Social Forecasting

Final Assignment

Thomas Chadefaux

Instructions

For your final model, you will need to analyse and forecast the evolution of a time series. Please choose *one* of three options (details of each option below):

- Option 1: Public transportation demand
- Option 2: Presidential Polls
- Option 3: Data of your choice.

Regardless of your chosen option, **your report needs to include the following**:

1. Brief introduction to the problem and your approach
2. Brief description of the method/combination and justification for this choice (both theoretical and technical)
3. If relevant, all estimated equations associated with constructing forecasts from this method
4. Report the MAPE and MAE for the training period and the validation period. You may also report other metrics if relevant.
5. Forecasts for the following periods as specified. If you choose option 3 below, split your data into a learning set (80% of the data) and a "future" set (20%).
6. A single figure showing the fit of the final version of the model to the entire period available in the data (i.e., in-sample fit. For options 1 and 2, you do not have access to the "future" data). Note that the figure may include subfigures but all must fit in a single panel. Presentation matters, so make sure your plots are easy to understand and convey the information as effectively as possible.

Total word limit (all included): 1,500. NB: this is a maximum, not a target. Please submit your assignment via Blackboard}.

Three options

Option 1: Public transportation demand

Problem Description

A public transportation company is expecting increased demand for its services and is planning to acquire new buses and extend its terminals. These investments require a reliable forecast of future demand. The company's has data on each 15-minute interval between 6:30 and 22:00 on the number of passengers arriving at the terminal. As a forecasting consultant, you have been asked to create a forecasting method that can generate forecasts for the number of passengers arriving at the terminal.

Available Data

Part of the historic information is available in the file publicTransport_part.csv. The file contains the worksheet "Historic Information" with known demand for a 3-week period, separated into 15-minute intervals. Your job is to predict the period from 22-Mar-05 at 6:30 to 24-Mar-05 at 22:00.

Option 2: Presidential Polls

Problem description

Prior to US elections, voters are regularly asked about their preference for one candidate or another. This matters in order to anticipate the likely winner of the election, but also for candidates to gain a better understanding of what works and what does not. As a forecasting consultant, you have been asked by Joe Biden to create a forecasting method that can generate forecasts for his future poll results.

Available Data

The historical information is stored in the file named "presidentialPreference_part.xlsx." This file includes polling data from Nate Silver's FiveThirtyEight, covering the period from July 1, 2020, to November 2, 2020, updated daily. Suppose you receive a request on October 13, 2020, to forecast the polling results for the next 20 days, which spans from October 14, 2020, to November 2, 2020. It is important to note that you are unable to observe actual events occurring during this forecast period, from October 14, 2020, to November 2, 2020. Therefore, this period should not be used for either training or testing your model; it should only be used for evaluating the accuracy of your predictions after the fact.

Please report your forecasts and associated performance.

Option 3: Data of your choice

You may also use data of your choice (see list below). Whichever data you choose, you should use only 80% of the data for learning/testing purposes, and provide forecasts for the remaining 20%.

Datasets:

The following are available datasets that are suitable for time series analysis (from Youseop Shin, Time Series Analysis in the Social Sciences (2017), Appendix two.).

Crime

- Disaster Center Crime Pages provide data on national and state-level crime rates in the U.S. from 1960 (http://www.disastercenter.com/crime/).
- Federal Bureau of Investigation's Crime Statistics provide data on violent crimes, property crimes, hate crimes, and other data on sworn officers and civilian employees in the U.S. (https://www.fbi.gov/stats- services/crimestats, https://www.fbi.gov/about-us/cjis/ucr).
- National Archive of Criminal Justice Data provides data on criminological and criminal justice and links to related websites (http://www.icpsr.umich.edu/icpsrweb/NACJD/index.jsp).
- RAND State Statistics contain social science databases on multiple topics, such as crimes, demographics, and economics, covering all 50 U.S. states (http://www.randstatestats.org).
- Transactional Records Access Clearinghouse provides data on spending and enforcement activities of the federal government: FBI, INS, DEA, Homeland Security, and IRS (http://tracfed.syr.edu/).
- UK CrimeStats provides monthly crime rates in the U.K. from 2010 (http://www.ukcrimestats.com/National_Picture/).

Demography

- American Community Survey is conducted by the U.S. Census Bureau to supplement the census data that is collected every tenth year. The survey collects data on an ongoing basis, January through December, to provide more current demographic, economic, and housing characteristics at geographic levels comparable to the decennial census (http://www.census.gov/programs- surveys/acs). Tables and maps can be generated from the American Factfinder site (http://factfinder.census.gov).
- British Household Panel Survey (BHPS) is an annual survey that measures social and economic characteristics at the individual and household level in Britain from 1991–2009 (waves 1–18). From wave 19, the BHPS became part of the United Kingdom Household Longitudinal Study (UKHLS). Users can match the BHPS wave 1–18 data to the UKHLS wave 2 data and onwards (https://discover.ukdataservice.ac.uk/series/?sn=200005).

- Canadian Socioeconomic Information Management database contains more than 52 million numeric time series covering a wide variety of social and economic indicators for Canada (http://www.statcan.gc.ca/start- debut-eng.html).

