Qingyuan Gao
qingyug3@uci.edu

Data Analytics and Modeling Task

1 Problem Statement and Assumptions

Vehicle electrification in the transportation sector is an emerging and effective solution to reduce carbon footprint and mitigate environmental and climatic impacts. Despite the fact that the U.S. has arguably lagged and is increasing the effort in decarbonization in recent years, EV in U.S. and especially in California has received increased attention from general public and stakeholder groups.

The Data Analysis and Modeling task is to develop a spatial temporal model for estimation and prediction of EV adoption (i.e., number of registered EVs) at the U.S. County level.

The data used are EV registration data, EV Chargers number and demographic data in 58 CA counties from 2012 to 2020. The demographic data is a 5-year estimate by ACS. The reasons for choosing these sets of data are:

1. The EV registration data recorded the new EV registration number in CA each year from 2010 to 2020
2. The EV charger number is from 2020-2022, in a quarterly format.
3. The ACS data recorded demographic data from 2009 to 2023. However, it was not until 2012 that ACS began to collect household level data. Datasets before 2012 only collected data on personal level.
4. There are 2 types of ACS dataset, the 1-year estimate and the 5-year estimate. However, the 1-year estimate only includes counties with populations larger than 65,000, which means 40-43 counties out of 58. The 5-year estimate includes all 58 counties. Therefore, the 5-year estimate dataset was used.

2 Data Summary and Description

Besides the EV registration data, demographic data ranging from population to housing rent is selected to be independent variables in the data analysis. The specific independent variables are: total number of households in the county, median age of people in a county, median household income, employment number, higher education rate, household size and median housing rent. The selection is based on the reflection of population and economic characteristics of the county, as well as the practice in previous research. Formers studies, such as S Wee et.al(2018), Prakobkaew et al(2021) and A Chhetri(2024) adopted a wide range of variables, including population & economics, distribution of EV chargers and geographical data of states or counties.

Qingyuan Gao
qingyug3@uci.edu

A summary of demographic data and EV registration number is on the table below. As the population, household numbers and employment are proportional to the size of the county, the median numbers are used. The rest of the summary data uses the mean value.

year	total population/ millions	median population/ thousands	median_age/ year
2012	37.33	179.07	38.53
2013	37.66	179.47	38.74
2014	38.07	179.99	38.93
2015	38.42	180.52	39.17
2016	38.65	181.11	39.28
2017	38.98	182.49	39.47
2018	39.15	183.44	39.61
2019	39.28	184.63	39.89
2020	39.35	185.46	39.95

year	total employment/millions	median employment/thousands	average employment rate
2012	16.61	75.63	0.45
2013	16.64	74.46	0.44
2014	16.89	73.92	0.44
2015	17.25	74.41	0.45
2016	17.58	75.56	0.45
2017	17.99	76.28	0.46
2018	18.31	77.72	0.47
2019	18.59	79.62	0.47
2020	18.65	79.95	0.47

year	median numbers of households/thousands	household size	median housing rent
2012	68.36	2.76	1051.12
2013	68.43	2.76	1068.57
2014	68.09	2.77	1077.48
2015	68.23	2.78	1084.10
2016	68.62	2.78	1118.62
2017	69.29	2.79	1164.10
2018	69.82	2.78	1203.40
2019	71.08	2.76	1260.48
2020	71.96	2.75	1314.24

Qingyuan Gao
qingyug3@uci.edu

year	total EV	median EV/ one county
2012	18054	25
2013	51849	58
2014	107983	132
2015	165121	249
2016	229005	382
2017	320644	605
2018	444612	789
2019	555347	1025
2020	623919	1252

Table 2.1, 2.2, 2.3,2.4 Summary of data

To better illustrate the distribution of EV numbers, map the EV number on the CA county.

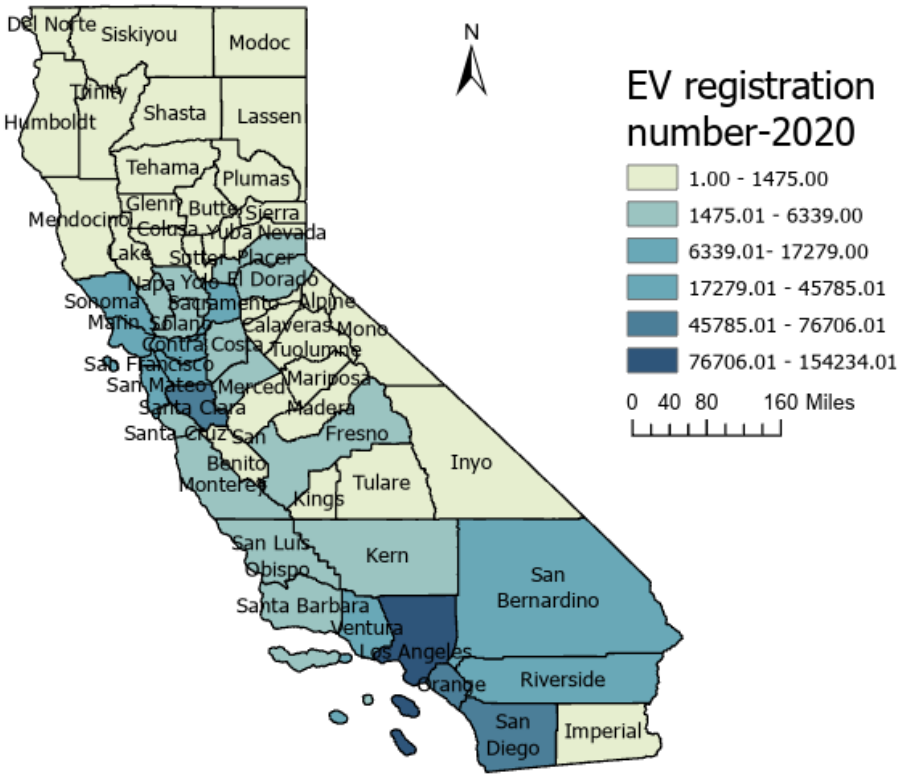

Figure 2.1 EV registration number in 2020

From the EV registration number in 2020, there is approximately a spatial cluster effect across the counties. Santa Clara and Los Angeles had the most EV registration, with the counties surrounding also had large EV numbers but smaller than that of them. For these 2 centers of EV registration, the counties nearby them and on the shore, like Sonoma, Orange and San Diego had more EV

Qingyuan Gao
qingyug3@uci.edu

registration than those inland counties. In contrast, counties in the north and in the mideast of CA had very small EV registration number.

