



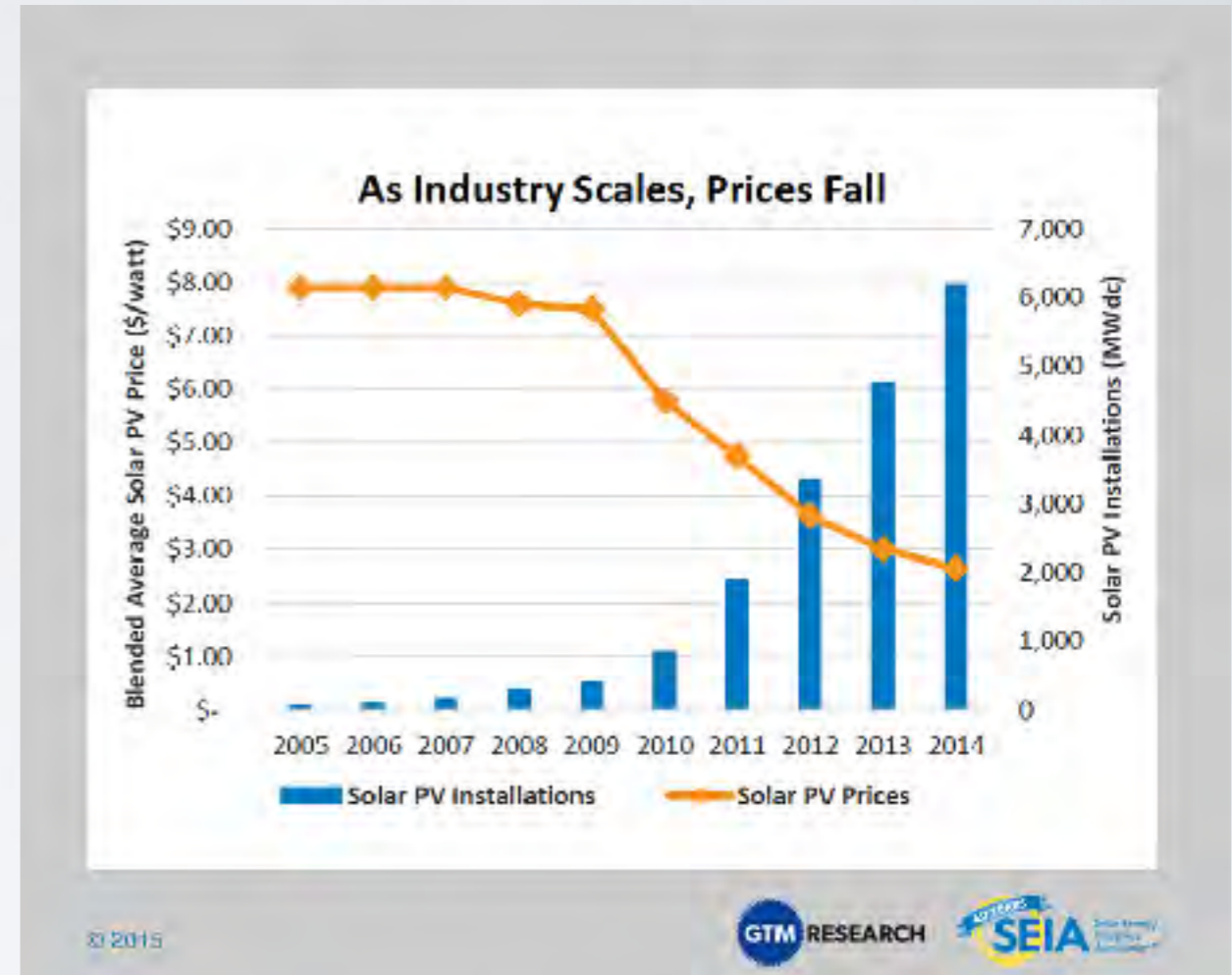
CAN I SOLAR?

HELPING YOU DECIDE IF SOLAR POWER IS RIGHT FOR YOU

Gabriel J. Michael

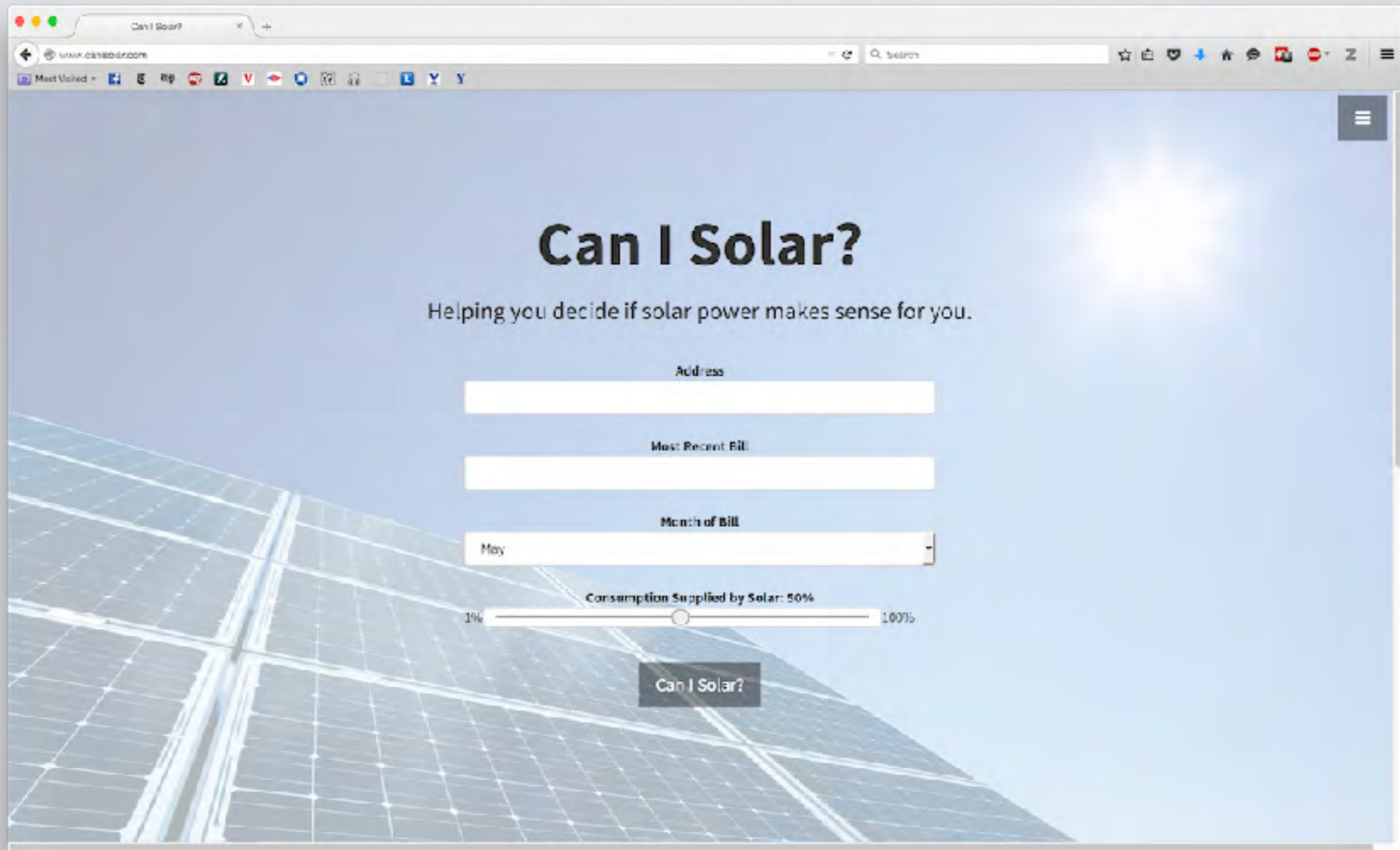
MOTIVATION

- Residential solar sector grew 51% from 2013 to 2014
- Projected market value of \$3.7 billion in 2015
- Complex decision with many variables
- Homeowners want to know:
 - How much money can I save?
 - When will I break even?



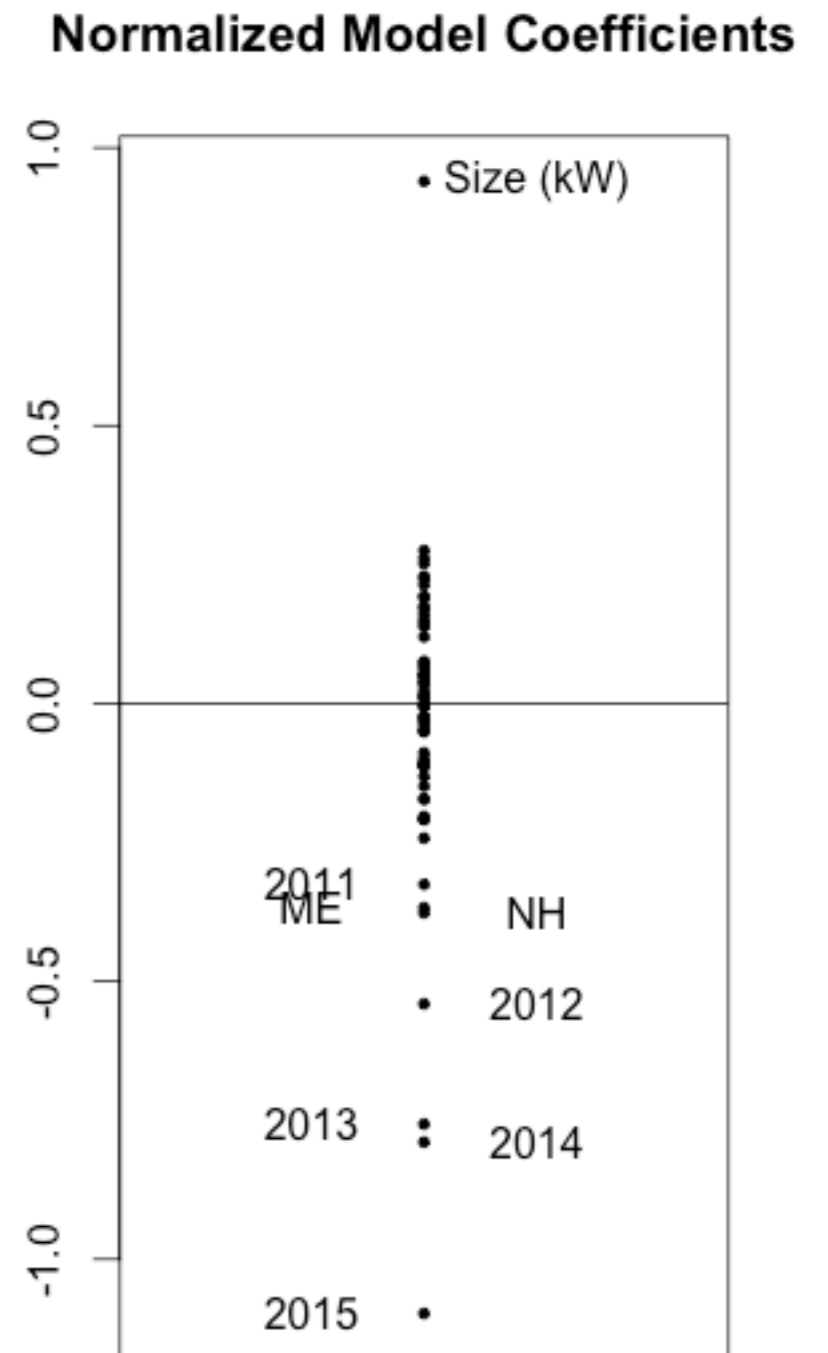
CAN I SOLAR?

