Manual cutoffs: adjust bw

Ricky Heinrich

2023-06-02

Introduction

Add `adjust = 1/2` to `density` functions to see effect. Effectively decreases the bandwidth of the kernel used by half. Tldr result: get many more 'suggested' cutoff points for each amenityprint("Cluster cutoff values"), as the distribution is 'less smooth'. May observe even more discernible 'clusters' in non-log kernel distribution. Much more interesting and suggestive that different `bw` values lead to vastly different results.

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.

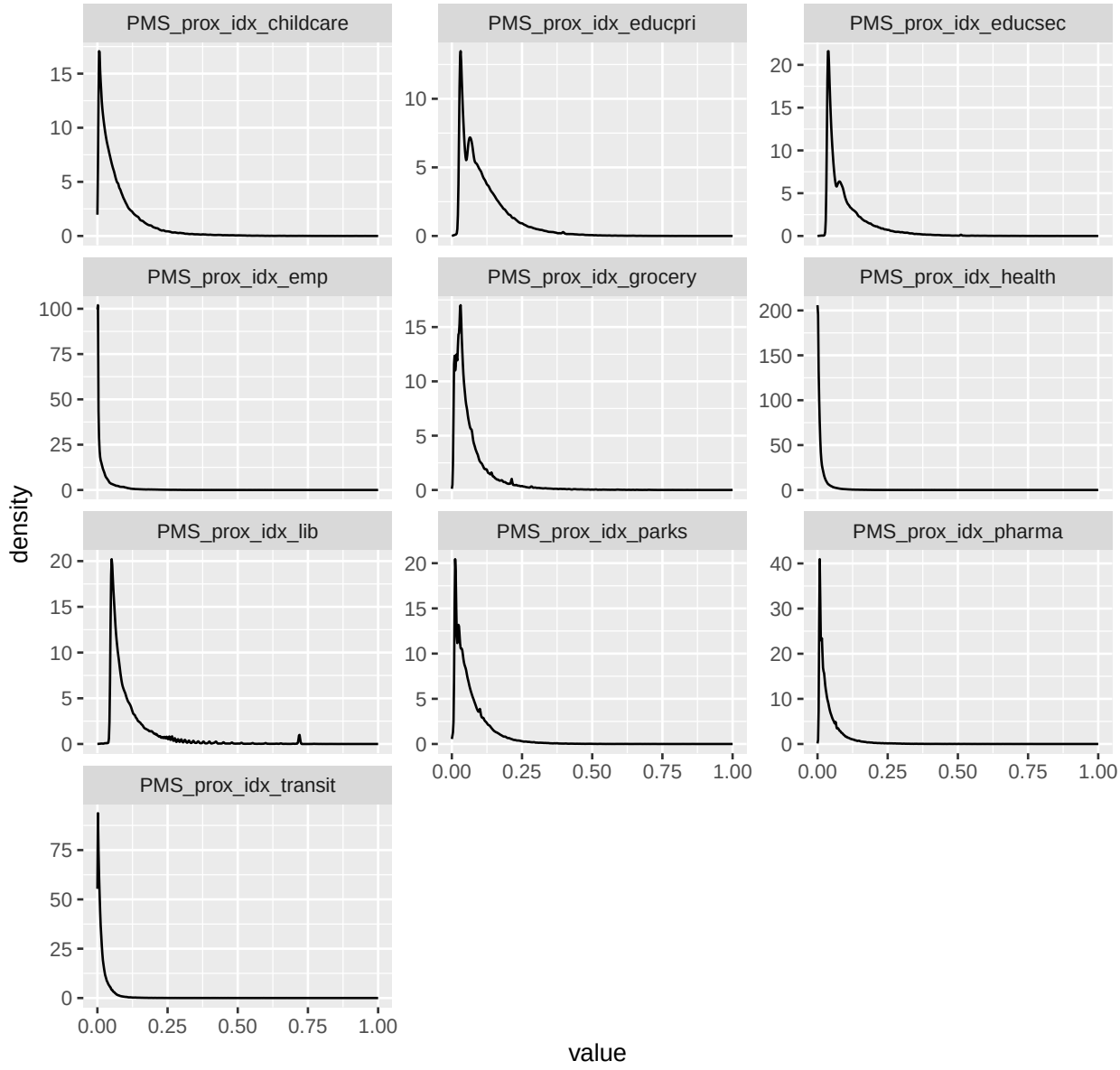

Figure 1: Distribution of proximity measures by amenity

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 -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.

In Figure 3, we only 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.

