MONASH University

MONASH
BUSINESS
SCHOOL

Forecasting Competitions

Rob J Hyndman

30 November 2018

Outline

1 Who’s the best forecaster?

2 Makridakis competitions

3 Feature-based forecasting

4 Kaggle and other competition platforms

5 Forecasting case studies

Outline

1. Who's the best forecaster?
2. Makridakis competitions
3. Feature-based forecasting
4. Kaggle and other competition platforms
5. Forecasting case studies

Who’s the best forecaster?

Forecast each of the following things:

1. Google closing stock price on 12 December 2018.
2. The seasonally adjusted estimate of total employment for Australia in November 2018.
3. The total runs scored on Boxing Day 2018 at the MCG.

For each of these, give a point forecast and an 80% prediction interval.

Who’s the best forecaster?

Forecast each of the following things:

1. Google closing stock price on 12 December 2018.
2. The seasonally adjusted estimate of total employment for Australia in November 2018.
3. The total runs scored on Boxing Day 2018 at the MCG.

For each of these, give a point forecast and an 80% prediction interval.

- How will we measure best point forecast?
- How will we measure best interval forecast?

Outline

1 Who’s the best forecaster?

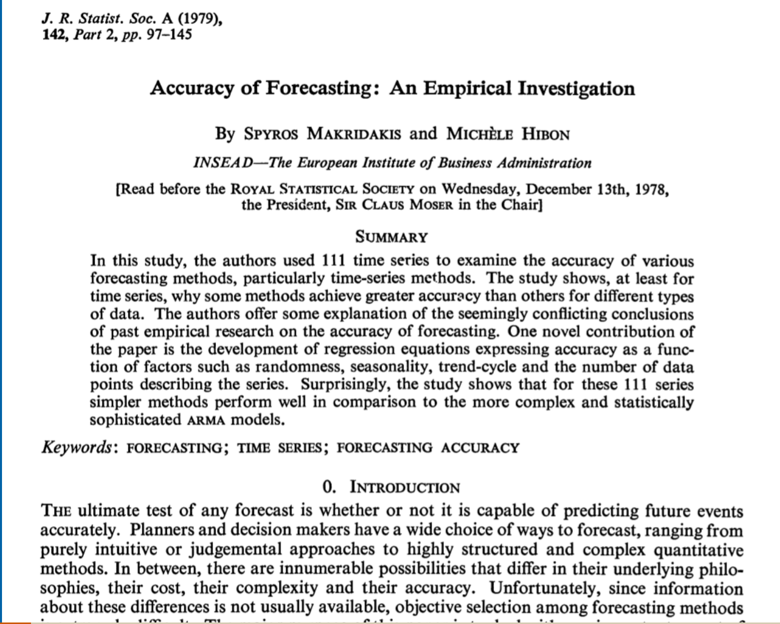
2 Makridakis competitions

3 Feature-based forecasting

4 Kaggle and other competition platforms

5 Forecasting case studies

Makridakis and Hibon (1979)



J. R. Statist. Soc. A (1979),
142, *Part* 2, *pp.* 97–145

Accuracy of Forecasting: An Empirical Investigation

By Spyros Makridakis and Michèle Hibon

INSEAD—The European Institute of Business Administration

[Read before the Royal Statistical Society on Wednesday, December 13th, 1978, the President, Sir Claus Moser in the Chair]

Summary

In this study, the authors used 111 time series to examine the accuracy of various forecasting methods, particularly time-series methods. The study shows, at least for time series, why some methods achieve greater accuracy than others for different types of data. The authors offer some explanation of the seemingly conflicting conclusions of past empirical research on the accuracy of forecasting. One novel contribution of the paper is the development of regression equations expressing accuracy as a function of factors such as randomness, seasonality, trend-cycle and the number of data points describing the series. Surprisingly, the study shows that for these 111 series simpler methods perform well in comparison to the more complex and statistically sophisticated ARMA models.

Keywords: FORECASTING; TIME SERIES; FORECASTING ACCURACY

0. Introduction

The ultimate test of any forecast is whether or not it is capable of predicting future events accurately. Planners and decision makers have a wide choice of ways to forecast, ranging from purely intuitive or judgemental approaches to highly structured and complex quantitative methods. In between, there are innumerable possibilities that differ in their underlying philosophies, their cost, their complexity and their accuracy. Unfortunately, since information about these differences is not usually available, objective selection among forecasting methods

Makridakis and Hibon (1979)

J. R. Statist. Soc. A (1979),
142, *Part* 2, *pp.* 97–145

Accuracy of Forecasting: An Empirical Investigation

By Spyros Makridakis and Michèle Hibon

INSEAD—The European Institute of Business Administration

[Read before the Royal Statistical Society on Wednesday, December 13th, 1978,
resident, Sir Claus Moser in the Chair]

Summary

used 111 time series to examine the
icularly time-series methods. The stud
hods achieve greater accuracy than othe
some explanation of the seemingly co
on the accuracy of forecasting. One n
ent of regression equations expressing
ndomness, seasonality, trend-cycle and
s. Surprisingly, the study shows that
well in comparison to the more comp
s.

SERIES; FORECASTING ACCURACY

0. Introduction

The ultimate test of any forecast is whether or not it is capable of predicting future events accurately. Planners and decision makers have a wide choice of ways to forecast, ranging from purely intuitive or judgemental approaches to highly structured and complex quantitative methods. In between, there are innumerable possibilities that differ in their underlying philosophies, their cost, their complexity and their accuracy. Unfortunately, since information about these differences is not usually available, objective selection among forecasting methods

Makridakis and Hibon (1979)

This was the first large-scale empirical evaluation of time series forecasting methods.

- Highly controversial at the time.
- Difficulties:
  - How to measure forecast accuracy?
  - How to apply methods consistently and objectively?
  - How to explain unexpected results?
- Common thinking was that the more sophisticated mathematical models (ARIMA models at the time) were necessarily better.

Makridakis and Hibon (1979)

> I do not believe that it is very fruitful to attempt to classify series according to which forecasting techniques perform "best". The performance of any particular technique when applied to a particular series depends essentially on (a) the model which the series obeys; (b) our ability to identify and fit this model correctly and (c) the criterion chosen to measure the forecasting accuracy. — *M.B. Priestley*

Makridakis and Hibon (1979)

I do not believe that it is very fruitful to attempt to classify series according to which forecasting techniques perform "best". The performance of any particular technique when applied to a particular series depends essentially on (a) the model which the series obeys; (b) our ability to identify and fit this model correctly and (c) the criterion chosen to measure the forecasting accuracy. — *M.B. Priestley*

… the paper suggests the application of normal scientific experimental design to forecasting, with measures of unbiased testing of forecasts against subsequent reality, for success or failure. A long overdue reform.

