University College Cork

CS3205 Lab Report 1

HDI Trends and Multidimensional Projections

Jack O'Connor

February 26, 2022

Contents

1 Introduction

1.1 List of Acronyms

UCC (University College Cork), HDR (Human Development Report), HDI (Human Development Index), GNI (Gross Naional Income), GDP (Gross Domestic Product), PCA (Principal Component Analysis), LSP (Least Squares Projection), ProjClus (Projection Clustering), t-SNE (t-distributed Stochastic Neighbour Embedding), CBR (Case-Based Reasoning).

1.2 The Report

This report was created to document my experience developing my data visualisation skills as part of UCC's CS3205 Data Visualisation module. It is divided into two separate parts, each tackling its own separate area of data visualisation. As well as documenting my experience exploring the specified datasets and the tools I chose to use, this report also served as an instructive exercise in typesetting and formatting an academic paper, which is sure to prove useful next year when I tackle my final year project.

Part 1 of this report deals with visualising survey data, namely in the form of the annual Human Development Report (HDR) dataset, to find attribute patterns using a wide variety of methods. Part 2 of this report uses the HDR Dataset as well as several other sample labeled datasets (which will be discussed in depth in the data description section) to create comprehensible projections into the 2-dimensional plane of the point-like datasets using several multi-dimensional scaling techniques.

As one of the primary goals of this report is to compare the different visualisation methods against each other, I have formatted the document into two or three columns where appropriate, such that multiple visualisations are visible on each page. However, this does come at the cost of reducing the size of each visualisation. To offset this issue each image in this report is hyperlinked to a full size version hosted on either GitHub or Tableau. I would suggest viewing this pdf report in your browser such that the need to switch applications when viewing full size images is eliminated.

The full project repository containing all scripts, datasets and other miscellaneous files needed to reproduce this report can be found here.

1.3 The Role of Visualisation

Humans being a strongly visually oriented species means visualisations are a key part of any data analysis. A well formulated visualisation can turn an incomprehensible raw dataset into a graphic full of valuable information for our highly optimised pattern seeking brains.

That does not mean that all visualisation techniques are suitable for all data analysis tasks. Care must be taken with the transformation of the raw data into visualisations that the resulting visual is actually meaningful, and not just misleading noise. Tasks 1 and 2 of this report are a good example of distinguishing when and when not to use different visualisation techniques.

Task 1 uses the HDR dataset which has a (compared to Task 2's datasets) relatively small number of attributes, each of which is meaningful in its own right i.e. corresponds to an attribute of a country which has a meaningful, physical interpretation such as a country's total population or gross domestic produce (GDP). Techniques which compare individual attributes directly against each other can expose correlations between attributes which might not be obvious at first.

Task 2 on the other hand uses highly multidimensional data such as images where each pixel of an image can be attribute of that data, or text documents where the words of each document are embedded into a vector space with hundreds or thousands of dimensions. Each individual attribute of these datasets on its own does not carry much weight in the context of an image or docuement and directly comparing them to each other is unlikely to yield any salient information. In such an instance it is much more useful to project each data point into a 2 or 3-dimensional space and seek more broad patterns between documents such as clustering and dissimilarity.

2 Data Description

2.1 Task 1 Datasets

I used two datasets to complete task 1, HDR 2020

2.2 Task 2 Datasets

Description of data.

3 Task 1

3.1 Free Exploration

In the free exploration section of the assignment I have mainly (but not exlusively) created visualisations which give a broad overview of the general distribution of HDI across countries.

Each of the following visualisations will include:

- A graphic
- A description of a pattern (or lack of)
- A hypothesis as to the cause of this pattern
- The dataset from which the data came

