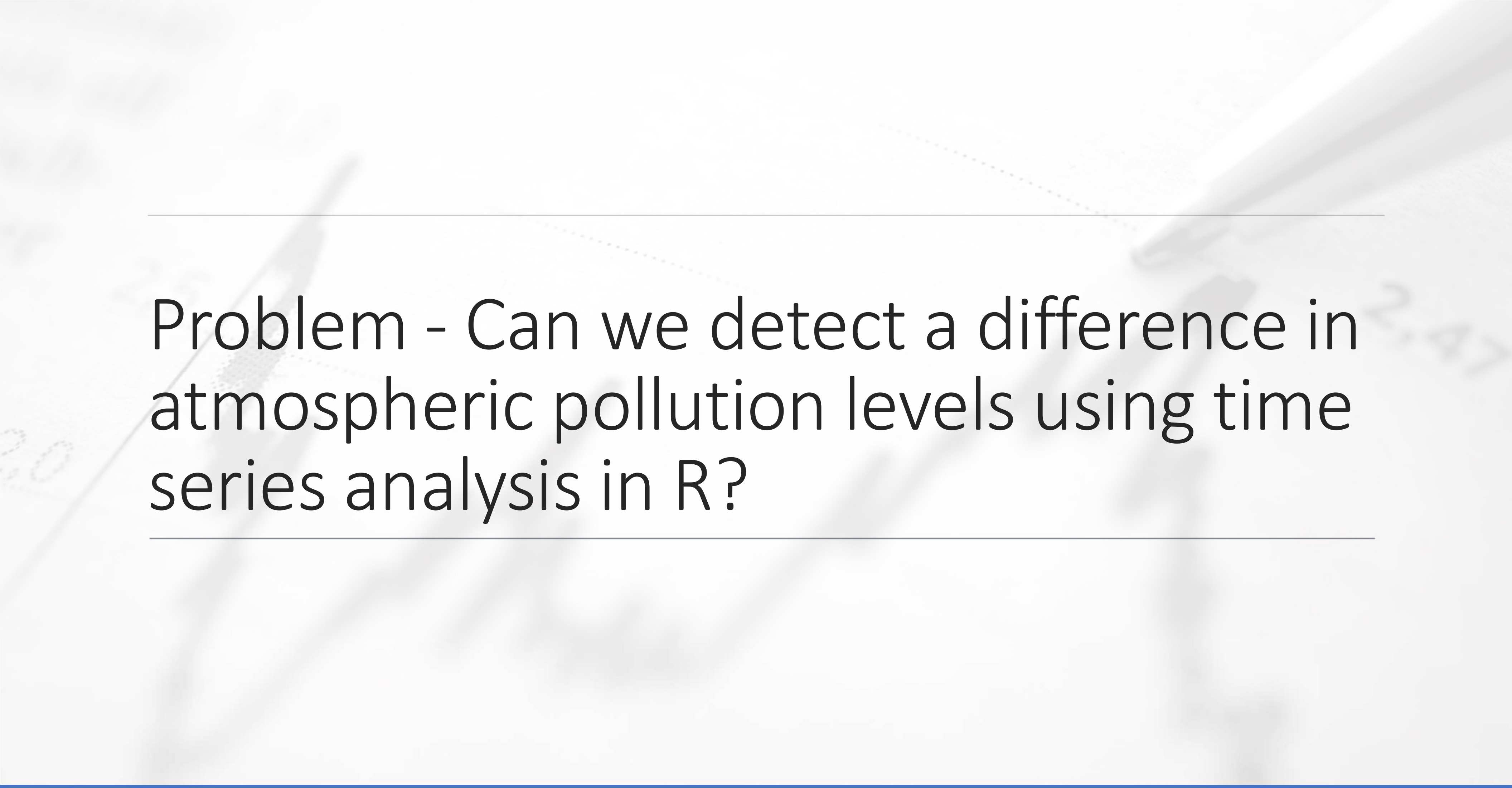
Means Testing Ozone Levels With Time Series in R

BY MAXWELL REESER



Problem - Can we detect a difference in atmospheric pollution levels using time series analysis in R?