— *F.H. Hansford-Miller*

Makridakis and Hibon (1979)

> Modern man is fascinated with the subject of forecasting — *W.G. Gilchrist*

Makridakis and Hibon (1979)

> Modern man is fascinated with the subject of forecasting — *W.G. Gilchrist*

> It is amazing to me, however, that after all this exercise in identifying models, transforming and so on, that the autoregressive moving averages come out so badly. I wonder whether it might be partly due to the authors not using the backwards forecasting approach to obtain the initial errors. — *W.G. Gilchrist*

Makridakis and Hibon (1979)

> I find it hard to believe that Box-Jenkins, if properly applied, can actually be worse than so many of the simple methods — *C. Chatfield*

Makridakis and Hibon (1979)

I find it hard to believe that Box-Jenkins, if properly applied, can actually be worse than so many of the simple methods — *C. Chatfield*

Why do empirical studies sometimes give different answers? It may depend on the selected sample of time series, but I suspect it is more likely to depend on the skill of the analyst and on their individual interpretations of what is meant by Method *X*. — *C. Chatfield*

Makridakis and Hibon (1979)

> I find it hard to believe that Box-Jenkins, if properly applied, can actually be worse than so many of the simple methods — *C. Chatfield*

> Why do empirical studies sometimes give different answers? It may depend on the selected sample of time series, but I suspect it is more likely to depend on the skill of the analyst and on their individual interpretations of what is meant by Method *X*. — *C. Chatfield*

> … these authors are more at home with simple procedures than with Box-Jenkins. — *C. Chatfield*

Makridakis and Hibon (1979)

> There is a fact that Professor Priestley must accept: empirical evidence is in *disagreement* with his theoretical arguments. — *S. Makridakis & M. Hibon*

Makridakis and Hibon (1979)

> There is a fact that Professor Priestley must accept: empirical evidence is in *disagreement* with his theoretical arguments. — *S. Makridakis & M. Hibon*

> Dr Chatfield expresses some personal views about the first author ... It might be useful for Dr Chatfield to read some of the psychological literature quoted in the main paper, and he can then learn a little more about biases and how they affect prior probabilities. — *S. Makridakis & M. Hibon*

Consequences of M&H (1979)

As a result of this paper, researchers started to:

- consider how to automate forecasting methods;
- study what methods give the best forecasts;
- be aware of the dangers of over-fitting;
- treat forecasting as a different problem from time series analysis.

Consequences of M&H (1979)

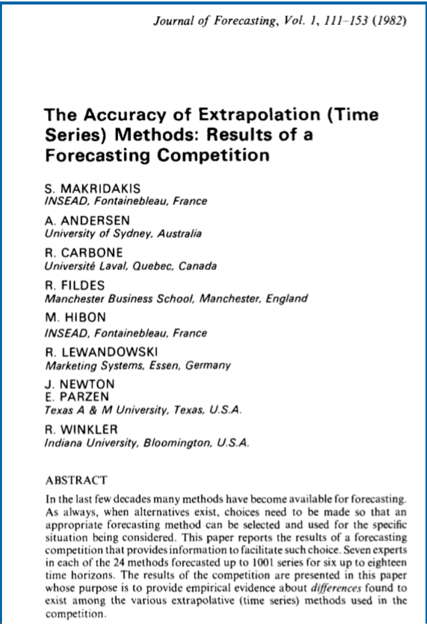
As a result of this paper, researchers started to:

- consider how to automate forecasting methods;
- study what methods give the best forecasts;
- be aware of the dangers of over-fitting;
- treat forecasting as a different problem from time series analysis.

Makridakis & Hibon followed up with a new competition in 1982:

- 1001 series from demography, industry, economics. Annual, quarterly, monthly.
- Anyone could submit forecasts (avoiding the charge of incompetence)

M-competition



Journal of Forecasting, Vol. 1, 111–153 (1982)

The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition

S. MAKRIDAKIS
INSEAD, Fontainebleau, France

A. ANDERSEN
University of Sydney, Australia

R. CARBONE
Université Laval, Quebec, Canada

R. FILDES
Manchester Business School, Manchester, England

M. HIBON
INSEAD, Fontainebleau, France

R. LEWANDOWSKI
Marketing Systems, Essen, Germany

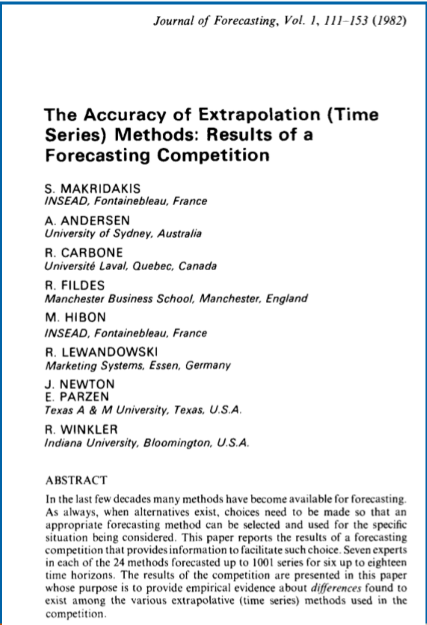
J. NEWTON
E. PARZEN
Texas A & M University, Texas, U.S.A.

R. WINKLER
Indiana University, Bloomington, U.S.A.

ABSTRACT

In the last few decades many methods have become available for forecasting. As always, when alternatives exist, choices need to be made so that an appropriate forecasting method can be selected and used for the specific situation being considered. This paper reports the results of a forecasting competition that provides information to facilitate such choice. Seven experts in each of the 24 methods forecasted up to 1001 series for six up to eighteen time horizons. The results of the competition are presented in this paper whose purpose is to provide empirical evidence about *differences* found to exist among the various extrapolative (time series) methods used in the competition.