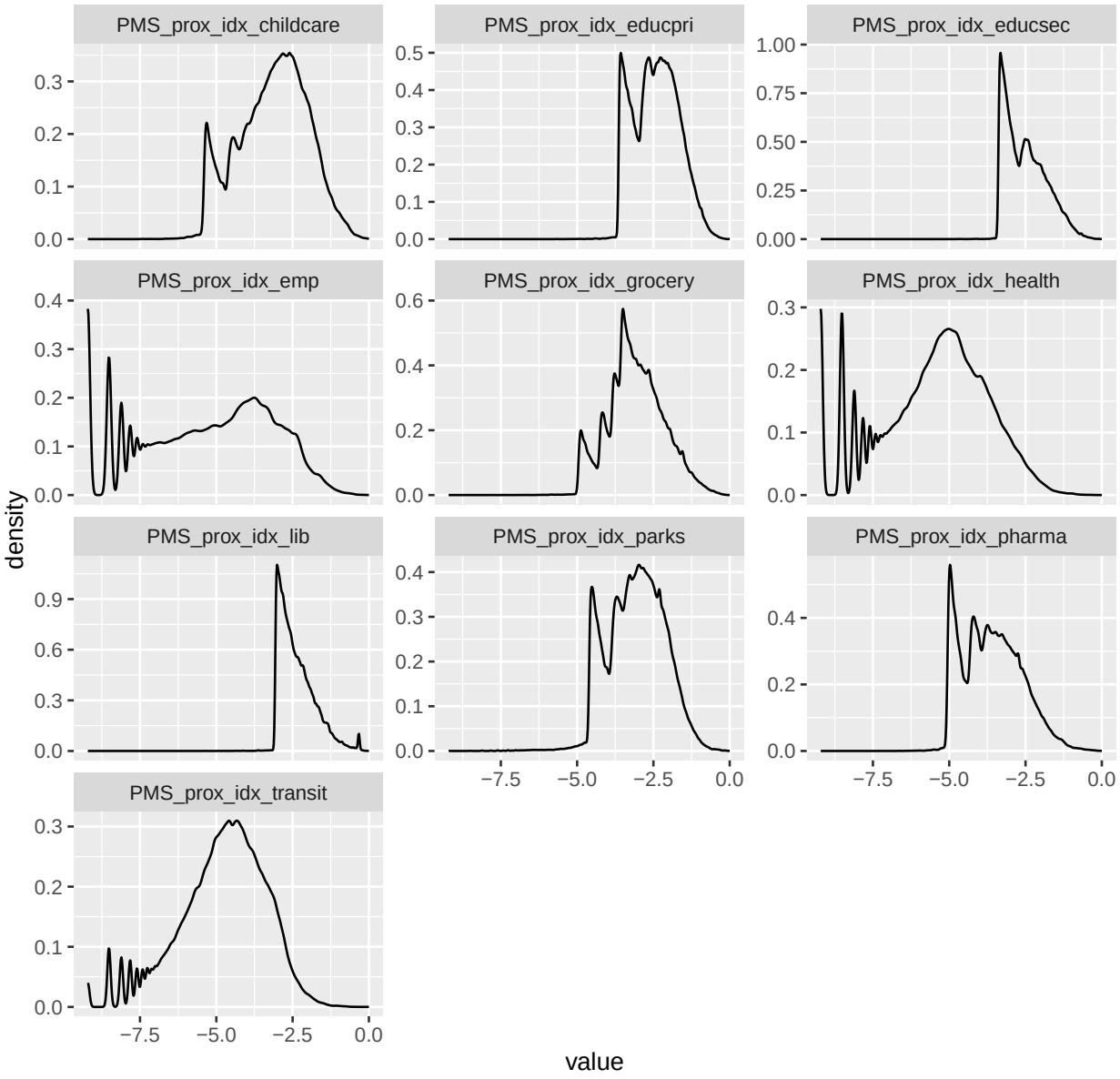

Figure 2: LOG TRANSFORMED(0.0001): Distribution of proximity measures by amenity

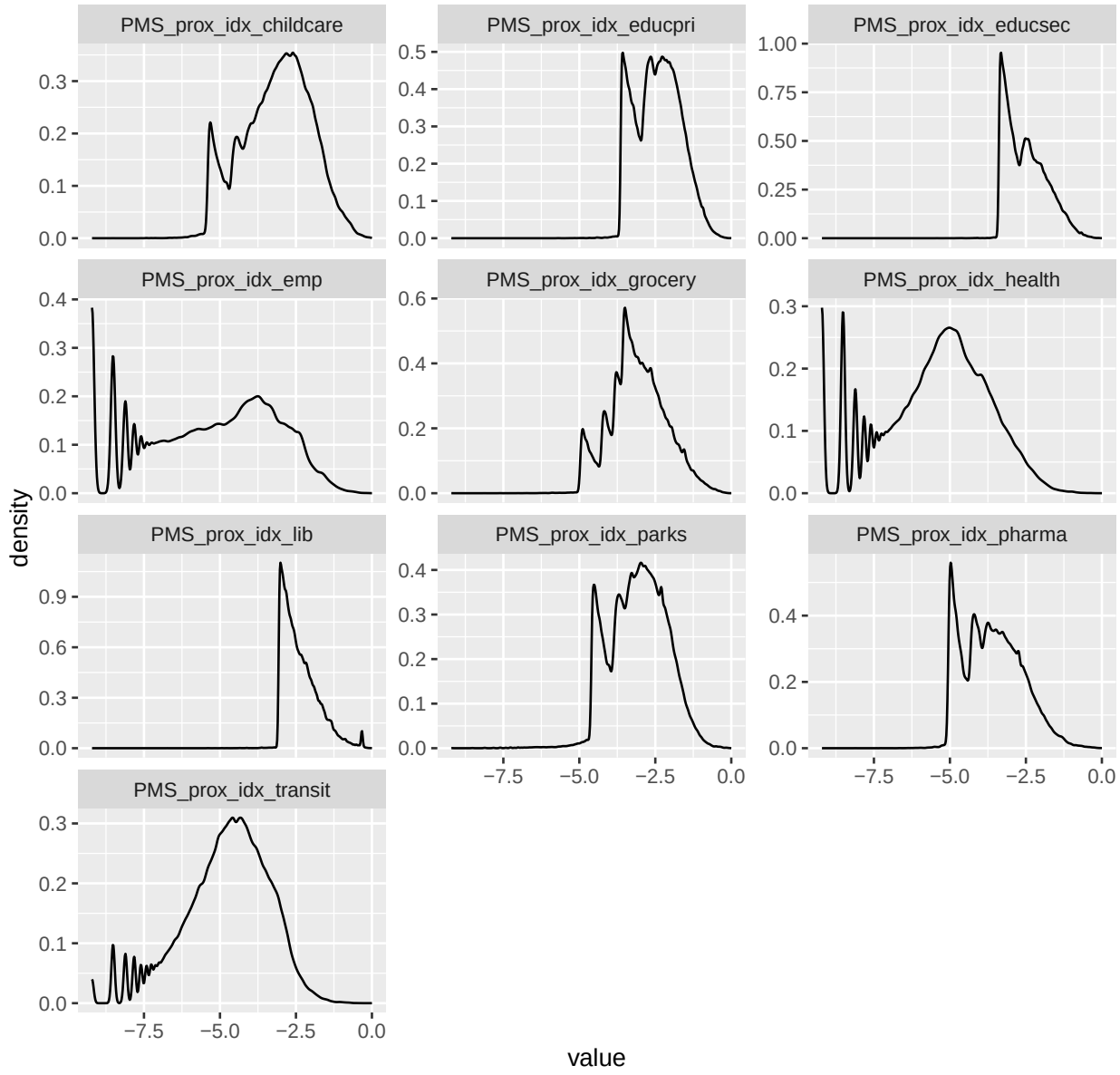

Figure 3: LOG TRANSFORMED(0.0001 in some): Distribution of proximity measures by amenity

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. We will conduct an indepth analysis of which minima should intuitively represent a cutoff value for each amenity.

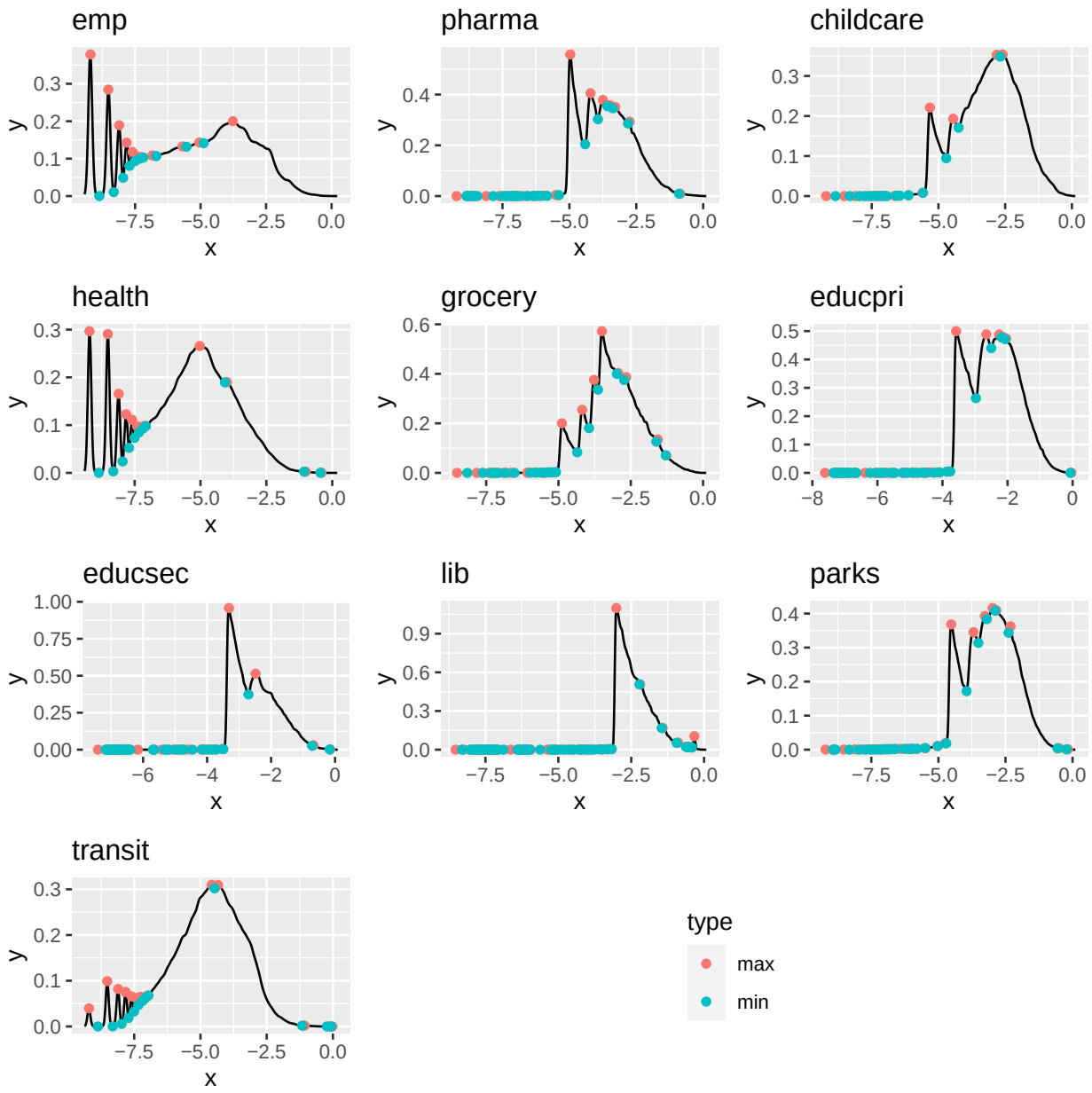

Figure 4: Location of minima and maxima

Employment

The Employment density distribution contains 10 minima. The following figure illustrates the density distribution with the minima ploted in blue and the maxima in red. Visually, we may not construe the third or the fourth minima as a cutoff value, as the peak in between is fairly small. As well, there are other areas in the curve that seem to plateau, and may be visually decent places for a cutoff value, but are not technically places where a minima is present.

As is, there would be 11 groups, corresponding to 10 cutoffs.

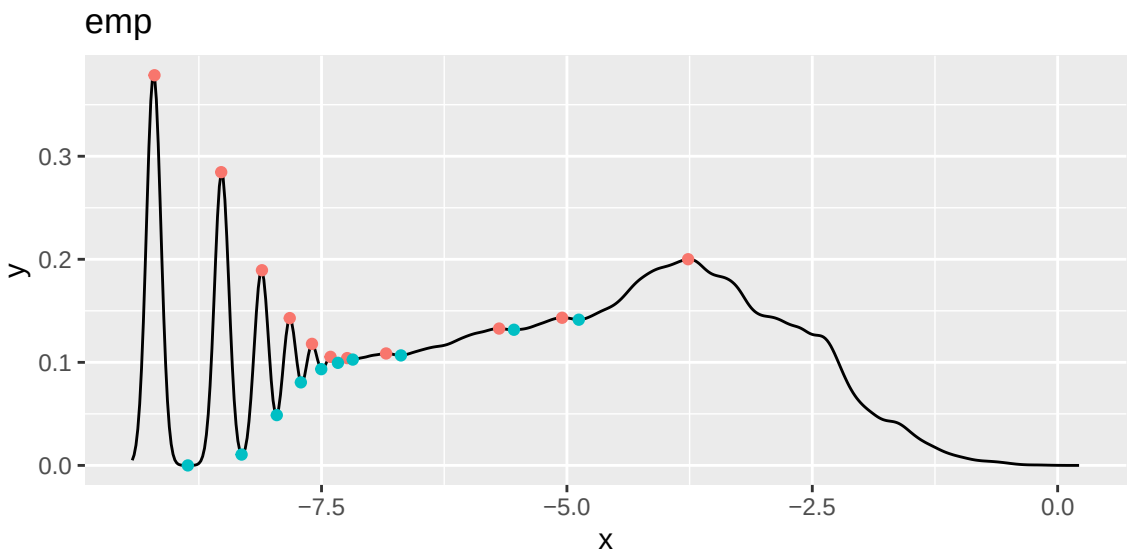

Figure 5: Employment density curve with minima and maxima

Pharmacies

In the case of Pharmacies, shown in Figure 6, there are many technical minima and maxima in an area that visually seems flat and have overall very low density. There is no doubt that these areas are not indicative specifically of density sparse regions, as the whole area is density sparse. The following plot, Figure 7, shows the difference between the density value of maxima-minima pairs (unidirectional). We see that for Pharmacies, the difference in the first 6 pairs is very small, as we can tell from the previous plot. The difference in density between the first maxima and the first minima, for example, is 5.6283242×10^{-5}, which is very small compared to the 8th (the 2nd visually discernible peak in Figure 6): 4.666643×10^{-5}. This may suggest that we should only use as cutoffs the minima that have a threshold difference with the neighbouring maxima. An appropriate threshold for Pharmacies may be a difference of 0.001.