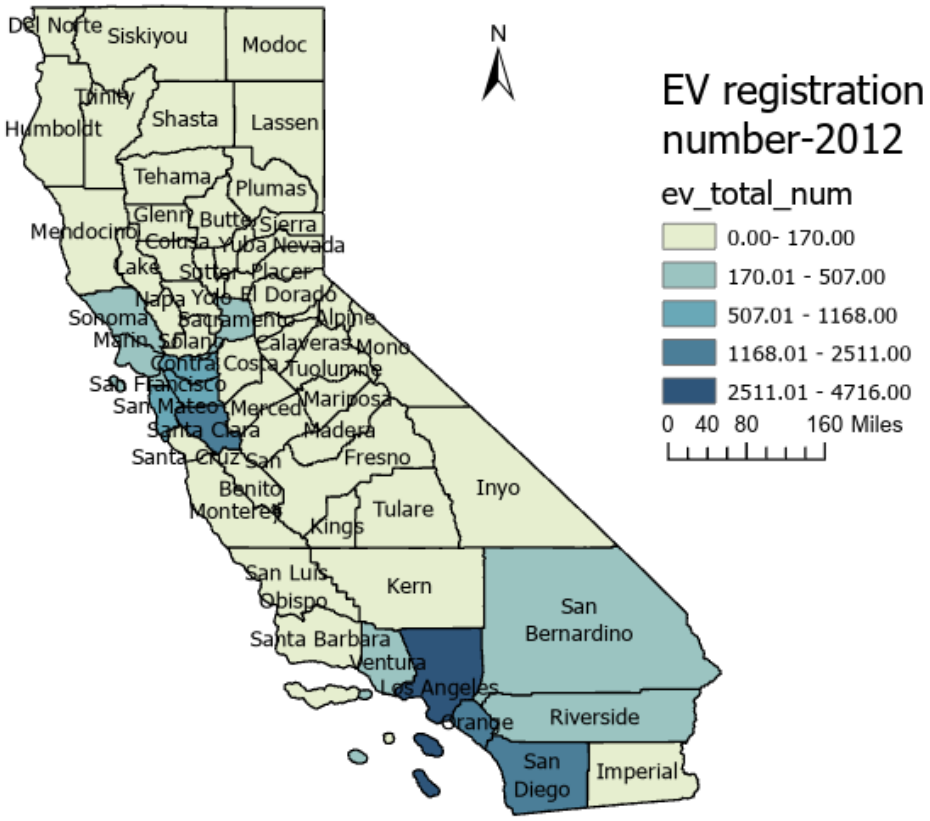

Figure 2.2 EV registration number in 2012

In comparison, back in 2012, the EVs were only clustered around Santa Clara and Los Angeles. Counties between these 2 centers, such as Santa Cruz, Monterey and San Luis had not witnessed many EV registrations. According to the dataset, only 30 of 58 counties had EV registration in 2012, and the rest counties did not have any EV registration. With years passing, more counties adopted EV and the expanding trend was approximately around these 2 centers.

3 Spatial Autocorrelation: Moran's I calculation

To evaluate whether there is spatial autocorrelation in the EV registration numbers, calculate the Global Moran's I of the full dataset and the local Moran's I for each county. The calculation was done in ArcGIS.

Qingyuan Gao
qingyug3@uci.edu

Year	Global Moran's I	z	p-value
2012	0.431	6.45	0.00
2013	0.401	6.12	0.00
2014	0.397	6.07	0.00
2015	0.397	6.03	0.00
2016	0.397	6.07	0.00
2017	0.411	6.32	0.00
2018	0.426	6.45	0.00
2019	0.440	6.64	0.00
2020	0.458	6.91	0.00

Table 3.1 Global Moran's I

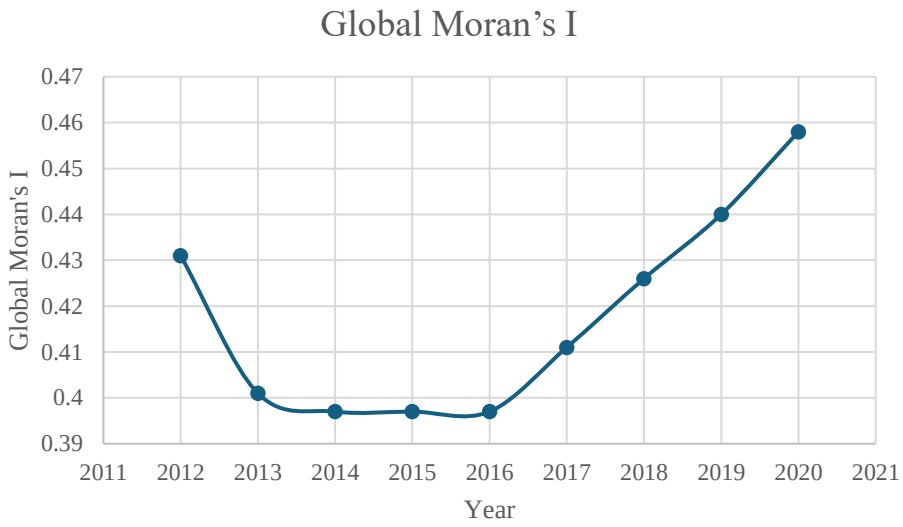

Figure 3.1 Global Moran's I over years

According to the Moran's I , the EV registration was highly clustered at the beginning of the study period, which is year 2012. Then, the EV registration dispersed from the year 2013 to year 2016. Starting from year 2017, the EV registration became increasingly clustered, and this trend still went on in year 2020.