3.1.1 HDI Histogram

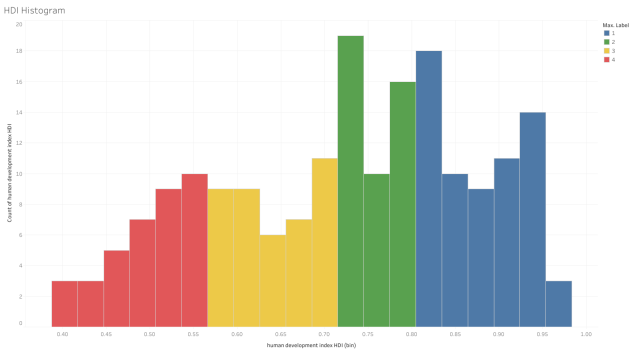

Figure 3.1: my hdi caption

3.1.2 HDI Global Heatmap

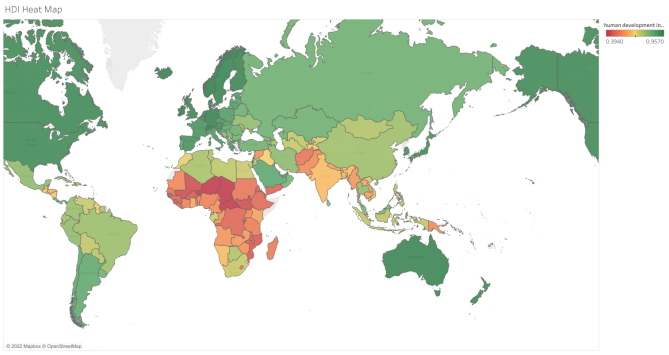

Figure 3.2: what a pretty map

3.1.3 Average HDI Trend

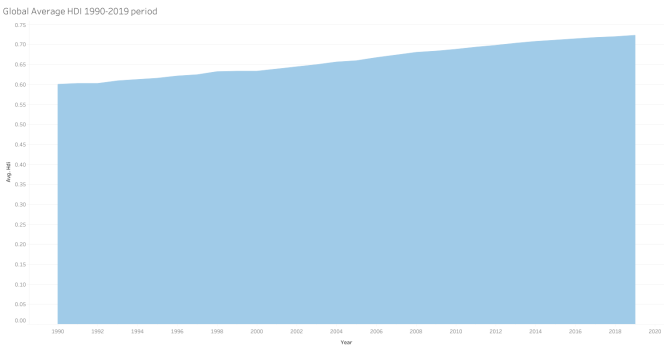

Figure 3.3: no dips

3.1.4 Min/Max HDI

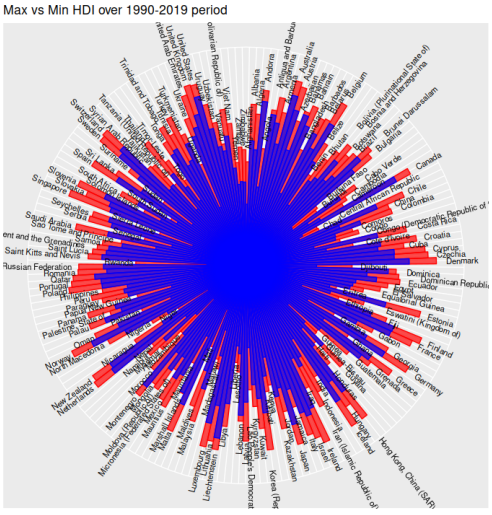

Figure 3.4: going in circles

3.1.5 Male/Female Mean Years Education

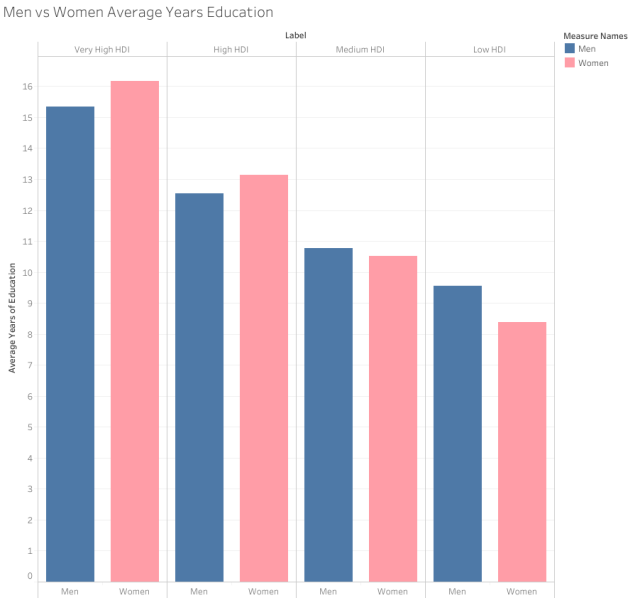

Figure 3.5: girl power

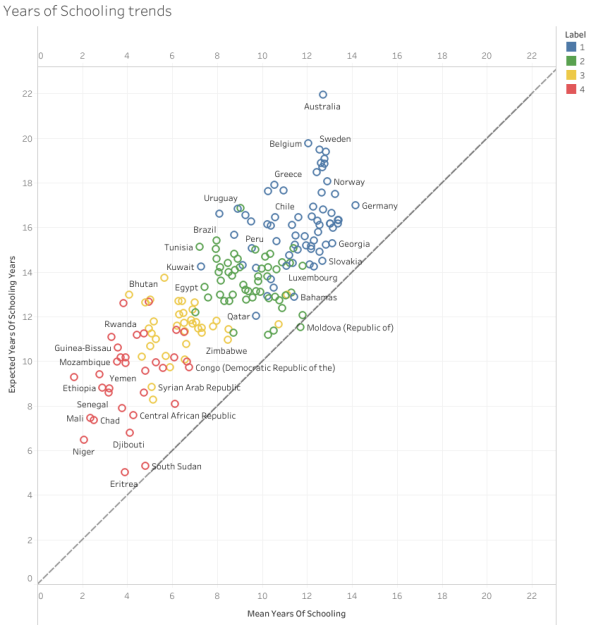

Figure 3.7: spurious value 2

3.2 Specific Observations

In the specific observations section of the assignment I have created visualisaions in line with the task list specifications.

