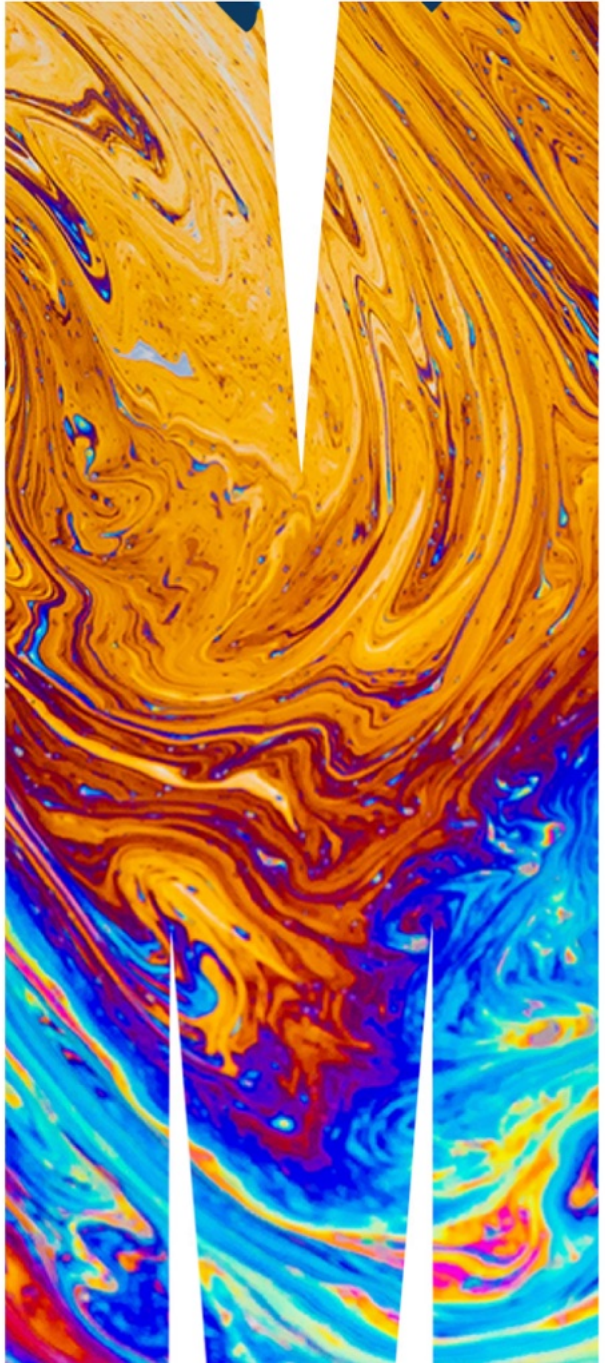
MONASH University

ETC5512: Wild Caught Data

Case study: US air traffic



Lecturer: *Lecturer: Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 3


Ready to take off with wild data?

This is a case study using the US airlines database

Motivation

American Statistical Association Statistical Graphics and Computing Sections 2009 Data Expo provided all of the commercial flight records for air travel in the USA from October 1987 to April 2008.

Questions provided

- When is the best time of day/day of week/time of year to fly to minimise delays?
- Do older planes suffer more delays?
- How does the number of people flying between different locations change over time?
- How well does weather predict plane delays?
- Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?

but participants could also decide for themselves what to analyse.

About the data

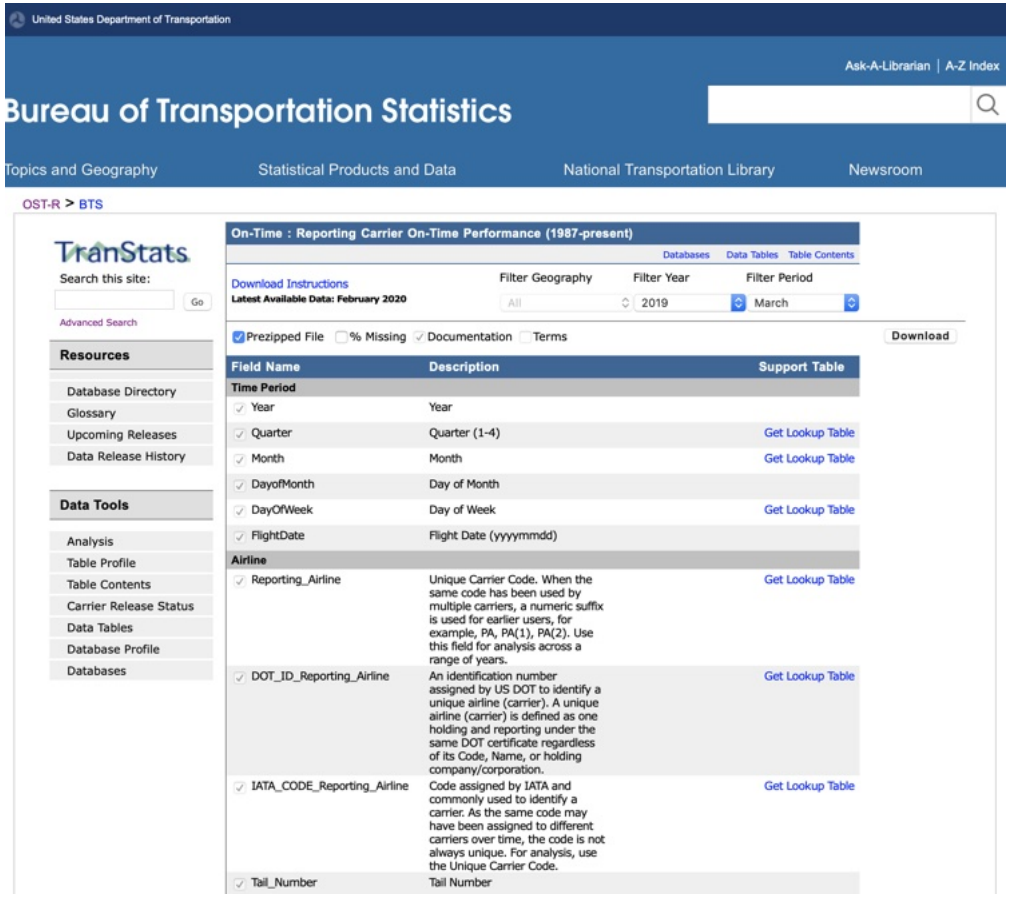
- nearly 120 million records
- 12Gb of space uncompressed
- 1.6Gb compressed

Organisers provided instructions on how to set up an **sqlite database**, and access from R.

Read about accessing databases from R at this RStudio site https://db.rstudio.com/databases/sqlite/ is a good starting place to read about working with a sqlite database.

The original data source

TranStats



- You can find the most current data at https://www.transtats.bts.gov/DataIndex.asp
- Look at the "On-Time Performance 1987-present" table.

- You can download data a month at a time
- There is a lag in records appearing on the site, currently of several months
- Data dictionary/explanations the variables
- Links at bottom of the site tells you what web site collects on you when you visit (Privacy Policy), but there is no clear license or policy on usage.

Accessing the data

- Data expo files: the data for the competition is still available because it was given a DOI: https://doi.org/10.7910/DVN/HG7NV7. 🦒
- Navigating the BTS web interface
  - What data is available
  - How do you download
  - Explanations of the records and variables
- R package `nycflights13`: provides a small domesticated data set. 🐨 This is a good way to *dip your toes in the water* with the airline data - try this as a start before working with the full data.

```
library(nycflights13)
data(airlines)
data(airports)
data(flights)
data(planes)
data(weather)
```

What does the data look like?

```
## # A tibble: 20 × 8
##    FL_DATE   OP_UNIQUE_CARRIER TAIL_NUM ORIGIN DEST  DEP_TIME ARR_TIME ARR_DE
##    <chr>     <chr>             <chr>    <chr>  <chr> <chr>    <chr>         <d
##  1 1/1/2023… 9E                N131EV   JFK    BGR   2056     2229
##  2 1/1/2023… 9E                N131EV   JFK    ORD   0941     1120
##  3 1/1/2023… 9E                N131EV   ORD    JFK   1524     1838
##  4 1/1/2023… 9E                N133EV   ABE    ATL   0601     0816
##  5 1/1/2023… 9E                N133EV   ATL    SGF   1125     1214
##  6 1/1/2023… 9E                N133EV   SGF    ATL   1354     1630
##  7 1/1/2023… 9E                N135EV   DTW    PVD   2132     2313
##  8 1/1/2023… 9E                N135EV   JFK    DTW   1252     1502
```

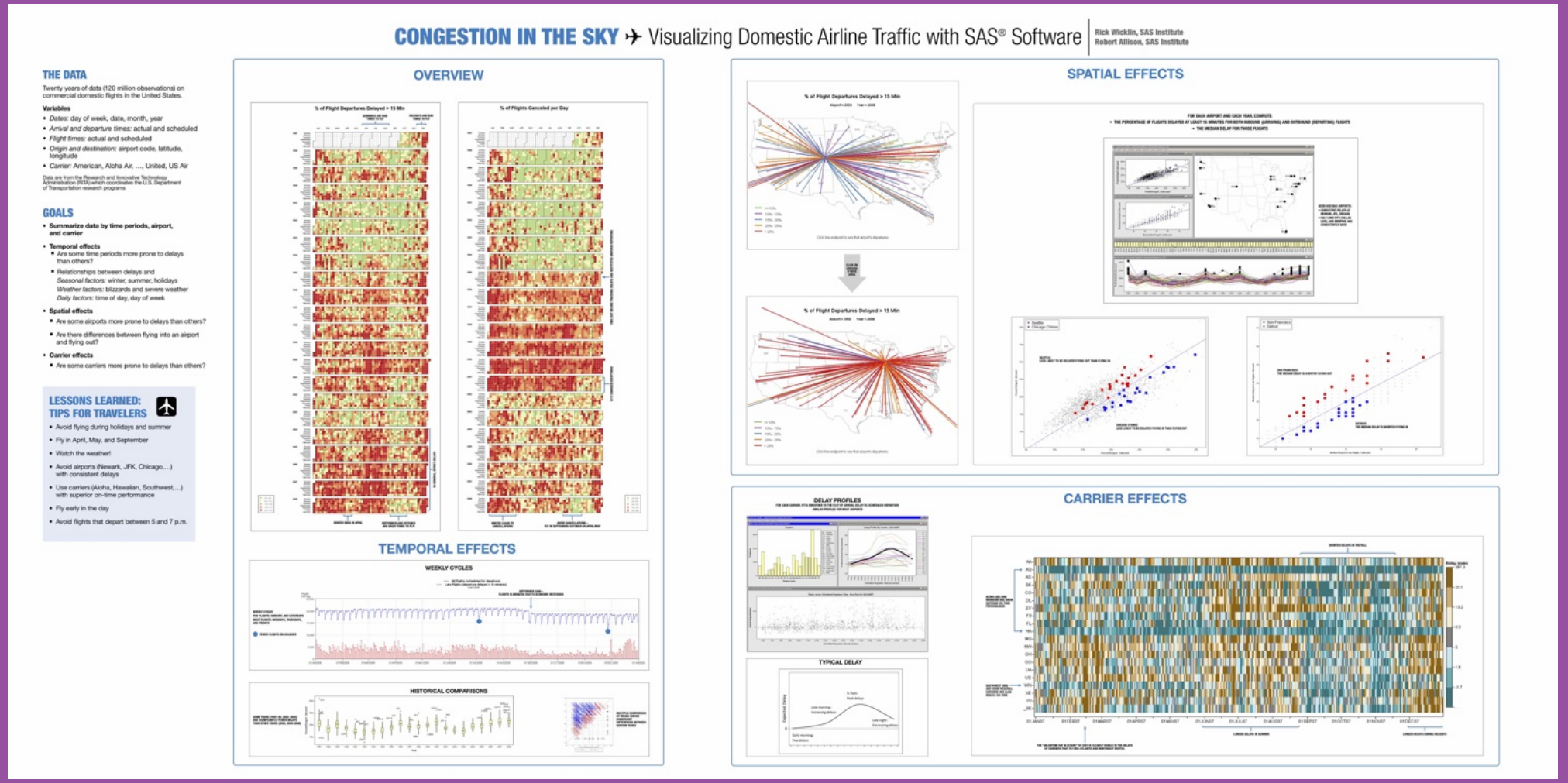
- What's in a row?
- What type of data collection is this? (e.g. experimental or observational? Census, survey sampling or occurrence?)

