User guide to UHERO's forecast processes

Peter Fuleky

4/10/23

Table of contents

1 About

This document describes the setup and components of the `forecastr` project. The focus of the project is forecasting using multi-equation behavioral models. The project encompasses data preparation, model selection (work in progress), external forecast generation, local forecast generation (planned), simulations (planned), and forecast distribution to a more granular scale.

1.1 Contents

Chapters 2-3 discuss the general setup of a collaborative project under version control. Chapter 4 deals with the setup file that configures the most general aspects of the `forecastr` project. Chapter 5 describes user defined helper functions for the `forecastr` project. Chapter 6 gives examples of best practices for time series manipulation.

2 Project setup

A project is self-contained in a folder (the project folder).

2.1 A quick overview:

TL;DR:
https://kdestasio.github.io/post/r_best_practices/

2.2 Basic ideas for a reproducible workflow

Set up your work in projects:
https://r4ds.hadley.nz/workflow-projects.html

Read section I in WTF: https://rstats.wtf/save-source.html

A good list of consideration for structuring projects: https://www.r-bloggers.com/2018/08/structuring-r-projects/

Use RStudio projects with sub-directories - R - R code. - data/raw - data external to the project. - data/processed - intermediate processed data. - notes - Rmd, and Rmd output, notes, papers, supporting documents, Rmd, etc. - output - reports, tables, etc. - output/plots - plots. - renv - used for library management (don't edit). - man - help files (don't edit).

Coding Conventions in R

Follow the `tidyverse` style guide: https://style.tidyverse.org/index.html

Add 4 dashes after a section header for it to show up in the outline:

```
# *************************
#  setup ----
# *************************
```

Package management with `renv`: https://rstudio.github.io/renv/articles/renv.html

Collaborating with `renv`: https://rstudio.github.io/renv/articles/collaborating.html

2.3 Produce output via R Markdown

Preference settings: https://bookdown.org/yihui/rmarkdown-cookbook/working-directory.html

Render results from R scripts via Rmd: 1) source your R code from within Rmd https://bookdown.org/yihui/rmarkdown-cookbook/source-script.html 1*) save output from R script and load it in Rmd setup chunk 2) only render important results in Rmd chunks

2.4 Dealing with credentials

Store your credentials and sensitive info in project specific .Renviron file (project's root directory, must end with `\n`):

```
udaman_token = "<your udaman token>"
udaman_user = "<your user name>"
udaman_pwd = "<your password>"
GITHUB_PAT = "<your github pat>"
```

Retrieve credentials on demand. Do not store/assign the retrieved credentials to a variable: `req <- httr::GET(url, httr::add_headers(Authorization = paste("Bearer", Sys.getenv("udaman_token"))))`

2.5 Additonal resources:

Look at the `targets` package for workflow automation: https://docs.ropensci.org/targets/index.html

Look at the Efficient R programming book: https://csgillespie.github.io/efficientR/

Also look at the Stanford Guide: http://dcl-workflow.stanford.edu

3 Git

Set up version control for a project.

3.1 A quick overview:

TL;DR: https://inbo.github.io/git-course/index.html

3.2 Git step by step

If you don't have Git, install it:
https://happygitwithr.com/install-git.html

Make sure .gitignore contains the following files:
.Renviron
.Rprofile

Introduce yourself to Git:
In the shell (Terminal tab in RStudio):
git config –global user.name 'Jane_Doe'
git config –global user.email 'jane@example.com'
git config –global –list

For more advanced tasks, use GitHub Desktop:
https://desktop.github.com

Generate and store your GitHub PAT (Personal Access Token):
https://happygitwithr.com/https-pat.html
https://docs.github.com/en/authentication/keeping-your-account-and-data-secure/creating-a-personal-access-token

Use one of three ways to add your project to GitHub:
Brand new project:
https://happygitwithr.com/new-github-first.html
Existing project without version control:
https://happygitwithr.com/existing-github-first.html

Existing project under local version control:
https://happygitwithr.com/existing-github-last.html

Troubleshooting if RStudio can't detect Git:
https://happygitwithr.com/rstudio-see-git.html

Git vocabulary:
https://happygitwithr.com/git-basics.html

Remote setups (try to stick to GitHub first discussed above):
https://happygitwithr.com/common-remote-setups.html

Useful Git workflows and dealing with conflicts:
https://happygitwithr.com/workflows-intro.html
https://happygitwithr.com/push-rejected.html
https://happygitwithr.com/pull-tricky.html

Additional resources:
https://happygitwithr.com/ideas-for-content.html

3.3 Suggested workflow

1) Initialize repository on GitHub.com under the UHERO account.

2) Clone it via RStudio project setup.

3) Commit changes, pull, resolve issues, push.
   3*) If work in a branch (create in RStudio), commit to branch, (pull) push to remote, pull request on GitHub.com from branch to main, merge, delete branch on GitHub.com.

See these detailed guides: https://inbo.github.io/git-course/index.html https://github.com/llendway/github_fo

Another research workflow based on Github: https://www.carlboettiger.info/2012/05/06/research-workflow.html

4 Setup of the forecastr project

The `setup.R` file contains general information used throughout the project. The contents are listed below.

