The role of news and social pressure in shaping U.S. public opinion on climate change

Data documentation

Masako Ikefuji

OSIPP, University of Osaka, Japan

Jan R. Magnus

Department of Econometrics & Data Science, Vrije Universiteit Amsterdam
and Tinbergen Institute, Amsterdam, The Netherlands

Zhehao Wang

OSIPP, University of Osaka, Japan

1 Introduction

We are interested in how news and social pressure affect U.S. public opinion on climate change. We choose U.S. as our research subject because U.S. is vast and full of heterogeneity. Furthermore, the climate policies of two main parties in U.S. (Democratic Party and Republican Party) are much different.

U.S. is divided into more than three thousand counties and county equivalents where individuals live in. In Section 2, we select 3,111 counties and county equivalents so that we can assemble data from several different sources across our entire study period – 1990 to 2024.

We attempt to use county-level data rather than national-level data to preserve the heterogeneity across U.S. However, our dependent variable – public opinion – is only available at national-level. For other factors that may influence the perception of individuals on climate change as well as the characteristics that capture the heterogeneity of counties – such as general election turnout, population, age structure, educational attainment, race, house price, and income – we are able to obtain county-level data. The general election data are based on four-year intervals, while other data are available with finer frequency. In Section 3, we collect all data as frequently as they are presented because the more frequent observed real data we have, the stronger our dataset is.

Then we have data at different time intervals, which can be treated as a missing data problem. In Section 4, we interpolate all variables to a monthly frequency. Finally, we are able to construct balanced panels for counties from 1990 to 2024, spanning 420 months.

In Section 5, we visualize historical information on climate disasters and news coverage of climate change.

2 Counties

The United States of America is a federation of 50 states, Washington D.C. (capital district), and 326 Indian land areas. The country is the world's third-largest land area with a whole population around 340 million at the moment. The 50 states are divided into over 3000 political subdivisions of states called counties. There are also several hundred county equivalents, including the District of Columbia (i.e. Washington D.C.) and county equivalents in Puerto Rico, which is an unincorporated territory of the United States. Although Puerto Rico legally belongs to U.S., it is exactly not a part of U.S. Due to this reason, we exclude Puerto Rico in this research. Additionally, we should note that Louisiana is divided as parishes, Alaska is divided as boroughs and census areas, while Maryland, Missouri, Nevada and Virginia have some independent cities. However, these areas can all be treated as counties.

We assemble date from various kinds of sources, spanning the year from 1990 to 2024, and sort all counties by FIPS (Federal Information Processing Standards) codes. We set the data of 2020 as a benchmark and clear up the data of other years to ensure the consistency and comparability. The manipulations are as follows:

1. Alaska (2) is divided into 19 organized boroughs and 1 unorganized borough. Currently, the United States Census Bureau divides the unorganized borough into 11 census areas. Therefore, there are 30 census areas in all for Alaska. However, there are 40 house and senate districts (election districts) for Alaska, which means that election districts in the data for election results do not correspond with the census areas in the data provided by the United States Census Bureau. Besides, we find that election results of some years include votes data of absentee, which refers to the vote casts submitted by mail or other methods, and thus we cannot identify these votes belong to which election district of Alaska. Nevertheless, absentee voting accounts for a large portion of the whole Alaska. Considering the uniqueness of Alaska, we treat Alaska as a whole county equivalent.

2. Broomfield (8014), Colorado was incorporated in Boulder County (8013) in 1961. In the 1900s, city leaders attempted to promote Broomfield becoming a consolidated city-county. Finally, Broomfield County was established as Colorado's 64th and smallest county on November 15, 2001. Therefore, in the data prior to 2001, we cannot find Broomfield, and we cannot know the portion occupied by Broomfield in Boulder County. Consequently, we decide to combine Broomfield and Boulder County in the data after 2001.

3. Connecticut was divided into eight counties. Since 2022, the Census Bureau began to adopt the nine planning regions as county-equivalents and apply new boundaries, names and codes for statistical data. The planning regions and former counties do not align because new planning regions may incorporate cities and towns of several former counties. Considering that the data for Connecticut after redistricting only stand for a small portion of the entire time span, we refer to the redistricting map and approximately define the following relationships among former counties and planning regions for simplicity:

   - Fairfield county (9001) = Western Connecticut Planning Region (9190) + Greater Bridgeport Planning Region (9120)

- Hartford county (9003) = 4/5 Capitol Planning Region (9110)
- Litchfield county (9005) = Northwest Hills Planning Region (9160)
- Middlesex county (9007) = Lower Connecticut River Valley Planning Region (9130)
- New Haven county (9009) = Naugatuck Valley Planning Region (9140) + South Central Connecticut Planning Region (9170)
- New London county (9011) = Southeastern Connecticut Planning Region (9180)
- Tolland county (9013) = 1/5 Capitol Planning Region (9110)
- Windham county (9015) = Northeastern Connecticut Planning Region (9150)

Access to the following link for more information:

`https://www.federalregister.gov/documents/2022/06/06/2022-12063/change-to-county-equivalents-in-the-state-of-connecticut`

4. District of Columbia (11001) is treated as a whole census area in the data given by the United States Census Bureau. However, there are 8 wards for it in general election. Thus, we merge the wards in the election data. Also, we include the values for federal ballots.

5. Because of the extremely small population, Kalawao County (15005), Hawaii does not have the same functions as other counties. Kalawao does not have elected government, and there is no election data for it. Thus, we delete Kalawao County, Hawaii in all data files.

6. Shannon County (46113), South Dakota was renamed Oglala Lakota (46102) in May, 2015. In order to unify the name of counties, we rename Shannon County of South Dakota as Oglala Lakota in all data files before 2015.

7. For some years in Loving County (48301), Texas, we find that the total votes is larger than the total population. For example, the population of Loving County in 2000 is 67 according to the Decennial Census, while the total votes of Loving County in 2000 presidential election is 156 according to official data, which is more than twice the population. Although the population could change throughout month, such a huge difference within the same year is still suspicious. Since we cannot

clarify the reason why this significant discrepancy happened, and the population of Loving County is only around 100 after 1990, we eliminate it in all data files.

8. Bedford was designated as an independent city in 1968, but continued to retain the county seat of Bedford County, Virginia. On July 1, 2013, its status returned to a town within Bedford County. Therefore, to unify all sources, we combine Bedford city (51515) with Bedford County (51019) in Virginia in all data files before 2013.

9. Clifton Forge (51560), Virginia was an independent city in the 2000 census. However, in 2001, Clifton Forge gave up its city status and reverted to a town in Alleghany County (51005), Virginia. Therefore, to unify all sources, we combine Clifton Forge with Alleghany County in Virginia in all data files before 2001.

10. South Boston (51780), Virginia was a town at the beginning, and it became an independent city in 1960. Then, it reverted to a town in Halifax County (51083), Virginia in 1995. Thus, we combine South Boston with Halifax County in Virginia in all data files before 1995.

Finally, we obtain 3111 counties and county equivalents in total.

3 Raw data

We collect national-level data for public opinion on climate change, as well as county-level data for political, demographic, economic, and geographic characteristics.

3.1 National level public opinion

Data for national public opinion on climate change are obtained from two sources, Gallup Poll Social Series (GPSS) and Yale Program on Climate Change Communication (YPCCC).

3.1.1 Gallup

Selected climate change related survey questions:

1. Worry about environment – enworry_gw (1989 $\sim$ 1991, 1997, 1999, 2000 $\sim$ 2004, 2006 $\sim$ 2024): I'm going to read you a list of environmental problems. As I read each one, please tell me if you personally worry

about this problem a great deal, a fair amount, only a little, or not at all. First, how much do you personally worry about – [READ AND ROTATE A-J]? E. The "greenhouse effect" or global warming / Global warming / Global warming or Climate Change

2. Global warming – gw_when (1997, 2001 ∼ 2024): Which of the following statements reflects your view of when the effects of global warming will begin to happen – [ROTATED: they have already begun to happen, they will start happening within a few years, they will start happening within your lifetime, they will not happen within your lifetime, but they will affect future generations, (or) they will never happen]?

3. Global warming – gw_cause (2001, 2003, 2006 ∼ 2008, 2010 ∼ 2024): And from what you have heard or read, do you believe increases in the Earth's temperature over the last century are due more to – [ROTATED: the effects of pollution from human activities (or) natural changes in the environment that are not due to human activities]?

4. Global warming – gw_threat (1997, 2001, 2002, 2006, 2008 ∼ 2010, 2012 ∼ 2024): Do you think that global warming will pose a serious threat to you or your way of life in your lifetime? (Yes, No)

3.1.2 YPCCC (2008 to 2023)

Selected climate change related survey questions:

1. happening: The world's climate may change as a result of global warming. Do you think that global warming is happening? (1. No 2. Don't know 3. Yes)

2. cause_recoded: Assuming global warming is happening, do you think it is . . . (1. Don't know 2. Other 3. Neither because global warming isn't happening 4. Caused mostly by natural changes in the environment 5. Caused by human activities and natural changes 6. Caused mostly by human activities)

3. worry: How worried are you about global warming? (1. Not at all worried 2. Not very worried 3. Somewhat worried 4. Very worried)

4. harm_personally: How much do you think global warming will harm: You personally (0. Don't know 1. Not at all 2. Only a little 3. A moderate amount 4. A great deal)

Figure 1 to Figure 4 show the trends of aforementioned survey questions conducted by GPSS and YPCCC. We are interested in trend rather than absolute values. According to these figures, regardless of which survey question, the results from Gallup and YPCCC exhibit consistent trends and synchronous changes in general.

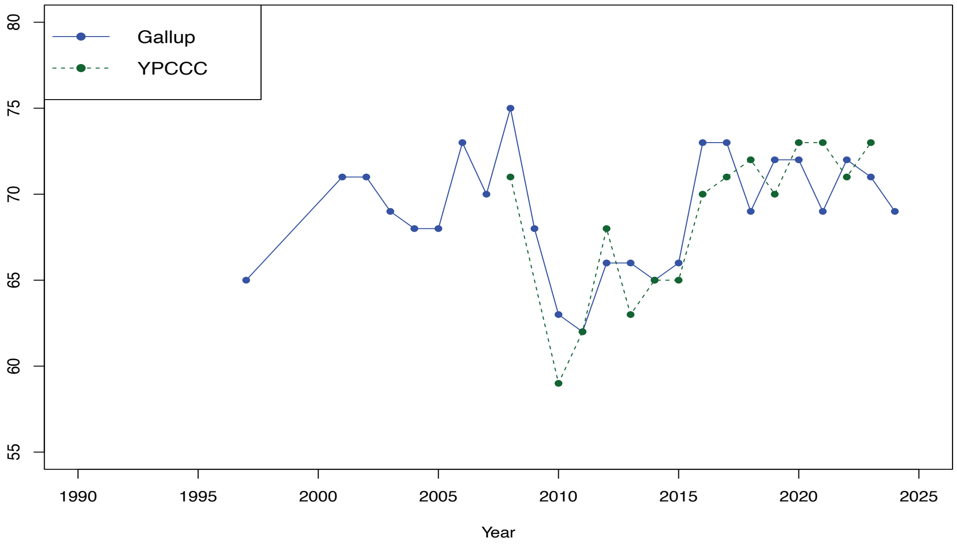

Figure 1: Percentage of respondents who think that global warming is happening

Data for public opinion on climate change can be obtained though the following links:

`https://www.gallup.com/home.aspx`
`https://osf.io/jw79p/`

3.2 County characteristics

All the county -level raw data that we need and their available corresponding years are summarized in Table 1.