A DATA-DRIVEN WEB APPLICATION



MODELING INSTALLATION COSTS

- Data on 400,000 installs obtained from National Renewable Energy Laboratory
- Cost of solar installations varies by:
 - size of the array
 - year of installation
 - location of installation
- Multiple linear regression provides good fit and is easily interpretable
- Also tried multilevel modeling and random forest regression

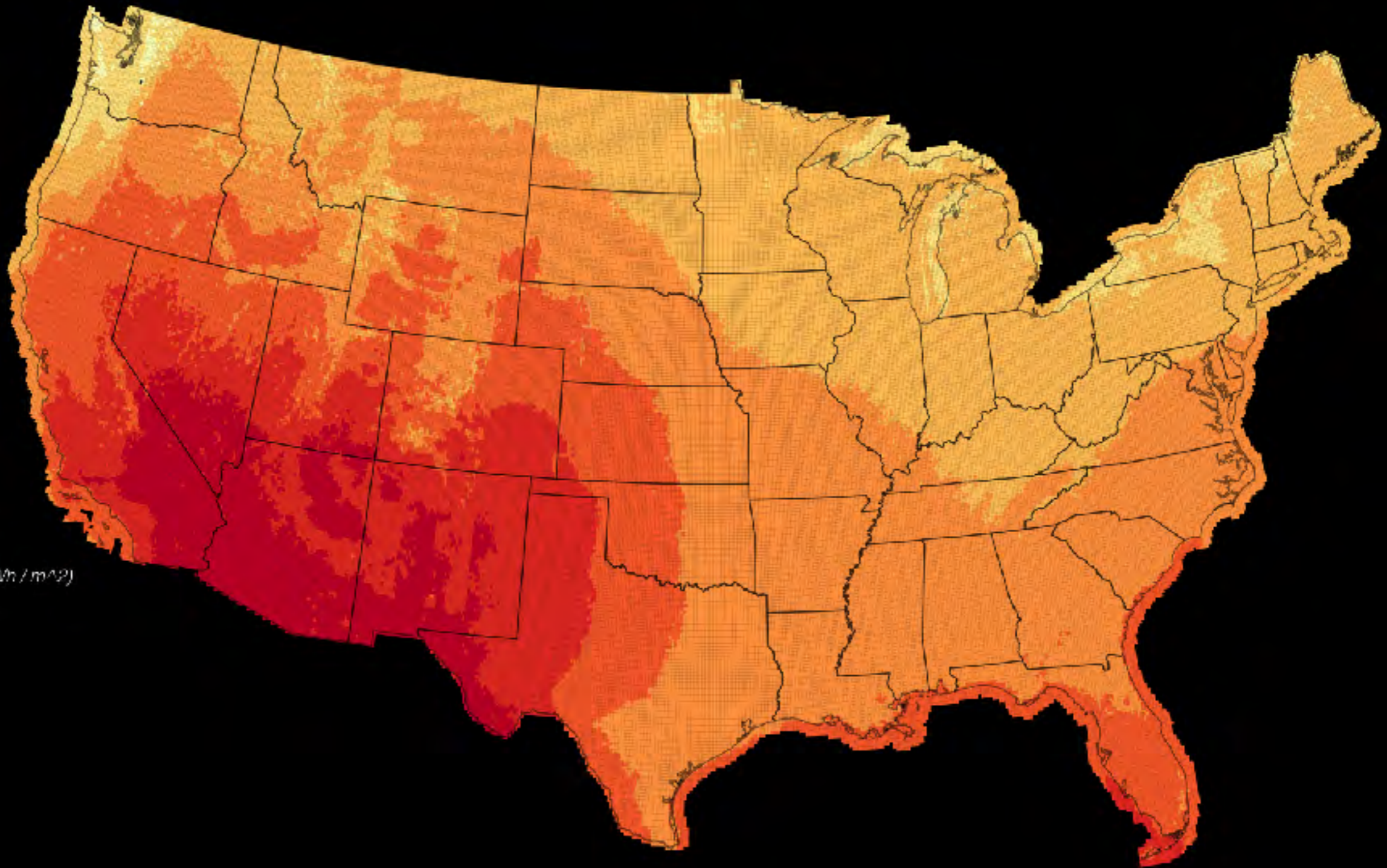


MODELING FUTURE ELECTRICITY PRICES

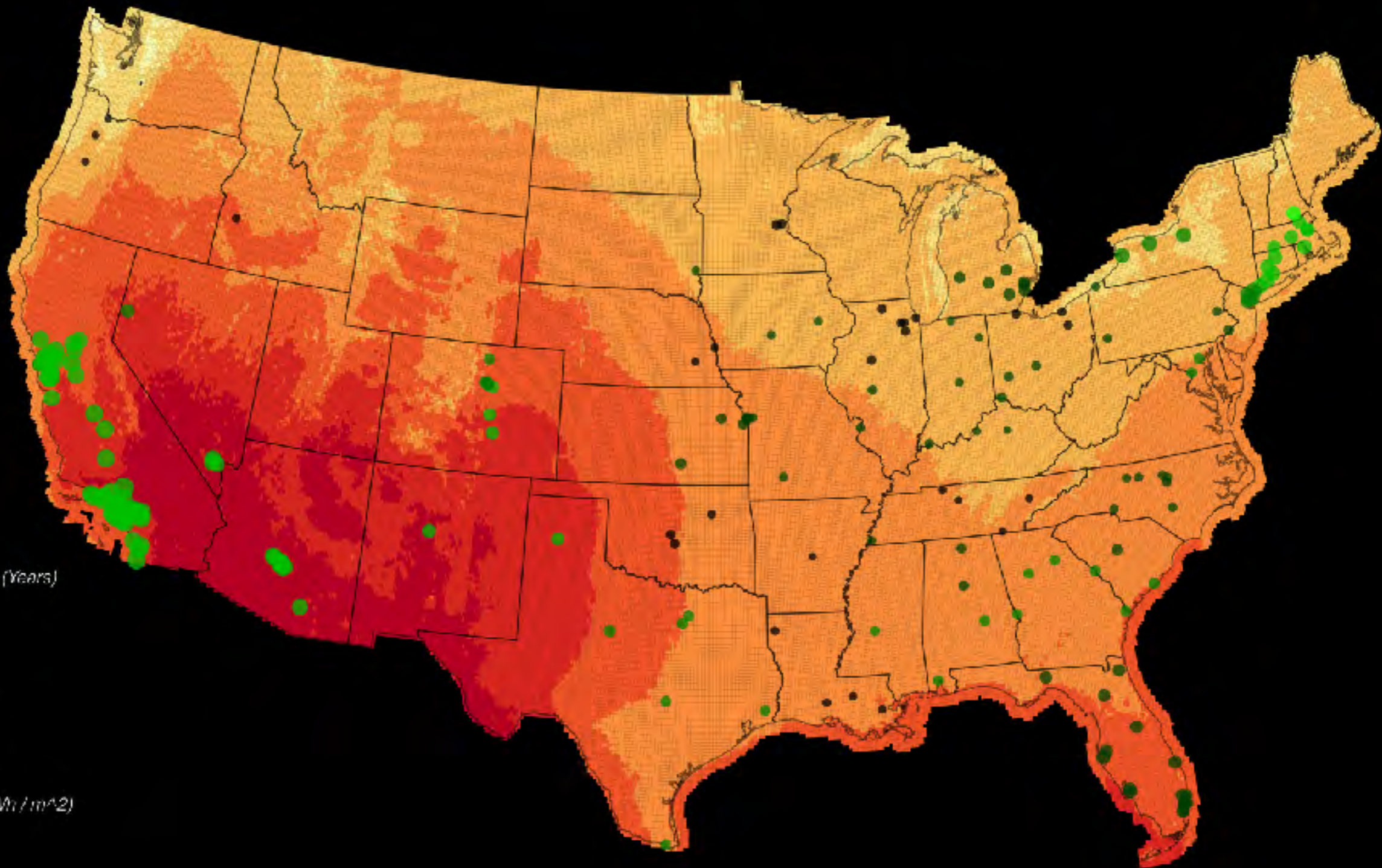
- 15 years of monthly historical electricity prices by state obtained from Energy Information Administration
- Prices and trends vary significantly by state, so no one model works best for all states
- Developed a pipeline to automatically test, validate, and select an appropriate time-series model for each state, e.g.:
 - linear
 - ARIMA
 - exponential smoothing



WHERE CAN I SOLAR?



WHERE CAN I SOLAR?



Legend

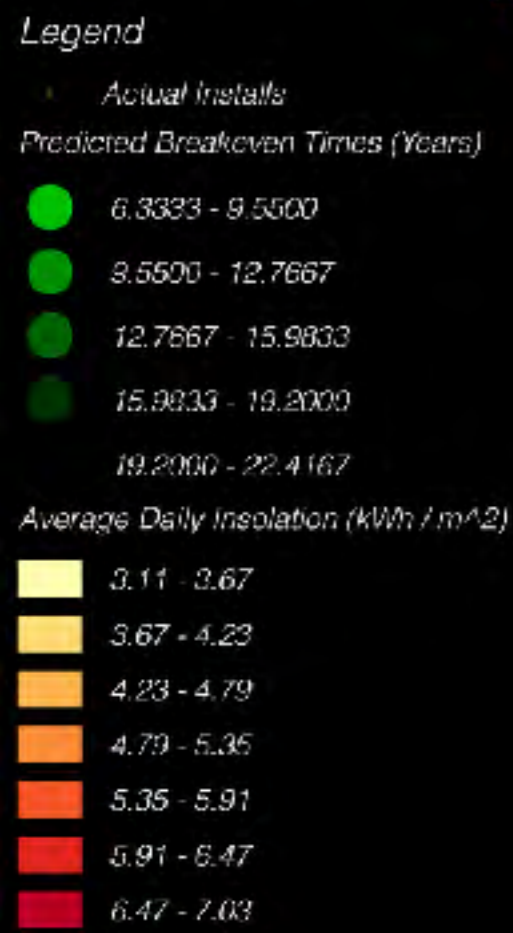
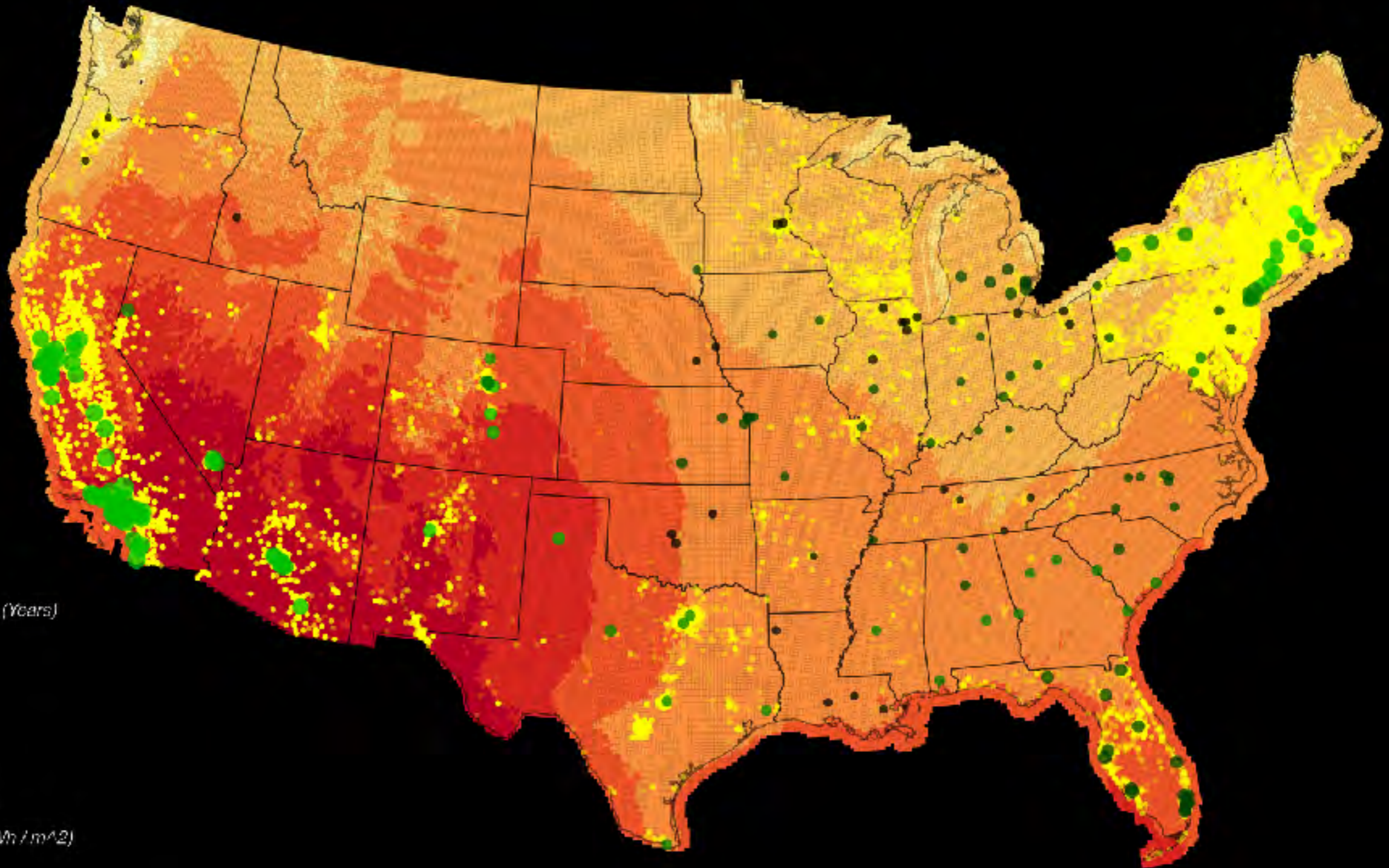
Predicted Breakeven Times (Years)

- 6.3333 - 9.5500
- 9.5500 - 12.7667
- 12.7667 - 15.9833
- 15.9833 - 19.2000
- 19.2000 - 22.4167

Average Daily Insolation (kWh/m²)

- 3.11 - 3.67
- 3.67 - 4.23
- 4.23 - 4.79
- 4.79 - 5.35
- 5.35 - 5.91
- 5.91 - 6.47
- 6.47 - 7.03

WHERE CAN I SOLAR?

