Households PM2.5 concentrations in rural and urban Peru

Josiah Kephart

October 24, 2016

Final Analysis

Preliminaries

Load libraries

```
#knitr::opts_chunk$set(error = TRUE)

#load libraries
library(data.table)
library(plyr)
library(dplyr)
library(stringi)
library(ggplot2)
library(xtable)
library(pander)
library(pheatmap)
library(RColorBrewer)
```

Load data and create new columns

Add filenames as new field to capture household ID from filename

Combine files into one dataset

```
#read filenames
urban_files <- list.files(path = "../Data/PDRUrbanData", pattern = ".CSV")
rural_files <- list.files(path = "../Data/PDRRuralData", pattern = ".CSV")

#create new column of filename to capture household ID; concatenate files into one
dataset
read_csv_urban <- function(urban_files){
  ret <- fread(file.path("../Data/PDRUrbanData",urban_files), showProgress = TRUE,
autostart = 15)
  ret$Source <- urban_files
  ret
}

read_csv_rural <- function(rural_files){
  ret <- fread(file.path("../Data/PDRRuralData",rural_files), showProgress = TRUE,
autostart = 15)
  ret$Source <- rural_files
  ret
}

#NOTE: warning messages are from final row of raw data, which is gibberish and appr
opriately dropped by fread

urbanraw <- ldply(urban_files, read_csv_urban)
ruralraw <- ldply(rural_files, read_csv_rural)

urbandata <- urbanraw #this will be the working urban dataset
ruraldata <- ruralraw #this will be the working rural dataset
```

Clean up and create new columns

```
#Fix inconsistent column names in raw data files
colnames(urbandata) <- c("Point","Date","Time","pmurban1","Source","pmurban2")
colnames(ruraldata) <- c("Point","Date","Time","pmrural1","Source","pmrural2")

urbandata$pmurban <- rowSums(urbandata[,c("pmurban1","pmurban2")], na.rm = TRUE)
ruraldata$pmrural <- rowSums(ruraldata[,c("pmrural1", "pmrural2")], na.rm = TRUE)

#Isolate information derived from the filename, build date fields
#"hid" = household ID
urbandata$hid <- stri_sub(urbandata$Source,7,12)
urbandata$yr <- stri_sub(urbandata$Source, 19,20)
urbandata$ndate <- paste(urbandata$Date,urbandata$yr)
urbandata$ndatetime <- paste(urbandata$ndate, urbandata$Time)
urbandata$datetime <- lubridate::dmy_hms(urbandata$ndatetime)
urbandata$hour <- stri_sub(urbandata$Time,1,2)
urbandata$hidtime <- paste(urbandata$hid,urbandata$ndate,urbandata$hour, sep = "_")
urbandata$hiddt <- gsub(" ", "", urbandata$hidtime)
urbandata$strtdate <- stri_sub(urbandata$Source,14,20)
urbandata$sampleid <- paste0(urbandata$hid,urbandata$strtdate)
urbandata <- select_(urbandata, "hiddt", "pmurban","sampleid", "hid","datetime")

ruraldata$hid <- stri_sub(ruraldata$Source,7,12)
ruraldata$yr <- stri_sub(ruraldata$Source, 19,20)
ruraldata$ndate <- paste(ruraldata$Date,ruraldata$yr)
ruraldata$ndatetime <- paste(ruraldata$ndate, ruraldata$Time)
ruraldata$datetime <- lubridate::dmy_hms(ruraldata$ndatetime)
ruraldata$hour <- stri_sub(ruraldata$Time,1,2)
ruraldata$hidtime <- paste(ruraldata$hid,ruraldata$ndate,ruraldata$hour, sep = "_")
ruraldata$hiddt <- gsub(" ", "", ruraldata$hidtime)
ruraldata$strtdate <- stri_sub(ruraldata$Source,14,20)
ruraldata$sampleid <- paste0(ruraldata$hid,ruraldata$strtdate)
ruraldata <- select_(ruraldata, "hiddt", "pmrural","sampleid", "hid","datetime")
```

Head of working datasets

```
head(urbandata,3)
```

```
##               hiddt pmurban      sampleid    hid            datetime
## 1 210005_14Oct13_15   0.022 21000514OCT13 210005 2013-10-14 15:11:22
## 2 210005_14Oct13_15   0.012 21000514OCT13 210005 2013-10-14 15:12:22
## 3 210005_14Oct13_15   0.001 21000514OCT13 210005 2013-10-14 15:13:22
```

```
head(ruraldata,3)
```

```
##               hiddt pmrural      sampleid    hid            datetime
## 1 222123_18Sep99_09   0.001 22212318SEP99 222123 1999-09-18 09:36:33
## 2 222123_18Sep99_09   0.055 22212318SEP99 222123 1999-09-18 09:37:33
## 3 222123_18Sep99_09   0.027 22212318SEP99 222123 1999-09-18 09:38:33
```