3.2.1 Spurious Values

Expected Population Growth

Figure 3.6: sexy visual spurious values

Mean vs Expected Years Schooling

3.2.2 Correlated Attributes

Fertility Rate vs Adolescent Birth Rate

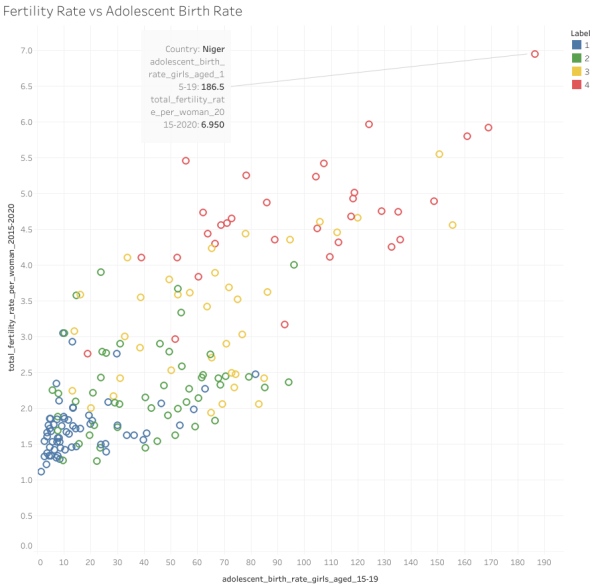

Figure 3.8: Early have children more children you have

HDI vs Gross National Income

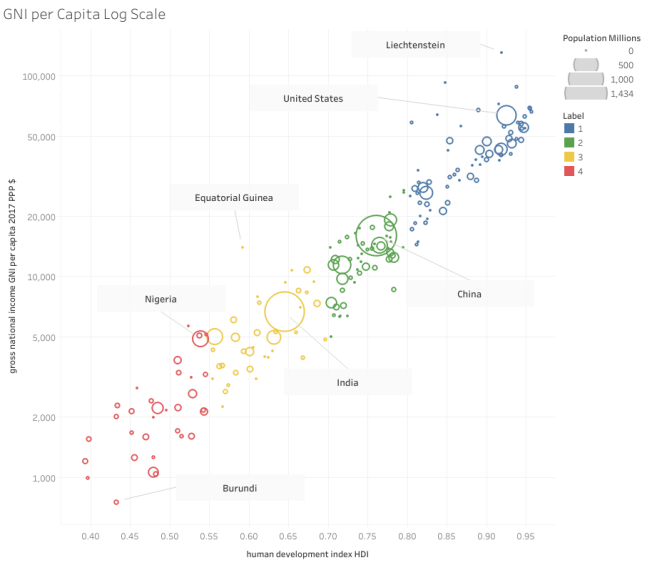

Figure 3.9: GNI accounts almost totally for HDI

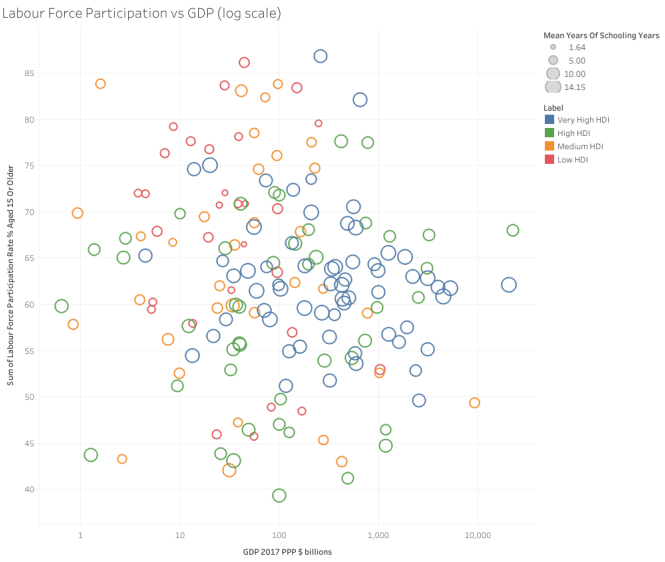

Figure 3.11: Surprising

3.2.3 Uncorrelated Attributes

Women Share of Seats in Government vs Average Years Education

