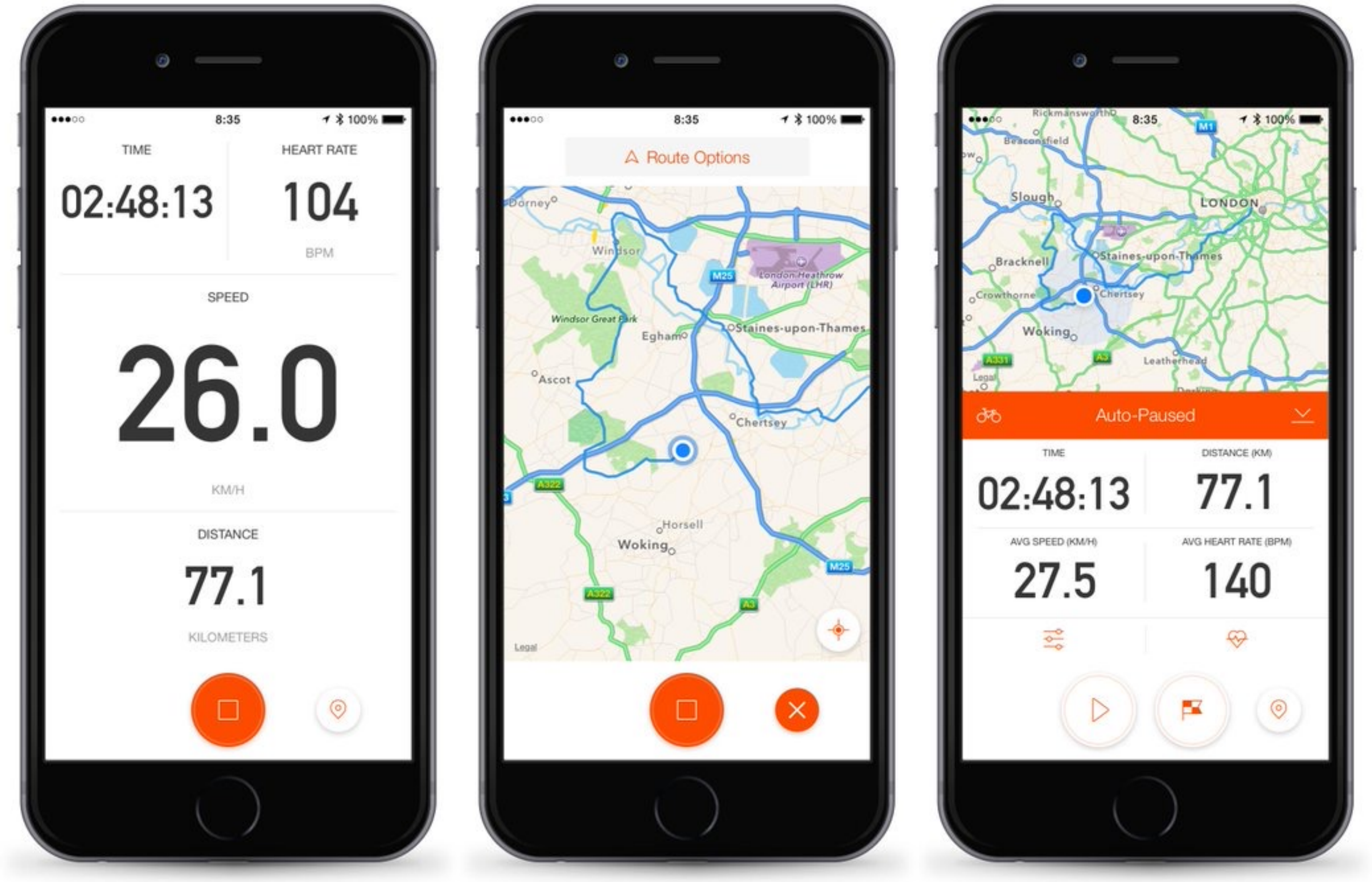
Predicting network wide cycling ridership with crowdsourced data

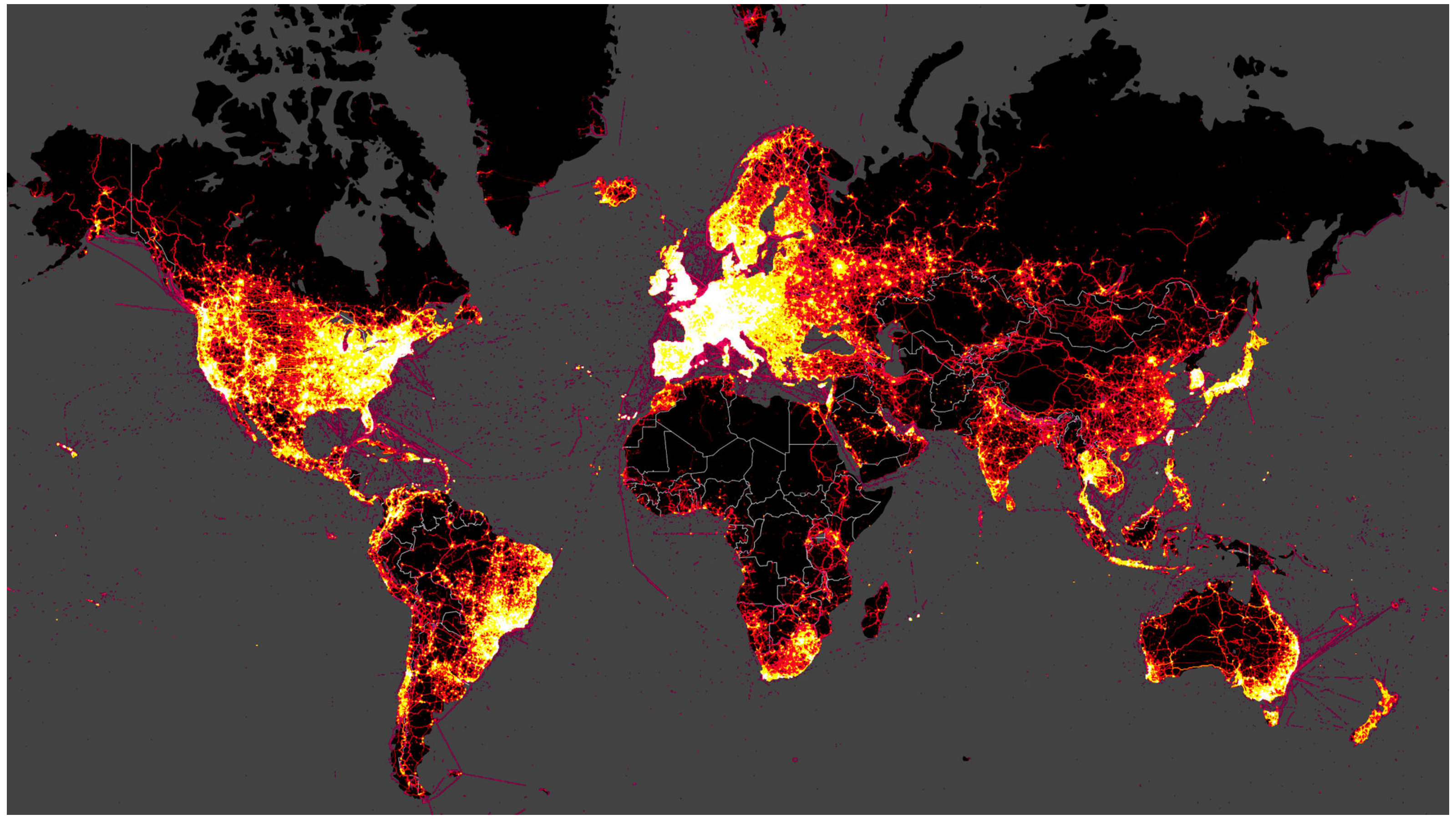
Dr Meead Saberi
UNSW Research Centre for Integrated Transport Innovation (rCITI)
@meeadsaberi
meead.saberi@unsw.edu.au

What problem are we aiming to solve?

- Growing awareness of health, environmental, and economic benefits of active transport have led to an unprecedented commitment to cycling policies.
- Lack of robust and continuously collected bicycling volume data is a substantive barrier to planning and safety studies, which require exposure data.
- We have a limited knowledge on how bicycling volumes flow throughout the network.