- County and City Data Books provides access to the 1944–2000 County and City Data Books, which include level of income, level of education, and unemployment rate (http://ccdb.lib.virginia.edu). QuickFacts at the U.S. Census Bureau site has up-to-date data for states, counties, and cities (http://www.census.gov/quickfacts/table/PST045215/00).

- EUROSTAT provides data on demography, economy, and finance at the European level (http://ec.europa.eu/eurostat/web/main/home).

- German Socio-Economic Panel is a yearly longitudinal study of private households, starting in 1984. The data provide information on all household members, including demography, employment, income, taxes, social security, education, attitudes, values, and personality (https://data.soep.de/topics).

- Minority Data Resource Center provides data on issues (crime, education, employment, discrimination, immigration, etc.) affecting racial and ethnic minority populations in the U.S. (http://www.icpsr.umich.edu/icpsrweb/RCMD/).

- Murray Research Archive provides data on human development and social change, especially data that illuminate women's lives and issues of concern to women (http://murray.harvard.edu/dataverse).

- National Archives and Records Administration provides access to U.S. federal government records related to agriculture, economic and financial statistics, demographics, labor, education, health and social services, attitudinal data, international issues, military, and the environment. Most of the records start in the 1960s, with some from as early as World War II (http://www.archives.gov/research/electronic-records/).

- National Center for Education Statistics provides data related to the condition of education in the U.S. and internationally (http://nces.ed.gov).

- PolicyMap provides data related to demographics, socioeconomics, mortgages and home sales, health statistics, jobs and employment, and so on, for geographies across the U.S., often at neighborhood scale, such as by census tract and block group (http://www.policymap.com/data.html).

- U.S. Census Bureau provides census data regarding population and government (http://www.census.gov/en.html).

Economy

- Bureau of Economic Analysis provides information on national economic growth and regional economic development, such as gross domestic product, personal income, consumer spending, and corporate profits, in the U.S.; and international economy, such as trade in goods and services, balance of payments, international investment, and multinational enterprises—some data going back to the 1940s (http://www.bea.gov).

- Bureau of Labor Statistics provides time series data on employment, prices, working conditions, and productivity in the U.S. and other countries (http://www.bls.gov/data/home.htm).
- China Data Online (Zhongguo shu ju zai xian) includes monthly and yearly reports on China's macroeconomic development, population, and economy at the county and city level, and statistical yearbooks (http://chinadataonline.org).
- Consumer Expenditure Survey collects data on the buying habits and expenditures of households in the U.S. by a quarterly interview survey and weekly diary survey from 1989 to 2014. Aggregate expenditure shares tables give total expenditures by category for all consumer units and percentages of the totals for various demographic groups (http://www.bls.gov/cex/csxashar.htm).
- Consumer Price Index is a measure of the average change in prices paid by urban consumers for goods and services in the U.S. (http://www.bls.gov/cpi).
- Federal Reserve Archive (FRASER) provides economic data such as U.S. federal budget and consumer price index (https://fraser.stlouisfed.org).
- Federal Reserve Board provides current and historical data associated with surveys such as the Survey of Consumer Finances and Survey of Small Business Finances (http://www.federalreserve.gov/econresdata/).
- Inter-American Development Bank provides data on economic performance and institutions in Latin America from 1960 by topic and by country (https://data.iadb.org/DataCatalog/Dataset).
- LABORSTA provides data on economic indicators such as employment, working conditions, income, economic stability, and trade unions for about 200 counties, with some series beginning in 1970 (http://laborsta.ilo.org).
- National Longitudinal Surveys collects data on labor market experiences of individuals in the U.S. over time (http://www.nlsinfo.org/dbgator/index.php3 and http://www.bls.gov/nls).
- Surveys of Consumers provides monthly, quarterly, and yearly indices of consumer sentiment, current economic conditions, and consumer expectations in the U.S. from 1946 onwards (http://www.sca.isr.umich.edu/tables.html).
- Tax Foundation provides data on tax rates, collections and overall tax burdens, comparisons among states, results of opinion polls on taxes, and so on, in the U.S. (http://taxfoundation.org/data).
- Union Membership and Coverage Database provides private- and public-sector labor union membership, coverage, and density estimates in the U.S. compiled from the monthly household Current Population Survey. Economy-wide estimates are provided beginning in 1973; estimates by state, detailed industry, and detailed occupation begin in 1983; and estimates by metropolitan area begin in 1986 (http://unionstats.gsu.edu).

Elections

- American National Election Studies collects election- related data every two years from 1948 onwards (http://electionstudies.org). It is mostly cross-sectional data and occasionally panel data. The same items (such as party identification, feeling thermometers, candidate evaluations, voting, and vote choices) are collected in every survey, and therefore we can build aggregate-level time series data. In addition, the ANES have merged all cross-section cases and variables for select questions from the ANES conducted since 1948 into a single file, the ANES Time Series Cumulative Data File (http://electionstudies.org/studypages/anes_timeseries_cdf/anes_timeseries_cdf.htm).