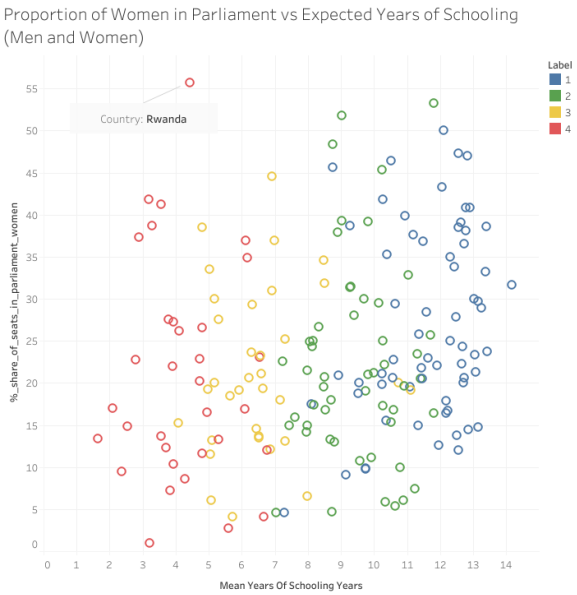

Figure 3.10: Surprising

Labour Force Participation vs GDP

3.2.4 Hypotheses for Previous Patterns

Health Expenditure ordered by Life Expectancy

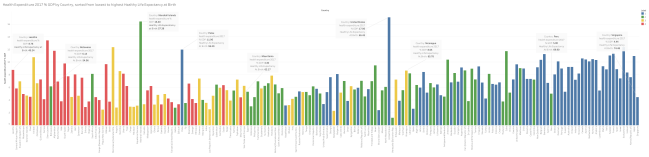

Figure 3.12: think of caption

3.2.5 Usefulness of All Visualisations

3.2.6 Alternative Visualisations

Min/Max HDI Alternative

Figure 3.13: too long for computer screen, can do horizontal line comparisons

Mean vs Expected Years Schooling

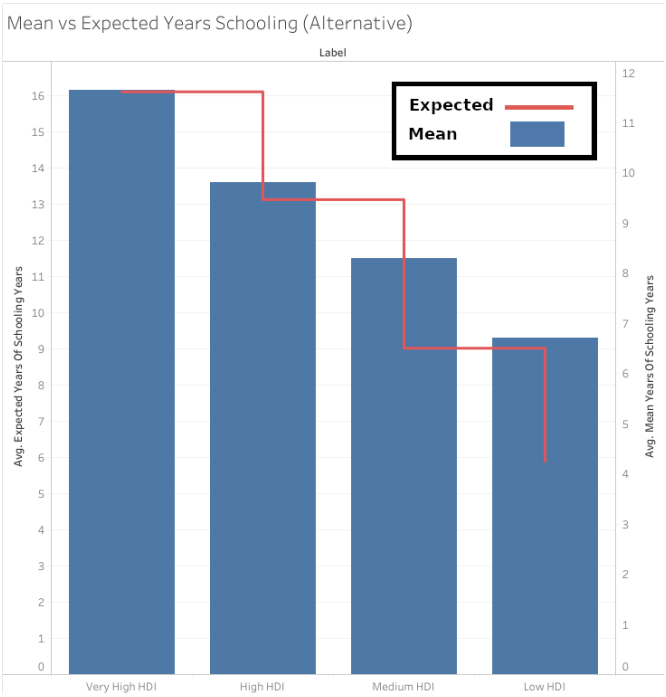

Figure 3.14: Lose individual countries

4 Task 2

Results for task 2.

