

Manual cutoffs

Ricky Heinrich

2023-06-06

Introduction

The Proximity Measures Database contains continuous measures for 10 amenities for a number of DB within a specific threshold. The distribution of these proximity measures is heavily right skewed, and there are for the most part no discernible clusters. The density distribution, with a default bandwidth, of each amenity is shown in Figure 1.

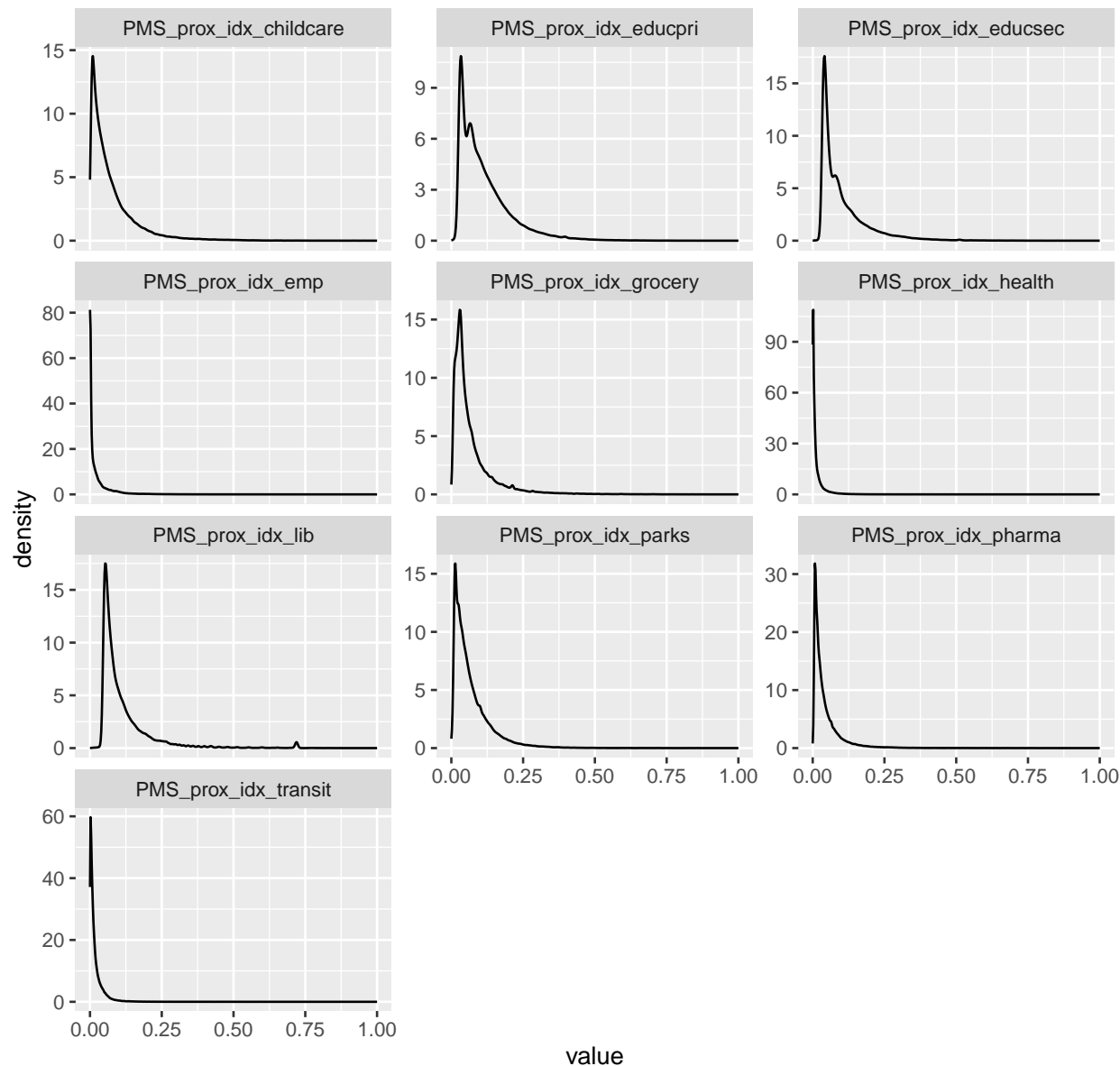


Figure 1: Distribution of proximity measures by amenity

When transforming the data, the inherent relationship between data points remain the same, but the new structure may reveal new insights. The most ‘famous’ transformation available is the log transform. It “can be used to make highly skewed distributions less skewed”. It may help “make patterns more visible”. A consideration to be aware of is that the log of 0 is -Inf. To account for proximity values of 0 in our dataset, we