4.1 Start with a clean slate

First remove all objects from global environment:
`rm(list = ls())`

If only some objects need to be removed, search for them via wildcards:
`rm(list = ls(pattern = glob2rx("*__*")))`

Detach all loaded packages:

```
if (!is.null(names(sessionInfo()$otherPkgs))) {
invisible(
  suppressMessages(
    suppressWarnings(
      lapply(
        paste("package:", names(sessionInfo()$otherPkgs), sep=""),
        detach,
        character.only = TRUE,
        unload = TRUE
        )
      )
    )
  )
}
```

4.2 Packages

The setup file clarifies its own location relative to the project root and loads the necessary packages.

Navigate within a project using the `here()` package. Start by specifying:
`here::i_am("R/setup.R")`

Then load necessary packages

```
library(here) # navigation within the project
library(conflicted) # detect conflicts across packages
library(tidyverse) # a set of frequently used data-wrangling tools
library(magrittr) # more than just pipes
library(lubridate) # dealing with dates
library(tsbox) # dealing with time series
# library(bimets)
```

Detect conflicts across packages and assign preferences

```
conflict_scout()
conflict_prefer("filter", "dplyr") # dplyr v stats
conflict_prefer("first", "dplyr") # dplyr v xts
conflict_prefer("lag", "dplyr") # dplyr v stats
conflict_prefer("last", "dplyr") # dplyr v xts
conflict_prefer("extract", "magrittr") # magrittr vs tidyr
```

Verify top level project directory
`here()`

4.3 Package descriptions

Use the `here` package to deal with file paths:
https://here.r-lib.org

Suppose you have a dataset in csv format. Use:
`readr::read_csv(here::here("<The subfolder where your csv file resides>", "<The CSV file.csv>"))`

Only load essential packages with many useful functions (don't load a whole package to access a single function).
Refer to individual functions in not loaded packages by `namespace::function()`

Resolve conflicts across multiple packages with `conflicted`:
https://conflicted.r-lib.org

Core `tidyverse` packages:
https://www.tidyverse.org

Non-core `tidyverse` packages (need to be loaded separately):
https://magrittr.tidyverse.org
https://lubridate.tidyverse.org

Time series tools in `tsbox` (learn them and use them, very useful). All start with `ts_`.
https://www.tsbox.help

Forecasting with multi-equation behavioral models (only load `bimets` if actually doing forecasts, no need for data manipulation):
https://cran.r-project.org/web/packages/bimets/index.html

`bimets` depends on `xts` (if not loaded, can access necessary functions via `xts::function()`):
https://cran.r-project.org/web/packages/xts/index.html

Prefer using `tsbox` and `tidyverse` functions whenever possible, but understand the components and behavior of `xts` objects: https://rc2e.com/timeseriesanalysis

4.4 Additional info in setup

Define project-wide constants:

```
bnk_start <- ymd("1900-01-01")
bnk_end <- ymd("2060-12-31")
```

Load user defined utility functions (details in next section):
`source(here("R", "util_funs.R"))`

5 User defined utility functions

Functions not available in existing packages are stored in util_funs.R. A pdf version of this document is available here.

5.1 AtoQ

Description:

```
Linear interpolation based on aremos command reference page 292
```

Usage:

```
AtoQ(ser_in, aggr = "mean")
```

Arguments:

```
ser_in: the xts series to be interpolated (freq = a)

aggr: interpolation method: aggregate via mean (default) or sum
```

Value:

```
interpolated xts series (freq = q)
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
```

5.2 explode_xts

Description:

```
Splitting of xts matrix to individual xts vectors (don't use,
pollutes global environment)
```

Usage:

```
explode_xts(xts_in)
```

Arguments:

```
xts_in: the xts matrix to be split into individual xts vectors
```

Value:

```
nothing (silently store split series in global environment)
```

Examples:

```
get_series_exp(74, save_loc = NULL) %>%
  ts_long() %>%
  ts_xts() %>%
  explode_xts()
rm(list = ls(pattern = glob2rx("*@HI.Q")))
```

5.3 find_end

Description:

```
Find the date of the last observation (NAs are dropped)
```

Usage:

```
find_end(ser_in)
```

Arguments:

```
ser_in: an xts series
```

Value:

```
date associated with last observation
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2060, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2018"] <- c(323127513, 325511184, 327891911)
find_end(`ncen@us.sola`)
```

5.4 find_start

Description:

```
Find the date of the first observation (NAs are dropped)
```

Usage:

```
find_start(ser_in)
```

Arguments:

```
ser_in: an xts series
```

Value:

```
date associated with first observation
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2017/2021"] <- c(325511184, 327891911, 330268840, 332639102, 334998398)
find_start(`ncen@us.sola`)
```

5.5 get_series

Description:

```
Download a set of series from udaman using series names
```

Usage:

```
get_series(ser_id_vec)
```

Arguments:

```
ser_id_vec: vector of series names
```

Value:

```
time and data for all series combined in a tibble
```

Examples:

```
get_series(c("VISNS@HI.M", "VAPNS@HI.M"))
```

5.6 get_series1

Description:

```
Download a single series from udaman using series name
```

Usage:

```
get_series1(ser_id)
```

Arguments:

```
ser_id: udaman series name
```

Value:

```
time and data for a single series combined in a tibble
```

Examples:

```
get_series("VISNS@HI.M")
```

5.7 get_series_exp

Description:

```
Download series listed in an export table from udaman
```

Usage:

```
get_series_exp(exp_id, save_loc = "data/raw")
```

Arguments:

```
exp_id: export id

save_loc: location to save the csv of the retrieved data, set to NULL
     to avoid saving
```