4.1 Corel Projections Comparison

4.1.1 LSP

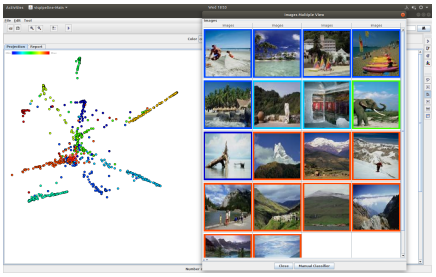

Figure 4.1: First projection

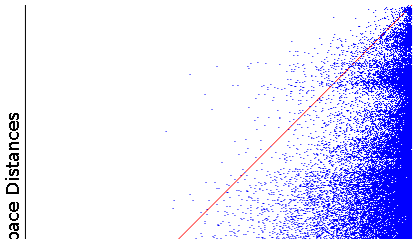

Figure 4.2: Stress Curve

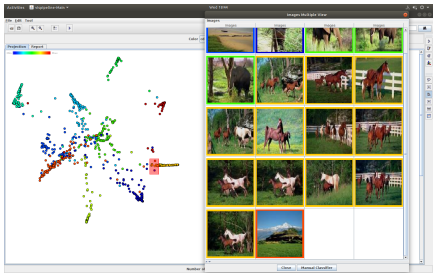

Figure 4.3: Second projection

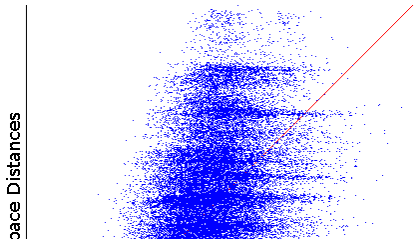

Figure 4.4: Stress Curve

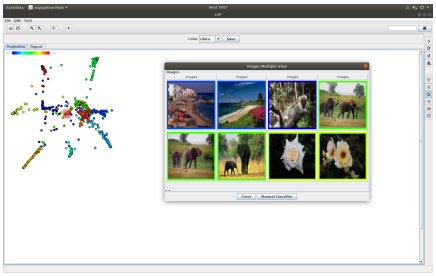

Figure 4.5: Third projection

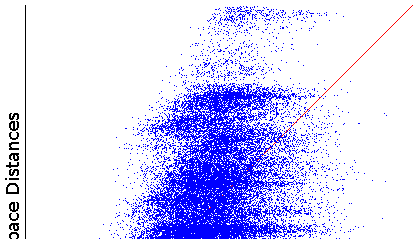

Figure 4.6: Stress Curve

4.1.2 t-SNE

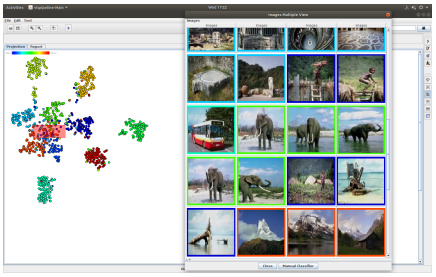

Figure 4.7: First projection

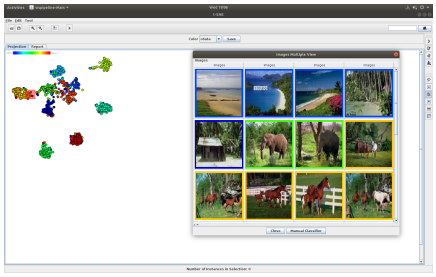

Figure 4.9: Second projection

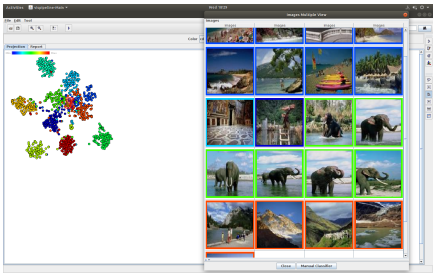