Despite the changes, the Moran's I were above zero throughout the study period, which indicates the EV registration has a spatial autocorrelation across different counties, and the EV registration numbers tend to highly clustered based on the geographical location of these counties.

Qingyuan Gao
qingyug3@uci.edu

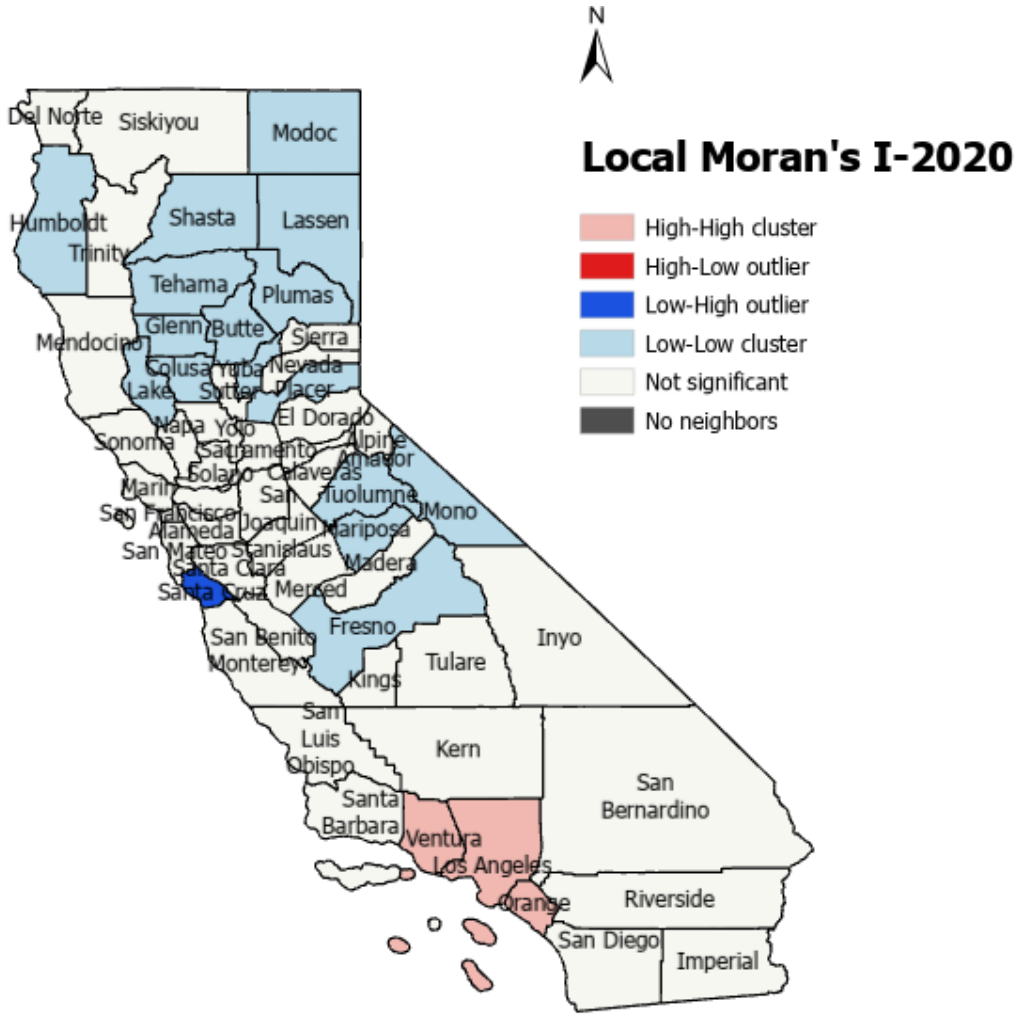

Figure 3.2 Local Moran's I in 2020

According to the local Moran's I in 2020, most counties with low EV registration number were clustered in the North of CA in 2020. 3 counties in the south, Ventura, Los Angeles and Orange were high in EV registration number and were clustered. Santa Cruz was an outlier with low EV registration but high EV registration in its neighbors.

In comparison, back in 2012, the distribution of EV registration is different.