M-competition



Journal of Forecasting, Vol. 1, 111–153 (1982)

The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition

S. MAKRIDAKIS
INSEAD, Fontainebleau, France

A. ANDERSEN
University of Sydney, Australia

R. CARBONE
Université Laval, Quebec, Canada

R. FILDES
Manchester Business School, Manchester, England

M. HIBON
INSEAD, Fontainebleau, France

R. LEWANDOWSKI
Marketing Systems, Essen, Germany

J. NEWTON
E. PARZEN
Texas A & M University, Texas, U.S.A.

R. WINKLER
Indiana University, Bloomington, U.S.A.

ABSTRACT

In the last few decades many methods have become available for forecasting. As always, when alternatives exist, choices need to be made so that an appropriate forecasting method can be selected and used for the specific situation being considered. This paper reports the results of a forecasting competition that provides information to facilitate such choice. Seven experts in each of the 24 methods forecasted up to 1001 series for six up to eighteen time horizons. The results of the competition are presented in this paper whose purpose is to provide empirical evidence about *differences* found to exist among the various extrapolative (time series) methods used in the competition.

Best method: DSES

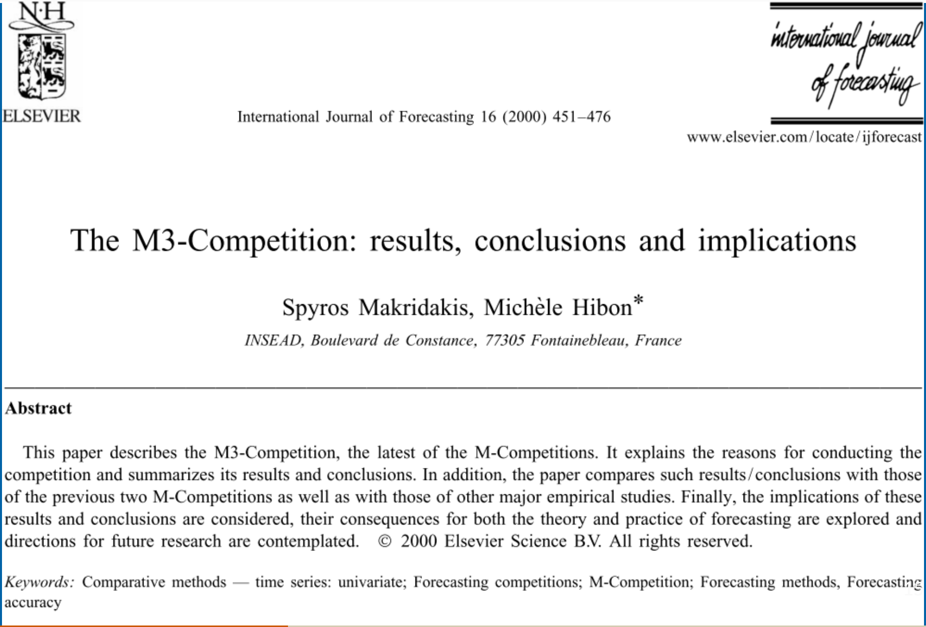
Classical multiplicative seasonal decomposition + Simple exponential smoothing applied to seasonally adjusted data, then reseasonalized.

M-competition

Main findings (taken from Makridakis & Hibon, 2000)

1. Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.
2. The relative ranking of the performance of the various methods varies according to the accuracy measure being used.
3. The accuracy when various methods are being combined outperforms, on average, the individual methods being combined and does very well in comparison to other methods.

M3 competition



ELSEVIER

International Journal of Forecasting 16 (2000) 451–476

international journal of forecasting

www.elsevier.com/locate/ijforecast

The M3-Competition: results, conclusions and implications

Spyros Makridakis, Michèle Hibon*

INSEAD, Boulevard de Constance, 77305 Fontainebleau, France

Abstract

This paper describes the M3-Competition, the latest of the M-Competitions. It explains the reasons for conducting the competition and summarizes its results and conclusions. In addition, the paper compares such results/conclusions with those of the previous two M-Competitions as well as with those of other major empirical studies. Finally, the implications of these results and conclusions are considered, their consequences for both the theory and practice of forecasting are explored and directions for future research are contemplated. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Comparative methods — time series: univariate; Forecasting competitions; M-Competition; Forecasting methods, Forecasting accuracy

M3 competition

> “The M3-Competition is a final attempt by the authors to settle the accuracy issue of various time series methods… The extension involves the inclusion of more methods/ researchers (in particular in the areas of neural networks and expert systems) and more series.”

- 3003 series
- All data from business, demography, finance and economics.
- Series length between 14 and 126.
- Either non-seasonal, monthly or quarterly.
- All time series positive.
- M&H claimed that the M3-competition supported the findings of their earlier work.

M3 results (recalculated)

Method	MAPE	sMAPE	MASE
Theta	17.42	12.76	1.39
ForecastPro	18.00	13.06	1.47
ForecastX	17.35	13.09	1.42
Automatic ANN	17.18	13.98	1.53
B-J automatic	19.13	13.72	1.54

M3 results (recalculated)

Method	MAPE	sMAPE	MASE
Theta	17.42	12.76	1.39
ForecastPro	18.00	13.06	1.47
ForecastX	17.35	13.09	1.42
Automatic ANN	17.18	13.98	1.53
B-J automatic	19.13	13.72	1.54

- Calculations do not match published paper.
- Some contestants apparently submitted multiple entries but only best ones published.

M3 competition

Theta

- A very confusing explanation.
- Shown by Hyndman and Billah (2003) to be average of linear regression and simple exponential smoothing with drift, applied to seasonally adjusted data.
- Later, the original authors claimed that their explanation was incorrect.

Forecast Pro

- A commercial software package with an unknown algorithm.
- Known to fit either exponential smoothing or ARIMA models using BIC.

M4 competition

- January – May 2018
- 100,000 time series: yearly, quarterly, monthly, weekly, daily, hourly.
- Point forecast and prediction intervals assessed.
- Code must be public
- 248 registrations, 50 submissions.

M4 competition

- January – May 2018
- 100,000 time series: yearly, quarterly, monthly, weekly, daily, hourly.
- Point forecast and prediction intervals assessed.
- Code must be public
- 248 registrations, 50 submissions.

