MONASH University

MONASH BUSINESS SCHOOL

Forecasting is not prophecy

Dealing with high-dimensional probabilistic forecasts in practice

Rob J Hyndman

What is it?

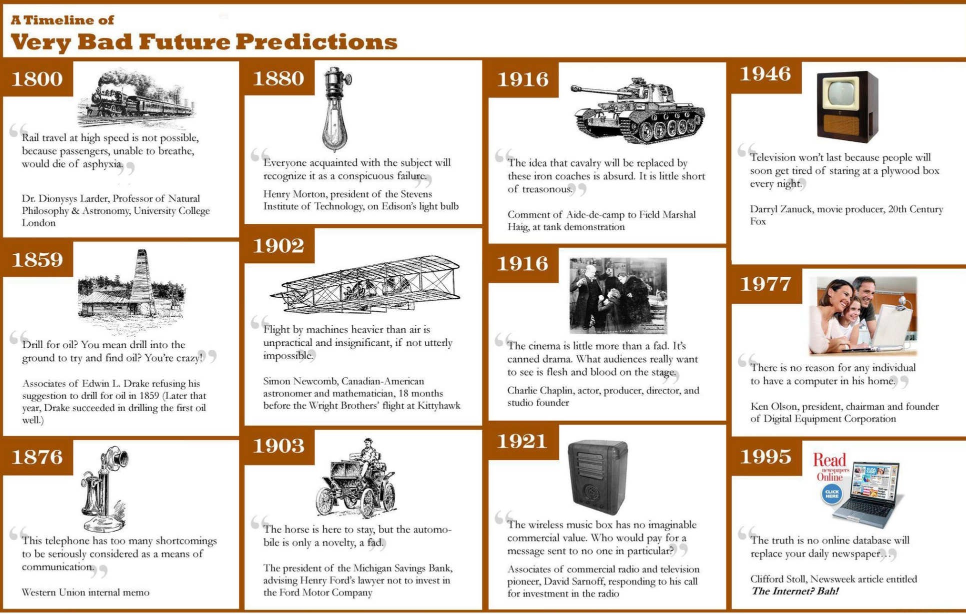
What is it?

Clay model of sheep's liver

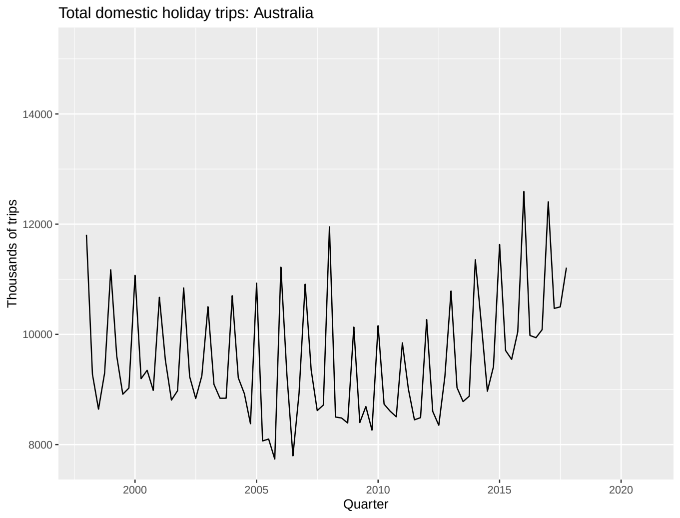
Used by Babylonian forecasters approximately 600 B.C.

Now in British Museum.

