County-level and point-level analysis of the relationship between political inclination and frequency of electric-vehicle charging stations in New England

by Anonymous

Abstract

Electric-vehicle charging stations play a key role in the adoption of electric vehicles. Therefore, it's important to understand how different factors are related to the frequency of charging stations in a region. Existing work has shown that U.S. states with the Republican candidate winning most votes for both the 2016 and 2020 elections tend to have fewer charging stations. Motivated by this study, we seek to answer whether this discovered relationship between political inclination and frequency of charging stations holds at two finer spatial scales, the county level and the point level, in the New England region of the United States. For the county-level analysis, we explored linear regression and four spatial regression models. For the point-level analysis, we explored the non-homogeneous Poisson process (NHPP). In both analyses, we used three county-level explanatory covariates constructed from the 2021 American Community Survey data, and one county-level explanatory covariate constructed from the 2020 U.S. presidential election vote counts for quantifying political inclination. The two final models, the linear regression model and the NHPP, both predict that counties with more Democratic votes than Republican votes tend to have more charging stations but disagree on the direction or discernibility of other explanatory covariates considered.

Keywords: clean energy, U.S. politics, spatial statistics

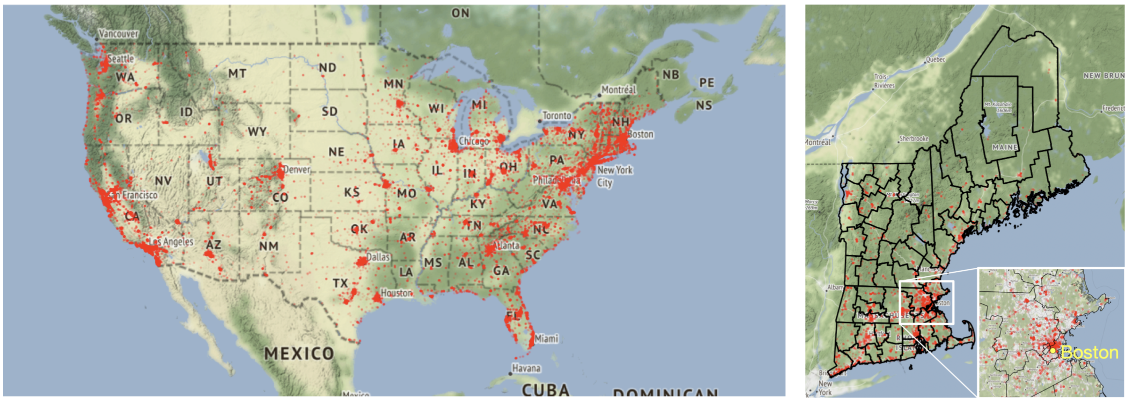
1 Introduction

As the world grapples with the challenges of global warming, transferring from traditional forms of transportation to more sustainable ones is becoming increasingly important. Among many forms of sustainable transportation, electric vehicles have gained significant traction over the last decade and can now be commonly spotted on roads. Compared to traditional vehicles that directly operate on fossil fuels, electric vehicles currently provide a similar amount of greenhouse gas emissions after taking into account their manufactoring process and the emissions due to electricity generation – this amount is projected to greatly reduce as electricity are generated from cleaner sources [13]. However, the adoption of electric vehicles heavily rely on an accessible network of charging stations. While many factors have been shown to affect/correlate with the frequency of charging stations in a region, the relationship between political inclination and their frequency remains underexplored.

To the best of our knowledge, [9] is the most relevant existing study on the relationship between political inclination and the frequency of charging stations. This study showed that, after accounting for median household income and highway density, U.S. states with the Republican candidate winning most votes for both the 2016 and 2020 elections tend to have fewer charging stations (page 11 of [9]). This result is perhaps not suprisingly: Democrats have been shown to have higher willingness for adopting electric vehicles [18] and a 2015-2016 survey of vehicle owners across U.S. shows that electric vehicles were indeed most popular among Democrats [4]. We also found studies that looked into the relationship between other factors and the frequency of charging stations. For instance, there tends to be a lower number/density of public charging stations in low-income, Black or Latino communities in New York City [10] and the state of California [8]. There are also studies that look at optimizing the distribution of charging stations to maximize certain utility metrics [6].

In our study, we seek to answer whether the relationship between political inclination and the frequency of charging stations discovered by [9] not only holds on the state level but also on two finer spatial scales, the county level and the point level. Our motivation is that changing the unit of aggregation can sometimes affect the outcome of inference (counties are much smaller than states), and modeling the variation of charging stations on a finer scale for a region of interest is by itself an interesting case study. Unlike [9], we focus on the New England region of the United States instead of the entire United States, although the same analysis can be easily extended. Methodologically, we follow [9] and use regression-style models to explain variations in the frequency of charging stations using a linear combination of explanatory covariates, including one that measures political inclination on a continuous scale rather than the binary scale used in [9] to allow for more fine-grained interpretations. We approached the county-level analysis with linear and spatial regression models and the point-level analysis with the non-homogenous Poisson process. To answer our research question, we interpreted the fitted coefficients of each model in the context of our data.

This paper is organized as follows. Section 2 outlines the data sources and the preprocessing steps for transforming the raw data into data that are ready for subsequent analyses. Section 3 (Methods) describes the exploratory and formal methods we used to analyze areal and point-level data. The results from these methods are then reported and interpreted in Section 4 (Results). Finally, in Section 5 (Discussion), we summarize main findings and discuss caveats and ideas for future work.



(a) (b)

Figure 1. (a) Locations (red points) of all in-operation public electric-vehicle charging stations in 48 U.S. states (not showing Hawaii and Alaska) by May 2023. We see more stations along the east and west coasts and in/near major cities. (b) Locations of charging stations subsetted to New England. Black polygons represent counties. We see a major cluster of charging stations near Boston.

2 Data

2.1 Data sources

To conduct the study, we relied on three sources of data. Our first source of data is an update-to-date (by April 10, 2023) point-level dataset containing locations of all in-operation public electric-vehicle charging stations in the United States and Canada (Figure 1a) downloaded from the United States Department of Energy's Alternative Fuels Data Center [15]. These locations are collected by the National Renewable Energy Laboratory from a variety of sources including trade media, Clean Cities coalitions, online Submit New Station forms, equipment manufacturers, and so on [14]. Since we are only interested in the frequency of charging stations within New England, we subsetted this point-level dataset to only charging stations within New England[1] (Figure 1b).