shift the distribution by $+0.0001$. This avoids the problem of $-\text{Inf}$ whilst maintaining the original distances between all values. The downsides of using a log transformation are [DOWNSIDES]. Figure 2 demonstrates the distribution of the log transformed proximity measures, where all the amenities' distributions were shifted by $+0.0001$. We can already visually identify more possible clusters.

We shifted the distribution by $+0.0001$ of the amenities that had a minimum value of 0. Grocery, educpri, educsec, and lib did not have values of 0 in their distribution and such were not shifted. The visual difference of the distributions between when $+0.0001$ is applied vs when it is not are imperceptible. For simplification in reproducibility, we will apply the distribution shift to all amenities.

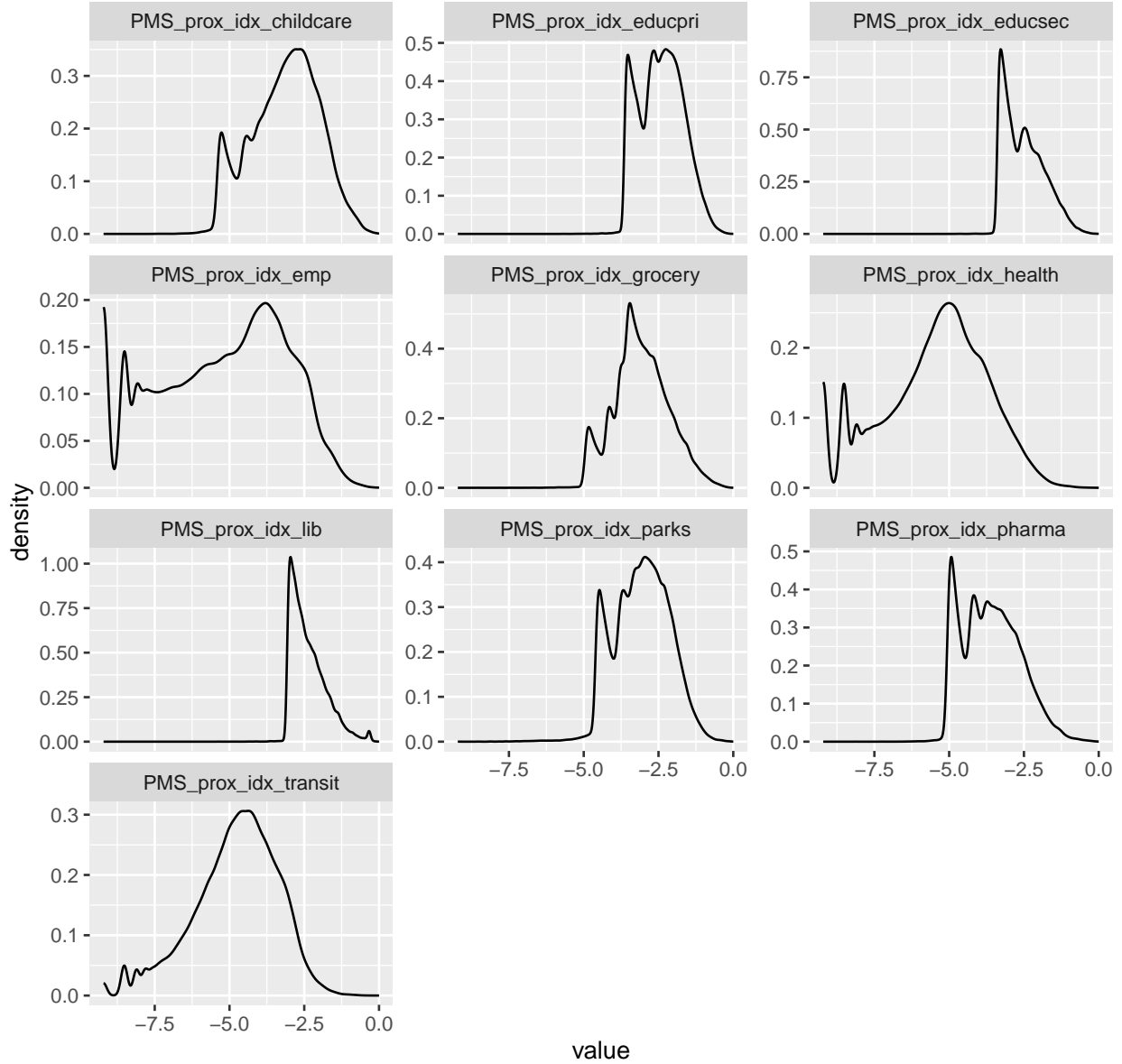
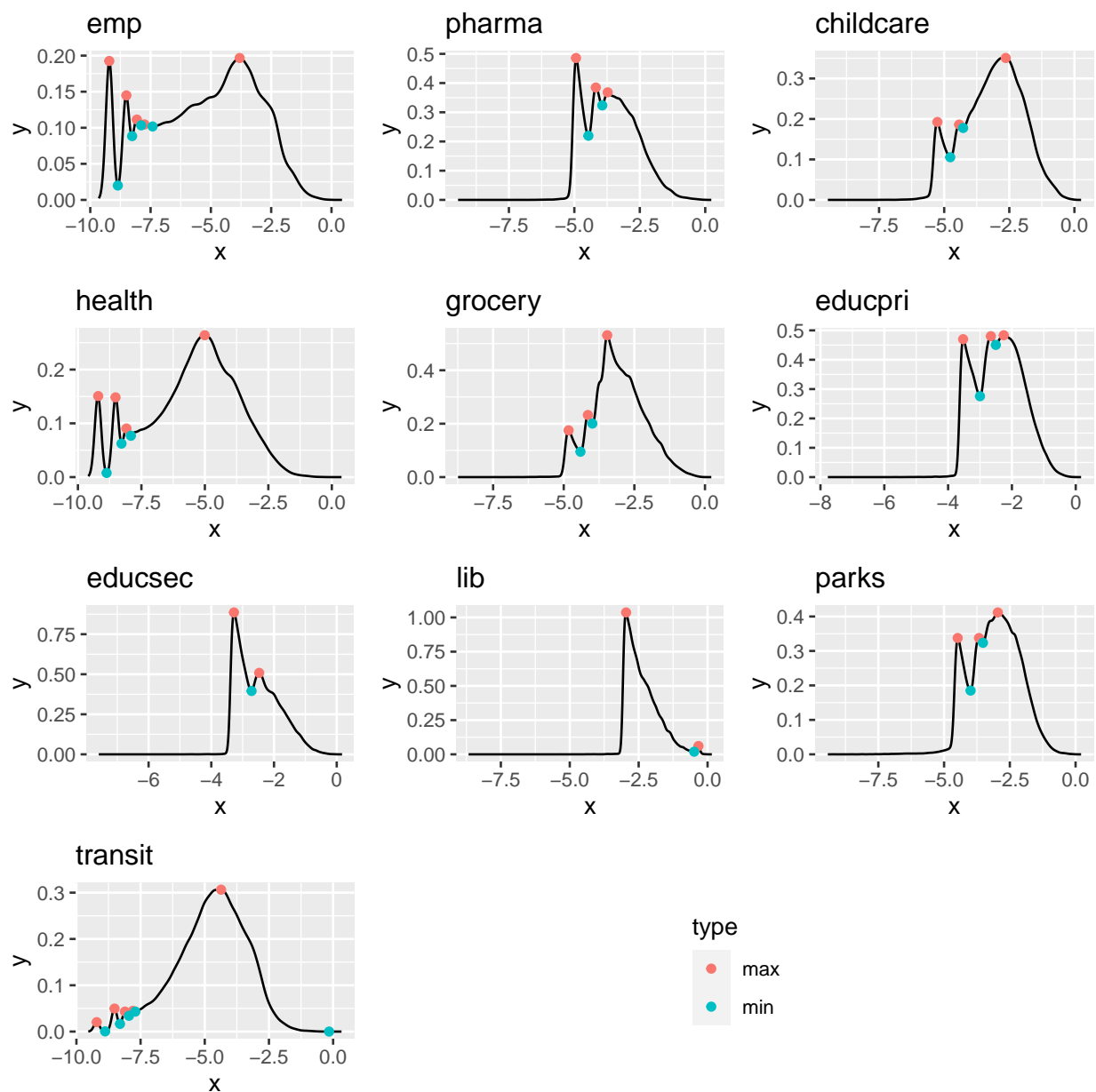