3.2.1 General presidential election votes

In the United States, several elections are conducted for government officials at the federal, state and local levels. For example, the presidential election is held every four years, which determines the president and the vice president

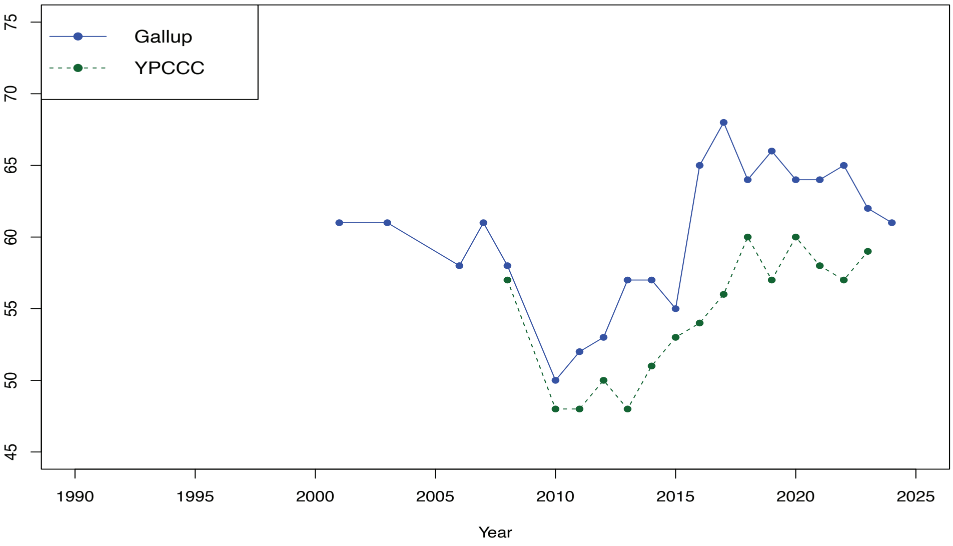

Figure 2: Percentage of respondents who think that global warming is caused mostly by human activities

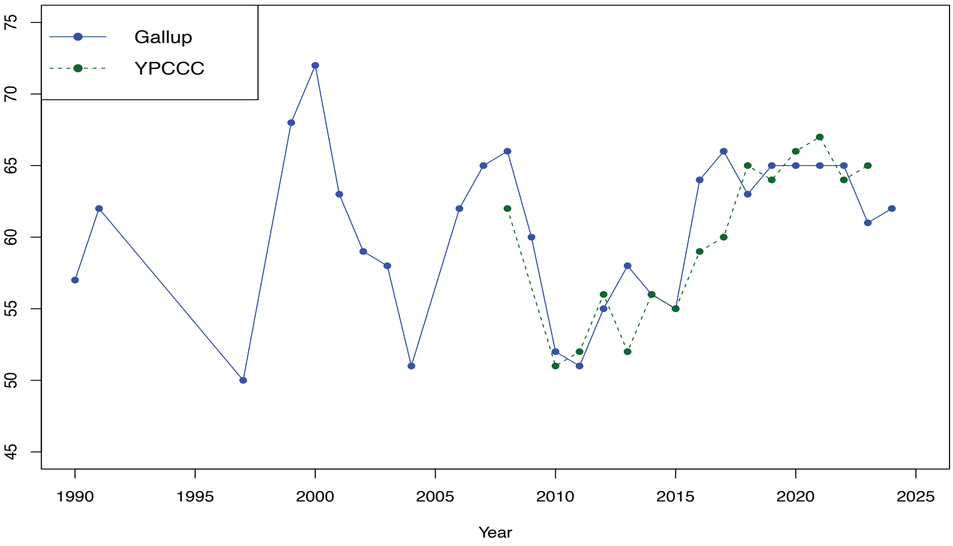

Figure 3: Percentage of respondents who are somewhat/very worried about global warming

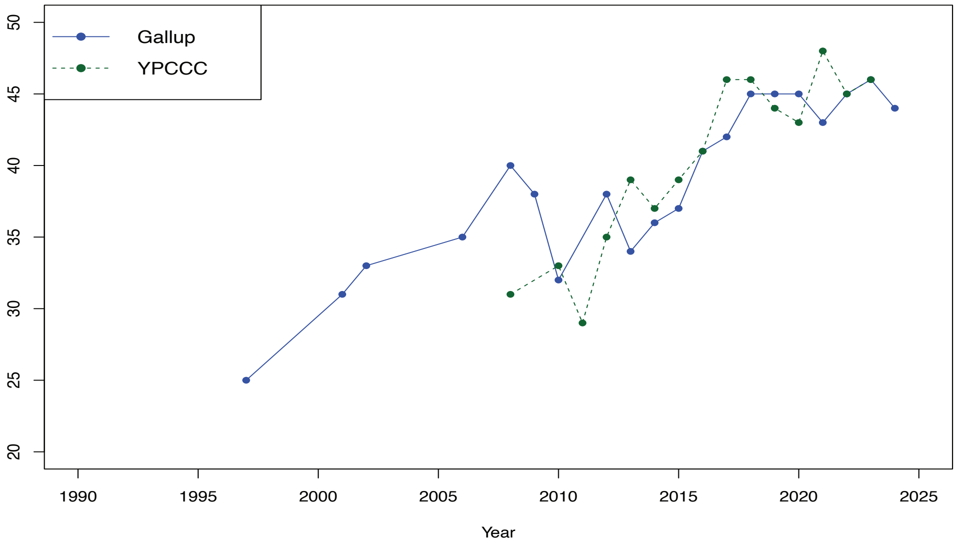

Figure 4: Percentage of respondents who think that global warming will harm them personally (in lifetime)

Table 1: County-level raw data

Item	Year
Political	
General presidential election votes	1990 to 2024, at four-year intervals
Demographic	
Total population	1990 to 2024, annually
Population by age group	1990 to 2024, annually
Population by race group	1990 to 2024, annually
25+ population by education group	1990 to 2005, 2010 to 2023, annually
Economic	
Median household income	1990, 1999, 2010 to 2023, annually
Median house price	1990, 2000, 2010 to 2023, annually
Labor force	1990 to 2024, annually
Geographic	
Land area	almost constant
Contiguity	constant

of the United States. On the other hand, the House of Representatives election and the Senate election take place every two years. In addition, there is gubernatorial election at state level and election for county government positions at local level. Here, we collect the data for the most representative and the most influential presidential election results since the presidential election is conducted simultaneously across the whole country in a unified election system, helping us to compare the political color of different counties more clearly during the same period.

Therefore, our goal is to get the presidential election data for 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020 and 2024. Considering the availability, the credit and whether the data could be obtained for free, we finally decide to use data given by Congressional Quarterly for state and county-level presidential election results from 1992 to 2020 once every four years.[1] However, Congressional Quarterly has not yet released the data for 2024 general election at this moment. So, we use the data provided by Github repository for the 2024 county-level presidential election results. These data are scraped from Fox News. We also compare the data from Congressional Quarterly and Github with some other sources, mainly Dave Leip's Atlas of U.S. Presidential Elections and Inter-university Consortium for Political and Social Research (ICPSR), to guarantee the accuracy and consistency of the data. Although Dave Leip's Atlas is widely used for presidential election related study, the spreadsheet for county-level election results need to be paid. While ICPSR is a very large social science data archive for free, the election data provided by ICPSR has the defect of a limited time span. For these reasons, we only use Dave Leip's Atlas and ICPSR as references.

County-level data for presidential election results given by Congressional Quarterly and Github repository can be obtained through the following links:

```
https://library.cqpress.com/elections/index.php
https://github.com/tonmcg/US_County_Level_Election_Results_
08-24/blob/master/2024_US_County_Level_Presidential_Results.csv
```

3.2.2 Total population by age and race group

Population estimates for counties by age and race come from the Population Estimates Program (PEP) of the United States Census Bureau. It provides intercensal population estimates for counties throughout each year. All the data are estimates based on the most recent decennial census. However, the data from 1990 to 1999 are provided in .txt format, and there are a small number of cases where, the population data for certain age and race groups

[1] By applying for a free trial, we can freely obtain the data.

in specific counties and years are corrupted. Considering such cases are very limited, we set these corrupted values to zero for simplicity.

Population data by age and race for each county can be obtained through the following links:

`https://www.census.gov/programs-surveys/popest/data/tables.html`

3.2.3 Educational attainment

Education levels among the population aged 25 and over capture the local educational attainment. Since the data for the year 2000, and the years 2010 to 2023 are directly available in the Decennial Census (DEC) and the American Community Survey (ACS), we simply combine them together. Further, ACS presents 1-year estimates data and 5-year estimates data. The former is collected over a 12-month period, while the latter is collected over a 60-month period. According to the guidance for data users provided by ACS, 5-year estimates are more reliable than 1-year estimates since 5-year estimates aggregate data over five years, especially for counties with small populations. Additionally, 5-year estimates provide data for all areas. By contrast, 1-year estimates only provide data for areas with populations over 65000 since the sample size of counties with small populations cannot ensure the accuracy of estimates. In this study, we need data for all counties and precision is more important than currency. Although 5-year estimates has the potential problem of failing to capture rapid changes between two consecutive years because it collects data over 5 years, and for continuously increasing variables, the 5-year estimates are often smaller than the 1-year estimates in the same release year, we do not want to omit counties with population below 65000. Therefore, we use the 5-year estimates rather than 1-year estimates. Regarding the data for the years before 2010, we refer to Bode (2011), in which he estimates annual data on educational attainment for over 3000 U.S. counties for the time span from 1990 to 2005. We download the data in Bode (2011) and made some modifications.

The data is missing for Alaska and Hawaii in the dataset given by Bode (2011). Therefore, we assemble the 1990 data for Alaska based on 1990 CP-2-3 from Census (page 32 of PDF). Similarly, we assemble the 1990 data for Hawaii based on 1990 CP-2-13 from Census(page 170 of PDF). In addition, Bode (2011) combines many counties in Virginia, making it hard to obtain specific data for each county. Thus, we manually compile the 1990 data for each county in Virginia based on 1990 CP-2-48 from Census(page 271 of PDF). Since the data is also missing for La Paz, Arizona (page 181 of PDF), Cibola, New Mexico (page 190 of PDF) and Menominee, Wisconsin (page 250 of PDF), we do the same thing to consummate the dataset. Consequently, the educational attainment data for these counties are missing for the years from 1991 to 1999 and 2001 to 2009.

