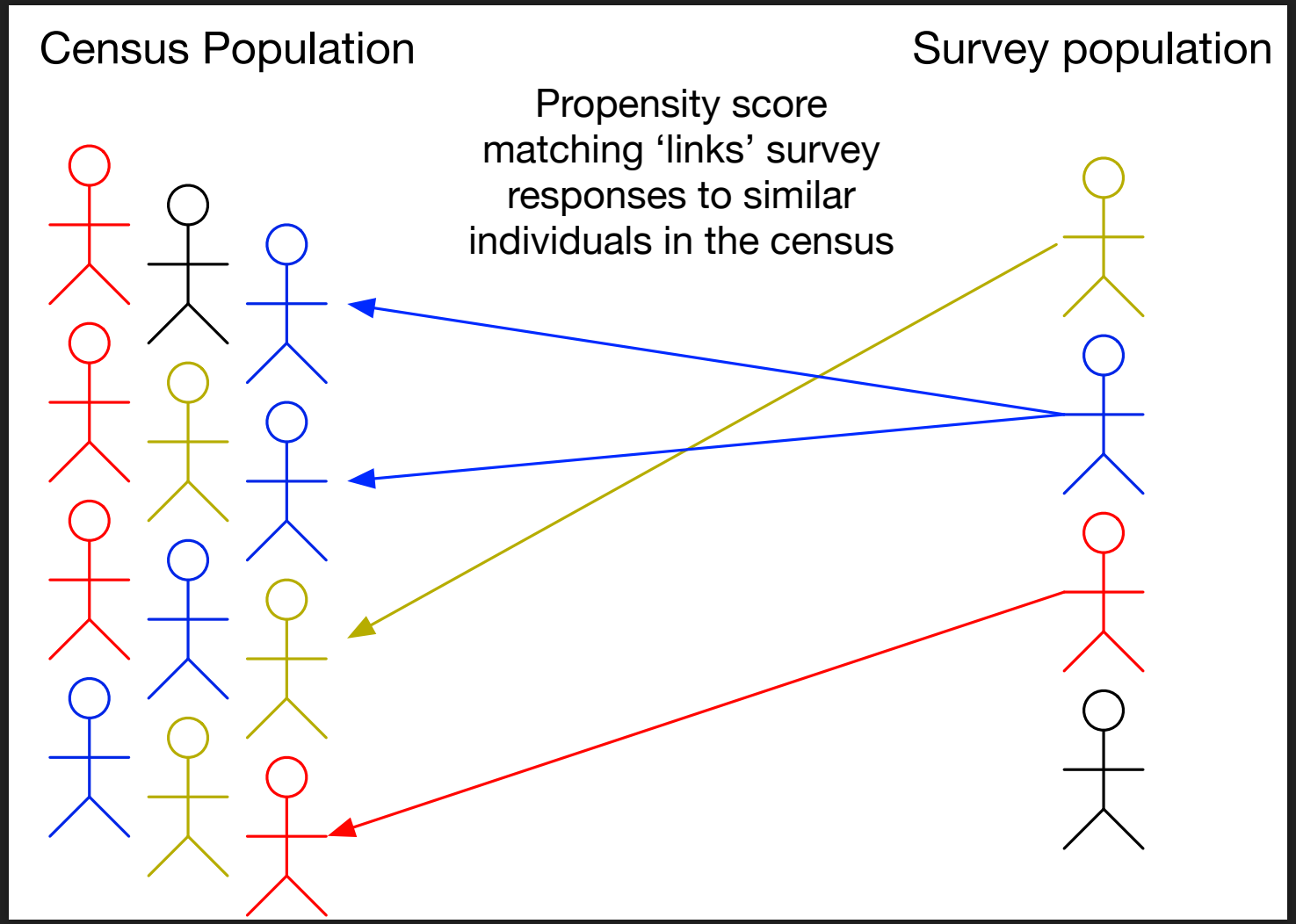
(Synthetic Populations)

- Can combine *aggregate* data with surveys to create individual-level *synthetic populations*
- Strong expertise @ University of Leeds

- But here we have the census micro-data so no need to create a synthetic population

Method: Propensity Score Matching (PSM)

Find individuals in the survey who are similar to individuals in the census



Propensity Score Matching

- First calculate the **propensity score**
- Common in medicine
- Converts *observational studies* (with non-random sampling) to *experimental studies*
- Tries to balance two groups — 'control' and 'treatment' — so that they have similar characteristics.
- Allows differences to be attributed to the effect of the treatment, rather than to differences in the two groups

Linking method

- **We don't care about a 'treatment', we just use the score as a way to link the two groups**