Our second source of data is the American Community Survey (ACS) data. The ACS is a "large demographic survey collected using mailed questionaires, telephone interviews, and from [U.S.] Census Bureau representatives to about 3.5 million household addresses annually" [17] and "covers U.S. residents in all 3141 counties in the 50 [U.S.] states, the District of Columbia and all 78 municipalities in Puerto Rico" [16]. Using the `tidycensus` package in R [19], we accessed the 2021 ACS database and downloaded the following county-level census variables for each county in New England: the size of its total population, sizes of its populations of different races (including White, Black, American Indian and Alaska Native, Asian, Native Hawaiian and other Pacific Islander, and others), sizes of its populations with different levels of higher education (including bachelor's degree, master's degree, doctorate degree and professional-school degree), the size of its unemployed population, and its median household income (in US dollars).

Our third and final source of data is the county-level vote counts for the 2020 U.S. presidential election of all counties in the United States [3]. Specifically, for each county, this dataset records separate vote counts for the Democratic party, the Republican party, the

[1] We excluded two counties in New England throughout our analysis: Dukes County and Nantucket County of Massachusetts. These two counties are islands that are not connected to the mainland via bridges and are considered as tourist destinations, so they could exhibit different spatial dynamics from the rest of New England. For the remainder of this paper, "New England" refers to the actual New England region *without* these two counties.

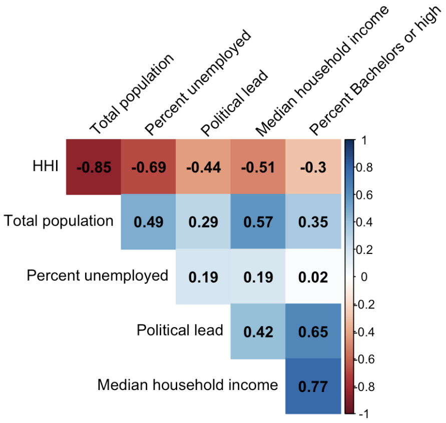


Figure 2. Correlations between pairs of candidate explanatory covariates.

Green party, the Libertarian party, and other parties. We also subsetted this dataset to only counties in New England.

2.2 Construction of new census-level covariates

Using raw census-level covariates and the point-level data mentioned in Section 2.1, we constructed new census-level covariates that are more informative. To calculate the number of charging stations per person of a county, which we refer to as the *rate* of charging stations, we divided the county's aggregated number of charging stations by the size[2] of its total population. To calculate the *percent unemployed* of a county, we divided the size of its unemployed population by the size of its total population. To calculate the *percent with Bachelor's degree or higher* of a county, we added the sizes of its populations with bachelor's degree, master's degree, doctorate degree and professional-school degree and divided this sum by the size of its total population. To calculate the *Herfindahl–Hirschman index* (HHI) [7] of a county, a measure of racial diversity (with higher values indicating lower racial diversity), we applied the following formula:

$$\text{HHI} = \frac{(\text{size of Race } 1)^2}{\text{size of total population}} + \frac{(\text{size of Race } 2)^2}{\text{size of total population}} + \cdots + \frac{(\text{size of Race } 6)^2}{\text{size of total population}},$$

where the six race categories we considered are (as in Section 2.1) White, Black, American Indian and Alaska Native, Asian, Native Hawaiian and other Pacific Islander, and other races. Finally, to calculate the *political lead* by the Democratic party over the Republican party in a county, we subtracted the number of Republican votes in the county from number of Democratic votes in the county and divided this difference by the total number of votes in the county. We chose to use this difference instead of just the percent of votes for the Democratic party or the percent of votes for the Republican party. This is because there is a non-negligible amount of votes for other parties and taking the difference allows us to focus on the political inclination towards the two major parties.

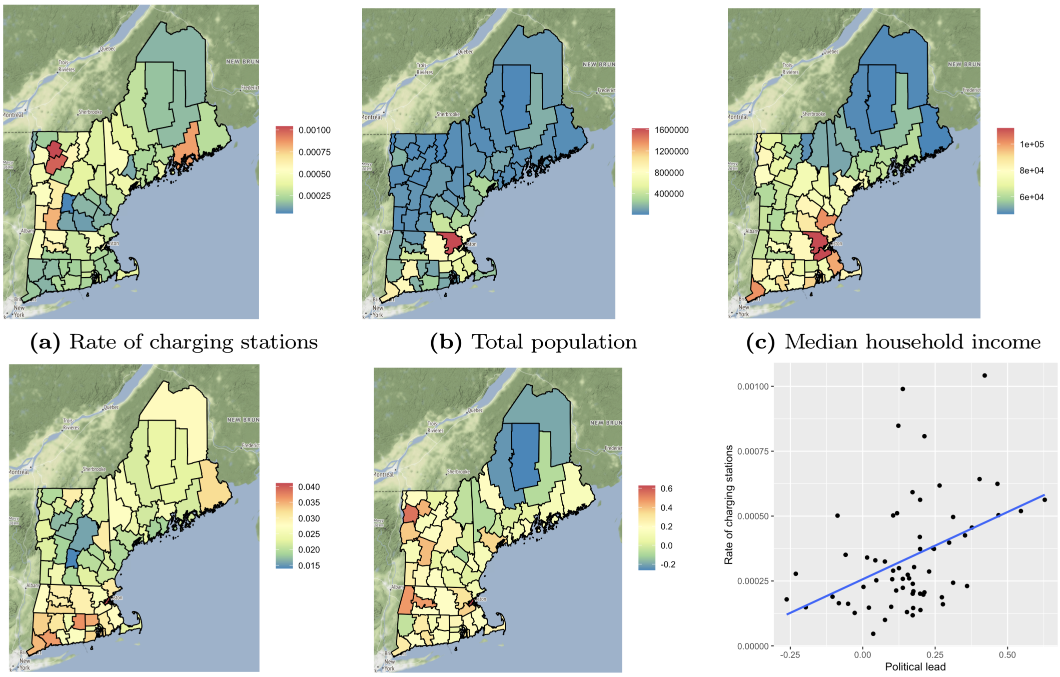
2.3 Variable selection

Among the six candidate explanatory covariates (including total population, median household income, percent unemployed, percent with Bachelor's or higher, HHI and political lead), we found some to be highly correlated with others (Figure 2). More specifically,

[2]. "size" means "number of people".

Covariate	*VIF before variable selection*	*VIF after variable selection*
Total population	4.89	1.90
Median household income	3.76	1.70
% unemployed	2.19	1.34
% with Bachelor's or higher	4.56	NA (removed)
HHI	7.88	NA (removed)
Political lead	2.61	1.23

Table 1. Variance-inflation factors (VIFs) before and after variable selection.