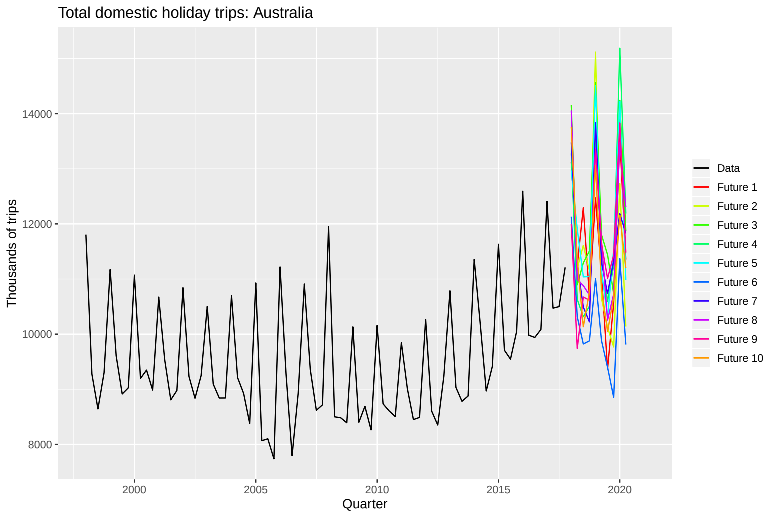
Modern prophecy



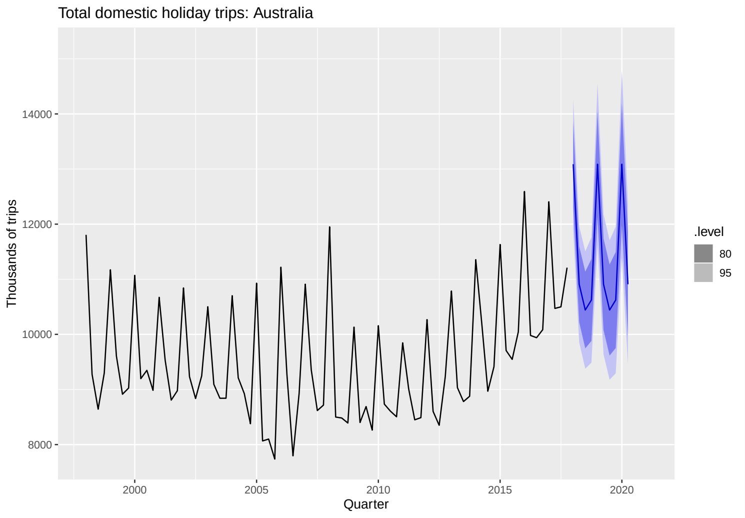
Probabilistic forecasting



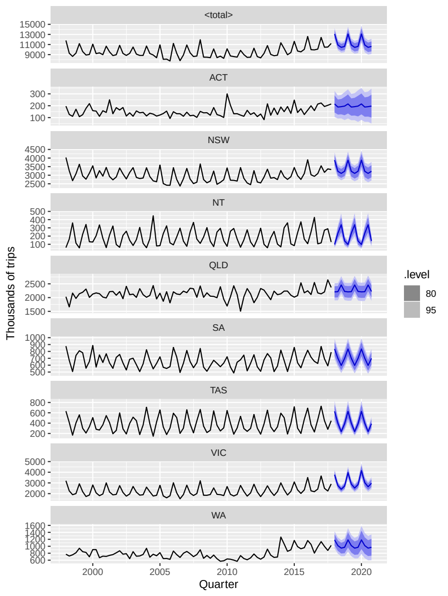
Probabilistic forecasting



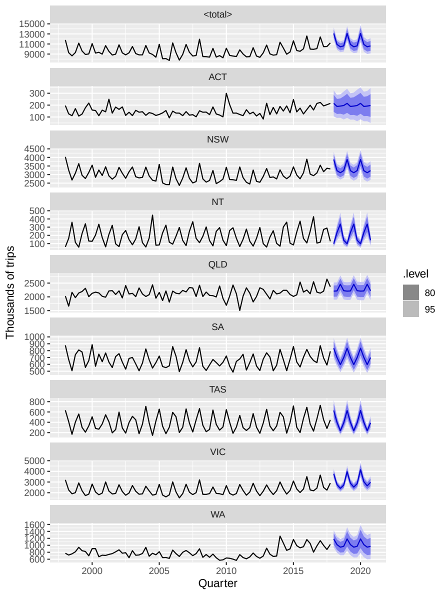
Probabilistic forecasting



Multivariate probabilistic forecasting



Multivariate probabilistic forecasting



These show the marginal distributions of the 9-dimensional distribution. In practice, we also want forecast for smaller geographical areas as well.

Who needs multivariate probabilistic forecasts?

- Tourism authorities forecasting visitor numbers to plan facilities.
- Manufacturing companies forecasting product demand to plan their supply chains.
- Call centres forecasting call volume to inform staff scheduling.
- Technology companies forecasting web traffic to maintain service levels.
- Energy companies forecasting electricity demand to prevent blackouts.

Challenges in multivariate probabilistic forecasting

1. How to produce high-dim probabilistic forecasts?
2. How to use and interpret high-dim probabilistic forecasts?
3. How to measure accuracy of high-dim probabilistic forecasts?