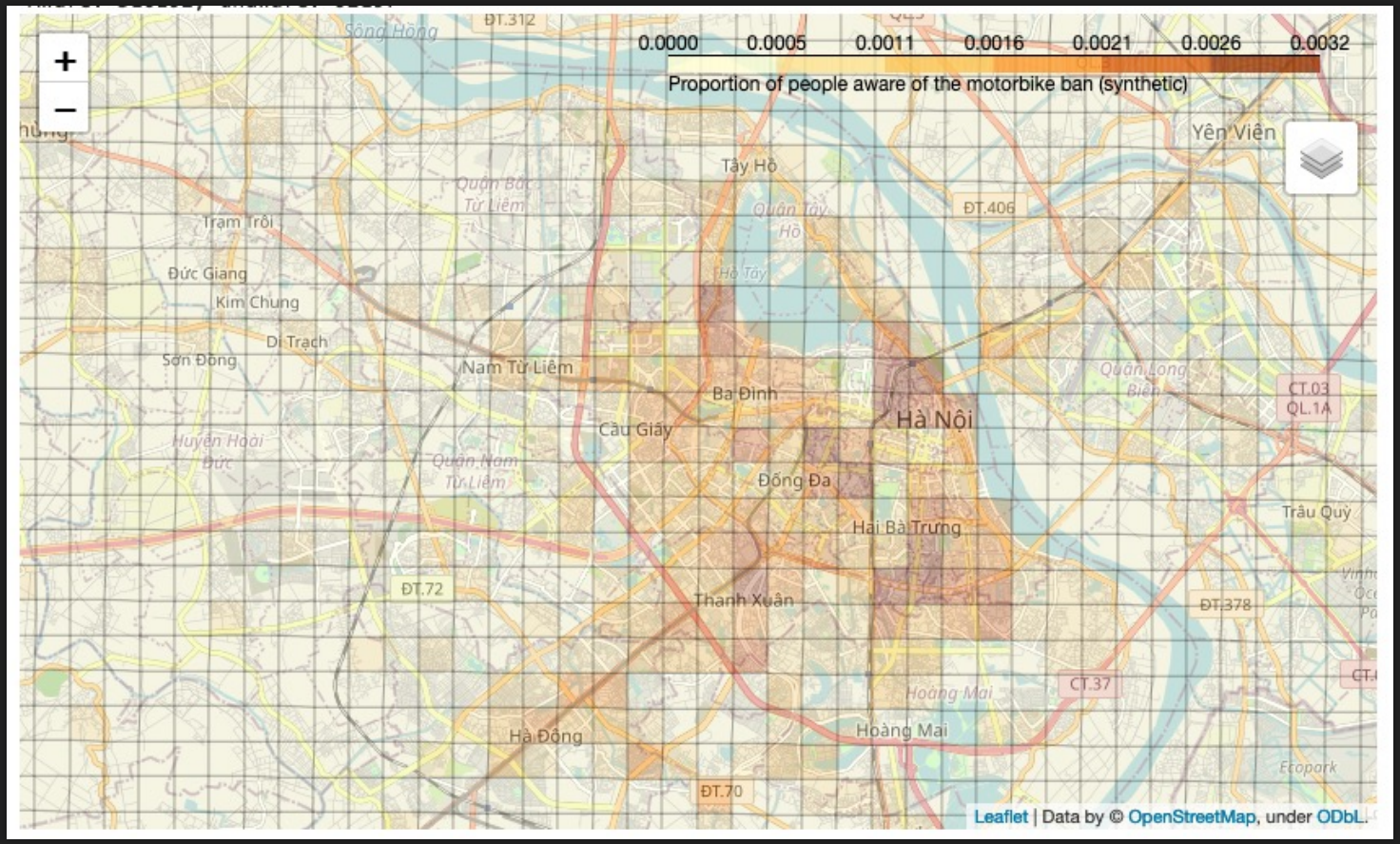
- Current shared attributes:
  - Sex
  - Age (6 groups)
  - House ownership (owned, rented, other)
  - Future work: more! (including geography)

Linking method

- Following Morrissey et al (2015) and Spooner (2021)
- 1: Assign treatment (census) and control (survey) groups
- 2: Calculate the propensity score
  - "probability of treatment assignment conditional on observed baseline characteristics" (Austin 2011)
  - "most often estimated using a logistic regression model, in which treatment status is regressed on observe baseline characteristics" (Austin 2011)
  - Here we use a logistic classifier in scikit-learn (Luvsandorj, Z., 2021).
- 3: Nearest-neighbour algorithm selects individuals in the survey who are close to those in the census
  - Using scikit-learn Nearest Neighbors class.

