GDA of *subway* (from **scorr**)

Background Annual passenger entries for New York subway stations from 1905 to 2006.

Aims How many passengers entered the subway system each year? Which years were the stations open? Which stations were busiest? Do stations have similar passenger patterns?

Source xmdv's website (davis.wpi.edu/xmdv/datasets/subway.html). They got the data from Michael Frumin's website (frumin.net/ation/), although the data are no longer available there, and Frumin got the data from Jeff Zupan of RPA.

Structure 423 observations on 104 variables (102 numeric, 1 name, 1 factor)

The following three plots show the development of passenger numbers over the years, the numbers of stations open each year, and the relationship between these two summaries. Some initial code is needed to aggregate.

```
library(dplyr); library(gridExtra)
data(subway, package="scorr")
subway <- data
subway <- within(subway, Station <- rownames(subway))
rownames(subway) <- NULL
allPassengers <- apply(select(subway, X1905:X2006), 2, sum)
nStations <- apply(select(subway, X1905:X2006), 2, function(x) sum(x>0))
sTot <- data.frame(Year=1905:2006, allPassengers, nStations)
a1 <- ggplot(sTot, aes(Year, allPassengers/1000000)) + geom_line() + ylim(0, 2000) +
  xlab("") + ylab("Passengers (Millions)")
a2 <- ggplot(sTot, aes(Year, nStations)) + geom_line() + ylim(0, 1000) +
  ylab("Stations") + xlab("")
a3 <- ggplot(sTot, aes(nStations, allPassengers/1000000)) + geom_point() + geom_path() +
  xlab("Stations") + ylab("Passengers (Millions)") + ylim(0,2000)
grid.arrange(arrangeGrob(a1, a2, nrow=2), a3, ncol=2, widths=c(3,4))
```

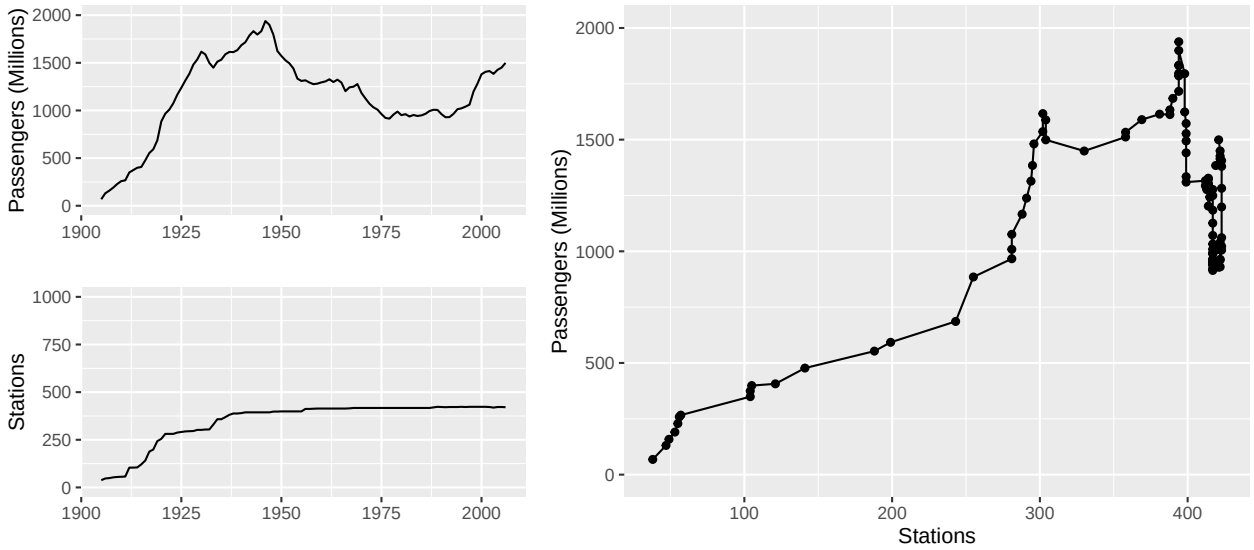

Fig 1: Total passenger numbers in the New York subway system from 1905 to 2006 (top left). Numbers of stations open in each year (lower left). Passenger numbers plotted against numbers of stations, joined in order of years (right). Both station and passenger numbers increased very fast in the first thirty years or so. Later there were few new stations and passenger numbers fell from their peak at the end of the 1940s to about half that in the mid 1970s. In the ten years from the mid 1990s passenger numbers rose again sharply.