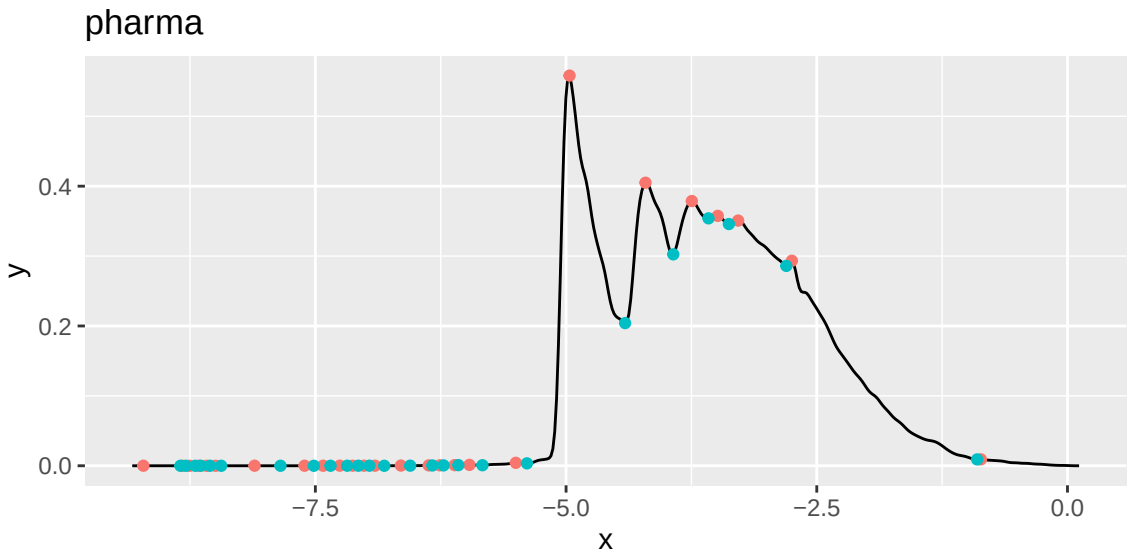

Figure 6: Pharmacies density curve with minima and maxima

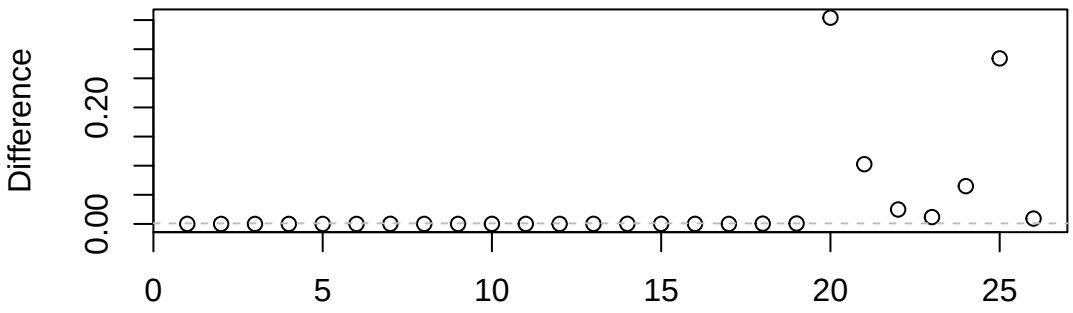

Figure 7: Difference between density value of a maxima-minima pairs, with suggested threshold = 0.001

Removing the pairs of where the difference is below the threshold values give us the following plot. We see that in this case, there would be 2 cutoff points giving 3 groups.

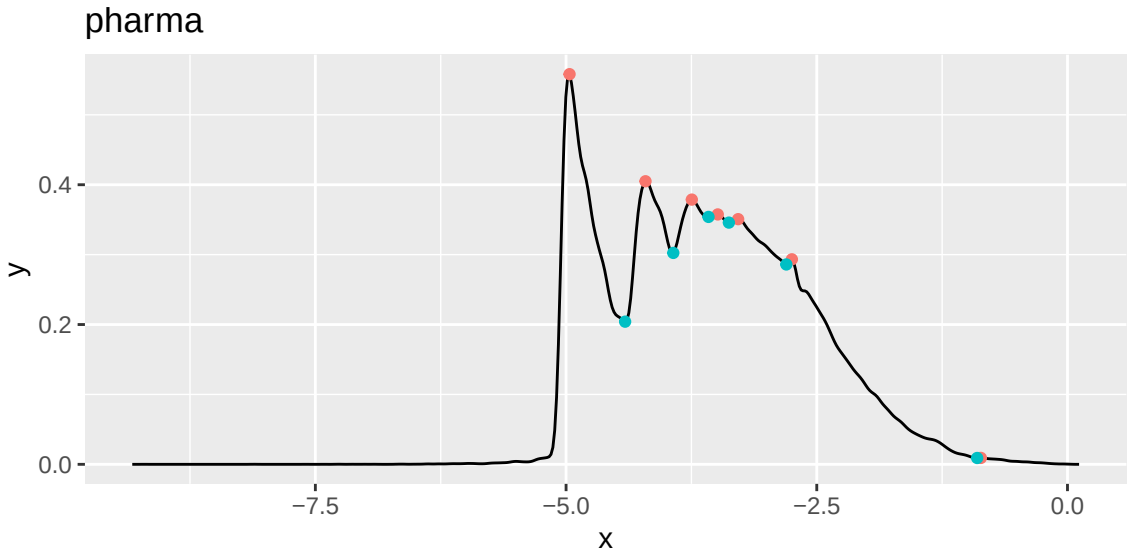

Figure 8: Density plot with suggested cutoff points in blue

Childcare

Similarly to Pharmacies, Childcare has indiscernible maxima and minima in some areas, as seen in Figure 9. There is even such an area at the top of the largest peak. Figure 10 shows the difference again, and we can see that the same threshold of 0.001 may be appropriate. Figure 11 shows the density with the values with a difference beneath this threshold removed, showing 2 cutoffs for 3 groups. The second 'maxima' at the top of the peak was retained due the methodology.

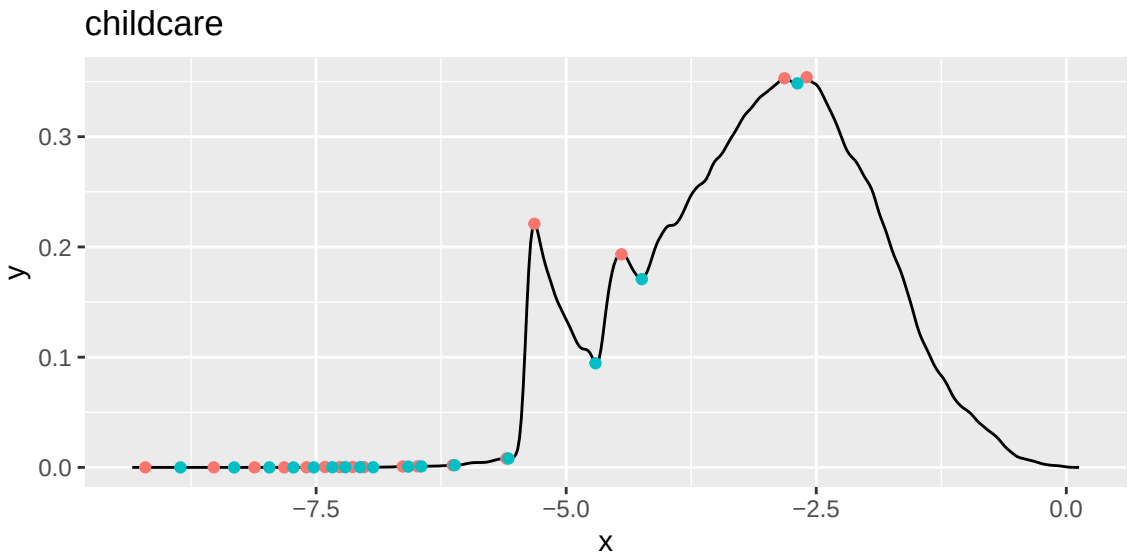

Figure 9: density curve with minima and maxima

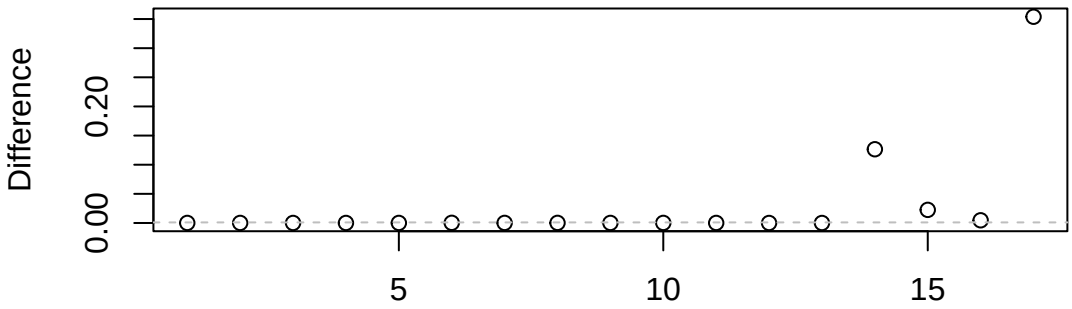

Figure 10: difference between density value of a maxima-minima pairs, threshold = 0.001