How would you start to process the data to answer ...

- When is the best time of day/day of week/time of year to fly to minimise delays?
- Are some carriers operating more efficiently?
- Do some carriers operate more broadly than others?
- Do older planes suffer more delays?

What did the prize winners do?

First prize


CONGESTION IN THE SKY ✈ Visualizing Domestic Airline Traffic with SAS® Software
Rick Wicklin, SAS Institute
Robert Allison, SAS Institute
THE DATA
Twenty years of data (120 million observations) on commercial domestic flights in the United States.
Variables
Dates: day of week, date, month, year
Arrival and departure times: actual and scheduled
Flight times: actual and scheduled
Origin and destination: airport code, latitude, longitude
Carrier: American, Aloha Air, ..., United, US Air
Data are from the Research and Innovative Technology Administration (RITA) which coordinates the U.S. Department of Transportation research programs
GOALS
Summarize data by time periods, airport, and carrier
Temporal effects
Are some time periods more prone to delays than others?
Relationships between delays and
Seasonal factors: winter, summer, holidays
Weather factors: blizzards and severe weather
Daily factors: time of day, day of week
Spatial effects
Are some airports more prone to delays than others?
Are there differences between flying into an airport and flying out?
Carrier effects
Are some carriers more prone to delays than others?
LESSONS LEARNED: TIPS FOR TRAVELERS
Avoid flying during holidays and summer
Fly in April, May, and September
Watch the weather!
Avoid airports (Newark, JFK, Chicago,...) with consistent delays
Use carriers (Aloha, Hawaiian, Southwest,...) with superior on-time performance
Fly early in the day
Avoid flights that depart between 5 and 7 p.m.
OVERVIEW
% of Flight Departures Delayed > 15 Min
% of Flights Canceled per Day
TEMPORAL EFFECTS
WEEKLY CYCLES
HISTORICAL COMPARISONS
SPATIAL EFFECTS
% of Flight Departures Delayed > 15 Min
% of Flight Departures Delayed > 15 Min
CARRIER EFFECTS
DELAY PROFILES
TYPICAL DELAY

Overview

Its good practice to show a useful view of entire data, to get a rough sense of major patterns.

Temporal trend

A major component of this data is traffic patterns over time.

Carriers

Are some carriers operating more widely, or more efficiently?

Spatial pattern

Airports are distributed across the country, explore how the traffic operates relative to this geography

Overview

THE DATA

Twenty years of data (120 million observations) on commercial domestic flights in the United States.

Variables

- *Dates:* day of week, date, month, year
- *Arrival and departure times:* actual and scheduled
- *Flight times:* actual and scheduled
- *Origin and destination:* airport code, latitude, longitude
- *Carrier:* American, Aloha Air, …, United, US Air

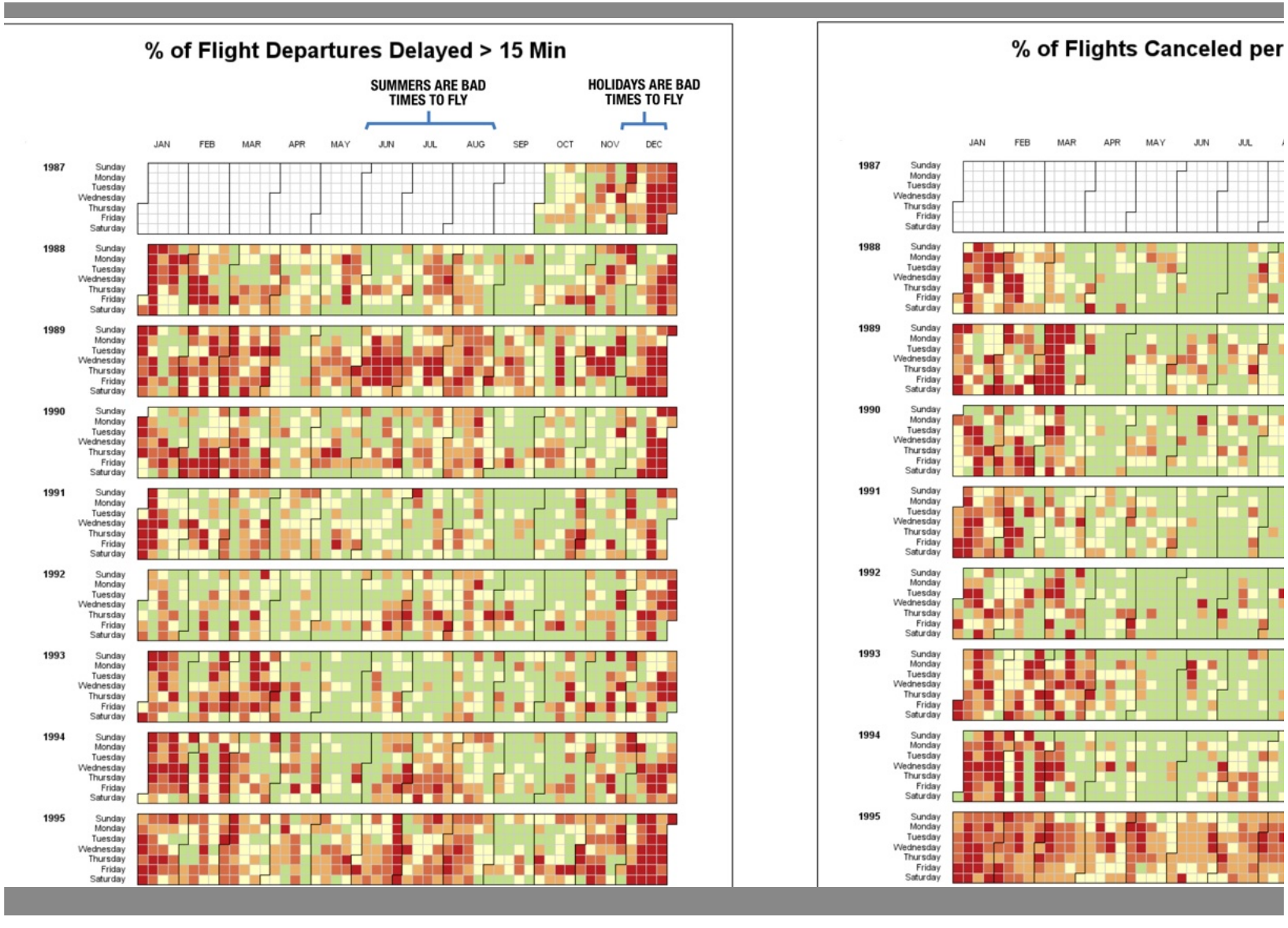
Data are from the Research and Innovative Technology Administration (RITA) which coordinates the U.S. Department of Transportation research programs

Overview

GOALS

- **Summarize data by time periods, airport, and carrier**
- **Temporal effects**
  - Are some time periods more prone to delays than others?
  - Relationships between delays and
    Seasonal factors: winter, summer, holidays
    Weather factors: blizzards and severe weather
    Daily factors: time of day, day of week
- **Spatial effects**
  - Are some airports more prone to delays than others?
  - Are there differences between flying into an airport and flying out?
- **Carrier effects**
  - Are some carriers more prone to delays than others?

Overview

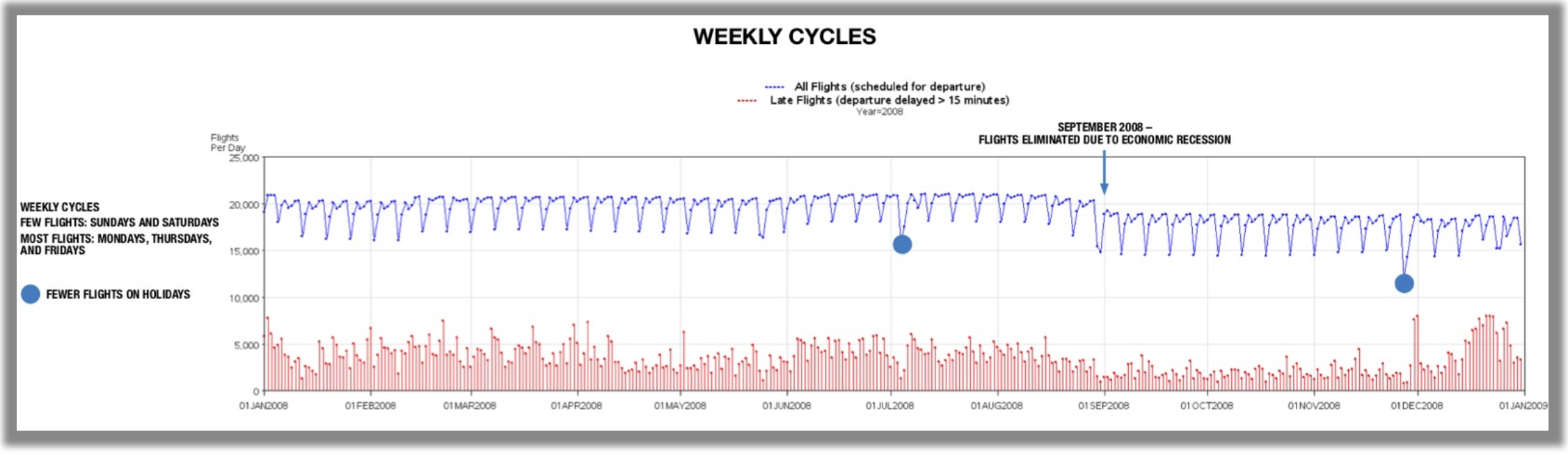

% of Flight Departures Delayed > 15 Min
SUMMERS ARE BAD TIMES TO FLY
HOLIDAYS ARE BAD TIMES TO FLY
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
1987 1988 1989 1990 1991 1992 1993 1994 1995
Sunday Monday Tuesday Wednesday Thursday Friday Saturday
% of Flights Canceled per
JAN FEB MAR APR MAY JUN JUL
1987 1988 1989 1990 1991 1992 1993 1994 1995
Sunday Monday Tuesday Wednesday Thursday Friday Saturday

Think about it

Delay was used in providing an overview.

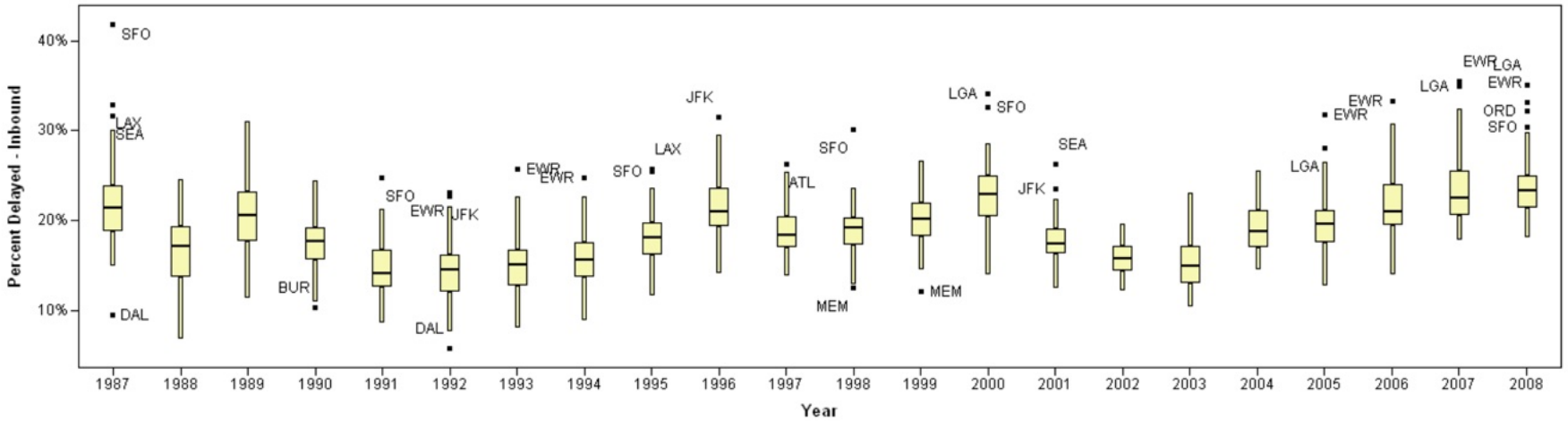
- What other aggregates could have been used?
- Why was delay chosen?

