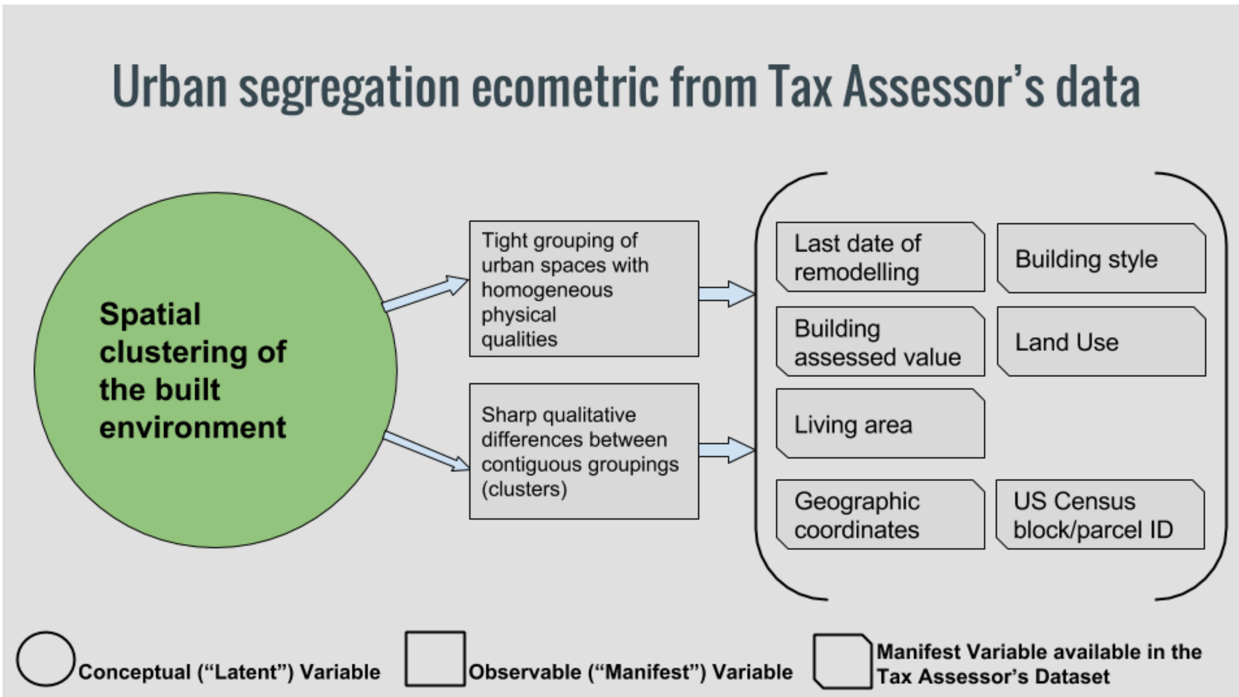
Tax Assessor's data: Building latent constructs

Hector Antonio Vazquez Brust

October 30, 2015

Last week, we propposed a couple of build environment ecometrics as instruments to measure urban segregation:



Now we'll start exploring how to define these ecometrics as arithmetics functions that can be coded in R.

First of all, let's examine the available variables in the TAx Assessor's dataset. Since we are looking for an index of heterogeneity of the built environment, good candidates for our purposes are:

- LU (Land Use)
- LAND_SF
- GROSS_AREA (Gross floor area for commercial properties.)
- LIVING_AREA (Total living area for residential properties)
- YR_BUILT
- R_BLDG_STYL (Residential building style)
- S_BLDG_STYL (Condo main building style)
- BRA_PD (Boston Redevelopment Authority Planning District)
- AV_LAND_PER_SF (Assessed land value per square foot)
- AV_BLDG_PER_SF (Assesed building value per square)

Let's begin with Land Use.