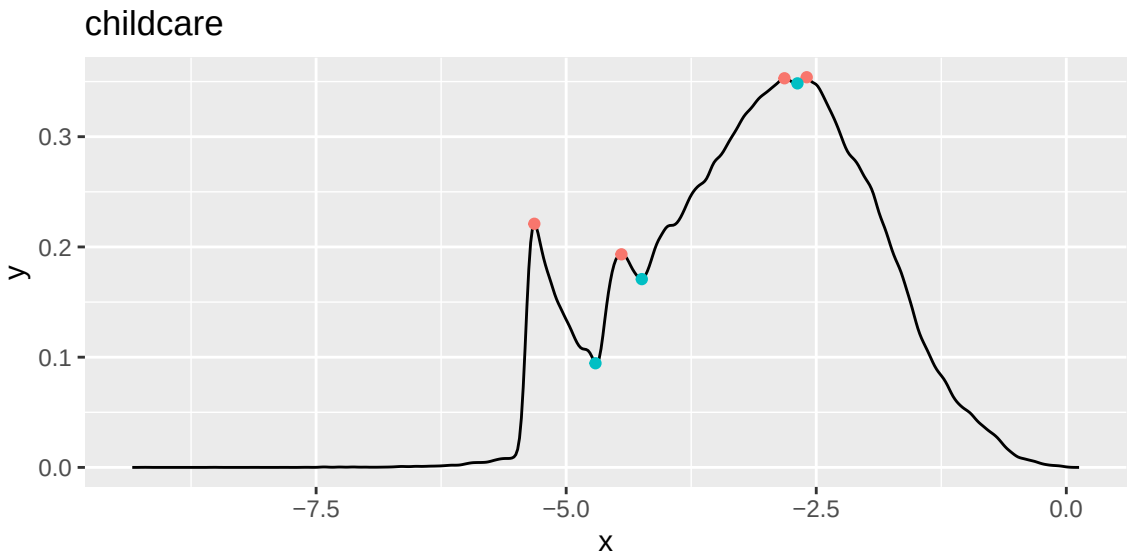

Figure 11: density plot with suggested cutoff points in blue

Healthcare

For Healthcare, a visual assessment of the density curve with the minima and maxima point doesn't reveal any point outside of expectation. The curve suggests 3 cutoff values, giving a total of 4 groups. For here on out, 'extra' plots are moved to the appendix.

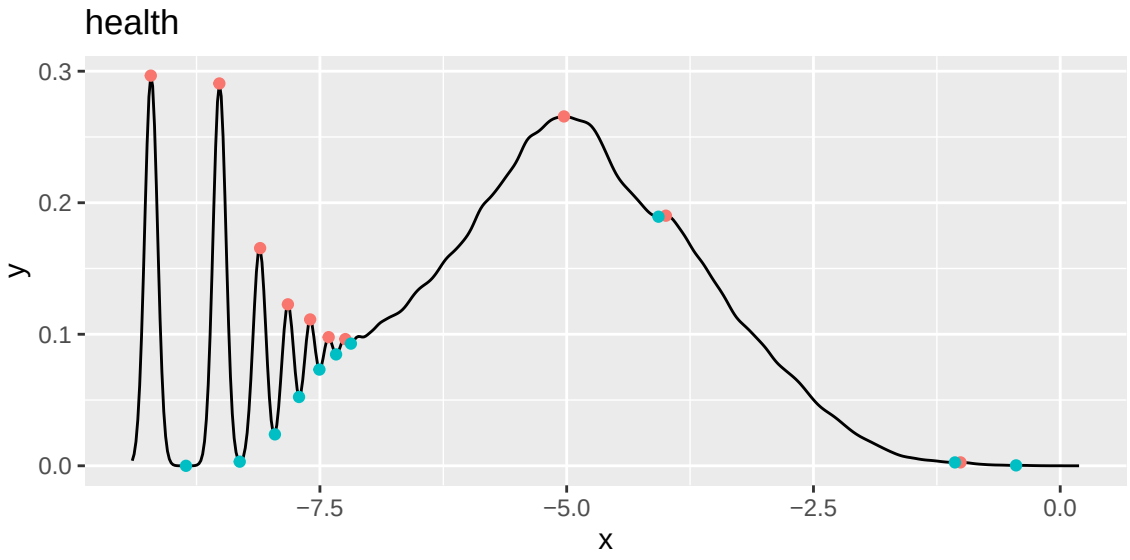

Figure 12: density plot with suggested cutoff points in blue

Grocery

Grocery has a similar start to Pharmacies and Childcare. Removing the points below the same threshold suggests 2 cutoff points, giving 3 groups. Visually, there are plateau areas, similarly to other amenities, that

may serve as decent additional cut off points.

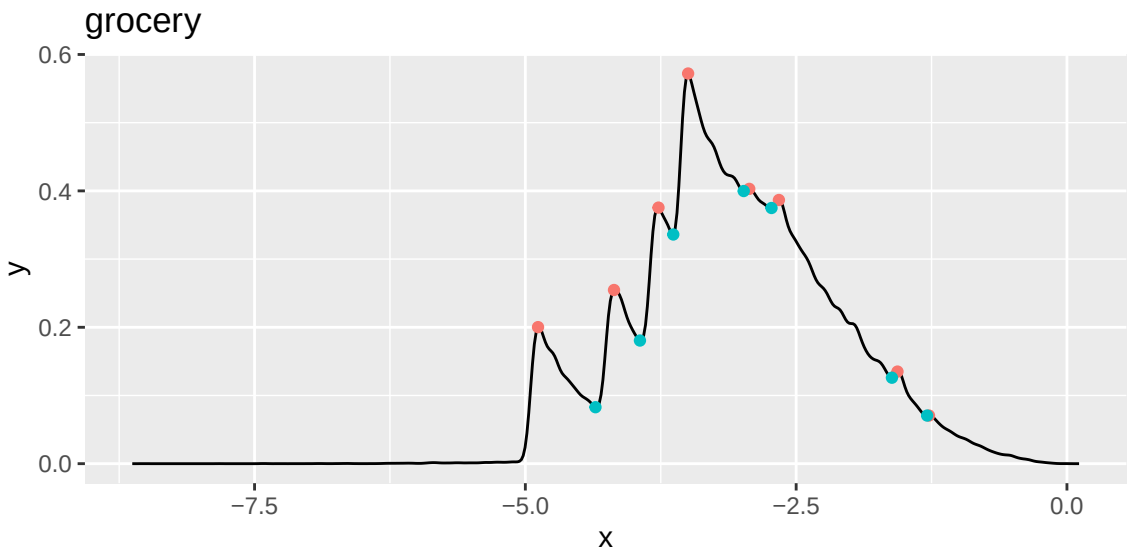

Figure 13: density plot with suggested cutoff points in blue

Primary Education

The final plot suggests 2 cutoff points, giving 3 groups.

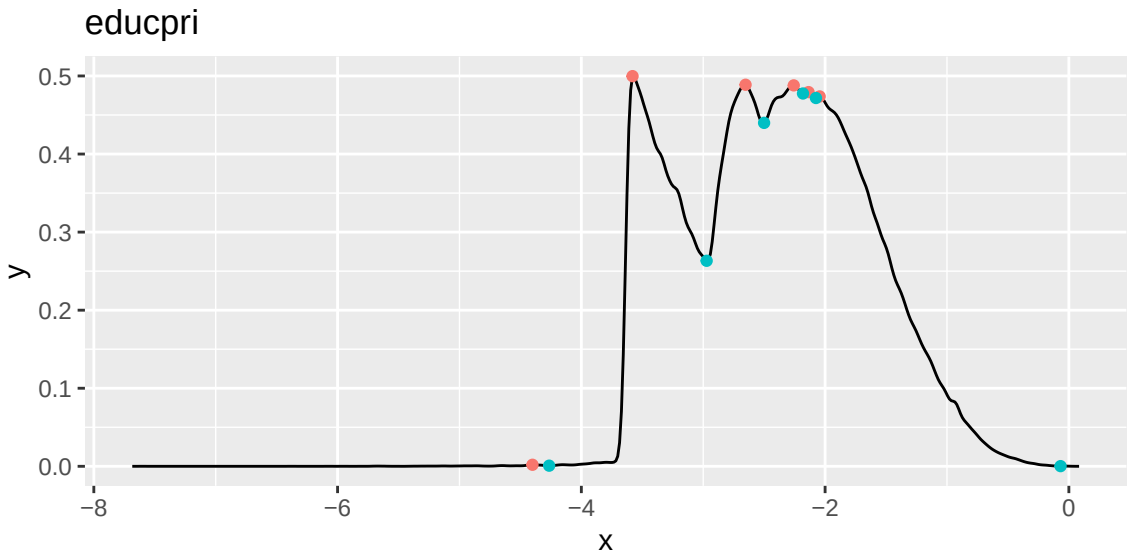

Figure 14: density plot with suggested cutoff points in blue

Secondary Education

For the first time, after removal of values below the threshold, only one cutoff point is suggested, giving two groups.