Winning methods

1. Hybrid of Recurrent Neural Network and Exponential Smoothing models
2. Forecast combination using xgboost to find weights

Outline

1 Who’s the best forecaster?

2 Makridakis competitions

3 Feature-based forecasting

4 Kaggle and other competition platforms

5 Forecasting case studies

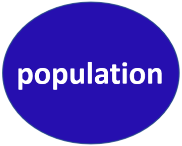
Forecast model selection

Features used to select a forecasting model

- length
- strength of seasonality
- strength of trend
- linearity
- curvature
- spikiness
- stability
- lumpiness
- first ACF value of remainder series
- parameter estimates of Holt's linear trend method
- spectral entropy
- Hurst exponent
- nonlinearity
- parameter estimates of Holt-Winters' additive method
- unit root test statistics
- first ACF value of residual series of linear trend model
- ACF and PACF based features

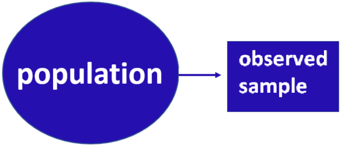
- calculated on both the raw and differenced series

FFORMS: Feature-based FORecast Model Selection

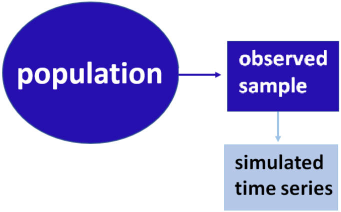


population

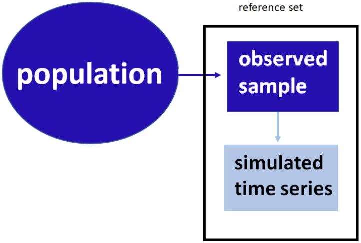
FFORMS: Feature-based FORecast Model Selection



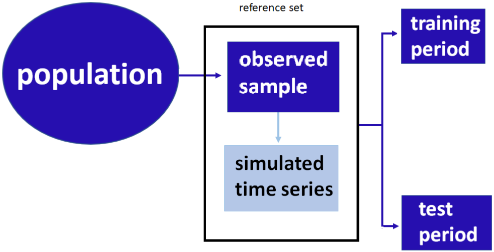
FFORMS: Feature-based FORecast Model Selection



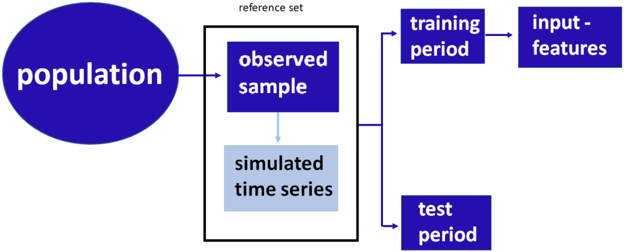
FFORMS: Feature-based FORecast Model Selection



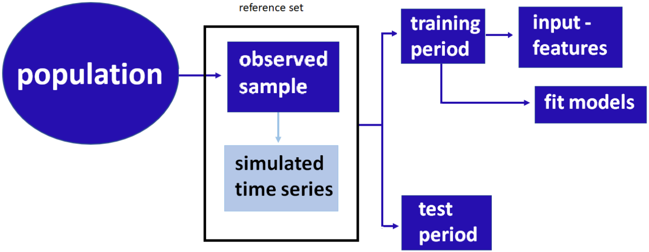
FFORMS: Feature-based FORecast Model Selection



FFORMS: Feature-based FORecast Model Selection



FFORMS: Feature-based FORecast Model Selection



population

reference set

observed sample

simulated time series

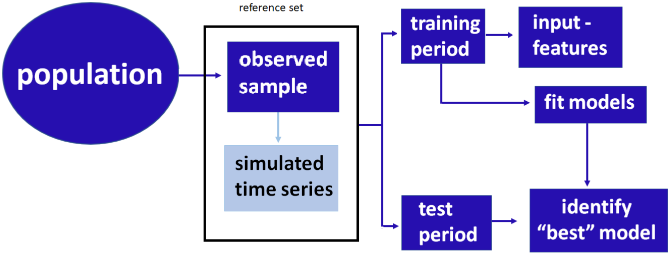
training period

input - features

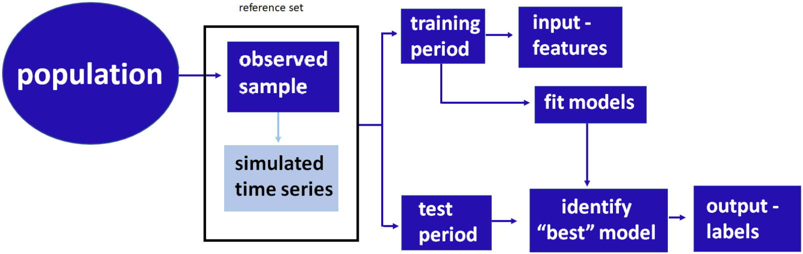
fit models

test period

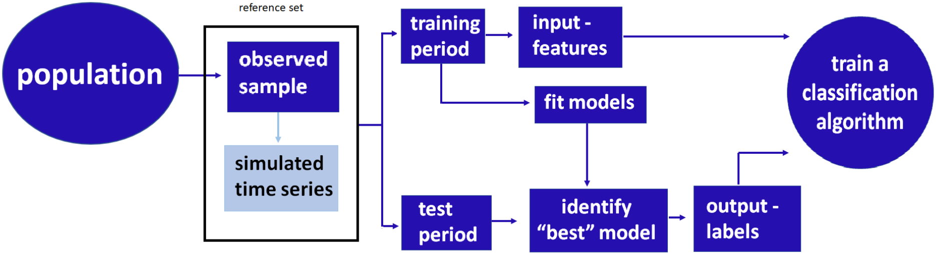
FFORMS: Feature-based FORecast Model Selection



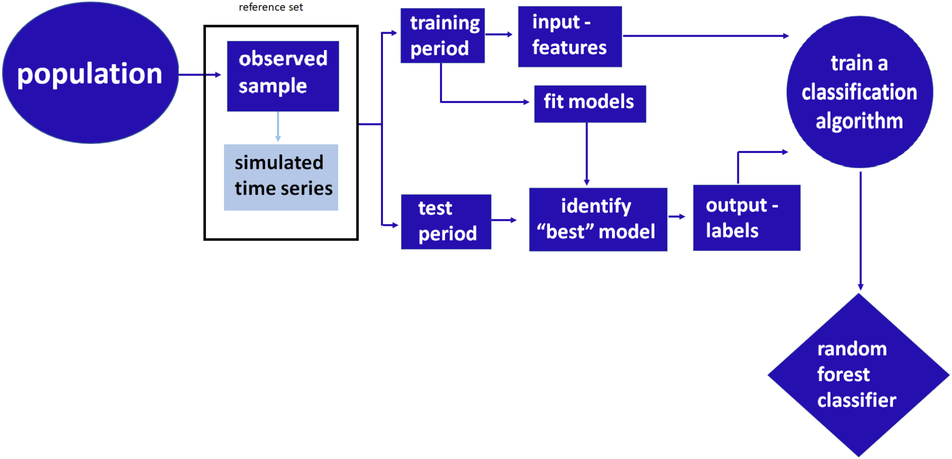
FFORMS: Feature-based FORecast Model Selection



