Manual cutoffs

Ricky Heinrich

2023-05-29

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 of each amenity is shown in Figure 1.

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. We can already visually identify more possible clusters.

A segmentation technique is to segment the distribution at select minima. Each minimum in the density curves represents a density sparse region, which may be a 'natural' break in the continuous measures.

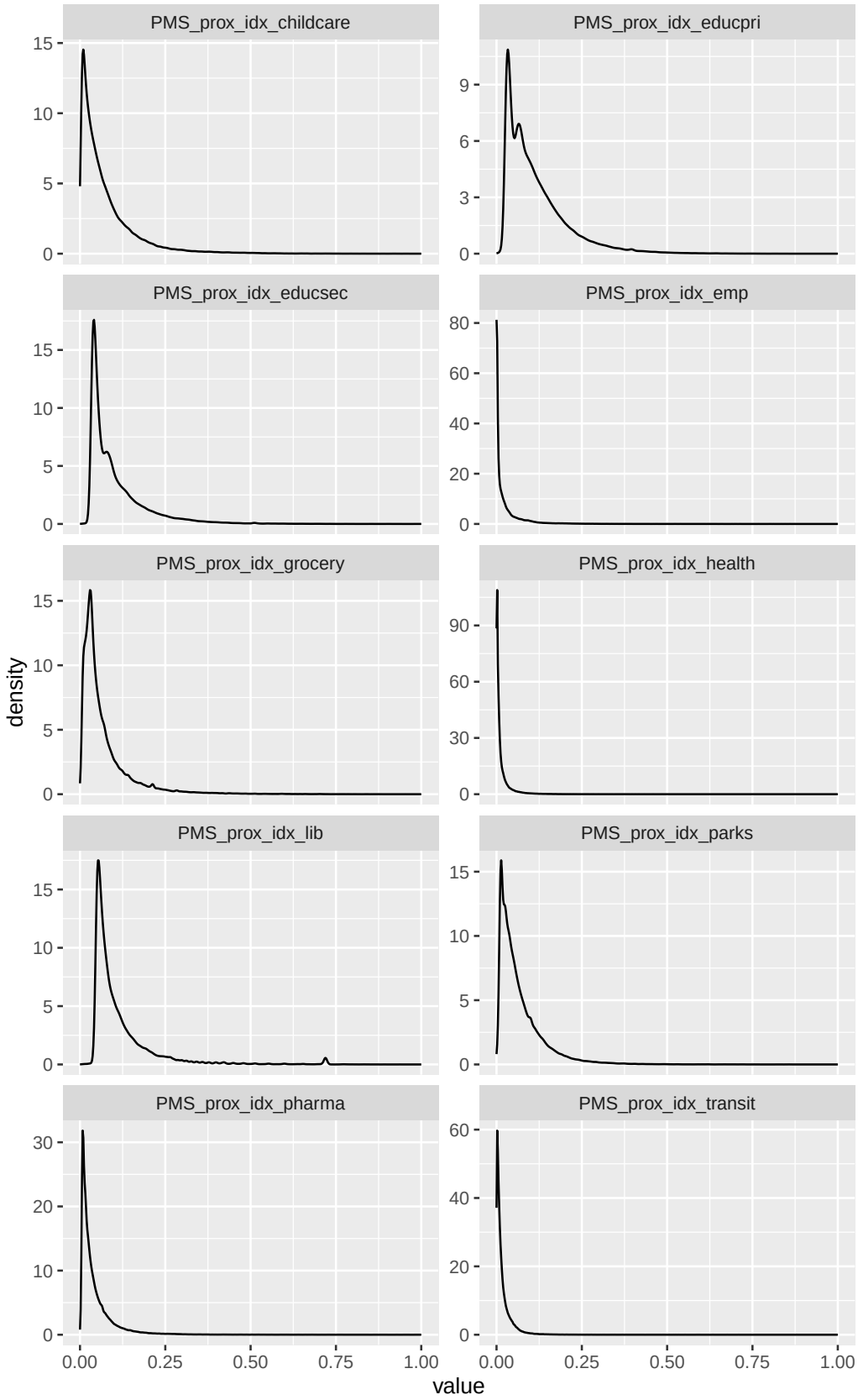

Figure 1: Distribution of proximity measures by amenity

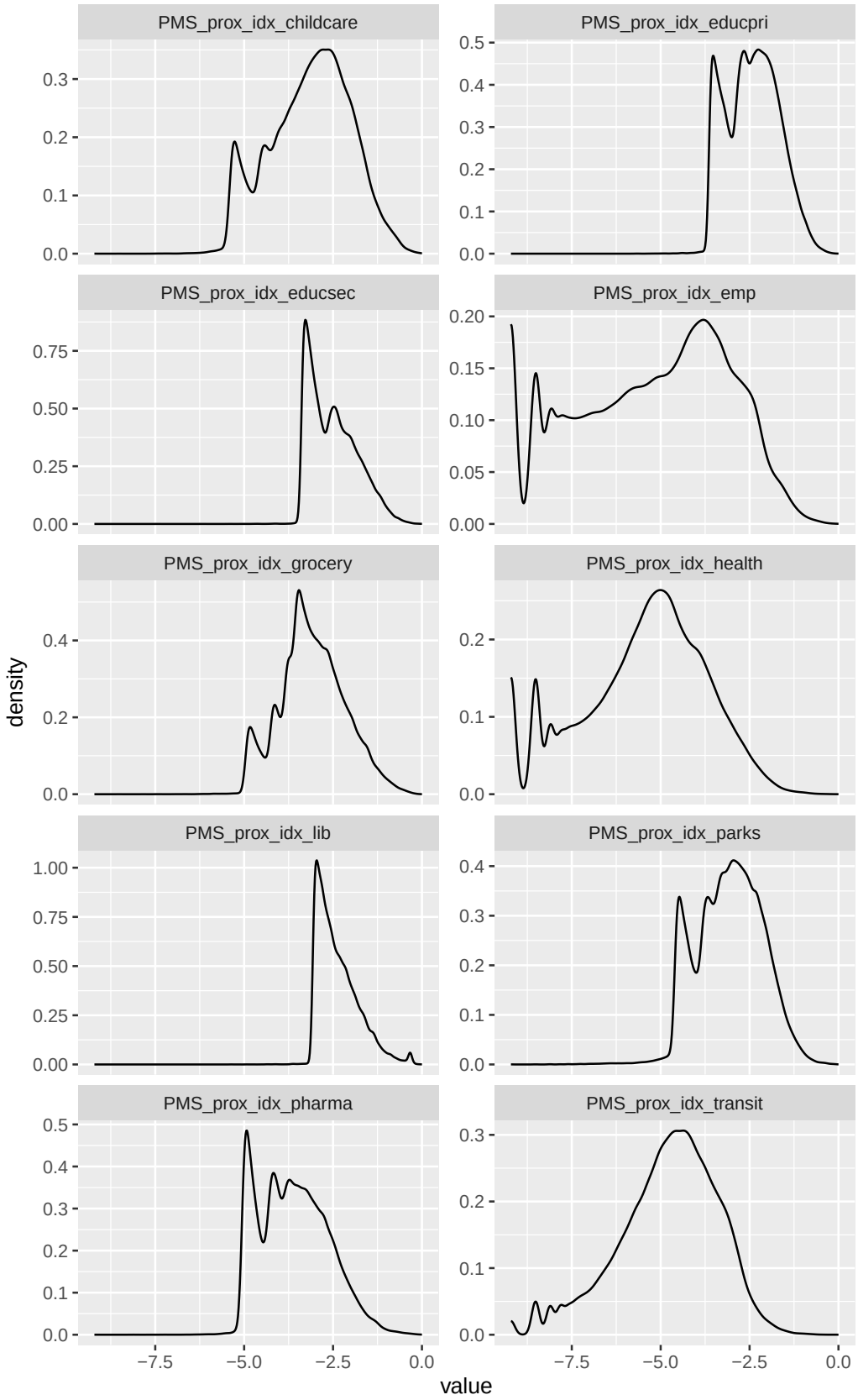

Figure 2: LOG TRANSFORMED(0.0001): Distribution of proximity measures by amenity

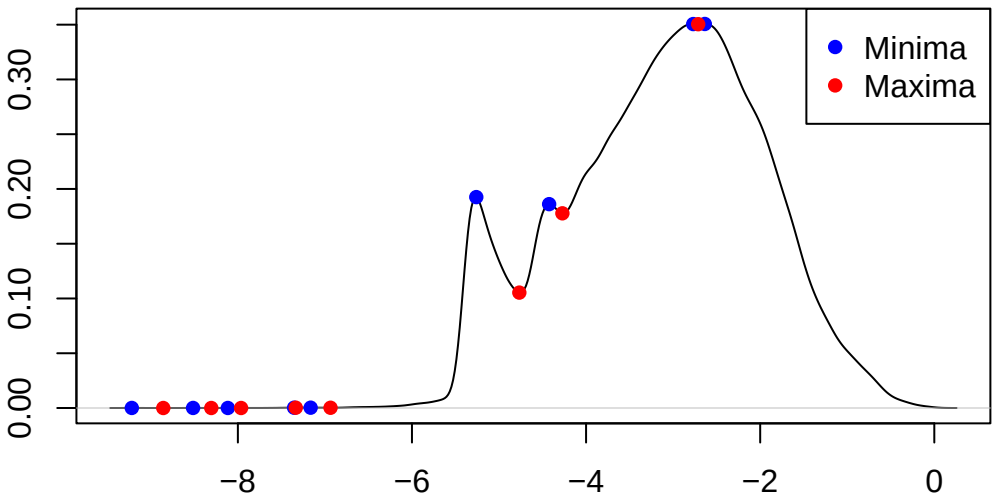

Figure 3: Location of Minima and Maxima