(a) Rate of charging stations

(b) Total population

(c) Median household income

(d) Percent unemployed

(e) Political lead

(f) Rate against political lead

Figure 3. (a) Choropleth of the rate or charging stations. (b-e) Choropleths of the four selected explanatory covariates. (f) An exploratory scatter plot of the rate of charging stations against political lead.

we found HHI to be highly negatively correlated with both total population and percent unemployed, and percent Bachelor's or higher to be highly positively correlated with both political lead and median household income. We also found high variance-inflation factors (see the first column of Table 1). Variance-inflation factors measure *multicollinearity*, i.e., how well each explanatory covariate can be predicted by a linear regression over other explanatory covariates; a value of one indicates no multicollinearity and higher values indicate stronger multicollinearity. Since we will be using linear combinations of explanatory covariates in the statistical models (Section 3), multicollinearity is particularly problematic because it decreases the statistical discernibility of model coefficients. To combat multicollinearity, we removed HHI and percent Bachelor's or higher, which noticeably reduced the variance-inflation factors for the four remaining covariates (see the second column of Table 1). The choropleths of the rate of charging stations and these four selected explanatory covariates are shown in Figure 3a-e. Figure 3f shows a scatter plot of the rate of charging stations against political lead; we see that these two covariates appear to be mildly positively correlated.

3 Methods

In this section, we discuss methods for modeling areal data as well as point-level data.

3.1 Areal data

3.1.1 Linear regression: an introductory model for areal data

We are interested in explaining variations in the county-level rate[3] of charging stations using three county-level covariates: median household income, political lead and percent unemployed. This data qualifies as *areal data* because both the response and the explanatory variables are aggregated quantities over each areal unit (e.g., each county) in a spatial window (e.g., New England). An introductory model for areal data is the *linear regression* model:

$$Y = X\beta + \varepsilon \quad \text{with} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma), \tag{1}$$

where Y_i is the response variable of areal unit i, the i-th row of X ($X_{i,:}$) contains the explanatory variables of areal unit i, β is the vector of coefficients, and σ is the standard deviation of residuals ε_i. In the context of our areal dataset, X in Equation 1 would have 4 columns representing the three explanatory variables plus a dummy variable of one and 65 rows representing the 65 counties in New England (excluding two islands); y in Equation 1 would also have 65 rows containing the rates of charging stations of 65 counties. As a result, β and ε would be column vectors with 4 and 65 rows respectively.

Importantly, the linear regression model assumes that the residuals are independently and identically distributed. However, this assumption might be violated when variations in the response variable are not fully captured by the selected explanatory variables. Therefore, after fitting the linear regression model, we conduct hypothesis tests on whether its residuals exhibit spatial association with the help of two test statistics, Moran's I and Geary's C, which we discuss in Section 3.1.3. If the tests indicate spatial association of the residuals, we must fit spatial regression models that capture such spatial association; these models are discussed at length in Section 3.1.4.

3.1.2 Neighborhood matrices

The *neighborhood* matrix, which we denote by W, is a fundamental concept underlying Moran's I, Geary's C, and the spatial regression models. Let n denote the number of areal units. W is a $n \times n$ matrix with all diagonal entries being zero and all off-diagonal entries falling between 0 and 1. Intuitively, $W_{i,j}$ quantifies the extent to which areal units i and j are considered as neighbors.

W can be defined in many styles. For example, in the spatial-adjacency style, we set $W_{i,j} = 1$ if the areal units i and j share a boundary and $W_{i,j} = 0$ otherwise; in the k-nearest-neighbor style, we set $W_{i,j} = 1$ if and only if the areal unit j is among the k closest neighbors of areal unit i by centroid distance. Finally, to create the row-standardized version of a neighborhood matrix, we simply divide all entries in each row by the sum of all entries in that row.

[3] We also considered using the log-transformed rate as the response variable in Section 4 (Results) and found that doing so gave more reasonable residuals.

3.1.3 Testing for spatial association of residuals with Moran's I and Geary's C

As we mentioned in Section 3.1.1, after fitting a linear regression model, we need to perform tests on whether there's spatial association among areal units in terms of their residuals. To do this, we set up the null hypothesis as "there is no spatial association in the residuals" and the alternative hypothesis as "there is spatial association in the residuals". However, we need test statistics to quantify the degree of spatial association.

One such statistic, the *Moran's I* statistic [12], is defined as $I = (n/\sum_i \sum_j W_{i,j})(\sum_i \sum_j W_{i,j}(\varepsilon_i - \bar{\varepsilon})(\varepsilon_j - \bar{\varepsilon}))/(\sum_i (\varepsilon_i - \bar{\varepsilon})^2)$, where we choose W to be the row-standardized spatial-adjacency neighborhood matrix, and $\bar{\varepsilon}$ is the mean of all residuals. I usually falls in $[-1, 1]$; it is positive when there's positive spatial association, negative when there's negative spatial association, and zero when there's no spatial association. To conduct the aforementioned test using Moran's I, we perform the following procedure known as *permutation testing*: we first simulate 50000 I values under the null hypothesis by randomly permuting the n residuals to n areal units 50000 times and computing an I value each time; we then calculate the p value as the percent of simulated I's that are greater[4] than the observed I. Finally, if $p < 0.05$, we say that Moran's I is statistically discernible from zero and reject the null that there's no (positive) spatial association; otherwise, we do not reject the null.

Another statistic, the *Geary's C* statistic [5], is defined as $C = ((n-1)\sum_i \sum_j W_{i,j}(\varepsilon_i - \varepsilon_j)^2)/(2(\sum_i \sum_j w_{i,j})\sum_i (\varepsilon_i - \bar{\varepsilon})^2)$, where W is again the row-standardized spatial-adjacency neighborhood matrix. By definition, $C \geq 0$. $C = 1$ when there's no spatial association, $C < 1$ when there's positive association, and $C > 1$ when there's negative spatial association. To conduct the aforementioned test using Geary's C, we follow the same procedure for Moran's I discussed above, except that the p value is now computed as the percent of simulated C's that are *less* than the observed C.

