# When machine learning does you dirty: Robustness to measurement error in environmental economics

Aaron Watt

October 2021

### Abstract

Spatially-dense machine-learning-predicted data are becoming more available and social scientists and regulators are availing themselves of it. However, these data are being treated as highly accurate and estimates of the measurement error in these predicted data are often excluded from analyses. Are pollution regulation policies and policy analyses sensitive to the inclusion of measurement error?

Previous work by Fowlie, Rubin, and Walker (2019) on the issue of measurement error in satellite data has explored two spatially dense ($<$1km$^2$ resolution) satellite-derived datasets that have been produced to estimate ground-level PM2.5 pollution. These data offer a feasible option for the EPA to regulate pollution without installing more expensive pollution monitors. Ground-level PM2.5 pollution concentration is estimated using aerosol optical depth (AOD) satellite data and various predictive models – PM2.5 is not directly measurable (yet) using current satellite instruments.

To help estimate measurement errors of these datasets, we can download PurpleAir data using their API.

Adoption of PurpleAir monitors is non-random so I need to develop and estimate a selection model into PurpleAir adoption. PurpleAir monitors are also known to be less accurate and might be systematically biased, so estimating PM2.5 using PurpleAir measurements and the sparse (but highly accurate) EPA monitors will be important.

## 1 Research Question

How much do prediction errors matter in pollution regulation? Does incorporating prediction errors of machine-learning-produced pollution data affect the policy categorization of areas without a pollution monitor?

## 2 Draft Project Outline

1. write the code to scrape the purple air data

2. apply a selection model to the purple air data (since monitor adoption is non-random)

3. write model of prediction errors of dense prediction of a spatially sparse, temporally dense variable using stats literature

4. generate descriptive stats for purple air data and estimates of measurement error using EPA monitors as ground truth

5. explore machine learning prediction of prediction errors using purple air data and using a leave-out group for validation testing

6. produce draft tables of changes in attainment status