Exploratory Data Analysis

How many measurements per sampling period

Used in first paragraph of results

```
#Calculate number and hours of measurements per sample: urban
samp_u <- length(unique(urbandata$sampleid))
logs_u <- length(urbandata$hiddt)

logs_per_samp_u <- (logs_u)/(samp_u)
logs_per_samp_u # Measurements per sample: urban
```

```
## [1] 1499.232
```

```
hrs_per_samp_u <- logs_per_samp_u/60
hrs_per_samp_u # Measurement-hours per sample: urban
```

```
## [1] 24.9872
```

```
#Calculate number of measurements per sample: rural
samp_r <- length(unique(ruraldata$sampleid))
logs_r <- length(ruraldata$hiddt)

logs_per_samp_r <- (logs_r)/(samp_r)
logs_per_samp_r # Measurements per sample: rural
```

```
## [1] 1470.539
```

```
hrs_per_samp_r <- logs_per_samp_r/60
hrs_per_samp_r # Measurement-hours per sample: rural
```

```
## [1] 24.50899
```

Summarize data by daily hour (mean, max, median, 95th percentile)

Extract ID and hour into new columns

```
quant_u <- "quantile(pmurban, probs=0.95)"
quant_r <- "quantile(pmrural, probs=0.95)"

funs_u <- c("mean", "max", "median",quant_u)
funs_r <- c("mean", "max", "median",quant_r)

sumdata_urban <- urbandata %>%
  group_by(hiddt) %>%
  summarise_at(vars(pmurban), funs_u) %>%
  mutate(hid = stri_sub(hiddt, 1,6)) %>%
  mutate(hr = stri_sub(hiddt, -2,-1))

sumdata_rural <- ruraldata %>%
  group_by(hiddt) %>%
  summarise_at(vars(pmrural), funs_r) %>%
  mutate(hid = stri_sub(hiddt, 1,6)) %>%
  mutate(hr = stri_sub(hiddt, -2,-1))

head(sumdata_urban,3)
```

```
## # A tibble: 3 × 7
##               hiddt        mean   max median quantile    hid    hr
##               <chr>       <dbl> <dbl>  <dbl>    <dbl>  <chr> <chr>
## 1 210005_14Oct13_15 0.004448980 0.027  0.003   0.0138 210005    15
## 2 210005_14Oct13_16 0.006000000 0.058  0.004   0.0147 210005    16
## 3 210005_14Oct13_17 0.004916667 0.018  0.003   0.0141 210005    17
```

```
head(sumdata_rural,3)
```

```
## # A tibble: 3 × 7
##               hiddt     mean    max median quantile    hid    hr
##               <chr>    <dbl>  <dbl>  <dbl>    <dbl>  <chr> <chr>
## 1 222009_05Jun15_06 0.451000  0.451  0.451   0.4510 222009    06
## 2 222009_05Jun15_07 9.253100 77.100  5.235  37.9750 222009    07
## 3 222009_05Jun15_08 1.601733 16.430  0.018   8.4435 222009    08
```

Summary statistics for hourly-summarized data

All based on median level per hour

Mean, median, 95th percentile, standard deviation of median levels per hour

```
#urban
urbanpm <- sumdata_urban %>%
  summarise(mean = mean(median))

urbanpm$median <- sumdata_urban %>%
  summarise(median = median(median))

urbanpm$perc95 <- sumdata_urban %>%
  summarise(perc95 = quantile(median,probs = 0.95))

urbanpm$sd <- sumdata_urban %>%
  summarise(sd = sd(median))

#rural
ruralpm <- sumdata_rural %>%
  summarise(mean = mean(median))

ruralpm$median <- sumdata_rural %>%
  summarise(median = median(median))

ruralpm$perc95 <- sumdata_rural %>%
  summarise(perc95 = quantile(median,probs = 0.95))

ruralpm$sd <- sumdata_rural %>%
  summarise(sd = sd(median))
```

Create table

```
sumstats <- bind_rows(urbanpm,ruralpm)
row.names(sumstats) <- c("Urban PM2.5","Rural PM2.5")
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
colnames(sumstats) <- c("Mean","Median","95th Percentile","Standard Deviation")

pander(sumstats)
```

	Mean	Median	95th Percentile	Standard Deviation
Urban PM2.5	0.01621484	0.005	0.048	0.132402
Rural PM2.5	0.66385067	0.003	2.066075	5.693289