Challenges in multivariate probabilistic forecasting

1. How to produce high-dim probabilistic forecasts?
2. How to use and interpret high-dim probabilistic forecasts?
3. How to measure accuracy of high-dim probabilistic forecasts?

- Users care about marginal distributions, but we need multivariate distribution to compute them.
- Forecast reconciliation provides a way to efficiently compute marginal distributions.
- Probability scoring allows us to measure distributional forecast accuracy.

Probability scoring

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1\{y \leq x\})^2 dx$$
$$= \text{E}_F|Y - y| - \tfrac{1}{2}\text{E}_F|Y - Y'|$$

- y is observation, F is estimated forecast distribution
- Y and Y' are iid with cdf F
- CRPS in same units as observations
- Expected CRPS minimized when observation y from same distribution as F.
- Maximizes sharpness subject to calibration.

Tidyverts R packages

tidyverts.org

tsibble

tsibbledata

feasts

fable

tsibble objects

```r
library(tsibble)
library(tsibbledata)
aus_holidays
```

```
## # A tsibble: 6,080 x 4 [1Q]
## # Key:       Region, State [76]
##    Quarter Region   State Trips
##      <qtr> <chr>    <chr> <dbl>
##  1 1998 Q1 Adelaide SA     224.
##  2 1998 Q2 Adelaide SA     130.
##  3 1998 Q3 Adelaide SA     156.
##  4 1998 Q4 Adelaide SA     182.
##  5 1999 Q1 Adelaide SA     185.
##  6 1999 Q2 Adelaide SA     135.
##  7 1999 Q3 Adelaide SA     136.
##  8 1999 Q4 Adelaide SA     169.
##  9 2000 Q1 Adelaide SA     184.
## 10 2000 Q2 Adelaide SA     134.
```

tsibble objects

```r
library(tsibble)
library(tsibbledata)
aus_holidays
```

```
## # A tsibble: 6,080 x 4 [1Q]
## # Key:       Region, State [76]
##     Index   Region   State Trips
##       <qtr> <chr>    <chr> <dbl>
##   1 1998 Q1 Adelaide SA     224.
##   2 1998 Q2 Adelaide SA     130.
##   3 1998 Q3 Adelaide SA     156.
##   4 1998 Q4 Adelaide SA     182.
##   5 1999 Q1 Adelaide SA     185.
##   6 1999 Q2 Adelaide SA     135.
##   7 1999 Q3 Adelaide SA     136.
##   8 1999 Q4 Adelaide SA     169.
##   9 2000 Q1 Adelaide SA     184.
##  10 2000 Q2 Adelaide SA     134.
```

tsibble objects

```r
library(tsibble)
library(tsibbledata)
aus_holidays
```

```
## # A tsibble: 6,080 x 4 [1Q]
## # Key:       Region, State [76]
##     Index    Keys                 Trips
##       <qtr> <chr>     <chr> <dbl>
##  1 1998 Q1 Adelaide SA     224.
##  2 1998 Q2 Adelaide SA     130.
##  3 1998 Q3 Adelaide SA     156.
##  4 1998 Q4 Adelaide SA     182.
##  5 1999 Q1 Adelaide SA     185.
##  6 1999 Q2 Adelaide SA     135.
##  7 1999 Q3 Adelaide SA     136.
##  8 1999 Q4 Adelaide SA     169.
##  9 2000 Q1 Adelaide SA     184.
## 10 2000 Q2 Adelaide SA     134.
```

tsibble objects

```r
library(tsibble)
library(tsibbledata)
aus_holidays
```

```
## # A tsibble: 6,080 x 4 [1Q]
## # Key:       Region, State [76]
##      Index   Keys              Measure
##        <qtr> <chr>     <chr> <dbl>
##  1 1998 Q1 Adelaide SA     224.
##  2 1998 Q2 Adelaide SA     130.
##  3 1998 Q3 Adelaide SA     156.
##  4 1998 Q4 Adelaide SA     182.
##  5 1999 Q1 Adelaide SA     185.
##  6 1999 Q2 Adelaide SA     135.
##  7 1999 Q3 Adelaide SA     136.
##  8 1999 Q4 Adelaide SA     169.
##  9 2000 Q1 Adelaide SA     184.
## 10 2000 Q2 Adelaide SA     134.
```

