Calculating Composite Demographic Indexes

July 27, 2022

Contents

Introduction

CBEP, like other National Estuary Programs will receive additional funding to support our programs via the "Bipartisan Infrastructure Law" signed into law last December.

EPA has recently released guidance for applying for those funds. A core component of the guidance is that overall, the NEP program should comply with the White House's "Justice 40" initiative, which requires that "at least 40% of the benefits and investments from BIL funding flow to disadvantaged communities."

EPA suggested that we use the National-scale EJSCREEN tools to help identify "disadvantaged communities" in our region. The EPA guidance goes on to suggest we focus on five demographic indicators:

- Percent low-income;
- Percent linguistically isolated;
- Percent less than high school education;

- Percent unemployed; and
- Low life expectancy.

This notebook examines the distributions of EPA's suggested demographic indicators in the Casco Bay Region.

Load Libraries

```
library(tidyverse)
#> -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
#> v ggplot2 3.3.6     v purrr   0.3.4
#> v tibble  3.1.7     v dplyr   1.0.9
#> v tidyr   1.2.0     v stringr 1.4.0
#> v readr   2.1.2     v forcats 0.5.1
#> -- Conflicts ------------------------------------------ tidyverse_conflicts() --
#> x dplyr::filter() masks stats::filter()
#> x dplyr::lag()    masks stats::lag()
library(GGally)
#> Registered S3 method overwritten by 'GGally':
#>   method from
#>   +.gg   ggplot2
library(readr)
```

Set Graphics Theme

This sets `ggplot()`graphics for no background, no grid lines, etc. in a clean format suitable for (some) publications.

```
theme_set(theme_classic())
```

Load Data

```
data_folder <- "Original_Data"
gis_folder <- "GIS_Data"
dir.create(file.path(getwd(), 'figures'), showWarnings = FALSE)
```

```
cb_indexes <- read_csv("cb_tracts_indexes.csv",
                       col_types = paste0('cdd--', rep('d', 23)))
```