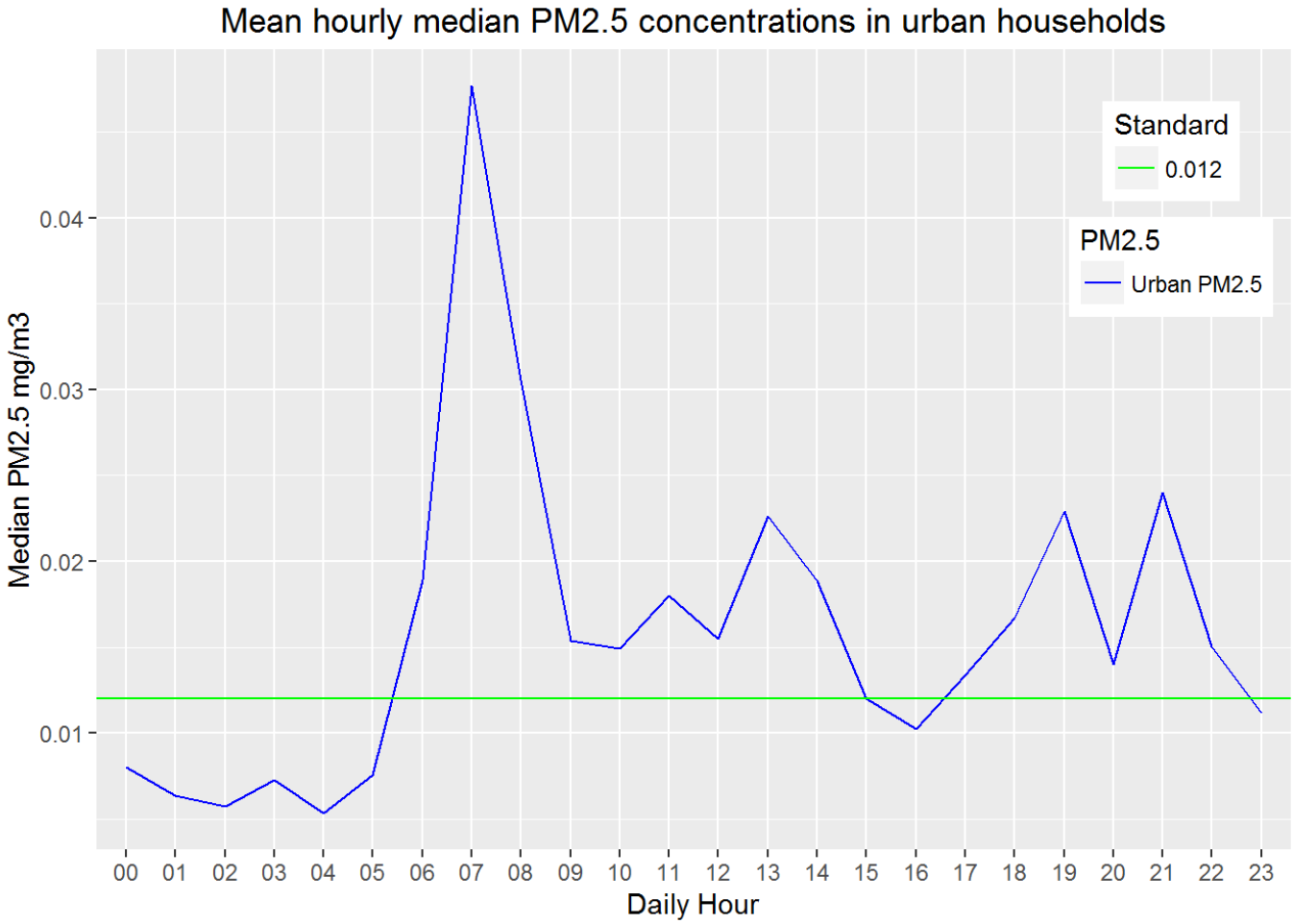
Summarize data by hour of day

```
hrurban <- sumdata_urban %>%
  group_by(hr) %>%
  summarise_at(vars(median),mean)

hrrural <- sumdata_rural %>%
  group_by(hr) %>%
  summarise_at(vars(median),mean)
```