tsibble objects

```r
library(tsibble)
library(tsibbledata)
aus_holidays
```

```
## # A tsibble: 6,080 x 4 [1Q]
## # Key:       Region, State [76]
##     Index   Keys          Measure
##       <qtr> <chr>     <chr> <dbl>
##  1 1998 Q1 Adelaide SA     224.
##  2 1998 Q2 Adelaide SA     130.
##  3 1998 Q3 Adelaide SA     156.
##  4 1998 Q4 Adelaide SA     182.
##  5 1999 Q1 Adelaide SA     185.
##  6 1999 Q2 Adelaide SA     135.
##  7 1999 Q3 Adelaide SA     136.
##  8 1999 Q4 Adelaide SA     169.
##  9 2000 Q1 Adelaide SA     184.
## 10 2000 Q2 Adelaide SA     134.
```

Domestic overnight holiday trips in thousands by state/region.

High-dimensional time series data

```
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) # 2 year test set
```

```
## # A tsibble: 5,472 x 4 [1Q]
## # Key:       Region, State [76]
##    Quarter Region   State Trips
##      <qtr> <chr>    <chr> <dbl>
##  1 1998 Q1 Adelaide SA     224.
##  2 1998 Q2 Adelaide SA     130.
##  3 1998 Q3 Adelaide SA     156.
##  4 1998 Q4 Adelaide SA     182.
##  5 1999 Q1 Adelaide SA     185.
##  6 1999 Q2 Adelaide SA     135.
##  7 1999 Q3 Adelaide SA     136.
##  8 1999 Q4 Adelaide SA     169.
##  9 2000 Q1 Adelaide SA     184.
## 10 2000 Q2 Adelaide SA     134.
## # ... with 5,462 more rows
```

Compute all aggregates

```
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips))
```

```
## # A tsibble: 6,120 x 4 [1Q]
## # Key:       State, Region [85]
##    State   Region  Quarter  Trips
##    <chr>   <chr>     <qtr>  <dbl>
##  1 <total> <total> 1998 Q1 11806.
##  2 <total> <total> 1998 Q2  9276.
##  3 <total> <total> 1998 Q3  8642.
##  4 <total> <total> 1998 Q4  9300.
##  5 <total> <total> 1999 Q1 11172.
##  6 <total> <total> 1999 Q2  9608.
##  7 <total> <total> 1999 Q3  8914.
##  8 <total> <total> 1999 Q4  9026.
##  9 <total> <total> 2000 Q1 11071.
## 10 <total> <total> 2000 Q2  9196.
```

Fit univariate models

```
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips))
```

```
## # A mable: 85 x 4
## # Key:     State, Region [85]
##    State Region                 mean     ets
##    <chr> <chr>                  <model> <model>
##  1 ACT   Canberra               <MEAN>  <ETS(M,N,A)>
##  2 ACT   <total>                <MEAN>  <ETS(M,N,A)>
##  3 NSW   Blue Mountains         <MEAN>  <ETS(M,N,M)>
##  4 NSW   Capital Country        <MEAN>  <ETS(A,N,N)>
##  5 NSW   Central Coast          <MEAN>  <ETS(M,N,M)>
##  6 NSW   Central NSW            <MEAN>  <ETS(M,N,A)>
##  7 NSW   Hunter                 <MEAN>  <ETS(A,N,A)>
##  8 NSW   New England North West <MEAN>  <ETS(M,N,N)>
##  9 NSW   North Coast NSW        <MEAN>  <ETS(A,N,A)>
```

Reconcile results

```r
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(ets = min_trace(ets))
```

```
## # A mable: 85 x 4
## # Key:     State, Region [85]
##    State Region                mean     ets
##    <chr> <chr>                 <model>  <model>
##  1 ACT   Canberra              <MEAN>   <ETS(M,N,A)>
##  2 ACT   <total>               <MEAN>   <ETS(M,N,A)>
##  3 NSW   Blue Mountains        <MEAN>   <ETS(M,N,M)>
##  4 NSW   Capital Country       <MEAN>   <ETS(A,N,N)>
##  5 NSW   Central Coast         <MEAN>   <ETS(M,N,M)>
##  6 NSW   Central NSW           <MEAN>   <ETS(M,N,A)>
##  7 NSW   Hunter                <MEAN>   <ETS(A,N,A)>
##  8 NSW   New England North West <MEAN>  <ETS(M,N,N)>
```