Temporal trend

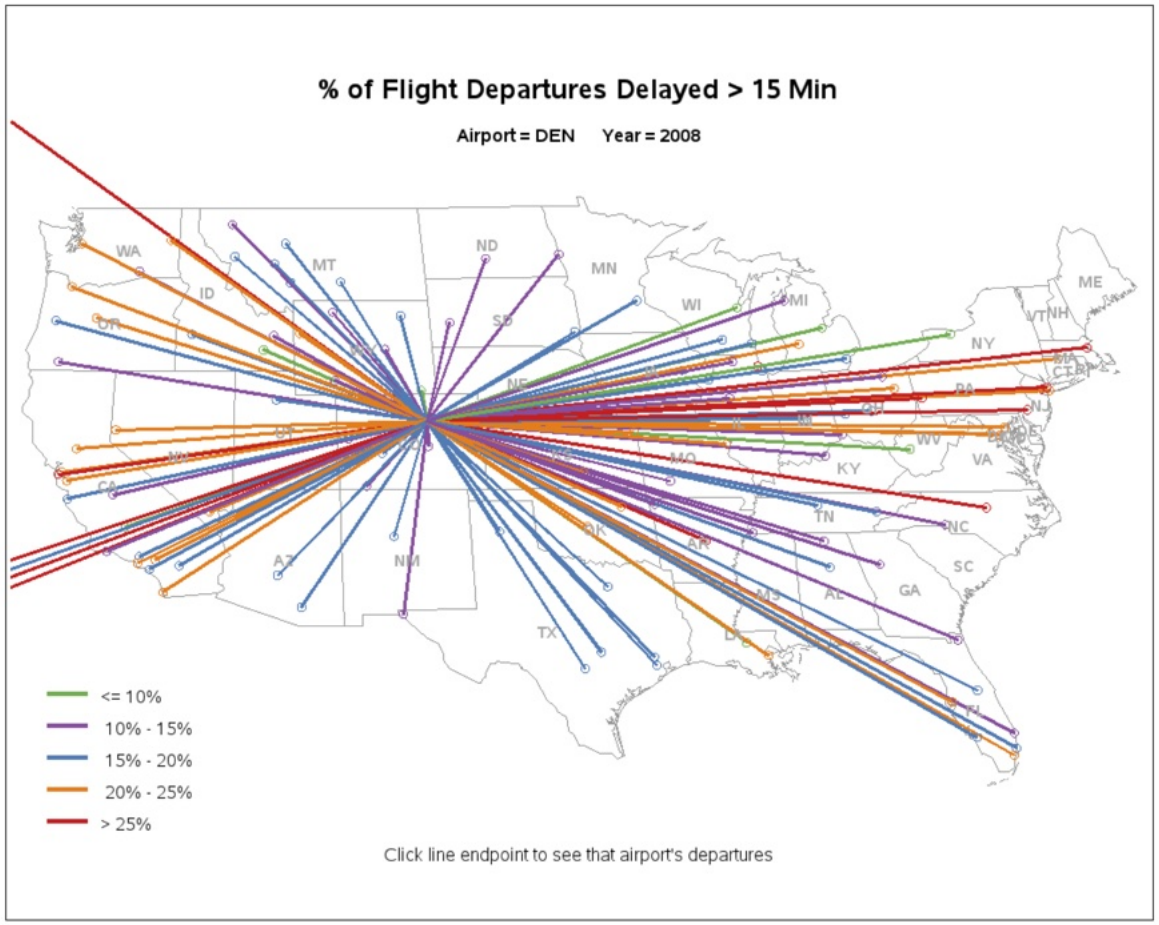


Temporal trend

SOME YEARS (1991–94, 2002–2003) HAD SIGNIFICANTLY FEWER DELAYS THAN OTHER YEARS (2000, 2006–2008)

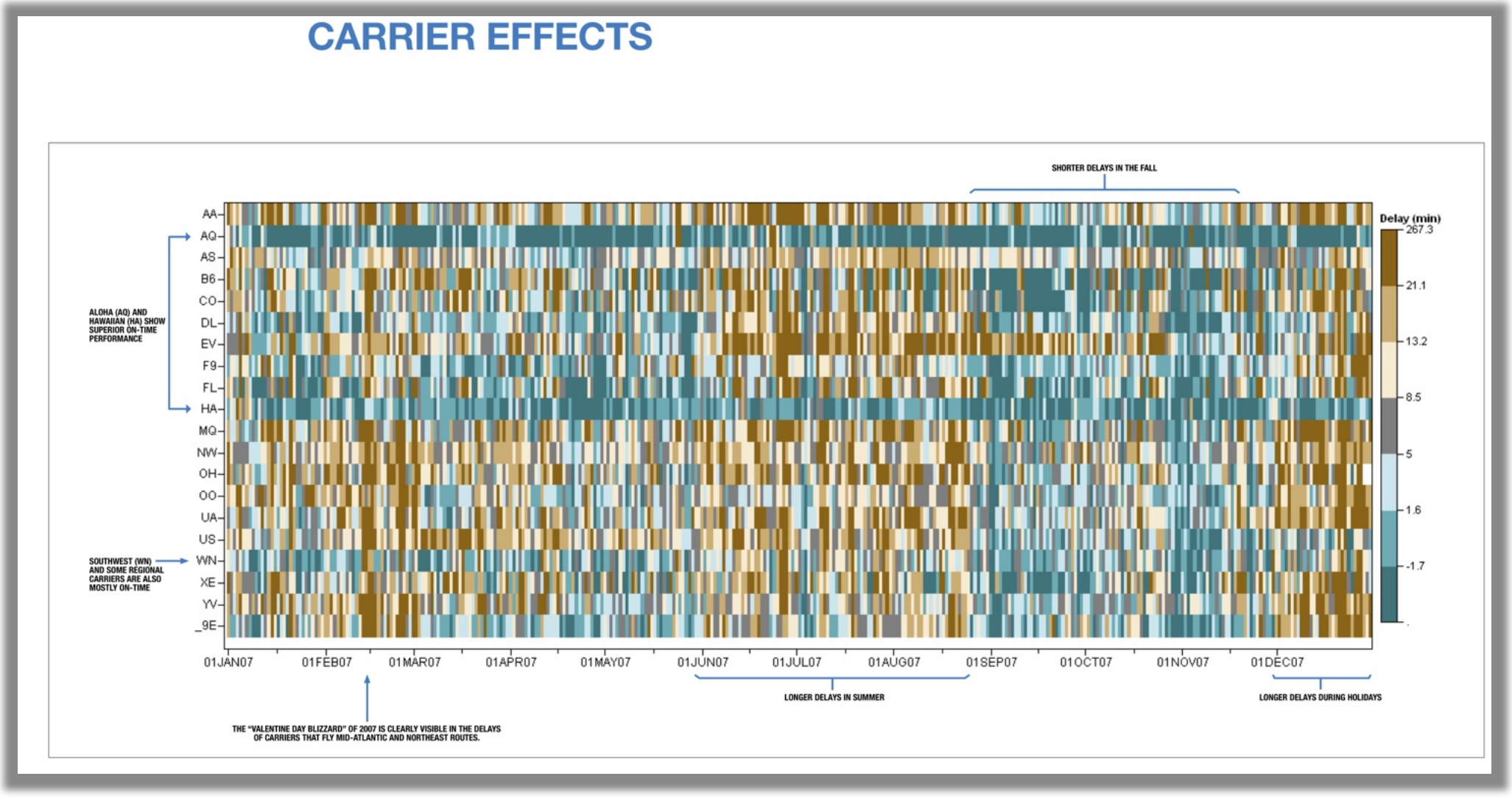


Spatial



CLICK ON CHICAGO O'HARE (ORD)

Carrier


CARRIER EFFECTS
SHORTER DELAYS IN THE FALL
ALOHA (AQ) AND HAWAIIAN (HA) SHOW SUPERIOR ON-TIME PERFORMANCE
SOUTHWEST (WN) AND SOME REGIONAL CARRIERS ARE ALSO MOSTLY ON-TIME
AA
AQ
AS
B6
CO
DL
EV
F9
FL
HA
MQ
NW
OH
OO
UA
US
WN
XE
YV
_9E
Delay (min)
267.3
21.1
13.2
8.5
5
1.6
-1.7
01JAN07
01FEB07
01MAR07
01APR07
01MAY07
01JUN07
01JUL07
01AUG07
01SEP07
01OCT07
01NOV07
01DEC07
LONGER DELAYS IN SUMMER
LONGER DELAYS DURING HOLIDAYS
THE "VALENTINE DAY BLIZZARD" OF 2007 IS CLEARLY VISIBLE IN THE DELAYS OF CARRIERS THAT FLY MID-ATLANTIC AND NORTHEAST ROUTES.

Highlights

LESSONS LEARNED: TIPS FOR TRAVELERS

- Avoid flying during holidays and summer
- Fly in April, May, and September
- Watch the weather!
- Avoid airports (Newark, JFK, Chicago,…) with consistent delays
- Use carriers (Aloha, Hawaiian, Southwest,…) with superior on-time performance
- Fly early in the day
- Avoid flights that depart between 5 and 7 p.m.

Second prize

Delayed, Cancelled, On Time, Boarding, ... Flying in the USA

Heike Hofmann, Di Cook, Chris Kielion, Barret Schloerke, Jon Hobbs, Adam Loy, Lawrence Mosley, David Rockoff, Yuanyuan Sun, Danielle Wrolstad, Tengfei Yin
Department of Statistics, Iowa State University

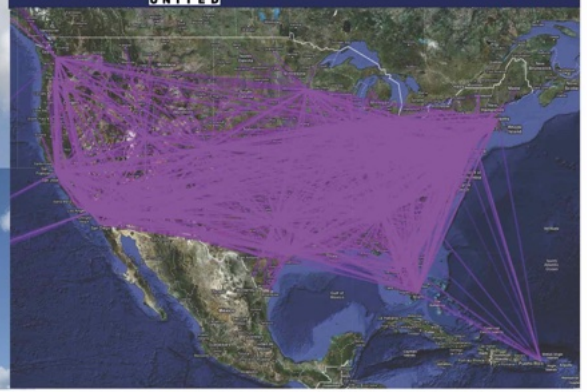
Data

The data are provided by Research and Innovative Technology Administration (RITA) and Bureau of Transportation Statistics (BTS). Arrival and departure details for 123 million commercial flights throughout the United States are recorded between October 1987 and December 2008, representing 29 commercial airlines and 3,376 airports. About 2.3 million flights were cancelled, 25 million flights were at least 15 minutes late.

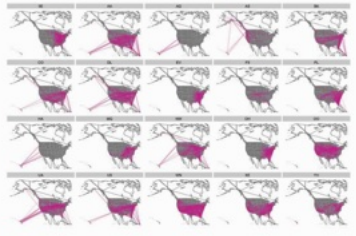
Additional Sources

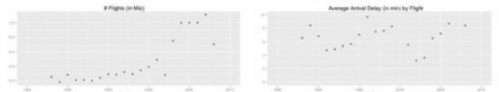
Additional BTS Data:
- Monthy fuel cost and consumption data by carrier
- Fleet information by carrier

Hourly weather details for each airport from Weather Underground at http://www.wunderground.com

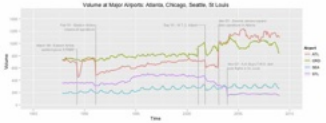


FLIGHTS OF '07
The maps above and below show all flights from 2007. Southwest Airlines (WN) operates without a hub system, the other airlines' hubs are prominent. Small carriers tend to operate locally.



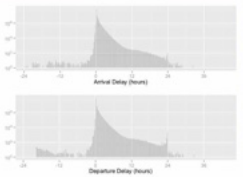


Flight volume is on the increase - dramatically so since 2000. Structural shifts (below) in the flight load for airports lead to minimal average delays (above, right) in '02 and '03. Delays have been on the increase since.