Qingyuan Gao
qingyug3@uci.edu

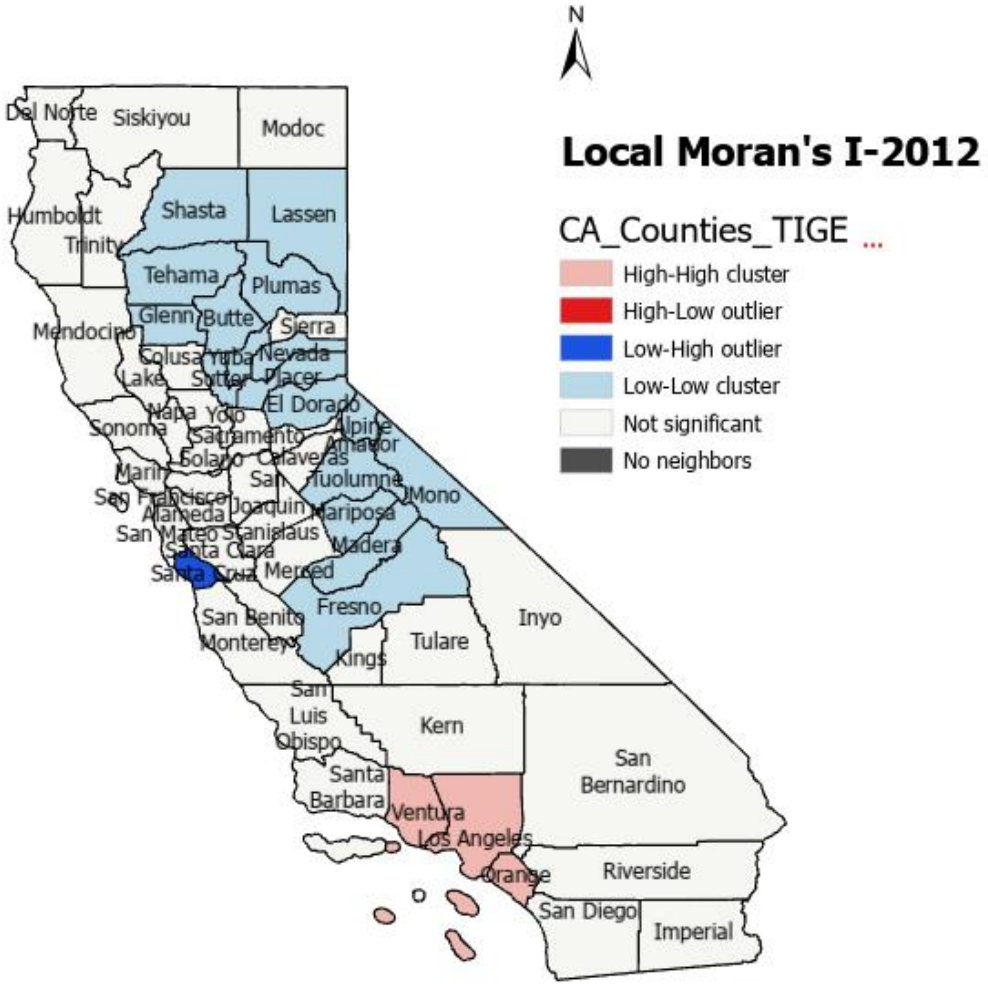

Figure 3.3 Local Moran's I in 2012

In 2012, the low EV registration counties were more clustered in the middle of CA. Counties in the north such as Modoc and Humbolt were not in the low-value cluster. Counties in the center CA like El-Dorado and Madera were within the low-value cluster in 2012 but not any more in 2020. Ventura, Los Angeles and Orange were still the high EV registration clusters.

4 Analysis of EV charger distribution

Previous sections mainly discussed the distribution of EV registration. Section 4 will discuss the distribution of EV chargers and its relationship to EV registration.