3.1.4 Spatial regression models

If we reject the null hypothesis of no spatial association among the linear regression residuals, then we must fit other models that take into account such spatial association. In this work, we focus on four alternatives: the spatial lag model, the spatial Durbin model, the spatial error model, and the CAR model. We will use the Akaike Information criteria (AIC) [1] to compare them (smaller AIC is better). For a model with fitted parameters θ, its AIC is defined as

$$\mathrm{AIC} = 2|\theta| - 2\ln \mathcal{L}(\theta)$$

where $|\theta|$ is the number of parameters in the model and $\mathcal{L}$ is the model's likelihood function. We see that the AIC is smaller when the fitted model is a better fit (as measured by $\mathcal{L}$) but also penalizes (i.e. by becoming larger) the number of parameters to discourage overfitting. In the remainder of this section, let n again denote the number of areal units and d denote the number of covariates.

The *spatial lag* model (page 305 of [2]) is defined as

$$Y = X\beta + \rho W Y + \varepsilon \quad \text{with} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma),$$

[4] In practice, we only look for evidence of positive association because it is the most common problem.

where $\rho \in \mathbb{R}$, W is some neighborhood matrix, and X and Y are the same X and Y defined in Section 3.1.1. Here, the response Y_i of areal unit i depends *not only* on its covariates $X_{i,:}$ *but also* on the weighted sum (or weighted average if W is row-standardized) of response over its neighboring areal units (also called the "lagged" response):

$$Y_i = X_{i,:}\beta + \rho W_{i,:} Y + \varepsilon_i = X_{i,:}\beta + \rho\left(\sum_{j=1}^{n} W_{i,j} Y_j\right) + \varepsilon_i. \quad (W_{i,:} \text{ denotes the } i\text{-th row of } W)$$

The *spatial Durbin* model (page 305 of [2]) is defined as

$$Y = X\beta + \rho W Y + W X \gamma + \varepsilon \quad \text{with} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma),$$

where γ is a d-by-1 matrix. This model extends the spatial lag model. Here, the response Y_i of areal unit i depends *not only* on its covariates $X_{i,:}$ and the weighted sum of response over its neighboring areal units *but also* on the weighted sums of covariates over neighboring areal units (also called the "lagged" covariates):

$$Y_i = X_{i,:}\beta + \rho\left(\sum_{j=1}^{n} W_{i,j} Y_j\right) + \sum_{k=1}^{d} \gamma_k (W_{i,:} X_{:,k}) + \varepsilon_i = X_{i,:}\beta + \rho\left(\sum_{j=1}^{n} W_{i,j} Y_j\right) + \sum_{k=1}^{d} \gamma_k \left(\sum_{j=1}^{n} W_{i,j} X_{j,k}\right) + \varepsilon_i,$$

where $\sum_j W_{i,j} X_{j,k}$ is the weighted sum of the k-th covariate over the neighbors of areal unit i.

The *spatial error* model (page 305 of [2]) is defined as

$$Y = X\beta + \lambda W (Y - X\beta) + \varepsilon \quad \text{with} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma),$$

where $\lambda \in \mathbb{R}$. This is similar to the spatial lag model except that $Y - X\beta$ are used in place of Y.

The *conditional autoregressive* (CAR) model (page 298 of [2]) is defined as

$$Y_i | Y_{j\sim i} \sim \mathcal{N}(X\beta + \lambda W (Y - X\beta), \ldots) \quad \text{with} \quad e_i | e_{j\sim i} \sim \mathcal{N}\left(\sum_{j\sim i} \frac{c_{i,j} e_j}{\sum_{j\sim i} c_{i,j}}, \frac{\sigma_{e_i}^2}{\sum_{j\sim i} c_{i,j}}\right),$$

where $Y_{j\sim i}$ and $\varepsilon_{j\sim i}$ denote the Y and ε of the neighbors of areal unit i; $c_{i,j}$'s are estimated from data. Both the spatial error model and the CAR model have the same mean for Y_i, but they specify the covariance structure of errors or residuals differently: in the CAR model, the covariance structure is setup so that Y_i depends on residuals of only the first-order neighbors of i; this property is known as *memorylessness*.

3.2 Point-level data

3.2.1 Motivation for modeling point-level data

While modeling data on the areal level is straightforward, it has some notable disadvantages. First, aggregating the number of charging stations across counties ignores the distribution of charging stations within counties. Doing so may also lead to the *modifiable areal unit* problem, where changing the unit of aggregation also changes the conclusions of parameter inference. To avoid these problems, we consider methods for directly modeling charging stations on a point level.

3.2.2 Homogeneous Poisson process

In a homogeneous Poisson process (HPP) (page 183 of [2]), we assume that points (e.g., charging stations) are generated within a spatial window W (e.g., New England) via the probabilistic model:

$$\begin{aligned} m &\sim \text{Pois}(\lambda \times |W|) \\ s_1, \dots, s_m &\overset{i.i.d}{\sim} \text{Unif}(W), \end{aligned}$$

where λ is the intensity[5], $|W|$ denotes the area of W and Unif(W) denotes a uniform distribution over W. Given a dataset of m points, one can show that $\hat{\lambda} = m / |W|$ is an unbiased estimator of λ.

Since points in a dataset generated by an HPP are independently and uniformly distributed across W, we say that this dataset exhibits "complete spatial randomness" (CSR). Datasets that do not exhibit CSR need to be modeled by more sophisticated models that capture the spatial variation of the intensity (i.e., no longer as a constant λ but rather as a spatially-varying function $\lambda(s)$ for $s \in W$); we discuss one introductory model of this kind in Section 3.2.4. Two common exploratory tools used to evaluate a point-level dataset's deviation from CSR are the G function and the F function; we discuss these in detail in Section 3.2.3. If these tools show that the locations of charging stations exhibit CSR, then there's no need to proceed with more sophisticated models.

3.2.3 Testing for complete spatial randomness with *G* and *F* functions

One tool for diagnosing CSR is the G function (page 179 of [2]). For an HPP with parameter λ, the probability that the Euclidean distance between a point generated by the HPP and its nearest-neighbor point generated by the HPP is below r can be shown to be

$$G(r) = 1 - \exp(-\lambda \pi r^2), \tag{2}$$

which is known as the theoretical G function. For a dataset of m observed points, the empirical G function is defined as the empirical cumulative distribution function of m distances, one for each observed point, where each distance is the Euclidean distance between an observed point and its nearest-neighbor observed point. To test for CSR, we plot the empirical G function of the dataset of interest together with (i) the theoretical G function under CSR (Equation 2 with $\lambda = m / |W|$) and (ii) the *simulation envelope* whose upper and lower bounds capture all empirical G functions of datasets (each with m points) sampled under CSR (i.e., independently and uniformly across W). We used a software that accomplishes these steps while also accounting for edge correction.