Figure 4.11: Third projection

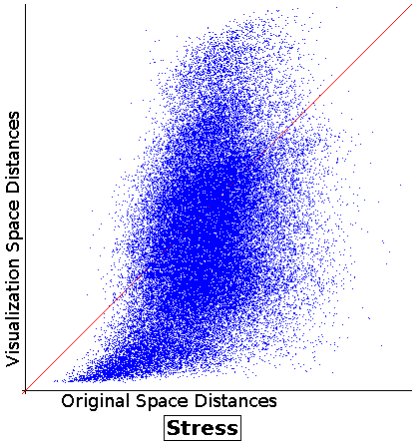

Figure 4.8: Stress Curve

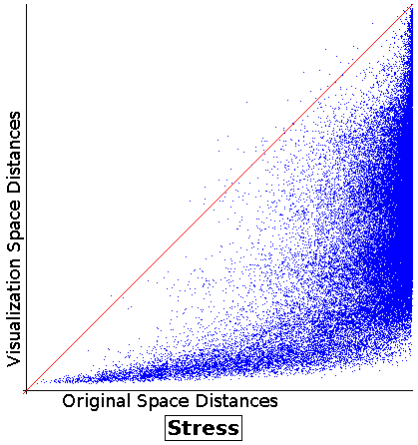

Figure 4.10: Stress Curve

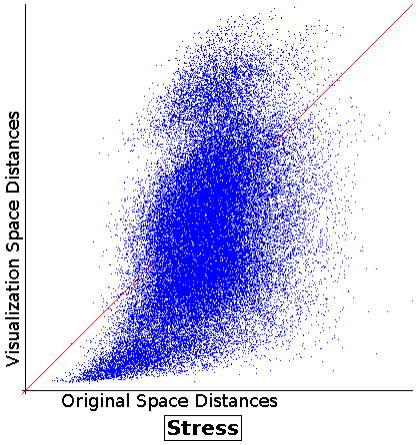

Figure 4.12: Stress Curve

4.1.3 PCA

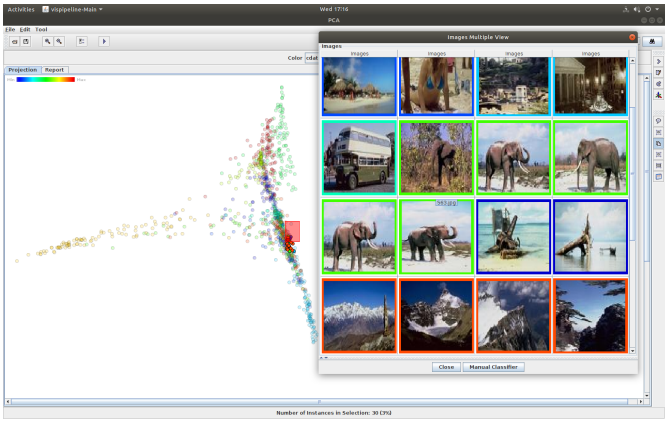

Figure 4.13: First projection

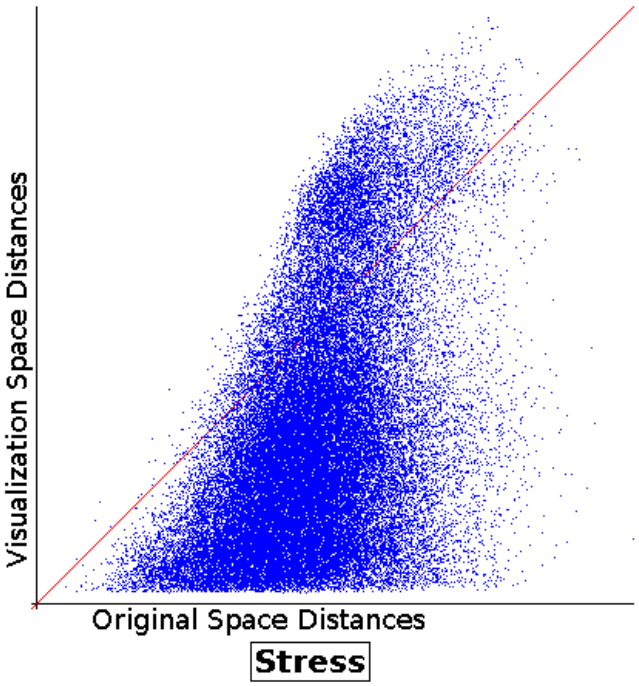

Figure 4.14: Stress Curve

4.2 CBR Projections Comparison

4.2.1 LSP

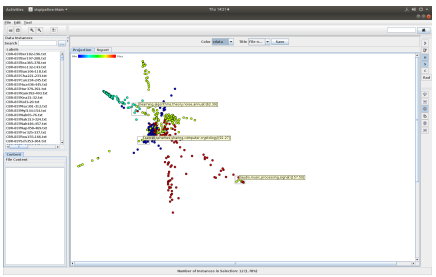

Figure 4.15: First projection

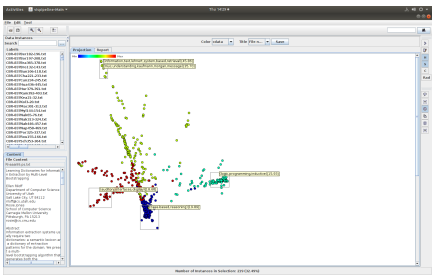

Figure 4.17: Second projection

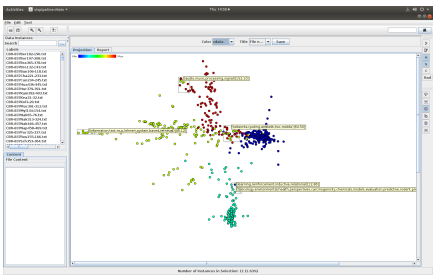