Value:

```
time and data for all series combined in a tibble
```

Examples:

```
get_series_exp(74)
get_series_exp(74, save_loc = NULL)
```

5.8 get_var

Description:

```
Construct a series name from variable components and retrieve the
series
```

Usage:

```
get_var(ser_in, env = parent.frame())
```

Arguments:

```
ser_in: a variable name (string with substituted expressions)

env: environment where the expression should be evaluated
```

Value:

```
variable
```

Examples:

```
ser_i <- "_NF"
cnty_i <- "HI"
get_series_exp(74, save_loc = NULL) %>%
  ts_long() %>%
  ts_xts() %$% get_var("E{ser_i}@{cnty_i}.Q")
```

5.9 make_xts

Description:

```
Create xts and fill with values
```

Usage:

```
make_xts(start = bnk_start, end = bnk_end, per = "year", val = NA)
```

Arguments:

```
start: date of series start (string: "yyyy-mm-dd")

end: date of series end (string: "yyyy-mm-dd")

per: periodicity of series (string: "quarter", "year")

val: values to fill in (scalar or vector)
```

Value:

```
an xts series
```

Examples:

```
make_xts()
make_xts(start = ymd("2010-01-01"), per = "quarter", val = 0)
```

5.10 p

Description:

```
Concatenate dates to obtain period
```

Usage:

```
p(dat1, dat2)
```

Arguments:

```
dat1: date of period start (string: yyyy-mm-dd)

dat2: date of period end (string: yyyy-mm-dd)
```

Value:

```
string containing date range
```

Examples:

```
p("2010-01-01", "2020-01-01")
```

5.11 pca_to_pc

Description:

```
Convert annualized growth to quarterly growth
```

Usage:

```
pca_to_pc(ser_in)
```

Arguments:

```
ser_in: the series containing annualized growth (in percent)
```

Value:

```
series containing quarterly growth (in percent)
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
ts_c(test1 %>% ts_pca() %>% pca_to_pc(), test1 %>% ts_pc())
```

5.12 pchmy

Description:

```
Calculate multi-period average growth
```

Usage:

```
pchmy(ser_in, lag_in = 1)
```

Arguments:

```
ser_in: name of xts series for which growth is calculated

lag_in: length of period over which growth is calculated
```

Value:

```
series containing the average growth of ser_in (in percent)
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
ts_c(pchmy(`ncen@us.sola`, lag_in = 3), ts_pc(`ncen@us.sola`))
ts_c(pchmy(test1, lag_in = 4), ts_pcy(test1), ts_pca(test1), ts_pc(test1))
```

5.13 plot_1

Description:

```
Interactive plot of a single variable with level and growth rate
```

Usage:

```
plot_1(
  ser,
  rng_start = as.character(Sys.Date() - years(15)),
  height = 300,
  width = 900
)
```

Arguments:

```
ser: time series to plot

rng_start: start of zoom range ("YYYY-MM-DD")

height: height of a single panel (px)

width: width of a single panel (px)
```

Value:

```
a dygraph plot
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
plot_1(`ncen@us.sola`, rng_start = "2017-01-01")
plot_1(test1, rng_start = "2017-01-01")
```

5.14 plot_comp

Description:

```
Three-panel plot of levels, index, and growth rates
```

Usage:

```
plot_comp(
  sers,
  rng_start = as.character(Sys.Date() - years(15)),
  rng_end = as.character(Sys.Date()),
  height = 300,
  width = 900
)
```

Arguments:

```
sers: a vector of series to plot

rng_start: start of the zoom range ("YYYY-MM-DD")

rng_end: end of the zoom range ("YYYY-MM-DD")

height: height of a single panel (px)

width: width of a single panel (px)
```

Value:

```
a list with three dygraph plots (level, index, growth)
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
plot_comp(ts_c(`ncen@us.sola`, test1), rng_start = "2017-01-01")
get_series_exp(74, save_loc = NULL) %>%
  ts_long() %>%
  ts_xts() %>%
  extract(, c("E_NF@HI.Q", "ECT@HI.Q", "EMN@HI.Q")) %>%
  plot_comp()
```

5.15 plot_comp_2

Description:

```
Two-panel plot of levels, index, and growth rates
```

Usage:

```
plot_comp_2(
  sers,
  rng_start = as.character(Sys.Date() - years(15)),
  rng_end = as.character(Sys.Date()),
  height = 300,
  width = 900
)
```

Arguments:

```
sers: a vector of series to plot

rng_start: start of the zoom range ("YYYY-MM-DD")

rng_end: end of the zoom range ("YYYY-MM-DD")

height: height of a single panel (px)

width: width of a single panel (px)
```