The dataset includes many zeros. These must be for years when the stations were not open and so firstly these have all been replaced by NA (while keeping a copy of the original dataset).

```
subway0 <- subway
subway[subway==0] <- NA
```

A missing value plot shows the patterns of missings:

```
library(extracat)
visna(subway[,1:102], sort="r", sort.method="optile")
```

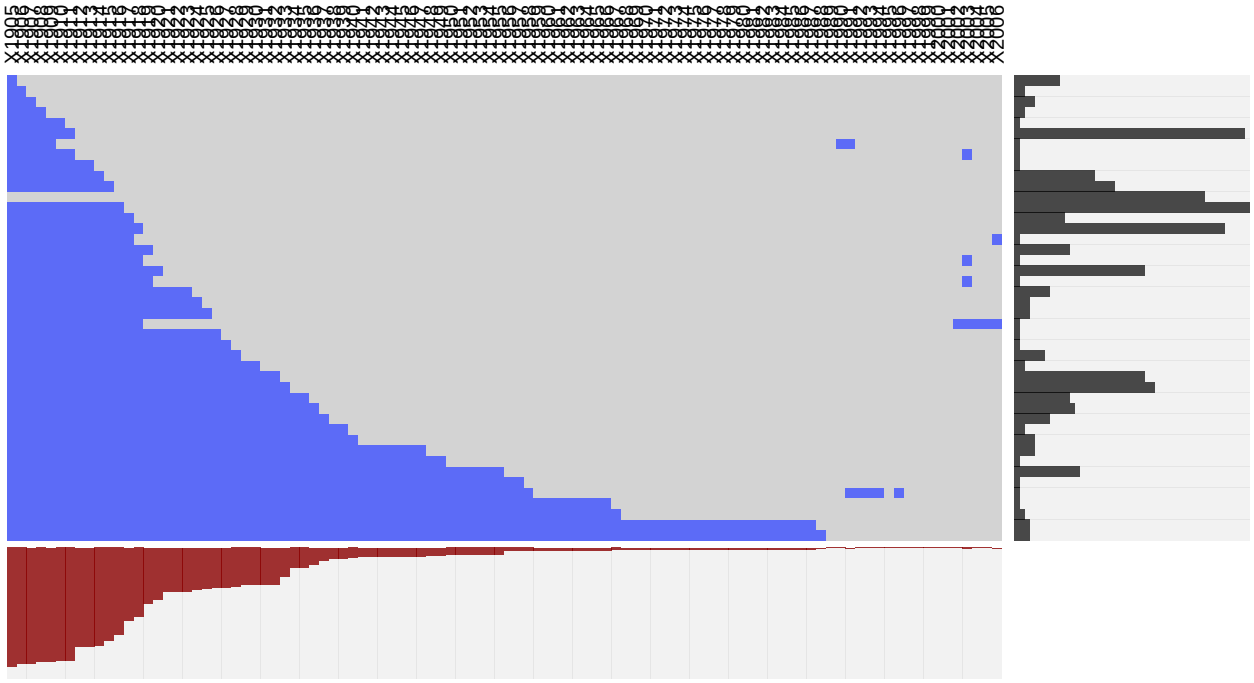

Fig 2: A plot of the years (columns) for which passenger numbers are missing. Stations which have the same sets of missing years are grouped together in the rows and the black bars in the right margin show the numbers of stations with those patterns. The all grey row (i.e., with no missings) represents the 38 stations with data for all of the years of this dataset. The isolated blue cells to the right are presumably years when stations were closed. For instance, the block of cells to the far right is because Cortlandt St had to be rebuilt after 9/11.

The numbers and names of the stations for a particular pattern can be found from code like these examples:

```
ww <- visna(subway[,1:102], sort="r", sort.method="optile", plot=FALSE)
rsum <- apply(ww, 1, sum) #numbers of years missing
rs <- attr(ww, "mar")$rm # number of stations in the rows
rs[rsum==0]
```

```
## [1] 38
```

```
subway %>% filter(is.na(X2006)) %>% select(Station)
```

```
##             Station
## 1  Cortlandt_St_(1)
## 2 Cortlandt_St_(RW)
```