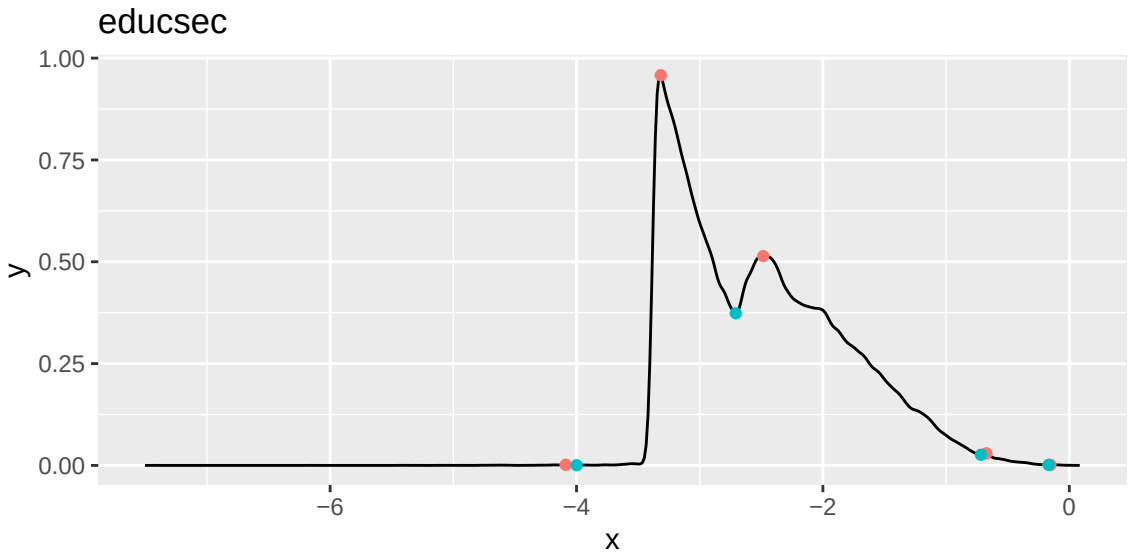

Figure 15: density plot with suggested cutoff points in blue

Libraries

The plot suggests 1 minima cutoff, giving 2 groups. Similarly as in other amenities, the case could be made that more clusters could be found in areas that plateau a bit.

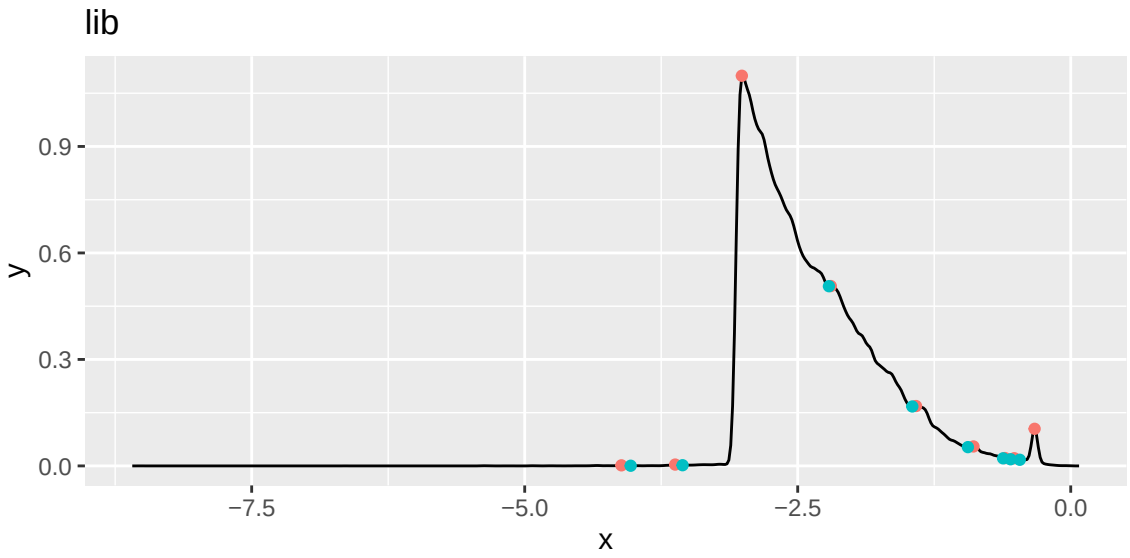

Figure 16: density plot with suggested cutoff points in blue

Parks

After removing values under the threshold, the curve suggests 2 cutoff locations, giving 3 groups.

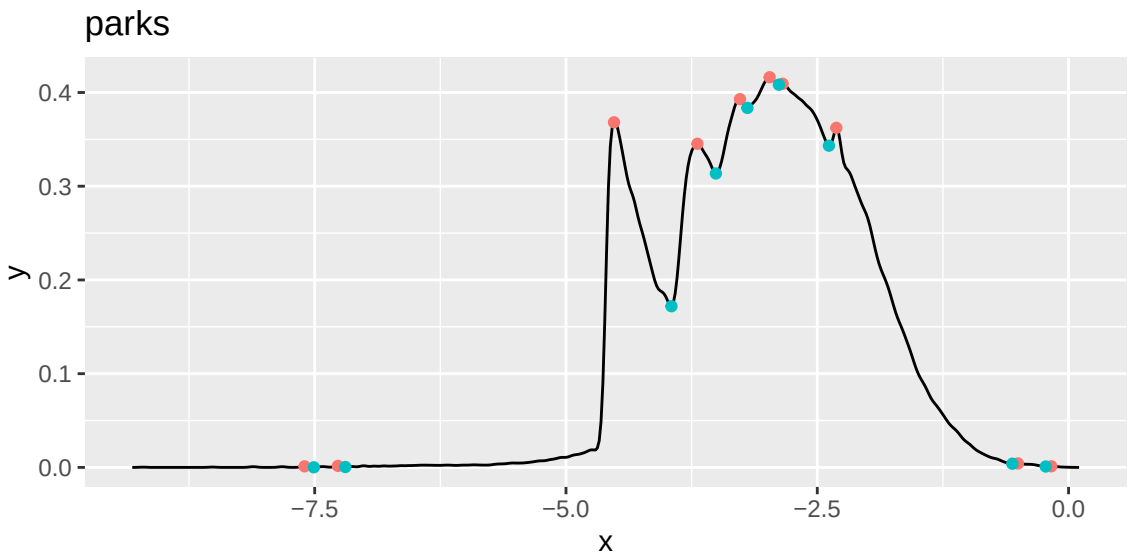

Figure 17: density plot with suggested cutoff points in blue

Transit

After removing all values below the same threshold of 0.001, Transit suggests one of the greatest number of groups: 5.

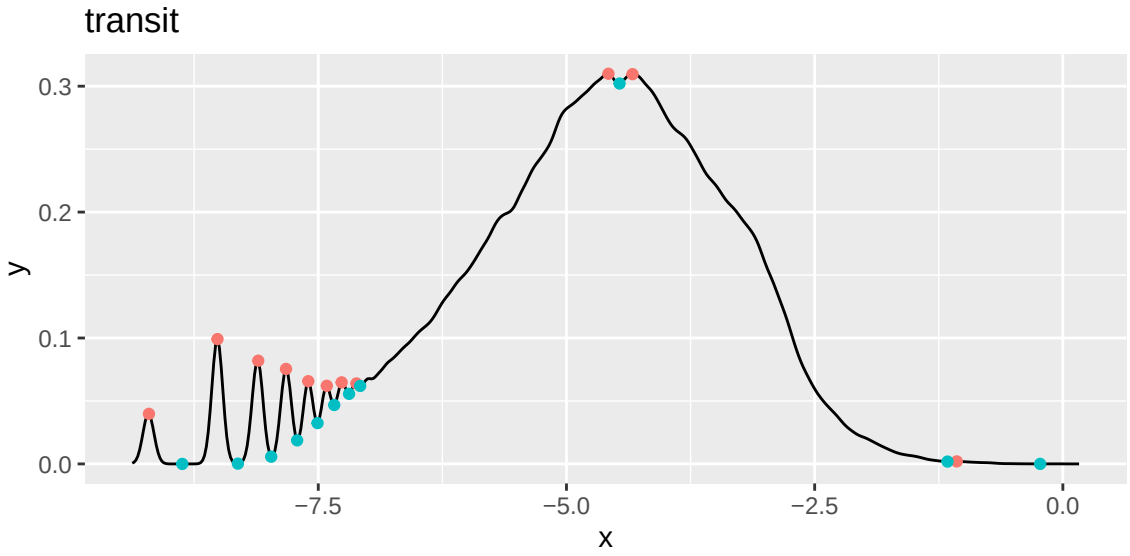

Figure 18: density plot with suggested cutoff points in blue

Summary

amenities	num_groups
PMS_prox_idx_emp	11
PMS_prox_idx_pharma	7
PMS_prox_idx_childcare	4
PMS_prox_idx_health	10
PMS_prox_idx_grocery	8
PMS_prox_idx_educpri	6
PMS_prox_idx_educsec	5
PMS_prox_idx_lib	9
PMS_prox_idx_parks	10
PMS_prox_idx_transit	11