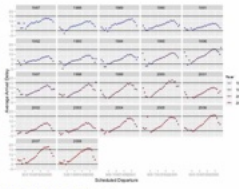


Large changes in airports' daily flight volume are triggered by different events, including strikes, FAA order, and seasonal fluctuations,

Delays

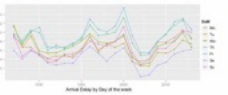


ARRIVAL & DEPARTURE
Most flights have 0 delay, with fewer and fewer flights having increasing delays. A secondary peak occurs at 24 hours suggests a limit of 24 hours delay is used by some carriers. Some data is likely incorrect, eg flights arriving 24 hours early.

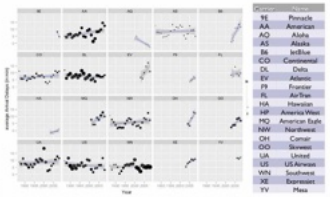


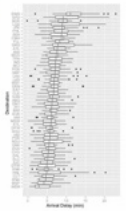
BY SCHEDULED DEPARTURE TIME & YEAR
Delays increase as day progresses. In 2001 they show an overall decrease, likely to be structural change, maybe FAA policy. Delays deteriorated again after 2003.

BY DAY OF WEEK
Best days to travel and avoid delays are Saturdays, and Tuesdays or Wednesdays. Fridays are bad for delays.



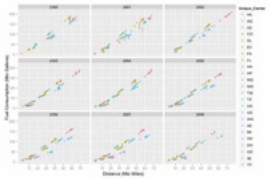
DELAYS BY CARRIER, YEAR
Increasing delays for small carriers, except for Aloha. Delta and US Airways are improving.



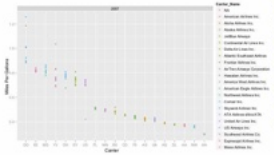


BY AIRPORT
EWR (Newark) is the worst. ORD (Chicago O'Hare) is not good, but also has high volume. DFW (Dallas-Fort Worth) is relatively good - high traffic but relatively small delay. Weather plays a huge role in delays - any kind of precipitation, high winds, or reduced visibility increases delays (scatterplots above)

Fuel Efficiency



FUEL USE
Consumption versus distance flown. Three groups of carriers are operating in 2000 - American Airlines is moving into the lead, Southwest is moving out of the middle group, closing up with AA and has overtaken AA on distance flown and efficiency 2007-8.



FUEL EFFICIENCY
Smaller carriers more efficient, probably due to use of smaller planes. Larger carriers American is one of the least efficient, Southwest is most efficient. Hawaiian Airlines very inefficient.

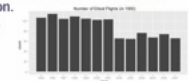
Ghosts of Flights

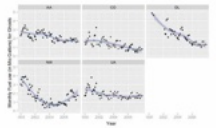
CAN WE SEE WHAT IS NOT THERE?
Planes have, for reasons such as maintenance, weather, or schedule fly empty between airports as so-called *Ghosts*. By tracking individual planes, we reveal their paths, including situations, where a plane lands in a different airport than where it takes off later, i.e. a ghost:

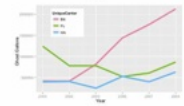
Example: US Airways Aircraft N-881 - Ghostflight from PIT to RIC (222 miles)

Year	Month	Day	DepTime	ArrTime	Origin	Dest	Diverted
1995	3	8	1102	1256	PIT	CVG	0
1995	3	8	1311	NA	CVG	PIT	1
1995	3	8	1913	2050	RIC	PIT	0
1995	3	8	2134	2300	PIT	MSY	0

Ghost Flight Totals: over 1 million flights since 1995, with an average distance between airports of 1000 miles, corresponding to about 1.5 million gallons of fuel. Since 2001 the number of ghost flights is cut, with major airlines also decreasing the fuel consumption. Smaller carriers are facing increasing costs from ghost flights.

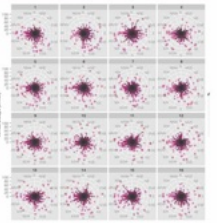






Crosswinds

All three runways in Phoenix are oriented East-West (see above). Do crosswinds affect air traffic?
On the right, arrival delays are plotted against permutations of wind direction (null plots). One plot shows the real data of the observed wind directions.



Can you spot which one?
If you can differentiate the true plot from the null plots, we can reject the null hypothesis of crosswinds being unrelated to arrival delays with a p-value of less than 0.0625 (= 1/16).

Believe it or not??

Racing Balloons?

Three of American Airlines' registered vehicles in the database are RAVEN hot air balloons. Based on the data, they cruise at an impressive average speed of 430 miles an hour. Fasten those hats!

I'm sorry, Sir, but your flight left 12h early'

According to the data, this was said to all passengers of 247 flights. Another 165 flights left at least 2h early.

"Within-City Hoppers"- if you're in a real time crunch

A total of 232,809 flights cover a distance of less than 50 miles. The shortest commercial flights occur between the New York airports La Guardia (LGA) and John F. Kennedy (JFK). The distance is 11 miles, which according to google.maps can be covered in 18 min by car, and according to data, takes 14 min of air time.

Tools

mySQL database on fast server (thanks to Ted Peterson)

R and packages
- ggplot2 (Hadley Wickham)
- DBI, RMySQL (David James, Jeff Horner)

Google Earth

supported by NSF grant # 0706949

Analysis overview

- Overview: flight paths over country
- Analysis:
  - Traffic patterns over time, including 911, and strikes, bankruptcies
  - Delays over time, and by day, hour
  - Airport efficiency
  - Carrier efficiency
  - Ghost flights: what's a ghost flight?
  - Mapping traffic spatially, and animating
- Curious findings

Believe it or not??

Racing Balloons?

Three of American Airlines' registered vehicles in the data-base are RAVEN hot air balloons. Based on the data, they cruise at an impressive average speed of 430 miles an hour. Fasten those hats!

I'm sorry, Sir, but your flight left 12h early'

According to the data, this was said to all passengers of 247 flights. Another 165 flights left at least 2h early.

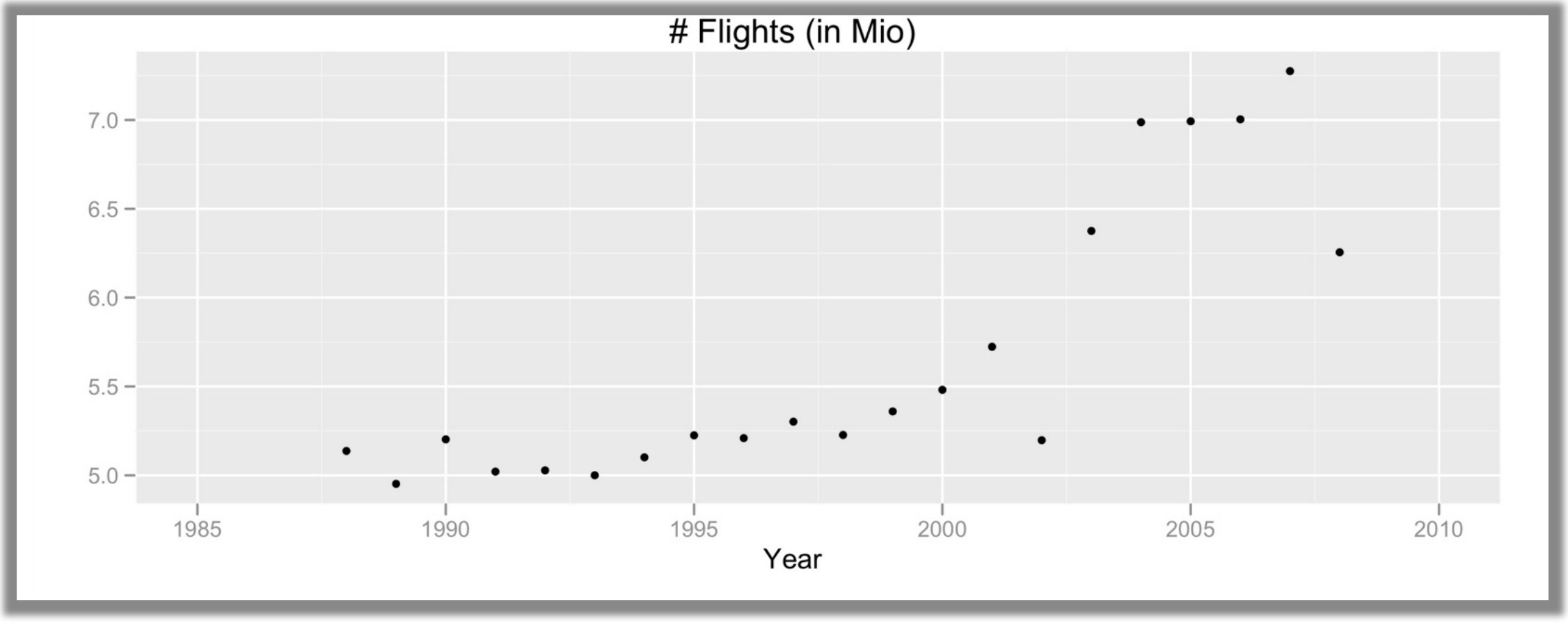
"Within-City Hoppers"- if you're in a real time crunch

A total of 232,809 flights cover a distance of less than 50 miles. The short-est commercial flights occur between the New York airports La Guardia (LGA) and John F. Kennedy (JFK). The distance is 11 miles, which accord-ing to google.maps can be covered in 18 min by car, and according to data, takes 14 min of air time.

As we work through the summary plots, think about

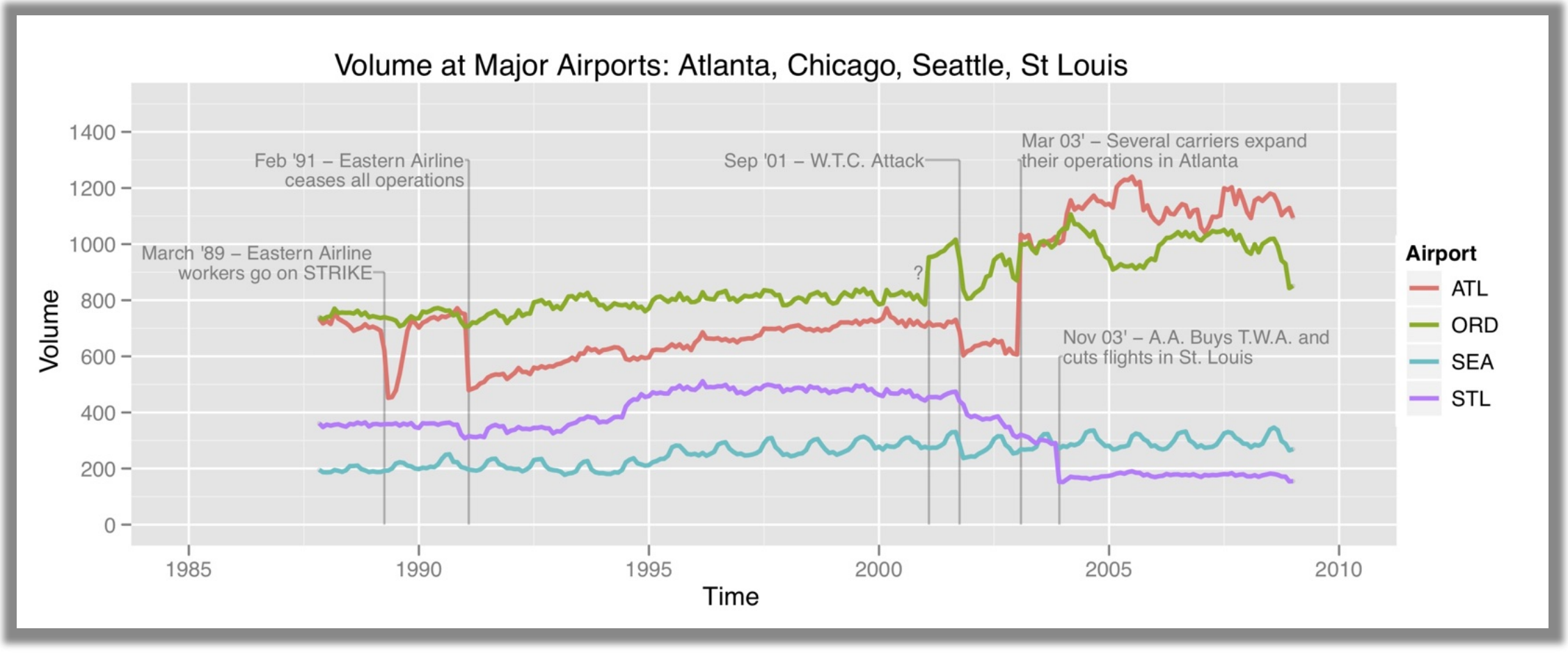
- what needs to be done to the data to get to this summary
- what do you learn from each display, what's expected, what's surprising
- what other ways might the same information be presented, or other calculations made