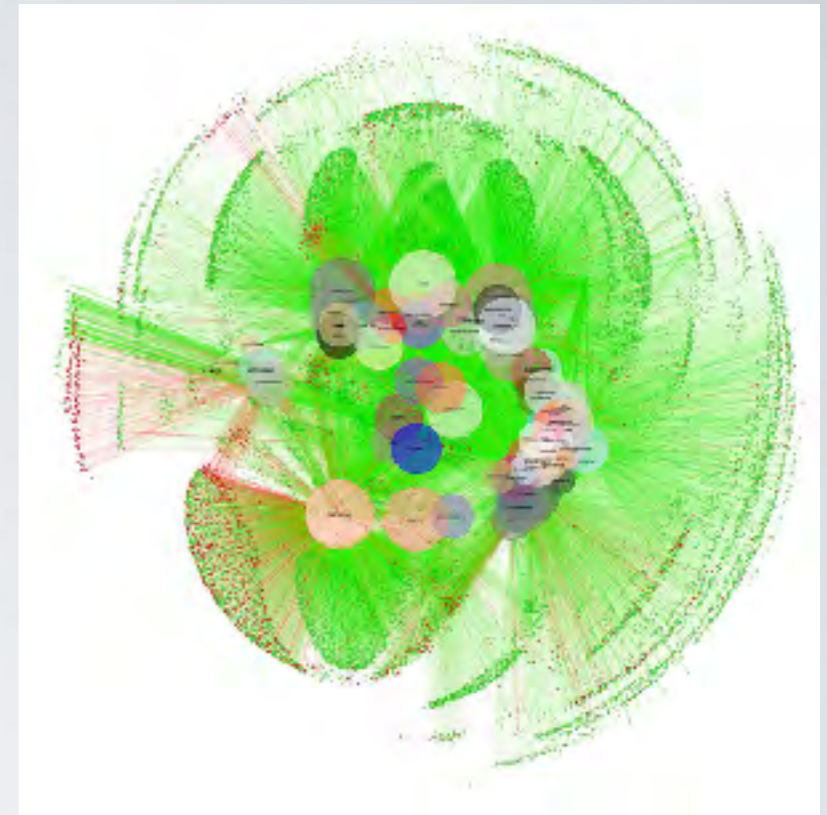


GABRIEL J. MICHAEL

- Ph.D., Political Science, George Washington University
 - Used survival regression to model countries' adoption of intellectual property laws
- Postdoc, Yale Law School
 - Used NLP with SVMs to classify tweets and regulatory comments on political topics



Urban explorer, electronics hobbyist



Visualization of Twitter users' connections and sentiment about net neutrality



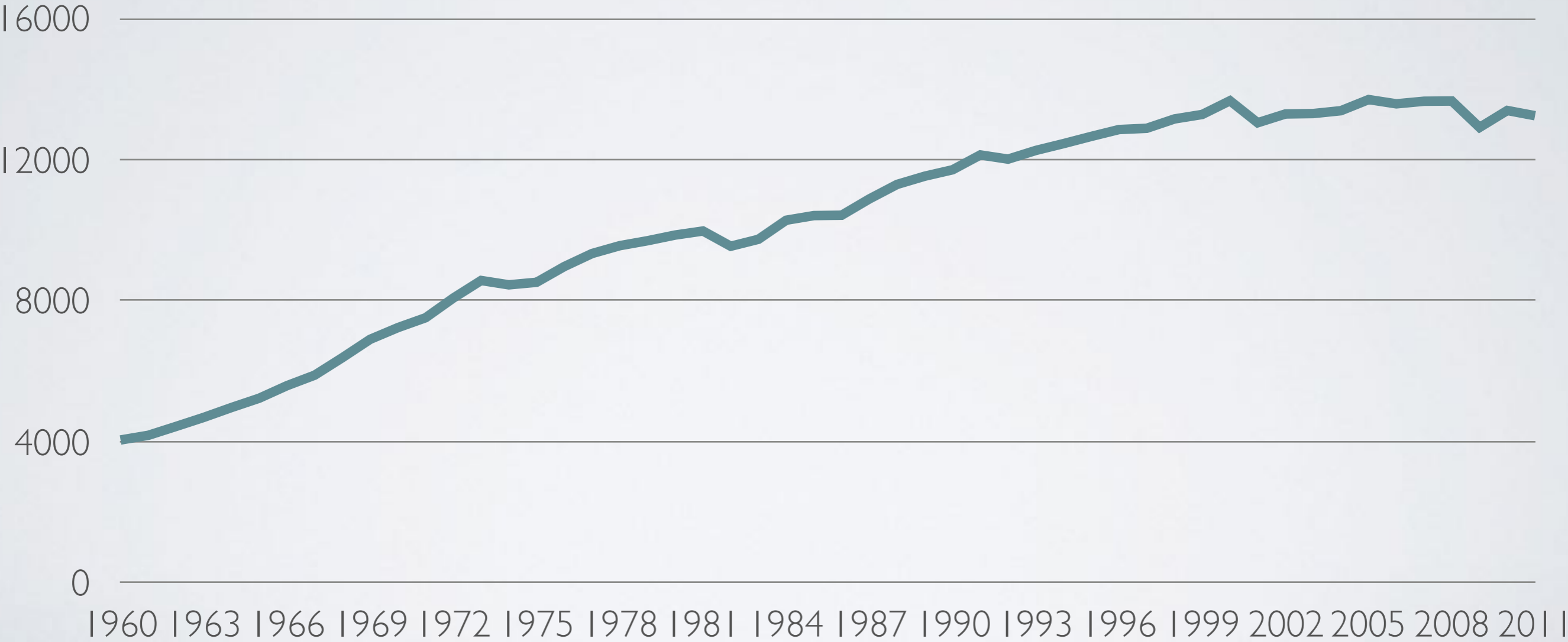
Exploring the since-demolished PEPCO Benning Generating Station, Washington, DC

MODELS OF INSTALLATION COSTS

	Simple Linear Regression	Multiple Linear Regression	Multilevel Model	Random Forest Regression
Model Form	$\log(\text{cost}) \sim \log(\text{size_kw})$	$\log(\text{cost}) \sim \log(\text{size_kw}) + \text{state} + \text{year}$	$\log(\text{cost}) \sim \log(\text{size_kw}) + (\log(\text{size_kw}) \mid \text{state}/\text{year_installed})$	$\log(\text{cost}) \sim \log(\text{size_kw})$
Notes		easy to interpret and explain	confidence and prediction intervals for multilevel models are difficult to interpret	scikit-learn's random forest regressor doesn't support factors, and the R packages are too slow
R ² or Pseudo R ²	0.81	0.89	0.89	0.93
10-fold CV MSE	0.089	0.053	0.050	0.050

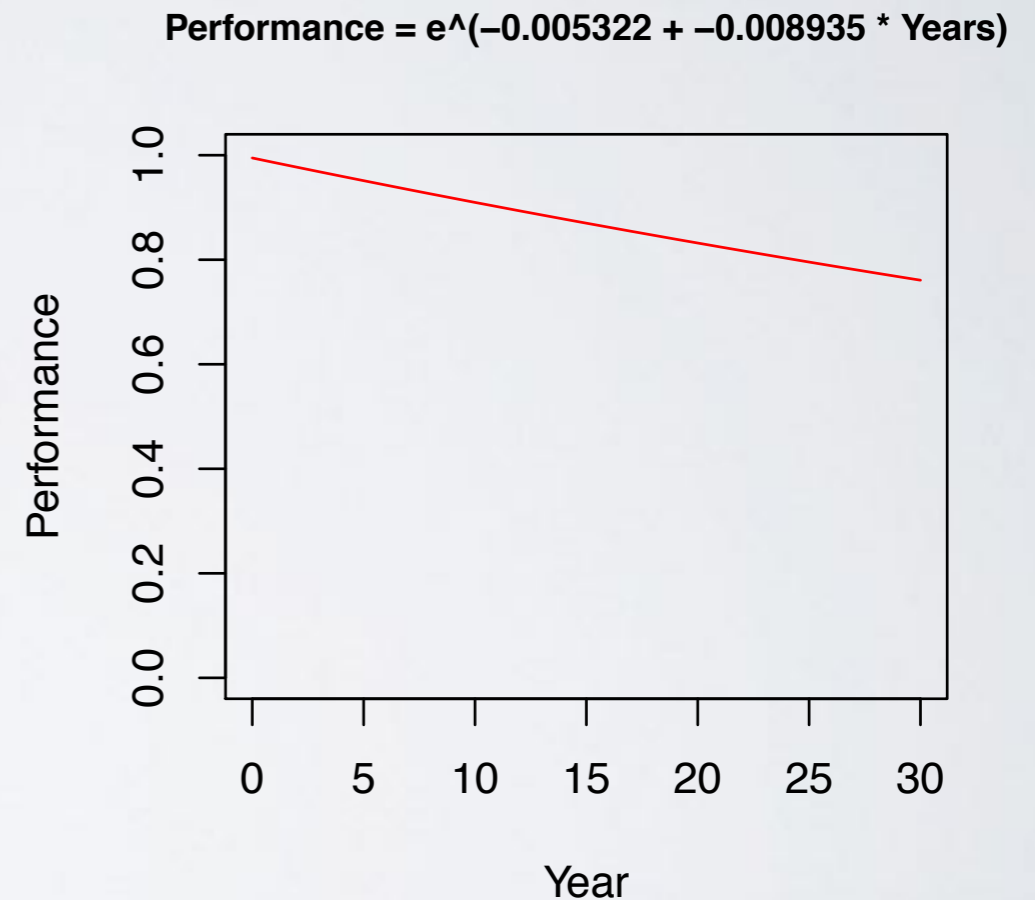
Per-capita electricity consumption has flattened and even declined in recent years