Generate plots

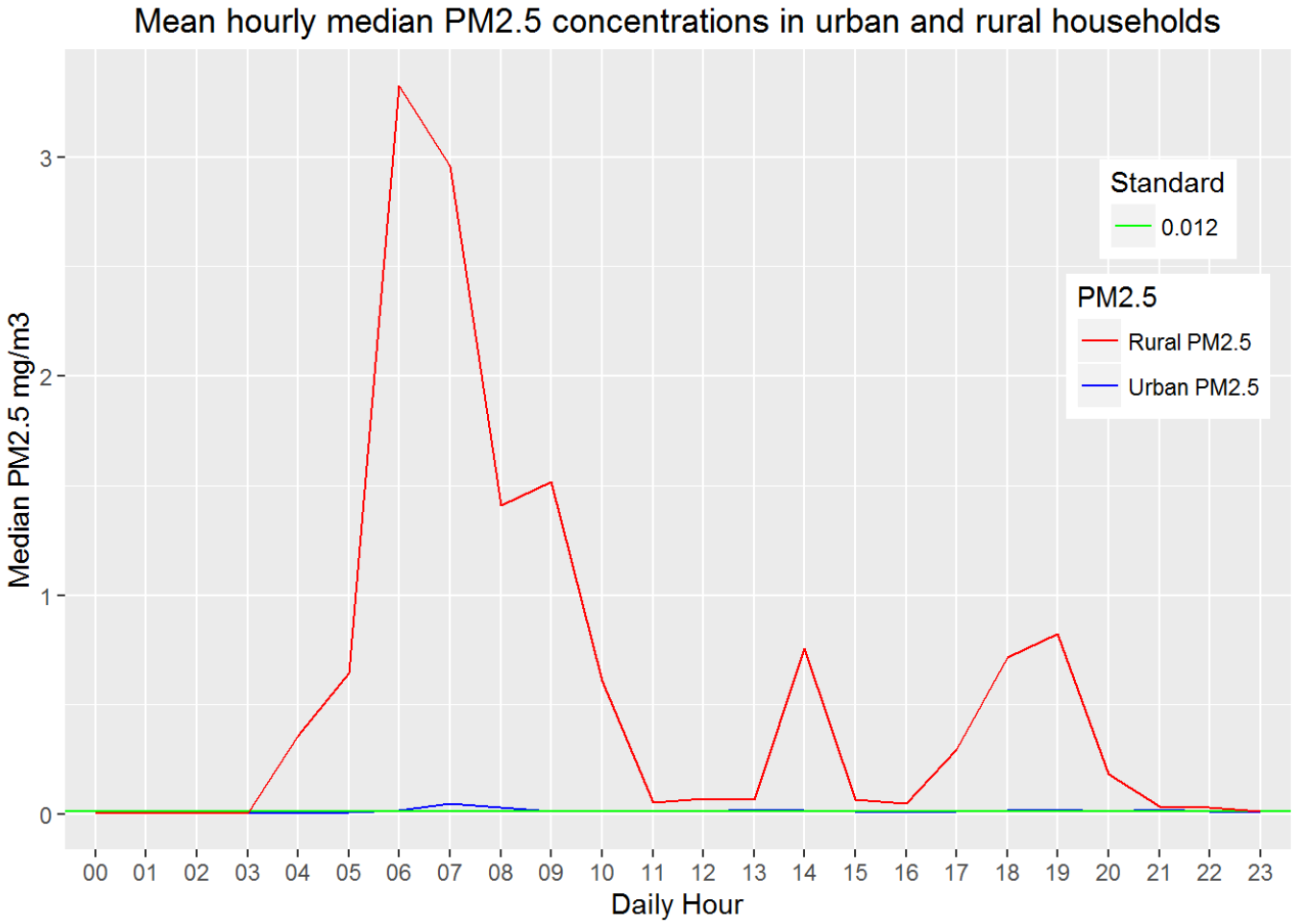
Line plots

```
#Plot mean hourly household median

p <- ggplot(data = hrurban, aes(x=hr,y=median, color = "Urban PM2.5", group = 1))
p <- p + geom_line(aes(hr,median, color="Urban PM2.5"))
Standard <- data.frame( x = c(-Inf, Inf), y = 0.012, Standard = factor(0.012) )
p <- p + geom_line(aes(x, y, linetype = Standard), Standard, colour = "green")
p <- p + ylab("Median PM2.5 mg/m3")
p <- p + xlab("Daily Hour")
p <- p + scale_color_manual(values = c("blue"))
p <- p + ggtitle("Mean hourly median PM2.5 concentrations in urban households")
p <- p + theme(legend.position = c(0.9,0.8))
p <- p + guides(colour = guide_legend(title = "PM2.5"))
p
```




```
## Saving 7 x 5 in image
```