Value:

```
 a list with two dygraph plots (level, index, growth)
```

Examples:

```
 `ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
   ts_xts()
 `ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
 test1 <- AtoQ(`ncen@us.sola`)
 plot_comp_2(ts_c(`ncen@us.sola`, test1), rng_start = "2017-01-01")
 get_series_exp(74, save_loc = NULL) %>%
   ts_long() %>%
   ts_xts() %>%
   extract(, c("E_NF@HI.Q", "ECT@HI.Q", "EMN@HI.Q")) %>%
   plot_comp_2()
```

5.16 QtoA

Description:

```
 Conversion from quarterly to annual frequency
```

Usage:

```
 QtoA(ser_in, aggr = "mean")
```

Arguments:

```
ser_in: the xts series to be converted (freq = q)

aggr: aggregate via mean (default) or sum
```

Value:

```
converted xts series (freq = a)
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
test2 <- QtoA(test1) # for stock type variables mean, for flow type variables sum
print(test1)
print(cbind(`ncen@us.sola`, test2))
```

5.17 QtoM

Description:

```
Interpolate a tibble of series from quaterly to monthly freq
```

Usage:

```
QtoM(data_q, conv_type)
```

Arguments:

```
data_q: tibble containing variables at quarterly freq

conv_type: match the quarterly value via "first", "last", "sum",
     "average"
```

Value:

```
tibble containing variables at monthly freq
```

Examples:

```
`ncen@us.sola` <- ts(NA, start = 2016, end = 2021, freq = 1) %>%
  ts_xts()
`ncen@us.sola`["2016/2021"] <- c(323127513, 325511184, 327891911, 330268840, 332639102, 334
test1 <- AtoQ(`ncen@us.sola`)
QtoM(ts_tbl(test1), "average")
ts_frequency(QtoM(ts_tbl(test1), "average") %>% ts_xts())
```

5.18 QtoM1

Description:

```
Interpolate a single series from quarterly to monthly freq
```

Usage:

```
QtoM1(var_q, ts_start, conv_type)
```

Arguments:

```
var_q: vector containing a single variable at quarterly freq

ts_start: starting period as c(year, quarter) e.g. c(2001, 1)

conv_type: match the quarterly value via "first", "last", "sum",
     "average"
```

Value:

```
vector containing a single variable at monthly freq
```

Examples:

```
QtoM1(test1, c(2010, 1), "average")
```

5.19 qtrs

Description:

```
Convert period in quarters to period months
```

Usage:

```
qtrs(nr_quarters)
```

Arguments:

```
nr_quarters: number of quarters in period (integer)
```

Value:

```
number of months in period
```

Examples:

```
qtrs(3)
ymd("2020-01-01") + qtrs(3)
```

6 Best practices for time series data manipulation

Use capital letters for series names. Special characters in variable names require putting the name between backticks (e.g. `N@US.A`). Eliminate special characters using a long tibble.

```
hist_q_mod <- hist_q %>%
  ts_tbl() %>%
  mutate(id = str_replace_all(id, c("@" = "_AT_", "\\." = "_")))

# revert back to udaman notation
hist_q <- hist_q_mod %>%
  ts_tbl() %>%
  mutate(id = str_replace_all(id, c("_AT_" = "@", "_Q" = "\\.Q")))
```

Use the `xts` format whenever possible. Observations in a multivariate `xts` can be accessed by time and series name in two ways: `mul_var_xts[time, ser_name]` or `mul_var_xts$ser_name[time]`.

Make sure all series are defined on the same range (default start = bnk_start, end = bnk_end). Take advantage of `make_xts()` (and its defaults, e.g. start and end period).

```
import_xts <- read_csv(here("data/raw", str_glue("{exp_id_a}.csv"))) %>%
  arrange(time) %>%
  ts_long() %>%
  ts_xts() %>%
  ts_c(
    temp = make_xts(per = "year") # temporary variable to force start and end in import_xts
  ) %>%
  extract(, str_subset(colnames(.), "temp", negate = TRUE)) # remove temp
```

Don't break up multivariate time series (think databank) into individual series in the global environment.

If referring directly to a series with a static name, use the `bank$series` notation (this can be used on both the right and the left hand side of the assignment, while `bank[, series]` can only be used for existing series in bank).

```
# find the last value in history
dat_end <- find_end(hist_q$N_AT_US_Q)
# same as
dat_end <- find_end(hist_q[, N_AT_US_Q])
```

Use `[p()]` to select a period in `xts` objects, otherwise use `ts_span()`.

```
# extend series with addfactored level
sol_q$N_AT_US_SOLQ <- hist_q$N_AT_US_Q[p("", dat_end)] %>%
  ts_bind(sol_q$NCEN_AT_US_SOLQ[p(dat_end, "")] +
    as.numeric(sol_q$N_AT_US_SOLQ_ADDLEV[dat_end]))

# addfactor for growth
sol_q$N_AT_US_SOLQ_ADDGRO[p(dat_end + qtrs(1), dat_end + qtrs(4))] <- -0.35

# extend history using growth rate
sol_q$N_AT_US_SOLQ <- sol_q$N_AT_US_SOLQ[p("", dat_end)] %>%
  ts_chain(ts_compound(sol_q$N_AT_US_SOLQ_GRO[p(dat_end, "")]))
```