Qingyuan Gao
qingyug3@uci.edu

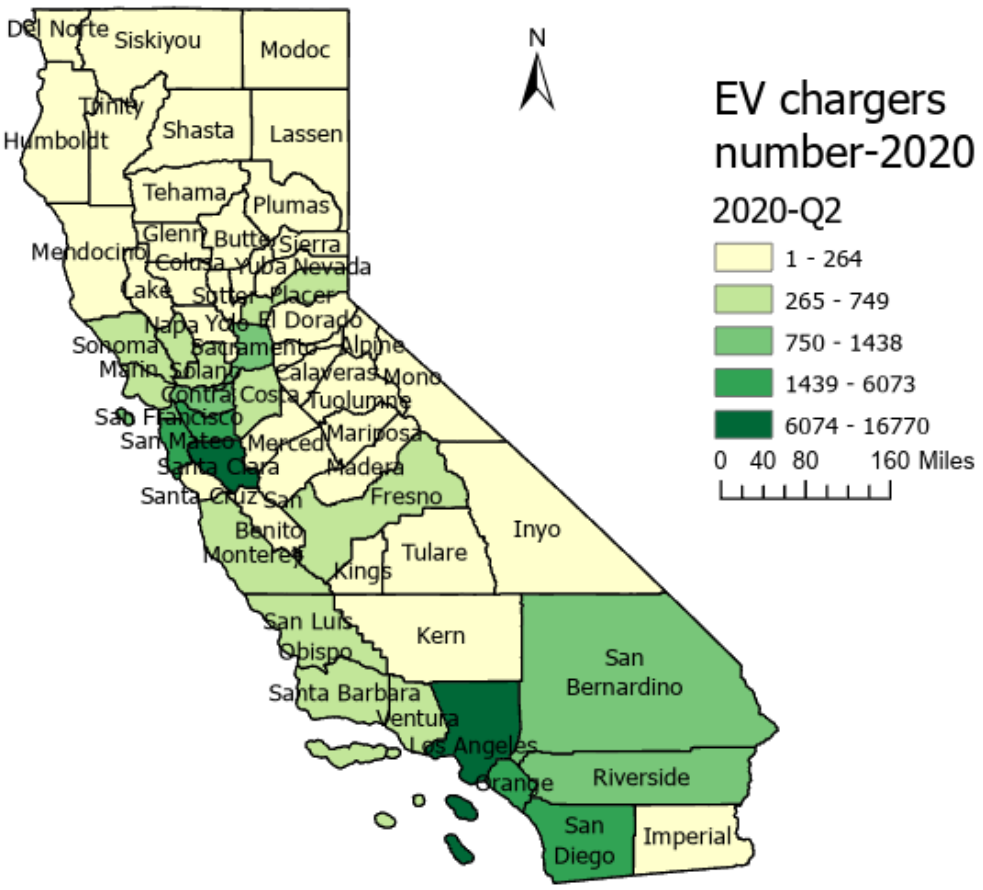

Figure 4.1 EV charger number in 2020

Qingyuan Gao
qingyug3@uci.edu

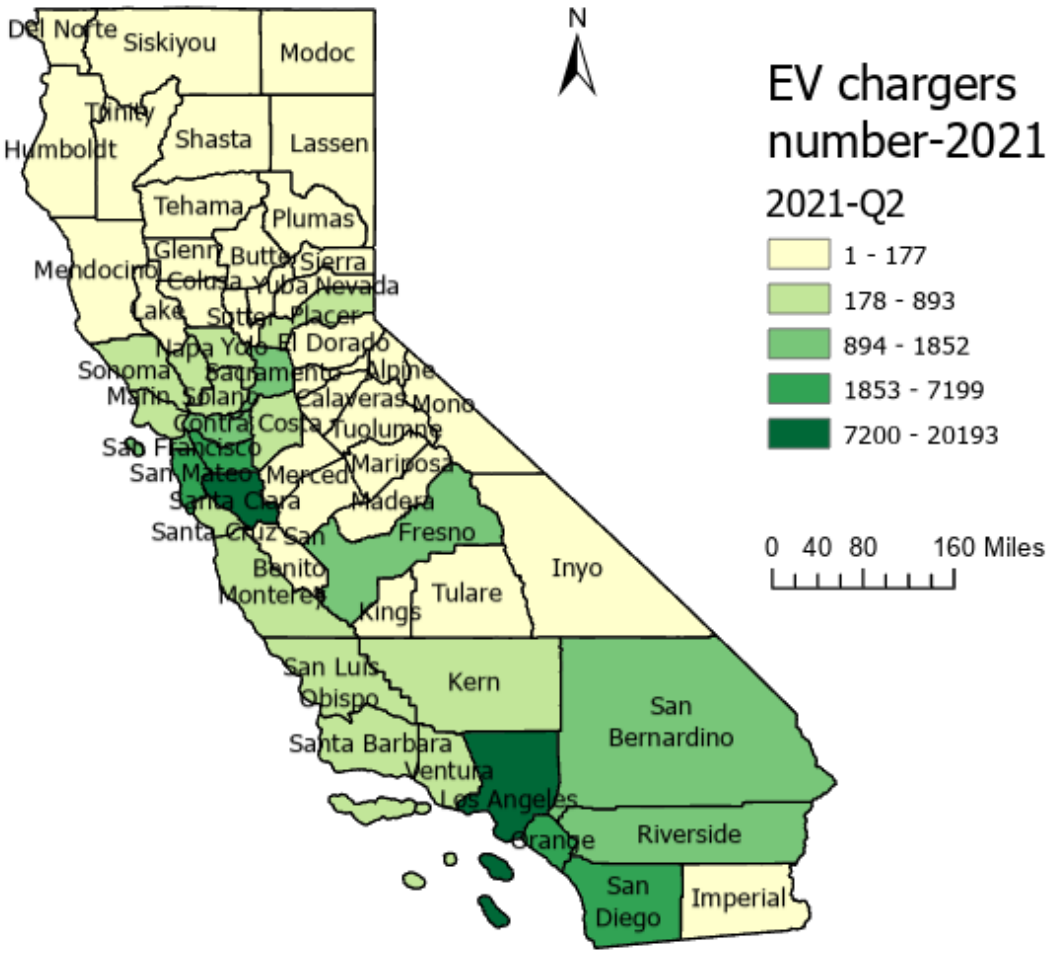

Figure 4.2 EV charger number in 2021

The distribution of EV chargers is very similar to the distribution of EV registration, with respect to relative number.

Besides the total number of EV chargers, the increase in EV charger number, or to say the newly built EV charger number can also be plotted.

Qingyuan Gao
qingyug3@uci.edu

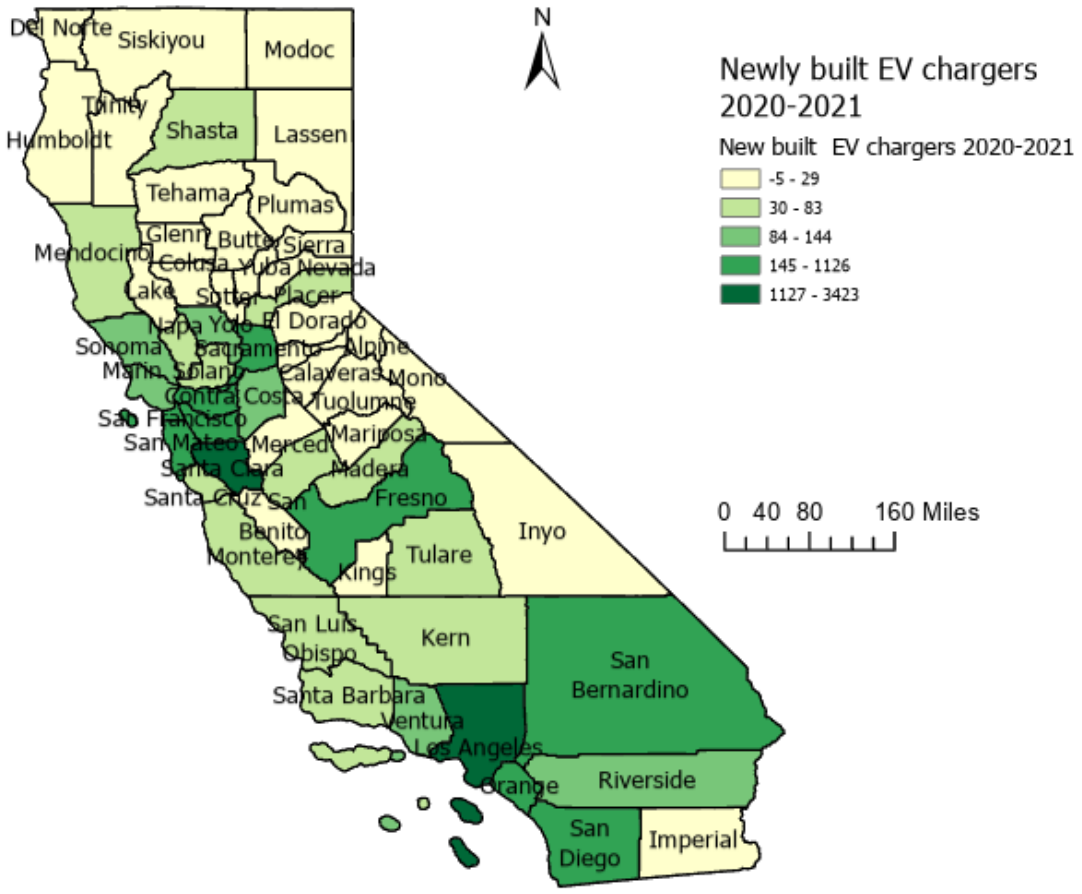

Figure 4.3 Newly built EV charger number 2020-2021

The newly built EV charger number has a similar distribution to the number of EV registration, but different in some counties. For example, Shasta and Mendocino in the north have a higher growth rate of EV chargers than other counties in the north. Though Fresno has a small number of EV registrations, its EV chargers number grew rapidly from 2020 to 2021.