Data for educational attainment can be obtained through the following links:

`https://data.census.gov/table?t=Educational%20Attainment&g=010XX00US$0500000`
`https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15351`
`https://data.census.gov/table/DECENNIALSF32000.P037?t=Educational%20Attainment&g=010XX00US$0500000&y=2000`
`https://www.census.gov/library/publications/1993/dec/cp-2.html`

3.2.4 Household median income

Since the data for income are related to value and involve a wide range of years, we have to adjust all values to a specific year. Thus, in order to adjust for inflation, all the income related data are adjusted to 2020 dollars using the appropriate CPI-U-RS (Consumer Price Index) adjustment factor. For FIPS 32009, 35039, 48243, and 48261, there is one missing data, respectively.

Data on household median and mean income are provided by DEC and ACS, which can be accessed through the following links:

`https://data.census.gov/table?t=Income%20(Households,%20Families,%20Individuals)&g=010XX00US$0500000`
`https://data.census.gov/table/DECENNIALSF32000.HCT012?t=Income%20(Households,%20Families,%20Individuals)&g=010XX00US$0500000&y=2000&d=DEC%20Summary%20File%203`
`https://www.census.gov/data/tables/time-series/dec/cph-series/cph-l/cph-l-123.html`

Note that for the areas that are combined with others, we calculated the mean value of their median income for simplicity. In addition, the data on household median income are available for 1990, 1999, and 2010 to 2023, while the data on household mean income are only available for years after 2010.

3.2.5 Median house price

Unlike income, house price is usually not adjusted for inflation, as house is not only a consumption good but also an asset. House price is typically influenced by both inflation and other factors such as demand and interest rate. Therefore, we just use nominal house price value to reflect the current characteristic of housing market.

Data on household median and mean income are also provided by DEC and ACS, which can be accessed through the following links:

`https://data.census.gov/table/ACSST5Y2020.S2506?t=Housing%20Value%20and%20Purchase%20Price&g=010XX00US$0500000&moe=false`
`https://data.census.gov/table/DECENNIALSF32000.H076?t=Housing%20Value%20and%20Purchase%20Price&g=010XX00US$0500000&y=2000&d=DEC%20Summary%20File%203`
`https://www.census.gov/library/publications/1992/dec/ch-1.html`

Note that we also calculated the mean value of median house price for those areas combined with others. The missing values are shown as follows:

- One missing data: 13125, 13141, 28055, 28063, 28157, 31009, 31115, 31117, 35003, 35011, 40025, 46102, 48023, 48105, 48283, 48327, 49009
- Two missing data: 13239, 30011, 48127
- Three missing data: 30079, 48033, 48137, 48243, 48311
- Four missing data: 30069, 48263
- Eight missing data: 46017
- Ten missing data: 48269
- Fourteen missing data: 48261

In the case of lacking over ten data points, interpolation results can be very unreasonable. Thus, for FIPS 48261 and 48269, we compute their median house price by taking the average of the median house prices of their respective neighboring counties. The adjacency information is as follows:

- 48261: 48047, 48215, 48273, 48489
- 48269: 48101, 48125, 48155, 48207, 48263, 48275, 48345, 48433

3.2.6 Labor force

Data for unemployment rate by county from 1990 to present are all available, provided by U.S. BUREAU OF LABOR STATISTICS, covering the number of employed and unemployed individuals:

`https://www.bls.gov/lau/tables.htm#cntyaa`

3.2.7 Land area

Land area is measured by square miles, which can be accessed through the following links:

`https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html`
`https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2020.html#list-tab-264479560`

3.2.8 Contiguity

We assume that communities (counties) may influence each other. Thus, we assemble geographic information to measure the extent that a county is influenced by other counties. Since the geography characteristic seldom changes over time, we simply use the latest 2023 county adjacency data:

`https://www.census.gov/geographies/reference-files/time-series/geo/county-adjacency.html`

It specifies adjacent counties for each county.

3.3 Power balance

The national power balance reflects the party control of president, house, and senate. The local power balance is determined by the local party share.

3.3.1 National power balance

Historical information on seat allocations of the House of Representatives and the Senate can be obtained through the following links:

`https://www.senate.gov/history/partydiv.htm`
`https://history.house.gov/Institution/Party-Divisions/Party-Divisions/`

3.3.2 Local power balance

Although we can collect the county-level general presidential election results, they are released every four years, but the interval is too long. Therefore, we aim to approximate two-year interval data for local power balance. The best we can do is to use the House election data at two-year intervals to estimate missing values. Specifically, we assume that the trend of presidential elections in a county is consistent with the trend of House elections in the corresponding state. And thus the same trend holds for all counties within a same state. By using the trend of whole state to estimate the midpoint between two presidential elections results for each county, we balance the differences across counties while incorporating the overall statewide common trend.

Data the House of Representatives elections are given by Congressional Quarterly (1988 to 2022) and CNN (2024):

`https://library.cqpress.com/elections/index.php`
`https://edition.cnn.com/election/2024`

We first aggregate district-level House election data to state level because congressional districts do not span states. Then we compare the aggregated state-level data with the raw state-level data for verification. The regions found to be problematic in cross check is shown as follows:

- 2022: Maine District 2
  Maine's ranked-choice voting system requires winning candidates to receive a majority vote. Candidates with the fewest ranked votes are eliminated, and their votes are transferred to the candidates ranked second on each voterís ballot. This process continues until a candidate receives a majority. Balloting in District 2 resulted in Round 2 counting, and is presented here as a runoff. We retained the first round result.

- 2020: Louisiana District 5
  Luke Letlow died December 29, 2020, and so did not take is seat in the 117th Congress. A special election was scheduled for March 20, 2021, to fill his seat. We used the first election result before Luke Letlow died.

- 2018: Maine District 2
  Maineís new ranked-choice voting system requires winning candidates to receive a majority vote. Candidates with the fewest ranked votes are eliminated, and their votes are transferred to the candidates ranked second on each voterís ballot. This process continues until a candidate receives a majority. Balloting in District 2 resulted in Round 2 counting, and is presented here as a runoff. We retained the first round result.

- 2016: Louisiana District 4
  There is a runoff. But we retained the result for jungle primary, because at the stage of jungle primary, there are several candidates from

the Republican Party and one candidate from the Democratic Party, wheras at the runoff, there is only one candidate from the two parties, respectively. The constituency have more choices at the first stage, which can better reflect their preferences for parties.

The data for state Louisiana in 2016 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 2014: Louisiana District 5 and 6
  As the same reason in 2016, we restained the result for jungle primary.

  The data for state Louisiana in 2014 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 2012: Louisiana District 3
  As the same reason in 2016, we restained the result for jungle primary.

  The data for state Louisiana in 2012 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 2010: Michigan District 2 $\sim$ 15
  In the district-level CQ House election data, District 2 $\sim$ 15 in Michigan is missing, we manually fill up it.

  Thus, the data for state Michigan in 2010 is also wrong in CQ statewide data. We recalculated it.

- 2006: Louisiana District 2, Texas District 23
  There is a runoff in District 2 of Louisiana, which is missing in the data provided by CQ House election. Therefore, we manually complemented the first round data for District 2 of Louisiana and recalculated the state-level result for Louisiana.

  There is a runoff in District 23 of Texas, and we retained the first round result.

- 2004: Louisiana District 3 and 7
  No candidate received the required majority in the first round voting held on Nov. 2, 2004. This run-off election was held between the top two finishers of the first-round vote. We restained the result for jungle primary.

The data for state Louisiana in 2004 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 2002: Louisiana District 5
  No candidate received the required majority in the first round voting held on Nov. 5, 2002. This run-off election was held between the top two finishers of the first-round vote. See Louisiana Primaries for results of the first-round of voting. We restained the result for jungle primary.

  The data for state Louisiana in 2002 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 2000: Virginia District 5
  Virgil Hamlin Goode Jr. was a Democrat before 2000, he switched to an independent in 2000, and then joined the Republican Party in 2002. Since he officially declared to left the Democratic Party before the House election, we still classified his votes as independent votes.

- 1996: Louisiana District 5 and 7
  Since no candidate received a majority of the vote in the September Election, a runoff election was held in November between the top two finishers. We restained the result for jungle primary.

  The data for state Louisiana in 1996 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- 1996: Missouri District 8
  Bill Emerson died in 1996 after he won the Republican primary. However, Missouri did not allow a replacement for the general election. As an alternative, Bill Emerson's wife Jo Ann Emerson participated in the election as an independent candidate, while she represented the Republican Party actually. Thus, we classified her votes in 1996 as republican votes.

- 1988: Louisiana
  The data for state Louisiana in 1988 is wrong in CQ statewide data. We calculated it by ourselves through the district-level data to revise it.

- District of Columbia
  Collected from Board of Elections (District of columbia) and Wikipedia.

- Vermont
  Bernie Sanders from Vermont always participated in Congressional elections as an independent candidate, but his relationship with the Democratic Party is very close, and he even competed in the Democratic Party presidential primaries in 2016 and 2020. Considering his great influence and the large number of votes he won in the House elections, we classified his votes as democratic votes in the years (1988 to 2004) he ran for the House elections.

After cleaning the data, we found cases where a party did not participate in the general House elections in any district within the state:

- District of Columbia: 2002, 2006, 2008, 2012, 2016, 2020.
- North Dakota: 2022.
- South Dakota: 2020, 2022.
- Vermont: 2008, 2016.

Even if one of the Republican Party and the Democratic Party does not have a candidate in a particular election year, it does not mean that there is no support for that party in that state, which means it is unreasonable to say the support rate for the absent party is equal to 0. We propose to deal with these cases as follows. In a particular year, if the Republican Party or the Democratic Party did not field candidates in all district House elections within a state, the support rate of the absent party for that year is treated as a missing value. We perform an OLS regression using data from all years in which both parties participated in the House elections to fit a linear trend. Then replace missing values with the corresponding values from the fitted line. This method ensures that the interpolated values do not influence the estimated line, as they lie on it. Figure 5 and 6 show fitted lines and interpolation results for democratic votes ratio and republican votes ratio, respectively. The black solid circles represent observed values, the blue lines represent fitted lines based on OLS regression, and the green hollow circles represent estimated votes ratios for years in which one party did not have a candidate.