Crowdsourced cycling data: Strava

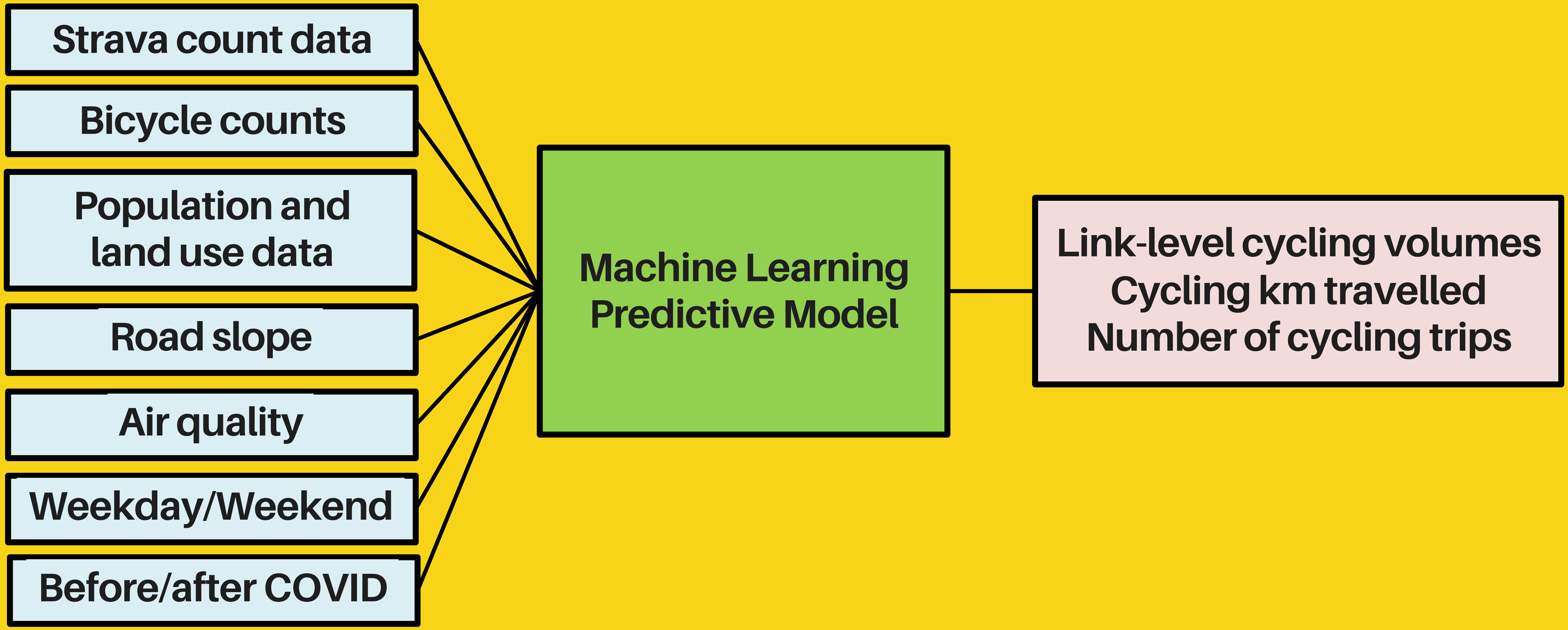




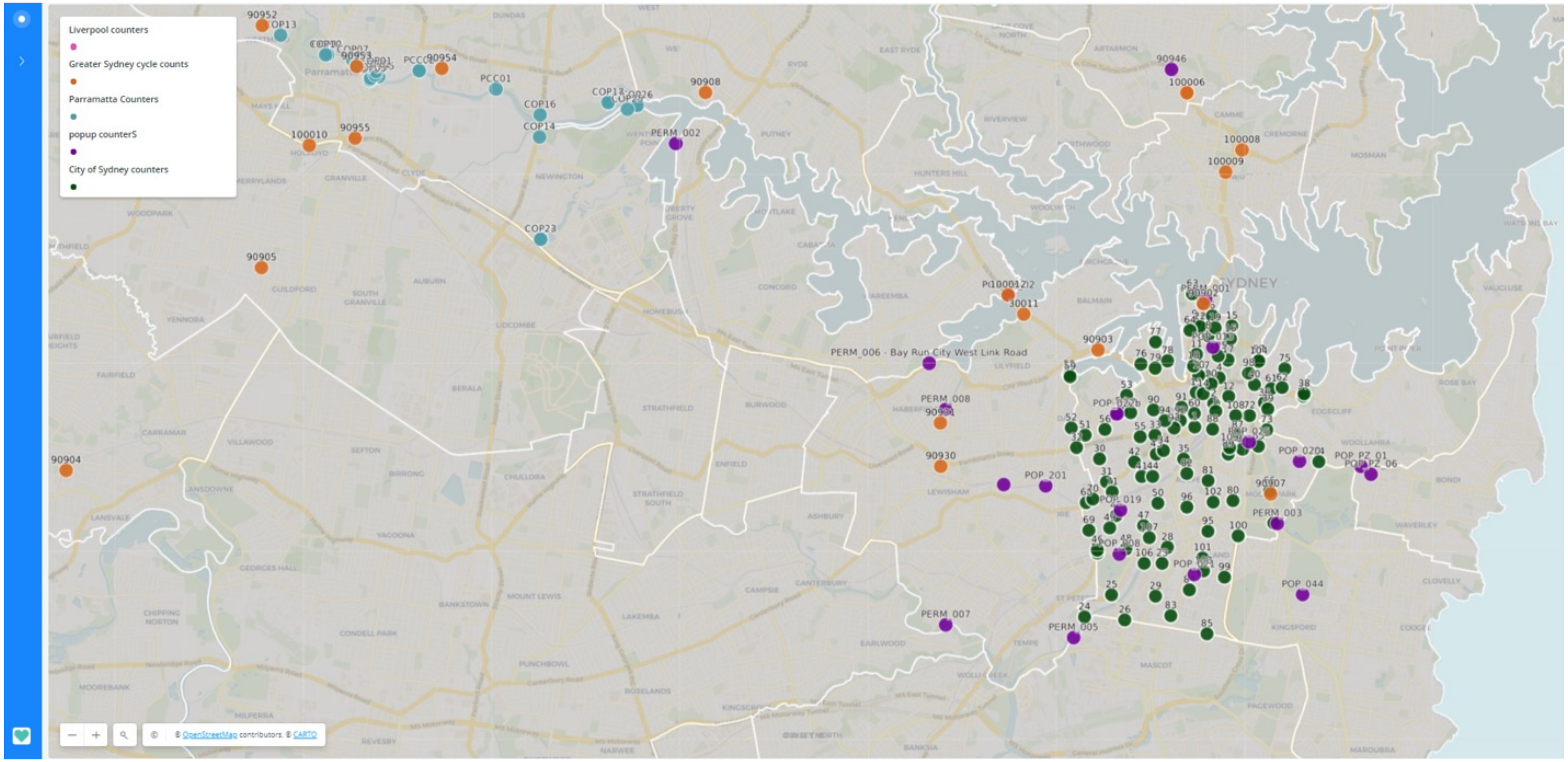
How to solve the sampling bias in data?

- Strava data is only a sample of bicycling ridership, biased toward people who use the Strava app.
- Strava users are disproportionately young adults (25-35 years in age) and male (Roy et al., 2019). Women, children, older adults, and low-income bicyclists are under sampled by Strava data.
- By integrating Strava data with multiple data sources it is possible to generate maps of predicted total bicycling volume that are more representative of all ages and abilities of bicyclists.