United States: kWh per capita



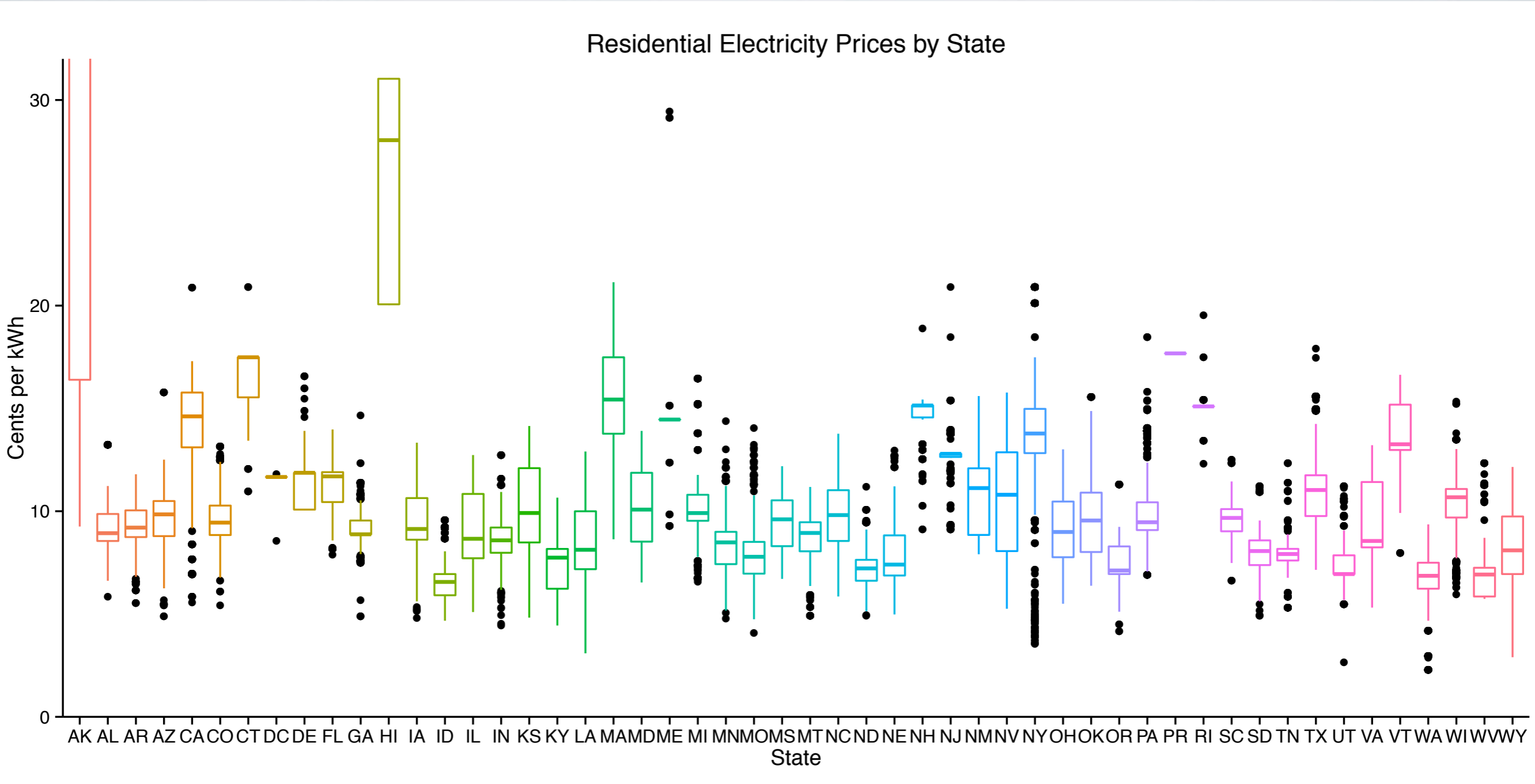
PHOTOVOLTAIC PERFORMANCE DECLINE OVER TIME

- Industry standard warranties offer guaranteed 90% output at 10 years, 80% output at 25 years
- I use a simple exponential decay curve to calculate performance in month 0 to month 360 (30 years)



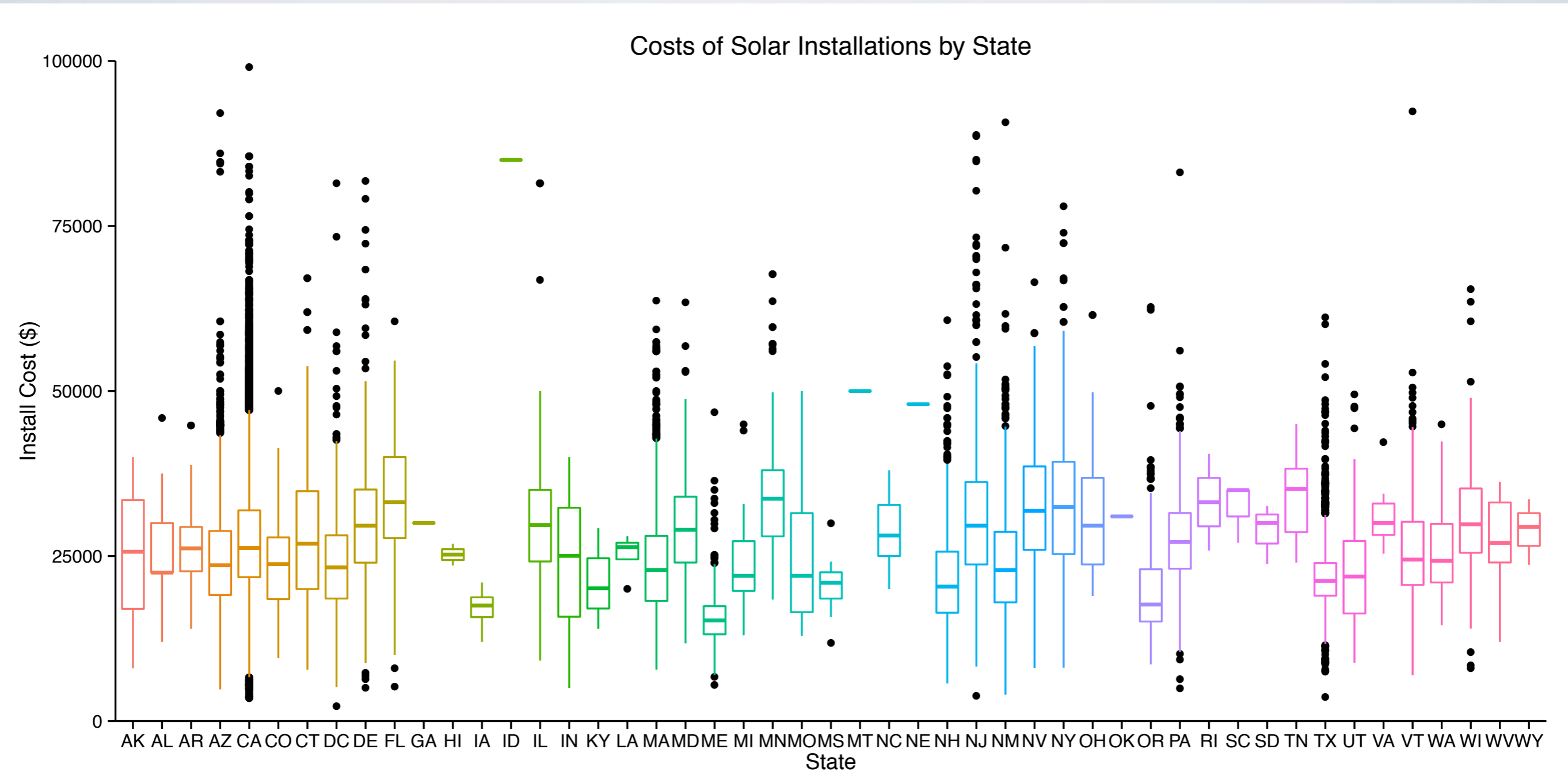
WITHIN VS BETWEEN GROUP VARIANCE IN ELECTRICITY PRICES

There is more variance between states than within states



WITHIN VS BETWEEN GROUP VARIANCE IN INSTALLATION COSTS (3 - 5 KW)

Significant variance between states, but also within states



BACKEND



- Python 3 + pandas for core classes and program logic



- R for modeling + rpy2 Python interface to R



- MySQL for storage of electricity consumption and price data, and solar installation cost/size data



mongoDB

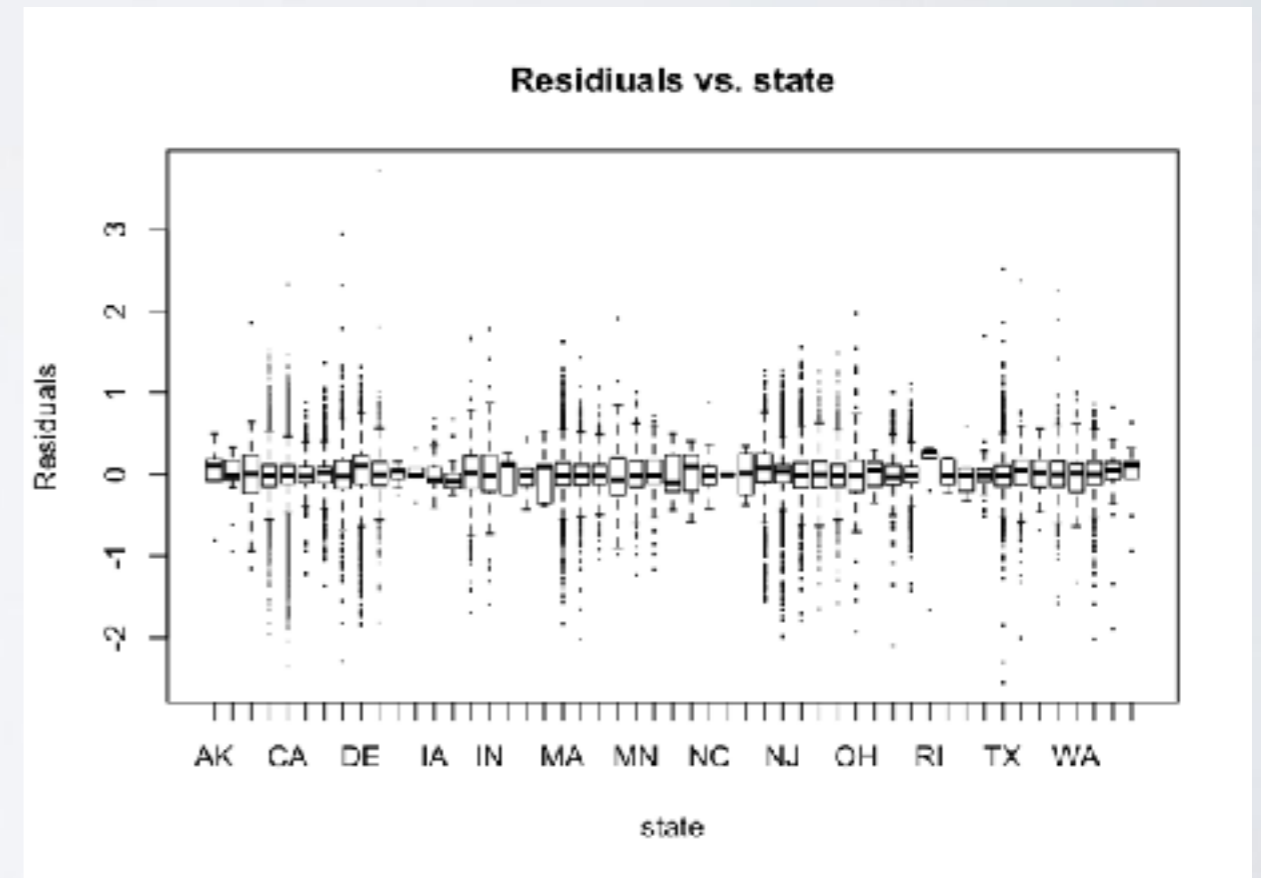
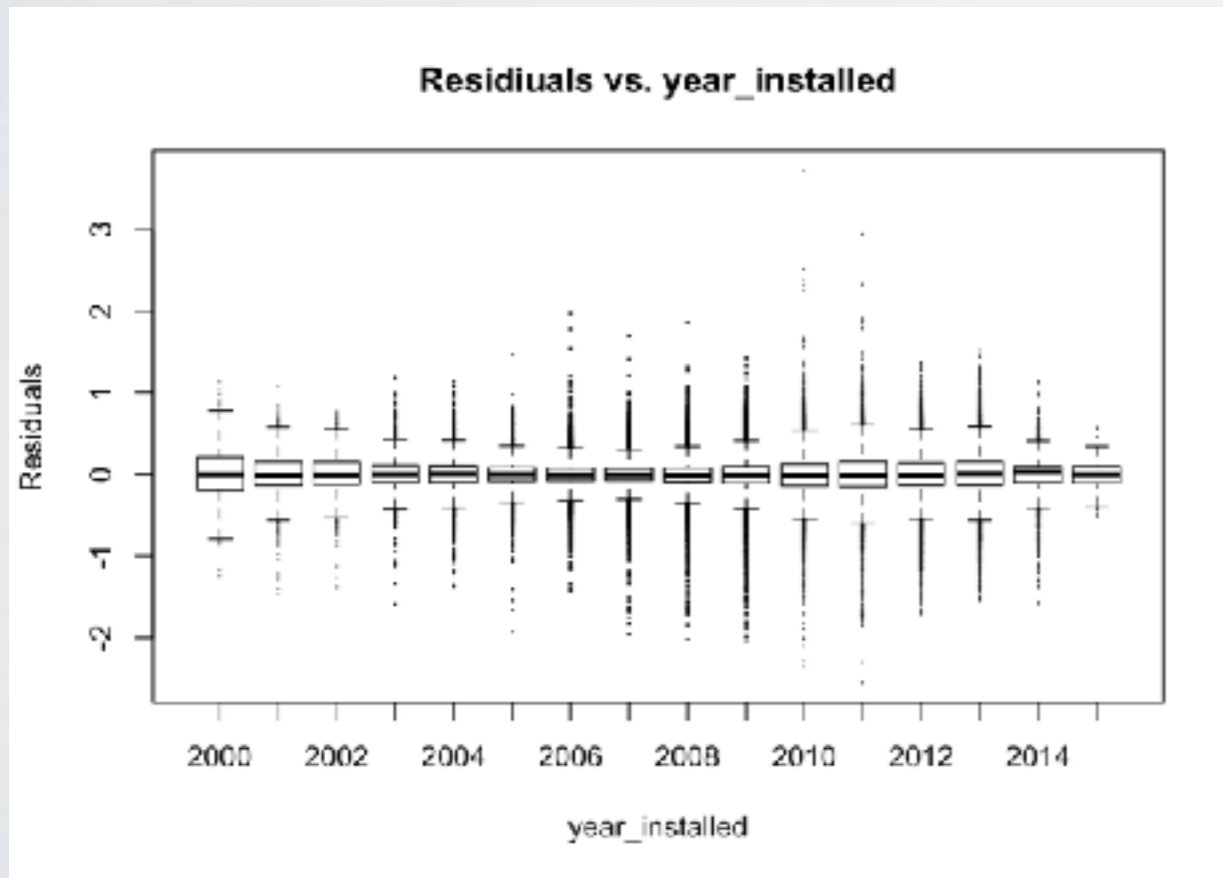
- MongoDB for storage and retrieval of geolocated insolation data