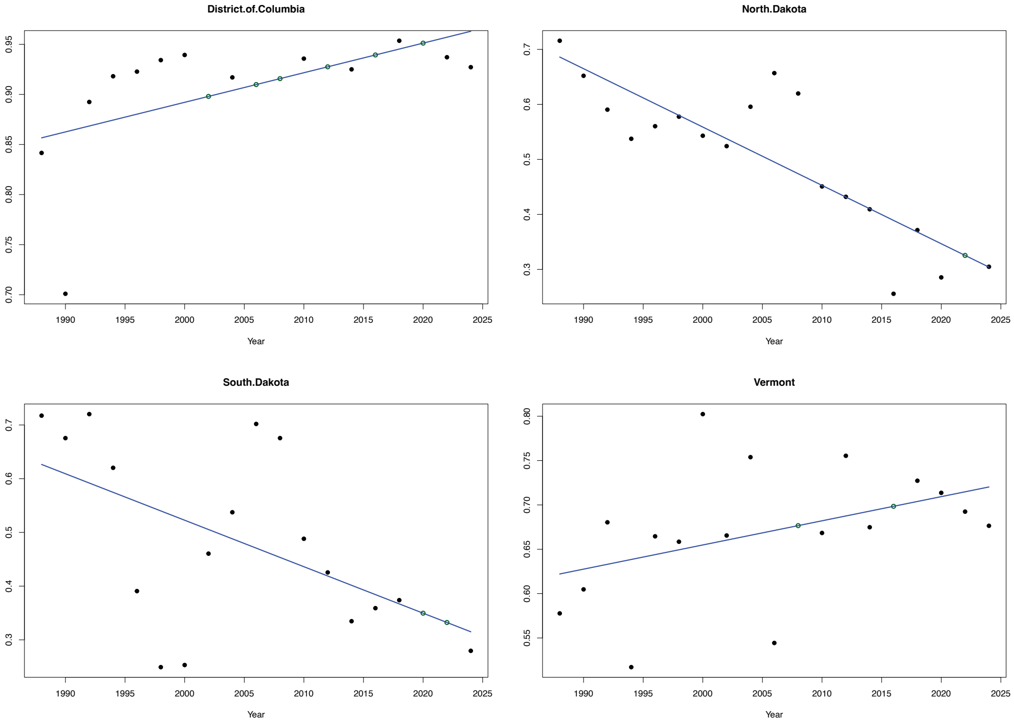

Figure 5: OLS and interpolation results for democratic votes ratio

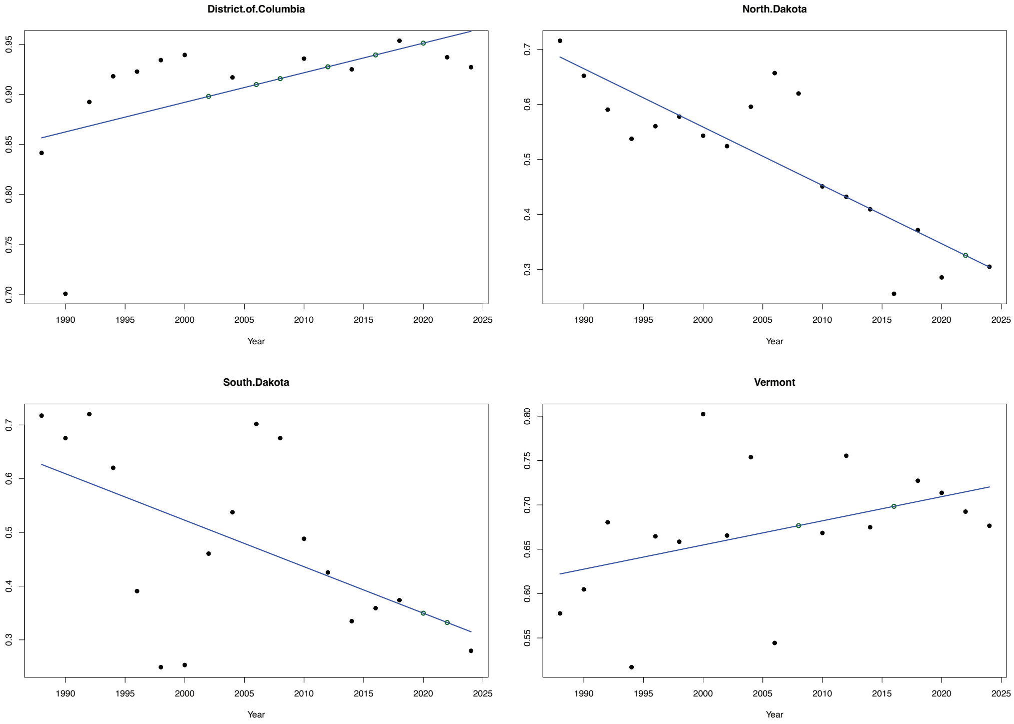

Figure 6: OLS and interpolation results for republican votes ratio

3.4 Summary statistics

Using the aforementioned raw data, we further obtain the following county-level variables.

3.4.1 Voter turnout

Voter turnout indicates the participation rate of an election. There are a variety of ways to measure turnout. Here, we simply divide total number of valid votes by total population to evaluate the universal suffrage of each county.

Table 2: Voter turnout (%)

	1992	1996	2000	2004	2008	2012	2016	2020	2024
Min.	7.30	6.71	8.02	10.89	14.22	12.15	13.83	16.69	19.16
1st Qu.	38.34	33.95	35.05	39.28	39.99	38.07	39.62	44.29	42.76
Median	43.40	38.84	39.73	44.09	44.89	43.35	44.66	49.89	48.29
Mean	43.41	38.95	39.85	44.26	44.80	43.45	44.53	49.68	48.36
3rd Qu.	48.49	43.55	44.75	49.54	50.02	48.89	49.82	55.02	54.20
Max.	94.63	80.80	77.58	82.35	90.72	90.40	86.24	95.85	86.19
Std. Dev.	7.68	7.24	7.24	7.66	7.63	8.16	7.89	8.31	8.53

3.4.2 Party share

We exclude votes for third parties, as their votes are typically negligible, and our analysis mainly focuses on the balance between Democratic and Republican parties. Therefore, party share is defined as the percentage of votes received by Democratic (or Republican) party relative to the total number of votes for the two major parties.

Table 3: Democratic share

	1992	1996	2000	2004	2008	2012	2016	2020	2024
Min.	0.11	0.12	0.07	0.07	0.05	0.04	0.03	0.03	0.04
1st Qu.	0.42	0.42	0.33	0.30	0.32	0.28	0.21	0.21	0.20
Median	0.50	0.50	0.41	0.39	0.42	0.38	0.30	0.31	0.29
Mean	0.50	0.50	0.41	0.39	0.42	0.39	0.33	0.34	0.32
3rd Qu.	0.57	0.57	0.49	0.47	0.51	0.48	0.42	0.43	0.41
Max.	0.90	0.90	0.91	0.91	0.93	0.94	0.96	0.95	0.93
Std. Dev.	0.11	0.12	0.12	0.13	0.14	0.15	0.16	0.16	0.16

The increase in standard deviation indicates that the differences in support for two parties (Republican or Democratic) among the counties are getting larger, revealing the phenomenon of political polarization.

3.4.3 Population density

Population density is measured by population per square miles. Specifically, we divide the population data from 1990 to 1999 by the 1990 land area data, from 2000 to 2009 by the 2000 land area data, from 2010 to 2019 by the 2010 land area data, and from 2020 to present by the 2020 land area data.

Table 4: Population density (per square miles)

	1990	2000	2010	2020
Min.	0.31	0.27	0.22	0.21
1st Qu.	16.32	17.49	17.67	17.02
Median	38.96	43.23	45.69	45.13
Mean	221.29	244.91	261.59	279.21
3rd Qu.	94.36	105.56	114.98	119.25
Max.	52407.33	67096.99	69591.20	74121.89
Std. Dev.	1437.44	1677.70	1735.93	1861.52

Apparently, population density has been increasing in general. However, the increasing standard deviation demonstrates that the disparity of population density among counties is becoming more and more large. The concentration of population in large counties and the sparsity of population in small counties may be an important reason.

3.4.4 Household median income

Income is adjusted to 2020 dollars to eliminate the effect of inflation, using the appropriate CPI-U-RS (Consumer Price Index) adjustment factor. The data for CPI can be obtained through:

`https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm`

The real income level increases continuously after 1990, but declines around 2010, reflecting the impact of Lehman shock in 2008. It begins to increase again after 2010. However, the increasing standard deviation implies an expansion of income inequality.

Table 5: Household median income (2020 dollars)

	1990	1999	2010	2020
Min.	16564.00	20305.00	23042.00	22292.00
1st Qu.	37884.50	46003.50	43953.00	45579.00
Median	43711.00	52398.00	50473.00	52713.00
Mean	46007.70	54867.27	52529.23	54885.28
3rd Qu.	51721.00	61046.00	58255.50	61333.25
Max.	114248.00	128770.00	137619.00	147111.00
Std. Dev.	12483.60	13740.04	13603.12	14594.57

3.4.5 Median house price

House prices are not adjusted for inflation, because a house is not only a consumption good but also an asset. House prices are typically influenced by both inflation and other factors such as demand and the interest rate.

Table 6: Median house price (dollars)

	1990	2000	2010	2020
Min.	10400.00	20100.00	35600.00	40600.00
1st Qu.	36800.00	57800.00	87400.00	117200.00
Median	45300.00	75700.00	114050.00	149800.00
Mean	53811.36	84385.59	140210.14	173603.18
3rd Qu.	59950.00	96500.00	159500.00	196300.00
Max.	487300.00	1000000.00	882500.00	1189100.00
Std. Dev.	33503.59	47675.59	87425.45	98155.54

It is obvious that house prices have been increasing since 1990, and also the differences across counties have been widening.

3.4.6 Unemployment rate

Divide labor force without work by total labor force, we obtain the unemployment rate. A relatively high unemployment rate around 2010 reflects the negative effect of Lehman shock.

Table 7: Unemployment rate (%)

	1990	2000	2010	2020
Min.	0.50	1.30	2.00	1.60
1st Qu.	4.20	3.20	7.20	5.20
Median	5.70	4.10	9.20	6.60
Mean	6.12	4.34	9.33	6.73
3rd Qu.	7.60	5.10	11.30	8.00
Max.	40.60	17.30	29.40	22.60
Std. Dev.	2.91	1.66	3.14	2.27

3.4.7 Demographic diversity

We follow Wang et al. (2021) and define the demographic diversities for race, education, and age as

$$s_{gk}(\tau) = 1 - \sum_i \left(\frac{g_{ik}(\tau)}{pop_k(\tau)} \right)^2 \qquad (g = r, e, a),$$

where r, e, and a represent race, education, and age, respectively, i denotes the group within variable g, and pop denotes total population. This is essentially the Simpson index (Simpson, 1949), which measures the level of diversity when individuals are grouped into categories.

Table 8: Race diversity

	1990	2000	2010	2020
Min.	0.00	0.01	0.01	0.02
1st Qu.	0.02	0.04	0.06	0.09
Median	0.07	0.10	0.13	0.17
Mean	0.15	0.18	0.20	0.22
3rd Qu.	0.25	0.28	0.31	0.34
Max.	0.66	0.73	0.73	0.73
Std. Dev.	0.16	0.16	0.16	0.16

Race diversity shows an increasing trend overall, whereas education diversity shows a decreasing trend. In addition, age diversity is quite stable.