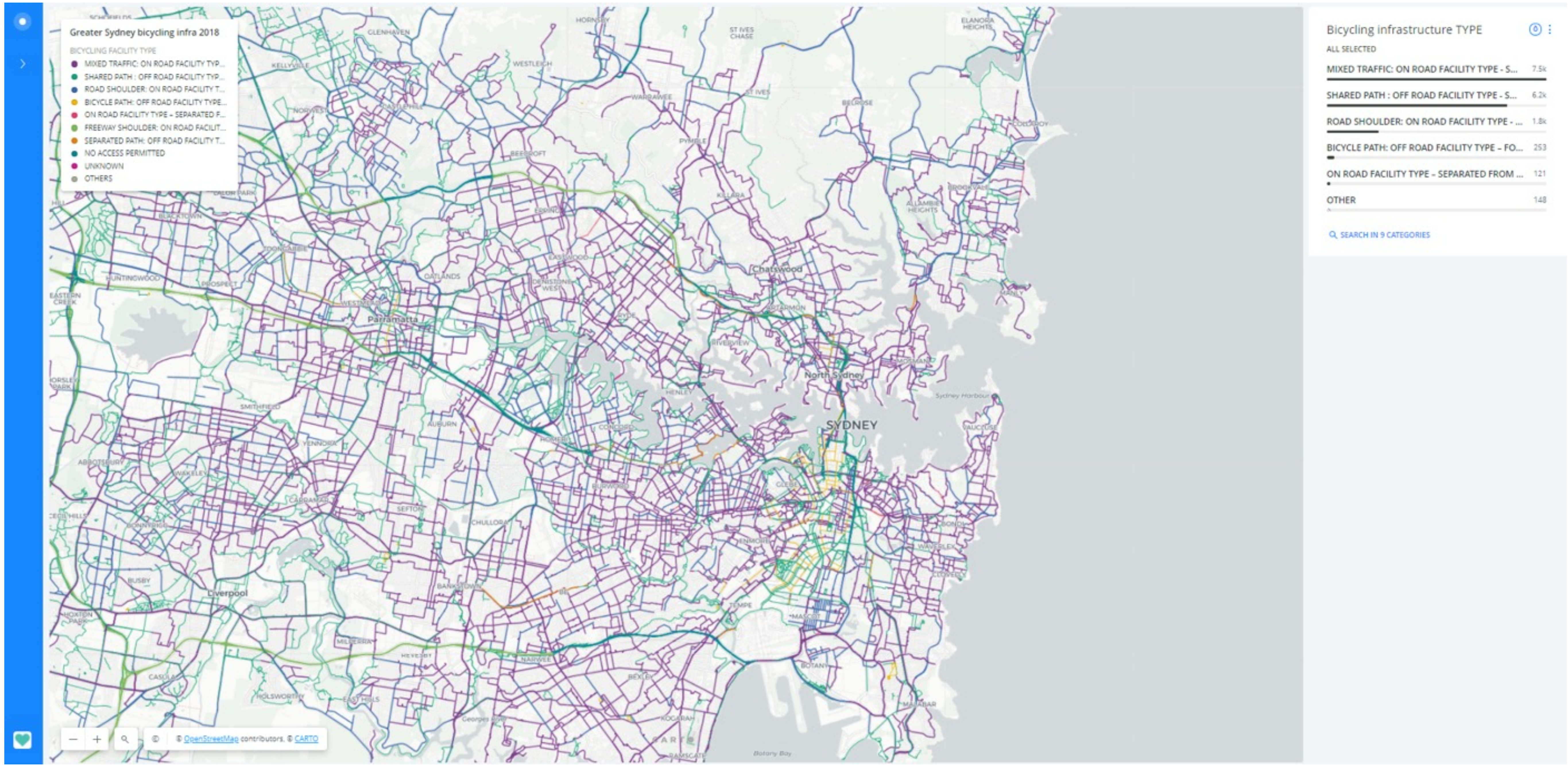
Cycling volume prediction model



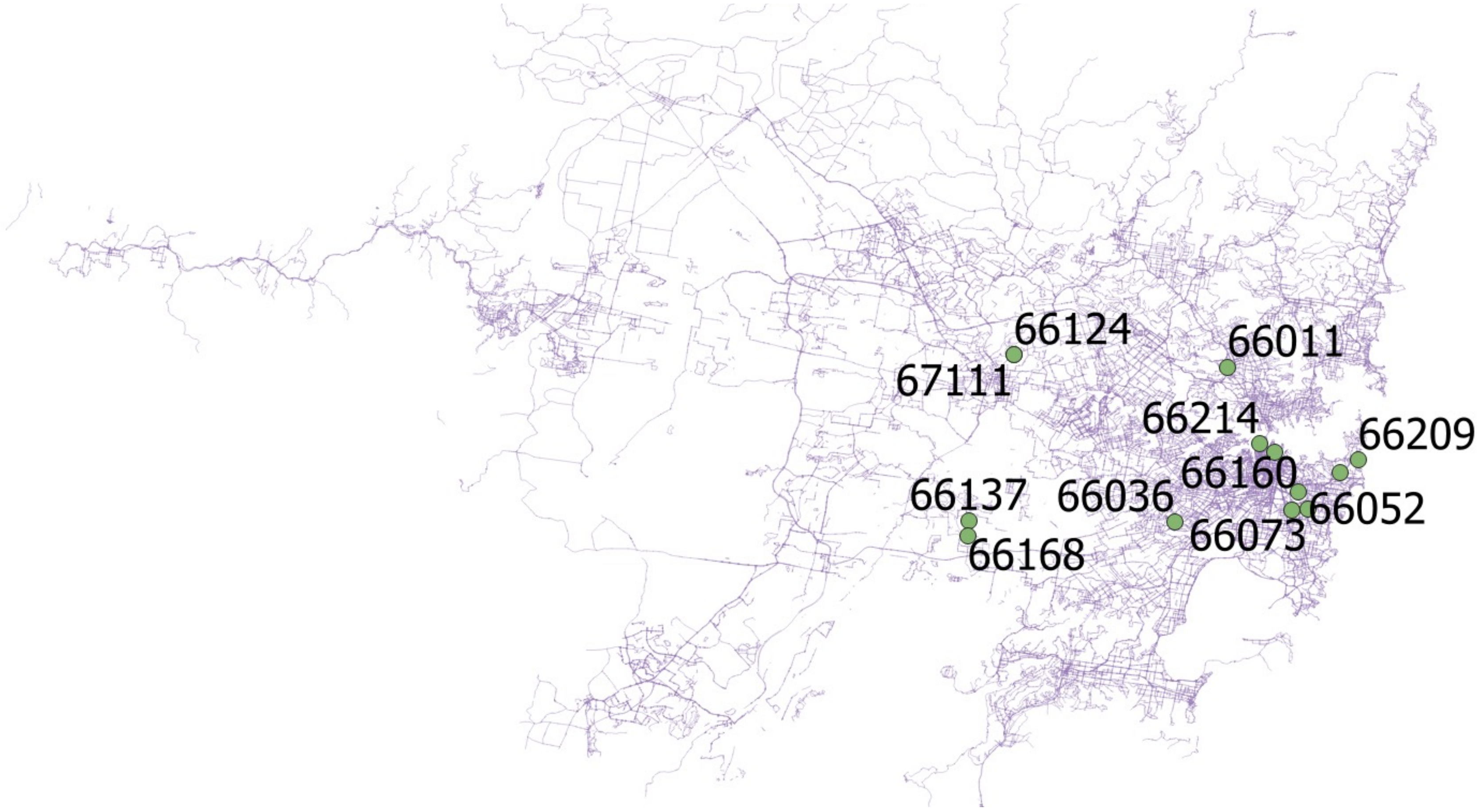
Location of official cycling counts



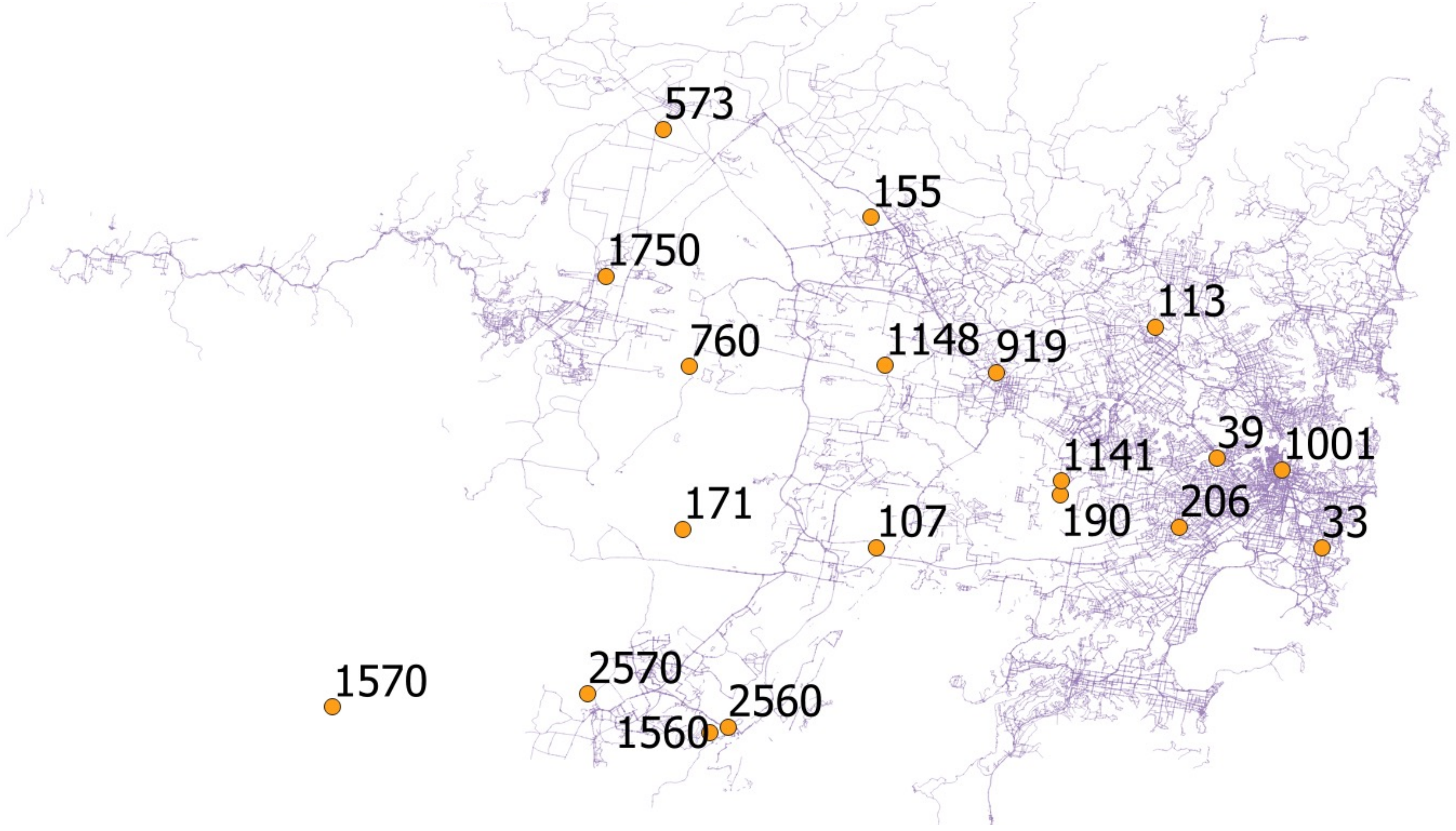