Cluster cut off values:

```
## [1] "PMS_prox_idx_emp"
##  [1] 0.0001416665 0.0002449192 0.0003505853 0.0004480988 0.0005515145
##  [6] 0.0006536468 0.0007602053 0.0012419127 0.0039282360 0.0076057831
## [1] "PMS_prox_idx_pharma"
## [1] 0.01214160 0.01962930 0.02788460 0.03416887 0.06058717 0.40632417
## [1] "PMS_prox_idx_childcare"
## [1] 0.009031823 0.014351423 0.068023147
## [1] "PMS_prox_idx_health"
##  [1] 0.0001424131 0.0002454503 0.0003506350 0.0004475420 0.0005501839
##  [6] 0.0006514448 0.0007569988 0.0170749412 0.3441201516 0.6393408831
## [1] "PMS_prox_idx_grocery"
## [1] 0.01287295 0.01940752 0.02640518 0.05058068 0.06537581 0.19874268 0.27506728
## [1] "PMS_prox_idx_educpri"
## [1] 0.01406477 0.05120818 0.08203794 0.11289269 0.12556912 0.93412279
## [1] "PMS_prox_idx_educsec"
## [1] 0.01833829 0.06673384 0.48797282 0.84523304
## [1] "PMS_prox_idx_lib"
## [1] 0.01778326 0.02859756 0.10925451 0.23443437 0.39001013 0.53836539 0.57617059
## [8] 0.62718190
## [1] "PMS_prox_idx_parks"
## [1] 0.0005484800 0.0007503324 0.0192404741 0.0299465613 0.0409675402
## [6] 0.0560444764 0.0921892765 0.5717702754 0.7967461397
## [1] "PMS_prox_idx_transit"
##  [1] 0.0001406045 0.0002461236 0.0003443861 0.0004472161 0.0005491284
##  [6] 0.0006495612 0.0007541559 0.0008435132 0.0115029755 0.3129091938
## [11] 0.7955635141
```

'Original' cut off values:

```
# i guess have to manually make this list ... ran the above for loop and copied the values here
num_of_clust_manual <- c(5, 3, 3, 4, 3, 3, 2, 2, 3, 5)

cut_offs_manual <- list(
  PMS_prox_idx_emp = c(0.0001423603, 0.0002573125, 0.0003743458, 0.0006010775),
  PMS_prox_idx_pharma = c(0.01152592, 0.01958168),
  PMS_prox_idx_childcare = c(0.008507409, 0.013950871),
```

```
  PMS_prox_idx_health = c(0.0001405736, 0.0002524638, 0.0003658077),
  PMS_prox_idx_grocery = c(0.01218361, 0.01856220),
  PMS_prox_idx_educpri = c(0.04975835, 0.08179011),
  PMS_prox_idx_educsec = c(0.06619424),
  PMS_prox_idx_lib = c(0.6150259),
  PMS_prox_idx_parks = c(0.01844893, 0.02954933),
  PMS_prox_idx_transit = c(0.0001390986, 0.0002481958, 0.0003512994, 0.0004514904, 0.8553858919)
)
```

Next Steps

- Investigate 'weight' of distribution for every suggested groups
- ~~Format table properly~~
- investigate change of bandwidth in kernel density
- change number points in density curve
- investigate with different transforms
- get cutoff values to send for plotting –> for each amenity get x-value of minima

Appendix: Extra Plots

Healthcare

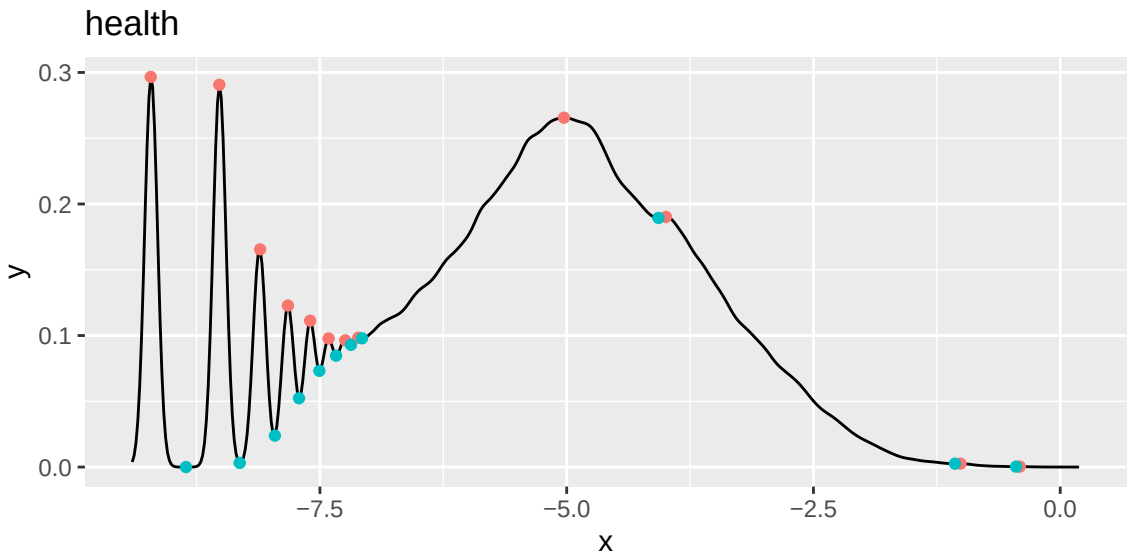

Figure 19: density curve with minima and maxima

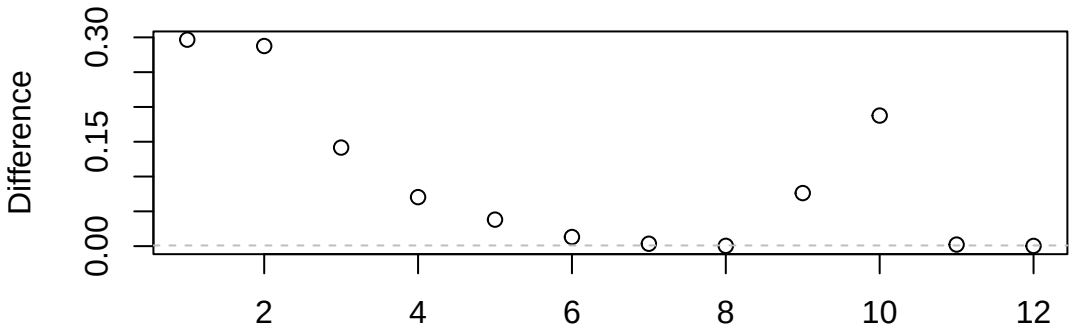

Figure 20: difference between density value of a maxima-minima pairs, threshold = 0.001