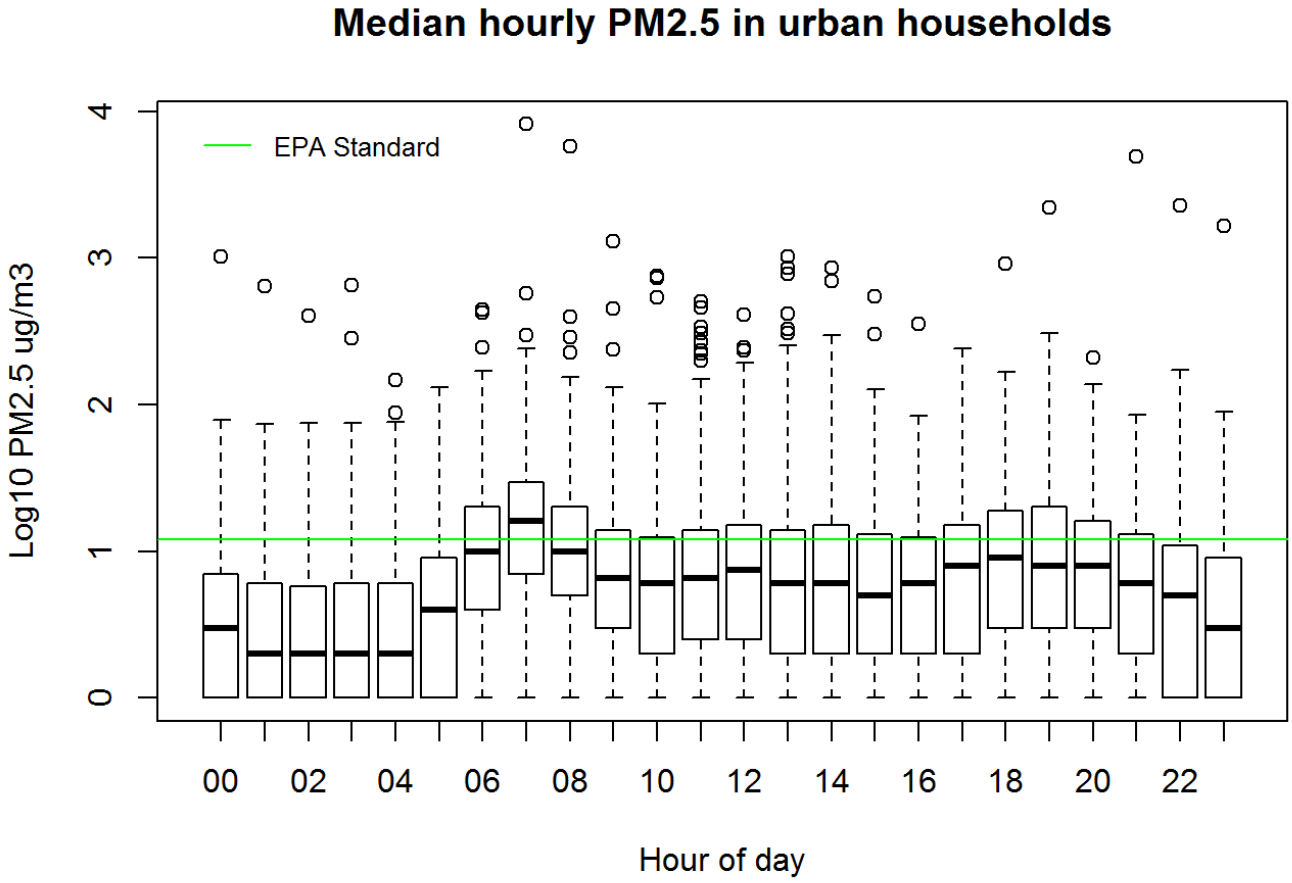
```
p + geom_line(data = hrrural, aes(hr,median, group = 1, color = "Rural PM2.5")) +
  scale_color_manual(labels = c("Rural PM2.5","Urban PM2.5"), values = c("red","bl
ue")) +
  theme(legend.position = c(0.9,0.7)) +
  ggtitle("Mean hourly median PM2.5 concentrations in urban and rural households")
```




```
## Saving 7 x 5 in image
```