Cycling infrastructure network



Climate and air quality data

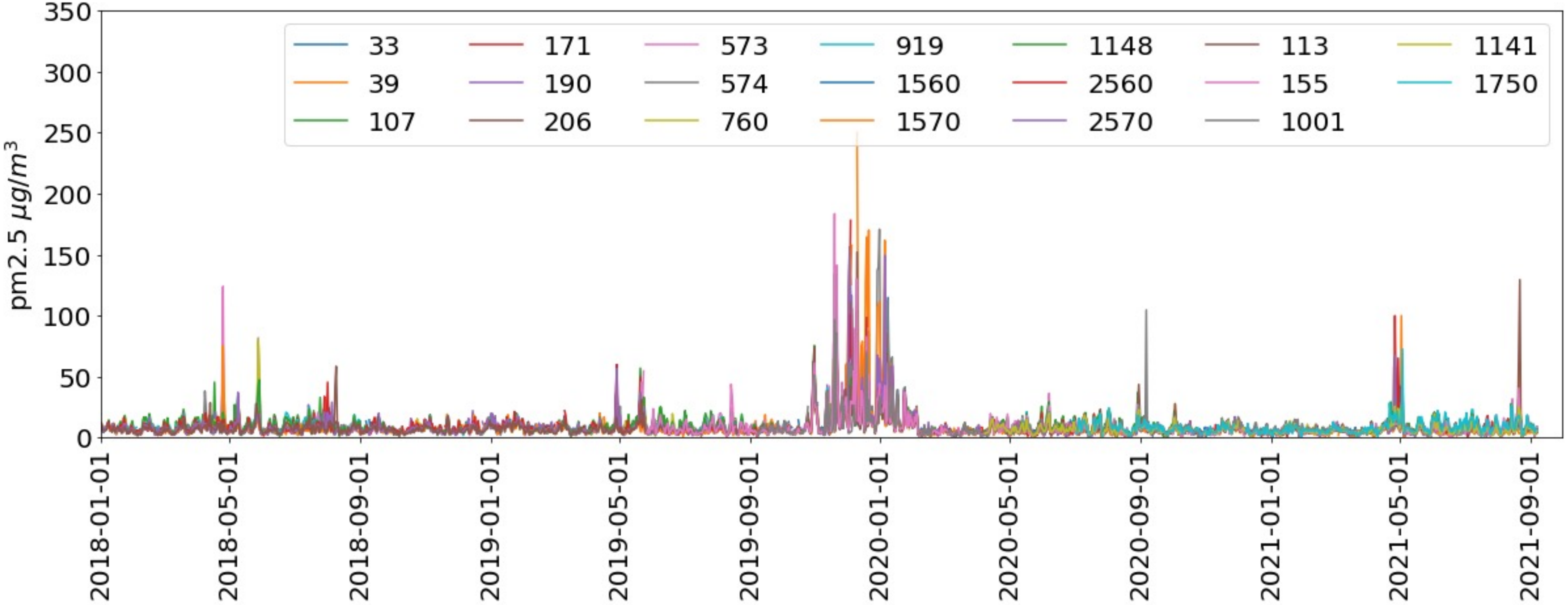


Weather stations

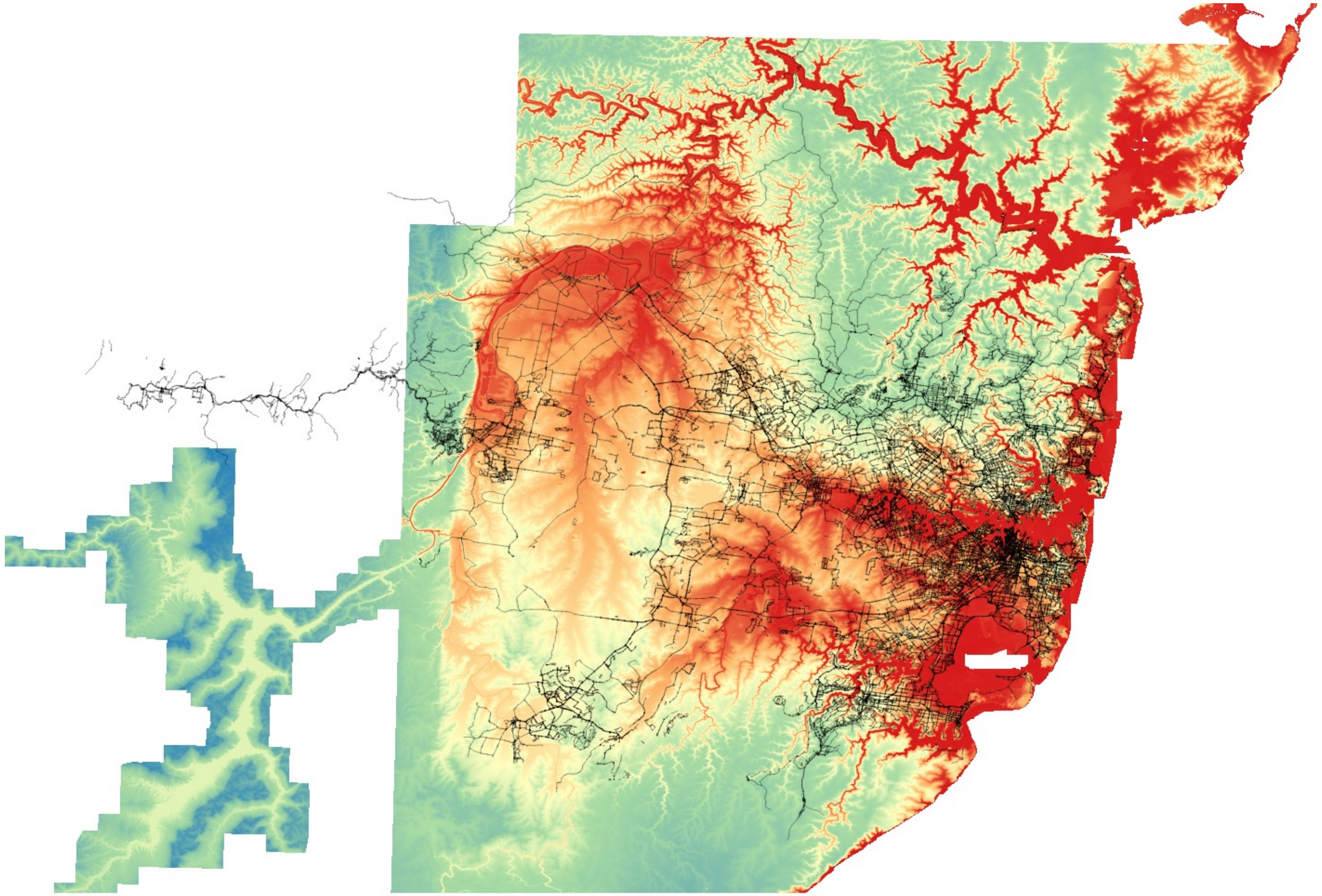


Air quality stations

Air quality data (PM 2.5)