Traffic patterns over time

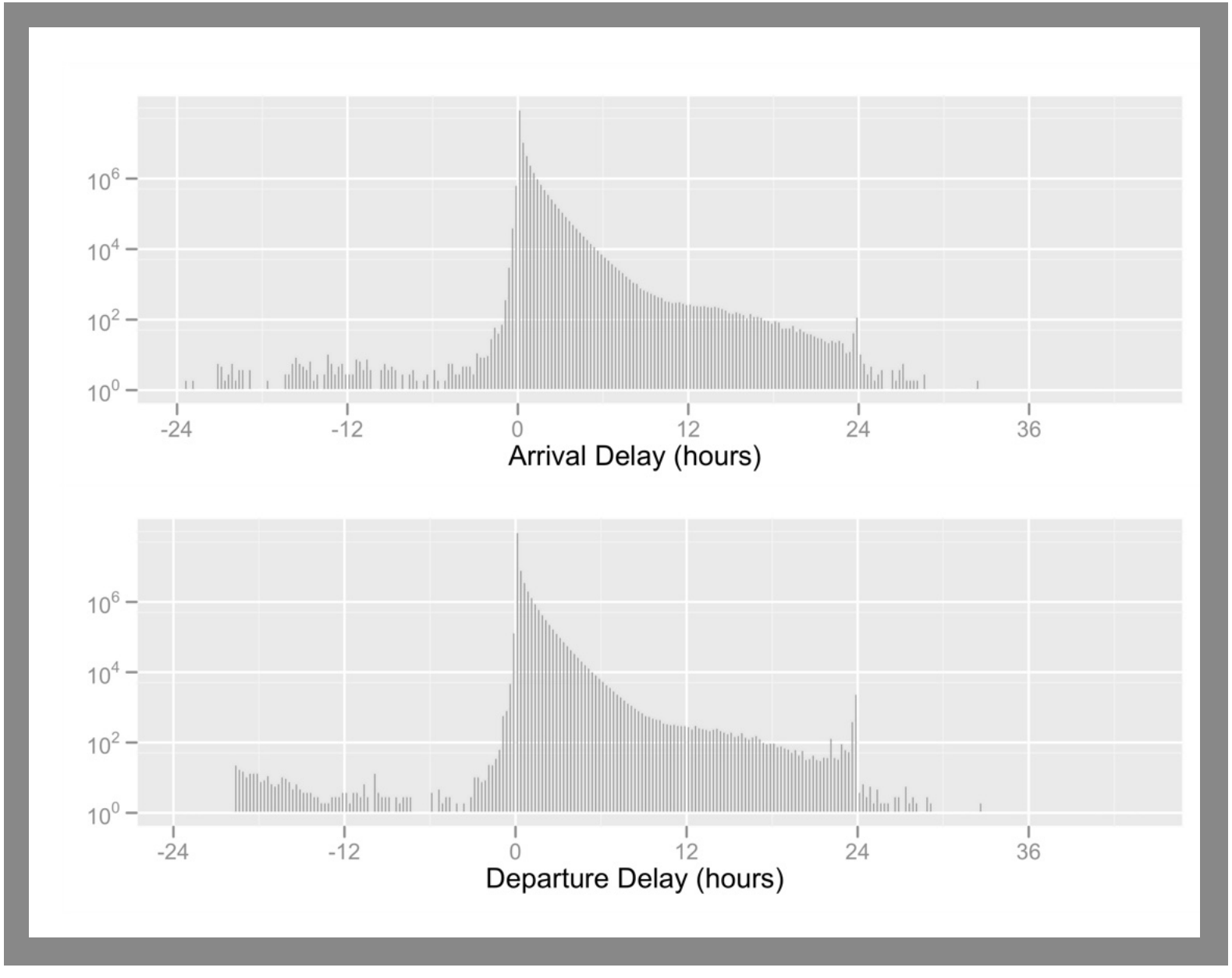


Number of flights in millions per year: steadily increasing volume until 2001, with a big drop in 2002. Volume recovered in 2003, and flattens 2004-7, with another drop in 2008. What happened in 2001? What was happening in 2008?

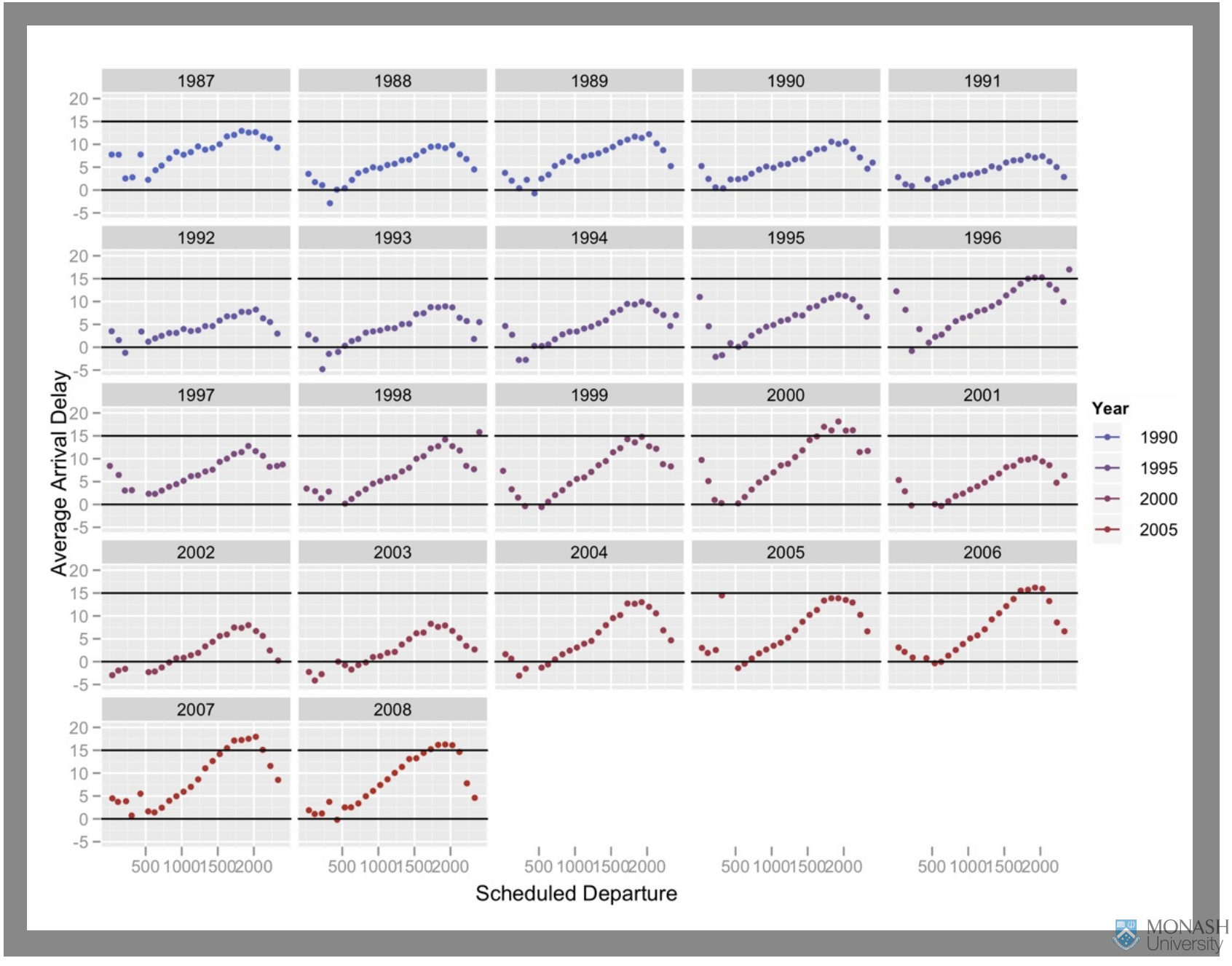
Traffic patterns at selected airports



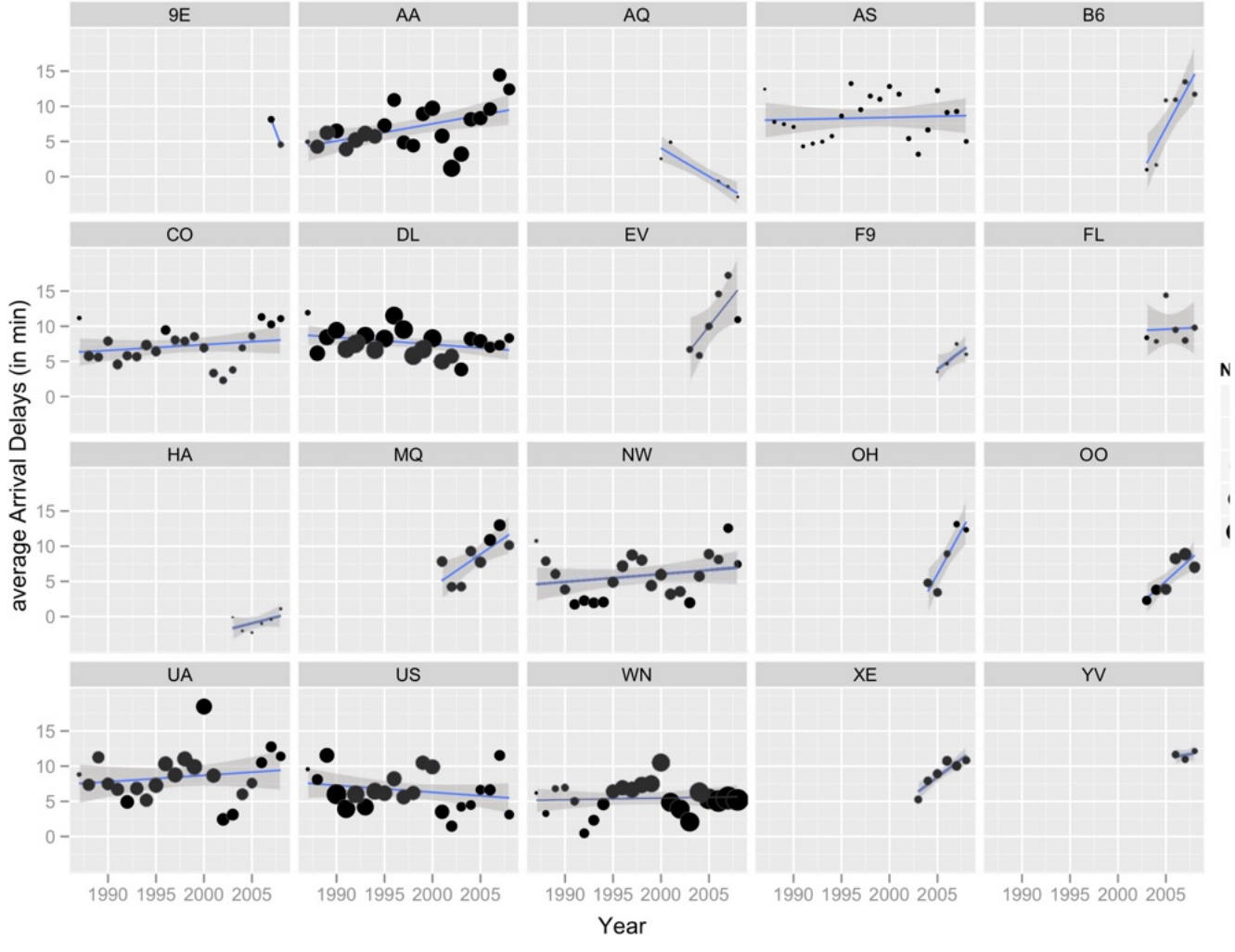
Delays


10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Arrival Delay (hours)
10^6
10^4
10^2
10^0
-24
-12
0
12
24
36
Departure Delay (hours)