Note that the AFDC Station Locator now counts the number of ports available to charge a vehicle rather than the number of connectors as previously counted. This results in a lower number of chargers. In addition, a phase-out of the 3G cellular network in Quarter 1 2022 has resulted in certain chargers being removed from the AFDC. These 2 problems make the EV chargers number lower than the actual count. Therefore, the study only selects data from 2020 to 2021 to perform the analysis.

Qingyuan Gao
qingyug3@uci.edu

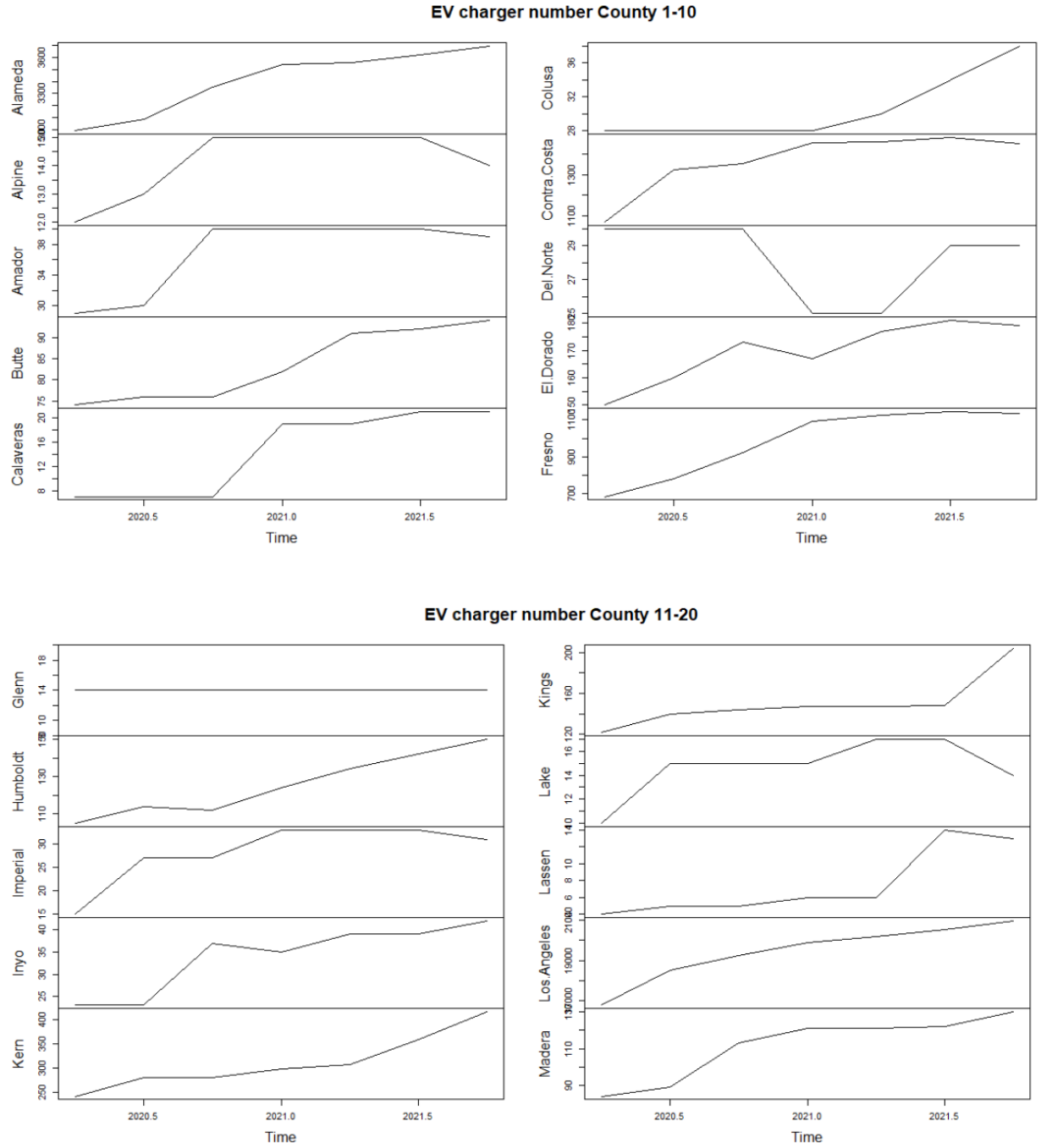

Figure 4.4,4.5 Time series data of EV charger number, county 1-20

Taking the first 20 counties for example, we can see that there are some differences between the EV chargers number patterns. But generally, there are quarterly and yearly trends for most of the counties. The EV charger number grows more rapidly in the 1st and 4th quarter of a year. Most counties' EV charger number increase more rapidly in 2020 than in 2021.

Qingyuan Gao
qingyug3@uci.edu

Suppose there is seasonality and trend in the data, which construct a time series data:

$$Y_t = S_t + T_t + \epsilon_t$$

Where Y_t is the EV charger number, S_t is the quarter effect(seasonality), T_t is the trend, and ϵ_t is the error.

