Project: The Forecasting Tourism 2010 Competition

EM1415

Marco Solari, 875475

Table of contents

1 Setup and Data Loading

1.1 Setup

```r
knitr::opts_chunk$set(
  echo = T,
  dev = "cairo_pdf"
)

libraries_list <- c(
  "tidyverse",
  "fpp3"
)

lapply(
  libraries_list,
  require,
  character.only = TRUE
)
```

```
Loading required package: tidyverse

-- Attaching core tidyverse packages ———————————————————————— tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
```

```
-- Conflicts ──────────────────────────────────────────── tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
Loading required package: fpp3

-- Attaching packages ──────────────────────────────────────────────── fpp3 0.5 --

v tsibble     1.1.3     v fable       0.3.3
v tsibbledata 0.4.1     v fabletools  0.3.3
v feasts      0.3.1

-- Conflicts ───────────────────────────────────────────────── fpp3_conflicts --
x lubridate::date()    masks base::date()
x dplyr::filter()      masks stats::filter()
x tsibble::intersect() masks base::intersect()
x tsibble::interval()  masks lubridate::interval()
x dplyr::lag()         masks stats::lag()
x tsibble::setdiff()   masks base::setdiff()
x tsibble::union()     masks base::union()

[[1]]
[1] TRUE

[[2]]
[1] TRUE
```

```r
theme_set(
  ggthemes::theme_tufte(
    base_size = 16,
    base_family = "Atkinson Hyperlegible"
  )
)
```

1.2 Loading Data

```r
data_main <- readr::read_csv(
  "Data/tourism_data.csv",
  show_col_types = F
)
```

```r
data_main %>% dim
```

```
[1]  43 518
```

```r
data_main %>% is.na() %>% sum
```

```
[1] 11668
```

1.3 Creating **tsibble**

```
tourism_full <- data_main %>%
  mutate(
    Year = 1965:2007
  ) %>%
  as_tsibble(
    index = Year
  )
```

```
tourism_full %>% interval
```

```
<interval[1]>
[1] 1Y
```

2 Assignment

2.1 Full Plot

> *Plot all the series (an advanced data visualization tool is recommended) - what type of components are visible? Are the series similar or different? Check for problems such as missing values and possible errors.*

```
tmelt <- reshape2::melt(tourism_full,id="Year")
```

```
tmelt %>%
  ggplot(
    aes(
      x = Year,
      y = value,
      colour = variable,
      group = variable
      )
    ) +
  geom_line(
    alpha = .8
  ) +
  scale_y_log10() +
  scale_color_viridis_d(
    option = "rocket"
  ) +
  theme(
    legend.position = "none"
    )
```

```
Warning: Removed 11668 rows containing missing values (`geom_line()`).
```