The interpretation of such a plot is straightforward. If the empirical G function of the observed point-level dataset lies above the simulation envelope, then observed points and their nearest-neighbor observed points are closer together than expected under CSR, indicating spatial clustering. If this observed G function lies below the simulation envelope, then observed points and their nearest-neighbor observed points are farther away than expected under CSR, indicating spatial repulsion. Finally, if this observed G function lies within the simulation envelope, then we say that the observed dataset exhibits CSR.

[5] Here λ can be interpreted as the average number of points per unit area in W. More formally, the intensity at a location $s \in W$, $\lambda(s)$, is defined as the limit of $n(A)/|A|$ as $|A| \to 0$, where A is a circle centered at s, $|A|$ is the area of A and $n(A)$ is the number of points in A. For an HPP, since $\lambda(s)$ is constant, these two definitions are equivalent.

Another tool for diagnosing CSR is the F function (page 181 of [2]). For an HPP with parameter λ, the probability that the Euclidean distance between an arbitrary point in W and its nearest-neighbor point generated by the HPP is below r can be shown to be

$$F(r) = 1 - \exp(-\lambda \pi r^2), \tag{3}$$

which is known as the theoretical F function. The empirical F function is defined similarly to the empirical G function except that each distance is now the Euclidean distance between an arbitrary point in W[6] and its nearest-neighbor observed point. To test for CSR, we also create a plot of the empirical F function of the dataset of interest together with (i) the theoretical F function under CSR (Equation 3 with $\lambda = m/|W|$) and (ii) the simulation envelope whose upper and lower bounds capture all empirical F functions of datasets (each with m points) sampled under CSR.

If the empirical F function of the observed point-level dataset lies above the simulation envelope, then arbitrary points and their nearest-neighbor observed points are closer together than expected under CSR, indicating less empty spaces and hence spatial repulsion among observed points. If this observed F function lies below the simulation envelope, then arbitrary points and their nearest-neighbor observed points are farther away than expected under CSR, indicating more empty spaces and hence spatial clustering among observed points. Finally, if this observed F function lies within the simulation envelope, then we say that the observed dataset exhibits CSR.

3.2.4 Non-homogeneous Poisson process: an introductory model for point-level data

If the observed point-level dataset does not exhibit CSR, we need another model that captures its spatially-varying intensity pattern. Here, we consider an introductory model, the non-homogeneous Poisson process (NHPP) (page 184 of [2]). In the NHPP, the intensity function is defined as

$$\lambda(s) = \exp(X(s)^T \beta), \tag{4}$$

where $X(s)$ the column vector of explanatory covariates at point $s \in W$ and β is again the column vector of coefficients. Here, $X(s)$ has five rows corresponding to the three covariates used in Section 3.1.1 and 3.1.4, total population (previously, in Section 3.1.1 and 3.1.4, total population was taken into account by using the rate of charging stations as the response), and a dummy variable of one.

This version of the intensity function seeks to explain all variations in intensity using only a linear combination of explanatory covariates and, since we only have access to the explanatory covariates on the county level (as discussed in Section 2), is constant within counties. In Section 5, we discuss these limitations at length and mention another point-level model that doesn't have these limitations. Nevertheless, the NHPP is a good starting point for modeling point-level datasets and can be easily extended to more complex point-level models.

Similar to the HPP, the NHPP model assumes the following probabilistic model: first, the number of points m is sampled from a Poisson distribution with rate parameter $\lambda(W) = \int_W \lambda(s)\, ds$; then, m points are independently and identically sampled according to the intensity function, which is like a unnormalized probability distribution over W. One can

[6] In practice, a grid of arbitrary points is created within W.

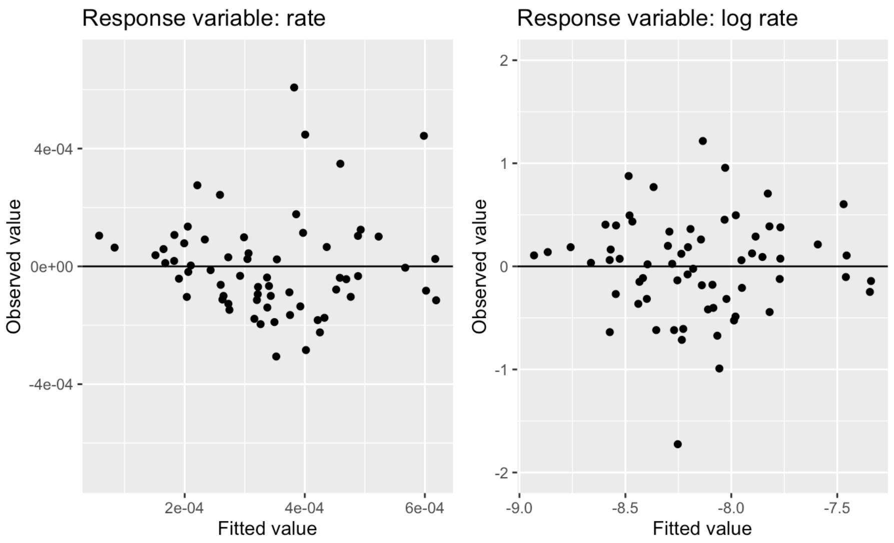


Figure 4. Linear regression residuals when the rate of charging stations is used as the response variable (left) and when the log rate of charging stations is used as the response variable (right).

Response	*Political lead included?*	*Moran's I*	*Geary's C*
log rate	Yes	0.071 ($p = 0.1471$)	0.905 ($p = 0.1529$)
log rate	No	**0.265** ($p = 8 \times 10^{-4}$)	**0.717** ($p = 0.00112$)
rate	Yes	0.019 ($p = 0.3185$)	0.950 ($p = 0.2913$)
rate	No	**0.263** ($p = 0.0018$)	**0.711** ($p = 0.00192$)

Table 2. Moran's I and Geary's C on linear regression residuals and their p values.

obtain an estimate of β by (approximately) maximizing this model's log-likelihood function with respect to β; the log-likelihood function can be found on page 188 of [2]. In practice, we rely on a software that implements this procedure.