Before we start our analysis, we'll fix a single parcel in the entire dataset with LU = "XX", a land use code that is not registered in the datasat dictionary. Since the parcel is owned by a church, we'll assume it is a Tax Exempt parcel.

```
TAdata[TAdata$LU == 'XX',]$LU <- "E"
```

Now, we can see that there are 17 different land uses codes in our dataset:

```
unique(TAdata$LU)
```

```
##  [1] R3 R2 R1 R4 C  RC A  E  RL CM CD CL I  CC CP EA AH
## Levels: A AH C CC CD CL CM CP E EA I R1 R2 R3 R4 RC RL XX
```

```
length(unique(TAdata$LU))
```

```
## [1] 17
```

So we can try counting how many land uses are present in each neighbourhood:

```
aggregate(LU~BRA_PD, data = TAdata, function(x) length(unique(x)))
```

```
##                  BRA_PD LU
## 1      Allston/Brighton 16
## 2  Back Bay/Beacon Hill 16
## 3               Central 16
## 4           Charlestown 16
## 5           East Boston 16
## 6        Fenway/Kenmore 16
## 7             Hyde Park 16
## 8          Jamaica Plain 17
## 9              Mattapan 15
## 10     North Dorchester 16
## 11           Roslindale 16
## 12              Roxbury 16
## 13         South Boston 16
## 14     South Dorchester 16
## 15            South End 16
## 16         West Roxbury 15
```

That wasn't very informative; every neighbourhood contains most of the land uses. (Jamaica Plain is the only neighborhood with the complete collection!). We'll need to develop a much more nuanced way to analize land use distribution to reach interesting conclusions.

What about land area? We can find the variance of land area as a measure of how diverse (in size) plots are in each neighborhood:

```
plot.size.variance <- aggregate(LAND_SF~BRA_PD, data = TAdata, var)
plot.size.variance[with(plot.size.variance, order(-LAND_SF)),]
```

```
##                  BRA_PD        LAND_SF
## 5           East Boston 81435130945419
## 13         South Boston   108326696891
## 9              Mattapan    93475221315
## 10     North Dorchester    88757732210
```

```
## 7              Hyde Park    49998965140
## 1       Allston/Brighton    28853142247
## 11            Roslindale    27728570599
## 16          West Roxbury    14842924471
## 6         Fenway/Kenmore     9283811205
## 8          Jamaica Plain     8553700629
## 4            Charlestown     5264503588
## 3                Central     1721754930
## 14     South Dorchester     1247604914
## 12               Roxbury      560740689
## 15             South End      482138424
## 2  Back Bay/Beacon Hill      331273436
```

East Boston has by far the largest plot size variance, and Back Bay the lowest.

Is the average land value in a neighbourhood somehow related to the plot size diversity?

We can find out if we aggregate the average land value per sqf and plot it agains plot size variance:

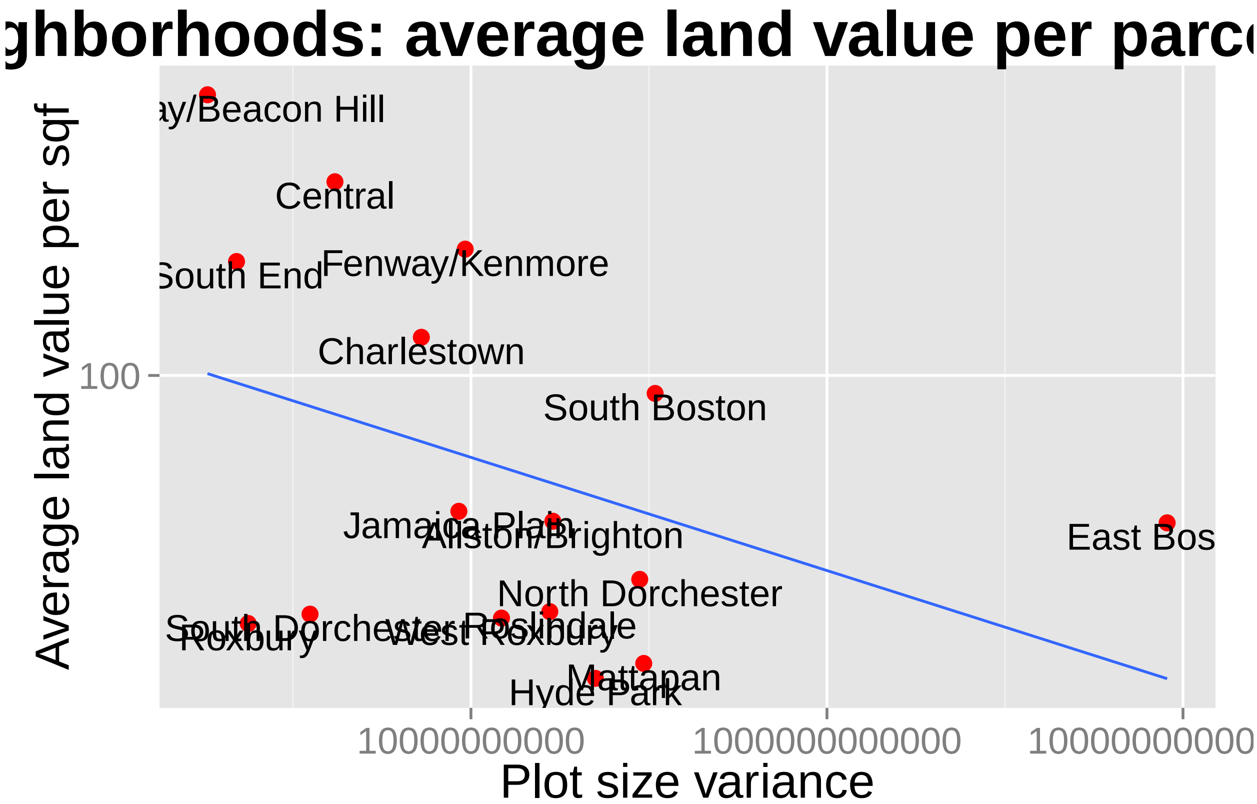
```
land.value.mean <- aggregate(AV_LAND_PER_SF~BRA_PD, data = TAdata, mean)

plot.variance.land.value <- merge(plot.size.variance, land.value.mean, by = 'BRA_PD')

names(plot.variance.land.value)[2:3] <- c("VARIANCE_LAND_SF","MEAN_AV_LAND_SF" )

library(ggplot2)

ggplot(plot.variance.land.value, aes(x = VARIANCE_LAND_SF, y = MEAN_AV_LAND_SF)) +
  geom_point(colour = 'red', size =3) +
  geom_smooth(method=lm, se=FALSE) +
  scale_x_log10() + scale_y_log10() + geom_text(aes(label=BRA_PD),vjust=1) +
  labs(title='Boston neighborhoods: average land value per parcel size variance',
       x='Plot size variance',y='Average land value per sqf') +
  theme(plot.title = element_text(size = 24, face="bold"),
        axis.title.x = element_text(size=18),
        axis.title.y = element_text(size=18),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=14))
```



The correlation is stronger than what I expected!

Let's now compare the mean age of buildings with their mean assessed value.