Grocery

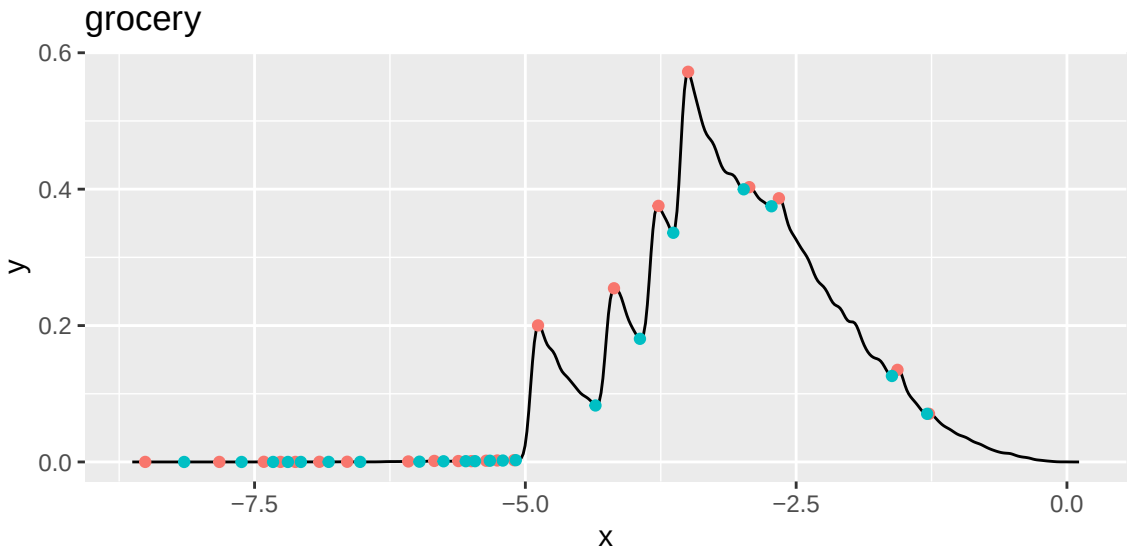

Figure 21: density curve with minima and maxima

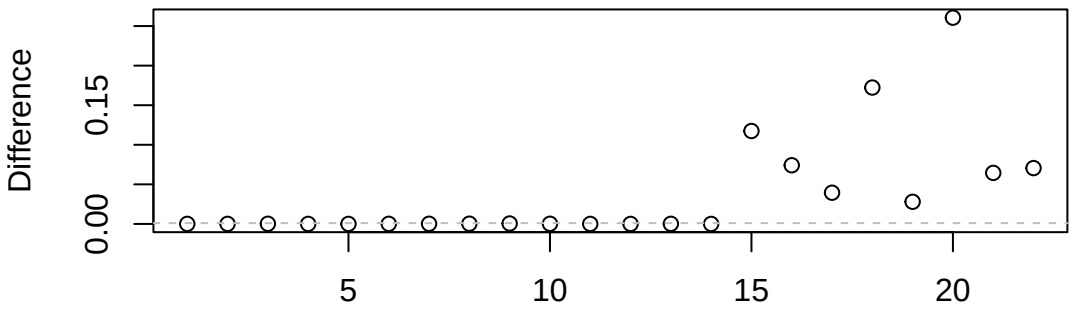

Figure 22: difference between density value of a maxima-minima pairs, threshold = 0.001

Primary Education

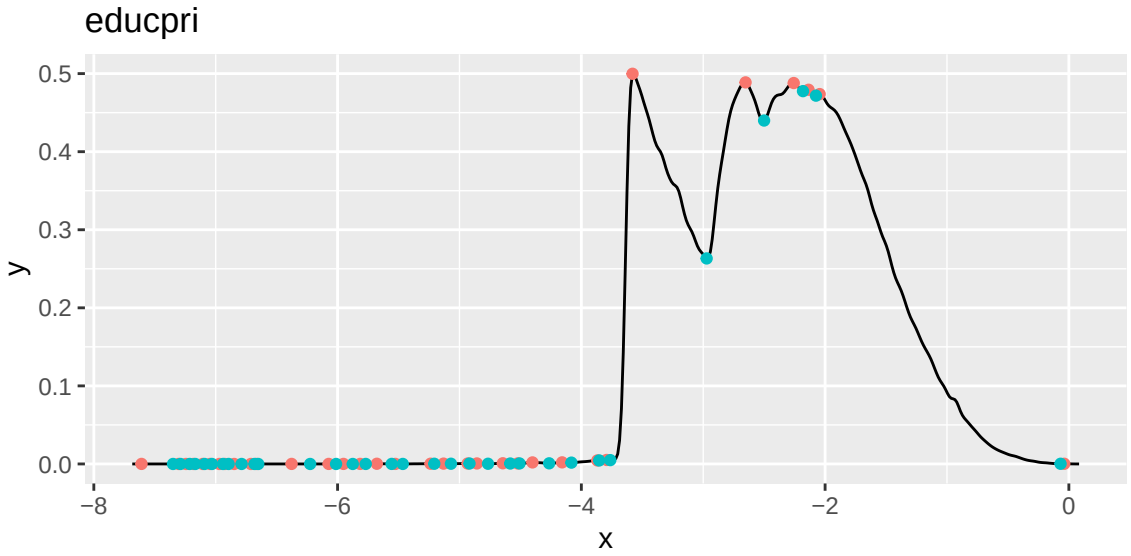

Figure 23: density curve with minima and maxima

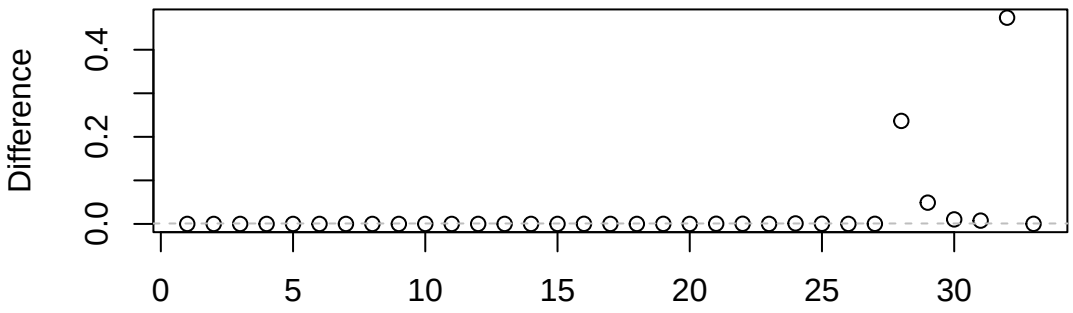

Figure 24: difference between density value of a maxima-minima pairs, threshold = 0.001

Secondary Education

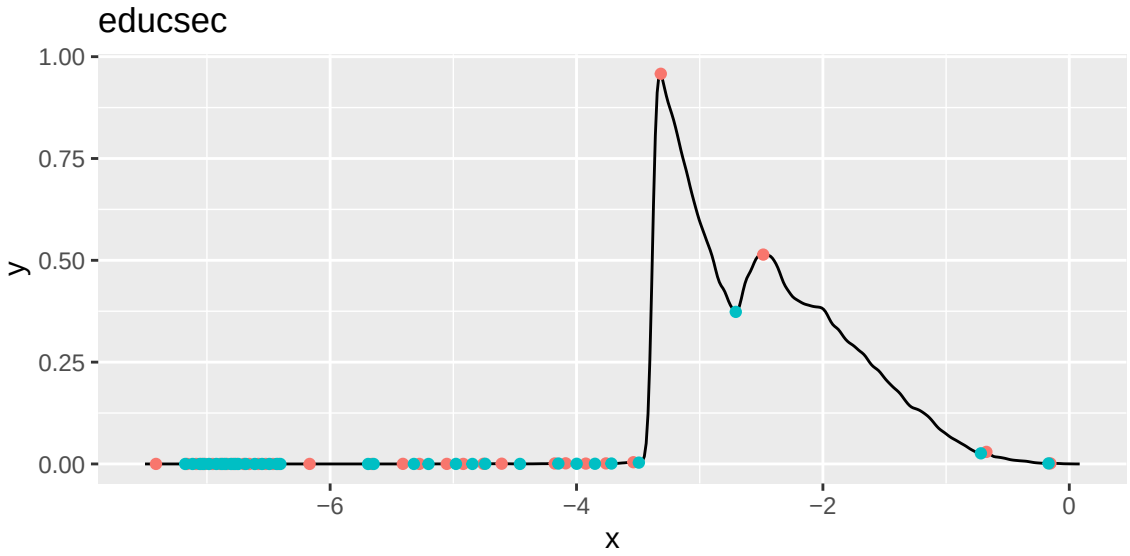

Figure 25: density curve with minima and maxima

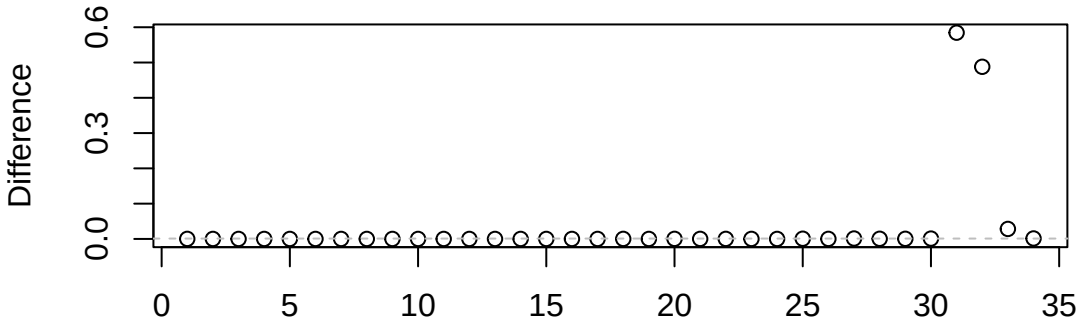

Figure 26: difference between density value of a maxima-minima pairs, threshold = 0.001