The variable 'labels' records in which of the four boroughs a station is located. The parallel coordinate function in *GGally*, ggparcoord, offers a quick way of plotting multiple time series recorded at the same time points. As it demands that missing values are imputed in some way or other, the original version of the dataset with zeros has been used. Figure 3 shows the time series for the 67 stations in the Bronx, coloured by station.

```
library(GGally)
ggparcoord(subway0[subway0$labels=="Bronx",], 1:102, scale="globalminmax", groupColumn=
  "Station") + xlab("Years from 1905 to 2006") + ylab("Annual passenger numbers") +
  theme(axis.text.x=element_blank(), legend.position = "none")
```

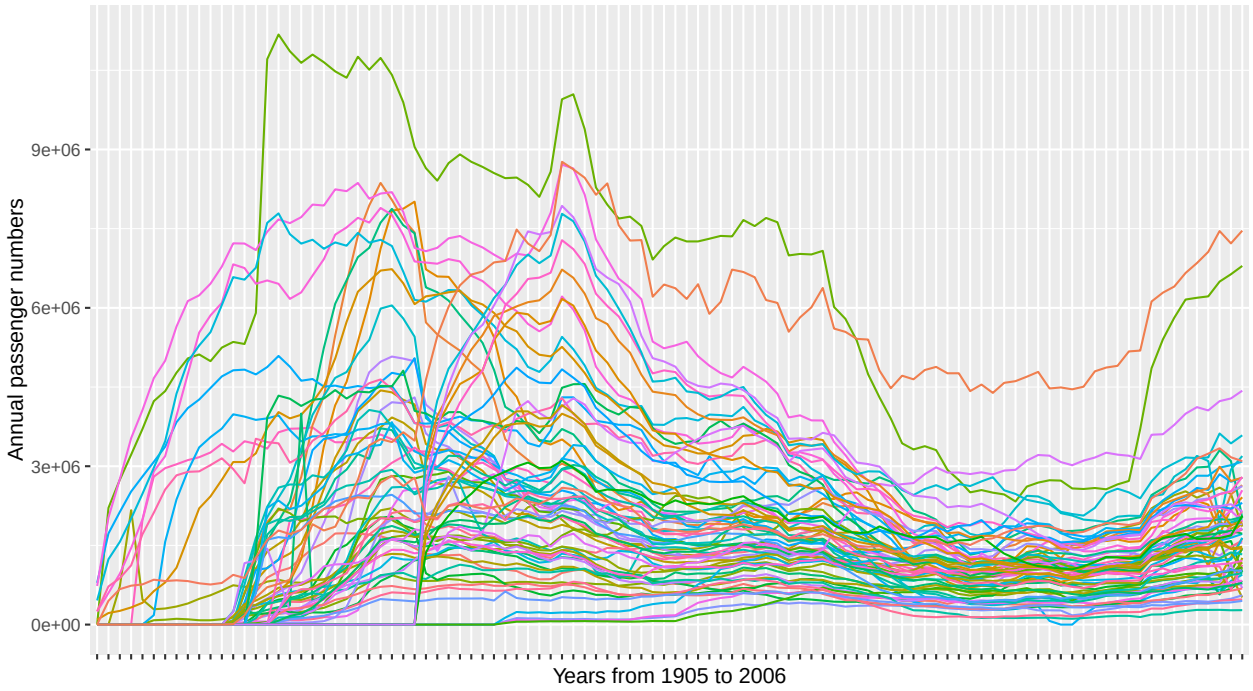

Fig 3: The time series of passenger data for each of the 67 stations in the Bronx. The two stations with the highest passenger numbers in 2006 had quite different patterns over the years. The isolated peak very early on for one station with otherwise low numbers is presumably an error. The group of lines rising sharply from zero in the 1930s must be a set of new stations, as are the the few low time series beginning at 1940.

Interestingly, the subway lines passing through a station are also recorded in the dataset, but in a rather awkward way within the station's name. While it is interesting to look at the patterns by lines, you have to bear in mind that the lines have been reorganised a few times (http://qz.com/549388/the-history-behind-new-york-citys-missing-subway-lines/).

While the static plot in Figure 4 gives an excellent first overview, many potentially interesting details are hidden. Interactive querying, zooming, highlighting, and alphablending would all help in spotting outliers and identifying patterns. Drawing smaller subsets of stations is fine, if you can work out which ones to draw.