4 Results

4.1 Areal data

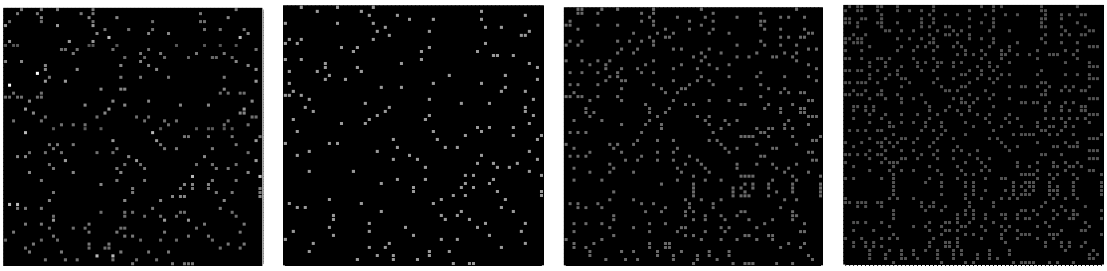
4.1.1 Linear regression

We fitted the linear regression model with median household income (normalized to $[0, 1]$), political lead and percent unemployed as the three explanatory variables and log rate of charging stations as the response variable. All variables are at the county level. The log transformation of the rate of charging stations was motivated by the observation that the residuals look more normally distributed after the transformation (Figure 4).

Moran's I and Geary's C of the residuals of this model and their respective p values associated with permutation tests are shown in the 1st row of Table 2. Since the p values are greater than 0.05, we fail to reject the null that there's no spatial association among the residuals. Interestingly, as shown by the 2nd row of Table 2, when political lead is removed from the model, the residuals exhibit spatial association, indicating that the spatial variability in political lead explains a lot of the spatial variability in the log rate of charging stations. It is also worth noting that these results are robust to whether we use log rate or rate as the response variable (see 3rd and 4th row of Table 2).

Coefficient	*Corresponding covariate*	*Value*	*95% CI*	*p-value*
β_0	1	**-7.200**	(-7.960, -6.439)	$<2\times10^{-16}$
β_1	median household income (normalized)	**-1.924**	(-2.943, -0.905)	0.000364
β_2	political lead	**2.193**	(1.389, 2.997)	9.51×10^{-7}
β_3	percent unemployed	-4.732	(-27.042, 17.578)	0.672973

Table 3. Coefficients of linear regression, their 95% confidence intervals and their p values.



Spatial adjacency 3-nearest-neighbor 7-nearest-neighbor 12-nearest-neighbor

Figure 5. The four row-standardized neighborhood matrices used in spatial regression. White represents the value one and black represents the value zero. As expected, as the number of nearest neighbors increases, the matrix becomes less sparse and the entries become darker, indicating smaller values.

Since the residuals of the linear regression model using log rate and all three covariates do not exhibit spatial association, the modeling assumption of the linear regression model is satisfied and we now interpret its coefficients. Table 3 shows the fitted coefficients, their 95% confidence intervals and their p values. Since only the coefficients associated with median household income and political lead are statistically discernible from zero, we only interpret these two coefficients. The model predicts that, while holding other covariates constant, each percent (i.e., 1% or 0.01) increase in political lead leads to a $2.193\times0.01 = 0.02193$ increase in the mean log rate, which corresponds to scaling the median of the rate by a factor of $e^{0.02193}\approx1.0222$. Intuitively, this means that counties with a bigger lead by Democratic votes over Republican votes tend to have more charging stations per person. Intuitively, this result makes sense because Democrats are more willingness to adopt electric vehicles [18] and a 2015-2016 survey of vehicle owners across US shows that electric vehicles were most popular among Democrats [4]. This result also aligns with a finding in [9]: states with the Republican candidate winning most votes for both the 2016 and 2020 elections tend to have fewer charging stations. On the other hand, median household income is negatively associated with the number of charging stations per person. This is counterintuitive because we'd expect communities that are more financially well-off to have more purchasing power for electric vehicles and better access to charging infrastructure; indeed, multiple prior work showed the opposite effect [8, 9, 10].

4.1.2 Spatial regression

Since the residuals of the linear regression model do not exhibit spatial association, we do not need to perform spatial regression. Nevertheless, we demonstrate its workflow in case the residuals of linear regression exhibit spatial association when applied to a different dataset. Table 4 shows the AIC scores of all combinations of the four spatial regression models discussed in Section 3.1.4 and four neighborhood matrices (Figure 5). These scores fall in a narrow range from 96.742 to 104.4, indicating that all models have similar quality. Among these 16 models, the best model is the spatial Durbin model with the 3-nearest-neighbor neighborhood matrix.

Neighborhood Matrix \ Model	*Spatial Lag*	*Spatial Durbin*	*Spatial Error*	*CAR*
Spatial adjacency (row-standardized)	**96.936**	98.772	99.845	100.93
3-*nearest neighbors (row-standardized)*	98.159	**96.742**	**99.307**	**100.5**
7-*nearest neighbors (row-standardized)*	98.65	102.27	100.3	101.06
12-*nearest neighbors (row-standardized)*	101.04	104.4	100.08	101.05

Table 4. The AIC scores for all combinations of four spatial regression models and four neighborhood matrices. The best AIC score per column is emphasized in bold. The best AIC score overall is underlined.

Coefficient	*Corresponding covariate*	*Value*	*95% CI*	*p-value*
β_0	1	**-5.458**	$(-7.627, -3.290)$	8.089×10^{-7}
β_1	median household income (normalized)	-1.189	$(-2.572, 0.194)$	0.09192
β_2	political lead	**2.180**	$(1.431, 2.929)$	1.183×10^{-8}
β_3	percent unemployed	-6.945	$(-34.078, 20.188)$	0.61592
γ_1	"lagged" median household income (normalized)	-1.653	$(-3.513, 0.208)$	0.08172
γ_2	"lagged" political lead	0.028	$(-1.312, 1.368)$	0.96747
γ_3	"lagged" percent unemployed	31.875	$(-2.335, 66.085)$	0.06783
ρ	"lagged" log rate	0.226	$(-0.059, 0.511)$	0.16359

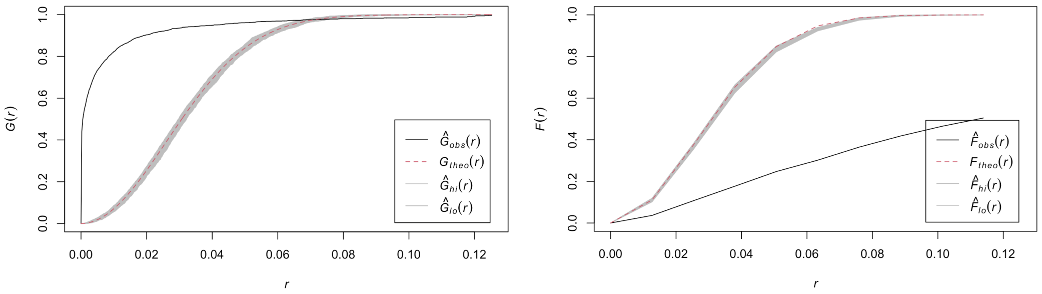
Table 5. Coefficients of the best spatial Durbin model, their 95% confidence intervals and their p values.

