Watts Up CA

Electricity Consumption and Forecasting in California

ADSP 31006
Time Series Analysis and Forecasting

Claire Boyd
Kathryn Link-Oberstar
Megan Moore
Eshan Prashar

Introduction

Overview: California Electricity Landscape

- **High Overall Consumption, Low Per Capita Consumption:**
  - California is the 3rd largest energy consumer in the nation but has the 2nd lowest per capita consumption.[1]

- **Transition to Carbon Neutrality requires more electricity from renewable sources:**
  - Goal for 100% renewable grid by 2040 (59% renewable as of 2022)[1]
  - Transition from Gas to Electricity puts more demand on the grid:[2]
    - i.e. 35% of new cars sold by 2026 must be electric, 100% by 2035.[2]
    - Estimated 3-fold increase in energy production required to meet these goals.[2]

- **Extreme temperatures put increased demand on the electrical grid, especially during the hot summers.**[2]

- **California leads the nation in energy efficiency standards:**[4]
  - Mandatory equipment and building standards
  - Voluntary customer programs (i.e. rebates)
  - Peak Pricing
  - Public Information Campaigns

- **Customers face some of the highest electricity costs in the nation.**[3]

In light of these competing forces influencing consumption, how has energy consumption changed over the last three decades?

. . .

How can we use these findings to anticipate consumption into the future?

Motivation

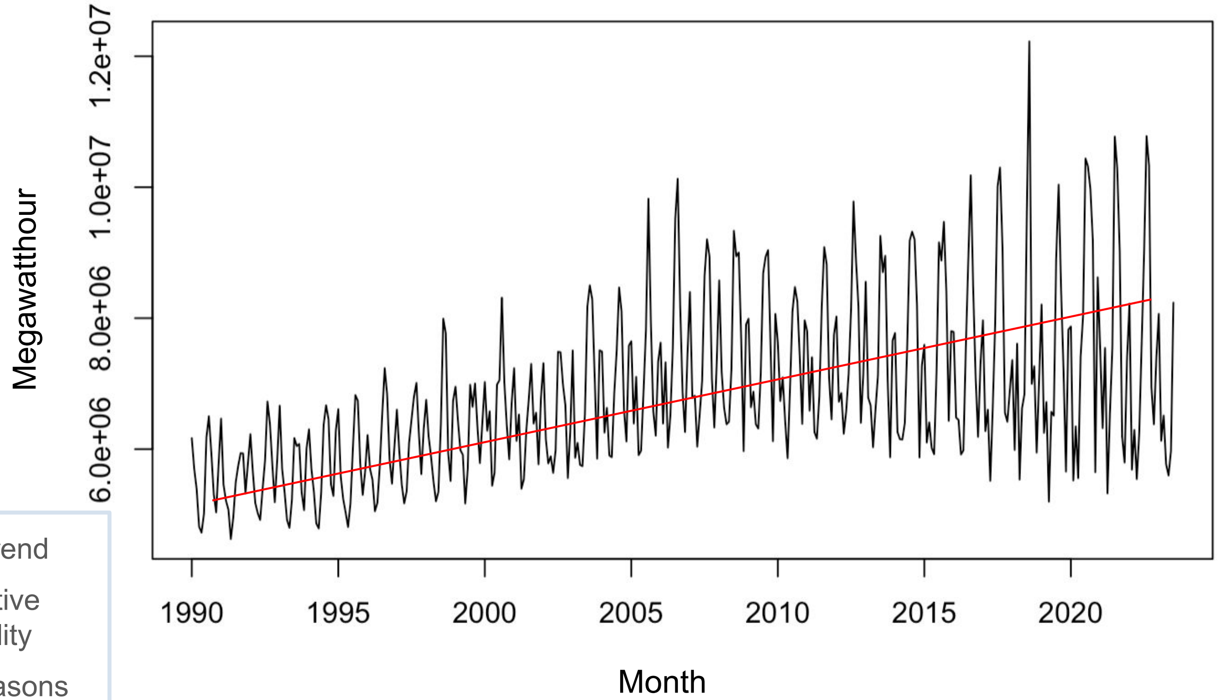
Evaluating California's residential electricity consumption may provide valuable insights into the potential efficacy and impact of electrification and and carbon neutrality efforts in other places.

- California has a history of leadership in environmental and energy policy, and ambitious goals for energy conservation and ambitious climate, electriciation and carbon neutrality history.

- Residential energy consumption:
    - Highlights factors impacting consumer behavior such as changing weather, public information, and rebates.
    - Energy supply & pricing has significant implications for equity and affordability.

Exploratory Analysis

Residential Energy Consumption California (Jan 1990 - June 2023)



Positive Trend

Multiplicative Seasonality

Monthly Seasons (frequency = 12)

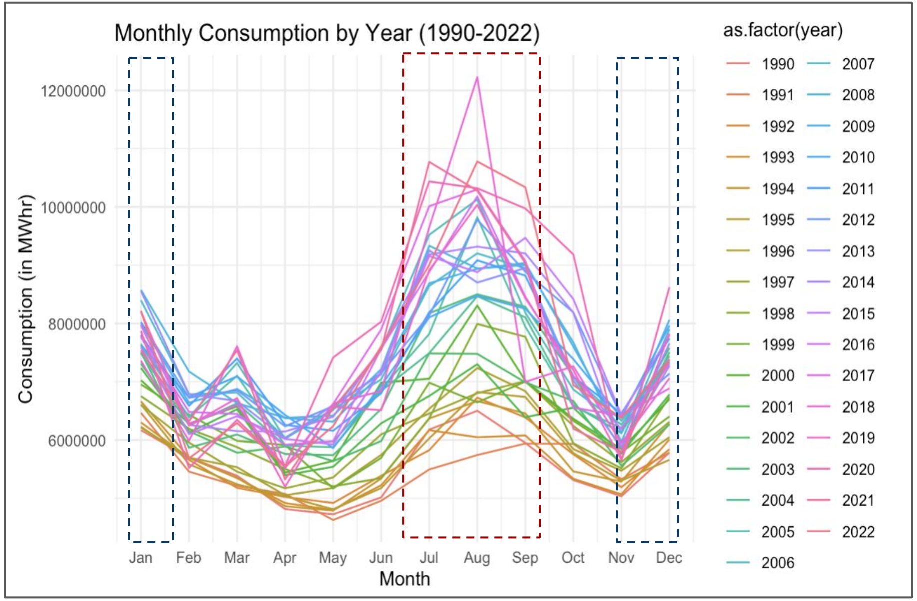
EIA-860 Annual Electric Generator Report, Monthly Data

Exploratory Analysis (1/3)

Overall Trends :

Overall values have increased across 3 decades; however increase is not uniform

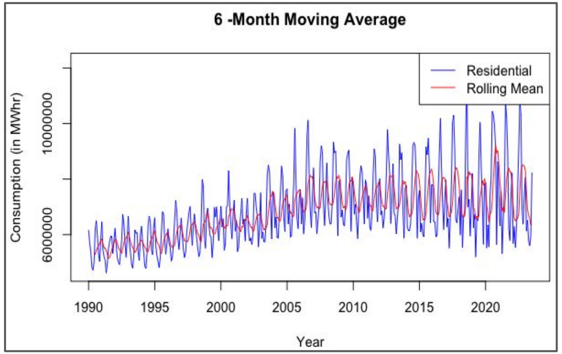
- Peak mean value in 2020, followed by 2008, and 2022. Peak monthly value in Aug 2018
- Monthly seasonal patterns:
  - Summer peaks during June-July-August
  - Winter peaks during Dec-Jan
  - Yearly lows seem to be in May