The `bank[,seriesname]` notation only works for *existing* `xts` series on the left of the assignment (it can also be used on the right). `seriesname` can be determined at runtime

```
# initialize the lhs series in the "bank"
hist_a$temp <- make_xts()
names(hist_a)[names(hist_a) == "temp"] <- str_glue("E{ser_i}_AT_{cnty_i}_ADD")

# calculate expression and assign to lhs
hist_a[, str_glue("E{ser_i}_AT_{cnty_i}_SH")] <-
  (hist_a[, str_glue("E{ser_i}_AT_{cnty_i}")] / hist_a[, str_glue("E_NF_AT_{cnty_i}")])
```

Alternatively, make multiple series in *bank* available by `%$%` and retrieve inividual series by `get_var()` on the right.

```
hist_a[, str_glue("E{ser_i}_AT_{cnty_i}_SH")] <- hist_a %$%
  (get_var("E{ser_i}_AT_{cnty_i}") / get_var("E_NF_AT_{cnty_i}"))
```

Bimets requires data in a particular `tslist` format. Convert `xts` to `tslist` using `ts_tslist()`.

```
# store series as tslist
hist_a_lst <- hist_a %>%
  ts_tslist() %>%

# convert series to bimets format
hist_a_bimets <- hist_a_lst %>%
  map(as.bimets)

# bimets strips the attributes, need to reset them for further manipulation by tsbox
hist_a <- hist_a_bimets %>%
  set_attr("class", c("tslist", "list")) %>%
  ts_xts()
```

For series collected in a `tslist` on the left of the assignment use the `bank[[seriesname]]` notation (it can also be used on the right). Here the lhs series `seriesname` does not need to exist, and it might easier to work with `tslist` than `xts` when variable names are determined at runtime.

```
# similar to above with a tslist variable
hist_a_lst[[str_glue("E{ser_i}_AT_{cnty_i}_ADD")]] <- hist_a_lst %$%
  (get_var("E{ser_i}_AT_{cnty_i}") - get_var("E_NF_AT_{cnty_i}"))
```

6.1 Harness the power of tsbox

Use the converter functions in `tsbox` to shift between various data types (`ts_tbl()`, `ts_xts()`, `ts_ts()`, `ts_tslist()`) and reshaping to the long and wide format (`ts_long()`, `ts_wide()`). `tsbox` further contains funtions for time period selection (`ts_span()`), merging and extension operations (`ts_c()`, `ts_bind()`, `ts_chain()`), transformations (`ts_lag()`, `ts_pc()`, `ts_pca()`, `ts_pcy()`, `ts_diff()`, `ts_diffy()`), and index construction (`ts_compound()`, `ts_index()`). Consider these before turning to solutions that are specific to the `xts`, `ts`, `dplyr` or `tidyr` packages.

7 Model selection

The model selection process can be run line-by-line from an R script directly (R/gets_model_select.R) or via sourcing an Rmd document (notes/gets_model_select.Rmd) which collects all model selection results in an easier to digest html file. Running the full script (source) takes about 1 minute.

7.1 Main user settings

- Start and end of period used for model selection.
- End of period used for estimation (selected model can be re-estimated for different sample).
- Start and end of quasi-forecast period (for model evaluation).
- Maximum number of lags considered in models.
- Response variable.
- List of predictors.

7.2 Data preparation (`tidyverse`)

- Download all series used in the model selection process from UDAMAN (about 500 rows and 1200 columns) and eliminate special characters from the series names.
- Log-transform all variables.
- Load (create) all indicators (dummies for impulse, level shift, seasonal) and trend.
- Combine all variables into a single dataset.

- Set date range for model selection.
- Generate 8 lags of predictors.
- Filter data set down to specific variables considered in a particular model, including trend and season dummies.

7.3 Model selection steps (`gets`)

https://cran.r-project.org/web/packages/gets/index.html
- Formulate a general unrestricted model.
- Run the gets (general to specific) model selection algorithm.
- Identify outliers in the relationship.
- Repeat gets model selection over specific model and outliers.
- Verify that no additional outliers arise due to greater model parsimony.
- If estimation period is shorter than model selection period, remove predictors containing zeros only (e.g. outlier past the end of estimation period).
- Re-estimate final model.
- Save model equation as a txt file (not plugging in estimated coefficients here to keep it general). If happy with the model, copy this equation into file containing all model equations.

7.4 Produce a quasi-forecast with the selected model (`bimets`)

https://cran.r-project.org/web/packages/bimets/vignettes/bimets.pdf
- Load model from txt file.
- Load data used by the model.
- Estimate the model (if estimation period ends before the last data point also run a Chow test of model stability).
- Simulate model.
- Evaluate simulation by plotting quasi-forecast and actual history.

8 Stochastic simulations

The model selection process stores a set of general equations (coefficient estimates are not plugged in) in a text file. Before simulation can commence, several steps need to take place: compile the system of equations, add data to the equations, estimate equations. `bimets` does not automatically adjust the sample for missing data points, so need to identify the time period with a rectangular sample for the estimation of each equation. For forecasting, deal with the ragged edge of the data by finding the last data point for each series and "exogenize" the series up to that point (use actuals in simulation).

8.1 Main user settings

- Start of forecast period.
- End of forecast period.
- End of estimation period.
- Maximum number of lags in models.