- Code on GitHub: <https://github.com/langelgjm/canisolar>

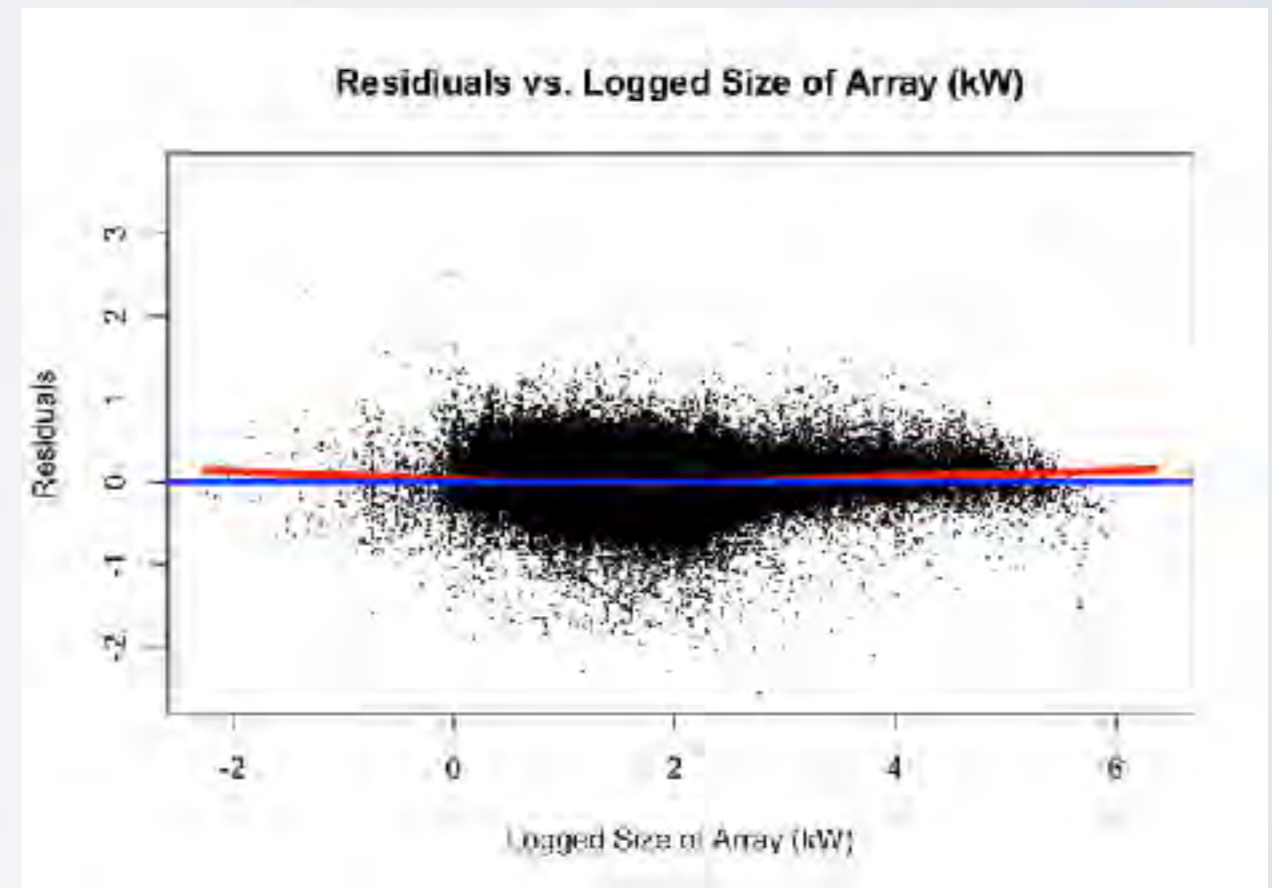
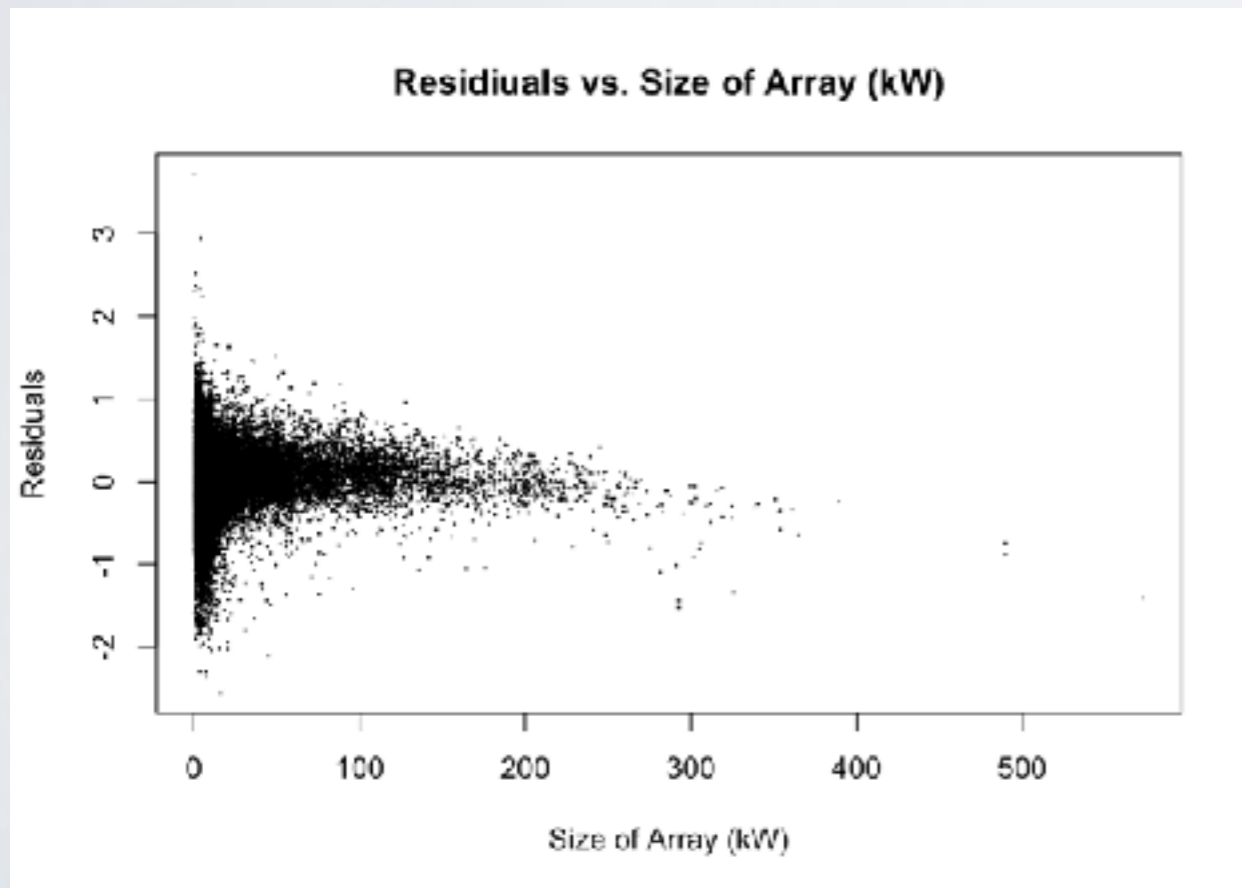
ASSUMPTIONS OF LINEAR REGRESSION