3.4.8 Education level

Education level here refers to the percentage of the population aged over 25 holding a bachelor degree or higher.

Obviously, education level has been increasing drastically.

Table 9: Education diversity

	1990	2000	2010	2020
Min.	0.40	0.32	0.30	0.29
1st Qu.	0.53	0.51	0.47	0.46
Median	0.56	0.54	0.51	0.49
Mean	0.55	0.54	0.51	0.49
3rd Qu.	0.58	0.57	0.54	0.53
Max.	0.66	0.69	0.64	0.65
Std. Dev.	0.03	0.05	0.05	0.05

Table 10: Age diversity

	1990	2000	2010	2020
Min.	0.85	0.86	0.85	0.84
1st Qu.	0.93	0.93	0.94	0.94
Median	0.94	0.94	0.94	0.94
Mean	0.94	0.94	0.94	0.94
3rd Qu.	0.94	0.94	0.94	0.94
Max.	0.94	0.94	0.94	0.94
Std. Dev.	0.01	0.01	0.01	0.00

Table 11: Education Level (%)

	1990	2000	2010	2020
Min.	3.69	3.95	3.61	1.48
1st Qu.	9.16	10.34	13.10	15.88
Median	11.76	13.40	16.82	20.20
Mean	13.49	15.21	19.01	22.61
3rd Qu.	15.61	17.64	22.57	26.73
Max.	53.42	57.11	70.91	79.14
Std. Dev.	6.58	7.15	8.66	9.69

3.4.9 Political diversity and political color

Since the election results in even years determine local power balance in the next two years, the data for political diversity and political color are one year behind the election.

Political diversity lies between 0 and 1, and it is defined as

$$s_{pk}(\tau) = 1 - 2|d_k(\tau) - 0.5|,$$

where $d_k(\tau)$ denotes the democratic votes ratio at time τ.

Table 12: Political diversity

	1991	2001	2011	2021
Min.	0.20	0.14	0.08	0.06
1st Qu.	0.74	0.66	0.50	0.42
Median	0.86	0.78	0.66	0.58
Mean	0.82	0.76	0.65	0.60
3rd Qu.	0.92	0.90	0.82	0.78
Max.	1.00	1.00	1.00	1.00
Std. Dev.	0.13	0.18	0.20	0.22

Local political color lies between -1 and 1. It is determined by both local and national power balance.

Table 13: Political color

	1991	2001	2011	2021
Min.	-0.74	-0.77	-0.84	-0.84
1st Qu.	-0.02	0.05	0.08	0.09
Median	0.12	0.20	0.25	0.28
Mean	0.11	0.19	0.23	0.24
3rd Qu.	0.25	0.35	0.40	0.42
Max.	0.70	0.82	0.81	0.82
Std. Dev.	0.21	0.23	0.25	0.27

The increase in median and mean value of political color indicates that the U.S. is turning to the right gradually.

4 Balanced panel

The next step is to construct a balanced panel. Since these data are provided at different frequencies, the low frequency data can be regarded as information with missing data.[2]

There are various interpolation methods that deal with missing data problem, such as linear interpolation, spline interpolation, Kalman filter, etc. Among them, the Kalman filter experts in handling dynamic time series data with noise and real-time updating by providing an optimal recursive

[2] Note that we do not interpolate data for political diversity and political color, because we want to emphasize the discontinuity of power balance, which means that they keep constant over each two-year period.

solution. The recursive solution uses all the data from the time series to update estimates continuously. Thus, the Kalman filter is capable of predicting the state of a dynamic model at the past, present and future. It is also able to capture characteristics and trends in time series data, as well as reducing the impacts of noise on interpolation results. Therefore, we chose Kalman filter as our tool to interpolate missing data.

4.1 Kalman filter

In the context of state-space model, recursive algorithm is used for state estimation, implemented through the Kalman filter. In brief, the algorithm involves three main steps. First, predict the new state based on the previous estimation. Second, bring in the latest observation data and calculate a correcting term, called the Kalman gain, that minimizes the prediction error. The Kalman gain reflects the lean of estimation, towards the predicted value or the observed value. Third, use the correcting term to update the predicted state. The Kalman filter process relies on the assumption of normal distribution for the underlying disturbance of system itself, the disturbance of observation, and the initial state. The variances of the three distributions are referred to as hyperparameters. Then, the likelihood function of model can be derived based on these initial hyperparameters, which is used to replace the hyperparameters by their maximum likelihood estimates (MLE). After the process of Kalman filter, state smoothing can be further implemented to obtain more accurate estimates of each state. The Kalman filter is a recursive algorithm that only relies on past and present data. On the other hand, state smoothing takes into account observations of entire time series, including future data. In other words, the Kalman filter first updates the state forward, and then state smoothing reevaluates the state backward.

In the next several subsections, we compare three state-space models and choose one model to implement Kalman filter.

4.2 Local level model

This subsection introduces a simple state-space model consists of observation equation and state equation. No seasonal or change rate of trend is present. y_t is a time series observed data, α_t refers to the unobserved value (i.e. real state) at time t. Observation error ϵ_t and state error η_t follow normal distribution with zero means and variances σ_ϵ^2 and σ_η^2, respectively. This gives the model:

$$
\begin{aligned}
y_t &= \alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2), \\
\alpha_{t+1} &= \alpha_t + \eta_t, \qquad \eta_t \sim N(0, \sigma_\eta^2).
\end{aligned}
$$

The updated state estimation at time t is given by:

$$
\begin{aligned}
a_{t|t} &= a_t + K_t(y_t - a_t), \\
P_{t|t} &= P_t(1 - K_t), \\
K_t &= \frac{P_t}{P_t + \sigma_\epsilon^2},
\end{aligned}
$$

where a_t and P_t denote the prior conditional distribution of α_t given vector $Y_{t-1} = (y_1, ..., y_{t-1})'$. $a_{t|t}$ and $P_{t|t}$ denote the posterior conditional distribution of α_t when the latest observation value y_t is taken in. K_t denotes the Kalman gain, which is equal to the ratio of prior error variance to prediction error variance. Note that when $\sigma_\epsilon^2 \to +\infty$, $K_t \to 0$, then $a_{t|t} \to a_t$, which means if the observation error is extremely large, the Kalman filter tends to believe the estimated value. Contrarily, when $\sigma_\epsilon^2 \to 0$, $K_t \to 1$, then $a_{t|t} \to y_t$, which means if the observation error is extremely small, the Kalman filter tends to believe the observed value.

Now consider the estimation of α_t not only based on past and present data, but also based on future data. Then, the distribution of α_t given $(y_1, ..., y_n)$ follows $N(\hat{\alpha}_t, V_t)$, where $\hat{\alpha}_t = E(\alpha_t|(y_1, ..., y_n))$ and $V_t = Var(\alpha_t|(y_1, ..., y_n))$. Define the state estimation error and prediction error as:

$$
\begin{aligned}
x_t &= \alpha_t - a_t, \\
v_t &= y_t - a_t = x_t + \epsilon_t,
\end{aligned}
$$

respectively. According to the multivariate regression theory, we have the conditional distribution of α_t:

$$
\begin{aligned}
\hat{\alpha}_t = E(\alpha_t|(y_1, ..., y_n)) &= a_t + \sum_{j=t}^{n} Cov(\alpha_t, v_j) F_j^{-1} v_j, \\
V_t = V(\alpha_t|(y_1, ..., y_n)) &= P_t - \sum_{j=t}^{n} [Cov(\alpha_t, v_j)]^2 F_j^{-1},
\end{aligned}
$$

where $F_t = P_t + \sigma_\epsilon^2$. Finally, the smoothed state and its variance are obtained through backwards recursion:

$$r_{t-1} = F_t^{-1} v_t + L_t r_t, \qquad \hat{\alpha}_t = a_t + P_t r_{t-1},$$
$$N_{t-1} = F_t^{-1} + L_t^2 N t, \qquad V_t = P_t - P_t^2 N t - 1,$$
$$L_t = 1 - K_t = \frac{\sigma_\epsilon^2}{F_t},$$

where r_{t-1} and N_{t-1} refer to the weighted sum of prediction errors and the weighted sum of inverse variances of prediction errors after time $t-1$, respectively. $t \in [n, 1]$, $r_n = 0$ and $N_n = 0$ since there are no observations after time n. By performing this series of calculation, the smoothed state values of each time point are able to be estimated, solving the problem of missing data.

Based on the process of Kalman filter and state smoothing provided above, we first implemented interpolation in the context of a local level model for each county, respectively. For example, presidential election data is provided at four-year intervals. Absolutely, we have the original data for election years to evaluate the political color of the public. We also want to know the information of county-level political color for non-election years.

To start up the algorithm of the Kalman filter, the conditional distribution for initial state is necessary. In R, when using dlm package, the mean value and variance of the initial state are set by default. More specifically, the dlm package sets the initial state mean value as zero, and the initial state variance as a large value since generally we have no information about the initial state. However, as the Kalman filter proceeds, the prediction converges to the real dynamic of the data, and thus the state variance also tends to stabilize. That is the reason why the exact value of initial state variance is not so restrictive.

The other important process in the Kalman filter is the update of hyperparemeter (variance of observation error σ_ϵ^2 and variance of state error σ_η^2) through MLE. Nevertheless, sometimes the estimated variance hyperparameters might lead to excessive smoothing due to the too large σ_ϵ^2 and too small σ_η^2. In cases with limited data or complex noise, this problem is likely to occur. Then manually adjusting these variances can help us to amend the interpolation results.

4.3 Local linear trend model

The local linear trend model makes it possible to account for the change rate of state, and the Kalman filter becomes more sensitive to the changes of data in early stages. To incorporate a trend component, add a slope term ν_t to local level model, then the local linear trend model can be expressed in the following form of equations:

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t, \qquad & \epsilon_t &\sim N(0, \sigma_\epsilon^2), \\ \alpha_{t+1} &= \alpha_t + \nu_t + \eta_t, \qquad & \eta_t &\sim N(0, \sigma_\eta^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, \qquad & \zeta_t &\sim N(0, \sigma_\zeta^2). \end{aligned}$$

When η_t and ζ_t are fixed to 0, we get:

$$\begin{aligned} y_t &= \alpha_t + \epsilon_t, \\ \alpha_{t+1} &= \alpha_t + \nu_t, \\ \nu_{t+1} &= \nu_t = \nu, \end{aligned}$$

where ν is a constant value. Thus, we can rewrite α_{t+1} and y_t as:

$$\begin{aligned} \alpha_{t+1} &= \alpha_0 + \nu(t+1), \\ y_t &= \alpha_0 + \nu t + \epsilon_t. \end{aligned}$$

Obviously, it is a classical regression model with a deterministic slope ν. When σ_η^2 and σ_ζ^2 are positive values, the trend level and slope of state equation is varying over time. Furthermore, the local linear trend model can also be written in the form of matrices:

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha_t \\ \nu_t \end{pmatrix} + \epsilon_t, \\ \begin{pmatrix} \alpha_{t+1} \\ \nu_{t+1} \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix}. \end{aligned}$$

Analogous to local level model, the Kalman filter and state smoothing process can also predict, update and smooth the data in local linear trend model. By contrast, since the local linear trend model captures both the level of state and rate of change, the Kalman filter becomes more sensitive to the fluctuations in data, especially for data in early stages. In consequence, the local linear trend model helps the Kalman filter and state smoothing to better estimate the state values.