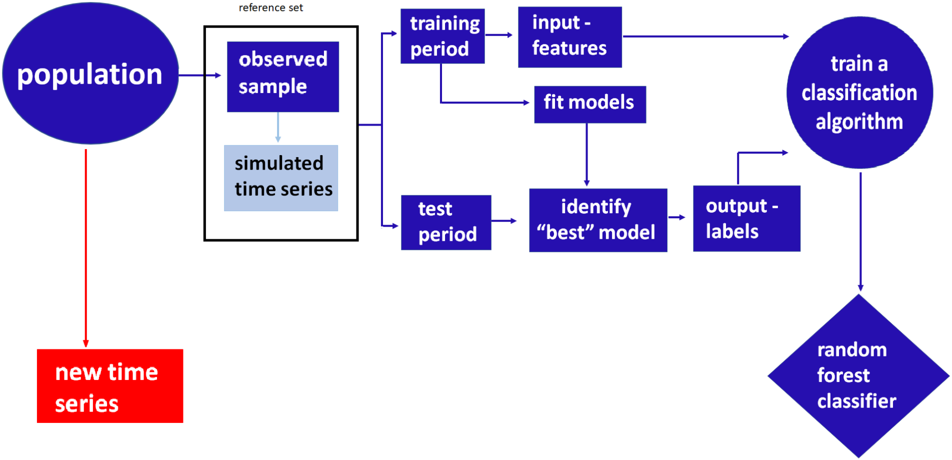
FFORMS: Feature-based FORecast Model Selection



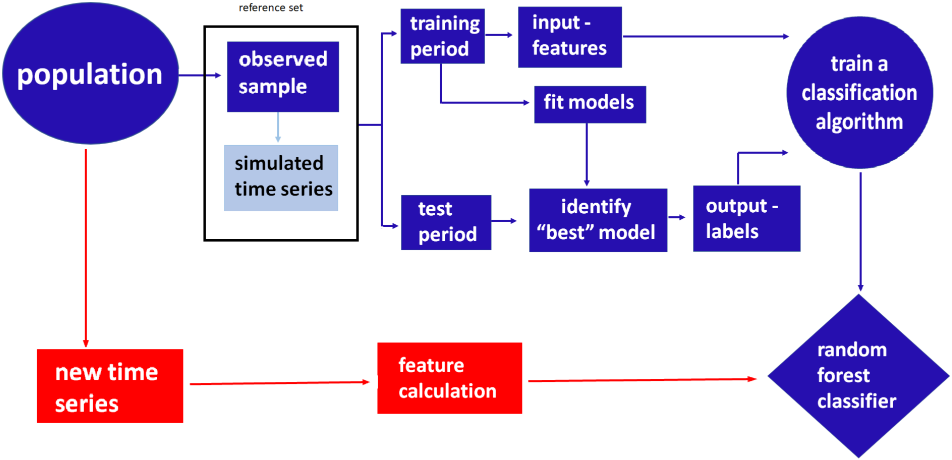
FFORMS: Feature-based FORecast Model Selection



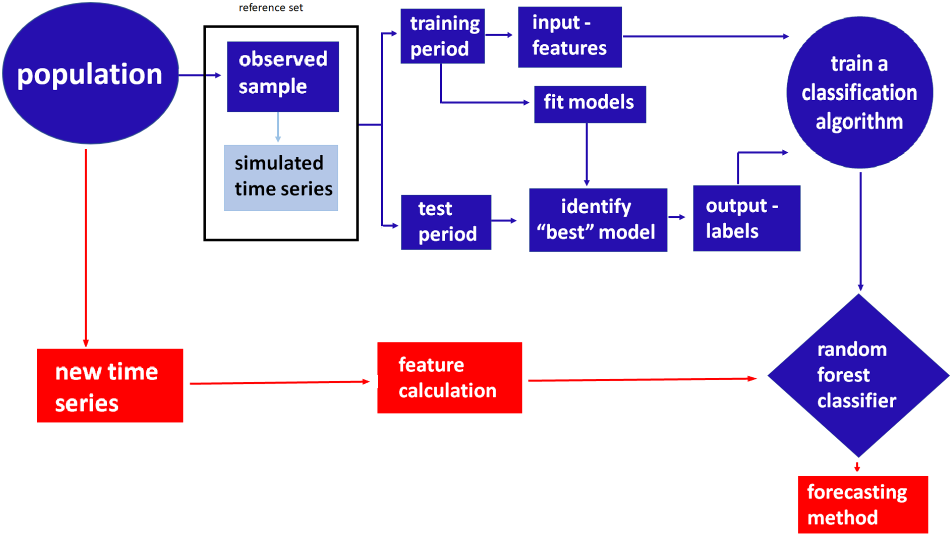
FFORMS: Feature-based FORecast Model Selection



FFORMS: Feature-based FORecast Model Selection



FFORMS: Feature-based FORecast Model Selection



Application to M competition data

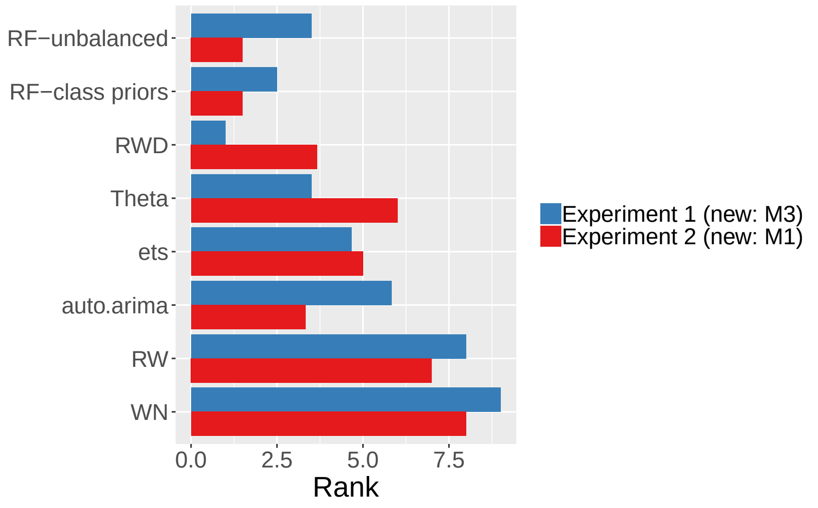
Experiment 1

	Source	Y	Q	M
Observed series	M1	181	203	617
Simulated series		362000	406000	123400
New series	M3	645	756	1428