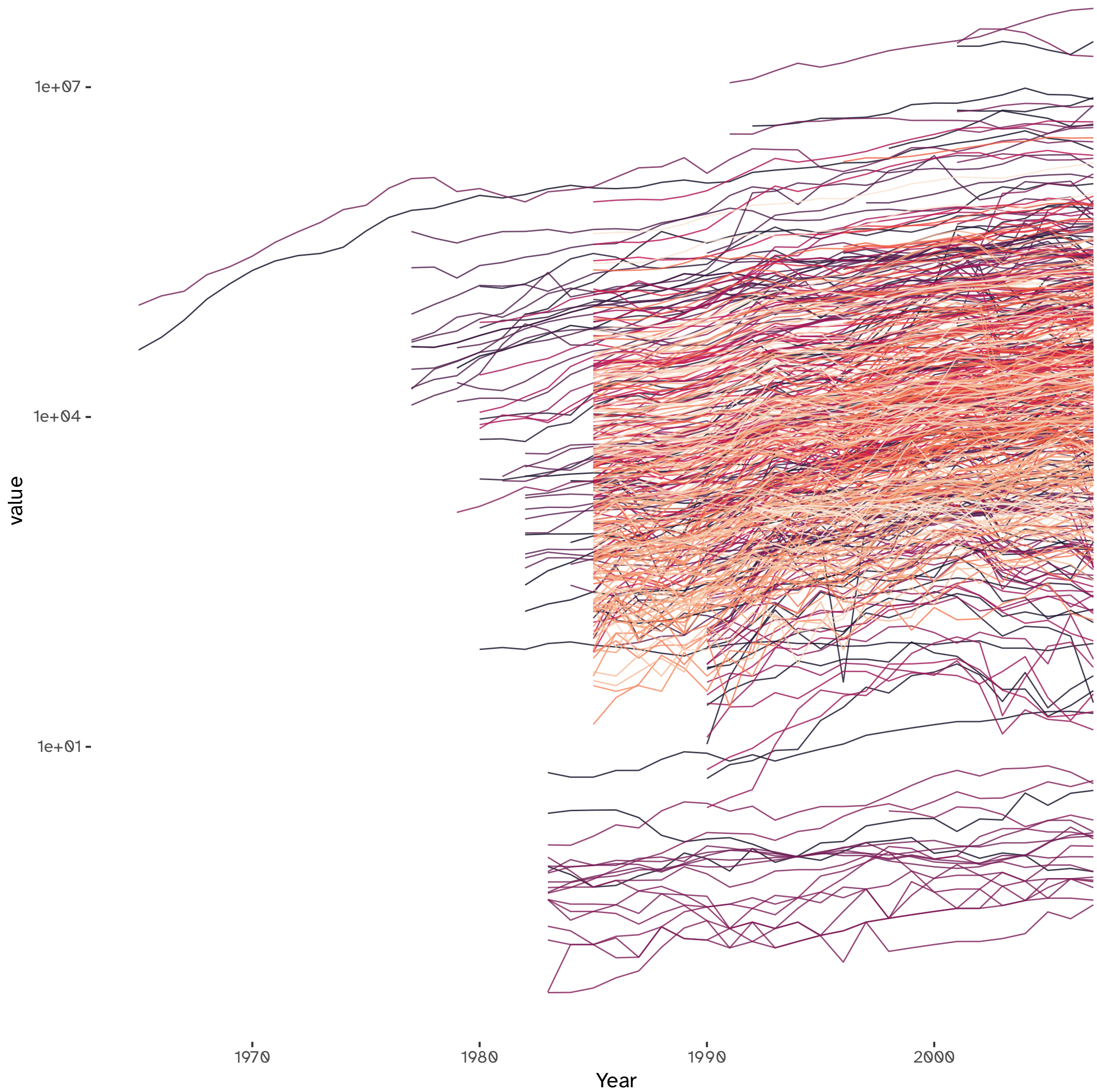

Figure 1: Printing a legend for 518 different series is not possible. However, color has been used only to differentiate the series and does not contain further information. Plotting the y-axis variable on the log scale was made necessary by the huge variation in the series values.

A check for NAs has already been made while loading data and it showed the presence of a large number of missing values. This is mostly to be attributed to the different starting time of the series. From the visualization we can definitely group different starting points: this suggests that another approach to visualisation might be successfully conducted by grouping series by their starting date.

Plotting all 518 series does not allow to spot details, such as the presence of seasonal patterns. However, a general upward trend is clear; moreover, we can spot some outliers and some clues about the presence of cyclicality in some of the series.

2.2 Creating Validation Set

> *Partition the series into training and validation, so that the last 4 years are in the validation period for each series. What is the logic of such a partitioning? What is the disadvantage?*

```r
tourism_train <- tourism_full %>%
  filter(Year < 2004)
tourism_validation <- tourism_full %>%
  filter(Year >= 2004)
```

2.3 Naïve Forecasts

Generate naïve forecasts for all series for the validation period. For each series, create forecasts with horizons of 1, 2, 3, and 4 years ahead ($F_{t+1}, F_{t+2}, F_{t+3}$, and F_{t+4}).

2.4 Choosing Measures

Which measures are suitable if we plan to combine the results for the 518 series? Consider MAE, Average error, MAPE and RMSE.

2.5 Computing MAPE

For each series, compute MAPE of the naive forecasts once for the training period and once for the validation period.

2.6 Computing MASE

The performance measure used in the competition is Mean Absolute Scaled Error (MASE). Explain the advantage of MASE and compute the training and validation MASE for the naive forecasts.

2.7 MAPE Pairs

Create a scatter plot of the MAPE pairs, with the training MAPE on the x-axis and the validation MAPE on the y-axis. Create a similar scatter plot for the MASE pairs. Now examine both plots. What do we learn? How does performance differ between the training and validation periods? How does performance range across series?

3 Ensemble Methods

The competition winner, Lee Baker, used an ensemble of three methods:

- *Naive forecasts multiplied by a constant trend*[1].
- Linear regression.
- Exponentially-weighted linear regression.

a. *Write the exact formula used for generating the first method, in the form* $F_{t+k} = ...(k = 1, 2, 3, 4)$,

b. *What is the rational behind multiplying the naive forecasts by a constant?*[2]

c. *What should be the dependent variable and the predictors in a linear regression model for this data? Explain.*

d. *Fit the linear regression model to the first five series and compute forecast errors for the validation period.*

e. *Before choosing a linear regression, the winner described the following process:*

[1] Global/local trend: "globally tourism has grown "at a rate of 6% annually."

[2] Hint: think empirical and domain knowledge.

”I examined fitting a polynomial line to the data and using the line to predict future values. I tried using first through fifth order polynomials to find that the lowest MASE was obtained using a first order polynomial (simple regression line). This best fit line was used to predict future values. I also kept the R^2 value of the fit for use in blending the results of the prediction.”

What are two flaws in this approach?

f. *If we were to consider exponential smoothing, what particular type(s) of exponential smoothing are reasonable candidates?*

g. *The winner concludes with possible improvements one being ”an investigation into how to come up with a blending ensemble method that doesn’t use much manual twerking would also be of benefit”. Can you suggest methods or an approach that would lead to easier automation of the ensemble step?*

h. *The competition focused on minimizing the average MAPE of the next four values across all 518 series. How does this goal differ from goals encountered in practice when considering tourism demand? Which steps in the forecasting process would likely be different in a real-life tourism forecasting scenario?*