4.4 Multivariate structural time series model

In the two previous subsections, we independently analyze the time series data for over 3000 counties, differing in whether the model includes a trend component. However, multivariate time series can also be analyzed simultaneously in a structural time series model. Moreover, there are two ways to incorporate common effects. One is called seemingly unrelated time series equations model (SUTSE model), the other is a model with explicit common factors.

In the case of SUTSE model, each time series seems to be modeled as in the univariate case like local level model, but it also considers common effects or interaction among each time series, which means the error terms are correlated across different series. The multivariate version of local level model that accounts for common effects is given by:

$$y_t = \begin{pmatrix} y_{1,t} \\ \vdots \\ y_{i,t} \end{pmatrix} = \begin{pmatrix} \alpha_{1,t} \\ \vdots \\ \alpha_{i,t} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \vdots \\ \epsilon_{i,t} \end{pmatrix} = \alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, \Sigma_\epsilon),$$

$$\alpha_{t+1} = \begin{pmatrix} \alpha_{1,t+1} \\ \vdots \\ \alpha_{i,t+1} \end{pmatrix} = \begin{pmatrix} \alpha_{1,t} \\ \vdots \\ \alpha_{i,t} \end{pmatrix} + \begin{pmatrix} \eta_{1,t} \\ \vdots \\ \eta_{i,t} \end{pmatrix} = \alpha_t + \eta_t, \qquad \eta_t \sim N(0, \Sigma_\eta).$$

The dimension of covariance matrices Σ_ϵ and Σ_η is $i \times i$. Assume that there is no correlation between observation errors, but there is correlation between state noise which captures the interaction between different time series. If we want to apply this model to the data with 3000 variables, we have to deal with a covariance matrix of state noise with a dimension 3000×3000, which is extremely cumbersome since it is a symmetric matrix but not a diagonal matrix. The number of parameters needed to be estimated in the covariance matrices of state noise is $\frac{3000\times(3000+1)}{2} = 4501500$.

Another way is to explicitly define the common factors to capture the common fluctuations. In this case, if we assume that there exists one common factor and each variable also holds its own independent specific factor for observation error, the number of parameters needed to be estimated in the covariance matrices of state noise decreases from 4501500 to 1 because the dimension of state noise for common factor is 1×1. That means explicitly define the common factors has the advantage of significantly simplifying the computation. A simple structural time series model with common factor is shown as follows:

$$\begin{aligned} y_t &= \mu + H\alpha_t^* + \epsilon_t, \qquad \epsilon_t \sim N(0, \Sigma_\epsilon), \\ \alpha_{t+1}^* &= \alpha_t^* + \eta_t^*, \qquad \eta_t^* \sim N(0, \Sigma_\eta^*). \end{aligned}$$

y_t is a $p\times 1$ vector that includes p variables at time t. μ is also a $p\times 1$ intercept vector, which refers to long-term average and reflects specific characteristics of each variable. α_t^* and η_t^* are $q \times 1$ vectors, representing common trends and state noise, respectively. $p \times 1$ vector ϵ_t denotes the specific observation error for each variable. Finally, H is a $p \times q$ matrix, called factor loading matrix. It quantifies the relationship between common factors and different observation variables.

We finally decide to adopt the local linear trend model for the Kalman filter and state smoothing. The local level model is very simple to structure, requiring the least computation. However, it is too simple to capture the dynamic characteristics and trends of data. By adding a trend component to the local level model, the local linear trend model enhances the sensitivity to dynamic changes in data, and the complexity of computation is still moderate. The multivariate structural time series model further involves the interactions and common trends between multivariate variables. Nevertheless, the computational cost is prohibitively high. In addition, with a relatively small sample size for each variable, it may be difficult for the multivariate time series model to capture the common factors. Since the data we collect have a relatively small sample size, but stable trends, we conclude that the local linear trend model is both efficient and appropriate for interpolation.

4.5 Monthly data

We interpolate all data to a monthly frequency since we wish to have as many observations as possible.

4.5.1 Public opinion on climate change

We use the data from Gallup since they are more abundant in terms of year and cover a larger range of time period. We conduct the Kalman filter and state smoothing based on a local linear trend model given in Subsection 4.3. Although the MLE can estimate the most probably hyperparameters σ_ϵ^2, σ_η^2 and σ_ζ^2, the results are derived from mathematical calculations. That means the hyperparameters estimated by MLE may not always be reasonable in real application, and consequently the results of the Kalman filter and state smoothing may not capture the underlying dynamic change of data well. Therefore, we experiment with several sets of σ_ϵ^2, σ_η^2 and σ_ζ^2, and finally assume that the variance of observation error $\sigma_\epsilon^2 \in [0, 25]$, the variance of state error $\sigma_\eta^2 \in [0, 0.05]$, and the variance of trend error $\sigma_\zeta^2 \in [0, 0.01]$. In specific, the observation error is not very large, so the model relatively trusts the observed data. Meanwhile, the level of public opinion is affected by a small short-term state noise, avoiding sudden great changes. Although the variance of state error is relatively small, its accumulation effect over time may lead to moderate changes. Furthermore, the long-term trend is influenced by an even smaller trend error. It makes sure that opinion varies slowly, without the existence of sharp reversals. Together, these constraints on variance ensure that the change of opinion is a smooth and gradual process.

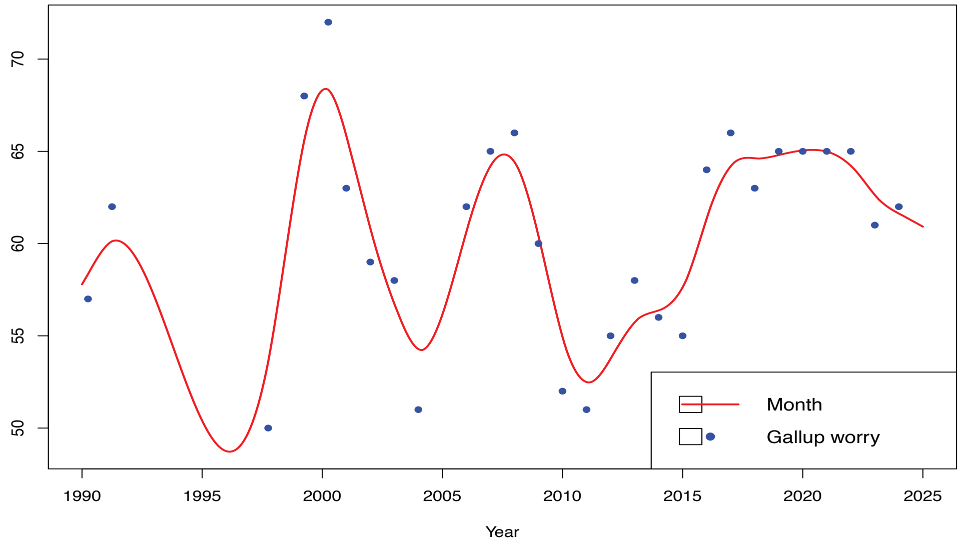

Figure 7: Percentage of respondents who are somewhat/worry worried about global warming

According to Figure 7, public opinion on climate change exhibits very large fluctuations before 2004. In specific, the large decrease from early 1990s to 1997 may be attributed to the tilt of the ruling party (Democratic) toward economic development, while the great increase between 1997 and 2000 may be attributed to the signing of the Kyoto Protocol. The ruling party (Democratic) signed this protocol, reflecting the shift of government attitude towards climate problems. Then, in 2000 the Republican Party came to the stage and withdrew from the Kyoto Protocol, leading to a significant decline in public opinion on worry about global warming / climate change. However, this trend quickly changed around 2005, which may be caused by the severe hurricane disasters during that time. Following the Lehman shock in 2008, the public became less concerned about climate change, instead they put more interest in economic issues, corresponding to the sharp decline observed around 2007. Public opinion on climate change returns to growth after 2010.

4.5.2 County characteristics

Since it is impossible to set separate variances of observation, state, and trend errors for all 3111 counties, we apply the same upper bound variances for these errors across counties for each variable, respectively. After a series of tests, the following sets of bounds lead to plausible interpolation data without outliers (see Table 14).

Table 14: Upper bound variances of observation, state, and trend error

variable	observation σ^2_ϵ	state σ^2_η	trend σ^2_ζ
voter turnout	1	0.1	0.05
population density	5000	1	0.01
median income	25000	900	100
house price	3.5×10^6	1500	160
unemployment	1	0.1	0.05
race diversity	1×10^{-7}	2×10^{-7}	7×10^{-7}
education diversity	1×10^{-7}	2×10^{-7}	7×10^{-7}
age diversity	1×10^{-7}	2×10^{-7}	7×10^{-7}
education level	1	0.1	0.05

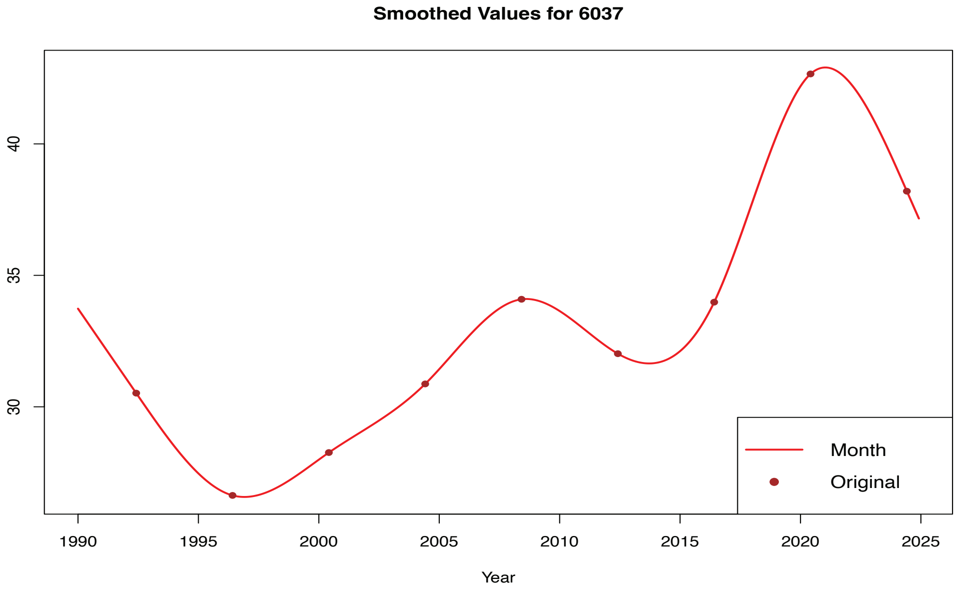

Figure 8: Interpolation for voter turnout (Los Angeles County)