Topography data



3 models

Linear regression

Very simple to understand and implement

Poisson regression

Appropriate for modelling count data

Decision tree regression

A widely used Machine Learning model

Spatial cross validation

- Linear and Poisson regressions are estimated from 100 trained different models (10 random splits of training and test set, 10-fold cross validation)
- Decision tree regressions are estimated from 1,000 different trained models (10 random splits of training and test set, 10-fold cross validation, 10 different hyperparameter sets)

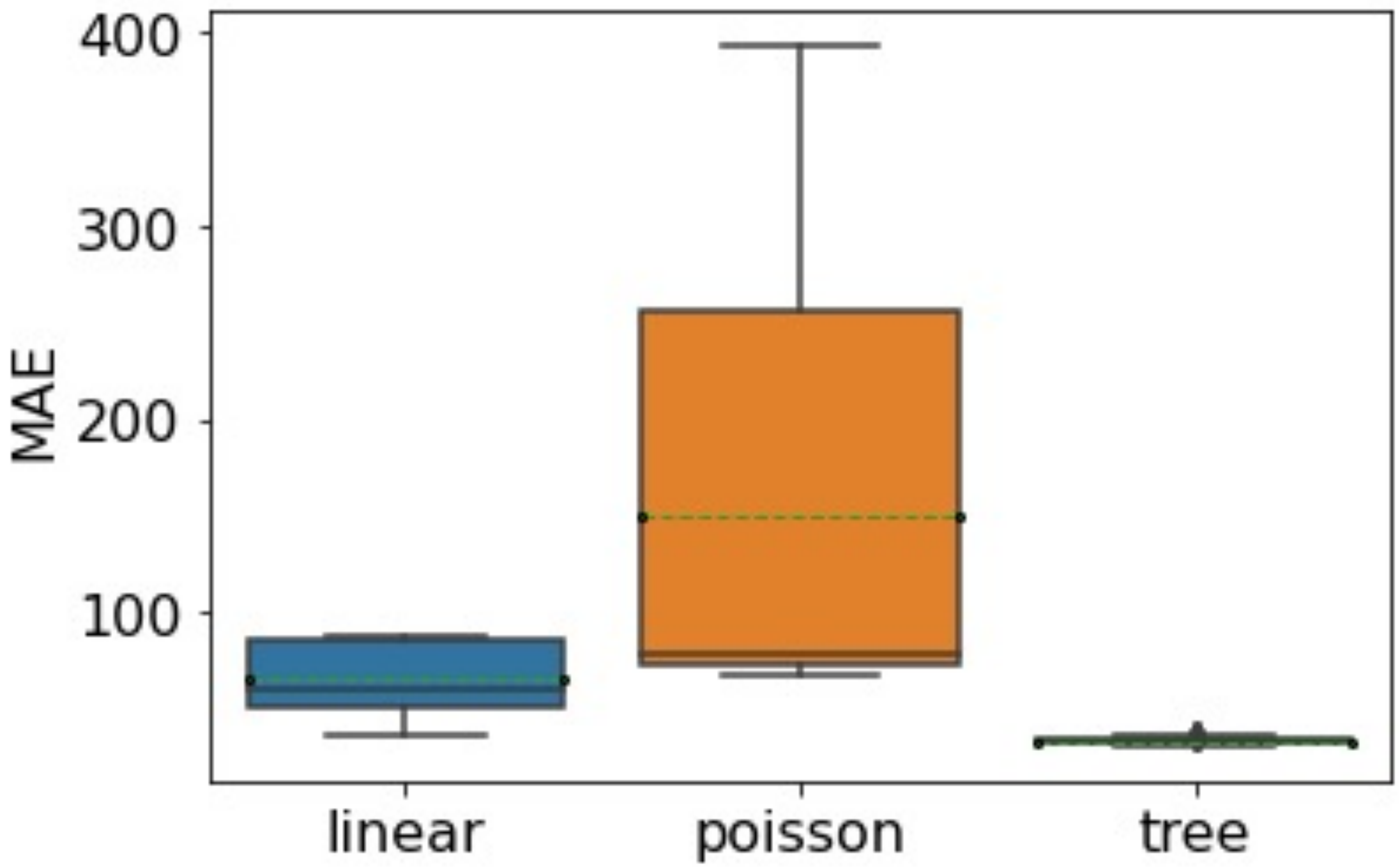
Model goodness of fit

Model	R^2	MAE	RMSE
Linear	0.64	37	50
Poisson	0.60	72	88
Tree	**0.82**	**31**	**59**

MAE: Mean Absolute Error
RMSE: Root Mean Squared Error

Variables remained in the models: Strava count, Population density, Land use entropy, Weekday, Slope, Pre-COVID and Air quality

Cycling models' predictive power



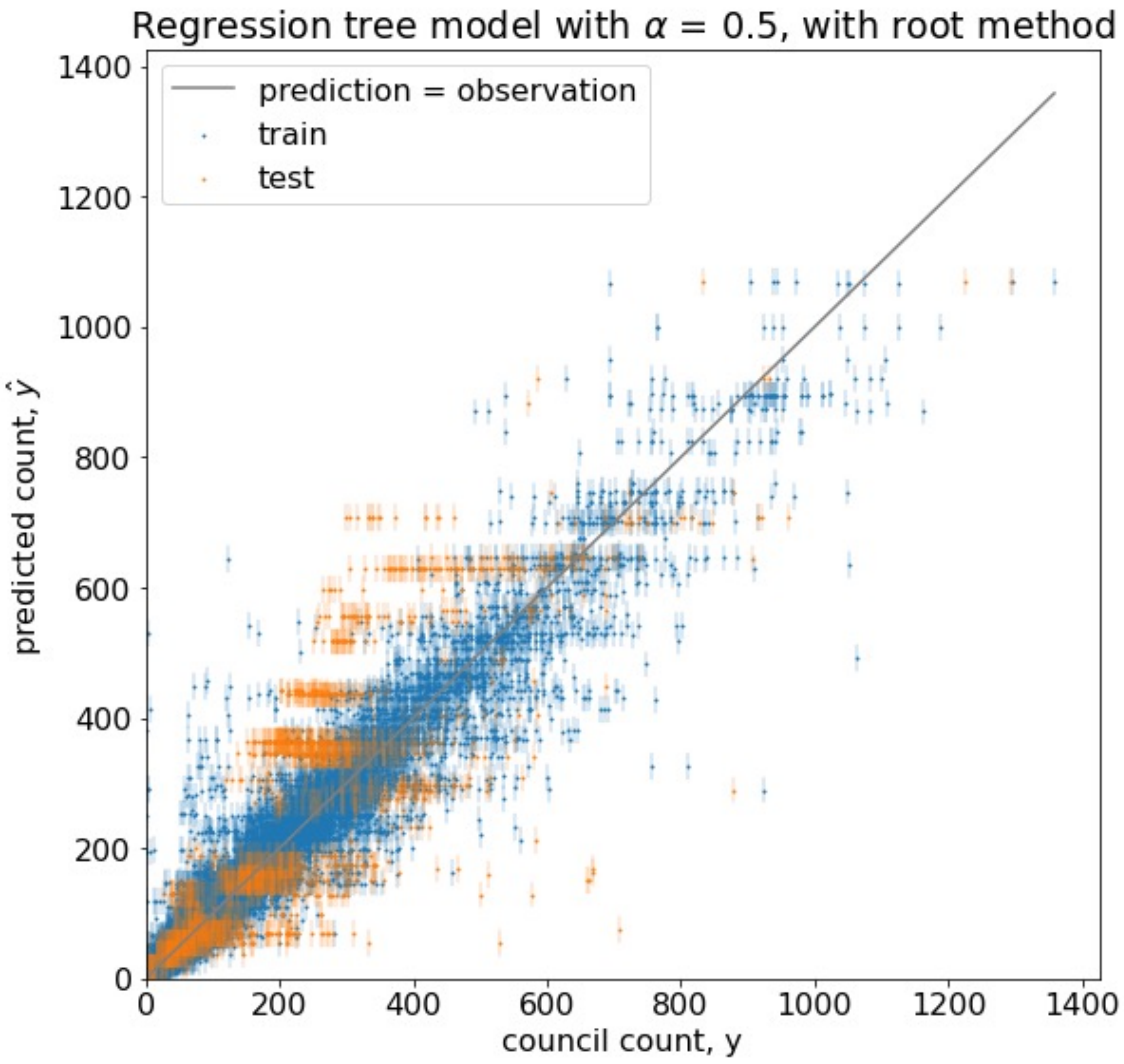
Estimating prediction intervals instead of point predictions

- The jackknife method estimates an interval centered at the predicted response of a test point, with the width of the interval determined by the quantiles of leave-one-out residuals.
- We use 100-fold cross validation instead of LOO to determine the residuals.