Libraries

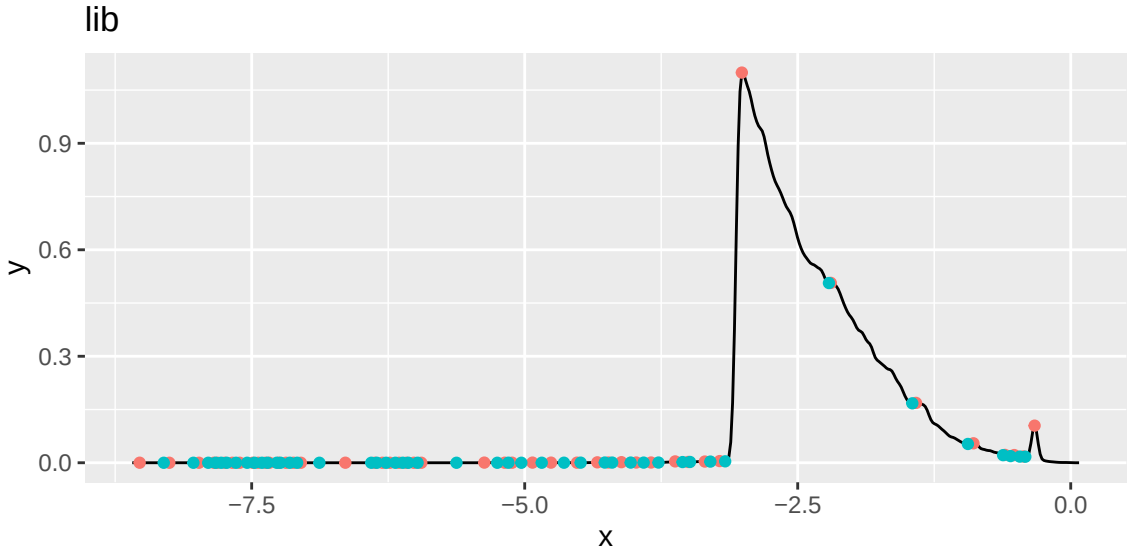

Figure 27: density curve with minima and maxima

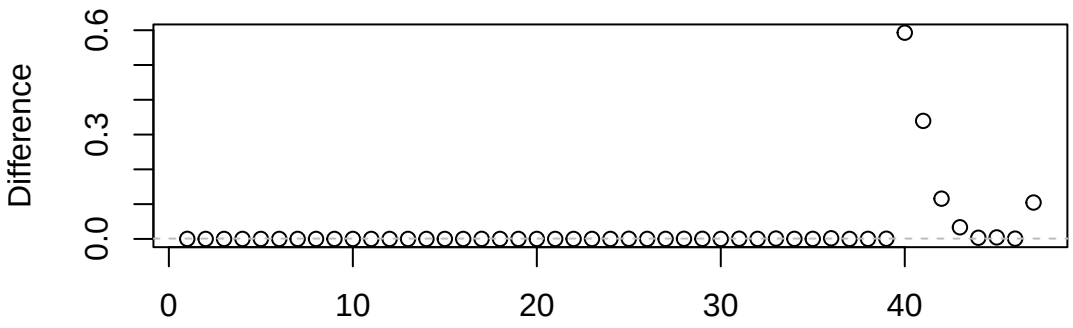

Figure 28: difference between density value of a maxima-minima pairs, threshold = 0.001

Parks

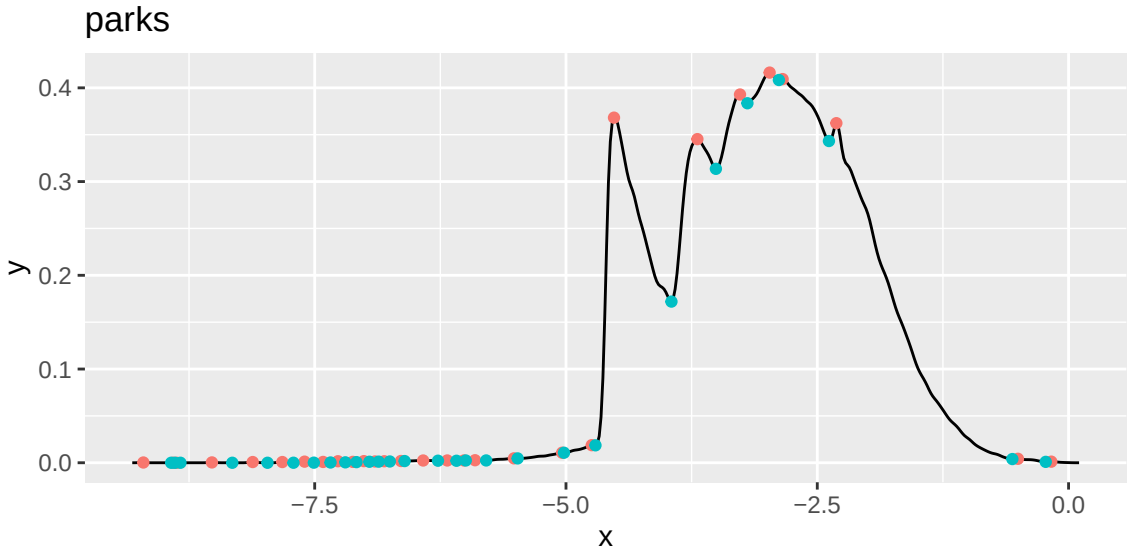

Figure 29: density curve with minima and maxima

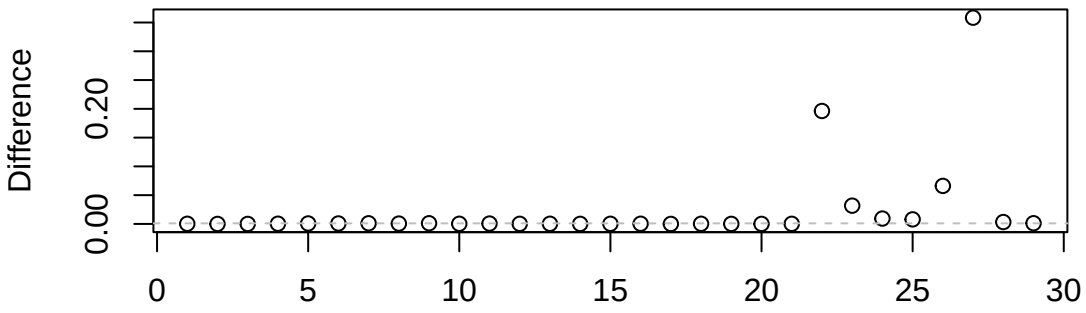

Figure 30: difference between density value of a maxima-minima pairs, threshold = 0.001

Transit

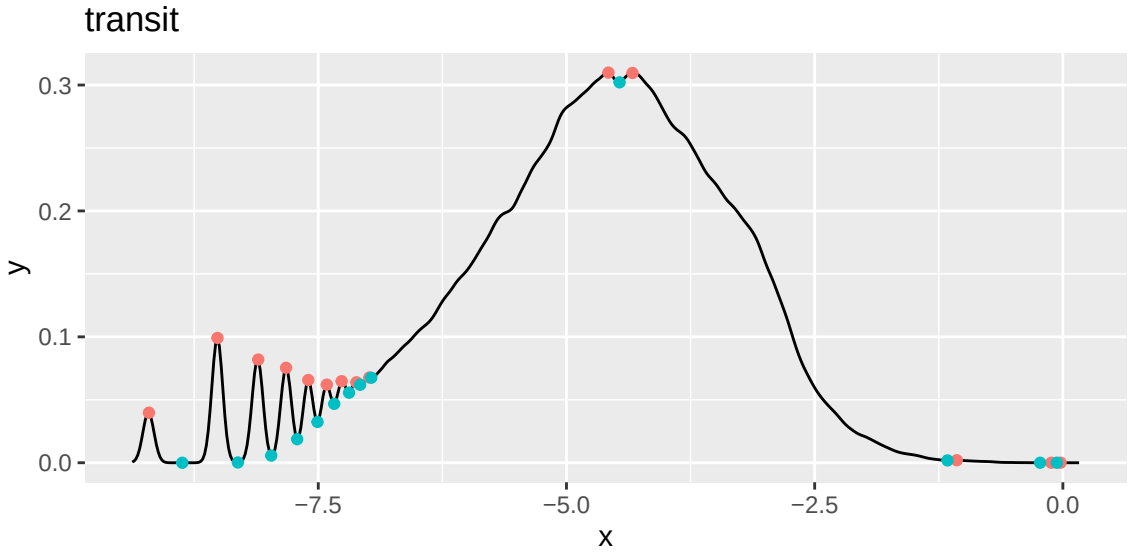

Figure 31: density curve with minima and maxima

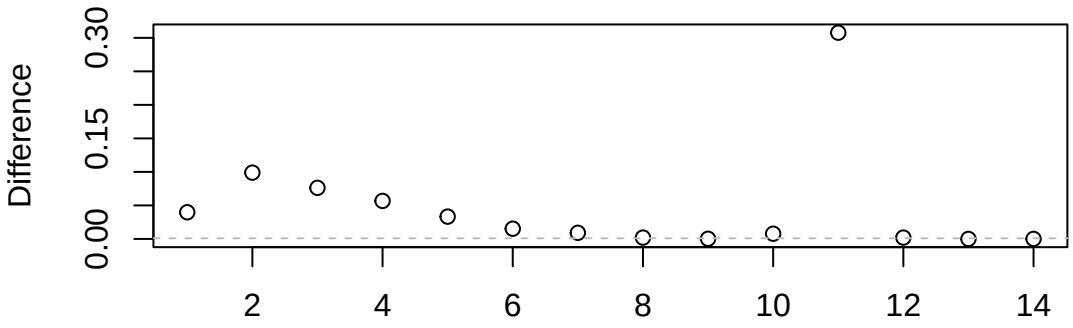

Figure 32: difference between density value of a maxima-minima pairs, threshold = 0.001