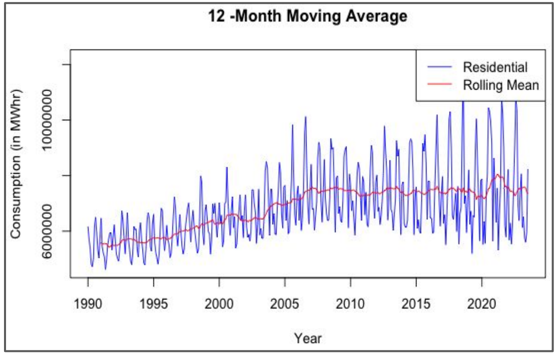


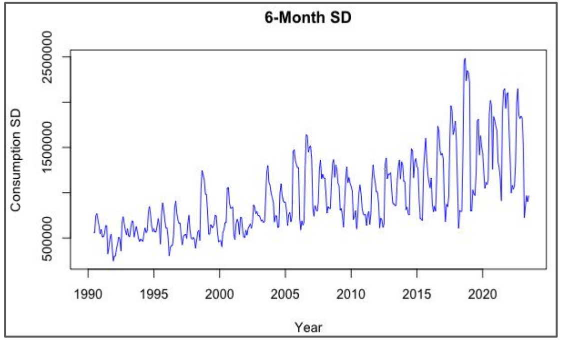
Exploratory Analysis (2/3)

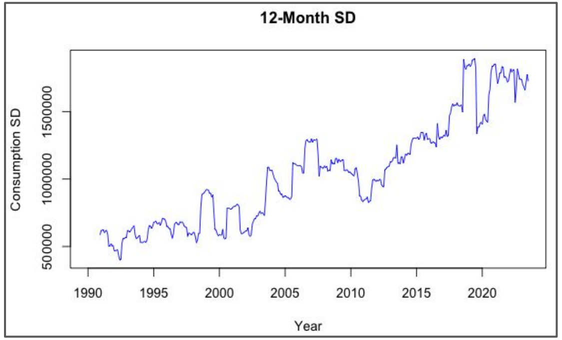
Seasonality:

- 12 monthly MA plot seems *wrinkle-free*: our starting point will be to assume yearly patterns in data
- However, the 12 month SD curve is still not quite smooth indicating that further analysis may be needed









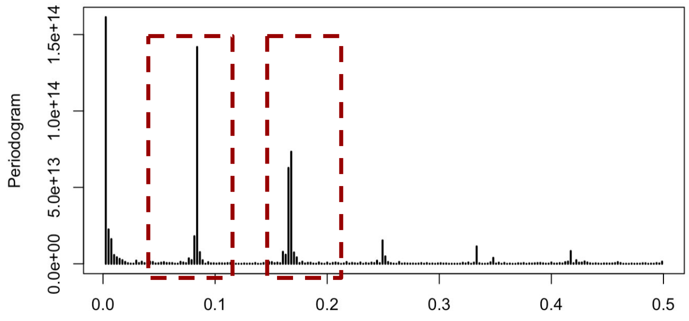
Exploratory Analysis (3/3)

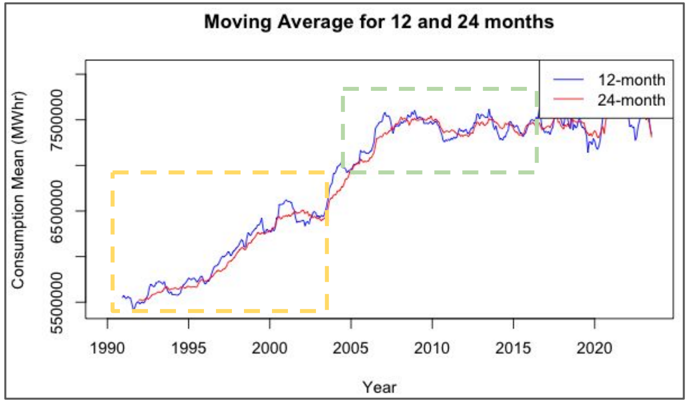
Seasonality (contd.) :

- We see additional seasonalities even in the periodogram

Different data generating processes:

- Visual inspection of moving averages indicates there could be multiple data generating processes roughly from 1990-2005, 2005-2018, and then 2013-Present
- In the early 2000s, increase in consumption is understandable; however the stabilization thereafter is difficult to model because of potentially multiple policy / behaviour causal factors





Baseline Models

Evaluation Metrics

1. **RMSE:** non-scaled metric (error is in magnitude of predicted variable units) that is calculable across different types of models.

2. **MAPE:** scaled metric (error is scaled as a percentage) that is calculable across different types of models.

3. **AICc:** Corrected AICc is useful because it normalizes for the sample size. This is especially useful to compare our long term and more recent time frames, but it is not used for regression.

Forecast time horizon (h): 12 observations (1 year)

Baseline Selection

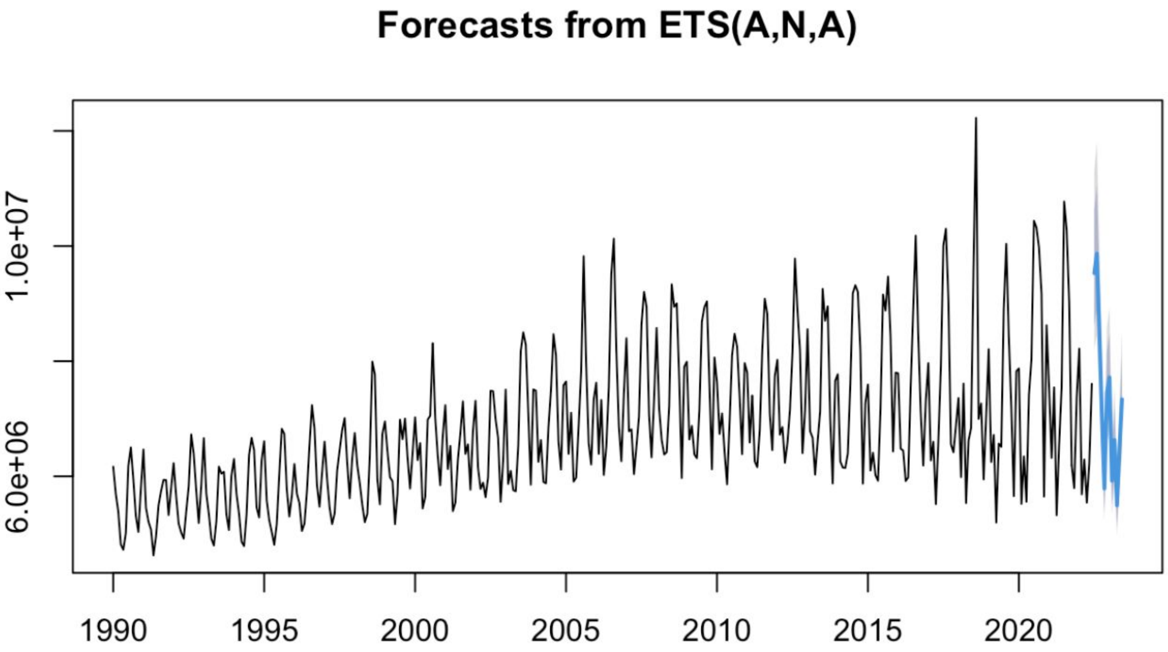
	Training Data				
	RMSE	**MAPE**	**AICc**	**Ljung-Box $p > 0.05$**	**KPSS $p > 0.05$**
Seasonal Naive	**562,981**	**5.52%**	**N/A**	2.109e-11 ❌	0.1 ✅
ETS	**466,623**	**4.58%**	**-10,893**	1.246e-2 ❌	0.1 ✅
TSLM (with trend and season)	**574,446**	**6.40%**	**N/A**	2.2e-2 ❌	0.01 ❌
Auto-arima	**461,164**	**4.58%**	**-1,003.24**	**1.955e-4** ❌	**0.1** ✅

