Vehicle type distribution and CO_2 emissions in California

Melissa Chen

December 2022

Abstract

We hope to identify a relation between the number of registered cars over the last 10 years based on fuel source (electric vehicles, plug-in hybrid vehicles, gasoline-powered vehicles...), electricity demand, and carbon emissions in California. Using this, we predict what carbon emissions in California would look like with different values of registered types of cars–on one extreme, what if all cars in California were EVs? What would carbon emissions look like with increased electricity consumption but decreased direct emissions from gasoline-fueled vehicles? [TODO]

1 Introduction

On 25 August, 2022, California Governor Gavin Newsom shared news that the state would take an exceptional step towards mitigating its carbon emissions: starting in 2035, 100% of all new vehicle sales must be zero-emission vehicles (ZEVs). Approved by the California Air Resource Board (CARB), this policy is the first of its scale in the United States, with the goal of drastically reducing direct carbon emissions from traditional combustion engine vehicles and eventually hitting zero emissions generated by personal transport for the average Californian. This new policy sets intermediary targets of 35% ZEV sales by 2026 and 68% by 2030, so doubling the 2022 figure of 17.7% within four years and almost quadrupling it within eight. A government investment of $10 billion to fund developments like charger installations and switching buses and other public transit vehicles out for ZEVs will aid the initiative. (**??**)

As California makes these changes, we hope to understand the resulting emissions reductions. On one hand, fewer combustion engine vehicles on the road means fewer greenhouse gases (GHGs) directly entering the atmosphere; on the other, the state of California will see a sharp increase in electricity demand over the next decade in order to power ZEVs, and the associated electricity generation will in turn engender greater emissions. This paper's goal is not to argue the effectiveness of the policy—though we trade one challenge for another (limiting carbon emissions from combustion engine vehicles for limiting carbon emissions in the process of energy generation), there is certainly more room for growth in the latter. Rather, this project aims to apply machine learning techniques to elucidate the relative difference in emissions if California's electric vehicle (EV) numbers hit certain milestones, as opposed to if they remain as they are.

The main steps taken throughout this project were: (i) gaining a greater understanding of relationships between contributors to CO_2 emitted during energy generation in the state of California, including vehicle type distribution, total electricity demand, and energy generation by resource, (ii) ascertaining the applicability of Random Forest and Support Vector Machines to the relation between these contributors for predicting CO_2 emissions from energy generation, and (iii) based on this relation, estimating CO_2 emissions reductions based on different distributions of vehicle types.

2 Background

2.1 Estimating CO_2 emissions

Accurately predicting regional greenhouse gas emissions poses a challenge for many reasons including the difficulty surrounding precise data collection on a large-scale, the breadth of contributors to GHG emissions, and identifying these factors' non-linear relations with emissions. For the scope of this paper, I chose to focus not on absolute predictions of CO_2 emissions, but the relative increase or decrease of these emissions based on current data. For that reason, I tolerate the use of estimates of state-wide emissions as a central dataset, described in section 2.2.

In existing literature, authors like **?** and **?** use support vector machines (SVMs) and modified SVMs for CO_2 emission estimation. In the context of corporate CO_2 emissions, **?** apply a meta-learners over base-learners like Elastic Net for increased prediction accuracy. [TODO]

2.2 Energy and emissions data

The California Independent System Operator (CAISO) is a non-profit independent grid operator which manages about 80% California's electric power system at the behest of the Federal Energy Regulatory Commission (FERC). (**?**) It participates in the Western Energy Imbalance Market (WEIM), through which 19 different power management and generation systems across the western U.S. buy and sell energy to meet their immediate consumer needs, allowing for low-cost, efficient energy management and easier integration of renewable energy. (**?**)

This paper uses three datasets collected from CAISO, all measurements for the entirety of the state taken at 5-minute intervals from 1 August, 2022, to 27 August, 2022: [TODO] (**?**)

- **Electricity generated by resource in megawatts (MW).**
  Resources include natural gas, large hydro (produced by facilities with a capacity greater than 30 MW), batteries, nuclear, coal, imports from other states, and other. Renewable resources include solar, wind, geothermal, biomass, biogas, and small hydro (produced by facilities with a capacity at 30 MW or less). These values are used in the below estimation of total CO_2 emissions.
- **Estimated CO_2 emissions resulting from energy production in megatons of CO_2 per hour ($mTCO_2/h$)**.
  CAISO's real-time estimated CO_2 emissions are calculated by adding estimated emissions generated from internal ISO dispatches and emissions from imports that serve ISO loads, excluding emissions from exports. (**?**) See Table 1.
- **Electricity demand in megawatts**.
  The amount of energy needed on the grid managed by CAISO at a given moment.