8.2 Data preparation

- Download all series used in the model selection process from UDAMAN (about 500 rows and 1200 columns) and eliminate special characters from the series names.
- Load (create) all indicators (dummies for impulse, level shift, seasonal) and trend.
- Combine all variables into a single dataset.

8.3 Simulation prep

- Compile model (load equations from text file and let `bimets` digest the info).
- Add variables to model.

- Set date range for estimation (`bimets` does not automatically drop periods with NA's).
- Set exogenization range to deal with ragged edge in simulation.

- Estimate model equations and save estimation results to text file for inspection.
- Set add factors.

8.4 Simulation

- Simulate model deterministically to obtain mean forecast.
- Extract forecast and combine it with history.
- Inspect the forecast via plots.
- Set parameters for stochastic simulations.
- Run stochastic simulation.
- Extract simulated paths and obtain deviations from the mean forecast.
- Inspect the paths via plots.

9 Notes

9.1 Project setup

Coding Conventions in R:

Basic ideas for a reproducible workflow:

Use RStudio projects with sub-directories

- R - R code.
- data/raw - data external to the project.
- data/processed - intermediate processed data.
- notes - Rmd, and Rmd output, notes, papers, supporting documents, Rmd, etc.
- output - reports, tables, etc.
- output/plots - plots.
- renv - used for library management (don't edit).
- man - help files (don't edit).

Preference settings:

9.2 Git

A quick overview: https://github.com/llendway/github_for_collaboration/blob/master/github_for_collaborat

9.3 Git step by step

If you don't have Git, install it:
https://happygitwithr.com/install-git.html

Make sure .gitignore contains the following files:
.Renviron .Rprofile

Introduce yourself to Git:
In the shell (Terminal tab in RStudio):
git config –global user.name 'Jane Doe'

git config –global user.email 'jane@example.com'
git config –global –list

For more advanced tasks, use GitHub Desktop:
https://desktop.github.com

Store your GitHub PAT (Personal Access Token):
https://happygitwithr.com/https-pat.html

Use one of three ways to add your project to GitHub:
Brand new project:
https://happygitwithr.com/new-github-first.html
Existing project without version control:
https://happygitwithr.com/existing-github-first.html
Existing project under local version control:
https://happygitwithr.com/existing-github-last.html

Troubleshooting if RStudio can't detect Git:
https://happygitwithr.com/rstudio-see-git.html

Git vocabulary:
https://happygitwithr.com/git-basics.html

Remote setups (try to stick to GitHub first discussed above):
https://happygitwithr.com/common-remote-setups.html

Useful Git workflows and dealing with conflicts:
https://happygitwithr.com/workflows-intro.html
https://happygitwithr.com/push-rejected.html
https://happygitwithr.com/pull-tricky.html

Additional resources:
https://happygitwithr.com/ideas-for-content.html

Suggested workflow:
1) Initialize repository on GitHub.com under the UHERO account.
2) Clone it via RStudio project setup.
3) Commit changes, pull, resolve issues, push. 3*) If work in a branch (create in RStudio), commit to branch, (pull) push to remote, pull request on GitHub.com from branch to main, merge, delete branch on GitHub.com.

Render results from R scripts via Rmd: 1) source your R code from within Rmd 2) only render important results in Rmd chunks

Use here() from the here package to write file paths

Suppose you have a dataset in csv format. Use readr::read_csv(here::here("The subfolder where your csv file resides", "The CSV file.csv"))

Do not use setwd() and rm(list = ls())

Do not save the workspace to the .Rdta file

Use library() not require()

Use version control (useful for recording changes between different versions of a file over time - see below for Git integration)

See the resources below:

Best Practices & Style Guide for Writing R Code: https://github.com/kmishra9/Best-Practices-for-Writing-R-Code

R Code – Best practices: https://www.r-bloggers.com/2018/09/r-code-best-practices/

R Best Practices by Krista L. DeStasio: https://kdestasio.github.io/post/r_best_practices/

Project-oriented workflow: https://www.tidyverse.org/blog/2017/12/workflow-vs-script/

R coding style best practices: https://www.datanovia.com/en/blog/r-coding-style-best-practices/

What They Forgot to Teach You About R by Jennifer Bryan and Jim Hester: https://rstats.wtf/save-source.html

Conflicted: a new approach to resolving ambiguity: https://www.tidyverse.org/blog/2018/06/conflicted/

Introduction to renv package: https://rstudio.github.io/renv/articles/renv.html#future-work-1

Row-oriented workflows in R with the tidyverse: https://github.com/jennybc/row-oriented-workflows#readme

Structuring R projects: https://www.r-bloggers.com/2018/08/structuring-r-projects/

Defensive Programming in R: https://bitsandbugs.io/2018/07/27/defensive-programming-in-r/#8

Nice R code: https://nicercode.github.io/blog/2013-04-05-projects/

Workflow basics: https://r4ds.had.co.nz/workflow-basics.html

Namespace package: https://r-pkgs.org/namespace.html

Writing R packages in RStudio: https://ourcodingclub.github.io/tutorials/writing-r-package/

It is dangerous to change state: https://withr.r-lib.org/articles/changing-and-restoring-state.html