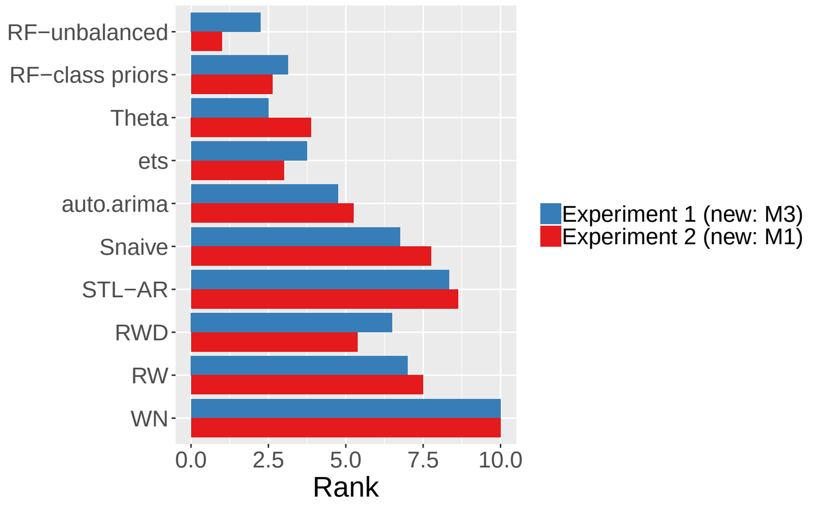
Experiment 2

	Source	Y	Q	M
Observed series	M3	645	756	1428
Simulated series		1290000	1512000	285600
New series	M1	181	203	617

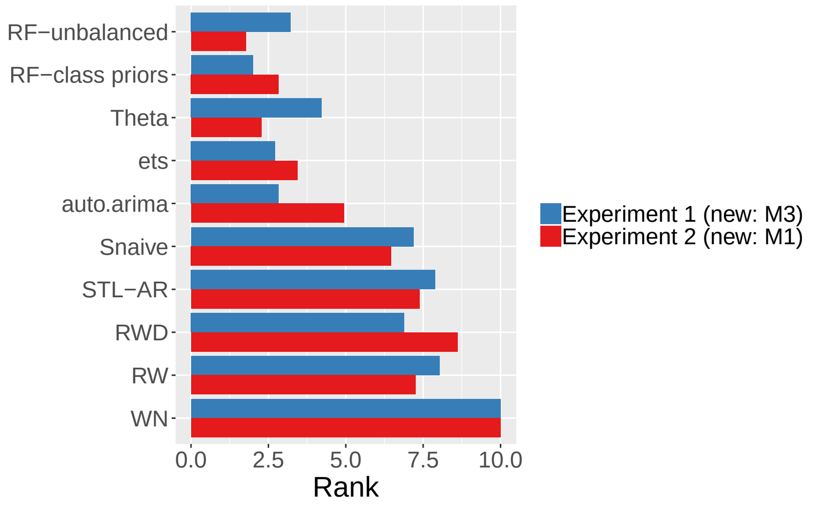
Results: Yearly



Results: Quarterly



Results: Monthly



FFORMA: Feature-based FORecast Model Averaging

- Like FFORMS but we use gradient boosted trees rather than a random forest.
- The optimization criterion is forecast accuracy not classification accuracy.
- The probability of each model being best is used to construct a model weight.
- A combination forecast is produced using these weights.
- **Came second in the M4 forecasting competition**

FFORMA: Feature-based FORecast Model Averaging

Models included

1. Naive
2. Seasonal naive
3. Random walk with drift
4. Theta method
5. ARIMA
6. ETS
7. TBATS
8. STLM-AR

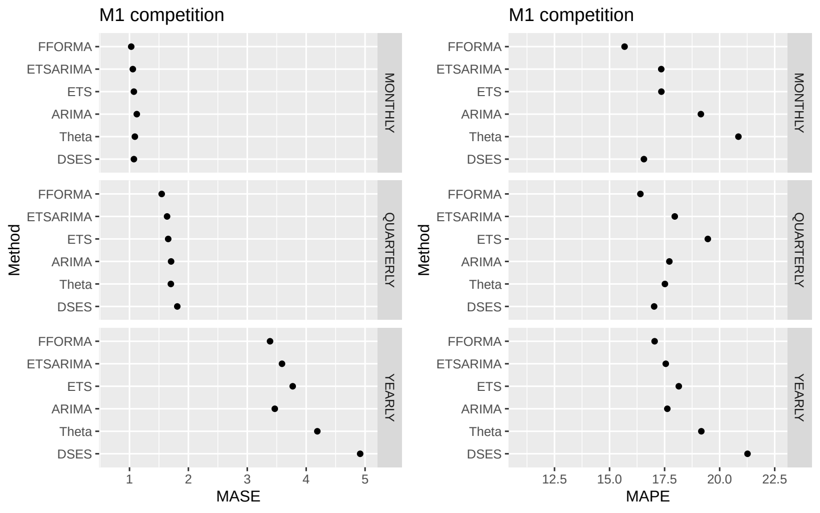
Are we getting better at forecasting?