Baseline - ETS

```
ets_model_1990 = ets(train_res_1990,lambda='auto')
```

Metrics for forecast:

RMSE: 466,623
MAPE: 4.5816
AICc: -10,892.98

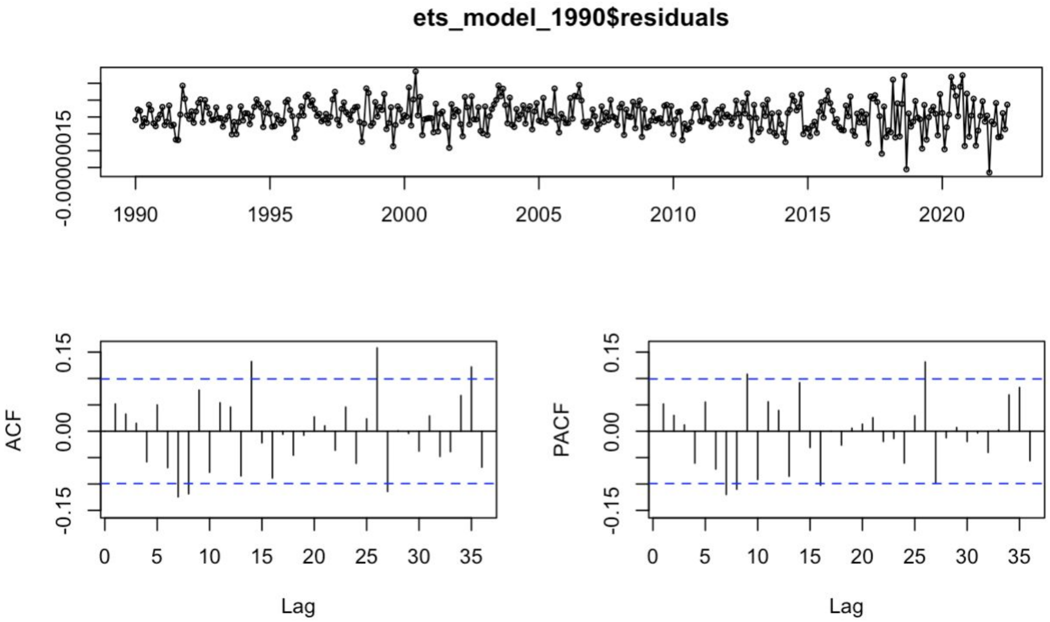


Analysis: Baseline - ETS

Observations:

Ljung Box: p-value = 1.246e-2; **Not white noise, still autocorrelation in residuals**

KPSS: p-value =0.1; **Data can be considered stationary**



Design Priorities

1. Model should include **secondary data** that is known to correlate with electricity consumption
2. Model should be able to handle **multiple levels of seasonality**

Next step model choices:

- Fourier Regression
- TSLM with independent variables
- ARIMA on residuals

Supplementary Data Exploration

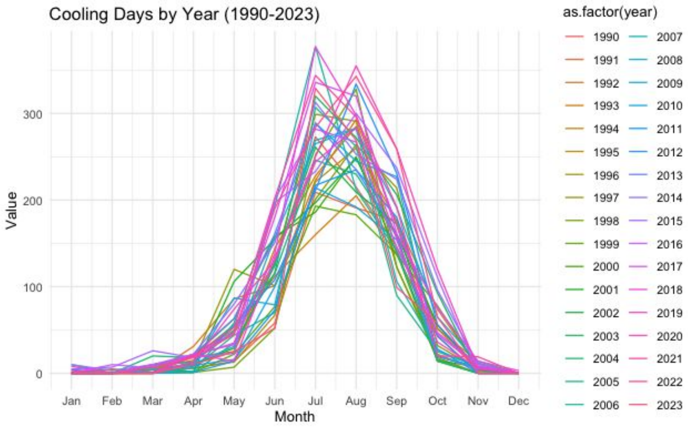
We also sourced **secondary data** that the field has indicated is correlated with electricity consumption:

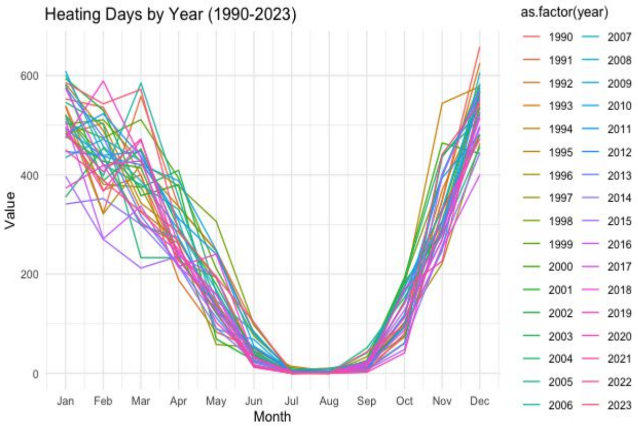
- *Weather*
  - *Monthly Temperature:* we used a variety of metrics to capture the fluctuations in temperature that might change energy consumption (extreme changes in weather may increase energy use)
    - Indicators: max temp (F), min temp (F), cooling degree days, heating degree days, average precipitation
- *Demand-Supply*
  - *Natural Gas Consumption: total monthly gas consumption (million cubic feet)*
  - *Electricity Prices:* average monthly price (cents/kWh)
  - *Number of residential electricity customers in California*

Highest correlated secondary data

Weather

- Cooling degree days (correlation of 62%)
- Heating degree days (correlation of -30%)

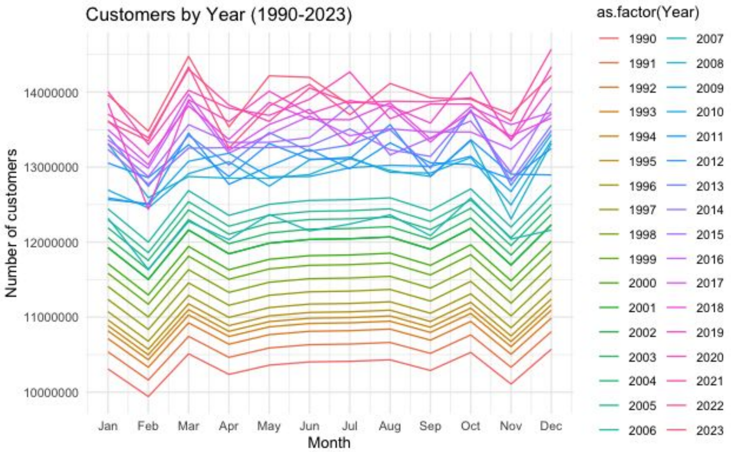


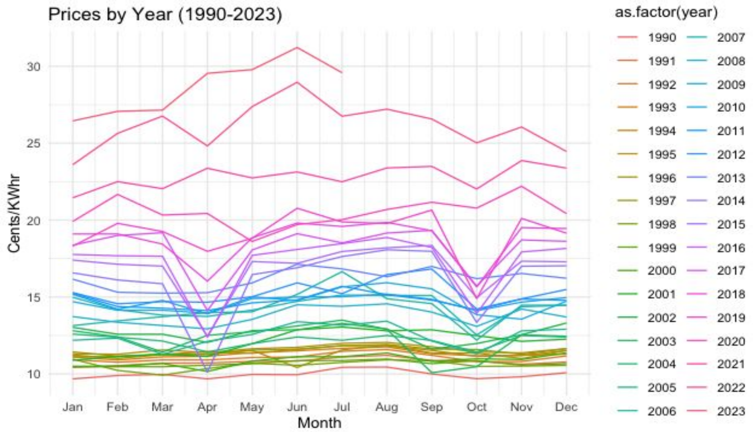


Highest correlated secondary data

Demand-Supply

- Customers (correlation of 57%)
- Prices (correlation of 45%)