The estimated CO_2 emissions dataset is also available at a monthly resolution for the 2014-2022 period. To match other datasets' resolutions, dataset values are sometimes compressed to hourly, daily, or monthly values, which are simply sums of all the reported values within each timeframe.

$$em_{ISO} = em_{dispatch} + em_{import} - em_{export}$$

em_{ISO}	total emissions produced in order to generate the energy circulated by CAISO.
$em_{dispatch}$	total CO_2 emissions produced through dispatches of ISO internal resources, measured in the above mentioned dataset. For each resource i, we multiply the amount of electricity generated through this resource e_i (MWh), CO_2 emission factor for that resource em_i (mTCO$_2$/MMBTU), and the resource heat rate $heat_i$ (MMBTU/MWh): $em_{dispatch} = \sum_i e_i \times em_i \times heat_i$
em_{import}	emissions generated in the process of producing energy imported through the WEIM. This is approximated as the product of the amount of energy imported and the unspecified emission rate established by CARB of 0.428 mTCO2/MWh.
em_{export}	emissions generated in the process of producing energy within the CAISO system that is exported through the WEIM. Calculated in a similar manner to $em_{dispatch}$

Table 1: CO_2 emissions estimation

The U.S. Energy Information Administration, under the greater U.S. Department of Energy, serves to collect and disseminate energy data and has made available the two below listed datasets. (**?**) These are the datasets originally referenced in this paper's abstract proposal and upon which I performed preliminary analyses. I continued on to use them as cross references with CAISO's data, as they measure similar values at a lower resolution:

- **Hourly electricity load in megawatts.**
  Reports for the state of California by "California-based balancing authorities."
- **Monthly electricity generation by energy source in thousands of megawatt hours**.
  Resource types include coal, natural gas, nuclear, conventional hydroelectric, wind, and solar.

2.3 Vehicle data

This paper analyzes light-duty vehicles registrations rather than their sale numbers, the true target of the new ZEV policy, due to data availability and a more salient relation to CO_2 ultimately emitted in the state. We look to the **?** for data on California's yearly registered vehicle numbers by different fuel types, enumerated in full in Table 2, over the 2010-2021 period. Monthly vehicle registrations are approximated by assuming that registrations for each vehicle type grow roughly linearly over the course of the year.

For context, in 2021, EVs numbered 522,445 or 1.745% of all light-duty vehicle registrations, a somewhat low number given that California leads the country in EV statistics by a huge margin: California's EV population accounts for more than 40% of the total number of EVs in the U.S., and it topped the state rankings of EV sales in 2021, outnumbering those of the next 10 states combined. (**?**)

Gasoline	Standard combustion engine vehicle	86.83%
Gasoline hybrid	Fueled by gasoline and electricity generated while driving, e.g., conventional Toyota Prius	4.34%
Flex fuel	Fueled by gasoline, ethanol, or a mix of both	4.04%
Diesel	Fueled by diesel, often a heavy combustion engine vehicle	1.97%
Electric **	All-electric vehicle, e.g., Tesla models	1.74%
Plug-in hybrid *	Fueled by both gasoline and electricity	1.02%
Fuel cell *	Fueled by electricity generated from a fuel cell of compressed hydrogen, e.g., Hyundai NEXO	0.03%
Natural gas	Fueled by compressed or liquefied natural gas	0.03%
Propane	Fueled by liquified petroleum gas, often used in fleet applications, e.g., school buses and police vehicles	0.00%

* Zero-emissions vehicle

* Uses electricity from being plugged in

Table 2: Vehicles by fuel type and the proportion of vehicle registrations in California that they represent in 2021

3 Methodology

3.1 Models

Lasso
Random Forest (RF) regression. As such RF regression is ill-suited to time series problems. Within the scope of this paper, we hope not to project into the future but to, within our current and historic context, understand energy generation and emissions trade-offs, making RF regression a reasonable tool.
Stacking

Developed by **?** and other colleagues, Support Vector Machines (SVM) algorithms identify a hyperplane dividing the n-dimensional space representing the n-feature input, separating the data points into two groupings (classes) such that its marginal distance from the two groupings is maximized. We refer to sample data points along the calculated margins as support vectors. This technique runs efficiently for both linear classification and non-linear classification through the application of the “kernel trick" A regression application of SVMs, called Support Vector Regression (SVR), exists [TODO]

3.2 Generating vehicle distribution and demand samples

In order to pass realistic test vehicle distribution data and demand values into our model, we start with vehicle distribution and use two different methods of generation.

The first uses RF regression on given monthly distributions over the 2010-2021 period based on the input value of electric vehicle numbers. Similarly to the main emissions model, we stack RF regression with a linear regression to mitigate lack of output breadth.

The second simply projects into the future the number of vehicles over time given current growth trends for each vehicle type through the application of linear regression.

4 Results

4.1 Initial discoveries?

We were quickly disabused of two original suppositions, the first that combustion engine vehicle registrations correlated strongly with CO_2 emissions and the second that ZEV and hybrid vehicle registration numbers impacted electricity demand. In an initial analysis, the correlation coefficients for both relations were insignificant. Indeed, regarding direct vehicular emissions, the statistic that transport counts for up to a third of total emissions (source) refers to the entirety of the automotive industry including production. With regard to ZEVs, the numbers remain too few at still less than 2% of all California vehicles to be significant

4.2 Results

List results etc

4.3 Reflection

Etc
Could be done just as well using averages

5 Conclusions

5.1 Future work

etc

5.2 Final thoughts

Etc