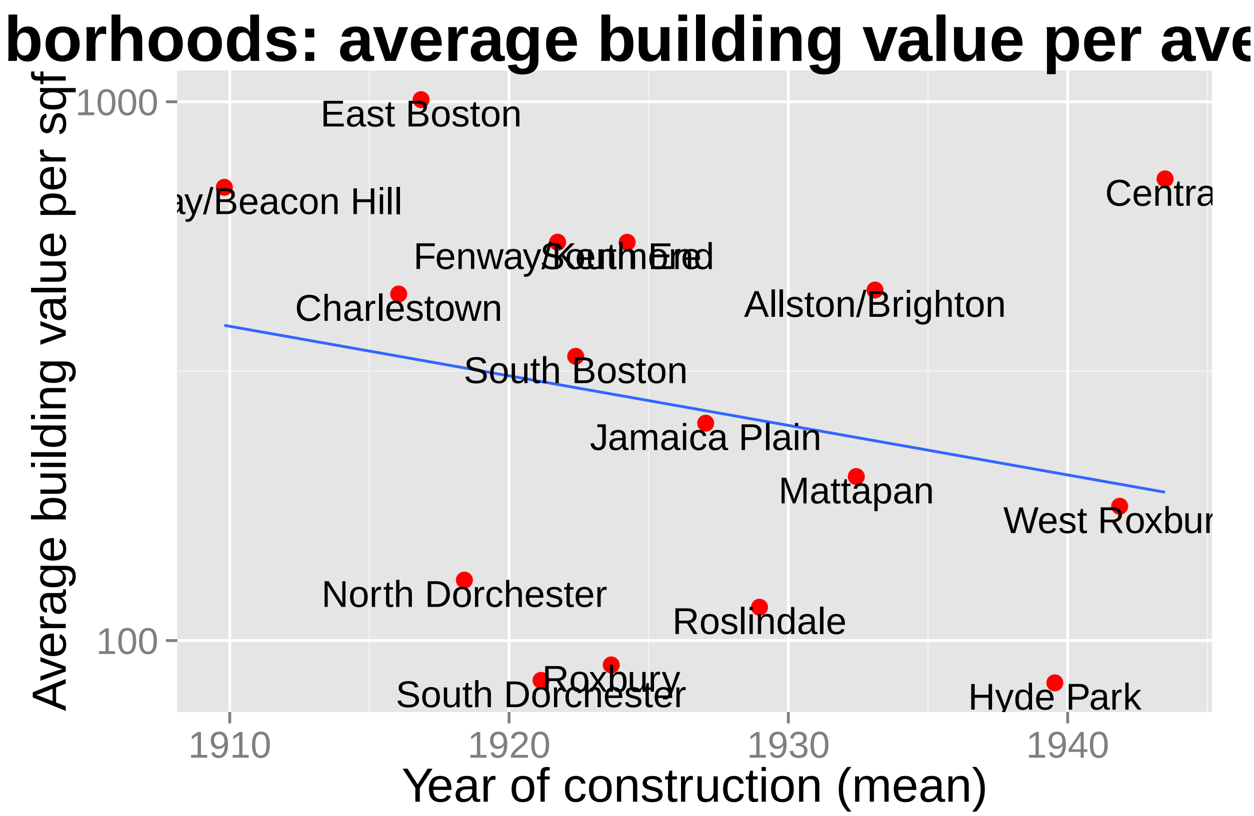
```
bldg.value.mean <- aggregate(as.integer(AV_BLDG_PER_SF)~BRA_PD, data = TAdata, mean)

year.mean <- aggregate(YR_BUILT~BRA_PD, data = TAdata, mean)

year.mean.bldg.value <- merge(year.mean, bldg.value.mean, by = 'BRA_PD')

names(year.mean.bldg.value)[2:3] <- c("MEAN_YR_BUILT","MEAN_AV_BLDG_SF" )

ggplot(year.mean.bldg.value, aes(x = MEAN_YR_BUILT, y = MEAN_AV_BLDG_SF)) +
  geom_point(colour = 'red', size =3) +
  geom_smooth(method=lm, se=FALSE) +
  scale_y_log10() + geom_text(aes(label=BRA_PD),vjust=1) +
  labs(title='Boston neighborhoods: average building value per average building age',
       x='Year of construction (mean)',y='Average building value per sqf') +
  theme(plot.title = element_text(size = 24, face="bold"),
        axis.title.x = element_text(size=18),
        axis.title.y = element_text(size=18),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=14))
```



Another interesting result: older building stocks tend to be more "valuable" that the more recent ones.