Delays, by year

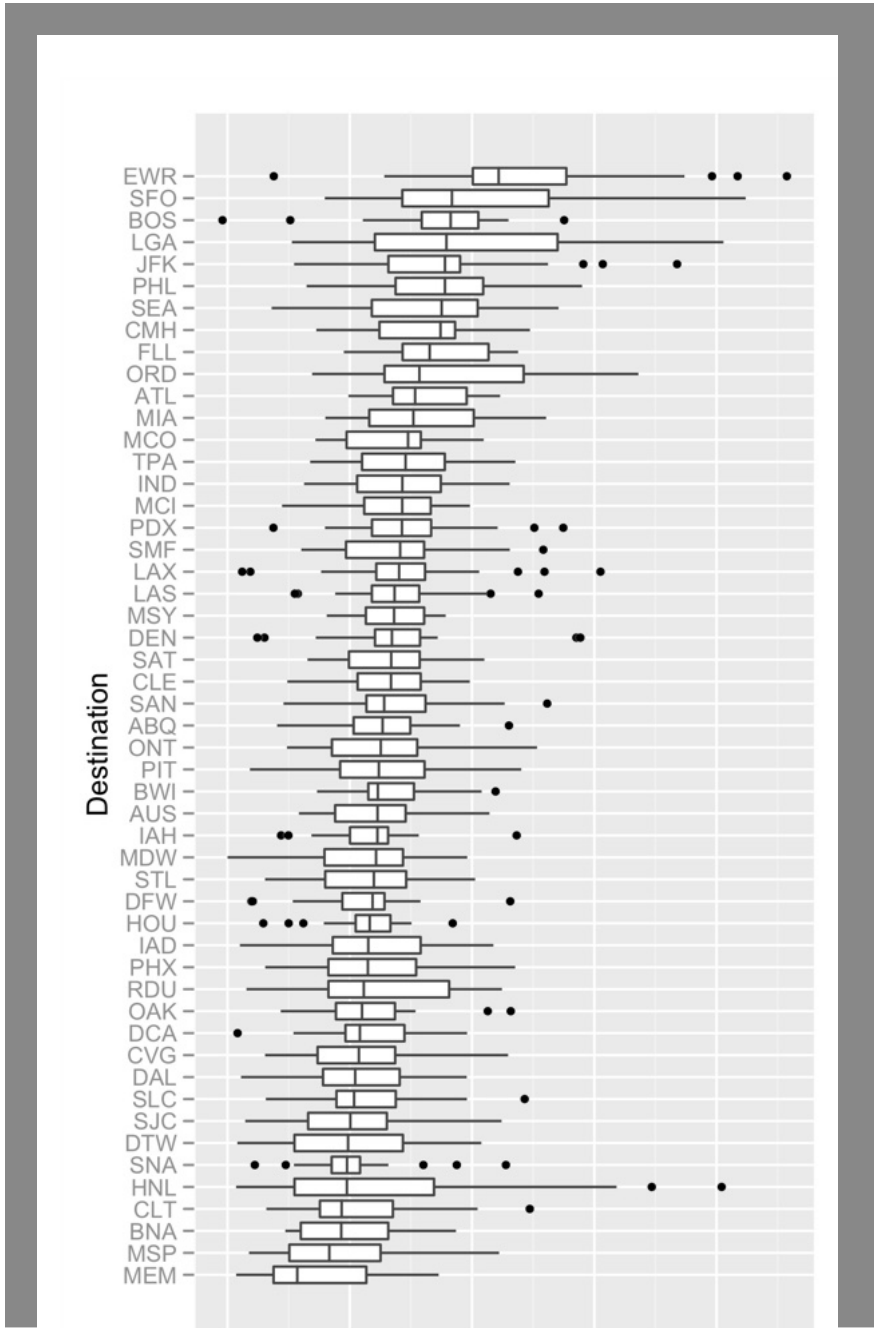


Delays, by carrier

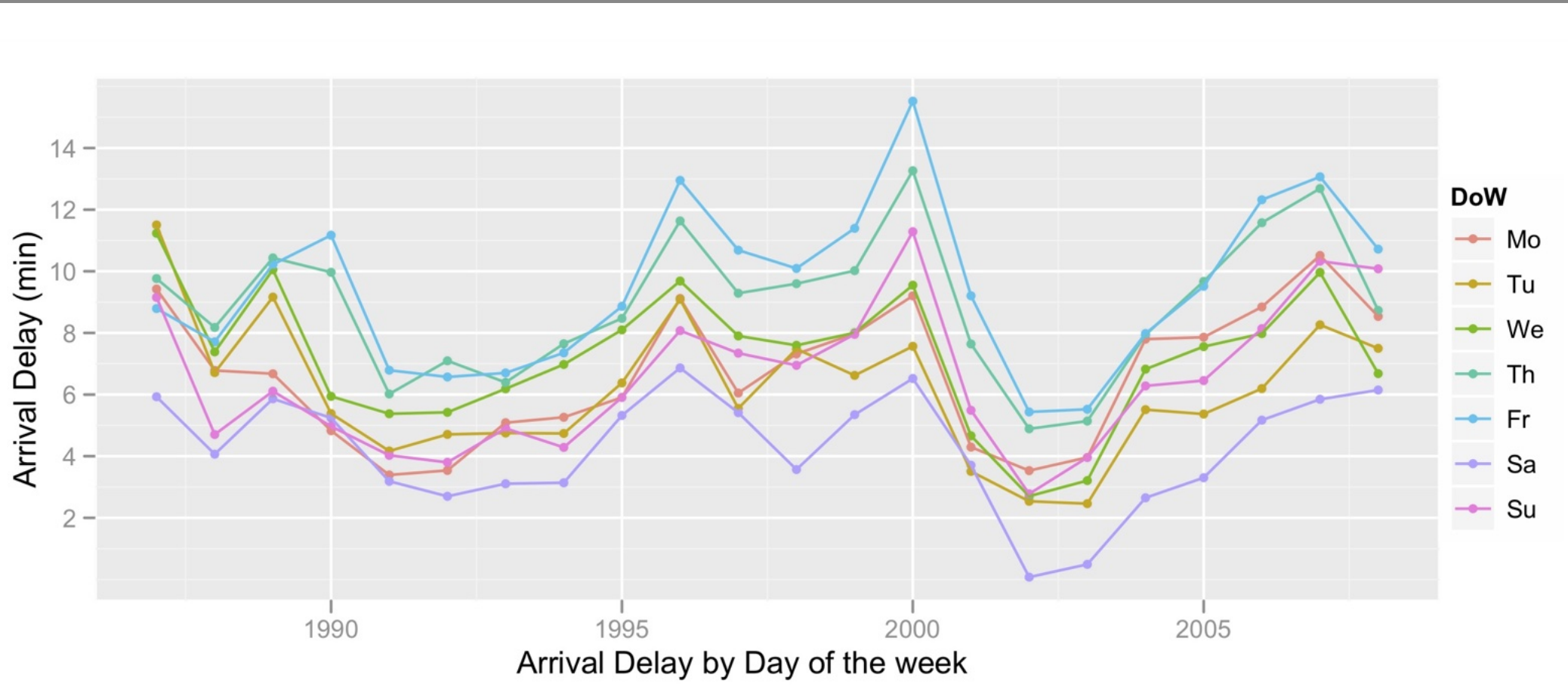


Carrier	Name
9E	Pinnacle
AA	American
AQ	Aloha
AS	Alaska
B6	JetBlue
CO	Continental
DL	Delta
EV	Atlantic
F9	Frontier
FL	AirTran
HA	Hawaiian
HP	America West
MQ	American Eagle
NW	Northwest
OH	Comair
OO	Skywest
UA	United
US	US Airways
WN	Southwest
XE	Expressjet
YV	Mesa