- Independence of errors



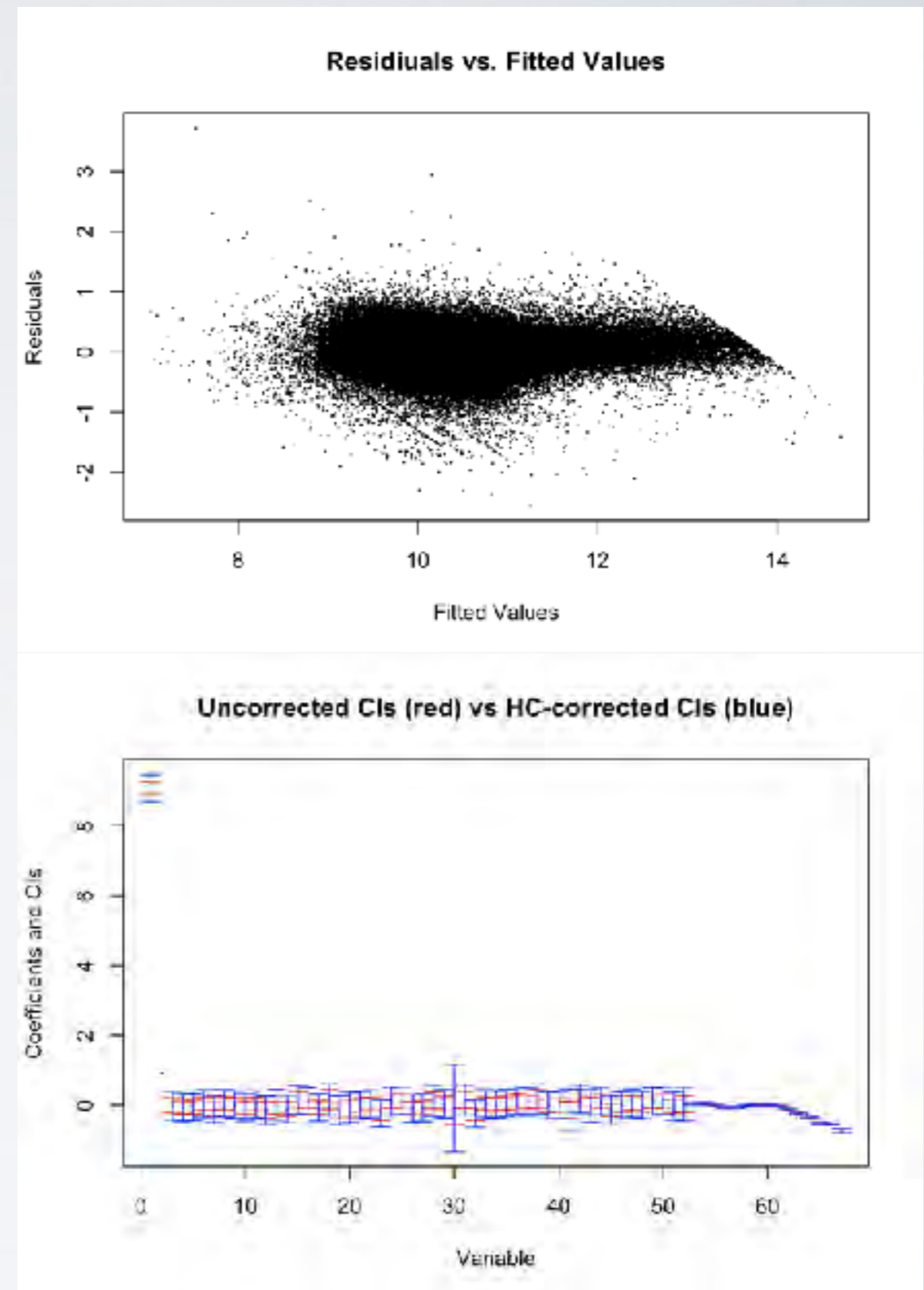
ASSUMPTIONS OF LINEAR REGRESSION

- Independence of errors



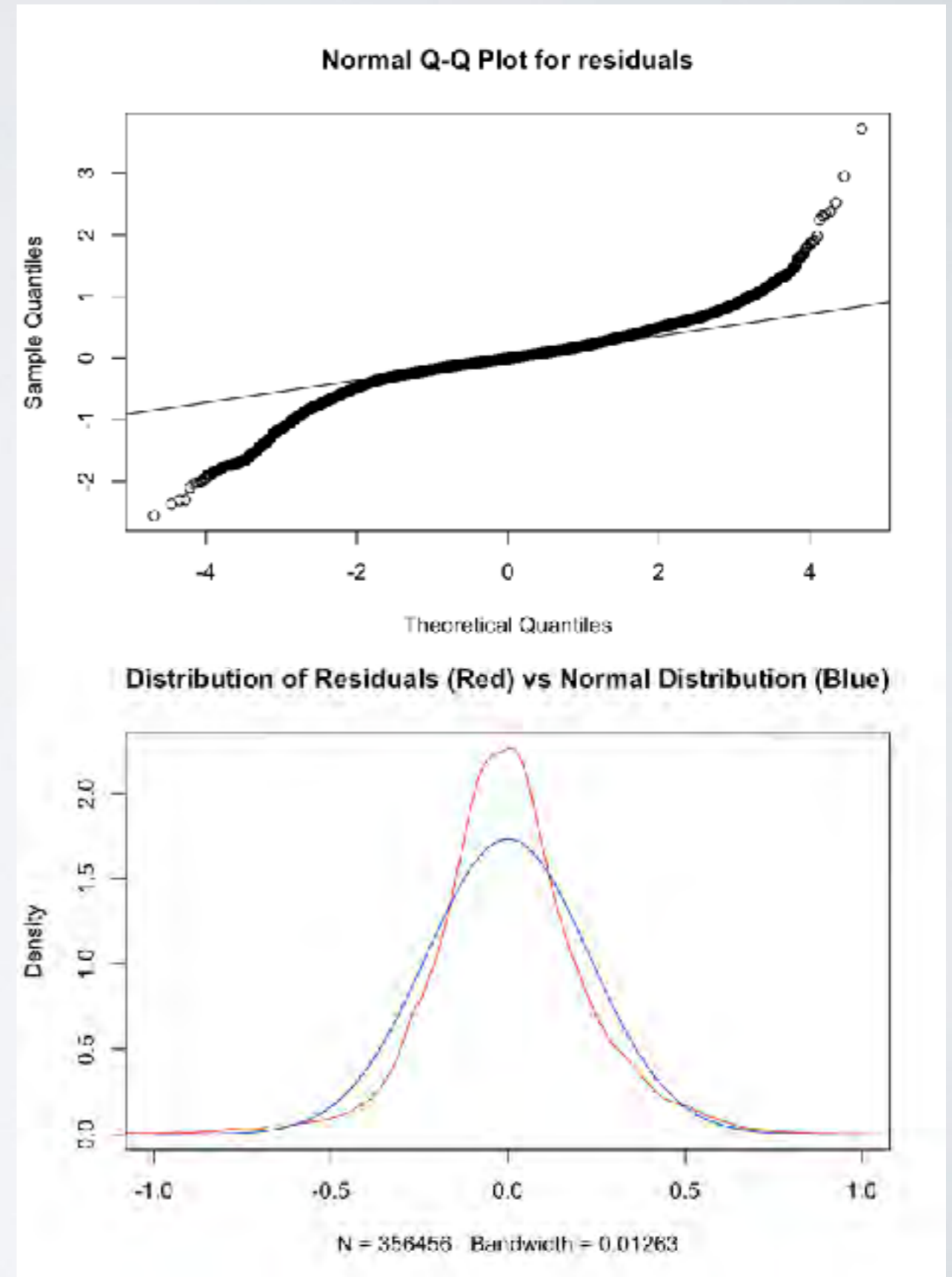
ASSUMPTIONS OF LINEAR REGRESSION

- Homoskedasticity
(constant variance of errors)
 - Some evidence of heteroskedasticity
- Could use robust standard errors for intervals, although the confidence intervals are not much wider



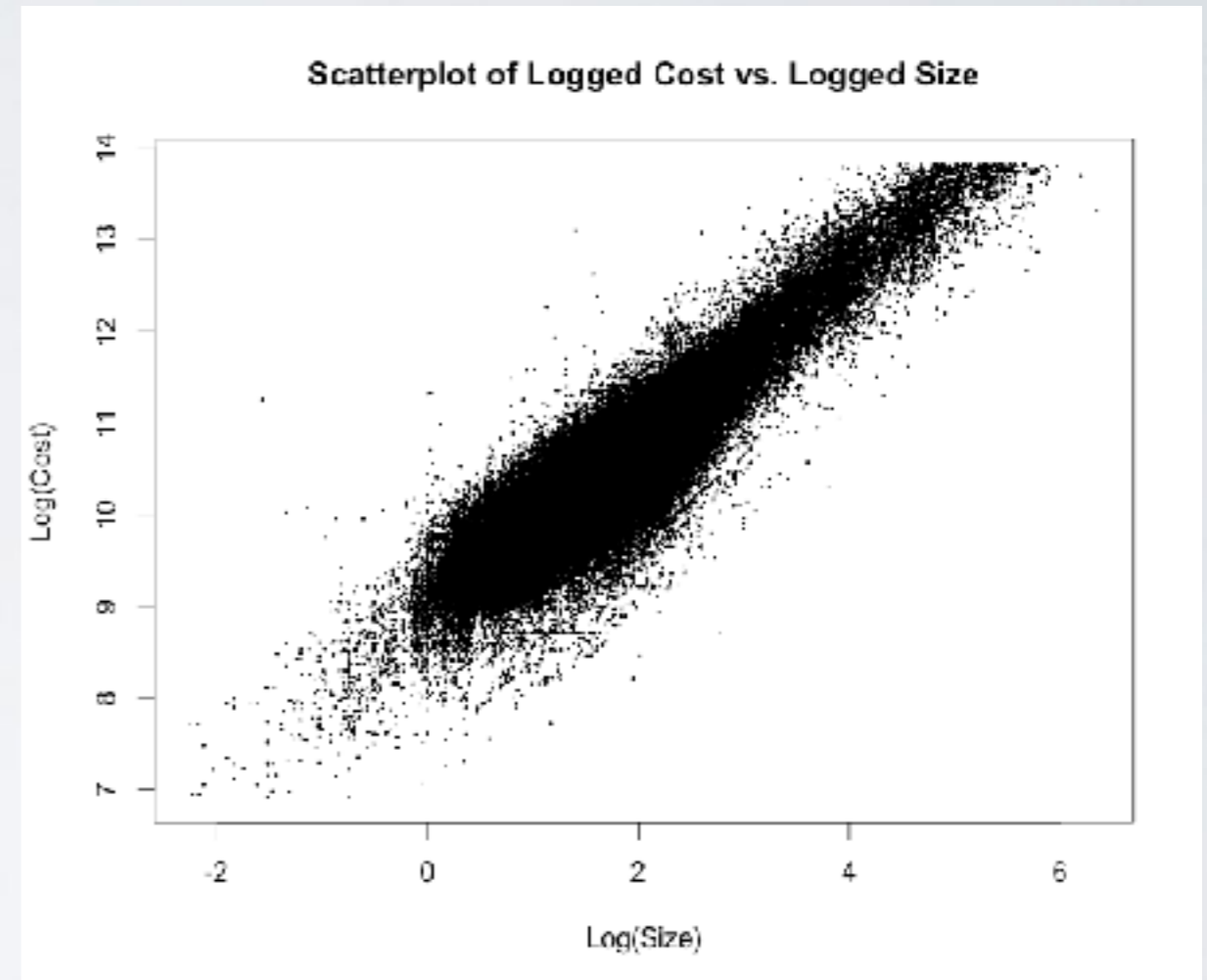
ASSUMPTIONS OF LINEAR REGRESSION

- Normality of residuals
 - Evidence of non-normal (heavy tailed) error distribution
 - This assumption only necessary for confidence intervals/p-values, not best linear unbiased estimates
 - Could use robust regression with t-distribution



ASSUMPTIONS OF LINEAR REGRESSION

- True linear relationship
 - True with simple regression of cost \sim size
- No significant multicollinearity
 - Variance inflation factors relatively low

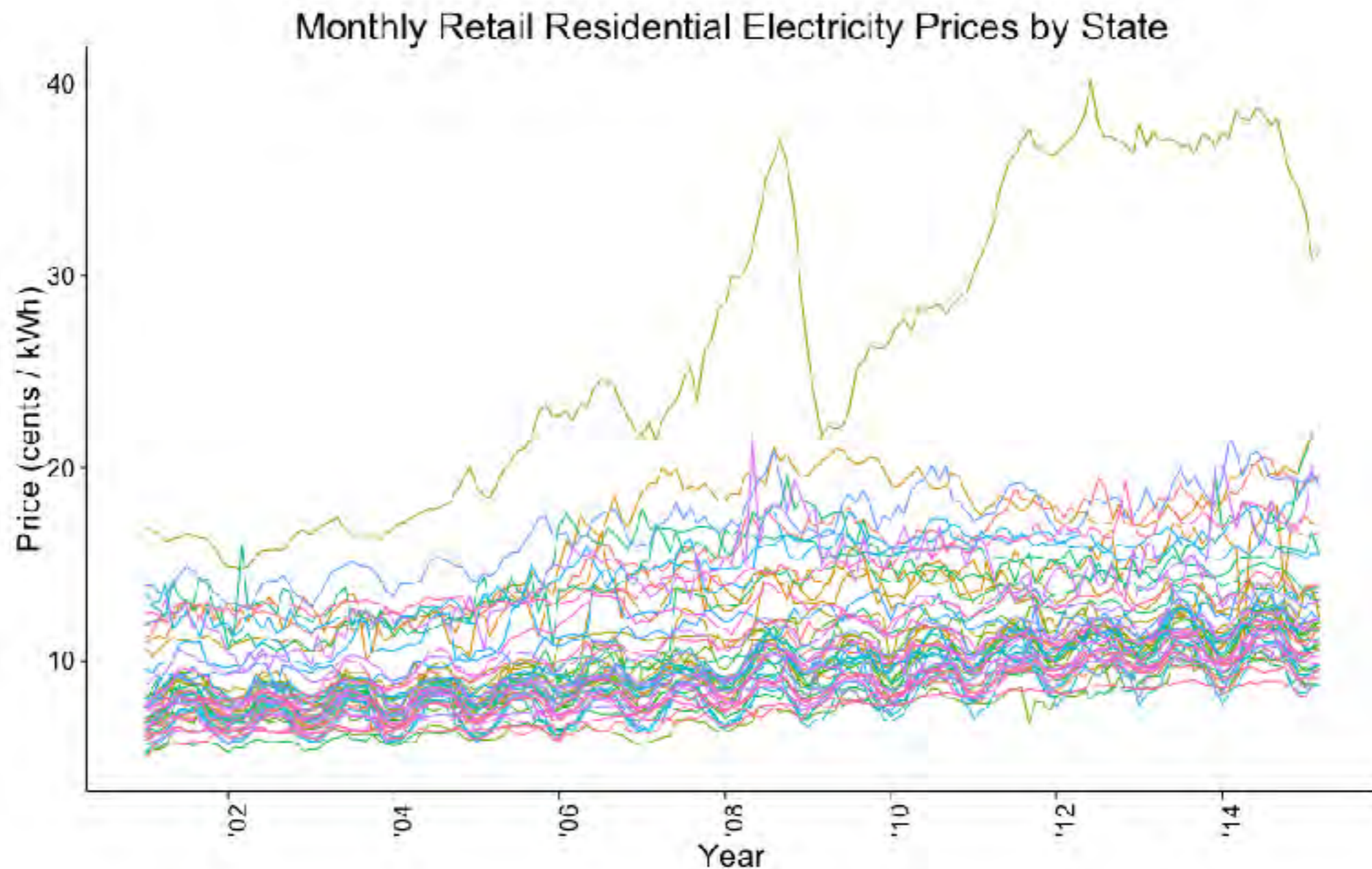


```
vif(mod_fe_state_year) # variance inflation factors
```

```
##          GVIF Df GVIF^(1/(2*Df))
## log(size)  1.092698  1      1.045322
## state      1.256685 50      1.002287
## year_installed 1.233002 15      1.007006
```