Constructed Models

2 Problems at hand

1

Explore relationships with other independent variables

2

Solving for multiple seasonalities

TSLM + Independent Variables

Regression on Independent variables + **ARIMA on errors**

Regression on Fourier Terms + ARIMA on errors

Regression **with non-linear terms** on Independent variables + ARIMA on errors

TSLM with independent variables

Metrics for training data:

RMSE: 474,536
MAPE: 4.98%
AICc: NA

Observations:

Seasonal coefficients for multiple lags and independent variables are significant

```
residential_ts_multiv_reg_1990 <- tslm(train_res_1990 ~ trend + season + train_customers_1990 +
                                        + train_cdays_1990 + train_hdays_1990, lambda = 0)
```

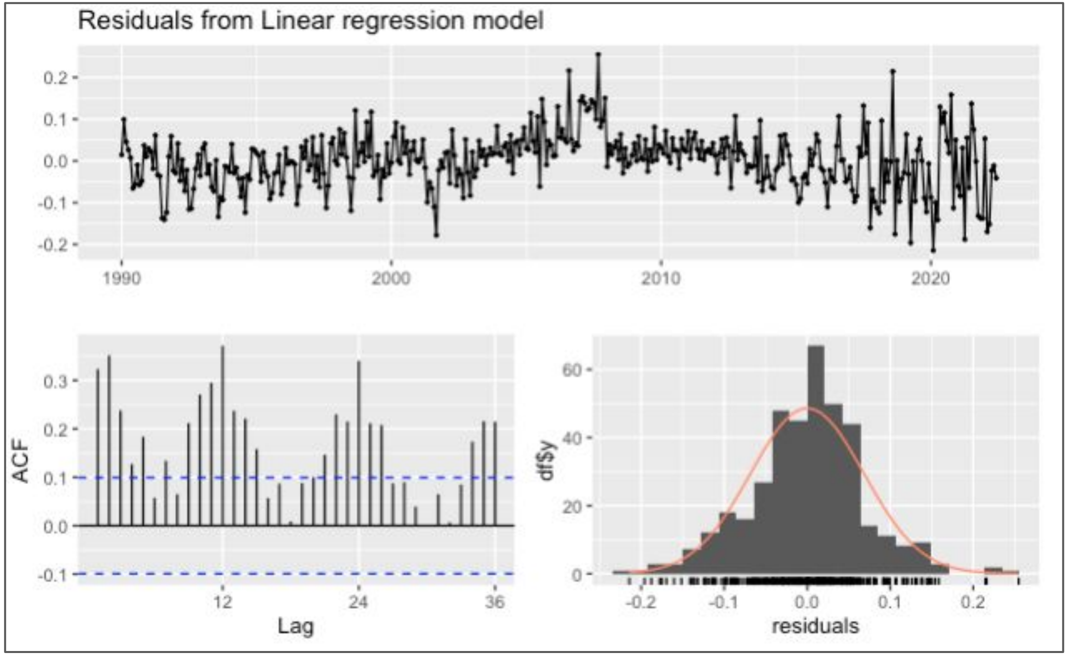
```
Coefficients:
                          Estimate     Std. Error t value             Pr(>|t|)
(Intercept)          13.53617743286  0.19597705185  69.070 < 0.0000000000000002 ***
trend                -0.00092296496  0.00017831502  -5.176       0.000000370752 ***
season2              -0.07032058867  0.01901149644  -3.699             0.000249 ***
season3              -0.16502544331  0.01871764146  -8.817 < 0.0000000000000002 ***
season4              -0.18990612582  0.02193596339  -8.657 < 0.0000000000000002 ***
season5              -0.18114014715  0.02770462472  -6.538       0.000000000204 ***
season6              -0.13860049072  0.03495308910  -3.965       0.000087831403 ***
season7              -0.10741194580  0.04375680798  -2.455             0.014553 *
season8              -0.04411074428  0.04397801619  -1.003             0.316501
season9               0.02714741574  0.03753448486   0.723             0.469969
season10             -0.05713364096  0.02990206599  -1.911             0.056809 .
season11             -0.12340076320  0.02087787088  -5.911       0.000000007681 ***
season12             -0.10696618346  0.01723515840  -6.206       0.000000001441 ***
train_customers_1990  0.00000018512  0.00000001858   9.965 < 0.0000000000000002 ***
train_cdays_1990      0.00125498686  0.00012876979   9.746 < 0.0000000000000002 ***
train_hdays_1990      0.00034553514  0.00006698733   5.158       0.000000405237 ***
```

Analysis: TSLM with independent variables

Observations:

However, our residuals still seem to be auto-correlated

TSLM only captures linear relationships and that might not be sufficient for our dataset



Regression with ARIMA errors

Metrics for training data:

RMSE: 414,933
MAPE: 4.1%
AICc: 10888

Observations:

Standard errors for the revised model seem reasonable

```r
# Regression with ARIMA on the errors
residential_reg_w_arima_err_1990 <- auto.arima(train_res_1990,stationary = FALSE,
                                               seasonal = TRUE,
                                               stepwise = TRUE,trace = TRUE,
                                               xreg = cbind(train_customers_1990, train_elec_prices_1990^2
                                                           ,train_elec_prices_1990,
                                       train_cdays_1990,train_cdays_1990^2, train_hdays_1990, train_hdays_1990^2))
```

```
Coefficients:
          ar1     ar2     ar3     ma1      ma2      sma1  train_customers_1990  train_elec_prices_1990^2
train_elec_prices_1990
      -0.2009  0.8312  0.2236  0.1580  -0.7724  -0.6466                0.5413                 -1878.676
32187.37
s.e.   0.0945  0.0609  0.0605  0.0904   0.0849   0.0435                0.1166                  2120.791
78284.27
      train_cdays_1990  train_cdays_1990^2  train_hdays_1990  train_hdays_1990^2
              4420.020              6.2961         -1886.311              4.0898
s.e.          3230.636              6.5695          2082.337              2.3521
```

```r
residential_reg_w_arima_err_v2_1990 <- auto.arima(train_res_1990,stationary = FALSE,
                                          seasonal = TRUE, # FALSE restricts to non-seasonal models
                                          stepwise = TRUE,trace = TRUE,
                                          xreg = cbind(train_customers_1990,train_cdays_1990^2, train_hdays_1990^2))
```