One obvious selection is to look in more detail at the two groups of new stations picked out in Figure 3.

```
ggparcoord(subway0[subway0$labels=="Bronx" & subway0$X1925==0,], 27:102,
           scale="globalminmax", groupColumn="Station") +
  xlab("Years from 1931 to 2006") + ylab("Annual passenger numbers") +
  theme(axis.text.x=element_blank(), legend.position = "none")
```

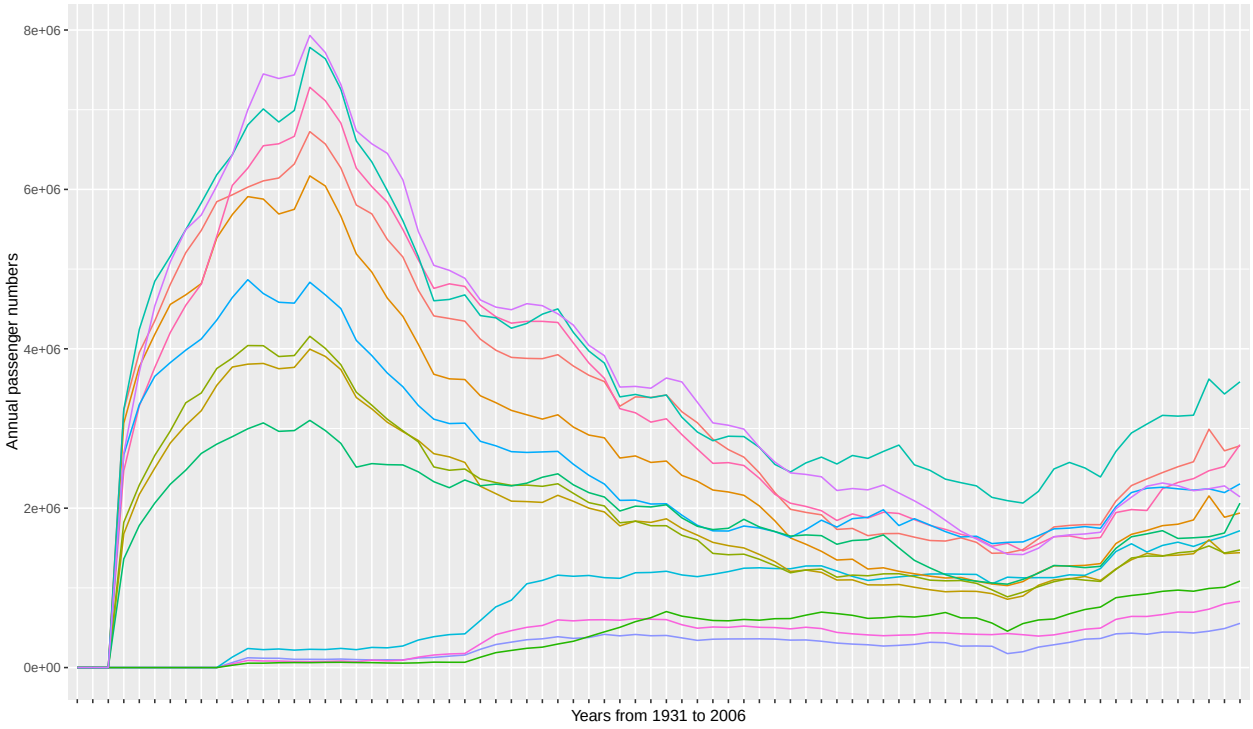

Fig 4: The time series of passenger data for the Bronx stations which came into subway service in 1934 and 1941. Each group has its own distinctive pattern.

Identifying outliers like the one mentioned in the caption to Figure 3 is difficult when there are so many series. An alternative approach is to transpose the data and look at boxplots of each station's data. An example is shown in Figure 5. Again, because there are so many stations, only the first 50 are shown.

```
subS <- data.frame(t(subway[,1:102]))
colnames(subS) <- subway$Station
library(tidyr)
msubS <- gather(subS)
ggplot(msubS[1:(102*50),], aes(key, value)) + geom_boxplot() + coord_flip()
```