	M1 competition			M3 competition		
Method	MAPE	sMAPE	MASE	MAPE	sMAPE	MASE
FFORMA	15.9	14.4	1.28	18.4	12.6	1.11
ETSARIMA	17.4	15.3	1.32	18.7	13.1	1.13
ETS	17.7	15.6	1.35	18.7	13.3	1.16
ARIMA	18.9	16.3	1.38	19.8	14.0	1.18
Theta	20.3	16.8	1.41	17.9	13.1	1.16
DSES	17.0	15.4	1.46	19.2	13.9	1.31

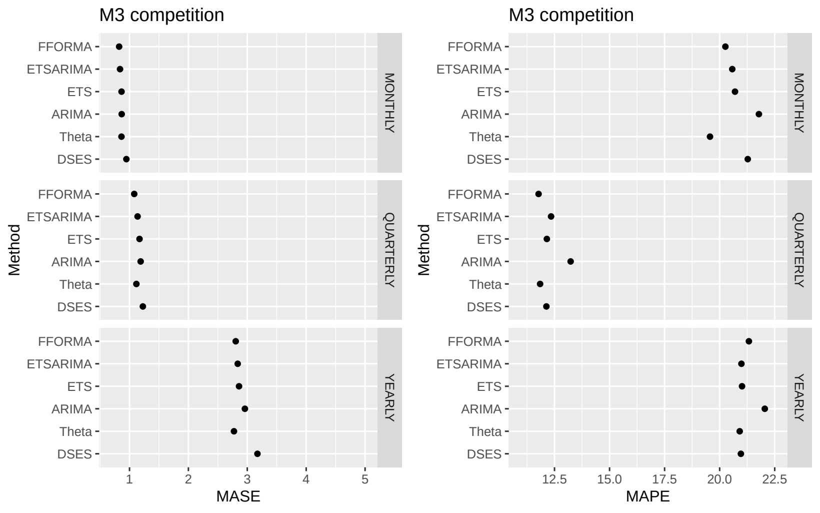
Are we getting better at forecasting?

- DSES did well as measured by MAPE and sMAPE on the M1 data, but very poorly everywhere else.
- While Theta did quite well on the M3 data, it performed poorly on the M1 data.
- FFORMA outperforms the other methods on every measure for the M1 competition, and on all but MAPE for the M3 competition.
- ETSARIMA is almost as good as FFORMA, and is much easier and faster to compute.

Are we getting better at forecasting?



Are we getting better at forecasting?



Outline

1 Who’s the best forecaster?

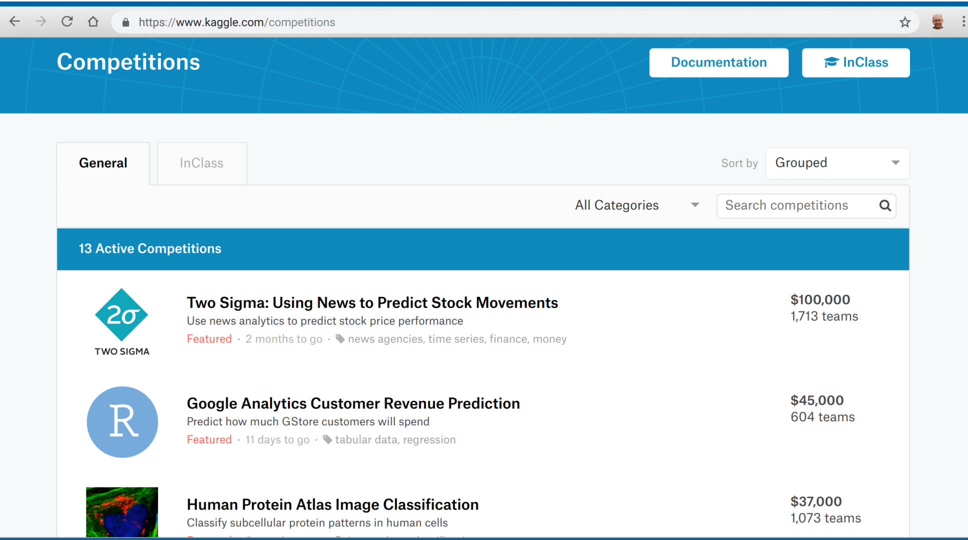
2 Makridakis competitions

3 Feature-based forecasting

4 Kaggle and other competition platforms

5 Forecasting case studies

Kaggle



https://www.kaggle.com/competitions

Competitions

13 Active Competitions

Two Sigma: Using News to Predict Stock Movements
Use news analytics to predict stock price performance
Featured · 2 months to go · news agencies, time series, finance, money
$100,000
1,713 teams

Google Analytics Customer Revenue Prediction
Predict how much GStore customers will spend
Featured · 11 days to go · tabular data, regression
$45,000
604 teams

Human Protein Atlas Image Classification
Classify subcellular protein patterns in human cells
$37,000
1,073 teams

Kaggle and Melbourne

- Started in Melbourne in 2010 by Anthony Goldbloom.
- 2010 tourism forecasting competition won by Jeremy Howard (also from Melbourne)
- Jeremy was Kaggle Chief Scientist, 2011–2013
- Chess rankings competition won by Alec Stephenson (CSIRO, Melbourne) in 2012
- Heritage health competition $500,000 prize winners included Phil Brierley (Melbourne consultant) in 2013.
- Kaggle was purchased by Google in 2017

Kaggle rip-offs

- ChallengerAI
- CrowdAI
- CrowdAnalytix
- DataFountain
- DrivenData

Outline

1 Who’s the best forecaster?

2 Makridakis competitions

3 Feature-based forecasting

4 Kaggle and other competition platforms

5 Forecasting case studies

CASE STUDY 1: Paperware company

Methods currently used

A 12 month average

C 6 month average

E straight line regression over last 12 months

G straight line regression over last 6 months

H average slope between last year's and this year's values. (Equivalent to differencing at lag 12 and taking mean.)

I Same as H except over 6 months.

K I couldn't understand the explanation.

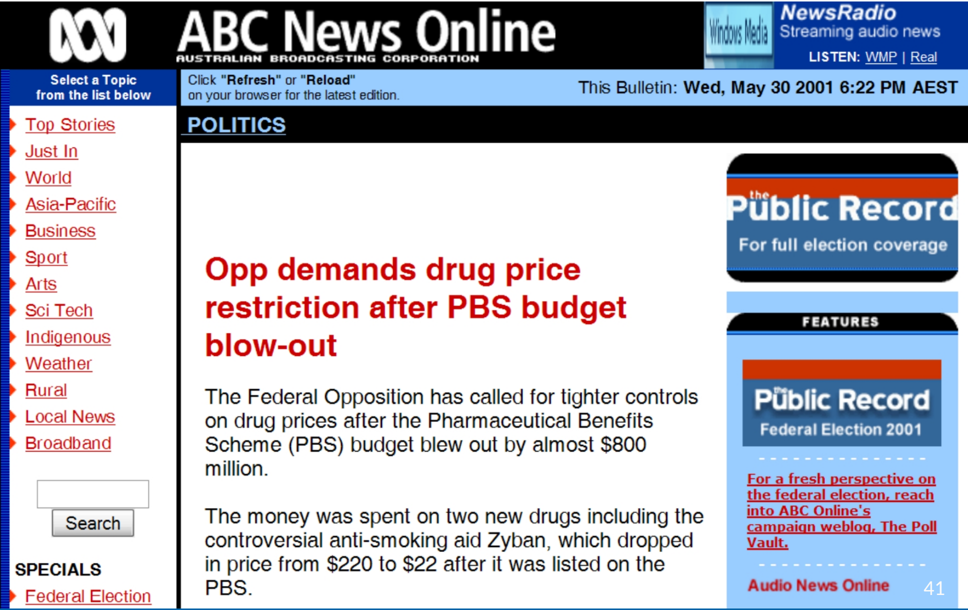
CASE STUDY 2: PBS

CASE STUDY 2: PBS

The Pharmaceutical Benefits Scheme (PBS) is the Australian government drugs subsidy scheme.

