Airline Data Analysis

Hamed Bastan-Hagh

6 July 2016

A step-by-step account of data work done for this analysis of airline data.

Data were not available via URL, so the files were downloaded on 7 Jul 2016 at 11:09. Changed them to xlsx format. Now read them into memory

```
library(xlsx)
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
d <- read.xlsx("combinedenplane.xlsx", 1, startRow = 3, endRow = 199)
dsum <- d %>%
    mutate(year = year(OBS_DATE)) %>%
    select(year, OBS_DATE:ENPLANE_D11) %>%
    group_by(year) %>%
    summarise(total = sum(ENPLANE))
dsum <- tbl_df(dsum)

domd <- read.xlsx("domesticenplane.xlsx", 1, startRow = 3, endRow = 199)
domsum <- domd %>%
    mutate(year = year(OBS_DATE)) %>%
```

```
    select(year, OBS_DATE:ENPLANE_D_D11) %>%
    group_by(year) %>%
    summarise(domestic = sum(ENPLANE_D))
domsum <- tbl_df(domsum)

intd <- read.xlsx("intenplane.xlsx", 1, startRow = 3, endRow = 199)
intsum <- intd %>%
    mutate(year = year(OBS_DATE)) %>%
    select(year, OBS_DATE:ENPLANE_I_D11) %>%
    group_by(year) %>%
    summarise(intl = sum(ENPLANE_I))
intsum <- tbl_df(intsum)

airdata <- cbind(dsum, domsum$domestic, intsum$intl)
airdata <- tbl_df(airdata)
names(airdata) <- c("year", "total", "domestic", "intl")
airdata
```

```
## Source: local data frame [17 x 4]
##
##     year  total domestic   intl
##    <dbl>  <dbl>    <dbl>  <dbl>
## 1   2000 669282   599565  69717
## 2   2001 625065   559617  65448
## 3   2002 616173   551898  64275
## 4   2003 647469   583293  64176
## 5   2004 703692   629768  73924
## 6   2005 738628   657261  81367
## 7   2006 744721   658363  86358
## 8   2007 769622   679168  90454
## 9   2008 743312   651709  91603
## 10  2009 703899   618051  85848
## 11  2010 720497   629538  90959
## 12  2011 730796   638247  92549
## 13  2012 736699   642289  94410
## 14  2013 743170   645679  97491
## 15  2014 762710   662831  99879
## 16  2015 798230   696027 102203
## 17  2016 189608   164931  24677
```

The totals for 2008-14 tally with those from the World Bank and the International Civil Aviation Organization, but differ slightly from those in the MarketLine report. The differences are small enough that we can attribute this to methodology or adjustments that MarketLine may have made based on their own information.

However the numbers for international and domestic are very different: the BTS numbers allocate much more of the total in each year to domestic flights. One clue as to why this might be comes from the instructions, which mention that "For the US and Canada, transborder passengers departing from either country are considered as part of the international segment". If the BTS allocate these to domestic that might explain the discrepancy.

I exchanged emailed with a librarian at the BTS and confirmed that those flights are categorised as domestic in the BTS numbers, which seems to explains the discrepancy. So to 'fix' the data I will use the 2015 numbers from the BTS for total passengers, and then allocate the proportions of domestic and international passengers to fit those in the MarketLine data, which averages at 79% domestic, 21% international.

```
## Add columns with imputed values for domestic and international passengers
## to the airdata table
airdata2 <- airdata %>%
    mutate(newdom = total * 0.79, newintl = total * 0.21) %>%
    select(year, total, newdom, newintl)
airdata2
```

```
## Source: local data frame [17 x 4]
##
##     year  total   newdom   newintl
##    <dbl>  <dbl>    <dbl>     <dbl>
## 1   2000 669282 528732.8 140549.22
## 2   2001 625065 493801.4 131263.65
## 3   2002 616173 486776.7 129396.33
## 4   2003 647469 511500.5 135968.49
## 5   2004 703692 555916.7 147775.32
## 6   2005 738628 583516.1 155111.88
## 7   2006 744721 588329.6 156391.41
## 8   2007 769622 608001.4 161620.62
## 9   2008 743312 587216.5 156095.52
## 10  2009 703899 556080.2 147818.79
## 11  2010 720497 569192.6 151304.37
## 12  2011 730796 577328.8 153467.16
## 13  2012 736699 581992.2 154706.79
## 14  2013 743170 587104.3 156065.70
## 15  2014 762710 602540.9 160169.10
## 16  2015 798230 630601.7 167628.30
## 17  2016 189608 149790.3  39817.68
```

We can also use the BTS data for the first quarter of 2016 to calculate the year-on-year growth vs. 2015.

```
d2015q1 <- d %>%
    mutate(year = year(OBS_DATE)) %>%
    mutate(month = month(OBS_DATE)) %>%
    select(year, month, OBS_DATE:ENPLANE) %>%
    filter(year == 2015 & month <= 3) %>%
    group_by(year) %>%
    summarise(total = sum(ENPLANE))

d2016q1 <- d %>%
    mutate(year = year(OBS_DATE)) %>%
    mutate(month = month(OBS_DATE)) %>%
    select(year, month, OBS_DATE:ENPLANE) %>%
    filter(year == 2016) %>%
    group_by(year) %>%
    summarise(total = sum(ENPLANE))

d2 <- cbind(t(d2015q1), t(d2016q1))
diff <- d2[2, 2] / d2[2, 1]
diff
```

```
##    total
## 1.052542
```

That means we've seen a c. 5% year-on-year increase from 2015 to 2016, and could expect about 840.17 million passengers in 2016, of which c. 663.735 million would be domestic and 176.436 million would be international.