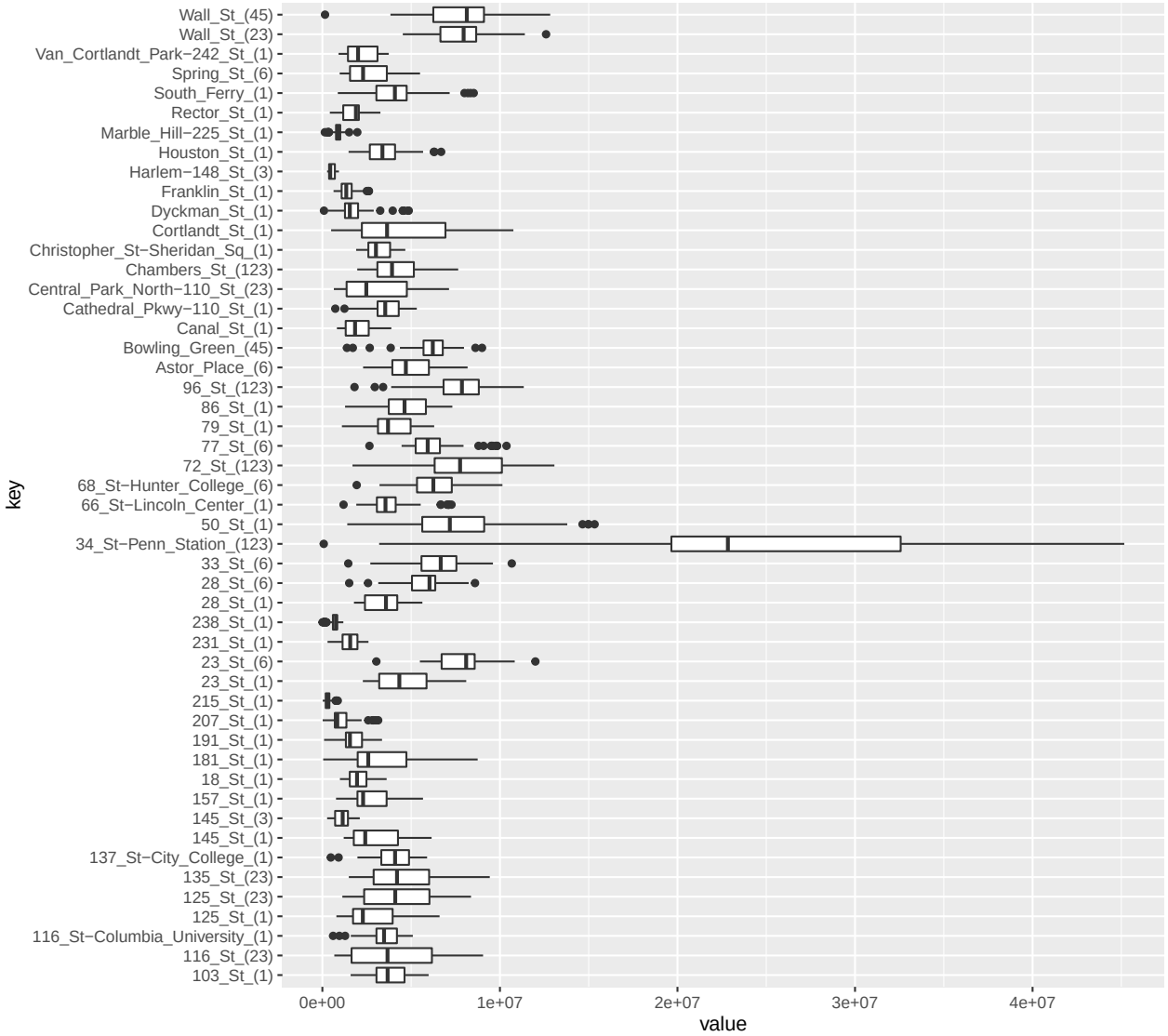

Fig 5: Boxplots of the yearly passenger numnbers for the first 50 stations listed in the dataset. Extreme individual outliers could be worth checking.

The transposed data could also be used to explore correlations between station passenger numbers over the years.

Further analyses of this dataset can be found on the webpage of *scorr* (http://mckennapsean.com/scorrplot/). It has also been used in a couple of papers by Dang and Wilkinson. Michael Frumin has a good discussion of local patterns and some nice spatial sparkline displays in his blog entry on the dataset (frumin.net/ation/2009/05/spark_it_up.html).

There are a few features of the dataset worth noting:

- There are no data for stations which are now permanently closed.
- In the form provided in *scorr* missing values are reported as 0, not ideal, especially when using statistics.
- The dataset treats the stations as cases (the rows) and the years as variables (the columns). Transposing the matrix to take the years as cases and the stations as variables can also be useful.
- The final variable is a label recording in which of four areas of New York the station is located. Of much more interest is surely what line(s) each station is on and that information is actually available in the Station variable, albeit in a rather inaccessible form.
- Frumin mashed the data using exact spatial locations of the stations and maps as backgrounds, a valuable approach for his small area analyses and displays.