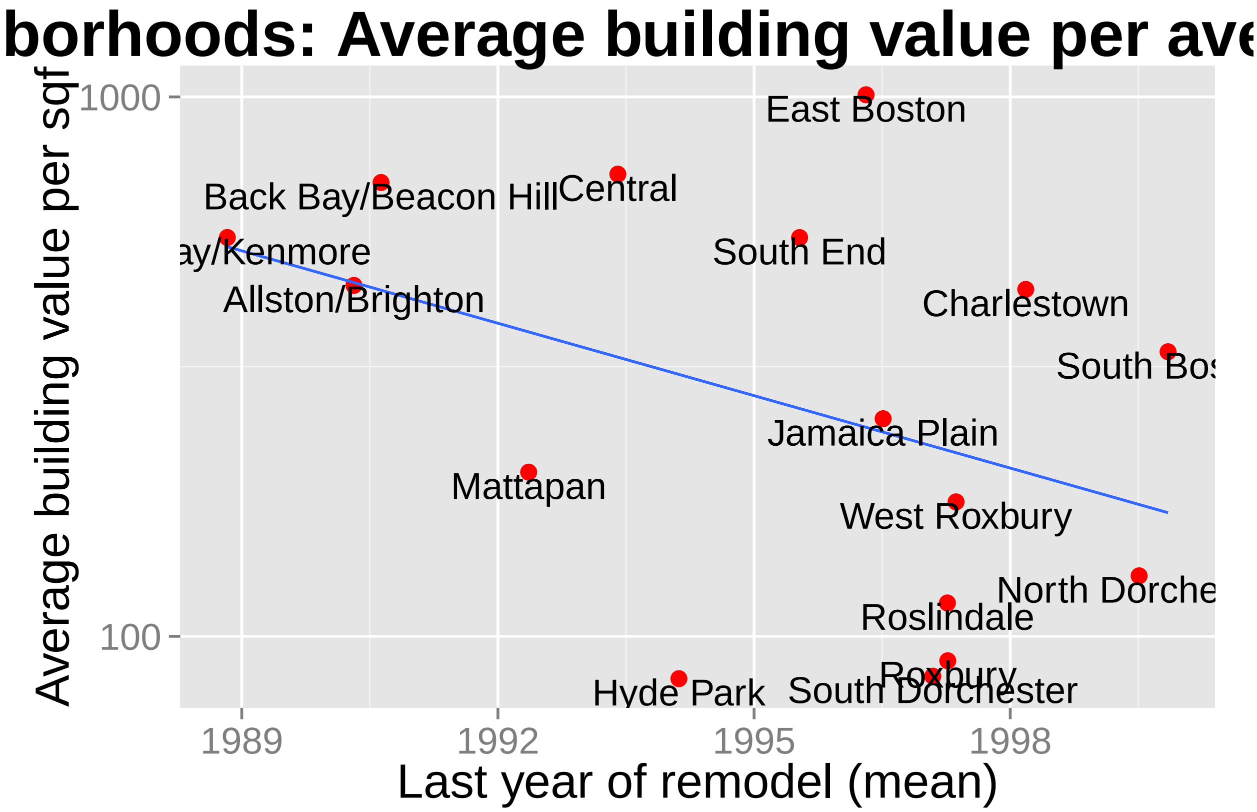
What about last year of remodelling and assessed building value?

```
remodel.year.mean <- aggregate(YR_REMOD~BRA_PD, data = TAdata, mean)

remodel.year.mean.bldg.value <- merge(remodel.year.mean, bldg.value.mean, by = 'BRA_PD')

names(remodel.year.mean.bldg.value)[2:3] <- c("MEAN_YR_REMODEL","MEAN_AV_BLDG_SF" )

ggplot(remodel.year.mean.bldg.value, aes(x = MEAN_YR_REMODEL, y = MEAN_AV_BLDG_SF)) +
  geom_point(colour = 'red', size =3) +
  geom_smooth(method=lm, se=FALSE) +
  scale_y_log10() + geom_text(aes(label=BRA_PD),vjust=1) +
  labs(title='Boston neighborhoods: Average building value per average last remodel',
      x='Last year of remodel (mean)',y='Average building value per sqf') +
  theme(plot.title = element_text(size = 24, face="bold"),
        axis.title.x = element_text(size=18),
        axis.title.y = element_text(size=18),
        axis.text.x = element_text(size=14),
        axis.text.y = element_text(size=14))
```



This seems unintuitive: The neighborhoods with higher building assessed values tend to be those with less recent remodels. And this correlation is stronger than that of building age and assessed value.

Also worth mentioning, a repeating outlier in all of these plots is East Boston.

We'll turn our attention again to land use. The dataset includes 17 different tipes of land use, whic we'll simplify into a more manageable group of nine: Residential, Commercial, Condo, Mixed Residential/Commercial, Agricultural, Industrial, Tax Excempt, and Tax Exempt by the Boston Redevelopmnet Authority (this last categories applies to parcels that are undergoing renovation projects)

```
simplify_LU <- function(LU) {
  if (LU %in% c("R1", "R2", "R3", "R4", "RL", "A")) {
      return("RESIDENTIAL")
    } else if (LU %in% c("CM", "CP")) {
      return("CONDO")
    } else if (LU == "CD") {
      return("CONDO_UNIT")
    } else if (LU == "RC") {
      return("MIX_RC")
    } else if (LU %in% c("CC", "C", "CL")) {
      return("COMMERCIAL")
    } else if (LU == "AH") {
      return("AGRICULTURAL")
    } else if (LU == "I") {
```

```
      return("INDUSTRIAL")
    } else if (LU == "E") {
      return("TAX_EXEMPT")
    } else if (LU == "EA") {
      return("TAX_EXEMPT_BRA")
    } else {
      return(NA)
    }
}


#Create a new column by applying the simplifyLU function
TAdata <- transform(TAdata, SIMPLIFIED_LU = sapply(LU, simplify_LU))
```