See this for details:

Barber, R.F., Candes, E.J., Ramdas, A. and Tibshirani, R.J., 2021. Predictive inference with the jackknife+. The Annals of Statistics, 49(1), pp.486-507.

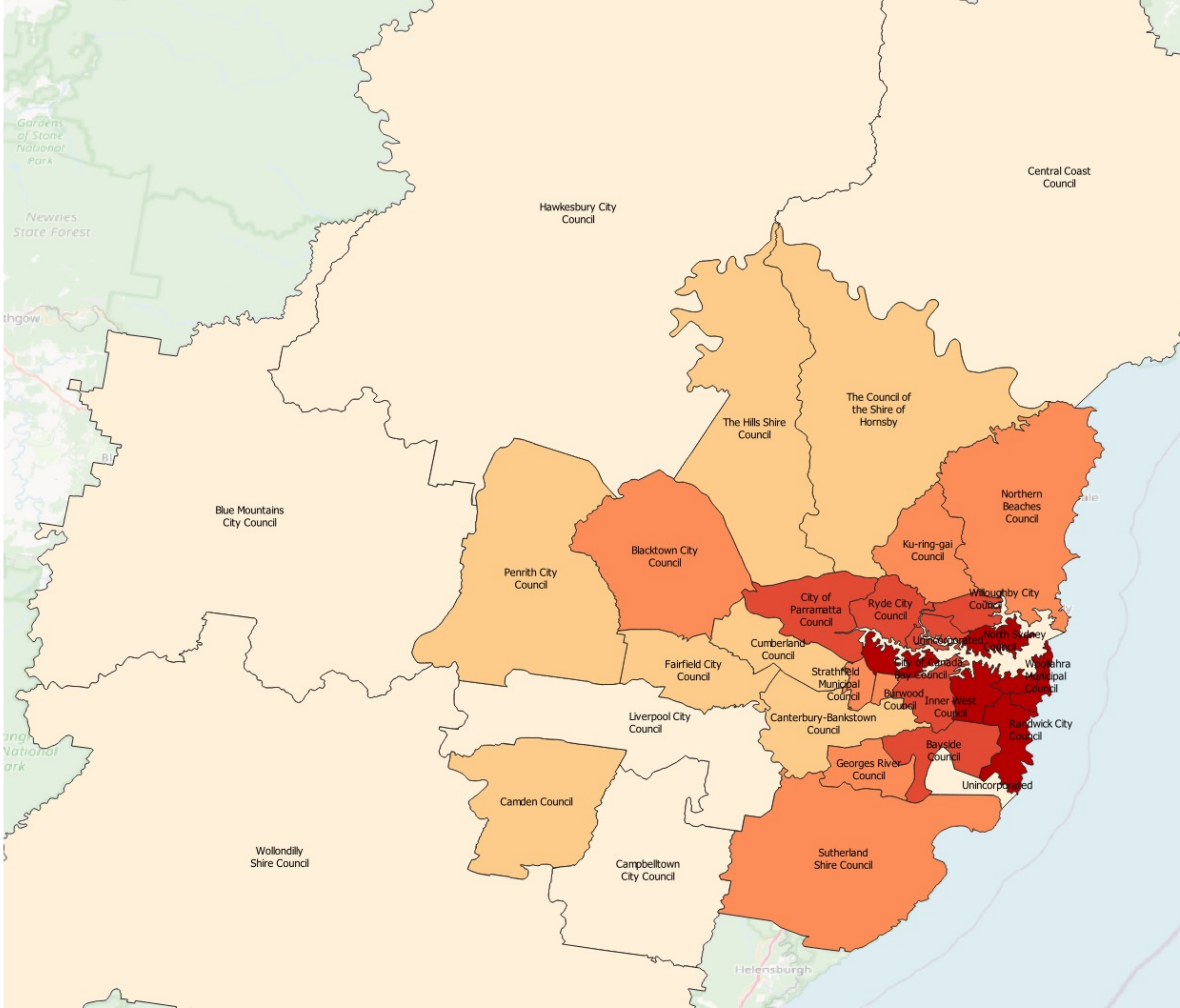
Predicted vs. observed cycling counts



Total number of trips per area

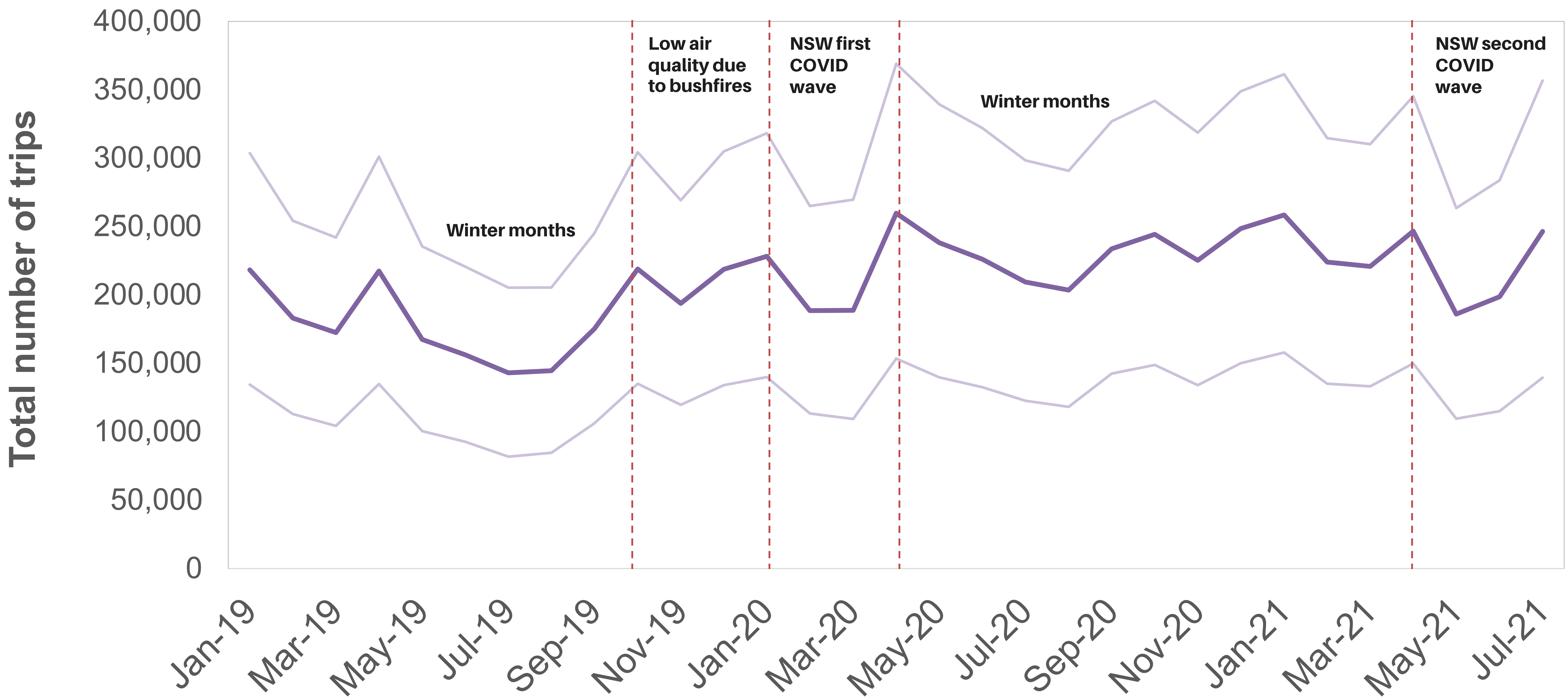
Number of trips/area(km^2)