Plot boxplots

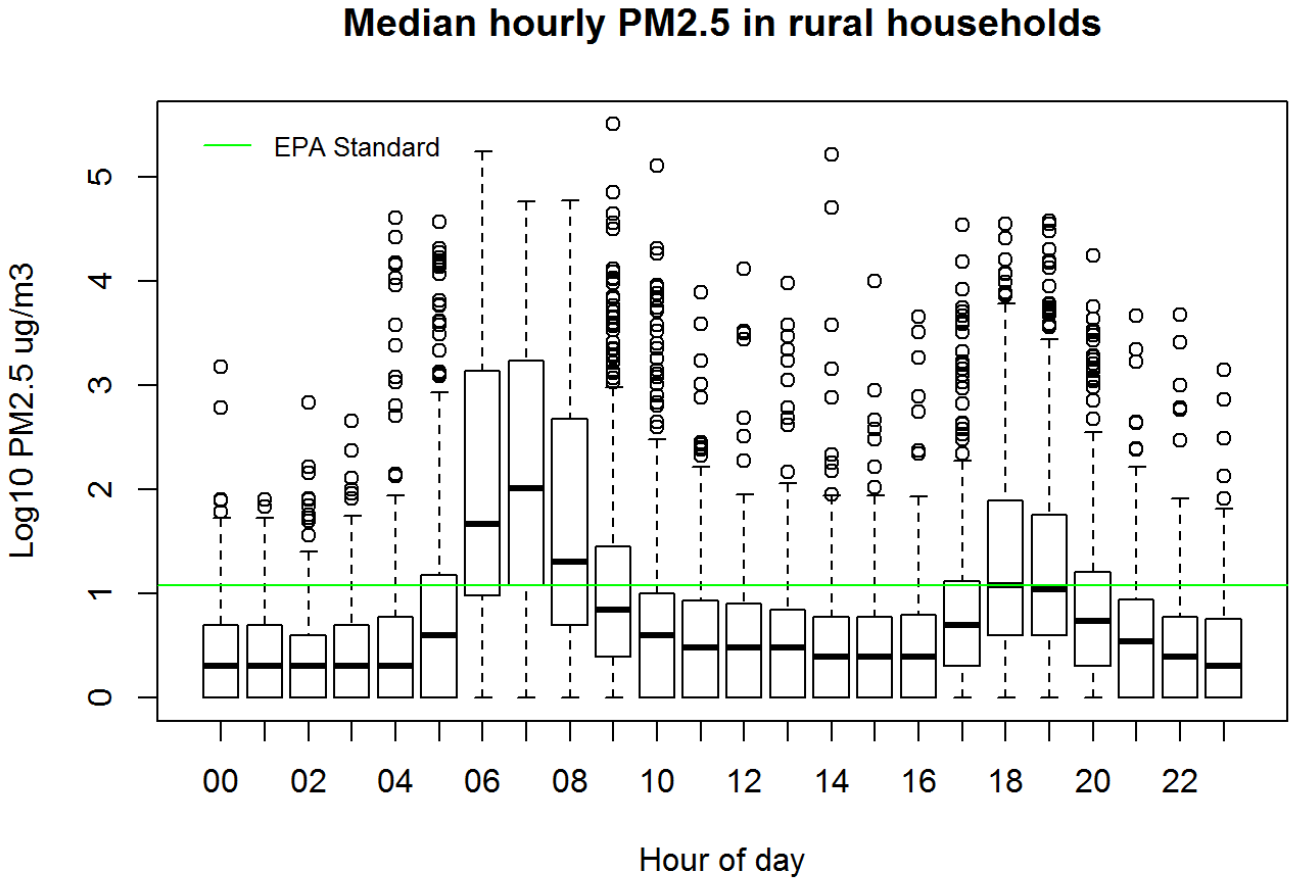
```
#Convert data to log base 10 for better visualization

sumdata_urban_log <- sumdata_urban %>%
  mutate(logurban = log10(median*1000+1))

sumdata_rural_log <- sumdata_rural %>%
  mutate(logrural = log10(median*1000+1))

#Plot urban box

boxplot(logurban ~ hr, data=sumdata_urban_log, main="Median hourly PM2.5 in urban
households", xlab="Hour of day",ylab="Log10 PM2.5 ug/m3")
abline(h=log10(0.012*1000), col="green")
legend("topleft", inset=.02, "EPA Standard", col="green",lty=1,horiz=TRUE, cex=0.8,
box.lty = 0)
```




```
## jpeg
##    3
```

```
## png
##   2
```

```
boxplot(logrural ~ hr, data=sumdata_rural_log, main="Median hourly PM2.5 in rural
households", xlab="Hour of day",ylab="Log10 PM2.5 ug/m3")
abline(h=log10(0.012*1000), col="green")
legend("topleft", inset=.02, "EPA Standard", col="green",lty=1,horiz=TRUE, cex=0.8,
box.lty = 0)
```

