**PPOL 670 Introduction to Data Science**

**Project Proposal**

Dingxuan Zhao

**Statement of problem**

Hybrid Electric Vehicle (HEV) and Battery Electric Vehicle (BEV) are possible solutions to reduce greenhouse gas (GHG) emission or air pollution. However, because of relatively higher price compared to Internal Combustion Engine (ICE) vehicles, subsidies are important incentives to purchasing decision. In United States, there are federal, state, and county-level subsidies, for consumers, producers, and infrastructure. Because subsidy is critical, it’s valuable to figure out the reason to promote subsidy policy, both in state and county government. It’s fair to assume that GDP and income per capita, and fiscal revenue are contributors, but how about the coefficient? Considering the partisanship towards global warming and GHG emission, will it influence subsidy? How about gasoline and electricity price? Some cultural factors such as driving preference, can we quantify them? In short, that is the

Someone may argue that the energy which push a car drive from A to B is constant, so EV only transfer GHG emission and pollution from car to power plant. The reason that EV can reduce GHG emission is that large dynamo has a higher thermal efficiency than car engine has. But how about air pollution? If we want to declare that EV can reduce pollution, we have to assume that the electricity comes from clear grid, but is it the truth? We can check it by comparing the pollution produced by cars and came from incremental electricity demand. Furthermore, we can check the environmental cost of EV subsidy, which means whether the highest EV subsidy is offered by the state/county with cleanest grid, and vice versa?

In short, we have two research questions:

**Research question 1**: What are the contributors of EV subsidy policies?

**Research question 2**: Will the EV subsidy policies help to solve air pollution?

**Data source**

EV sales

Advanced Technology Vehicle Sales Dashboard[[1]](#endnote-1)

EV subsidies

State-level: A Strata report[[2]](#endnote-2) offered a summary of state subsidy, some subsidies have been outdated but it also offered an index to detailed policies. For example, if we want to find the California subsidy details, we can get it in U.S. Department of Energy Alternative Fuels Data Center[[3]](#endnote-3).

County-level: Holland[[4]](#endnote-4) offered an index of county-level subsidies.

Energy price, electricity generation source (clean grid or not)

U.S. Energy Information Administration API[[5]](#endnote-5)

Economic data

State/County GDP/income: BEA[[6]](#endnote-6)

State/County fiscal revenue: each state offer both state and county level budget document. For example, county of Maryland’s data can be found in Maryland Association of Counties[[7]](#endnote-7).

**Plan to obtain data**

Some data can be easily downloaded. For the other, for example, some county-level EV sales/subsidies are not in an existing table, but are published on the website or existing data visualization without original data, then we can use web scraping.

**Methods**

It can be expected that every dataset is wide table, so I will delete observation with missing values, transform them to a long table and aggregate different states/counties data within same topic into one data frame.

The correlation between amount of subsidy and independent variables will be checked by supervised learning. Subsidy data will be divided into training and test sets, regressed with independent variables, then use K-Fold cross-validation to generate estimates for the test error.

For subsidy’s contribution to reduce air pollution, a ratio between total amount of subsidy and total pollution produced by grid is expected. A dichromatic county-level map will be used to illustrate the ratio, greener means higher ratio (more subsidy offered in cleaner grid) while redder means lower ratio.

**What constitutes success**

For research question 1, a successful answer means a series of factors that promote the subsidy policies made by state and local government. The coefficient between independent and dependent variables should be statistically meaningful. For research question 2, the data visualization mentioned above is expected.

A “successful” project means a comprehensive methodology to answer two research questions has been built, including web scraping, data cleaning, supervised learning, and data visualization. But it doesn’t I will finish these works for every county in 50 states, considering the data accessibility, the county-level research may be limited to some focus states, so long as they are sufficient to show that the methodology works.

1. <https://autoalliance.org/energy-environment/advanced-technology-vehicle-sales-dashboard/> [↑](#endnote-ref-1)
2. Bosworth. R., Patty. G., *The current state if electric vehicle subsidies: economic, environmental, and distributional impacts.* <https://strata.org/pdf/2017/ev-full.pdf> [↑](#endnote-ref-2)
3. <https://afdc.energy.gov/laws/state_summary?state=CA> [↑](#endnote-ref-3)
4. Holland. S. P., Mansur. E. T., Muller. N. Z., Yates. A. J., *Environmental Benefits from Driving Electric Vehicles?* https://energy.umich.edu/te3/wp-content/uploads/sites/2/2018/09/mansur\_holland\_muller\_yates\_2015\_Te3.pdf [↑](#endnote-ref-4)
5. <https://www.eia.gov/opendata/register.php> [↑](#endnote-ref-5)
6. <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas> [↑](#endnote-ref-6)
7. <https://www.mdcounties.org/167/Budget-Finance> [↑](#endnote-ref-7)