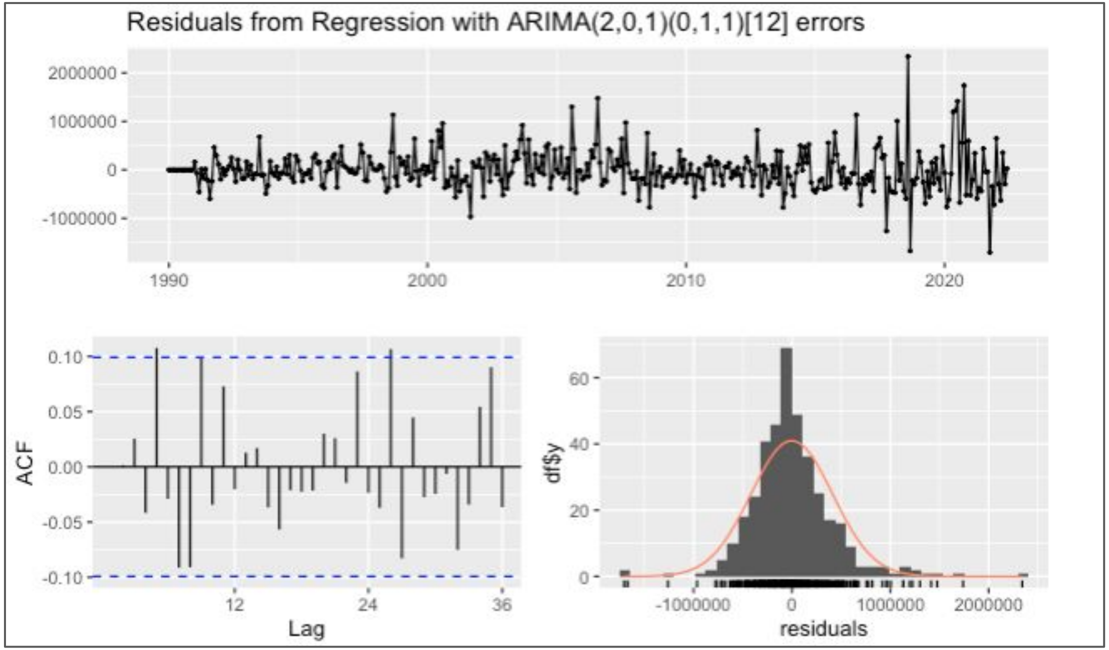
```
Coefficients:
         ar1     ar2      ma1     sma1  train_customers_1990  train_cdays_1990^2  train_hdays_1990^2
      0.7925  0.1476  -0.8286  -0.6866                0.5509             16.1744              1.4689
s.e.  0.0879  0.0632   0.0736   0.0399                0.1107              1.9219              0.4687
```

Analysis: Regression with ARIMA errors

Observations:

Residuals are not auto-correlated; adding quadratic terms seem to have worked

However, our residuals are still not normally distributed - there are some extreme values; which indicates potential heteroscedasticity



Fourier & ARIMA errors

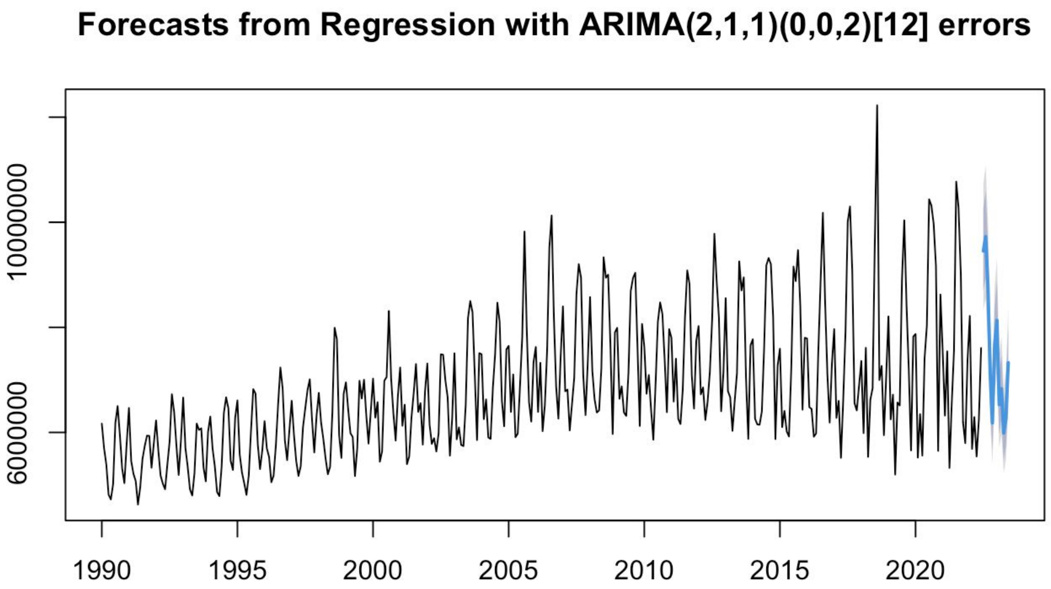
```
residential_fourier_1990 <- auto.arima(train_res_1990, xreg = cbind(fourier(train_res_1990,5)),
                                        seasonal = TRUE, lambda = 0)
```

Metrics for training data:

RMSE: 462,546
MAPE: 4.77%
AICc: -1,007

Observations:

Despite many seasonal patterns being observed, this did not seem to capture more information than previous models.

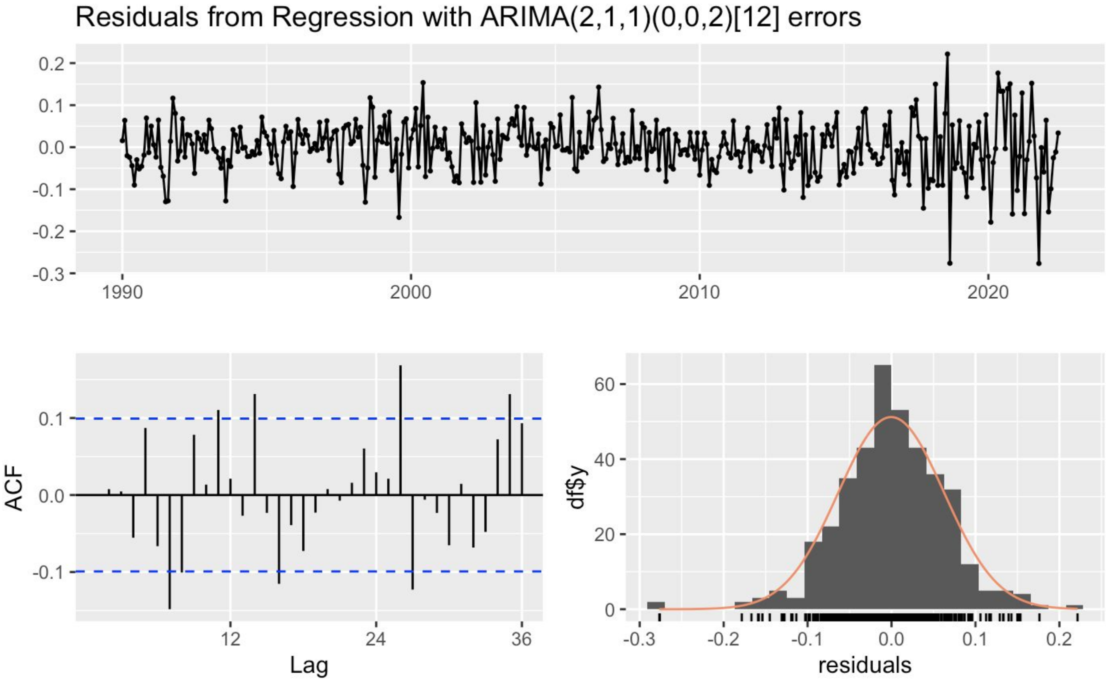


Analysis: Fourier & ARIMA errors

Observations:

Stationarity not achieved.