Figure 4.19: Third projection

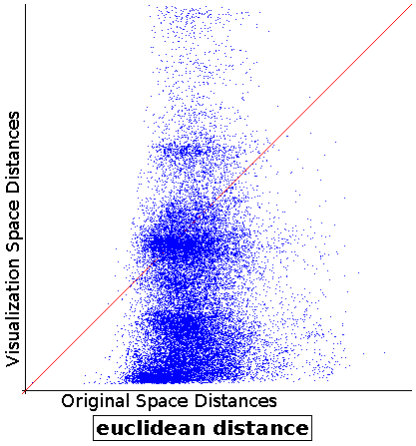

Figure 4.16: Stress Curve

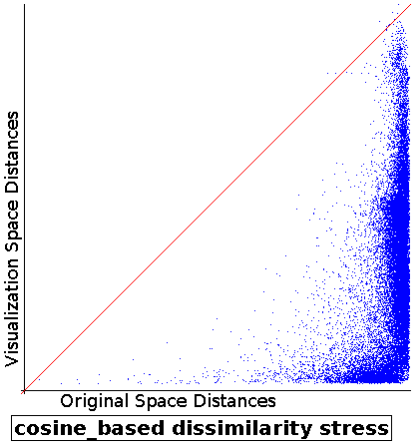

Figure 4.18: Stress Curve

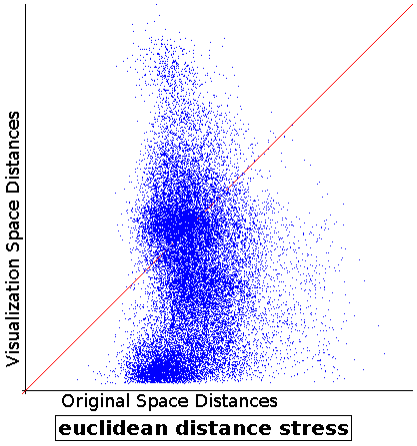

Figure 4.20: Stress Curve

4.2.2 t-SNE

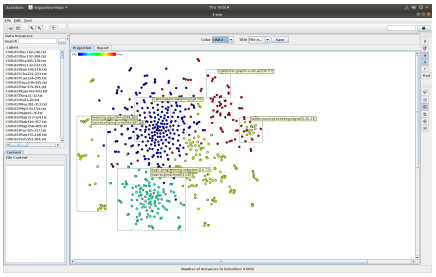

Figure 4.21: First projection

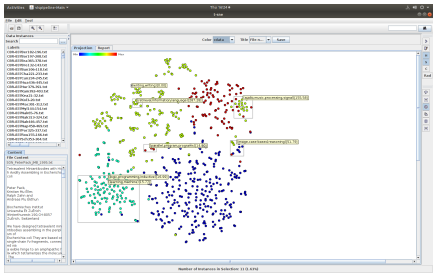

Figure 4.23: Second projection

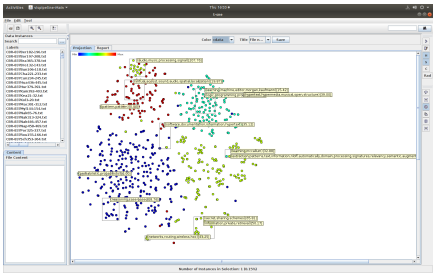

Figure 4.25: Third projection

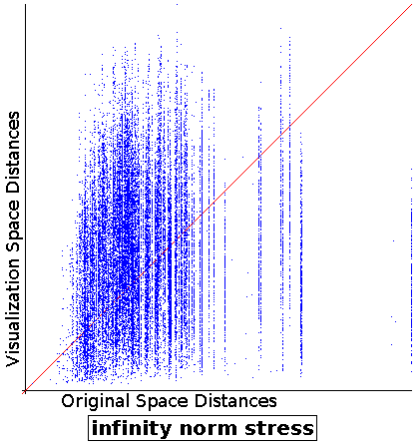

Figure 4.22: Stress Curve

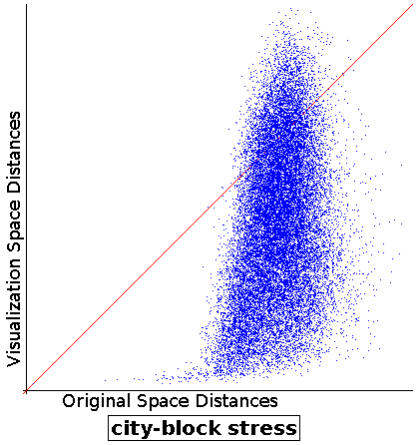

Figure 4.24: Stress Curve

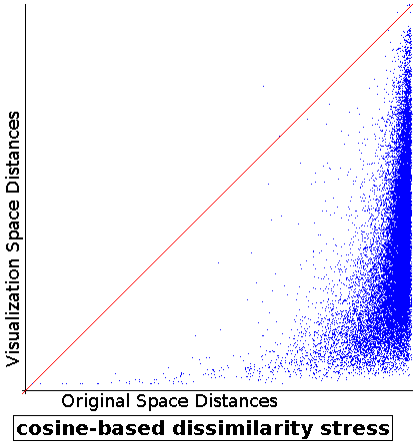

Figure 4.26: Stress Curve