Table 3 emits several interesting observations. First, AIC scores of CAR models are all above 100, while the other three spatial regression models all have two or more scores below 100. This indicates that spatial memory is important for modeling the log rates, and only utilizing spatial information of first-order neighbors is sub-optimal. Second, for the spatial lag, spatial Durbin and spatial error models, AIC score tends be higher when the number of neighbors is large. This implies that, for a county, spatial variables from further-away counties tend to be less relevant for predicting its log rate, and averaging over them when creating "lagged" covariates tend to do more harm than good.

Finally, we provide interpretations for the best spatial Durbin model. For a sanity check, we confirm that the residuals of this best model do not exhibit spatial association (Moran's $I = 0.05$ with $p = 0.21$; Geary's $C = 0.91$ with $p = 0.16$). Table 5 shows its fitted coefficients and their p values.

From Table 5, we see that only the intercept and the coefficient for political lead are statistically discernible from zero. The value of coefficient for political lead (2.180) in this spatial Durbin model is similar to that in the linear regression model (2.193), indicating some level of agreement between the two models. While holding other covariates constant, this spatial Durbin model predicts that each percent (i.e., 1% or 0.01) increase in political lead leads to a $2.18 \times 0.01 = 0.0218$ increase in the mean log rate, which corresponds to scaling median of the rate by a factor of $e^{0.0218} \approx 1.0220$. The implications of this are similar to what we described earlier for linear regression.

All other coefficients are not statistically discernible from zero. This includes ρ, the coefficient of the weighted average of log rate over three nearest counties, and γ_1, the coefficient of the weighted average of the median household income over three nearest counties. This suggests that spatial information from neighboring counties are not important predictors of log rate, which makes sense since linear regression residuals did not exhibit spatial association. Nevertheless, including these "lagged" covariates yields a slightly better AIC (96.742) than linear regression (99.078).



(a) G function of charging stations (black line) against G functions of datasets generated under CSR (gray band)

(b) F function of charging stations (black line) against F functions of datasets generated under CSR (gray band)

Figure 6. Exploratory plots for diagnosing CSR.

4.2 Point-level data

4.2.1 Testing for complete spatial randomness

Before modeling, we test whether locations of charging stations exhibit CSR using G and F functions discussed in Section 3.2.3. Figure 6a shows the (empirical) G function of charging stations against G functions of datasets generated under CSR. We see that, for a large range of r values, the observed G function lies above the simulation envelope. This means that, for each radii in this range, the proportion of charging stations with their nearest-neighbor charging stations in this radii is higher than expected under CSR. In other words, charging stations and their nearest-neighbor charging stations are closer than expected under CSR, indicating clustering of charging stations.

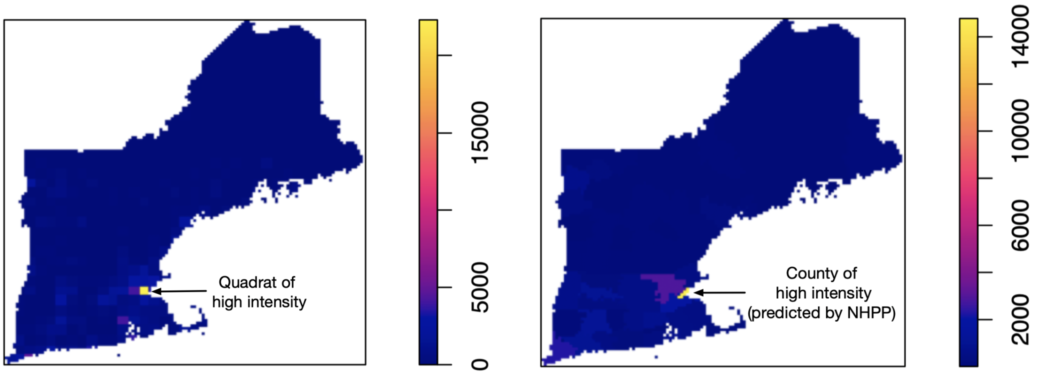
We also compared the F function of charging stations against F functions of datasets generated under CSR in Figure 4b. We see that, for almost all r values, the observed F function lies below the simulation envelope. This means that, for each radii, the proportion of arbitrary locations with their nearest-neighbor charging stations within this radii is lower than expected under CSR. In other words, arbitrary locations and their nearest-neighbor charging stations are farther away than expected under CSR, indicating more empty spaces and hence clustering of charging stations.

4.2.2 Non-homogeneous Poisson process

The fact that locations of charging stations do not exhibit CSR motivates us to move beyond the HPP and perform further modeling with the NHPP model. We fitted the NHPP model with total population (normalized to $[0, 1]$), median household income (normalized to $[0, 1]$ as in linear and spatial regression), political lead and percent unemployed as the four explanatory covariates. Recall that all these covariates are only available at the county level. This means that $X(s)$ in Equation 4 is a 5-dimensional vector (with one dummy dimension being one and the other four dimensions being the values of the four explanatory covariates) and is constant within each county.

Coefficient	*Corresponding covariate*	*Value*	*95% CI*	*p-value*
β_0	1	0.238	$(-0.029, 0.506)$	$\geq$0.05
β_1	median household income (normalized)	**1.818**	$(1.513, 2.124)$	<0.001
β_2	political lead	**4.774**	$(4.582, 4.966)$	<0.001
β_3	percent unemployed	**102.883**	$(97.133, 108.633)$	<0.001
β_4	total population (normalized)	**1.716**	$(1.560, 1.871)$	<0.001

Table 6. Coefficients of NHPP, their 95% confidence intervals and their p values.



(a) Intensity surface from quadrat counts (35-by-35 grid)

(b) Intensity surface from fitted NHPP

Figure 7. Comparison between the intensity surface obtained from quadrat counts (by dividing the number of charging stations in each quadrat by its area) and the intensity surface obtained from the fitted NHPP.