- Atlas of US presidential Elections aggregates official election results from all 50 states and D.C. (http://uselectionatlas.org). Canadian Election Study provides data on election participation and vote choices in Canadian since 1965 (http://www.queensu.ca/cora/ces.html).
- Cooperative Congressional Election Study is a 50,000+ person national stratified sample survey administered by YouGov/Polimetrix in the U.S. every year from 2005 to 2012. The survey consists of two waves in election years. In non-election years, the survey consists of a single wave conducted in the early fall (http://projects.iq.harvard.edu/cces).
- Election Assistance Commission provides U.S. national and state-level voter registration and turnout statistics for federal elections from 1960 (http://www.eac.gov/research/election_administration_an d_voting_survey.aspx).
- Federal Election Commission provides data related to campaign finance (http://www.fec.gov/about.shtml). An FTP site for FEC data files back to 1978 is available (ftp://ftp.fec.gov/FEC). Council of European Social Science Data Archives provides access to important resources of relevance to the European social science research agenda. It also provides links to the data centers of member countries: Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Lithuania, Netherlands, Norway, Slovenia, Sweden, Switzerland, and United Kingdom (http://cessda.net/).
- Cross-National Time-Series Data Archive is a dataset of annual data for social science research from 1815 onwards for over 200 countries. It includes data related to domestic conflict usage, economy, elections, legislative process, government revenue and expenditure, military, population, and education (http://www.databanksinternational.com).
- Global Peace Index is a measure of the relative position of nations' and regions' peacefulness in terms of the level of safety and security in society, the extent of domestic and international conflict, and the degree of militarization since 2007 (http://economicsandpeace.org).
- Kansas Event Data System uses automated coding of English-language news reports to generate political event (conflicts and mediation) data focusing on the Middle East, Balkans, and West Africa (https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10713).
- Minorities at Risk Project monitors the status and conflicts of 284 politically active ethnic groups in all countries with a current population of at least 500,000 throughout the world from 1945 to the present (http://www.cidcm.umd.edu/mar).
- OECD DATA provides data on economy, education, finance, government, health, society, and so on for countries in the Organisation for Economic Co-operation and Development

(https://data.oecd.org/).

- UNdata pools major UN databases and several international databases (such as LABORSTA and OECD DATA), so users do not have to move from one database to another to access different types of information (http://data.un.org/).
- UNESCO Institute for Statistics provides data on education, science and technology, culture, and communication for more than 200 countries (http://www.uis.unesco.org).
- UN Population Information Network provides a list of links to national and nongovernmental agencies that provide data on demographics, population, and family planning (http://www.un.org/popin/other4.htm).
- Vote World archives datasets of roll-call voting from legislative bodies throughout the international community, including the U.S. House of Representatives and Senate, the European Parliament, and the United Nations (http://voteworld.berkeley.edu).
- American Presidency Project provides archives that contain datasets on presidents' relationship with congress, popularity, and so on. It also provides document archives related to the study of the presidency, from which researchers can build their own time series data sets, for example regarding the presidents' issue priority and the degree of going public (http://www.presidency.ucsb.edu).
- Center for Responsive Politics (OpenSecrets.org) offers data related to money in U.S. politics, such as campaign expenditures and contributions (https://www.opensecrets.org).
- Keith Poole's Data Download Page (http://voteview.com/dwnl.html) and Tim Groseclose's Interest Group Score Page (http://www.sscnet.ucla.edu/polisci/faculty/groseclose/Adj .Int.Group.Scores) provide measures of U.S. legislators' ideological predispositions. These two measures are substitutes for interest groups' ratings of individual legislators' voting records (such as American Conservative Union and Americans for Democratic Action ratings) that are not comparable over time.
- Pew Research Center makes its data related to various topics (such as U.S. Politics & Policy, Journalism & Media, Religion & Public Life, Global Attitudes & Trends, and Social & Demographic Trends) available to the public for secondary analysis after a period of time (http://www.pewresearch.org/data/download-datasets).
- Policy Agendas Project collects and organizes data (such as Gallup's "most important problem," policy moods, and federal budget) from various archived sources to trace changes in the national policy agenda and public policy outcomes since the Second World War (http://www.policyagendas.org).
- Presidential Data Archive (Presidency Research) provides datasets that include presidential support scores and survey reports from Gallup, CBS/New York Times, Tyndal, Wirthlin, and Pew (http://presdata.tamu.edu).

Public Opinion Public opinion surveys provided by the following organizations repeatedly contain the same items, and therefore we can build time series data regarding various topics.

- Canadian Opinion Research Archive collects opinion surveys dating back to the 1970s (http://www.queensu.ca/cora).
- Gallup provides surveys that track public opinions on various political, economic, and social issues in more than 160 countries (http://www.gallup.com).
- Inter-university Consortium for Political and Social Research provides mostly cross-sectional survey data and occasionally panel data about various topics in the U.S. and abroad (http://www.icpsr.umich.edu/icpsrweb/ICPSR).
- Roper Center for Public Opinion Research provides social science data, specializing in data from public opinion surveys on a vast range of topics in U.S. and abroad, from the 1930s (http://ropercenter.cornell.edu/).