- Many drugs bought from pharmacies are subsidised to allow more equitable access to modern drugs.
- The cost to government is determined by the number and types of drugs purchased. Currently nearly 1% of GDP.
- The total cost is budgeted based on forecasts of drug usage.

CASE STUDY 2: PBS



ABC News Online
AUSTRALIAN BROADCASTING CORPORATION
Windows Media
NewsRadio
Streaming audio news
LISTEN: WMP | Real
Select a Topic from the list below
Click "Refresh" or "Reload" on your browser for the latest edition.
This Bulletin: Wed, May 30 2001 6:22 PM AEST
POLITICS
Top Stories
Just In
World
Asia-Pacific
Business
Sport
Arts
Sci Tech
Indigenous
Weather
Rural
Local News
Broadband
Search
SPECIALS
Federal Election
the Public Record
For full election coverage
FEATURES
Public Record
Federal Election 2001
For a fresh perspective on the federal election, reach into ABC Online's campaign weblog, The Poll Vault.
Audio News Online

Opp demands drug price restriction after PBS budget blow-out

The Federal Opposition has called for tighter controls on drug prices after the Pharmaceutical Benefits Scheme (PBS) budget blew out by almost $800 million.

The money was spent on two new drugs including the controversial anti-smoking aid Zyban, which dropped in price from $220 to $22 after it was listed on the PBS.

CASE STUDY 2: PBS

- In 2001: $4.5 billion budget, under-forecasted by $800 million.
- Thousands of products. Seasonal demand.
- Subject to covert marketing, volatile products, uncontrollable expenditure.
- Although monthly data available for 10 years, data are aggregated to annual values, and only the first three years are used in estimating the forecasts.
- All forecasts being done with the *FORECAST* function in MS-Excel!

CASE STUDY 3: Car fleet company

Client: One of Australia’s largest car fleet companies

Problem: how to forecast resale value of vehicles? How should this affect leasing and sales policies?

CASE STUDY 3: Car fleet company

Client: One of Australia's largest car fleet companies

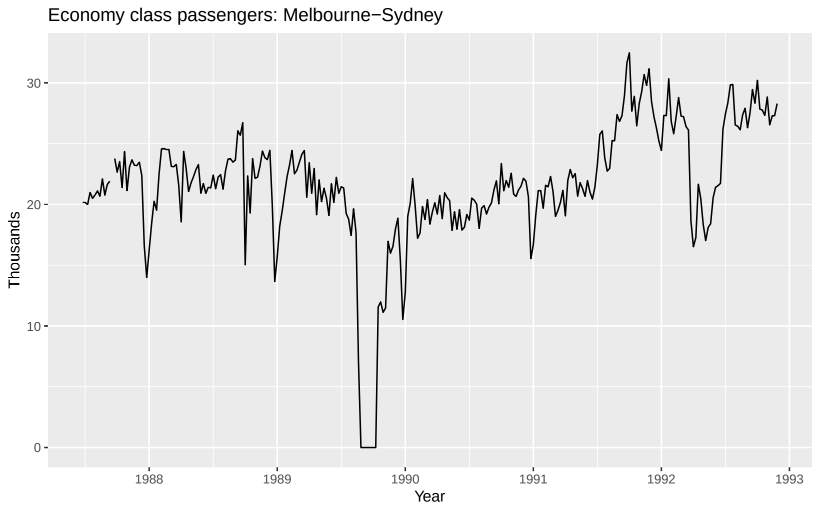
Problem: how to forecast resale value of vehicles? How should this affect leasing and sales policies?

Additional information

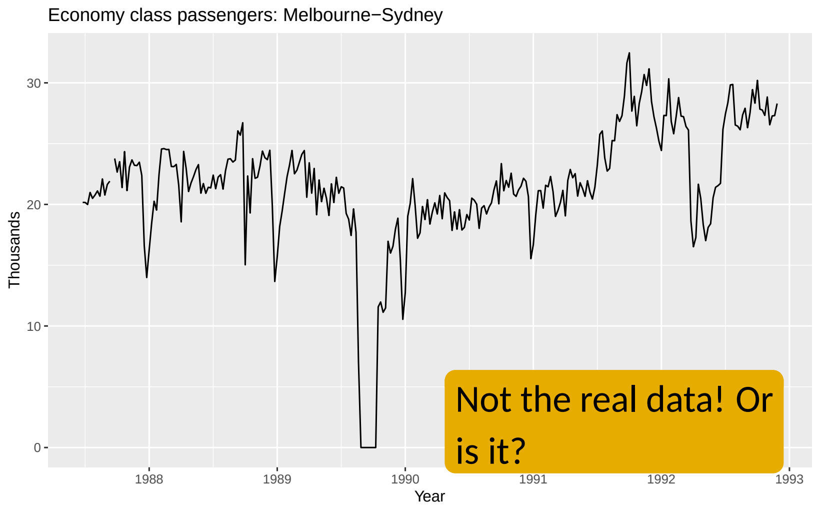
- They can provide a large amount of data on previous vehicles and their eventual resale values.
- The resale values are currently estimated by a group of specialists. They see me as a threat and do not cooperate.

CASE STUDY 4: Airline

CASE STUDY 4: Airline



CASE STUDY 4: Airline



Not the real data! Or is it?

CASE STUDY 4: Airline

Problem: how to forecast passenger traffic on major routes?

Additional information

- They can provide a large amount of data on previous routes.
- Traffic is affected by school holidays, special events such as the Grand Prix, advertising campaigns, competition behaviour, etc.
- They have a highly capable team of people who are able to do most of the computing.