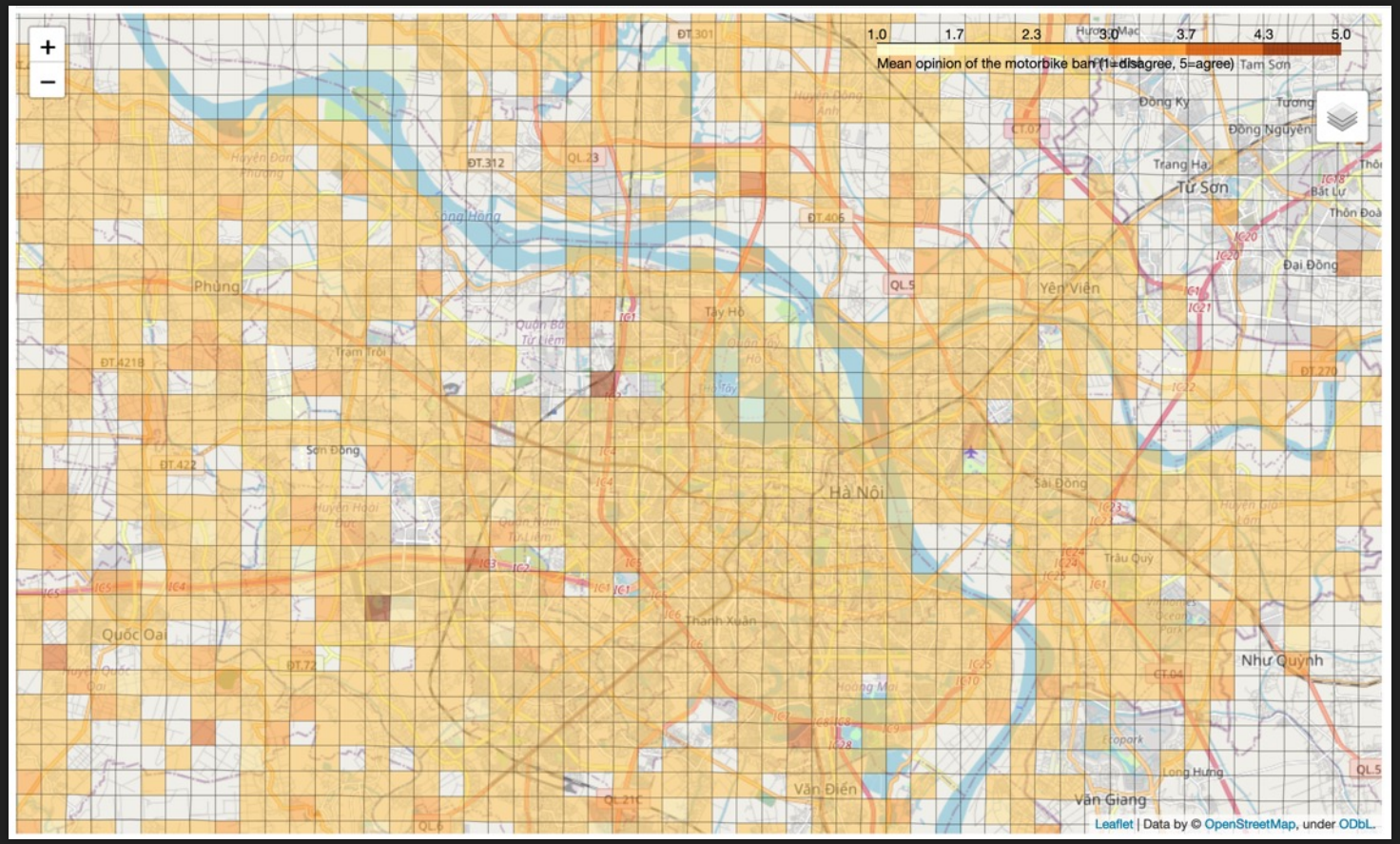
Preliminary Results (1)

Awareness of a possible motorbike ban



Preliminary Results (2)

Opinion on the possible ban



Summary & Conclusions

- Better understand residents' transport opinions and behaviours
- Use propensity score matching to up-scale a travel survey
- Explore awareness and opinion on a motorbike ban
- **CAVEAT**: Currently too few factors considered, links between the census and the survey are not sufficiently nuanced

Next steps:

- Improve census-survey link to be more detailed
- Take spatial location into account
- Other features of the survey to explore: e.g. aspirational vehicle ownership, journeys, public transport, etc.