The targets R Package User Manual: https://books.ropensci.org/targets/

Github and R:

Install git on the R system from here: https://git-scm.com/downloads

Go to RStudio → Global Options → Git/SVN → Make sure the box "Enable version control interface for RStudio projects" is checked

Tell RStudio where your Git executable is in the Git/SVN under Global Options

Create a new project in R (make sure the check box "Create a git repository" is checked)

Create a new task file in R (New File → Rscript) and save it as a .R file

To use Git version control on the .R file we need to commit that file

To commit a file with Git in RStudio go to the Git tab in the top right pane in R → Select one or more files by checking the box

Checking the box means that it is ready to be committed

To actually commit the file click the "Commit" button (will open up a commit window)

Include a commit message then click on the second "Commit" button

For collaboration on Github:

Load the usethis package and type in ?use_github in the R console

In the Authentication section, click on GitHub personal access token (PAT)

Click on the button to generate a new token

Put a Note and use repo permission for your token and then click on "Generate token"

Copy the token ID number (needs to be stored)

Type in edit_r_environ() in the R console and then type in GITHUB_PAT ='copy and paste token ID number here'

In R console type in use_github(protocol ='https', auth_token = Sys.getenv("GITHUB_PAT"))

Run it and will ask if you are sure. Select 3

This will create a Github repository and will set up the syncing

Another way to collaborate on Github (easier so follow this!):

Go to http://github.com and create an account

Create a new repository and give it a name (click "Add a README file)

Go to R → Install the usethis package and include library(usethis) → Type in use_git_config(user.name = "Your Name on the GitHub account", user.email = "Your email address on the GitHub account")

In the newly created repository, click the "Code" button on GitHub. Copy the URL under the "Clone with HTTPS"

Go to R → New Project → Version Control → Git → Repository URL (copy and paste the HTTP URL from your Github repository) - this will connect what's on the cloud on Github to your computer (also called cloning your repository)

Can start a new R script and would be able to see the Git tab in R

Can commit and include a commit message (will add the files to your depository)

Need to push to fully make the changes go through and to show up on your GitHub account

Under the History tab you would be able to see the changes you made and committed

Can link the SSH keys from settings on your account and into R under the Git/SVN tab (have to create a SSH RSA key if it has not been created already)

If there is a merge conflict when collaborating on making simultaneous changes together then pull first and then fix the merge conflict. Then can commit by finalizing on which changes to keep by eliminating the "====" and "»»" and push it out. The other person will have to pull in the changes in her hand.

Creating a new branch will allow you to do things on your own. Click on the branch button to create a new branch and name it. A new branch will allow you to make changes on it and work separately on it. The other person will have to pull to see the new branch and your changes on it. In this way, we can work independently when working together at the same time. Then will have to merge the independent branches.

Open a pull request by clicking on the Compare and pull request button on the Github site to merge the separate branches together. Can delete your separate branch if desired. Then go to R and pull the changes down.

For .Renviron have to use specific user credentials such as user name, password, Github and udaman tokens.

The .Rprofile can be ignored in gitignore if there is a problem with different paths across Macs and PCs.

Resources:

Happy Git and GitHub for the useR: https://happygitwithr.com/

Github for collaboration: https://github.com/llendway/github_for_collaboration/blob/master/github_for_col

My research workflow, based on Github: https://www.carlboettiger.info/2012/05/06/research-workflow.html

Collaborating with renv: https://rstudio.github.io/renv/articles/collaborating.html

R style guide: http://adv-r.had.co.nz/Style.html

UHERO R style guide:

Use block letters for R file names (because the NAS file server is case sensitive)

Comment your code

Time Series Modeling:

Forecasting: Principles and Practice (3rd ed) by Rob J Hyndman and George Athanasopoulos : https://otexts.com/fpp3/index.html

An Introduction to Statistical Learning (1st ed): https://www.statlearning.com

Manipulating Time Series Data in R with xts & zoo: https://rstudio-pubs-static.s3.amazonaws.com/288218_117 https://rpubs.com/mpfoley73/504487 Time Series in R, The Power of xts and zoo: https://ugoproto.github.io/ugo_r_doc/time_series_in_r_the_power_of_xts_and_zoo/ xts Cheat Sheet: Time Series in R: https://www.r-bloggers.com/2017/05/xts-cheat-sheet-time-series-in-r/

R For Data Science Cheat Sheet by DataCamp: https://s3.amazonaws.com/assets.datacamp.com/blog_assets/x

Evaluate the R packages: gets, ARDL, etc.