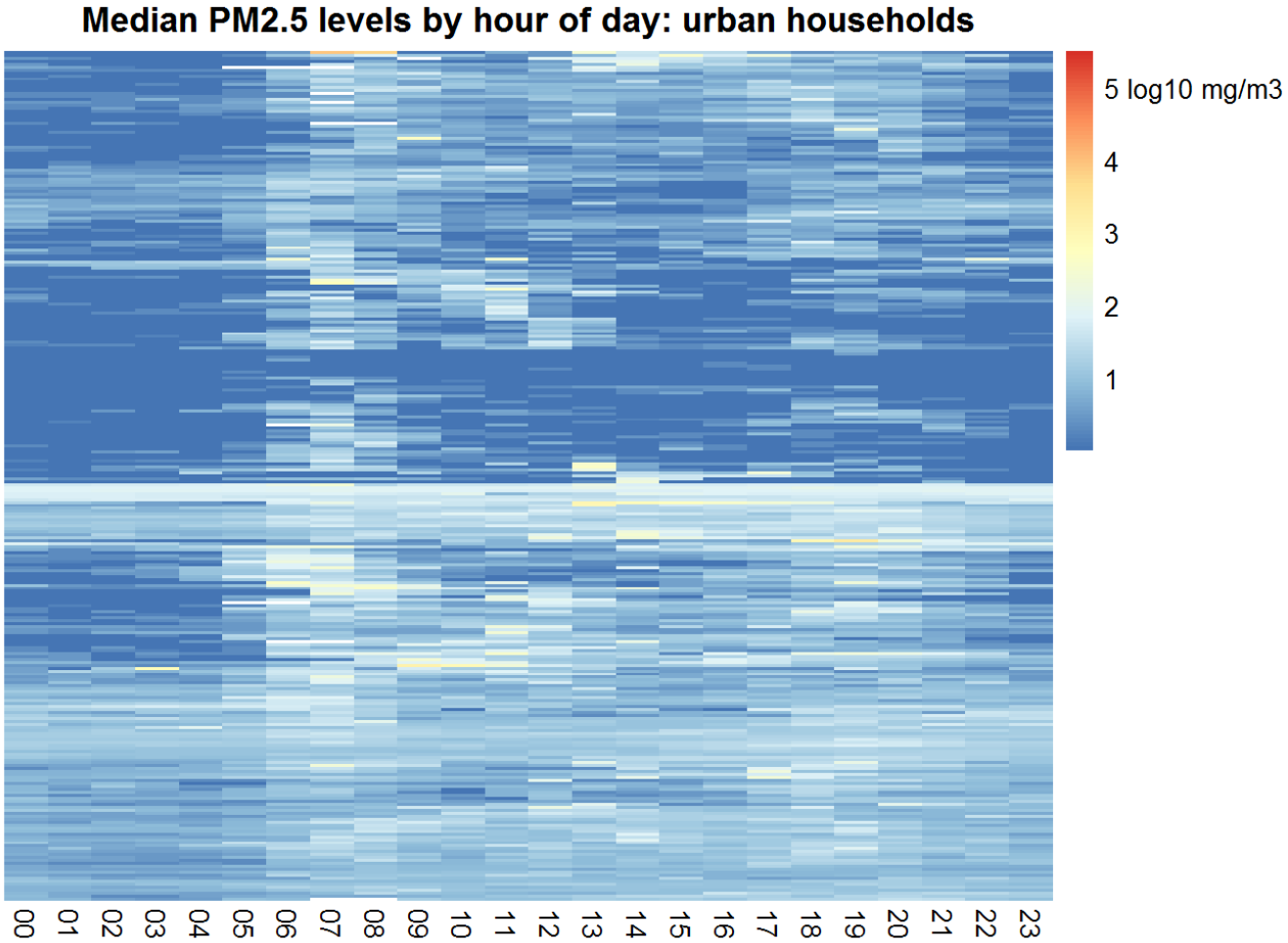



```
## jpeg
##    3
```

```
## png
##   2
```