Metrics

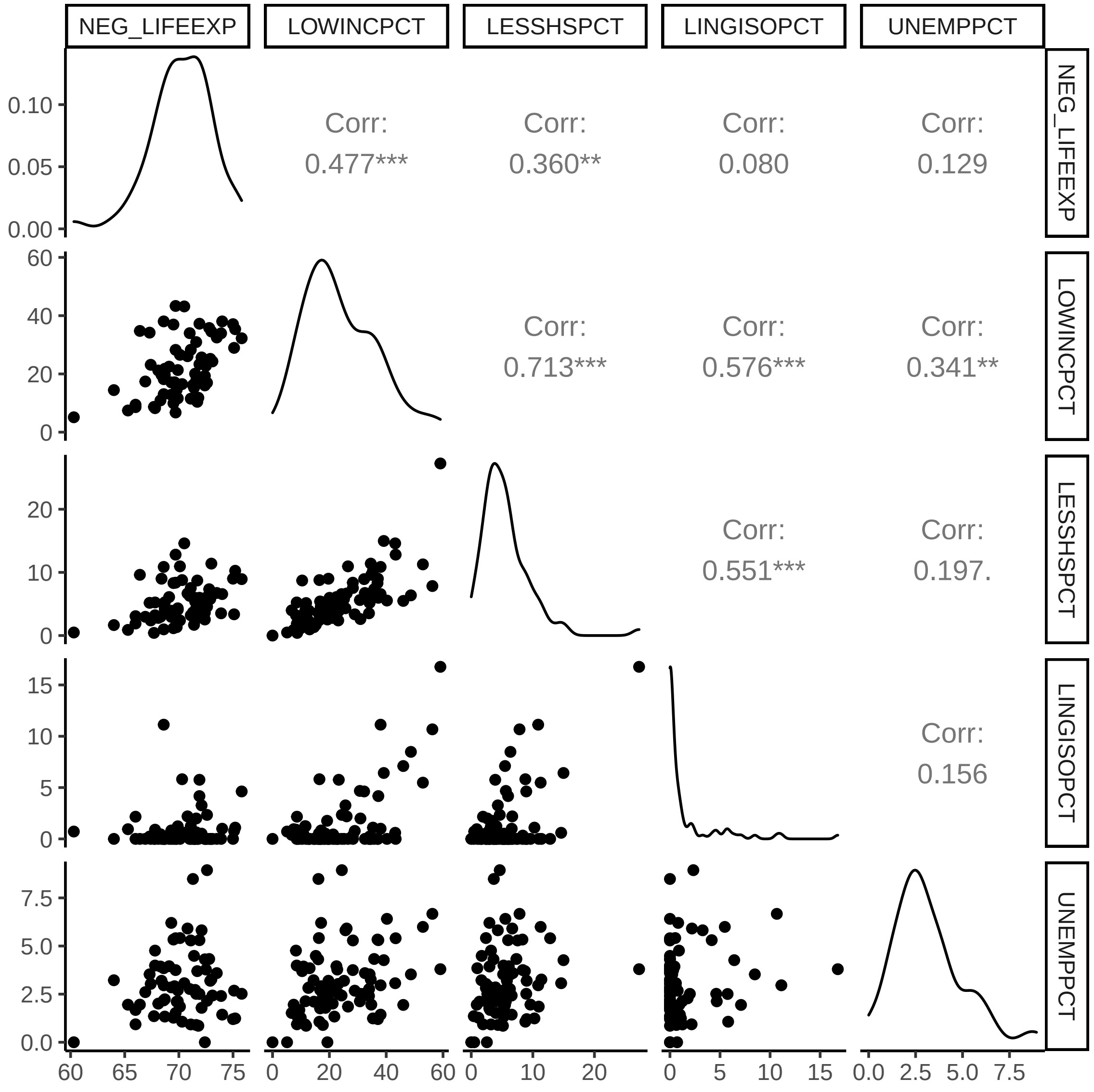
Summaries

```
summary(cb_indexes[,4:10])
#>     LIFEEXP        LIFEEXP_SE      NEG_LIFEEXP      LOWINCPCT
#>  Min.   :74.20   Min.   :1.104   Min.   :60.30   Min.   : 0.00
#>  1st Qu.:77.90   1st Qu.:1.611   1st Qu.:68.60   1st Qu.:15.24
#>  Median :79.80   Median :1.871   Median :70.20   Median :21.17
#>  Mean   :79.73   Mean   :2.007   Mean   :70.27   Mean   :23.81
#>  3rd Qu.:81.40   3rd Qu.:2.173   3rd Qu.:72.10   3rd Qu.:33.94
#>  Max.   :89.70   Max.   :3.990   Max.   :75.80   Max.   :59.07
#>  NA's   :13      NA's   :13      NA's   :13
#>    LESSHSPCT        LINGISOPCT        UNEMPPCT
#>  Min.   : 0.000   Min.   : 0.000   Min.   :0.000
#>  1st Qu.: 2.968   1st Qu.: 0.000   1st Qu.:1.951
#>  Median : 4.976   Median : 0.000   Median :2.720
#>  Mean   : 5.598   Mean   : 1.468   Mean   :3.092
#>  3rd Qu.: 6.742   3rd Qu.: 1.100   3rd Qu.:3.935
#>  Max.   :27.251   Max.   :16.765   Max.   :8.939
#>
```

Note that more than 50% of our census tracts report no linguistic isolation. In our region, this is a surrogate largely for immigrants in Portland.

Pairs Plot

```
cb_indexes %>%
  select( "NEG_LIFEEXP", "LOWINCPCT", "LESSHSPCT", "LINGISOPCT","UNEMPPCT" ) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
```



Data (except life expectancy) is not normally distributed, especially for those metrics that have mostly low values. That is not unexpected for percents. which are bounded below by zero, and are a transformation of count data.

Adding or averaging raw values will lead to indexes dominated by the sub-indexes with the largest variance.