ARIMA on residuals captures a remaining component of seasonality that was not addressed by Fourier



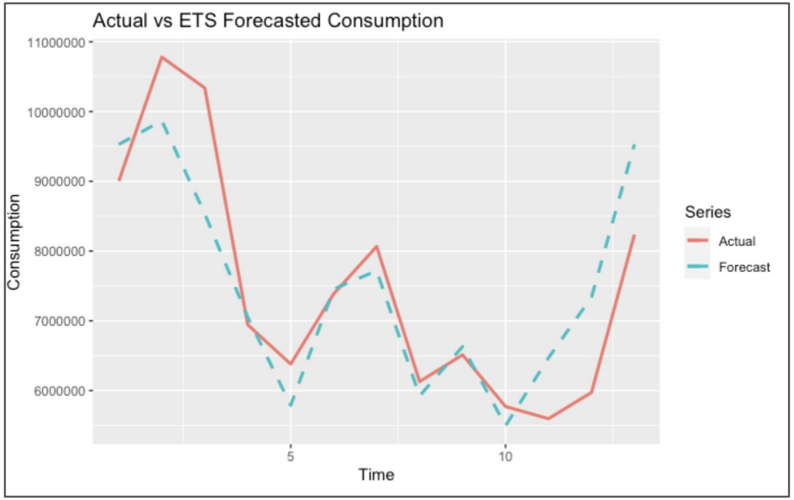
Constructed Model Selection

	Model fit of <u>train</u> data			Examining residuals		
	RMSE	MAPE	AICc	Ljung-Box > 0.05	KPSS > 0.05	Shapiro- wilk > 0.05
Baseline Model: ETS	**466,622**	**4.58%**	**-10,892**	**1.23e-3** ❌	**0.1** ✅	**0.01** ✅
TSLM with independent variables, trend, season	**474,536**	**4.97%**	**NA**	**2.2e-16** ❌	**1.14e-2** ❌	**3.30e-3**❌
Regression with all independent vars & ARIMA errors	**397,216**	**3.94%**	**10,865**	**6.796e-1** ✅	**0.1** ✅	**2.84e-12** ❌
Regression with select variables with ARIMA errors	**414,933**	**4.09%**	**10,888**	**1.7e-1** ✅	**0.1**✅	**4.31e-13**❌
Fourier series	**462,546**	**4.77%**	**-1,007**	**8.11e-4**❌	**0.1** ✅	**5.4×10−5** ❌

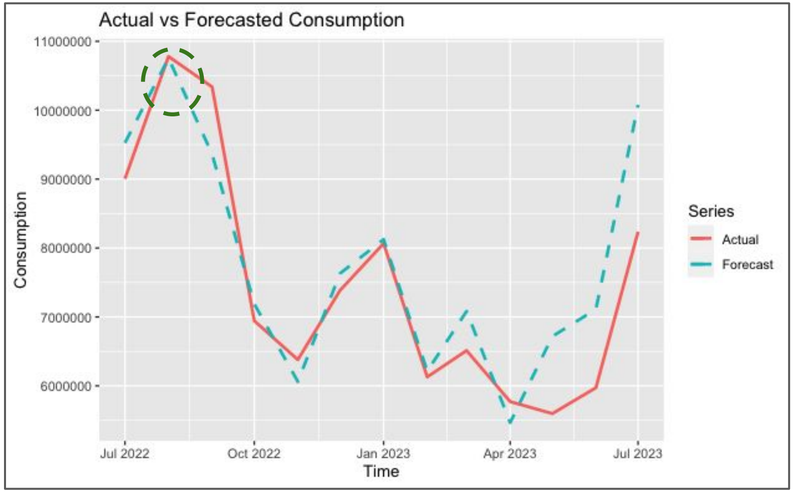
Model Performance

	Model fit of test data	
Model Name	**RMSE**	**MAPE**
Baseline: ETS	**796,074**	**8.02%**
Regression (w/ select variables) with ARIMA errors	**775,461**	**8.06%**
Fourier Regression + ARIMA	**729,852**	**7.50%**

Forecast of test data: ETS vs. ARIMAX

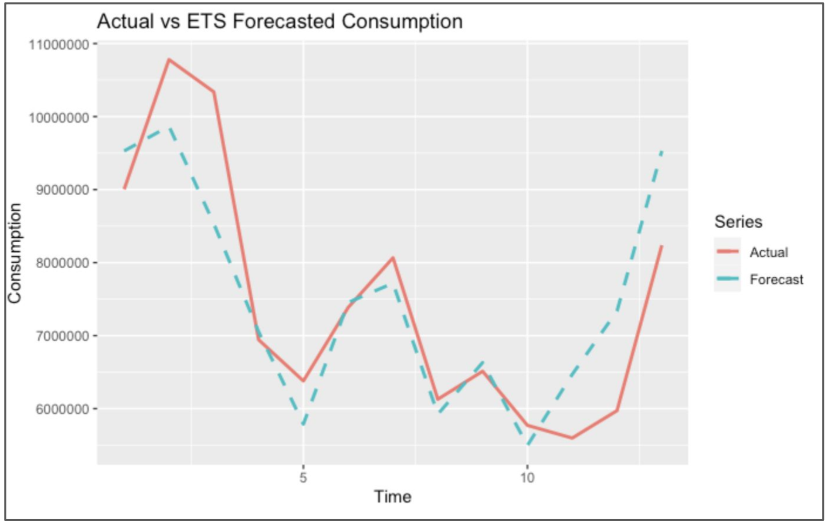


Base ETS Model

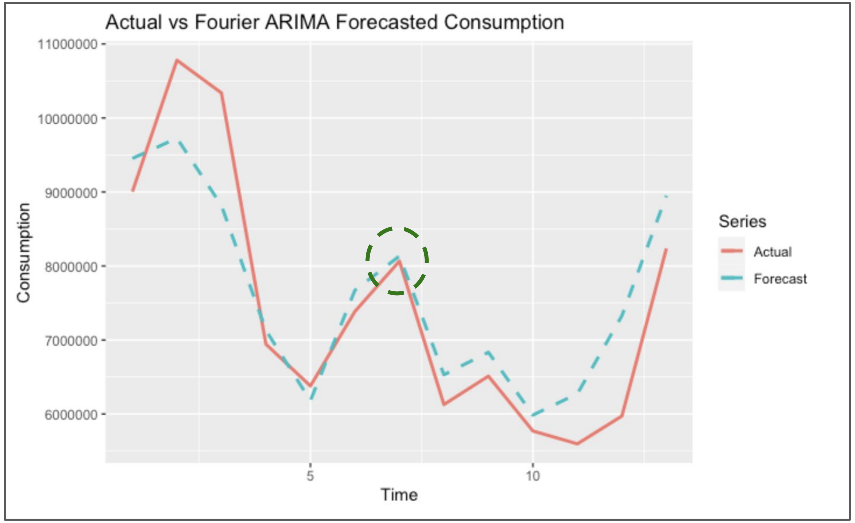


Regression with ARIMA errors: Peak value of August is really close

Forecast of test data: ETS vs. Fourier ARIMA



Base ETS Model



Fourier ARIMA: Better approximates winter/spring values

Concluding Thoughts

Core Takeaways

1. Multiple discrete time frames that seem to have clearly different trends and seasonalities
2. Many other variables factor into consumption, likely more than we were able to incorporate, and also in non-linear ways
3. There are multiple seasonalities at play

Potential Next Steps

If we had more time we would explore

1. **Dealing with potentially multiple data generating processes:** Subsetting time frames and recombining to capture trends that seem to not proceed through the full time series
2. **Independent variables:** Features engineering based on research
3. **Model Validation: K-fold cross validation**
4. **Intervention Analysis:** Looking at patterns pre and post Covid

