Topological Data Analysis of Big Spatio-Temporal Urban Data

Krasen Samardzhiev

University of Liverpool

Introduction

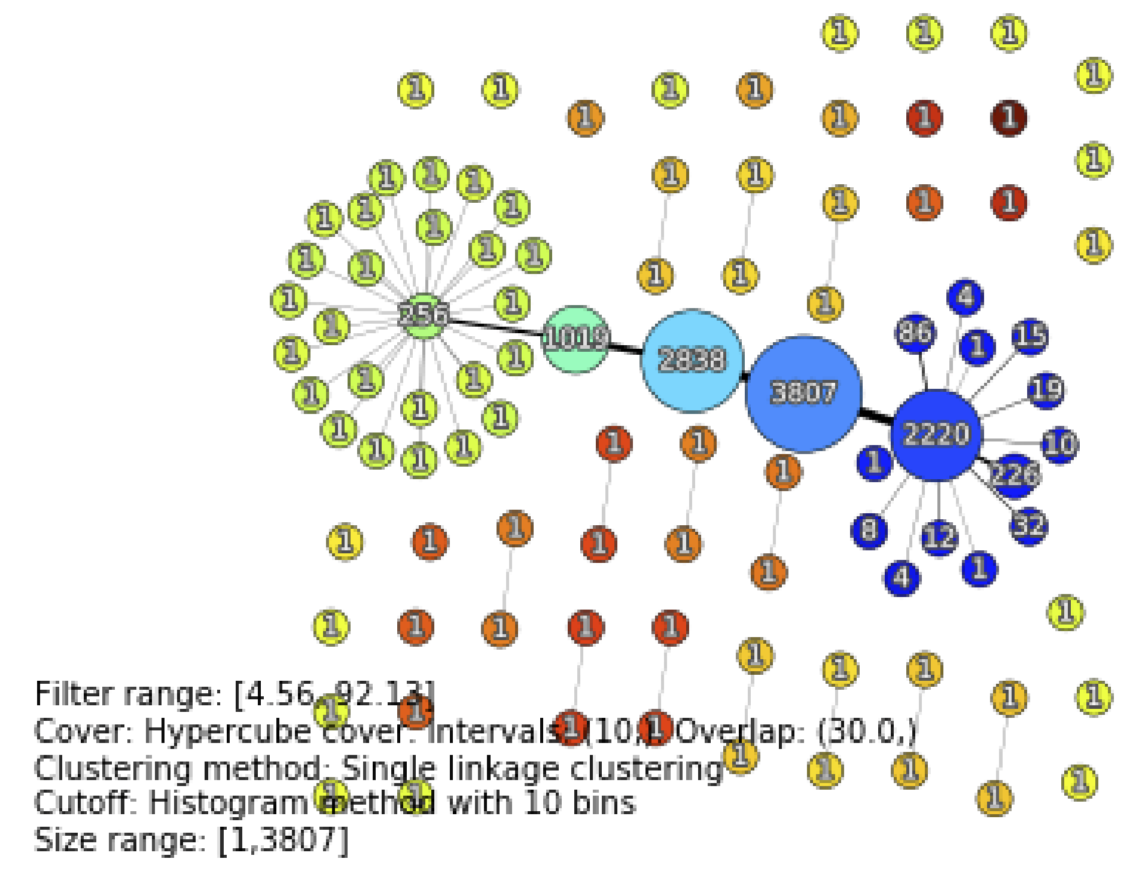
Topological data analysis (TDA) is a new and expanding field, focused on applying insights from mathematical topology, which studies spaces up to continuous deformations, to data. Knowing the coarse ’shape’ of the dataset space is beneficial in modeling. For example, harmonic or quasi-harmonic signals can form a circle in higher dimensional space and topological methods can be used to detect those. Furthermore, the number of connected components and their relationships (clusters) gives important information about how the dataset is connected. The main goal of my project is to bring the techniques of TDA into to the domain of urban data analysis.