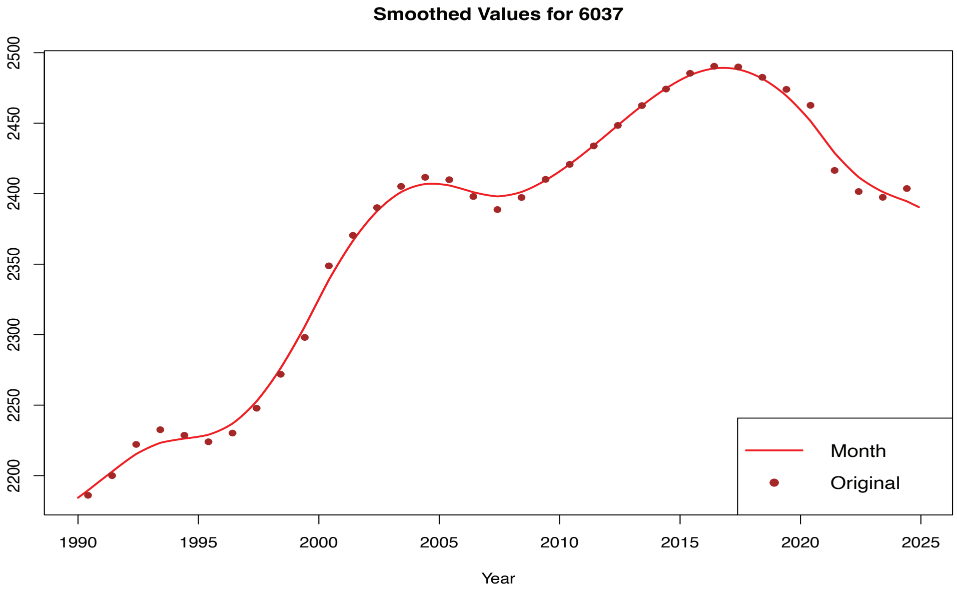

Figure 9: Interpolation for population density (Los Angeles County)

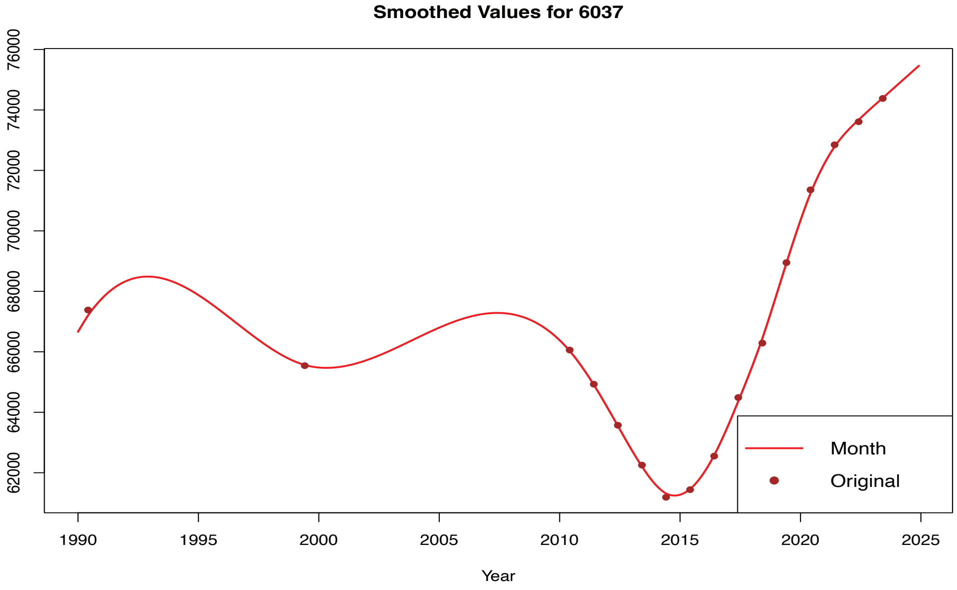

Figure 10: Interpolation for median income (Los Angeles County)

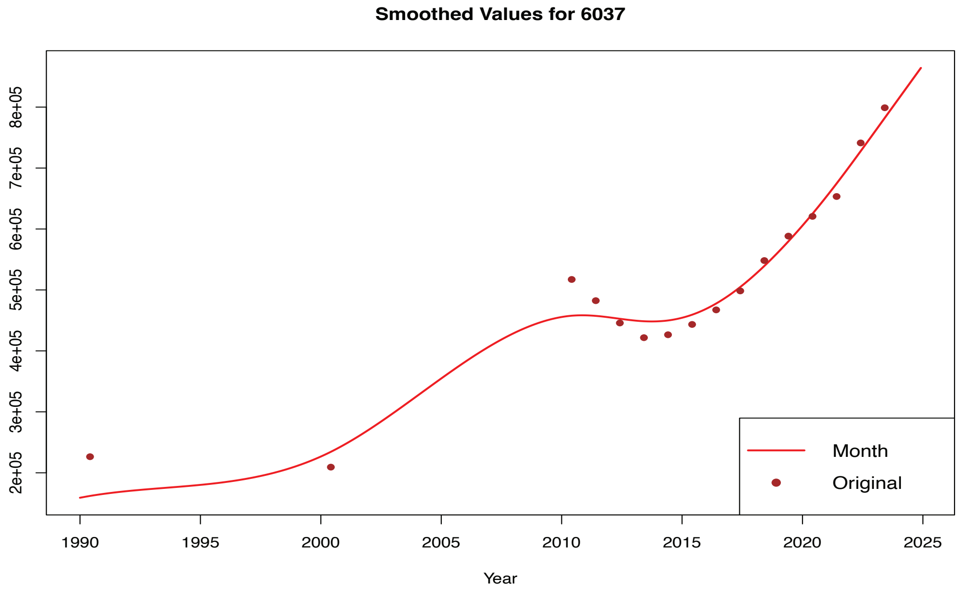

Figure 11: Interpolation for median house price (Los Angeles County)

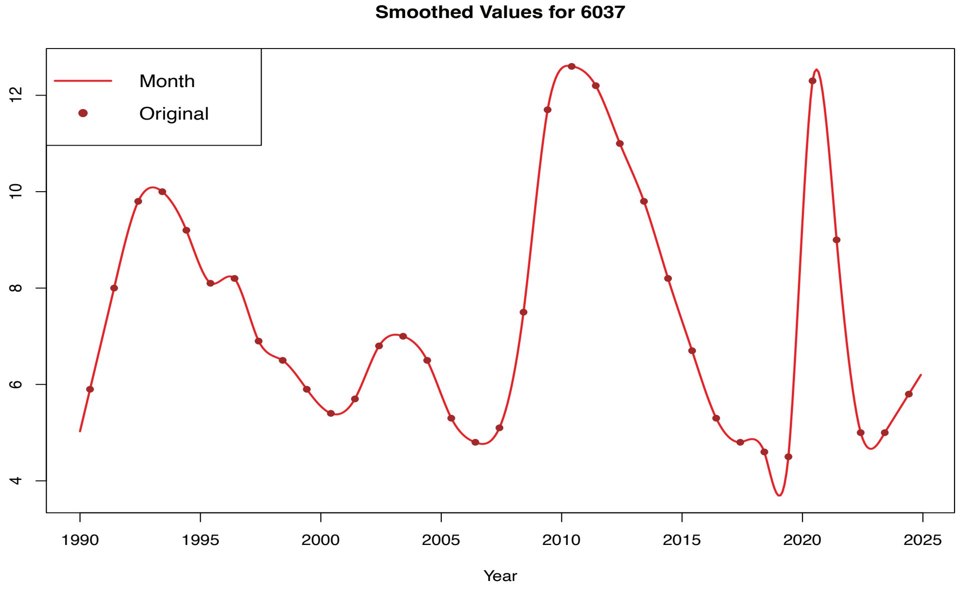

Figure 12: Interpolation for unemployment (Los Angeles County)

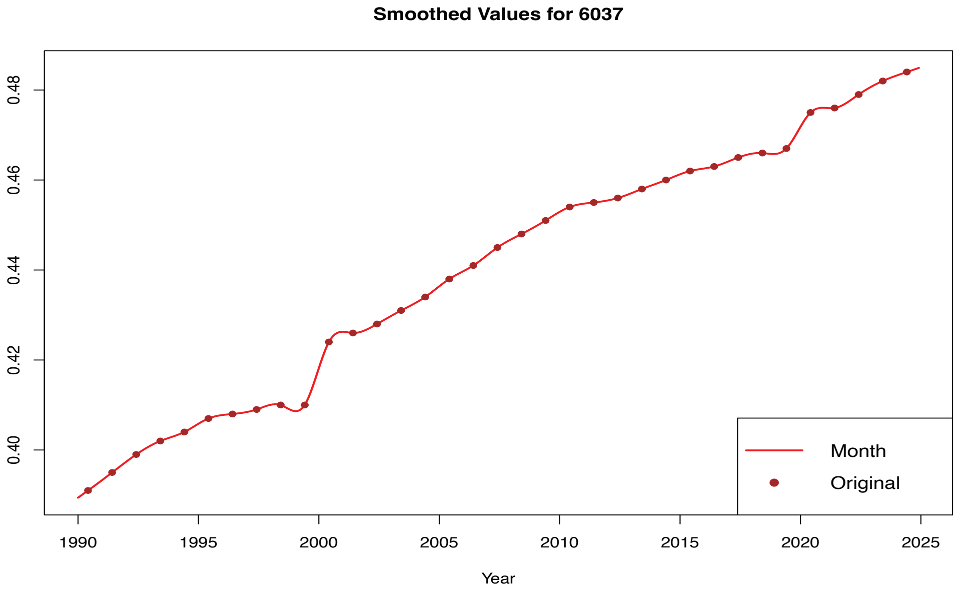

Figure 13: Interpolation for race diversity (Los Angeles County)