Produce probabilisic forecasts

```
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(ets = min_trace(ets)) %>%
  forecast(h = "2 years")
```

```
## # A fable: 1,360 x 6 [1Q]
## # Key:     State, Region, .model [170]
##    State Region   .model Quarter Trips .distribution
##    <chr> <chr>    <chr>    <qtr> <dbl> <dist>
##  1 ACT   Canberra mean   2016 Q1  146. N(146, 1521)
##  2 ACT   Canberra mean   2016 Q2  146. N(146, 1521)
##  3 ACT   Canberra mean   2016 Q3  146. N(146, 1521)
##  4 ACT   Canberra mean   2016 Q4  146. N(146, 1521)
##  5 ACT   Canberra mean   2017 Q1  146. N(146, 1521)
##  6 ACT   Canberra mean   2017 Q2  146. N(146, 1521)
##  7 ACT   Canberra mean   2017 Q3  146. N(146, 1521)
```

Compare against test set

```r
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(coherent = min_trace(ets)) %>%
  forecast(h = "2 years") %>%
  accuracy(aus_holidays)
```

```
## # A tibble: 170 x 4
##    .model Region                 State  CRPS
##    <chr>  <chr>                  <chr> <dbl>
##  1 ets    Canberra               ACT    34.6
##  2 ets    <total>                ACT    34.6
##  3 ets    Blue Mountains         NSW    15.1
##  4 ets    Capital Country        NSW    15.0
##  5 ets    Central Coast          NSW    11.5
##  6 ets    Central NSW            NSW    15.8
##  7 ets    Hunter                 NSW    47.9
```

Probability scoring

```
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(coherent = min_trace(ets)) %>%
  forecast(h = "2 years") %>%
  accuracy(aus_holidays) %>%
  spread(key=.model, value=CRPS)
```

```
## # A tibble: 85 x 4
##    Region                     State   ets   mean
##    <chr>                      <chr> <dbl>  <dbl>
##  1 Adelaide                   SA    24.1   24.4
##  2 Adelaide Hills             SA     2.90   4.18
##  3 Alice Springs              NT     6.08  10.4
##  4 Australia's Coral Coast    WA    14.9   35.0
##  5 Australia's Golden Outback WA     9.93  17.2
##  6 Australia's North West     WA     7.99  18.0
```

Probability scoring

```r
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(coherent = min_trace(ets)) %>%
  forecast(h = "2 years") %>%
  accuracy(aus_holidays) %>%
  spread(key=.model, value=CRPS) %>%
  mutate(RelCRPS = coherent/mean)
```

```
## # A tibble: 85 x 5
##    Region                      State   ets   mean RelCRPS
##    <chr>                       <chr> <dbl>  <dbl>   <dbl>
##  1 Adelaide                    SA    24.1   24.4    0.988
##  2 Adelaide Hills              SA     2.90   4.18   0.694
##  3 Alice Springs               NT     6.08  10.4    0.586
##  4 Australia's Coral Coast     WA    14.9   35.0    0.427
##  5 Australia's Golden Outback  WA     9.93  17.2    0.579
```

Probability scoring

```r
aus_holidays %>%
  filter(Quarter <= yearquarter("2015 Q4")) %>%
  aggregate_key(State/Region, Trips=sum(Trips)) %>%
  model(mean = MEAN(Trips), ets = ETS(Trips)) %>%
  reconcile(coherent = min_trace(ets)) %>%
  forecast(h = "2 years") %>%
  accuracy(aus_holidays) %>%
  spread(key=.model, value=CRPS) %>%
  mutate(RelCRPS = coherent/mean) %>%
  arrange(RelCRPS)
```

```
## # A tibble: 85 x 5
##    Region                           State   ets   mean RelCRPS
##    <chr>                            <chr> <dbl>  <dbl>   <dbl>
##  1 <total>                          WA     37.6  177.    0.212
##  2 South Coast                      NSW    26.1  101.    0.259
##  3 Australia's South West           WA     27.5  102.    0.268
##  4 Katherine Daly                   NT      3.59  11.8   0.304
```

Best forecast

Challenges in multivariate probabilistic forecasting

1. The fable package produces high-dim probabilistic forecasts?
2. We can focus on marginal distributions for interpretation, while needing the full distribution for calculation.
3. Forecast reconciliation provides a way to efficiently compute marginal distributions.
4. Probability scoring allows us to measure distributional forecast accuracy.

Acknowledgements

Mitchell O'Hara-Wild

Earo Wang

tidyverts.org

robjhyndman.com