Appendix

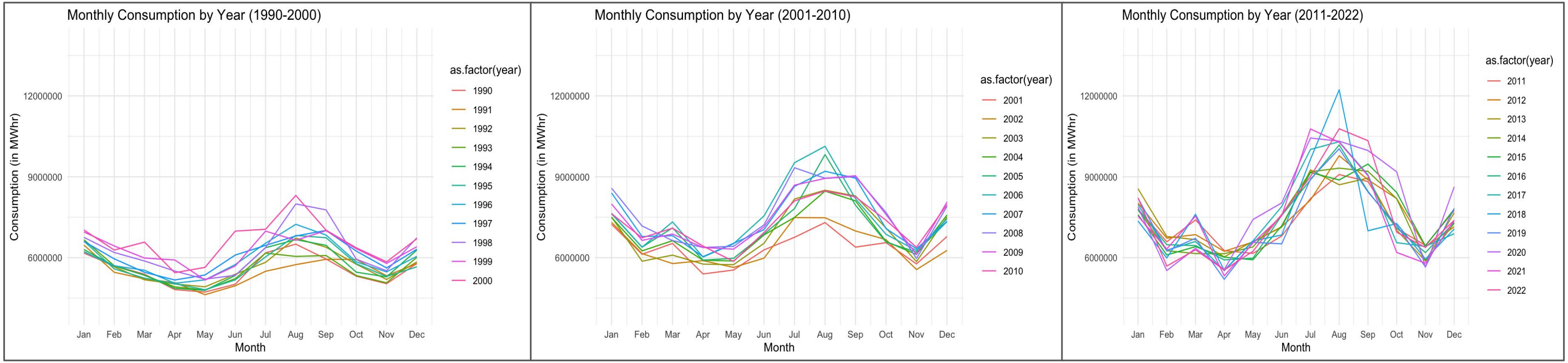
Team Member	Project Contributions
Claire	Constructing baseline models, building PPT/ QA on other model selection (evaluating residuals for baselines/constructed)
Kathryn	Data Cleaning & Loading; Exploratory Analysis (BC, SW, Differencing, etc.); Baseline ARIMA model; Train/Test Metrics; Background Research; Building PPT
Megan	Initial temperature data, regression with ARIMA error, Fourier transform with ARIMA error
Eshan	Gathered supplemental data, Exploring independent variables (customers, heating/cooling days, natural gas, price), creating/tuning Regression models, plots for overall trends in y and x values

Sources

1. https://www.eia.gov/state/print.php?sid=CA#41
2. https://calmatters.org/environment/2023/01/california-electric-cars-grid/
3. https://www.publicadvocates.cpuc.ca.gov/-/media/cal-advocates-website/files/reports/230224-public-advocates-office-2022-electric-rates-report.pdf
4. https://www.cpuc.ca.gov/-/media/cpuc-website/files/uploadedfiles/cpuc_public_website/content/news_room/fact_sheets/english/regulating-energy-efficiency-0216.pdf

All code used to for this analysis is publicly available here: https://github.com/meganhmoore/watts-up-ca

Residential Energy Consumption across decades (1990-2022)

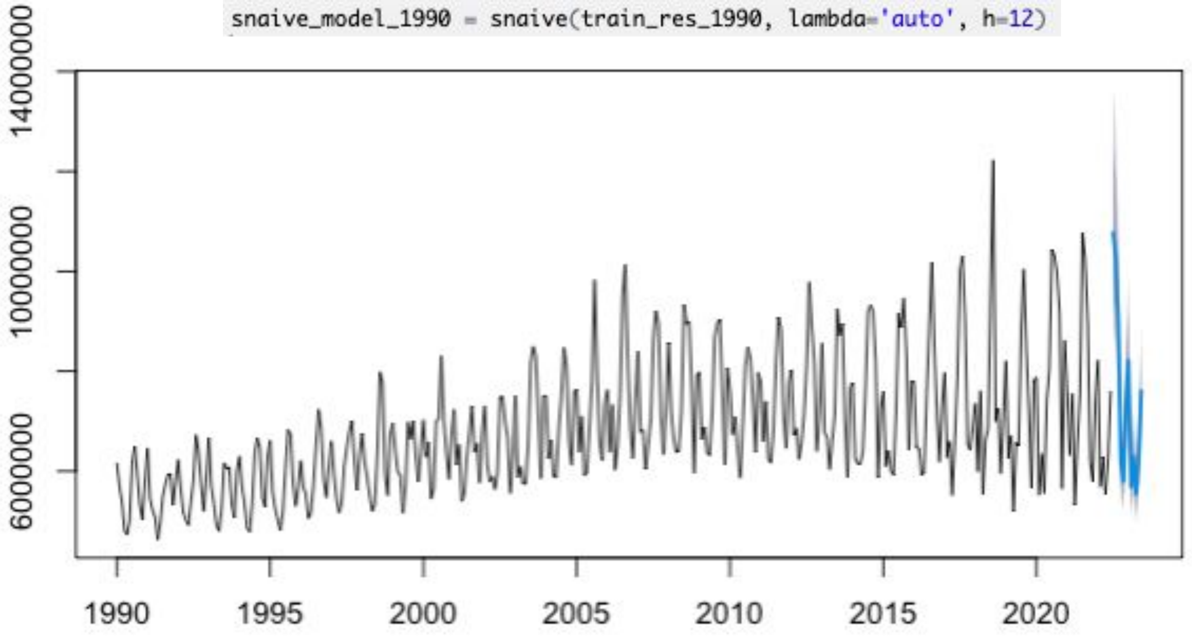


Baseline - Seasonal Naive

Metrics for forecast:

RMSE: 562,980.8
MAPE: 5.521388
AICc: NA

```
snaive_model_1990 = snaive(train_res_1990, lambda='auto', h=12)
```

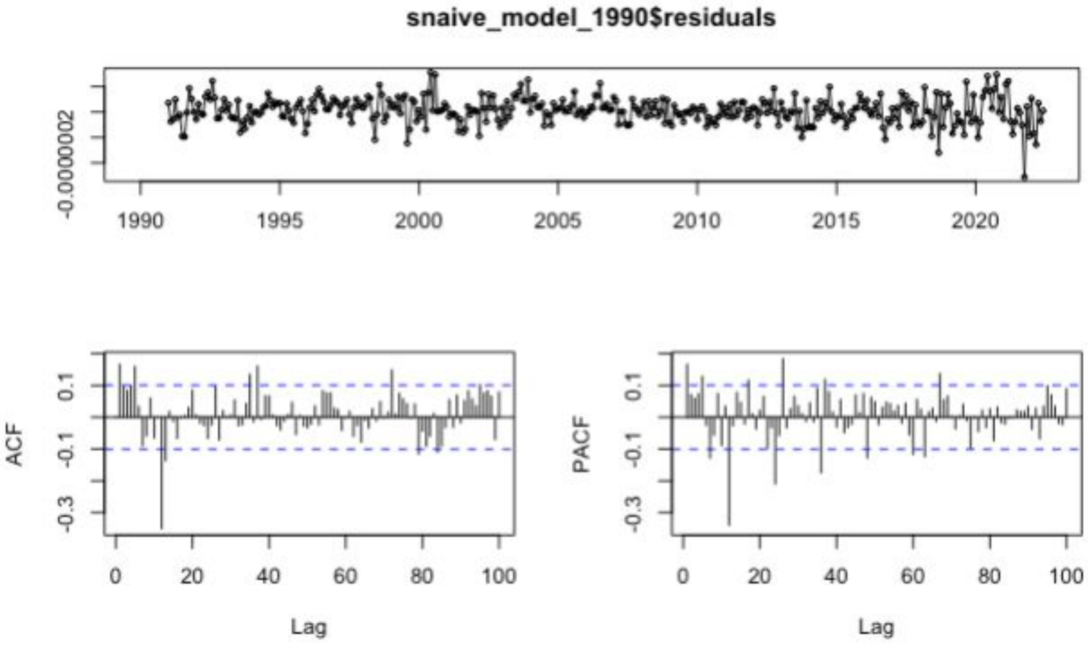


Analysis: Baseline - Seasonal Naive

Observations:

Ljung Box: p-value = 0.0000000001282; **Not white noise, still autocorrelation in residuals**

KPSS: p-value =0.1; **Data can be considered stationary**

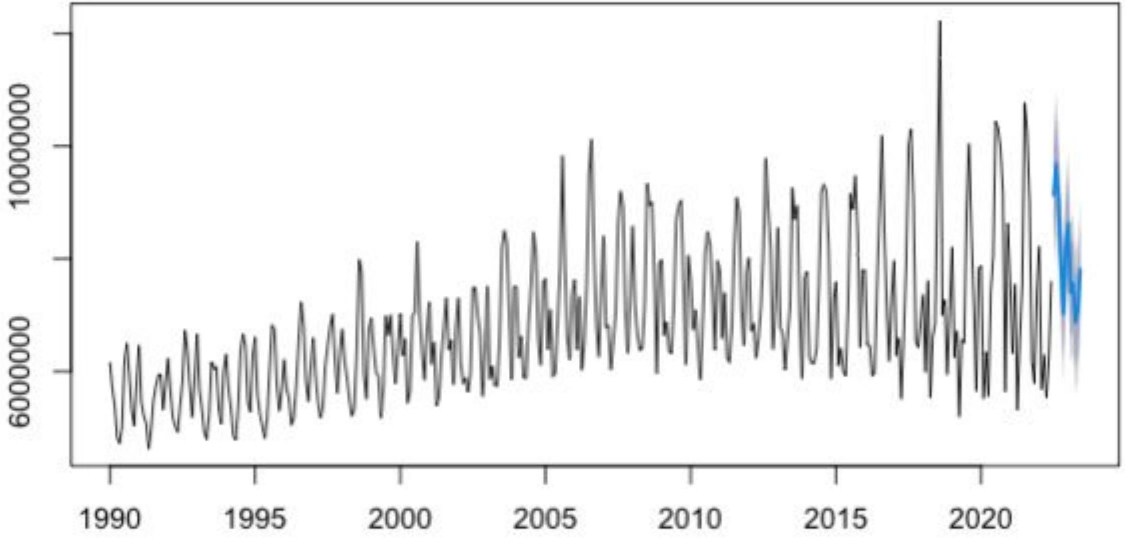


Baseline: Time Series Linear Regression

Metrics for training data:

RMSE: 610,919.6
MAPE: 6.877424
AICc: NA

```
tslm_model_1990 = tslm(train_res_1990 ~ trend + season)
```

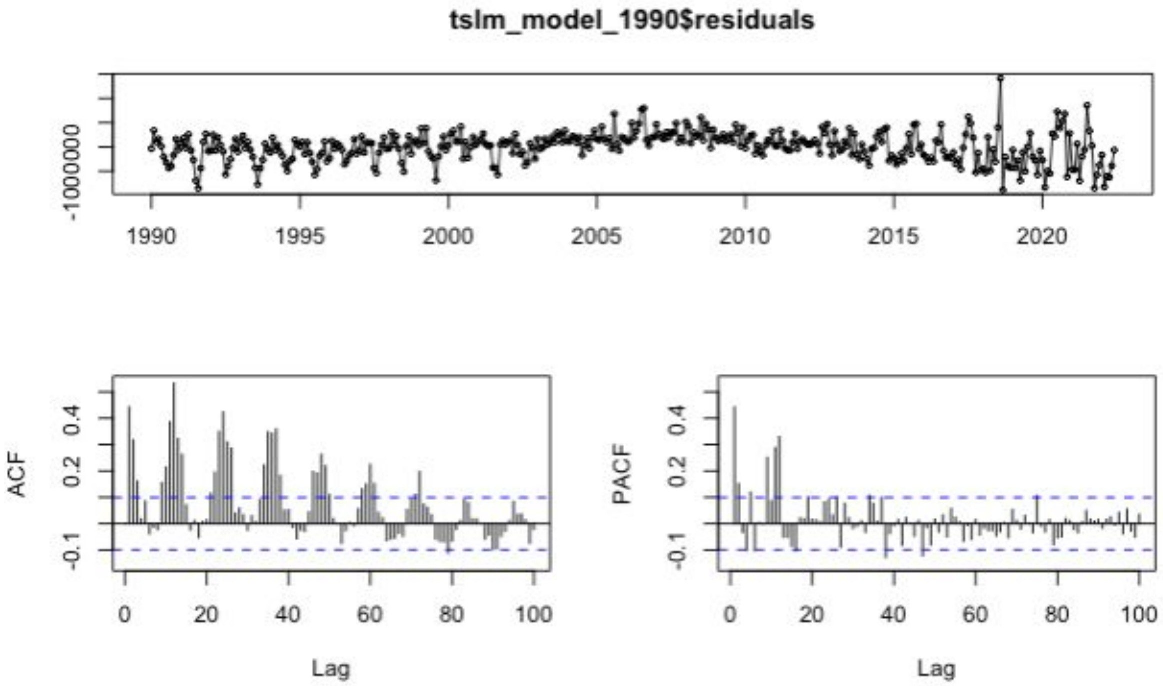


Analysis: Baseline - TS Linear Regression

Observations:

Ljung Box: p-value = 0.00000000000000022 ; **Not white noise, still autocorrelation in residuals**

KPSS: p-value =0.01096; **Not stationary, mean or variance are still a function of time.**



Baseline: Auto Arima

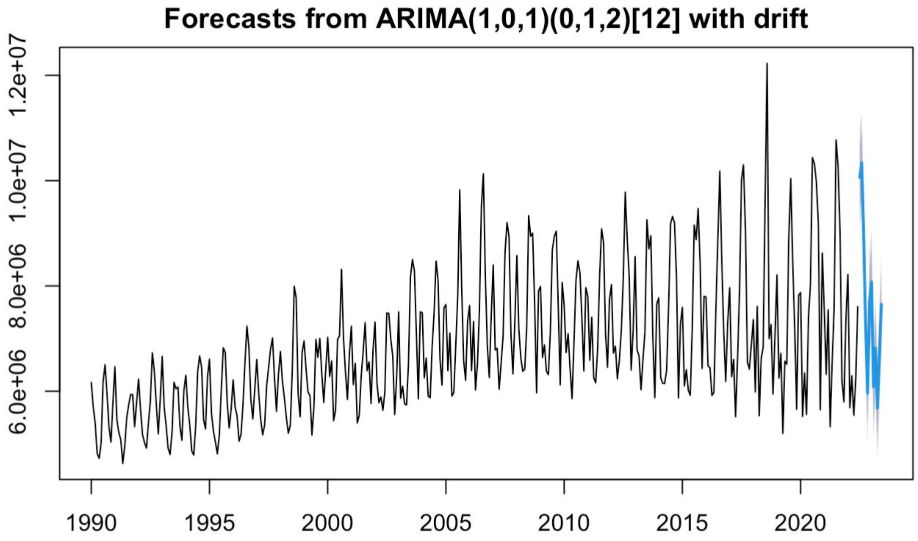
Metrics for training data:

RMSE: 465890.8
MAPE: 4.657618
AICc: 10970.39

Observations:

Good RMSE & MAPE, but AICc is high

```
residential_auto_arima_1990 <- auto.arima(train_res_1990)
```



Analysis: Baseline - Auto Arima

Observations:

Ljung Box: p-value = 0.01733; **Not white noise, still autocorrelation in residuals**

KPSS: p-value =0.06162; **Data can be considered stationary**