The gets package is used for Multi-path General-to-Specific (GETS) modelling of the mean and/or variance of a regression, and Indicator Saturation (ISAT) methods for detecting structural breaks in the mean.https://cran.r-project.org/web/packages/gets/index.html

The ARDL package creates complex autoregressive distributed lag (ARDL) models providing just the order and automatically constructs the underlying unrestricted and restricted error correction model (ECM). It also performs the bounds-test for cointegration as described in Pesaran et al. (2001). https://cran.r-project.org/web/packages/ARDL/index.html https://github.com/Natsiopoulos/ARDL

Tidy tools for time series modeling under tidyverts: https://tidyverts.org - The fable package applies tidyverse principles to time series modeling used for forecasting: https://fable.tidyverts.org/ - The tsibble package provides a tidy data structure for time series: https://cran.r-project.org/web/packages/tsibble/index.html - The tsibbledata package provide a different types of datasets in the tsibble data structure: https://cran.r-project.org/web/packages/tsibbledata/index.html - The tsibbletalk package introduces shared key to the tsibble, to easily {crosstalk} between plots on both client and server sides (i.e. with or without shiny): https://cran.r-project.org/web/packages/tsibbletalk/tsibbletalk.pdf- The feasts package provides a collection of features, decomposition methods, statistical summaries and graphics functions for the analysing tidy time series data: https://cran.r-project.org/web/packages/feasts/index.html - The fable.prohphet package provides an interface allowing the prophet forecasting procedure to be used within the fable framework: https://cran.r-project.org/web/packages/fable.prophet/vignettes/intro.html

The xts or Extensible Time Series package provides an extensible time series class, enabling uniform handling of many R time series classes : https://cran.r-project.org/web/packages/xts/index.html xts: Extensible Time Series: https://cran.r-project.org/web/packages/xts/vignettes/xts.pdf

Think about dummies, breaks, outliers

Figure out how bimets deals with ragged edge, add-factors, goal search

The bimets is an R package developed with the aim of easing time series analysis and building up a framework that facilitates the definition, estimation and simulation of simultaneous equation models: https://cran.r-project.org/web/packages/bimets/index.htmlbimets - Time Series And Econometric Modeling In R: https://github.com/cran/bimets https://cran.r-project.org/web/packages/bimets/vignettes/bimets.pdf

Structural Equation Models (SEM): https://rviews.rstudio.com/2021/01/22/sem-time-series-modeling/

Look at tidy models

The tidymodels package is a collection of packages for modeling and machine learning using tidyverse principles: https://www.tidymodels.org

Port the Gekko code into R: http://t-t.dk/gekko/

Look at DiagrammeR package, also the Gantt charts it can produce

https://rich-iannone.github.io/DiagrammeR/

A Beginner's Guide to Learning R:

A (very) short introduction to R: https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Rstudio Education: https://github.com/rstudio-education

Remaster the tidyverse: https://github.com/rstudio-education/remaster-the-tidyverse

Introduction to R and Rstudio: https://jules32.github.io/2016-07-12-Oxford/R_RStudio/

An intro to R for new programmers: https://rforcats.net

fasteR: Fast Lane to Learning R!: https://github.com/matloff/fasteR

RStudio Cheatsheets: https://rstudio.com/resources/cheatsheets/

R for Data Science: https://r4ds.had.co.nz

Data wrangling, exploration, and analysis with R: https://stat545.com

R Markdown: The Definitive Guide: https://bookdown.org/yihui/rmarkdown/

Data Visualization with R: https://rkabacoff.github.io/datavis/

Modern R with the tidyverse: https://b-rodrigues.github.io/modern_R/

R Cookbook, 2nd Edition: https://rc2e.com

Advanced R by Hadley Wickham: http://adv-r.had.co.nz

UC Business Analytics R Programming Guide: http://uc-r.github.io/descriptive

R Programming for Data Science: https://bookdown.org/rdpeng/rprogdatascience/

Hands-On Programming with R: https://rstudio-education.github.io/hopr/

Efficient R programming: https://csgillespie.github.io/efficientR/index.html

R for Fledglings: http://www.uvm.edu/~tdonovan/RforFledglings/index.html

R Intermediate Level (includes applications):

Advanced Statistical Computing: https://bookdown.org/rdpeng/advstatcomp/

Feature Engineering and Selection: A Practical Approach for Predictive Models: http://www.feat.engineering/index.html

Advanced Quantitative Methods: https://uclspp.github.io/PUBLG088/index.html

Principles of Econometrics with R: https://bookdown.org/ccolonescu/RPoE4/

Modern Data Analysis for Economics: https://jiamingmao.github.io/data-analysis/Resources/

Data Science for Economists: https://github.com/uo-ec607/lectures

Data Science for Psychologists: https://bookdown.org/hneth/ds4psy/10-time.html

Rewriting R code in C++: https://adv-r.hadley.nz/rcpp.html

Writing R Extensions: https://cran.rstudio.com/doc/manuals/r-devel/R-exts.html

Other R packages for data analysis:

The data.table package is used for fast aggregation of large data (e.g. 100GB in RAM), fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns, friendly and fast character- separated-value read/write: https://cran.r-project.org/web/packages/data.table/

The mlr3 (Lang et al. 2019) package and ecosystem provide a generic, object-oriented, and extensible framework for classification, regression, survival analysis, and other machine learning tasks for the R: https://mlr3book.mlr-org.com

purr package tutorial: https://jennybc.github.io/purrr-tutorial/

Data Visualization with R:

Data Analysis and Visualization Using R: http://varianceexplained.org/RData/

Data Analysis and Visualization in R for Ecologists: https://datacarpentry.org/R-ecology-lesson/

Data Visualization with R by Rob Kabacoff: https://rkabacoff.github.io/datavis/

R Graphics Cookbook, 2nd edition: https://r-graphics.org

ggplot2: elegant graphics for data analysis: https://ggplot2-book.org