We provide the fitted values, confidence intervals and p values of the coefficients in Table 6. Interestingly, the coefficients of all covariates are statistically discernible from zero (in the positive direction) while the coefficient of the intercept is not. Median household income, political lead, percent unemployed and total population at the county level all have positive effects on the spatial intensity of charging stations in New England. Intuitively, this means that counties with higher median household income, a bigger lead by Democratic votes over Republican votes, a higher percent unemployed and more people tend to have more charging stations. The positive effect from political lead aligns with the result of linear regression and [9]. The positive effect from median household income makes sense intuitively because we'd expect wealthier communities to have more electric vehicles and better access to charging infrastructure, and also aligns with prior work [8, 9, 10]. The positive effect from total population seems logical since we'd expect counties with more residents to have more electric vehicles which rely on a larger network of charging stations. Finally, the positive effect from percent unemployed is the least intuitive among all explanatory covariates.

To qualitatively evaluate the reliability of these coefficient estimates, we compare the intensity surface obtained from the fitted NHPP (Figure 7b) against the intensity surface obtained from quadrat counts (Figure 7a). We see that both intensity surfaces look similar and the fitted NHPP successfully captures the region of high intensity. However, this high-intensity region is an outlier compared to the rest of New England and therefore may heavily influence the coefficient estimates.

5 Discussion

This project was motivated by the research question: what is the relationship between political inclination and the frequency of electric-vehicle charging stations in the New England region of the United States? To approach this question, we performed areal analysis, which models the county-level log rate of charging stations as a linear function of county-level covariates (which includes the covariate *political lead* as a measure of political inclination), and point-level analysis, which models the log intensity of charging stations as a linear function of county-level covariates. To answer the research question, we interpreted the coefficients associated with the covariates in the fitted models.

The final model from areal analysis, the linear regression model, and the final model from point-level analysis, the NHPP, gave similar results for the relationship between political inclination and the frequency of charging stations. Both of these models showed that counties with a bigger lead by Democratic votes over Republican votes tend to have more charging stations after accounting for other covariates. This means the relationship between political inclination and the frequency of charging stations on the state level discovered by [9] also holds on the county level and point level, at least in New England. However, the two models yielded conflicting results for the discernability and direction of effect of other covariates: percent unemployed is discernible in the NHPP but is not in linear regression; median household income is discernible in the positive direction in linear regression but is discernible in the negative direction in the NHPP.

The linear regression model and the spatial regression models agreed on the discernability and direction of effect of the three covariates. This was expected because the linear regression residuals didn't display spatial association, meaning that spatial regression models are technically not required for our areal dataset despite giving slightly better AIC than the linear regression model. Nevertheless, we decided to describe, fit and interpret spatial regression models and we note that the areal-level analysis described in this paper can generalize to other areal datasets for which linear regression residuals are spatially correlated.

The linear regression model and the NHPP each has its own limitations. As discussed in Section 3.2.1, while the linear regression model and other areal models are simple to fit and interpret, they ignore the distribution of charging stations within counties and could potentially suffer from the modifiable areal unit problem, where the unit of aggregation influences the conclusions of inference. Models for point-level data, on the other hand, bypass these problems by directly modeling locations of charging stations on a point level. However, as briefly mentioned in Section 3.2.4, the NHPP we considered is a very introductory and limited point-level model because it assumes the intensity function to be fully determined by the explanatory covariates. As a result, its intensity function is constant on the county level because our covariates are only available on the county level. Additionally, as discussed in Section 4.2.2, while it successfully captured the county with unusually high intensity and produced a reasonable intensity surface overall, its coefficient estimates could be heavily influenced by that unusual county; for this reason, the coefficient estimates of the linear regression model appear to be more reliable than those of the NHPP in the context of this project.

One possibility for future work is to use the log-Gaussian Cox process [11], which adds a smooth-varying Gaussian process to the intensity function of the NHPP, for overcoming the aforementioned limitations of the NHPP. The added Gaussian process allows the intensity to vary within areal units even when covariates are constant within areal units and also takes into account intensity patterns not captured by the covariates themselves. Another direction of future work could be performing the same analysis conducted in this project for other regions in the United States; comparing coefficient estimates across different regions could lead to valuable insights on how political inclination affects the frequency of charging stations differently in different regions.

References

[1] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*, pages 199–213, 1998.

[2] Roger S Bivand, Edzer J Pebesma, Virgilio Gomez-Rubio, and Edzer Jan Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.

[3] MIT Election Data and Science Lab. County Presidential Election Returns 2000-2020. https://doi.org/10.7910/DVN/VOQCHQ.

[4] Andrew Farkas, Hyeon-Shic Shin, Seyedehsan Dadvar, Jessica Molina et al. Electric vehicle ownership factors, preferred safety technologies and commuting behavior in the united states. 2017.

[5] Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

[6] Diego-Alejandro Giménez, Anabela Ribeiro, Javier Gutiérrez-Puebla, and Antonio Pais-Antunes. Charging-stations for electrical vehicles: analysis and model to identify the most convenient locations. *Computer-based Modelling and Optimization in Transportation*, pages 101–111, 2014.

[7] Albert O Hirschman. *National power and the structure of foreign trade*, volume 105. Univ of California Press, 1980.

[8] Chih-Wei Hsu and Kevin Fingerman. Public electric vehicle charger access disparities across race and income in california. *Transport Policy*, 100:59–67, 2021.

[9] Levente Juhász and Hartwig H Hochmair. Spatial and temporal analysis of location and usage of public electric vehicle charging infrastructure in the united states.

[10] Hafiz Anwar Ullah Khan, Sara Price, Charalampos Avraam, and Yury Dvorkin. Inequitable access to ev charging infrastructure. *The Electricity Journal*, 35(3):107096, 2022.

[11] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

[12] Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[13] Rachael Nealer. Cleaner cars from cradle to grave: how electric cars beat gasoline cars on lifetime global warming emissions. *JSTOR Sustainability Collection*, 2015.

[14] U.S. Department of Energy. About the alternative fueling station data. https://afdc.energy.gov/stations#/find/nearest?show_about=true (Accessed on June 4, 2023).

[15] U.S. Department of Energy. Alternative fuels data center - data downloads. https://afdc.energy.gov/data_download (Accessed on April 10, 2023).

[16] U.S. Department of Health and Human Services. American community survey (acs). https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/american-community-survey-acs (Accessed on June 4, 2023).

[17] U.S. Bureau of Labor Statistics. American community survey (acs) questions and answers. https://www.bls.gov/lau/acsqa.htm (Accessed on June 4, 2023).

[18] Nicole D Sintov, Victoria Abou-Ghalioum, and Lee V White. The partisan politics of low-carbon transport: why democrats are more likely to adopt electric vehicles than republicans in the united states. *Energy Research & Social Science*, 68:101576, 2020.

[19] Kyle Walker and Matt Herman. *Tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*. 2023. R package version 1.4.1.