	0 - 18
	18 - 32
	32 - 96
	96 - 186
	186 - 463

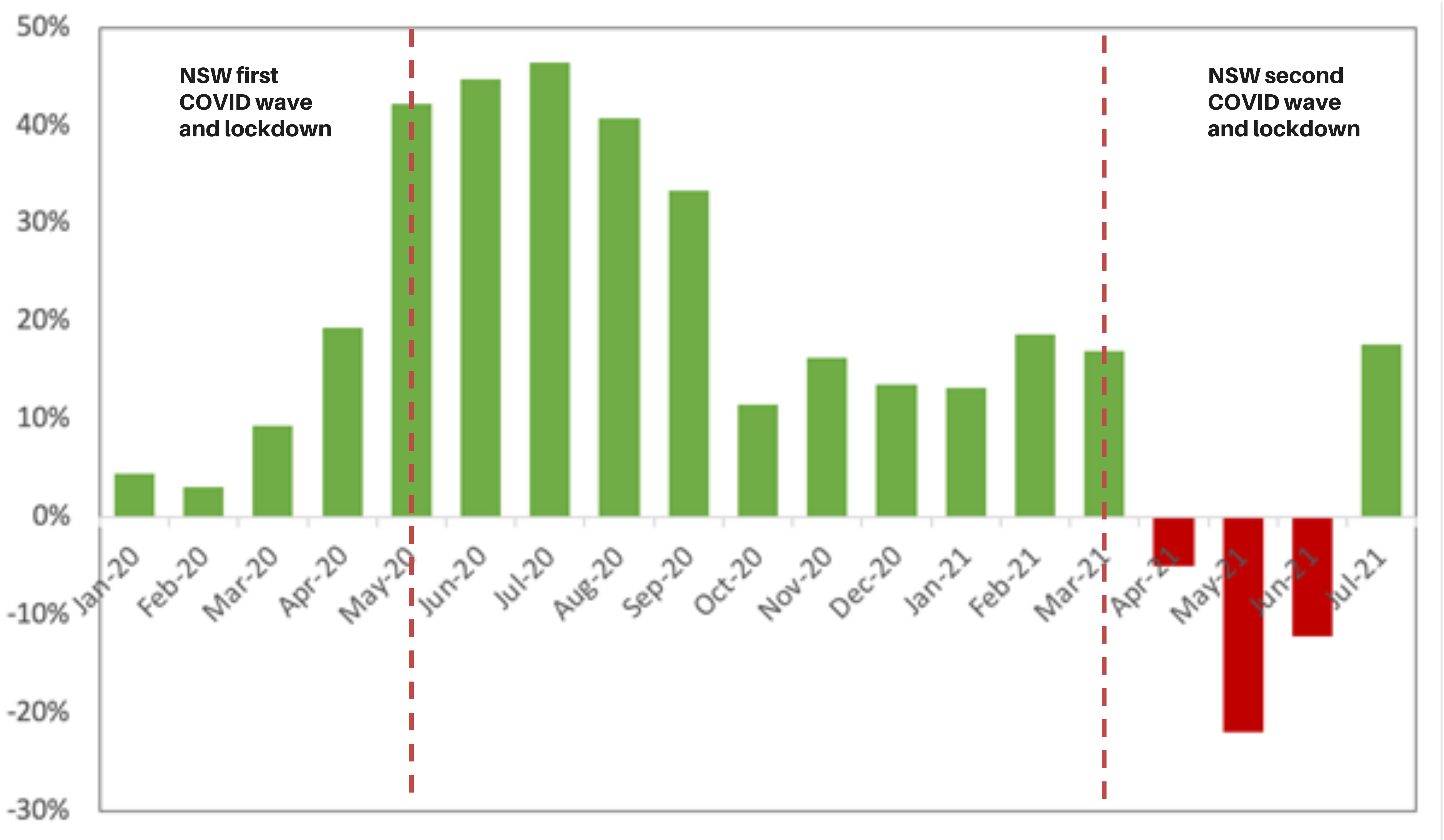


Note: Average cycling trip travel distance is assumed 4.7 km.