According to the decomposition result in R, the trend , quarter effect and error term of the first 10 counties are listed below

	Alameda	Alpine	Amador	Butte	Calaveras
trend 2020-Q4	3313.50	14.13	36.13	79.13	11.50
trend 2021-Q1	3452.63	14.75	38.75	83.25	14.75
trend 2021-Q3	3563.50	14.88	39.88	87.50	18.25

	Q1	Q2	Q3	Q4	
quarter effect	5.92	-8.22	-3.68	5.97	

error	Alameda	Alpine	Amador	Butte	Calaveras
2020-Q4	33.53	-5.10	-2.10	-9.10	-10.47
2021-Q1	85.45	-5.67	-4.67	-7.17	-1.67
2021-Q3	4.72	8.34	8.34	11.72	8.97

Qingyuan Gao
qingyug3@uci.edu

	Colusa	Contra Costa	Del Norte	El Dorado	Fresno
trend 2020-Q4	28.25	1349.88	28.13	165.88	924.38
trend 2021-Q1	29.25	1418.75	27.38	171.88	1025.25
trend 2021-Q3	31.25	1450.50	27.13	175.25	1097.25

quarter effect					
error	Colusa	Contra Costa	Del Norte	El Dorado	Fresno
2020-Q4	-6.22	-0.85	-4.10	1.15	-9.35
2021-Q1	-7.17	30.33	-8.30	-10.80	60.83
2021-Q3	6.97	18.72	6.09	9.97	37.97

Table 4.1, 4.2 Time series analysis of EV charger numbers

From the table, except for Del Norte which has a decreasing trend in EV charger numbers, most counties have a growing trend in EV charger numbers. The seasonality, or the quarterly effect, is positive in the 1st quarter and the 4th quarter, which means the actual EV charger number will be larger than the trend. In comparison, the quarterly effect is negative in the 2nd and 3rd quarters, so the actual EV charger number will be lower than the trend. The quarter effect is especially large in the 2nd quarter, that the EV charger number will be lower than the trend by about 8.

5 Analysis of EV Registration Distribution

Similar to Section 4, time-series analysis can be done to the EV registration.

Qingyuan Gao
qingyug3@uci.edu

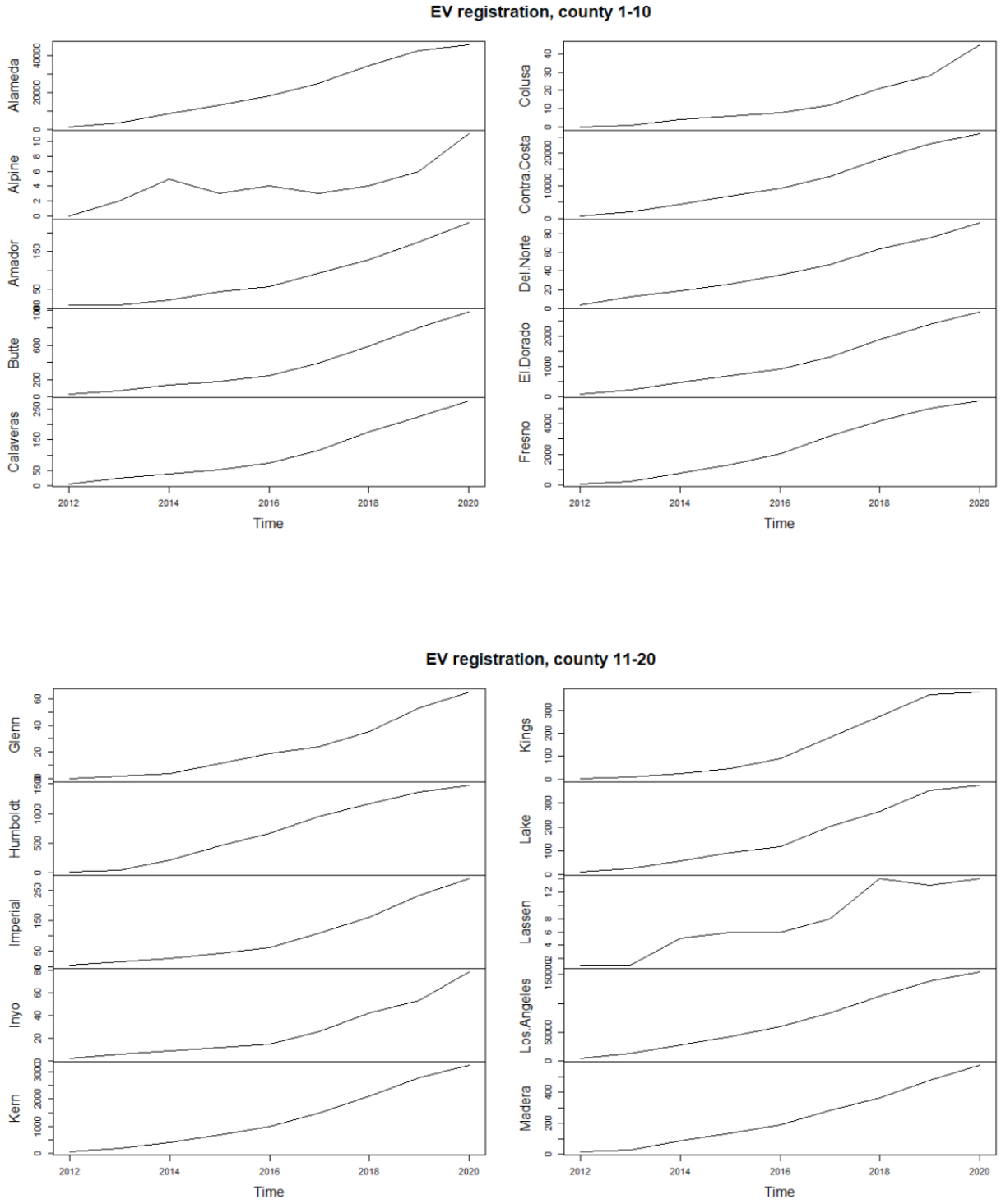

Figure 5.1,5.2 EV registration of county 1-20

Qingyuan Gao
qingyug3@uci.edu

According to the figures, most counties show an increasing growing rate in EV registration number from 2012 to 2018. However, in recent years from 2019-2020, this growing rate becomes lower.