Plot heat maps

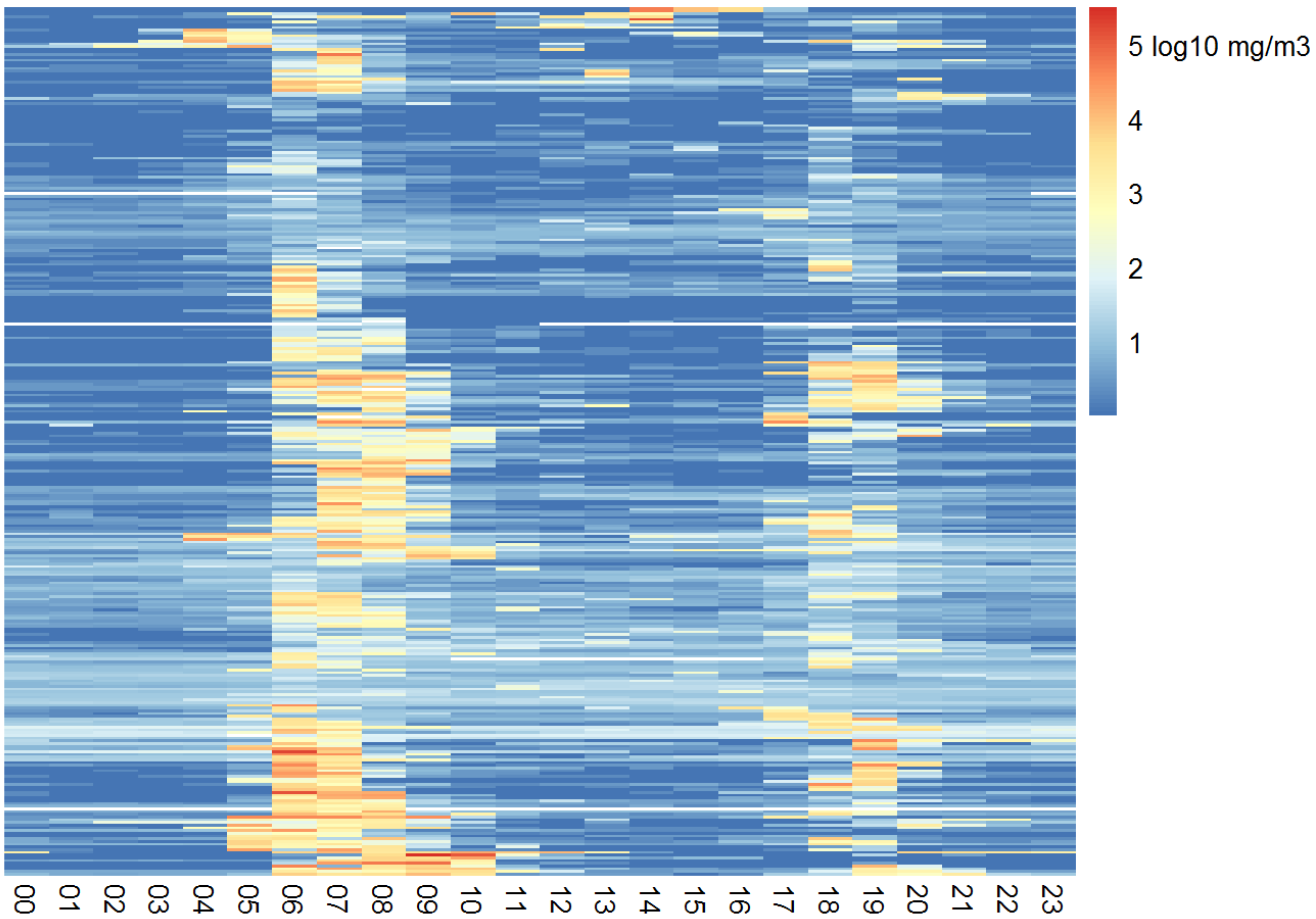
```
#reformat data to wide format for heatmap
num_r <- as.data.frame(sumdata_rural_log)
nums_r <- num_r %>%
  select(hid, hr, logrural)
wide_r <- reshape(nums_r, idvar = "hid", timevar = "hr", direction = "wide")

num_u <- as.data.frame(sumdata_urban_log)
nums_u <- num_u %>%
  select(hid, hr, logurban)
wide_u <- reshape(nums_u, idvar = "hid", timevar = "hr", direction = "wide")

#drop and reorder columns for plotting, reformat as matrix
wide2_r <- wide_r %>%
  select(-hid)
wide3_r <- wide2_r[,c(19:24,0:18)]
ruralmat <- as.matrix(wide3_r)

wide2_u <- wide_u %>%
  select(-hid)
wide3_u <- wide2_u[,c(10:24,0:9)]
urbanmat <- as.matrix(wide3_u)

#Assign column labels, breaks, colors for plots

collabels <- c("00","01","02","03","04","05","06","07","08","09",(10:23))
heatbrks <- c(1:5)
heatlbs <- c("1","2","3","4","5 log10 mg/m3")

color = colorRampPalette(rev(brewer.pal(n = 7,
                                        name = "RdYlBu")))(100)
n = length(color)
x = c(urbanmat, ruralmat)
breaks = seq(min(x, na.rm = TRUE),
             max(x, na.rm = TRUE),
             length.out = n +
            1)

#Plot heat maps
pheatmap(urbanmat,
         cluster_rows = TRUE, cluster_cols = FALSE,
         treeheight_row = 0,
         labels_col = collabels,
         legend = TRUE,
         legend_breaks = heatbrks,
         breaks = breaks,
         legend_labels = heatlbs,
         color = colorRampPalette(rev(brewer.pal(n = 7, name ="RdYlBu")))(100),
main="Median PM2.5 levels by hour of day: urban households",
         silent = FALSE,
         show_rownames = FALSE)
```




```
## jpeg
##     3
```

```
## png
##   2
```

```
pheatmap(ruralmat,
         cluster_rows = TRUE, cluster_cols = FALSE,
         treeheight_row = 0,
         labels_col = collabels,
         legend = TRUE,
         legend_breaks = heatbrks,
         legend_labels = heatlbs,
         color = colorRampPalette(rev(brewer.pal(n = 7, name ="RdYlBu")))(100),
         breaks = breaks,
         main="Median PM2.5 levels by hour of day: rural households",
         silent = FALSE,
         show_rownames = FALSE)
```


Median PM2.5 levels by hour of day: rural households
5 log10 mg/m3
4
3
2
1
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23

```
## jpeg
##     3
```

```
## png
##   2
```