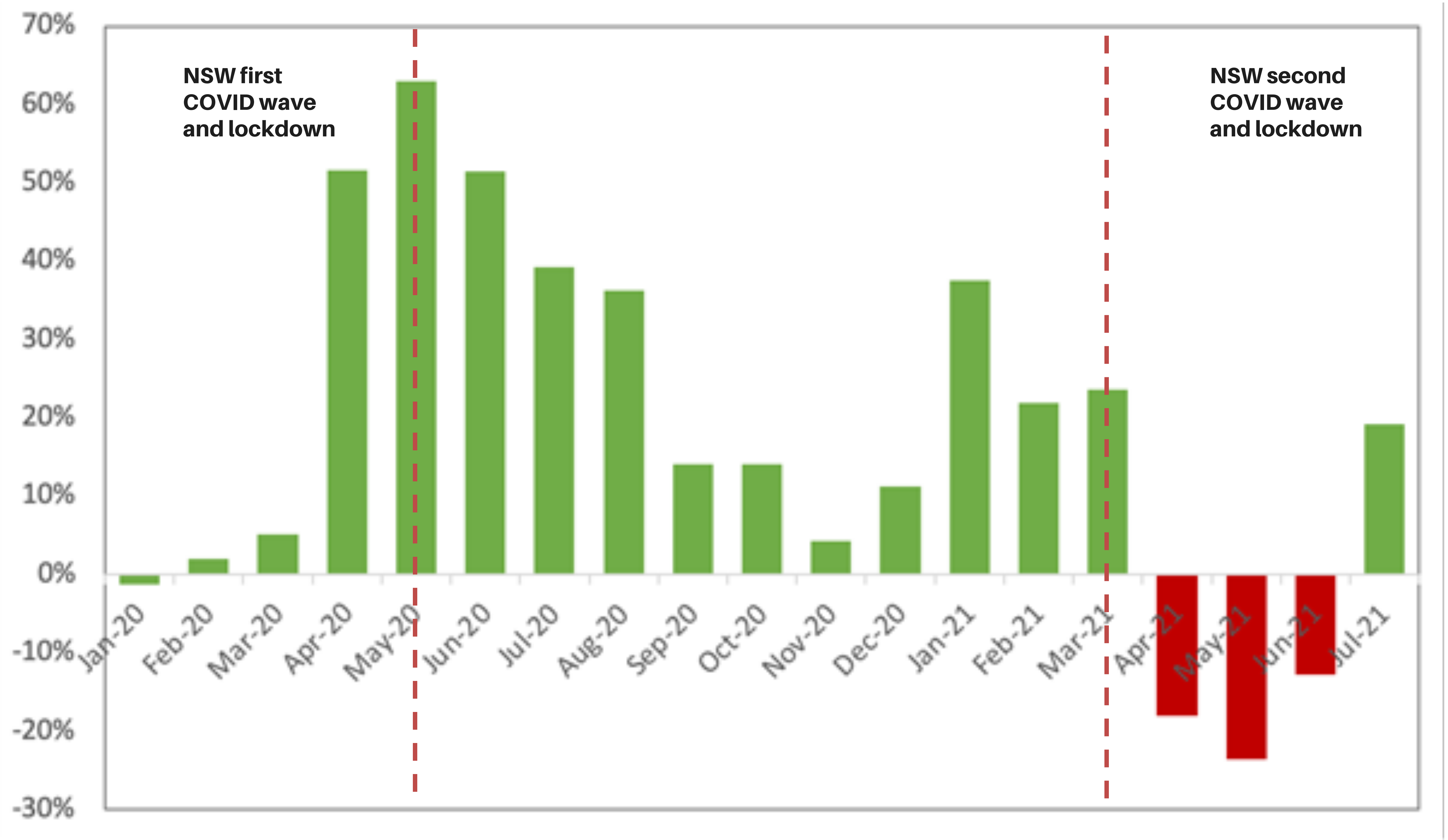
Average weekday trips



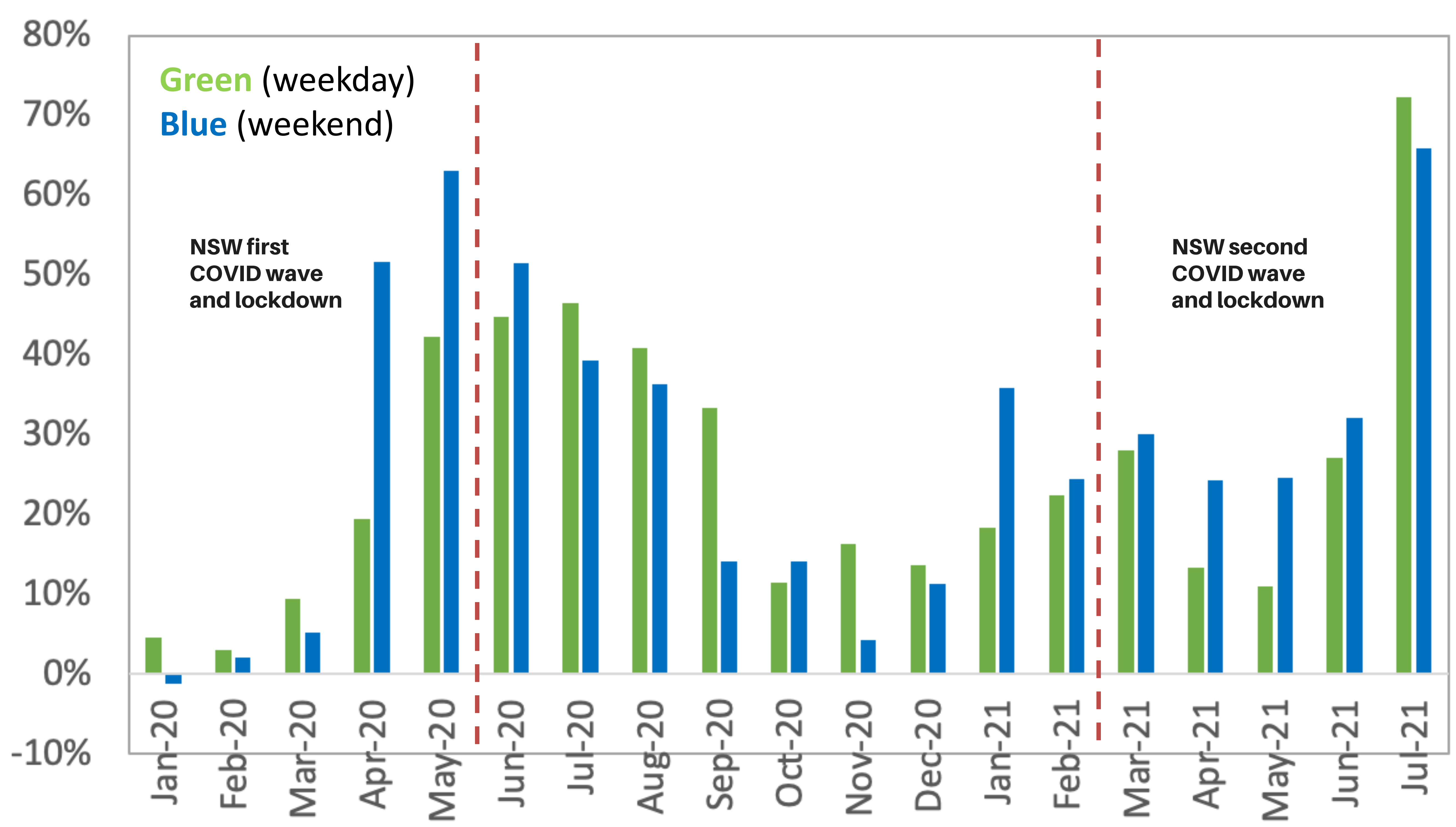
YOY percentage change (weekday)

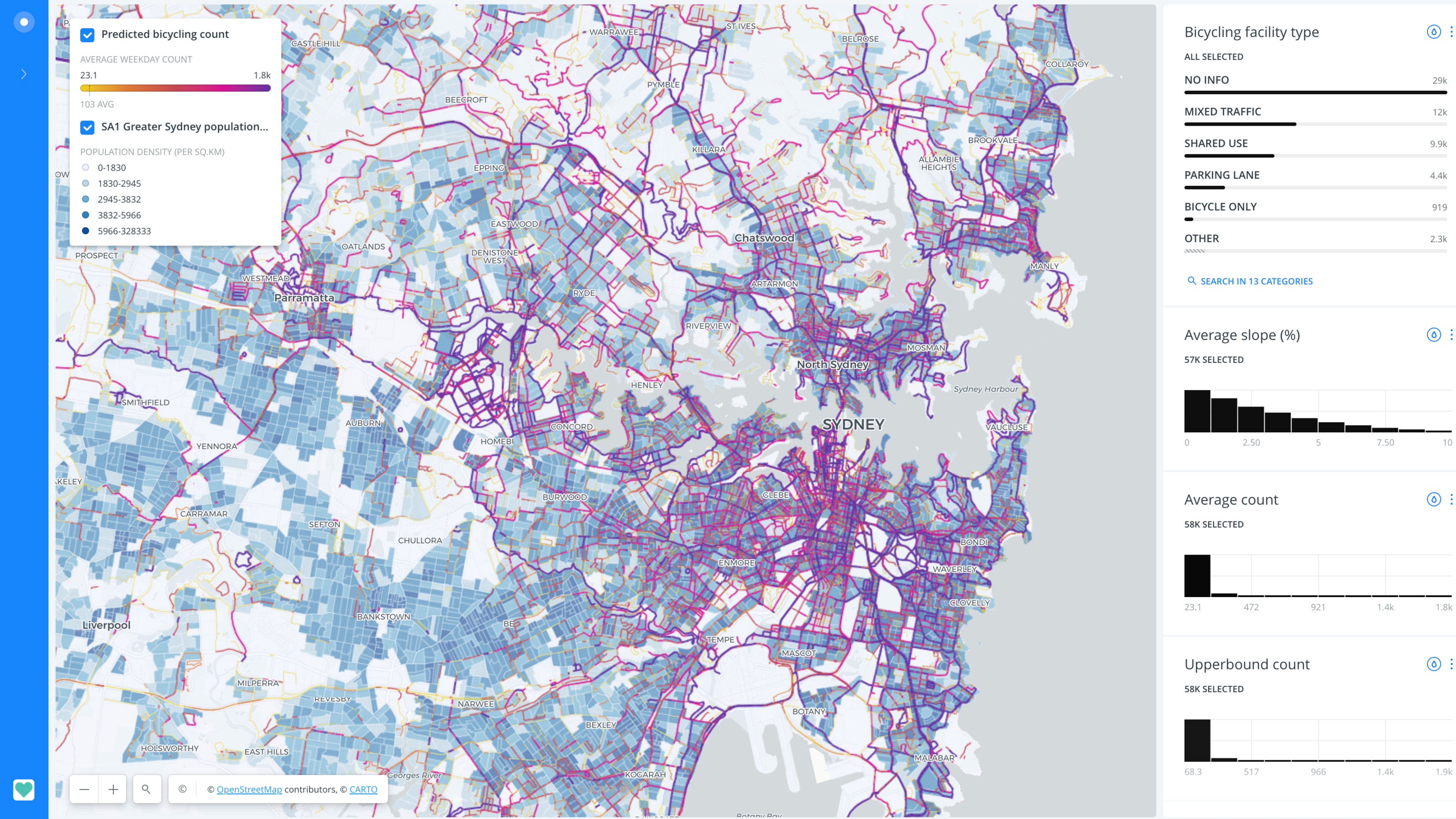


YOY percentage change (weekend)



Percentage change (compared to 2019)




Predicted bicycling count
AVERAGE WEEKDAY COUNT
23.1
1.8k
103 AVG
SA1 Greater Sydney population...
POPULATION DENSITY (PER SQ.KM)
0-1830
1830-2945
2945-3832
3832-5966
5966-328333
WARRAWEE
ST IVES
BELROSE
COLLAROY
CASTLE HILL
PYMBLE
BEECROFT
KILLARA
BROOKVALE
ALLAMBIE HEIGHTS
EPPING
EASTWOOD
Chatswood
OATLANDS
DENISTONE WEST
PROSPECT
WESTMEAD
Parramatta
RYDE
ARTARMON
MANLY
RIVERVIEW
MOSMAN
North Sydney
HENLEY
Sydney Harbour
SMITHFIELD
AUBURN
CONCORD
SYDNEY
VAUCLUSE
YENNORA
HOMEBUSH
BURWOOD
GLEBE
CARRAMAR
SEFTON
CHULLORA
BONDI
ENMORE
WAVERLEY
CLOVELLY
BANKSTOWN
Liverpool
TEMPE
MASCOT
MILPERRA
REVESBY
NARWEE
BOTANY
BEXLEY
HOLSWORTHY
EAST HILLS
MALABAR
Georges River
KOGARAH
Botany Bay
© OpenStreetMap contributors, © CARTO
Bicycling facility type
ALL SELECTED
NO INFO 29k
MIXED TRAFFIC 12k
SHARED USE 9.9k
PARKING LANE 4.4k
BICYCLE ONLY 919
OTHER 2.3k
SEARCH IN 13 CATEGORIES
Average slope (%)
57K SELECTED
0
2.50
5
7.50
10
Average count
58K SELECTED
23.1
472
921
1.4k
1.8k
Upperbound count
58K SELECTED
68.3
517
966
1.4k
1.9k

Conclusions

1. Appropriate statistical models can address the sampling bias of crowdsourced data

2. Crowdsourced data can be used along with other variables including official counts, land use and population data to estimate network wide cycling traffic

3. The quality and quantity of the input data are key (not surprisingly).