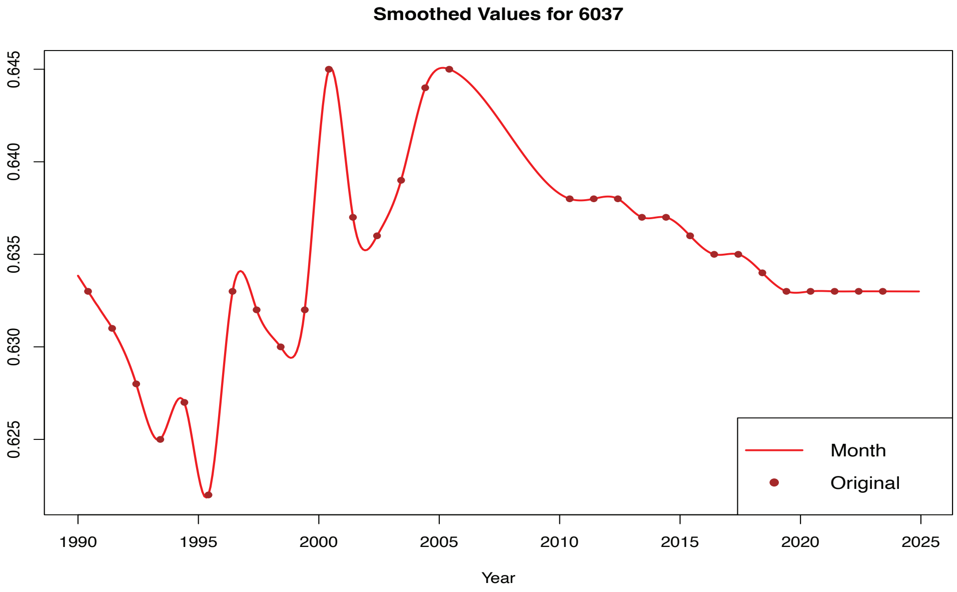

Figure 14: Interpolation for education diversity (Los Angeles County)

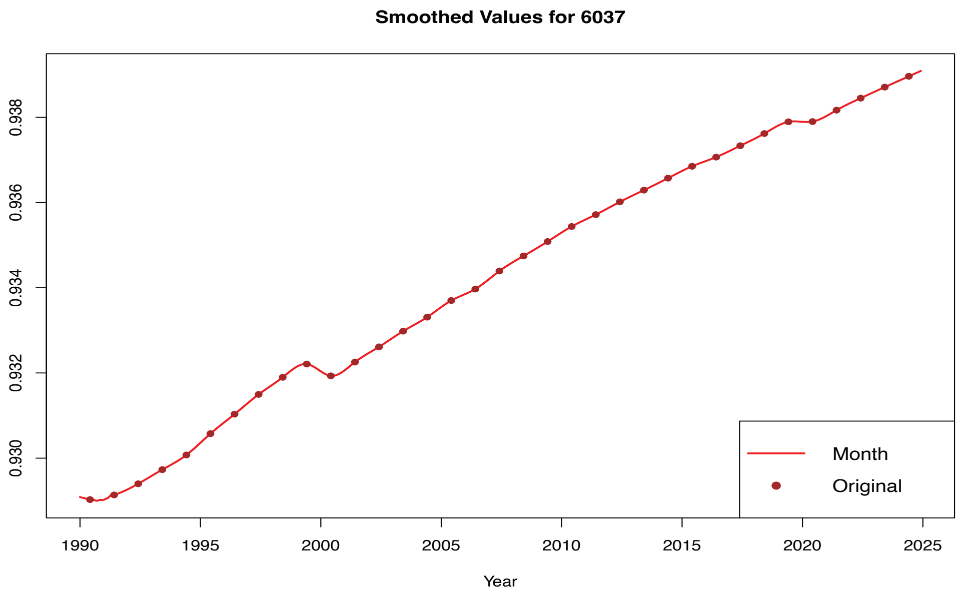

Figure 15: Interpolation for age diversity (Los Angeles County)

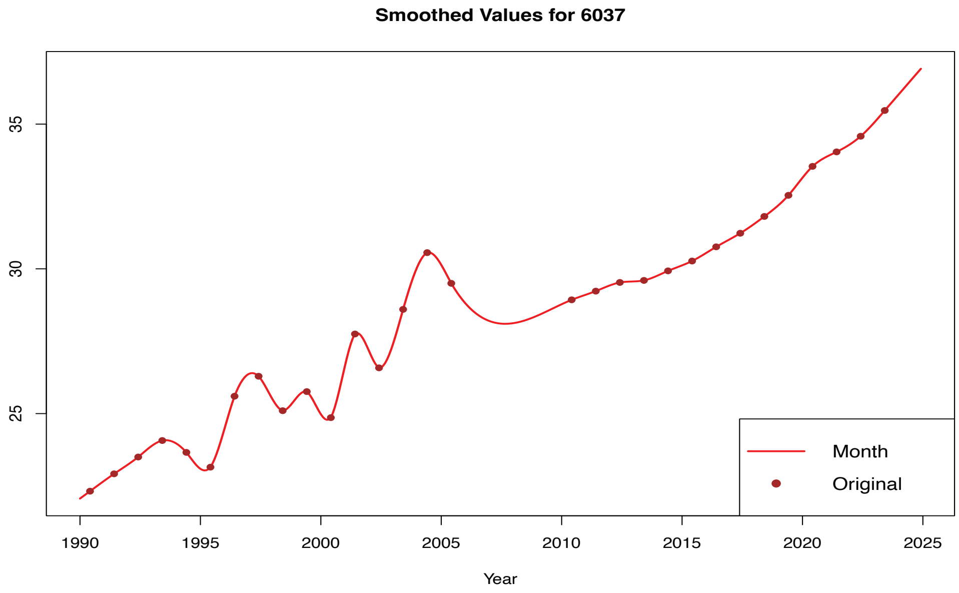

Figure 16: Interpolation for education level (Los Angeles County)

We take Los Angeles County (FIPS 6037) as an example and show the monthly interpolation results for each of the variables, as shown in Figure 8 – 16.

5 Climate disasters and news coverage of climate change

We use the data for the U.S. billion-dollar climate disasters, as well as the data for news coverage of climate change to identify severe climate issues, and then generate news information, including number of news and objective importance of each news item.

The data for U.S. climate disasters (see Figure 17) is provided by National Centers for Environmental Information (NCEI):

`https://www.ncei.noaa.gov/access/billions/state-summary/US`

And the data for the news coverage of climate change (see Figure 18) is provided by Media and Climate Change Observatory (MeCCO):

`https://mecco.colorado.edu/tv/index.html`

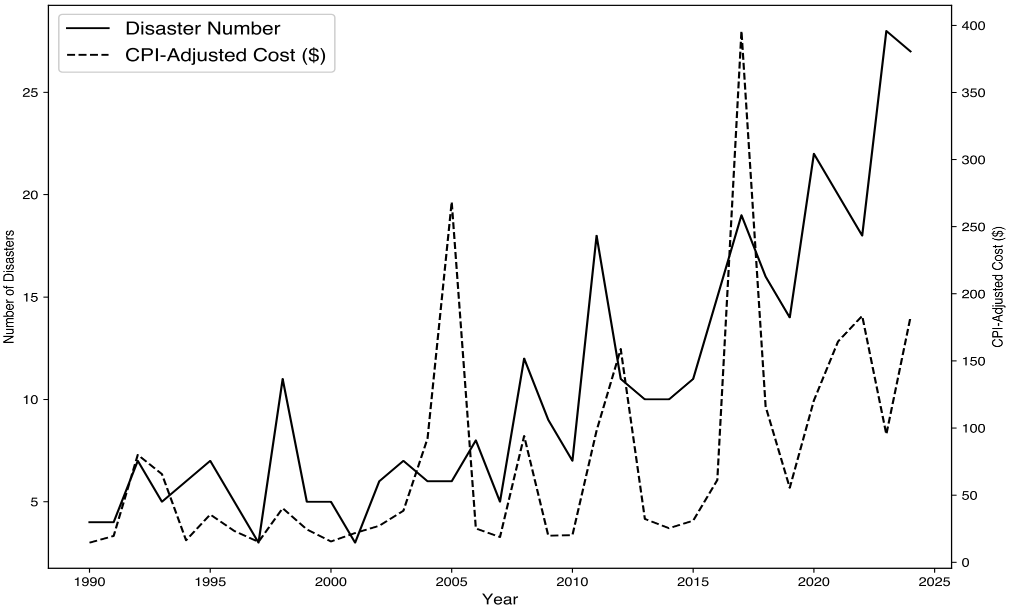

Figure 17: Number and losses of the U.S. billion-dollar disasters

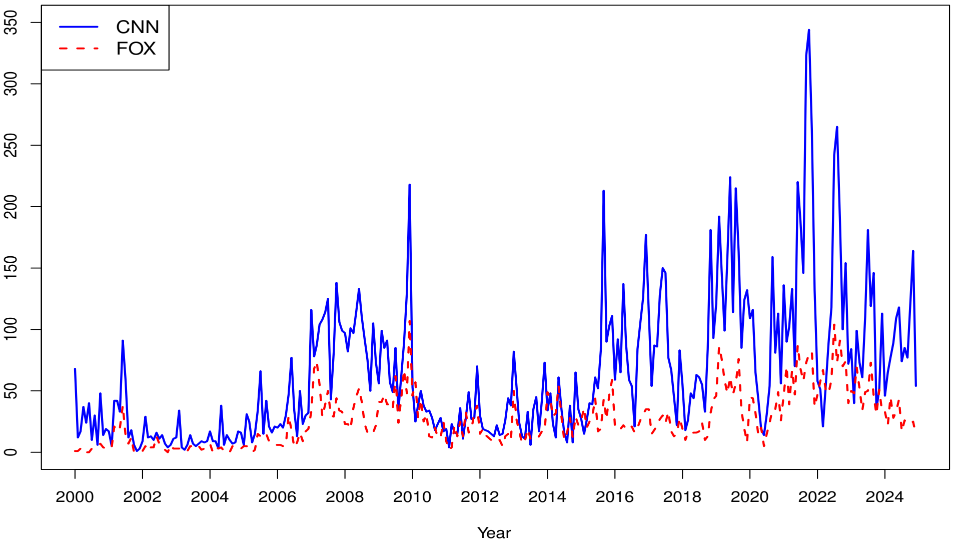

Figure 18: Monthly TV news coverage of climate change in the USA

References

Bode, E. (2011). Annual educational attainment estimates for US counties 1990–2005. *Letters in Spatial and Resource Sciences*, 4(2), 117–127.

Bowman, K., E. O'Neil, and H. Sims (2016). Polls on the environment, energy, global warming and nuclear power. *AEI Public Opin. Stud., Apr 21.Washington, DC: Am.Enterp. Inst.*

Durbin, J. and S.J. Koopman (2012). Time series analysis by state space methods (Vol. 38). *OUP Oxford.*

Egan, P.J. and M. Mullin (2017). Climate change: US public opinion. *Annual Review of Political Science*, 20(1), 209–227.

Jalles, J.T. (2009). Structural time series models and the kalman filter: a concise review.

Mariano, R.S. and Y. Murasawa (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of applied Econometrics*, 18(4), 427–443.

Matsumoto, N. (1990). Interpolation Method for Missing Data in Time Series of Groundwater Level and Barometric pressure. *Journal of the Seismological Society of Japan*, 43–2.

Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163, 688.

Wang, X., G. Deltas, M. Khanna, and X. Bi (2021). Community pressure and the spatial redistribution of pollution: The relocation of toxic-releasing facilities. *Journal of the Association of Environmental and Resource Economists*, 8(3), 577–616.