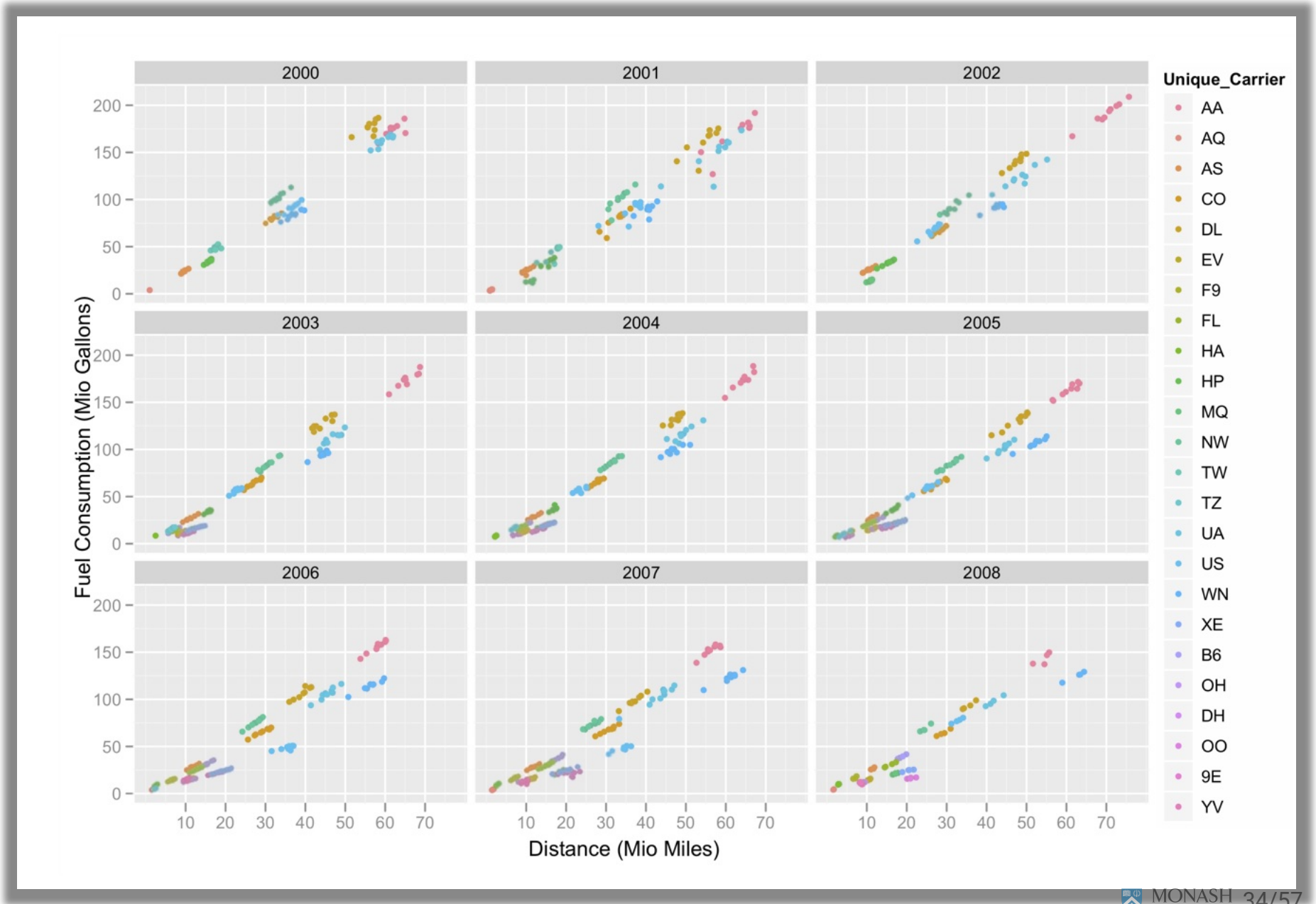
Delays, by airport



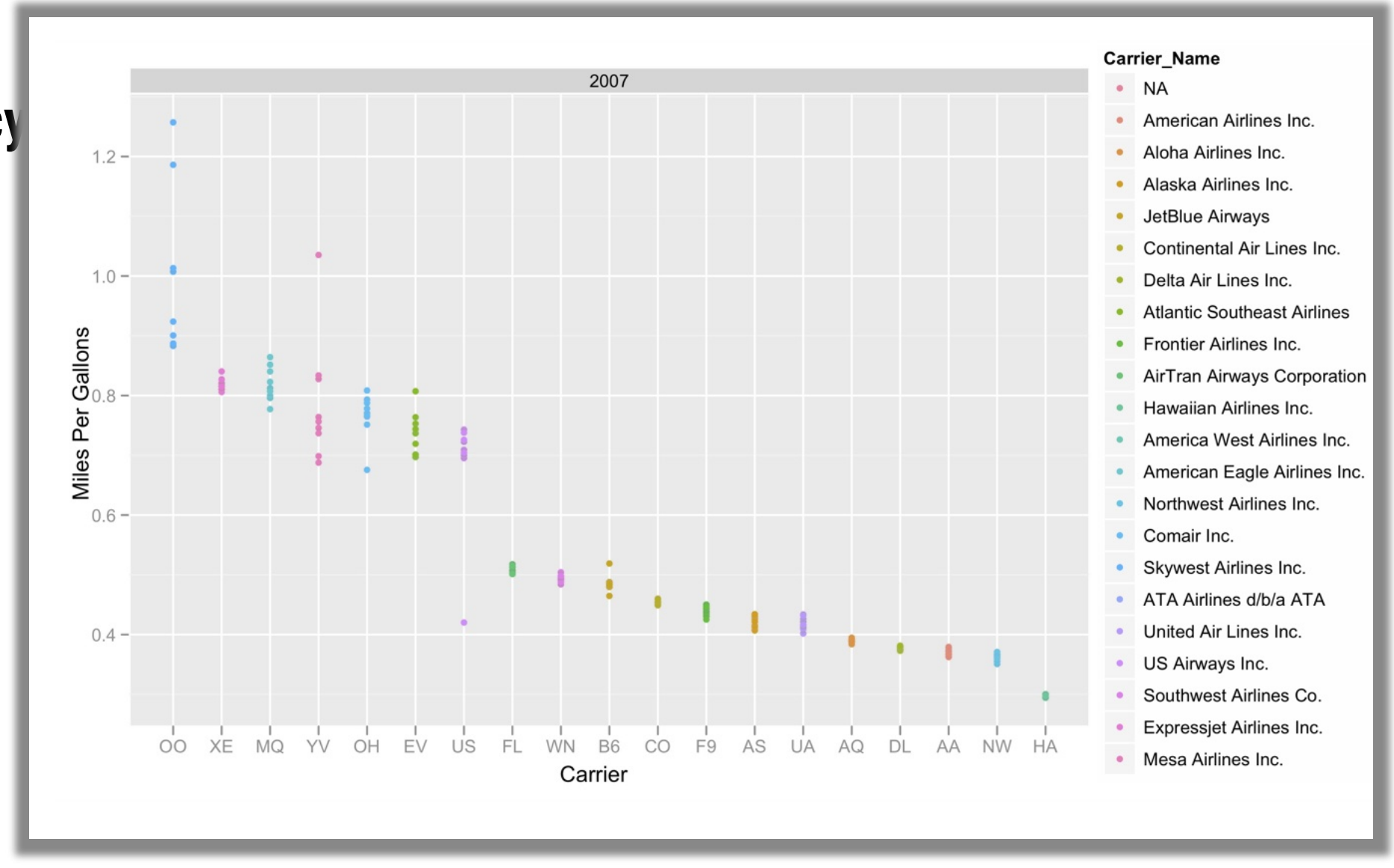
Delays, by day



Fuel use by carrier



Fuel efficiency



Ghost flights

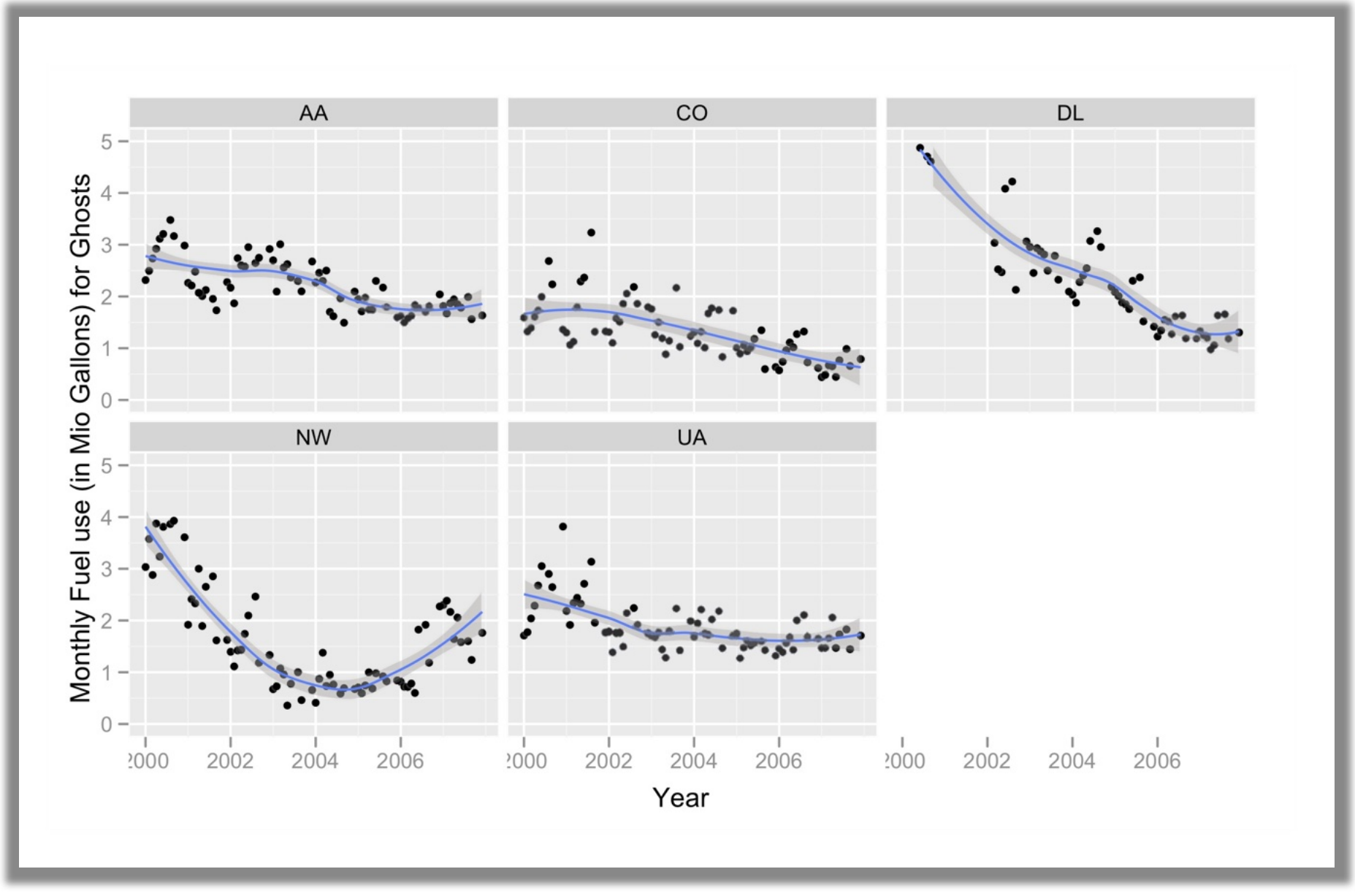
CAN WE SEE WHAT IS NOT THERE?

Planes have, for reasons such as maintenance, weather, or schedule fly empty between airports as so-called *Ghosts*. By tracking individual planes, we reveal their paths, including situations, where a plane lands in a different airport than where it takes off later, i.e. a ghost:

Example: US Airways Aircraft N-881 - Ghostflight from PIT to RIC (222 miles)

Year	Month	Day	DepTime	ArrTime	Origin	Dest	Diverted
1995	3	8	1102	1256	PIT	CVG	0
1995	3	8	1311	NA	CVG	PIT	1
1995	3	8	1913	2050	RIC	PIT	0
1995	3	8	2134	2300	PIT	MSY	0

Ghost flights, wasted fuel



What tools were used and why

A subset of the analysis materials including data and code can be downloaded from the paper site

- sqlite database: Inspired by the guidelines provided by the organisers we created a `mysql` database, on a central server that all team members could access with a password. Each person accessed the data through R.
- R packages: `RMySQL`, `DBI`, `ggplot2`

A brief introduction to working with databases

Databases

Working from these notes

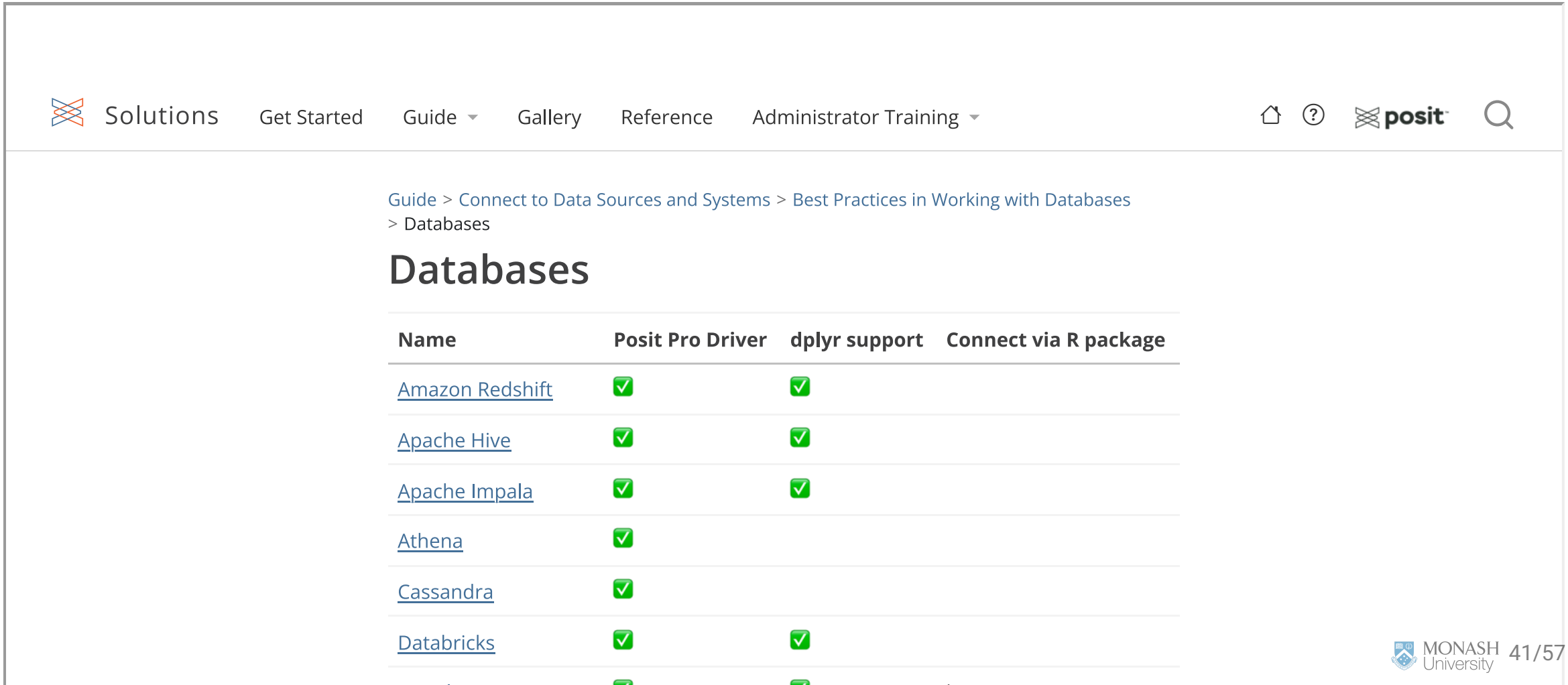
https://db.rstudio.com/databases/sqlite/

Why should I use a database?