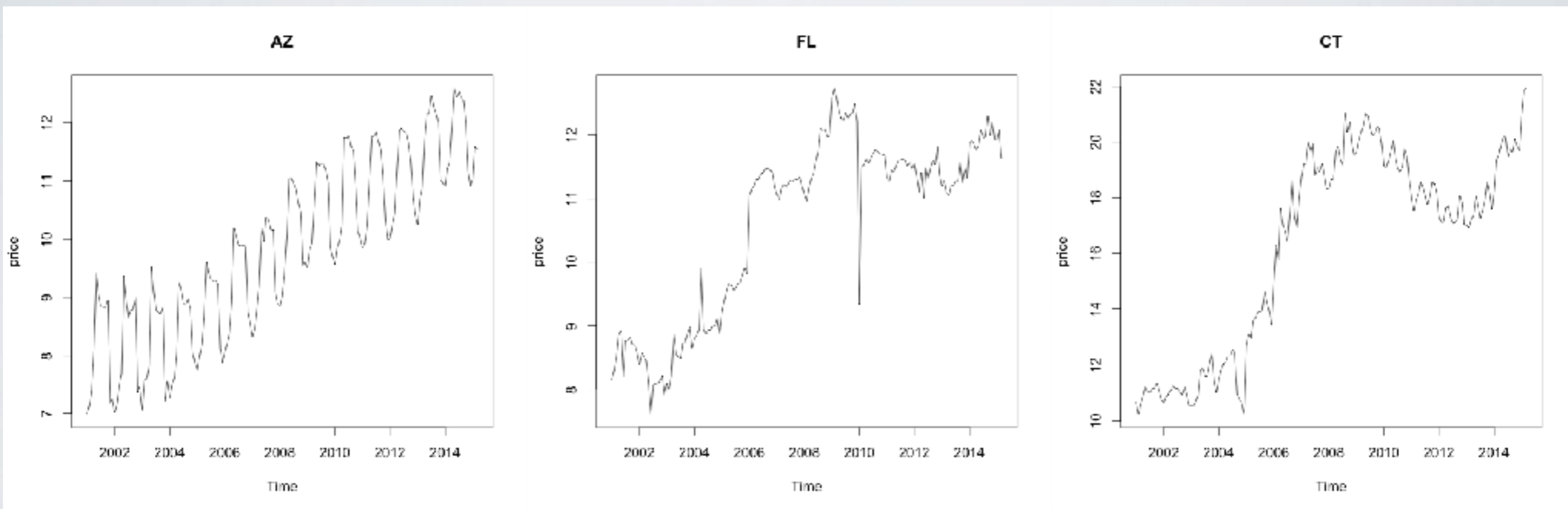

TIME SERIES MODELING

- No other predictors (time is the only variable)
- Strong a priori reason to believe most states will have an increasing, roughly linear trend in future electricity prices, often with seasonality



TIME SERIES MODELING

- States vary significantly from one another in historical prices, trends, and seasonality

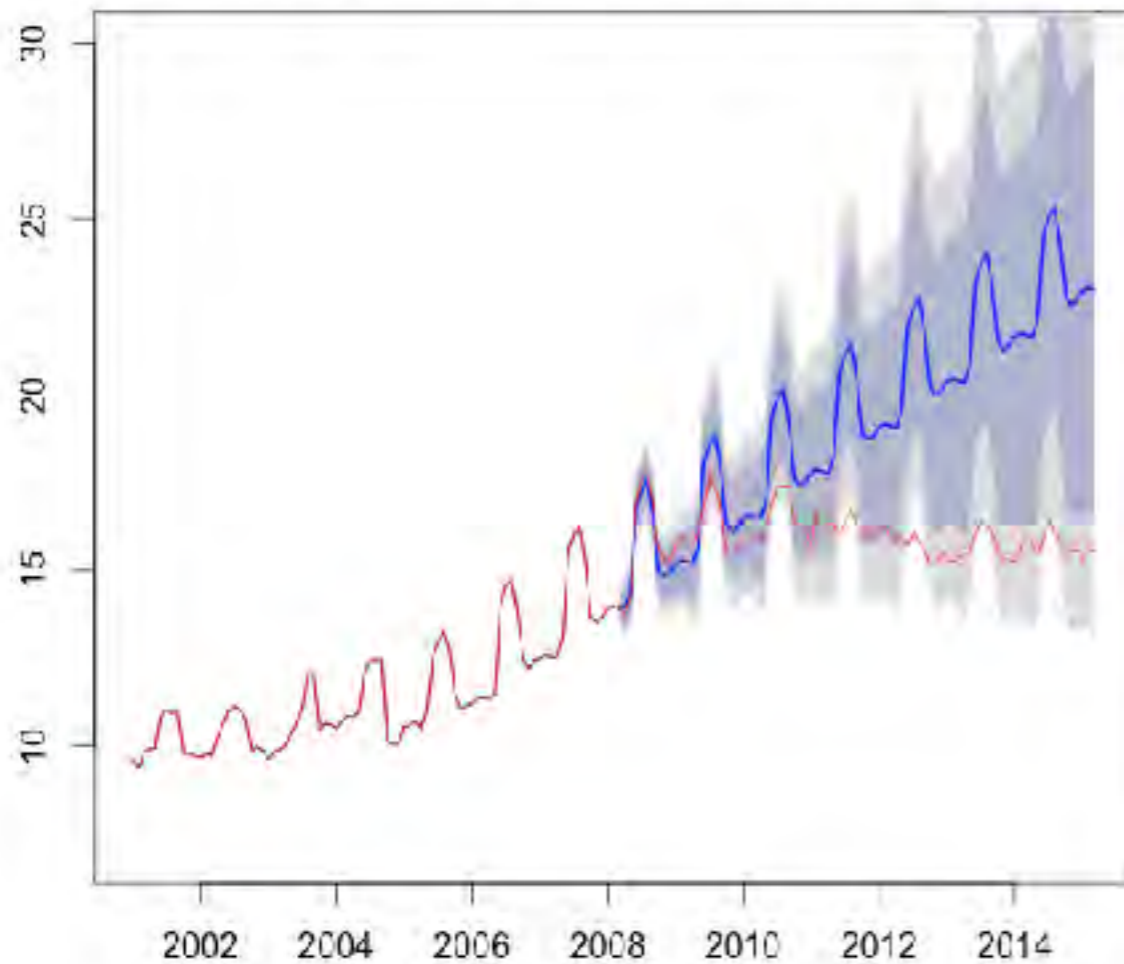


- We cannot expect the same model to perform well for all states!

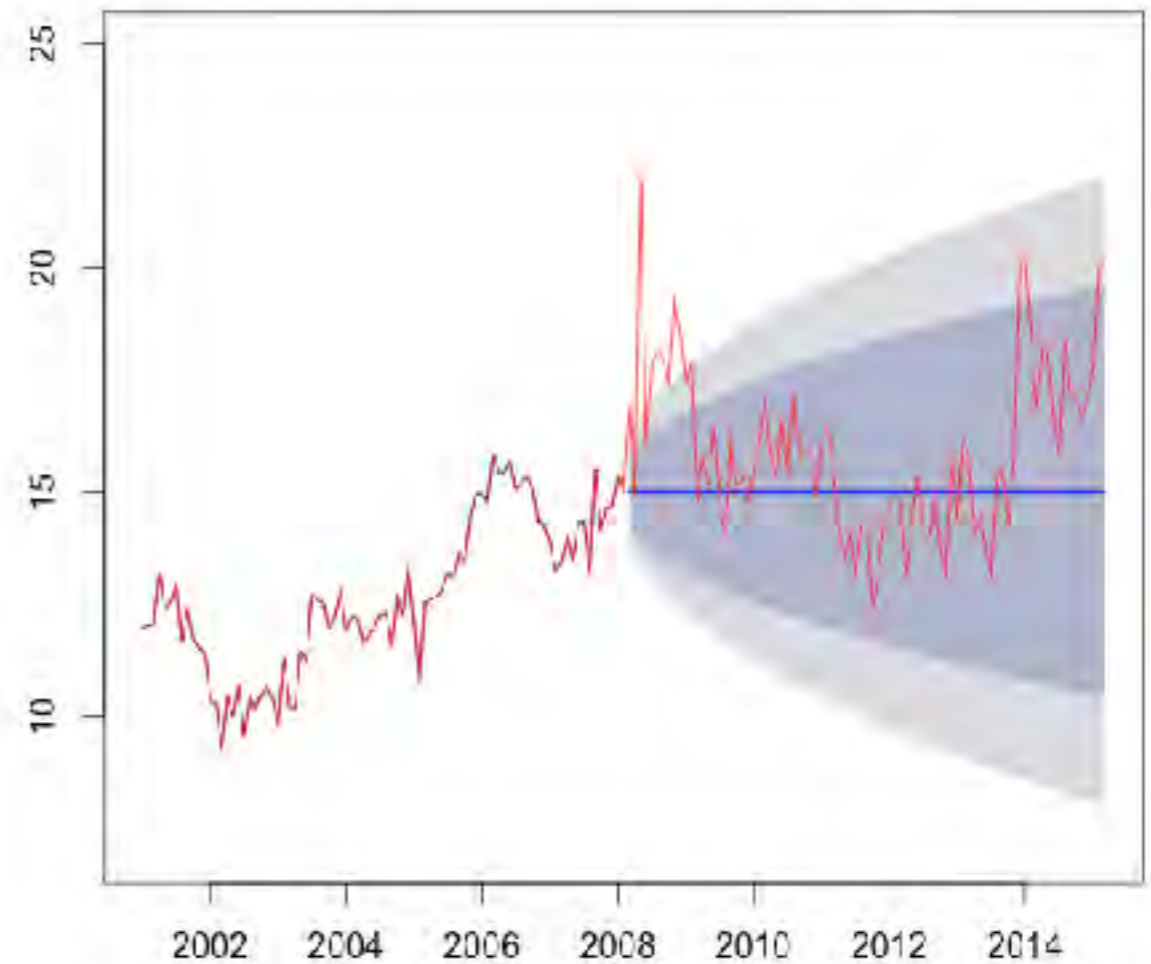
TIME SERIES MODELING

- Automatic model fitting is a bad idea for long term forecasts

New Jersey Auto-ARIMA forecast (blue) vs. actual (red)



Rhode Island Auto-ETS forecast (blue) vs. actual (red)



LONG TERM FORECASTING: A SOLUTION

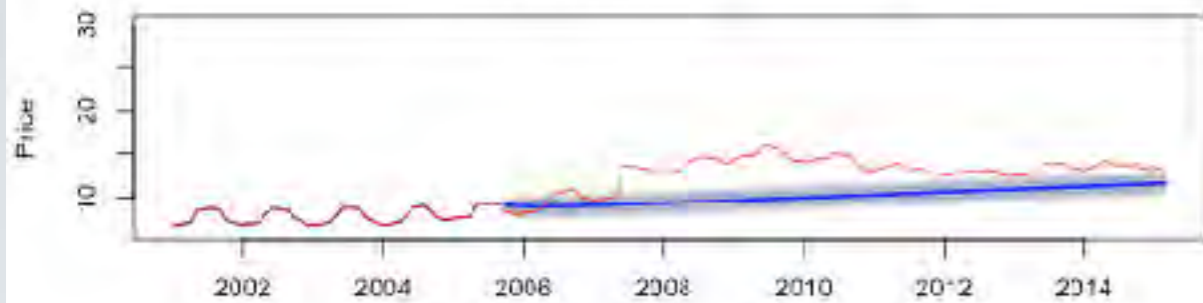
1. Create a handcrafted list of 7 possible models (1 linear, 4 ARIMA, and 2 exponential smoothing)

	Parameters	Seasonal Parameters	Note
Linear	n/a	n/a	
ARIMA	(1,0,0)	None	include drift
ARIMA	(1,1,0)	None	include drift
ARIMA	(1,0,0)	(1,0,0)	
ARIMA	(1,0,0)	(1,1,0)	
Exponential Smoothing	M	M	no damping
Exponential Smoothing	A	A	no damping

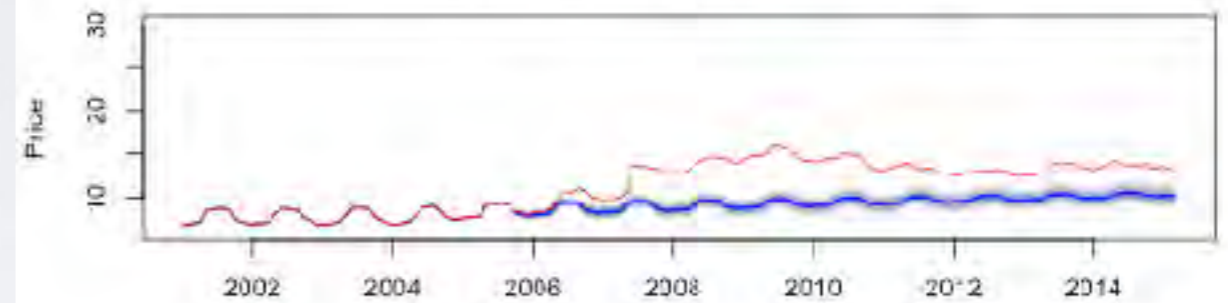
LONG TERM FORECASTING: A SOLUTION

2. Train each model on 1/3, 1/2, & 2/3 of historical data; test on the respective remaining proportion of historical data (2 models shown)