Based on the EV registration patterns, assume that the EV registration follows a logistic pattern. EV number in 2012 is the base number. In 2020, the new registered EV approximately reaches the max, so it can be predicted that the new registered EV number will decrease in the following years over 2020, then the EV number will reach the max in approximately 2030. This can be justified by checking the cumulative EV registration number.

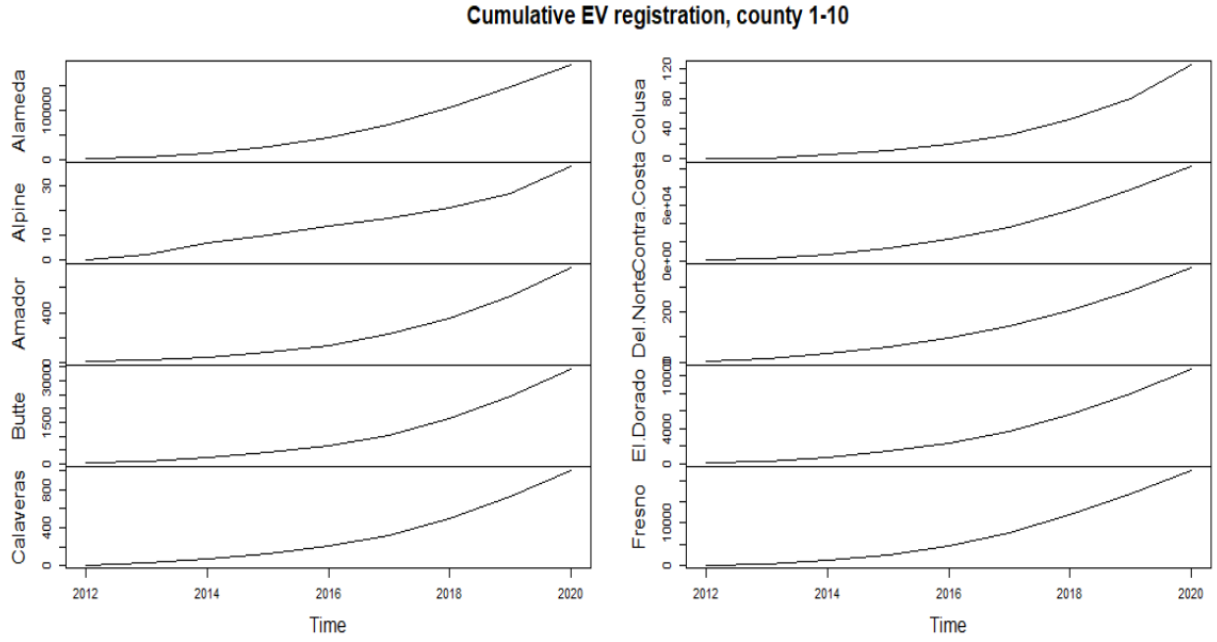

Figure 5.3 Cumulative EV registration of county 1-10

Qingyuan Gao
qingyug3@uci.edu

6 Spatial-temporal modeling

The study used panel data from year 2012 to year 2020, and each subset of data contained the EV registration number of each county and their demographic data, including population, income, education, household features and housing rent.

As there is spatial autoregression within the data, a spatial-temporal model is needed to interpret the changes in EV registration across different counties

A spatial lag model can be implemented to address the spatial autoregression, which is

$$Y = X\beta + \tau + \varepsilon$$

Where Y is the EV registration numbers, X is the vector of demographic variables, β is the parameters to estimate, τ is the error term that is spatially correlated, ε is the error term that is spatially independent.

The model can be estimated using *spmodel* package in R

The final specified model is:

$$EV_{total_{num}} = \beta_0 + \beta_1 total\_num\_hh + \beta_2 hh\_size + \beta_3\, median\_age +$$
$$+ \beta_4 median\_hh\_income + \beta_5 employment + \beta_6 higher\_education\_percentage$$
$$+ \beta_7 median_{housing_{rent}} + \tau + \varepsilon$$

Qingyuan Gao
qingyug3@uci.edu

The modeling result is shown in the table below:

```
Coefficients (fixed):
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.070e+04  9.334e+03   2.218 0.026571 *
total_num_hh              -4.654e-01  1.414e-02 -32.904  < 2e-16 ***
hh_size                   -6.140e+03  1.993e+03  -3.081 0.002066 **
median_age                -3.270e+02  1.169e+02  -2.798 0.005148 **
median_hh_income          -2.521e-02  5.484e-02  -0.460 0.645760
employment                 3.417e-01  9.851e-03  34.688  < 2e-16 ***
higher_education_percentage  2.405e+01  7.502e+01   0.321 0.748532
median_housing_rent        1.120e+01  2.889e+00   3.876 0.000106 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.8094

Coefficients (exponential spatial covariance):
       de        ie     range
1.881e+07 1.881e+07 1.022e+00
```

Table 6.1 Spatial regression model result

In conclusion, number of households, household size, median age , employment and housing rent significantly influence the EV registration number. The increase in housing rent and employment contribute to the increase in EV registration, while the number of households, household sizes and people's age will negatively impact on the EV numbers.

There is also more complicated model which contains the spatial lag term, which is

$$Y = X\beta + \delta WY + \varepsilon$$

Where W is the weight matrix given by the k-nearest neighbor counties, δ is the corresponding coefficients of the lag term. The model can be constructed using the *splm* package, but there is no time for the model scripting. I hope I can investigate this model in the future.