Figure 2: LOG TRANSFORMED(0.0001): Distribution of proximity measures by amenity

Segmenting via minima

A segmentation technique is to segment the distribution at select minima of the density distribution. Each minimum in the density curves represents a density sparse region, which may be a ‘natural’ break in the continuous measures. Figure 4 provides an overview of where maxima and minima are located in the density curves of every amenity. We see that there are a lot of points that are by definition local minima, but are not fully indicative of density sparse regions. We can limit which minima are representative of density sparse regions by only including those who have a threshold difference between themselves and surrounding maxima. The results of the indepth analysis of which minima should intuitively represent a cutoff value for each amenity are displayed in the following plot, and summarized in the table.

Summary



	num_groups
emp	5
pharma	3
childcare	3
health	4
grocery	3
educpri	3
educsec	2
lib	2
parks	3
transit	5

Cut off values:

```
## [1] "PMS_prox_idx_emp"
## [1] 0.0001423603 0.0002573125 0.0003743458 0.0006010775
## [1] "PMS_prox_idx_pharma"
## [1] 0.01152592 0.01958168
## [1] "PMS_prox_idx_childcare"
## [1] 0.008507409 0.013950871
## [1] "PMS_prox_idx_health"
## [1] 0.0001405736 0.0002524638 0.0003658077
## [1] "PMS_prox_idx_grocery"
## [1] 0.01218361 0.01856220
## [1] "PMS_prox_idx_educpri"
## [1] 0.04975835 0.08179011
## [1] "PMS_prox_idx_educsec"
## [1] 0.06619424
## [1] "PMS_prox_idx_lib"
## [1] 0.6150259
## [1] "PMS_prox_idx_parks"
## [1] 0.01844893 0.02954933
## [1] "PMS_prox_idx_transit"
## [1] 0.0001390986 0.0002481958 0.0003512994 0.0004514904 0.8553858919
```

Essentially, for every amenity, add a new column (so 10 new cols today) that outlines which group it belongs to, according to the cutoff points.

How many DBs in each group

```
## [1] "emp"
##
##           1         2         3         4         5    <NA>
## [1,] 29831 22179 14893 26887 329812 66074
## [1] "pharma"
##
##           1         2         3    <NA>
## [1,] 41305 30505 106711 311155
## [1] "childcare"
##
##           1         2         3    <NA>
## [1,] 26274 18663 199027 245712
```

```
## [1] "health"
##
##           1           2           3           4    <NA>
## [1,] 14556 14259    8086 263564 189211
## [1] "grocery"
##
##           1           2           3    <NA>
## [1,] 11600 10766 118697 348613
## [1] "educpri"
##
##           1           2           3    <NA>
## [1,] 57009 45865 122485 264317
## [1] "educsec"
##
##           1           2    <NA>
## [1,] 63449 77764 348463
## [1] "lib"
##
##           1           2    <NA>
## [1,] 111546 1109 377021
## [1] "parks"
##
##           1           2           3    <NA>
## [1,] 42995 31335 159738 255608
## [1] "transit"
##
##           1           2           3           4           5           6    <NA>
## [1,] 1014 2474 2082 1923 173810 2 308371
```

summary statistics of IoR

Next Steps

- assign 'group' to each
- number of DB for each clusters
- find way to format in table nicely -> makes sense Jonah's way per amenity
- summary statistics of IoR for each cluster
- population summary statistics
- CMA type summary statistics