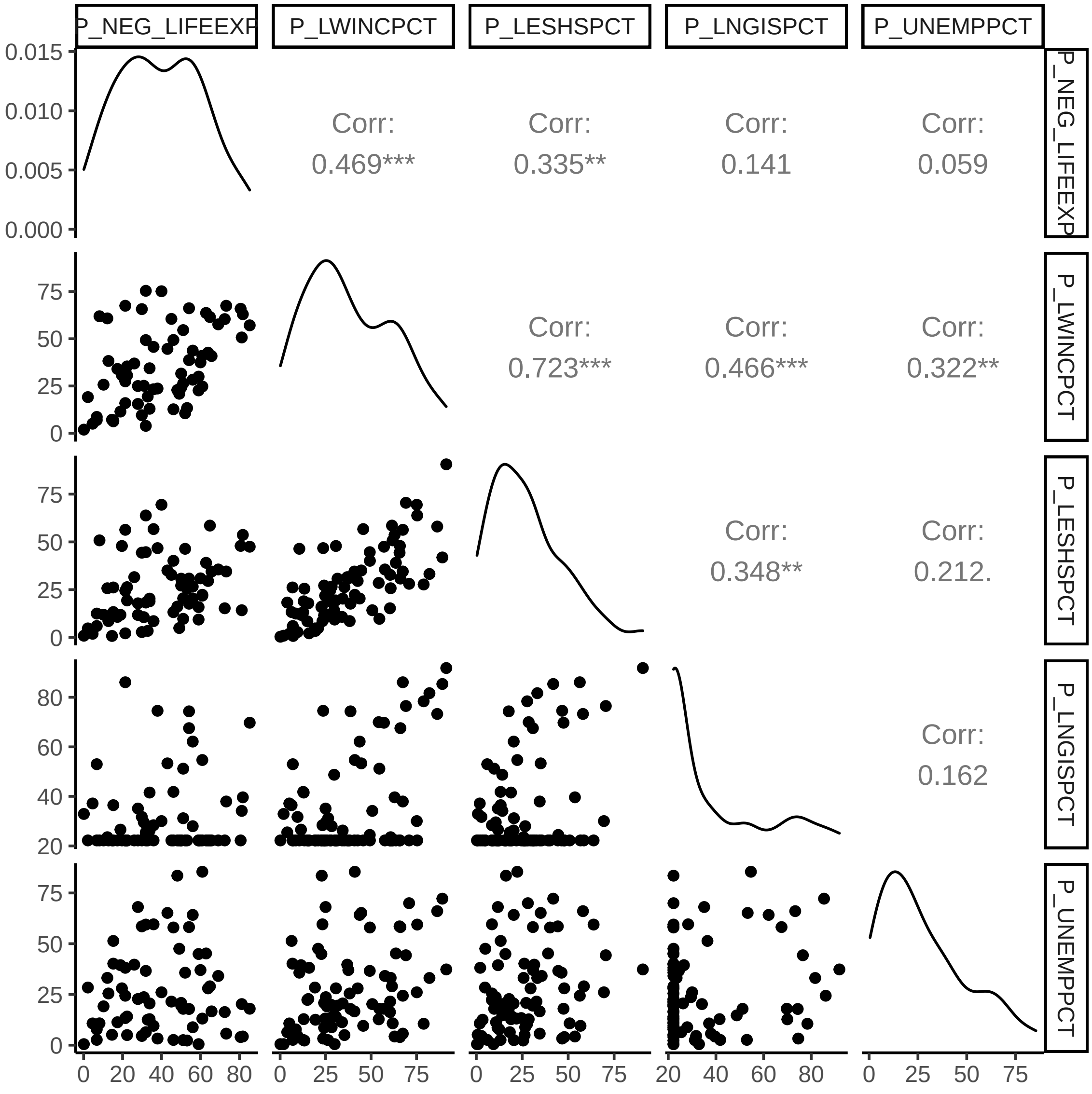
National Percentiles

Summaries

```
summary(cb_indexes[,11:15])
#>  P_NEG_LIFEEXP       P_LWINCPCT        P_LESHSPCT        P_LNGISPCT
#>  Min.   : 0.1236   Min.   : 0.1501   Min.   : 0.3721   Min.   :22.21
#>  1st Qu.:21.4028   1st Qu.:20.9051   1st Qu.:11.8659   1st Qu.:22.21
#>  Median :36.9231   Median :34.0198   Median :24.4703   Median :22.21
#>  Mean   :39.6818   Mean   :38.2241   Mean   :26.8084   Mean   :35.17