Data Acquisition & Cleaning

- EPA has daily pollution data
  - Website: https://aqs.epa.gov/aqsweb/airdata/download_files.html
- Chose the period between January 1st 1996 and December 31st 2004
- Not all of it is pristine for database insertion
  - Required some manual cleaning
- Used PostgreSQL to further process data
  - Get county averages
  - Determine locations with consistent data collection

Fitting Models

- Used time series (ts) function in R, and nnetar to train
  - Tested across a large number of different splits for the data
  - Performed means test on 51 counties for each of 31 days worth of predicted data
- Best splits mostly have last 1-3 years in right group
- Significant differences are always negative
  - If a difference exists, ozone levels dropped

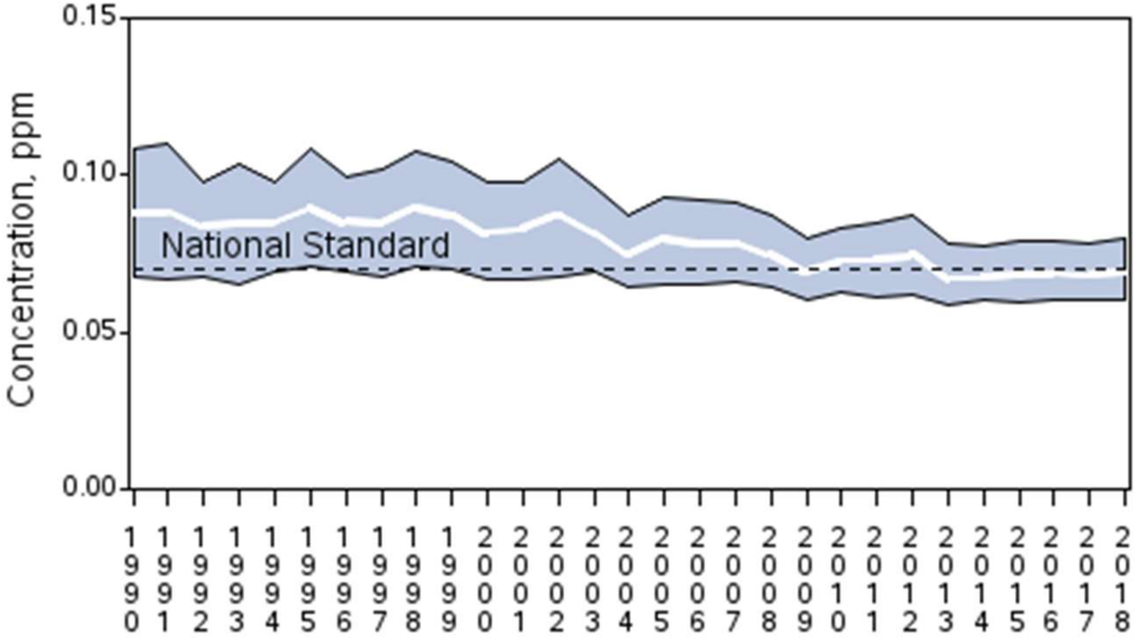
Model Split Name	Proportion Significantly Different	Graphical Representation of split
Even_split	0.355	🟩🟩🟩🟩🟧🟧🟧🟧
25_75	0.226	🟩🟩🟧🟧🟧🟧🟧🟧
75_25	0.516	🟩🟩🟩🟩🟩🟩🟧🟧
7_to_1	0.935	🟩🟩🟩🟩🟩🟩🟩🟧
1_to_7	0.387	🟩🟧🟧🟧🟧🟧🟧🟧
Even_split_4_year_gap	0.0645	🟩🟩🟦🟦🟦🟦🟧🟧
66_33_2_year_gap	0.0645	🟩🟩🟩🟩🟦🟦🟧🟧
Late_even_split	0.935	🟦🟦🟦🟦🟩🟩🟧🟧
Late_75_25	0.935	🟦🟦🟦🟦🟩🟩🟩🟧
Late_25_75	0.935	🟦🟦🟦🟦🟩🟧🟧🟧
Left_625	0.484	🟩🟩🟩🟩🟩🟧🟧🟧

Conclusion

- Using time series to conduct means tests was effective
- Research online indicates that ozone levels did in fact drop off steeply after 2002
  - EPA chart at left
  - Consistent with the results from the models

Ozone Air Quality, 1990 - 2018

(Annual 4th Maximum of Daily Max 8-Hour Average)
National Trend based on 414 Sites



1990 to 2018 : 21% decrease in National Average