By far the two most common ways of analyzing data topologically are computing topological signatures using persistent homology and creating visual summaries of the data for qualitative analysis using Mapper. Mapper is a method that outputs a n-dimensional complex that captures the underlying high dimensional structure of the data, such as loops, flares and connected components. The outputs from the quantitative approach and the possible feature engineering from the qualitative approach can also be used as inputs for other machine learning techniques. There have been successful applications in various fields such as neuroscience, sensors networks, time series analysis and others.

WAC analysis

Workplace Area Characteristics (WAC) data is a subset of United States Census data and consists of counts of how many workers of a certain type, as well as the total number of jobs available there are in a census block. Mapper was identified as the most promising TDA approach to the problem of analysing this dataset, due to its ability to find interesting sub-groups in data from various domains. The goal was to find interesting clusters that other methods missed. To that end multiple experiments were performed using different parameters, to identify sub-populations in the WAC dataset. The results suggested that successful applications of this algorithm rely heavily on domain expertise and that mapper cannot be used as a substitute for classical geodemographic methods. Additionally, mapper might be best suited for finding interesting connections traditional methods miss, but in already extensively studied datasets, rather than unfamiliar ones.

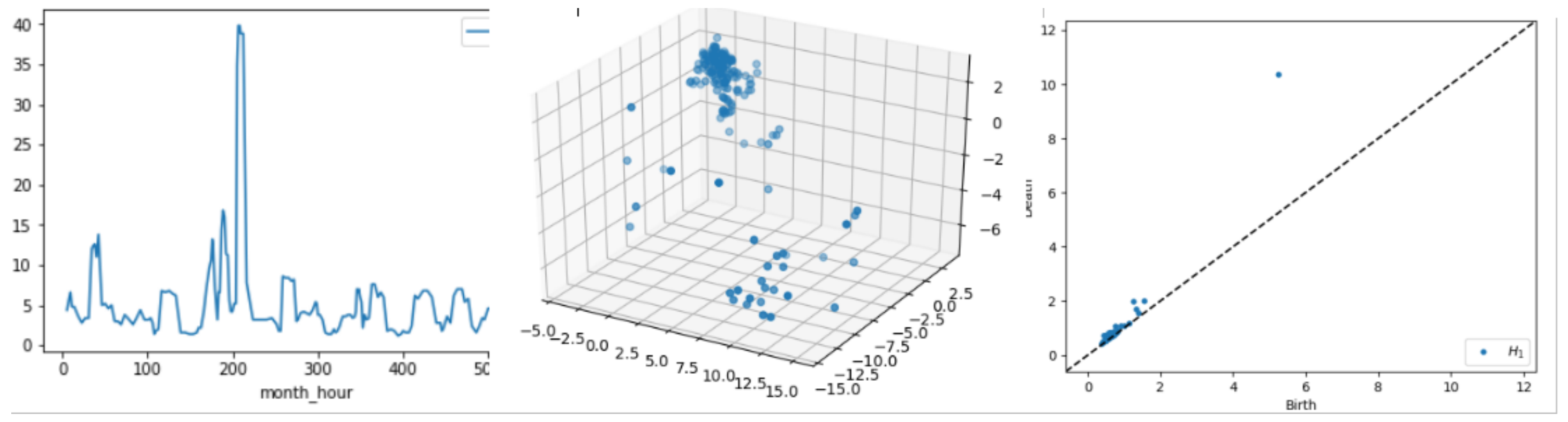
```
<matplotlib.collections.CircleCollection at 0x7ff27c072518>)
```



Filter range: [4.56, 92.13]
Cover: Hypercube cover. Intervals: (10,) Overlap: (30.0,)
Clustering method: Single linkage clustering
Cutoff: Histogram method with 10 bins
Size range: [1,3807]

Movement data analysis

In general, movement data is used to learn something about the objects themselves. A different approa is to focus on analyzing the areas through which the objects move. In this project movement data will aggregated by time and area and the resulting time series will be analysed. The goal is to see whether TI techniques such as persistent homology and sl1ding windows can be used to improve on already existing gorithms for time series classification. The focus will be on finding and clustering together areas that exhi similar periodic behaviour. Currently, baseline and sl1ding window methods are implemented and were tes on aggregated skyhook GPS data. A variety of other datasets will be used for future experiments.


month_hour
Death
Birth
H_1