With the help of our new simplified land use code, we'll create a table showing what percentage of total parcel area each land use represents by neighborhood:

```
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(tidyr)

# Remove Condo Units as we are now interested in neighborhood total areas and CU areas should not be su
TAdata.no.condo.units <- filter(TAdata, LU != "CD")

# Group occupied area by neighborhood and land use
NB.LAND.USE <- summarise(group_by(TAdata.no.condo.units, BRA_PD,SIMPLIFIED_LU),
                         extension = sum(as.numeric(LAND_SF), na.rm = TRUE))

# Remove parcels with no neighborhood iformation
NB.LAND.USE <- filter(NB.LAND.USE, !is.na(BRA_PD))

# "Spread" the land use variables as one long row per neighborhood
NB.LAND.USE <- spread(NB.LAND.USE, SIMPLIFIED_LU, extension)

# Replace NAs with 0s
NB.LAND.USE[is.na(NB.LAND.USE)] <- 0

# Add a total area column
NB.LAND.USE <- transform(NB.LAND.USE, TOTAL = rowSums(NB.LAND.USE[2:9]))

# Tranform the lan use coverage areas into percentages
```

```
NB.LAND.USE[2:9] <- sapply(NB.LAND.USE[2:9], function(x) { round(x / NB.LAND.USE[10] * 100, 2)})

NB.LAND.USE
```

```
##                 BRA_PD AGRICULTURAL COMMERCIAL CONDO INDUSTRIAL MIX_RC
## 1      Allston/Brighton         0.00       6.79  8.37       3.97   1.86
## 2  Back Bay/Beacon Hill         0.00      26.01 20.68       0.00   4.31
## 3               Central         0.00      25.18 14.74       0.24   9.11
## 4           Charlestown         0.00      11.70  6.91       1.66   3.81
## 5           East Boston         0.00       0.38  0.03       0.03   0.01
## 6         Fenway/Kenmore        0.00       6.08  1.77       0.12   1.95
## 7             Hyde Park         0.00       7.26  2.73       1.85   0.26
## 8         Jamaica Plain         1.18       2.77 15.13       0.36   0.91
## 9              Mattapan         0.00       1.03  0.92       0.03   0.11
## 10     North Dorchester         0.00       6.72  1.64       2.55  33.13
## 11           Roslindale         0.00       3.74  6.24       0.23   0.75
## 12              Roxbury         0.00       9.39  5.20       2.80   2.01
## 13         South Boston         0.00       8.80  1.49       5.42   0.58
## 14     South Dorchester         0.00      10.40  5.89       2.61   1.06
## 15            South End         0.00       8.73 12.90       1.08   9.46
## 16         West Roxbury         0.00       7.07 13.52       0.59   1.53
##    RESIDENTIAL TAX_EXEMPT TAX_EXEMPT_BRA      TOTAL
## 1        10.75      67.95           0.32  319319123
## 2        15.45      33.28           0.27   19908543
## 3         2.16      47.25           1.32   63027035
## 4         5.12      70.72           0.07   76310228
## 5         0.27      99.27           0.02 6702505015
## 6         3.54      85.51           1.04   71393569
## 7        31.21      56.56           0.13  163641099
## 8        21.56      54.70           3.39  121879594
## 9        12.61      83.49           1.82  283717621
## 10        7.53      48.38           0.07  167940704
## 11       35.90      52.51           0.63  105783830
## 12       49.41      27.34           3.84   85787053
## 13        2.62      79.78           1.31  341360518
## 14       51.28      28.38           0.39   98054422
## 15       18.79      45.37           3.68   33976488
## 16       37.79      39.51           0.00  160763950
```