#>  3rd Qu.:56.0140   3rd Qu.:60.3845   3rd Qu.:35.5721   3rd Qu.:39.62
#>  Max.   :85.2650   Max.   :91.3387   Max.   :90.6131   Max.   :91.81
#>  NA's   :13
#>    P_UNEMPPCT
#>  Min.   : 0.4998
#>  1st Qu.:10.7103
#>  Median :20.8290
#>  Mean   :27.3289
#>  3rd Qu.:39.4257
#>  Max.   :85.4525
#>
```

Pairs Plot

```
cb_indexes %>%
  select( c(P_NEG_LIFEEXP:P_UNEMPPCT)) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
```



Indexes

Summaries

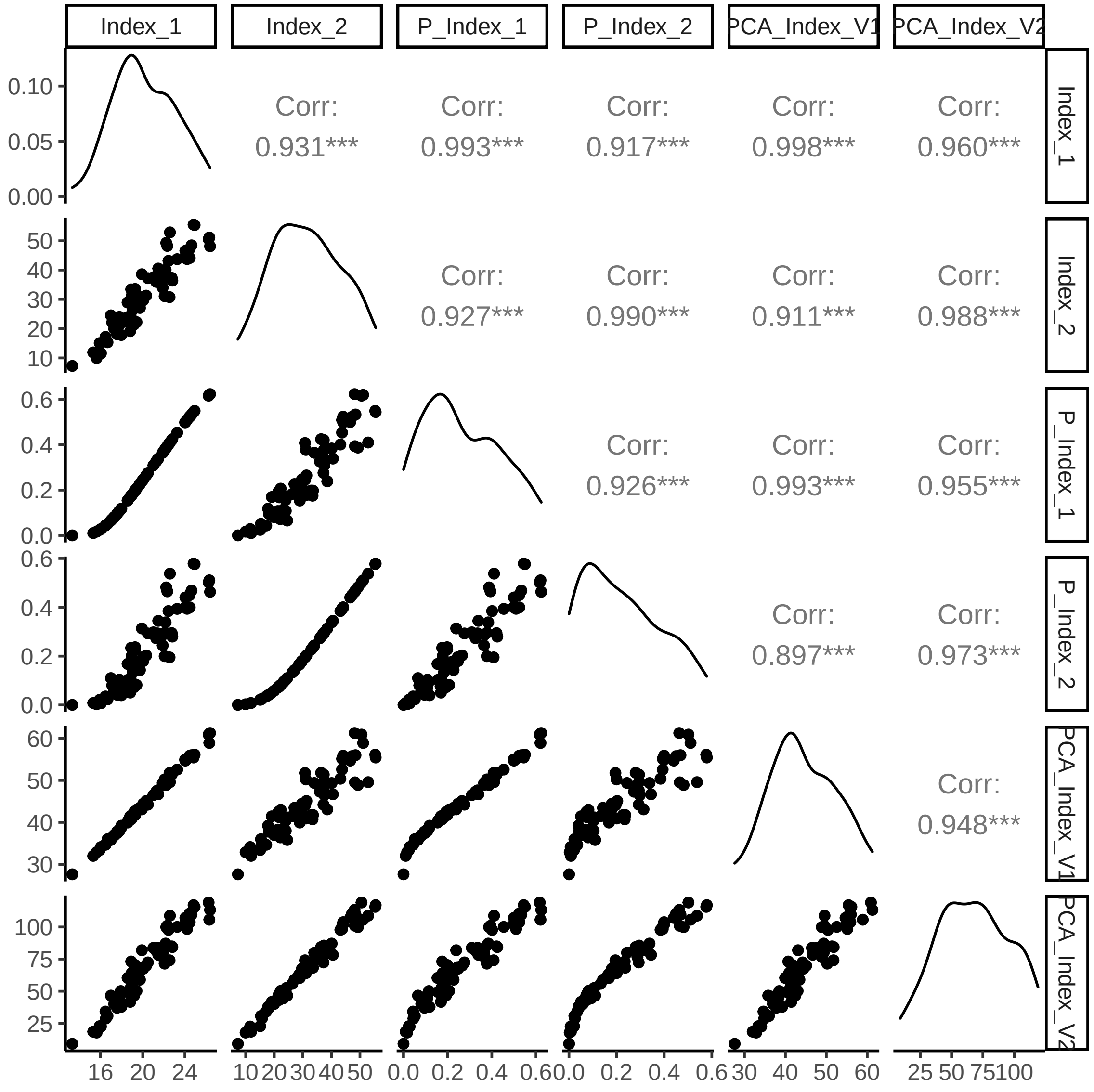
As a reminder:

- Index 1 is a mean of the raw scores
- index 2 is a mean of national percentiles.(This may be what EPA wants)
- P_Index_1 is national percentiles of Index 1
- p_Index_2 is national percentile of index 2
- PCA_Index_V1 is the first PCA axis of a (scaed) PCA on raw scores
- PCA_Index_V2 is the first PCA Axis of a (Scaeld)PCA on national percentiles.

```
summary(cb_indexes[,21:26])
#>     Index_1         Index_2         P_Index_1          P_Index_2
#>  Min.   :13.32   Min.   : 7.267   Min.   :0.000089   Min.   :0.000447
#>  1st Qu.:17.96   1st Qu.:22.031   1st Qu.:0.117287   1st Qu.:0.079465
#>  Median :19.70   Median :30.944   Median :0.224335   Median :0.197891
#>  Mean   :20.25   Mean   :31.614   Mean   :0.266640   Mean   :0.224731
#>  3rd Qu.:22.36   3rd Qu.:40.198   3rd Qu.:0.395778   3rd Qu.:0.339868
#>  Max.   :26.38   Max.   :55.501   Max.   :0.623803   Max.   :0.578988
#>  NA's   :13      NA's   :13       NA's   :13         NA's   :13
#>   PCA_Index_V1    PCA_Index_V2
#>  Min.   :27.62   Min.   :  9.075
#>  1st Qu.:39.03   1st Qu.: 46.599
#>  Median :43.05   Median : 68.495
#>  Mean   :44.66   Mean   : 68.219
#>  3rd Qu.:49.83   3rd Qu.: 85.892
#>  Max.   :61.28   Max.   :119.030
#>  NA's   :13      NA's   :13
```

Pairs Plot

```
cb_indexes %>%
  select( Index_1:PCA_Index_V2) %>%
  ggpairs(progress = FALSE)
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values

#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Warning: Removed 13 rows containing non-finite values (stat_density).
#> Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
#> Removed 13 rows containing missing values
#> Warning: Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Removed 13 rows containing missing values (geom_point).
#> Warning: Removed 13 rows containing non-finite values (stat_density).
```


Index_1
Index_2
P_Index_1
P_Index_2
PCA_Index_V1
PCA_Index_V2
Corr:
0.931***
Corr:
0.993***
Corr:
0.917***
Corr:
0.998***
Corr:
0.960***
Corr:
0.927***
Corr:
0.990***
Corr:
0.911***
Corr:
0.988***
Corr:
0.926***
Corr:
0.993***
Corr:
0.955***
Corr:
0.897***
Corr:
0.973***
Corr:
0.948***
0.10
0.05
0.00
50
40
30
20
10
0.6
0.4
0.2
0.0
0.6
0.4
0.2
0.0
60
50
40
30
100
75
50
25
16
20
24
10
20
30
40
50
0.0
0.2
0.4
0.6
0.0
0.2
0.4
0.6
30
40
50
60
25
50
75
100