- The data is too large to load into memory, ie work directly with it in R
- Database can make more efficient calculations
- Only load the data needed for specific analysis tasks

Connecting to an existing database

The packages `DBI`, `RMySQL`, `RPostgreSQL`, `RSQLite`, `bigrquery`, `odbc` enable connection to many different types of databases. The package `dbplyr` enables tidy style access to the databases.



Set up connection, using SQLite

```
# Set up connection
library(DBI)
library(RSQLite)
con <- dbConnect(RSQLite::SQLite(), "flights_database")
```

This creates the link between R and the database.

Suppose we want to set up a database

We want to add some data.

One month of air traffic data is quite manageable in an R session. We can use this to get started and to practice.

```
# Download a month of data and read into R
library(tidyverse)
flight_data_from_csv <- read_csv("../data/On_Time_Reporting_Carrier_On_Time_Perf
```

Practice adding data to the database

Add the flight data to our SQLite database. For this you can use

```
copy_to(con, flight_data_from_csv, "flights",
  temporary = FALSE,
  indexes = list(
    c("FlightDate",
    "Reporting_Airline",
    "Tail_Number",
    "Origin",
    "Dest"
  )
))
```

Here we named the data we uploaded "flights".

Note setting up the indexes makes it faster to process data on the database.

There are other approaches we can use to add data to a database using DBI functions

```
dbWriteTable(con, "flights", flight_data_from_csv)
dbListTables(con)
```

You can use the function tbl() to pull data from the database. Below shows an example, where we practice retrieving the flights data we uploaded before.

```
flight_data_from_db <- tbl(con, "flights")
flight_data_from_db
```

Getting data from the database

Using `tbl()` lets you look at what is stored, but does not pull the data back into memory yet. For that you need `collect()`.

This allows us to do operations on our data and identify the subset we are interested in before we pull the data back into computer memory.

```
subset_of_flight_data <- flights_data_from_db  %>%
  filter(DayofMonth==1) %>%
  select(DayofMonth, Origin, Dest) %>%
  collect()
subset_of_flight_data
```

or using dbSendQuery() with SQL

```
dbListFields(con, "flights")
res <- dbSendQuery(con, "SELECT * FROM
                   flights WHERE DayofMonth=1")
firstday <- dbFetch(res)
```

Add a table on airport details

Information about airport location and details is found in a different table at the BTS site: https://www.transtats.bts.gov/Fields.asp?Table_ID=288 . We will download this and add to our database to use for plotting flights on a map.

```
airport_data_from_csv <- read_csv("data/402312038_T_MASTER_CORD.csv") %>%
  select(-X29)
copy_to(con, airport_data_from_csv, "airports",
  temporary = FALSE
  )
dbListTables(con)
```

Its easy to forget what variables are in the table

You can check this with

```
dbListFields(con, "airports")
```

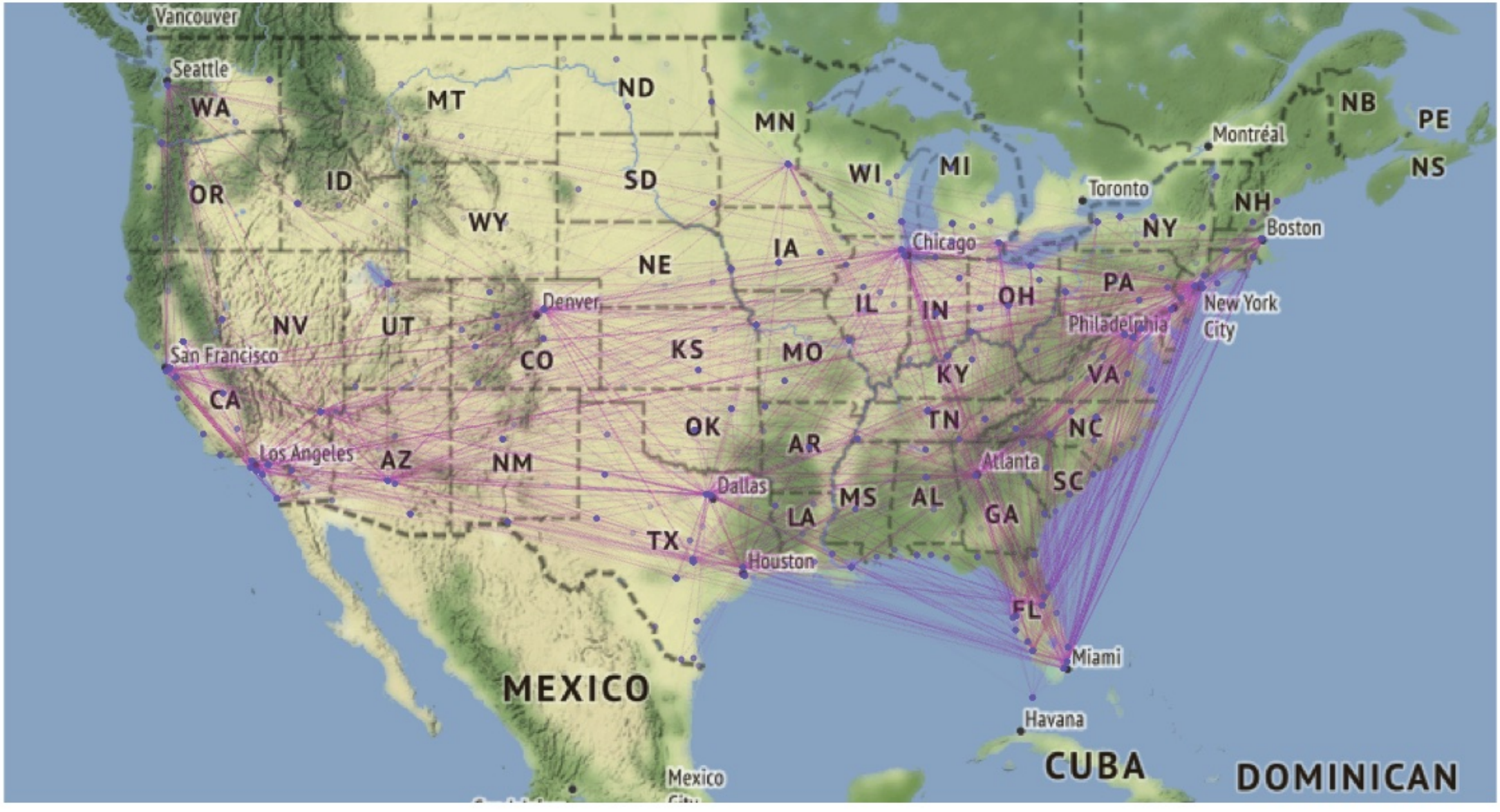
Make a map of flights for Feb 1

```
airport_locations <- tbl(con, "airports") %>%
  filter(AIRPORT_IS_LATEST == 1, AIRPORT_COUNTRY_CODE_ISO == "US") %>%
  select(AIRPORT, DISPLAY_AIRPORT_NAME, LONGITUDE, LATITUDE) %>%
  collect()

feb1_flights <- feb1 %>%
  left_join(airport_locations, by=c("Origin" = "AIRPORT")) %>%
  rename(Origin_lon = LONGITUDE, Origin_lat = LATITUDE,
         Origin_name = DISPLAY_AIRPORT_NAME) %>%
  left_join(airport_locations, by=c("Dest" = "AIRPORT")) %>%
  rename(Dest_lon = LONGITUDE, Dest_lat = LATITUDE,
         Dest_name = DISPLAY_AIRPORT_NAME)
```

```
# OLD CODE / WILL REQUIRE API KEY NOW TO RUN
# REFER TO TUTORIAL CODE INSTEAD
library(ggmap)
usa_bbox <- c(-130, # min long
              20, # min lat
              -60, # max long
              50) # max lat
usa_map <- get_map(location = usa_bbox, source = "osm")
ggmap(usa_map)
```

```
# OLD CODE - WILL REQUIRE API KEY TO GET MAP
# REFER TO TUTORIAL CODE INSTEAD
library(ggthemes)
ggmap(usa_map) + geom_segment(data=feb1_flights,
                      aes(x=Origin_lon,
                          xend=Dest_lon,
                          y=Origin_lat,
                          yend=Dest_lat),
                      colour="#9651A0", alpha=0.01) +
  geom_point(data=feb1_flights, aes(x=Origin_lon, Origin_lat),
             colour="#746FB2", alpha=0.1, size=1) +
  theme_map()
```


Vancouver
Seattle
WA
MT
ND
MN
NB
PE
Montréal
NS
OR
ID
SD
WI
MI
Toronto
NH
WY
NY
Boston
IA
Chicago
NE
PA
Denver
IL
IN
OH
New York City
Philadelphia
NV
UT
San Francisco
CO
KS
MO
KY
VA
CA
TN
NC
OK
AR
Los Angeles
AZ
NM
Atlanta
SC
Dallas
MS
AL
GA
LA
TX
Houston
FL
Miami
MEXICO
Havana
CUBA
Mexico
DOMINICAN

Animating flights for one day

Data Visualization and Statistical Graphics in Big Data Analysis: Video 1 from Annual Reviews on Vimeo.

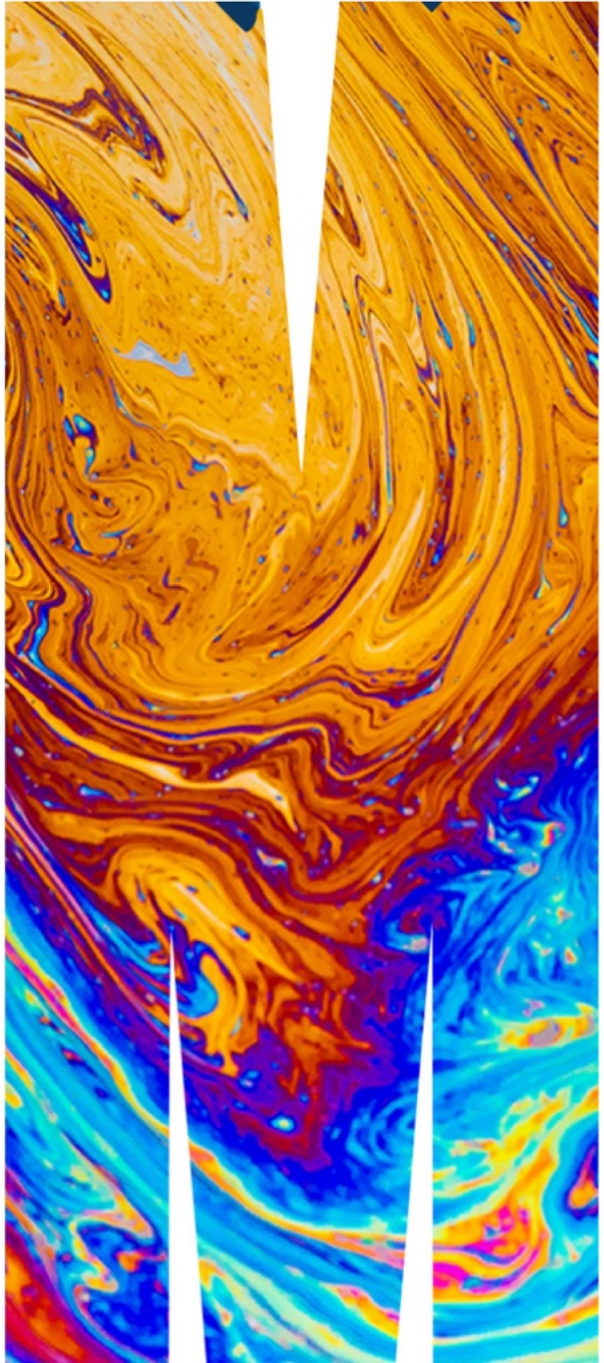
Code is here

Summary

Working with wild data can be daunting!

1. Start with questions that might be answered using the data.
2. Map out a pipeline to process the data, to address the question.
3. Think about what might be expected, so results can be "externally validated".

Slides originally developed by Professor Di Cook and maintained by Dr Kate Saunders



Lecturer: *Lecturer: Kate Saunders*

Department of Econometrics and Business Statistics

ETC5512.Clayton-x@monash.edu

Week 3