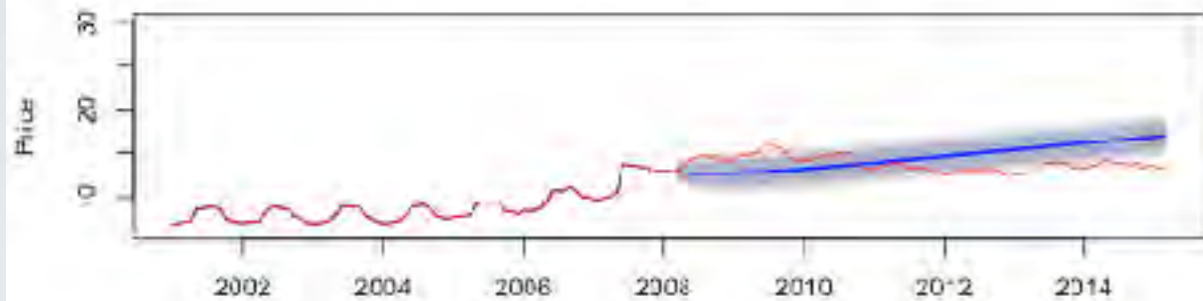
MD, ARIMA(1,0,0) trained on 1/3



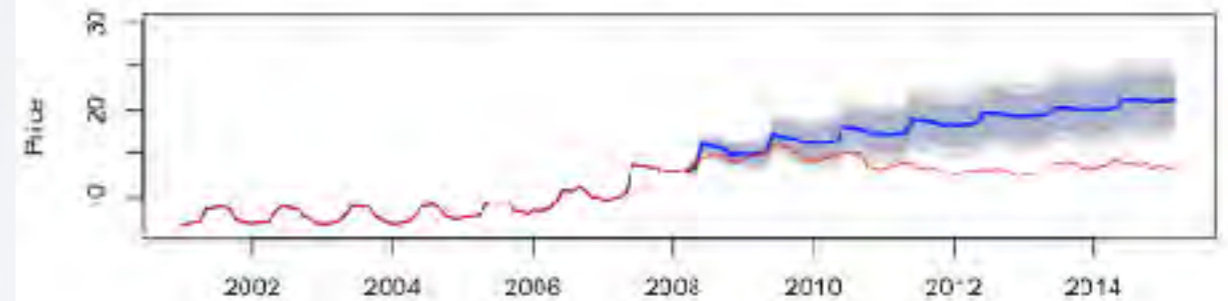
MD, ARIMA(1,0,0)(1,0,0) trained on 1/3



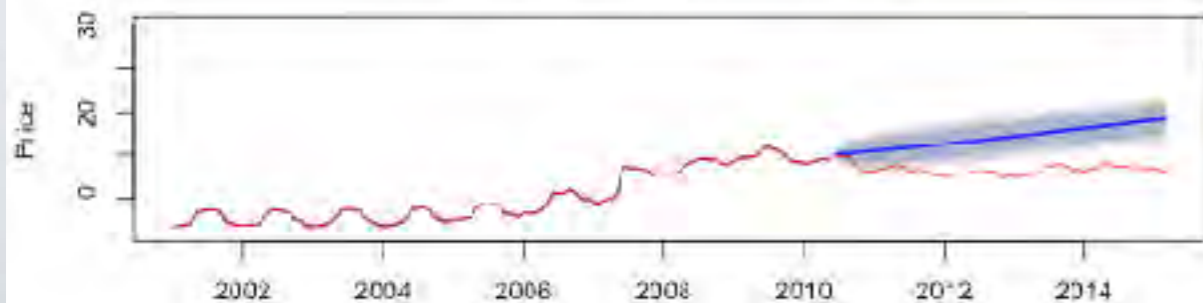
MD, ARIMA(1,0,0) trained on 1/2



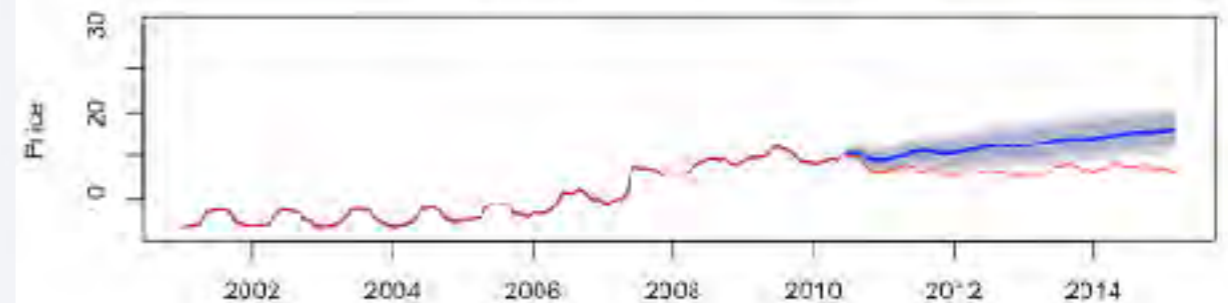
MD, ARIMA(1,0,0)(1,0,0) trained on 1/2



MD, ARIMA(1,0,0) trained on 2/3



MD, ARIMA(1,0,0)(1,0,0) trained on 2/3

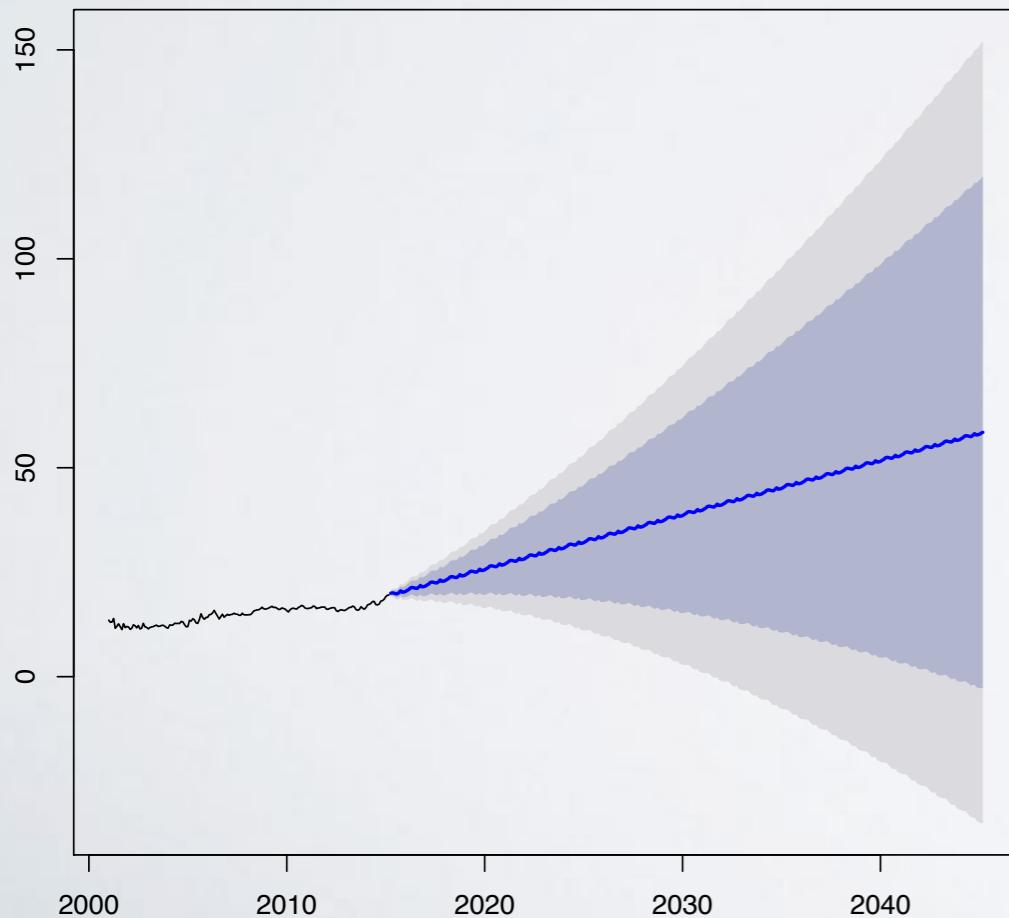


LONG TERM FORECASTING: A SOLUTION

3. Select the model with the lowest MSE across all tests
4. Repeat for every U.S. state + DC
5. Sanity check the resulting models

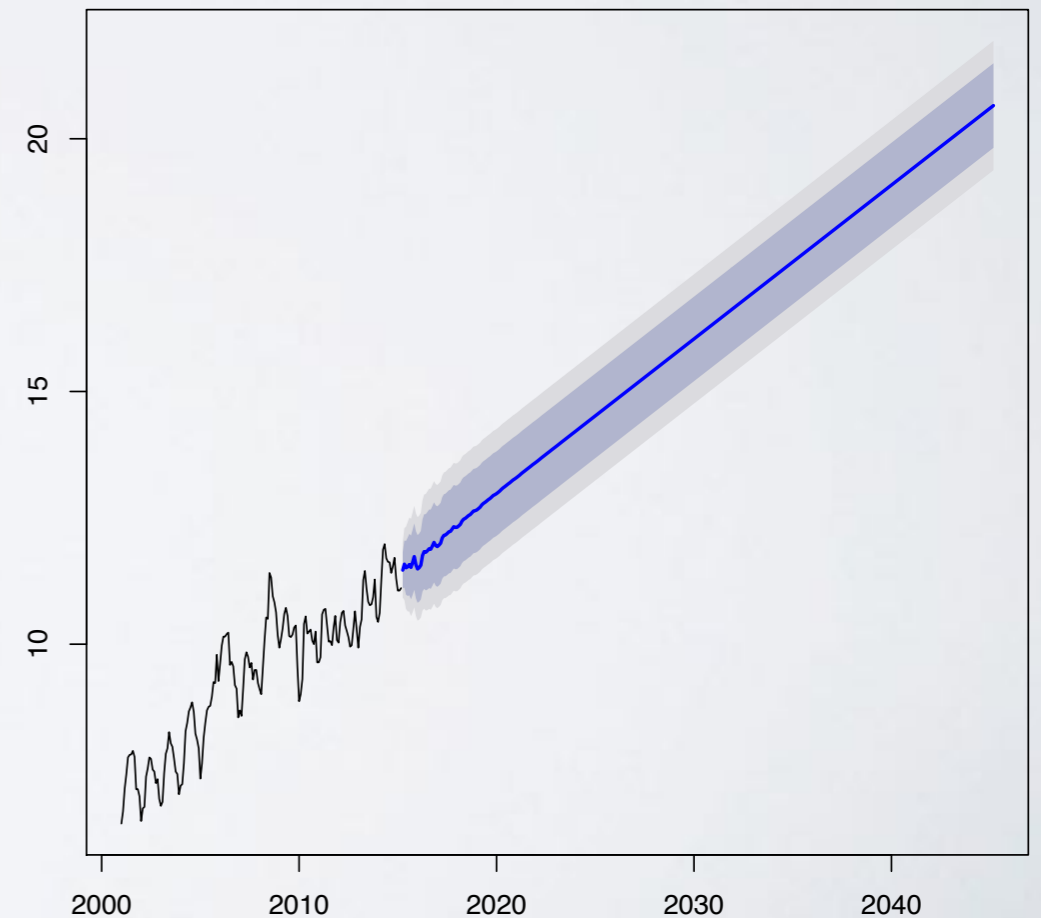
NH

Forecasts from ETS(A,A,A)



MS

Forecasts from ARIMA(1,0,0)(1,0,0)[12] with non-zero mean



ARIMA

- AR = autoregressive
 - Linear regression of current values against prior values
- I = integrated
 - We difference the series to achieve a stationary series
 - Alternatively one could fit a curve and use the differenced series
- MA = moving average
 - Linear regression of current values against prior errors

EXPONENTIAL SMOOTHING

- Single exponential smoothing:
 - Smoothed values a weighted function of prior observations and prior smoothed values
- Double exponential smoothing
 - Accounts for trends in the data
- Triple exponential smoothing:
 - Also accounts for seasonality

RANDOM FORESTS

- Ensemble of decision trees
- Decision trees prone to overfitting
- By growing multiple trees and taking a randomized subset of predictors for each tree, we avoid overfitting