4.2.3 ProjClus

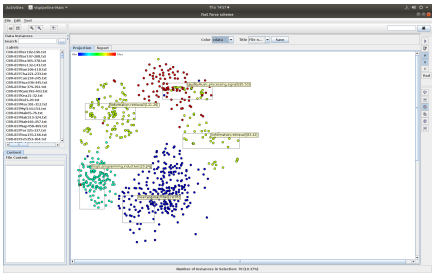

Figure 4.27: First projection

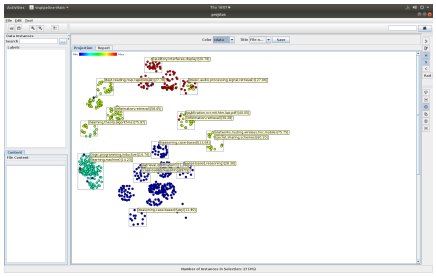

Figure 4.29: Second projection

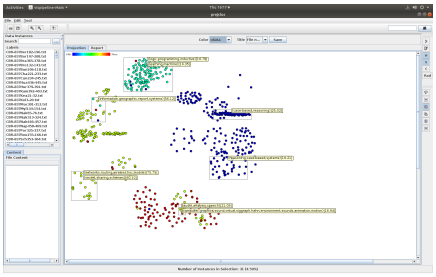

Figure 4.31: Third projection

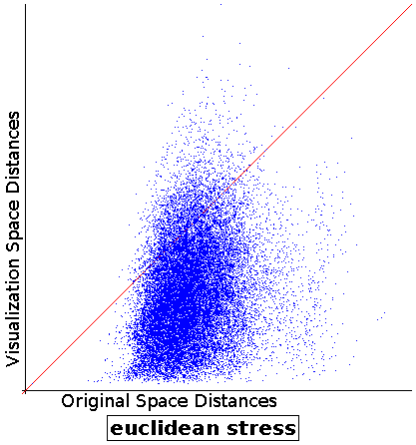

Figure 4.28: Stress Curve

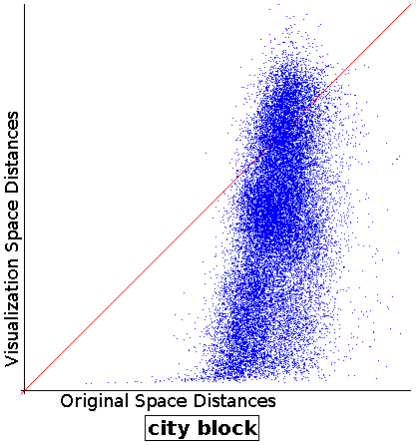

Figure 4.30: Stress Curve

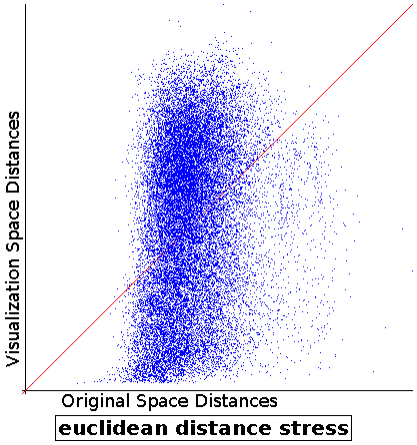

Figure 4.32: Stress Curve

4.3 Medical Images Projections Comparison

4.3.1 LSP

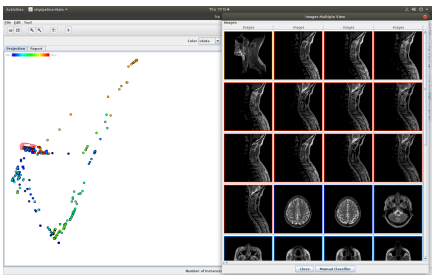

Figure 4.33: First projection

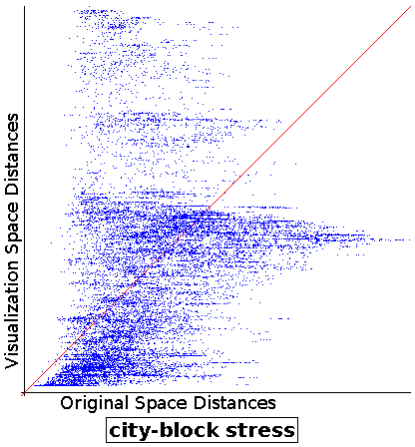

Figure 4.34: Stress Curve

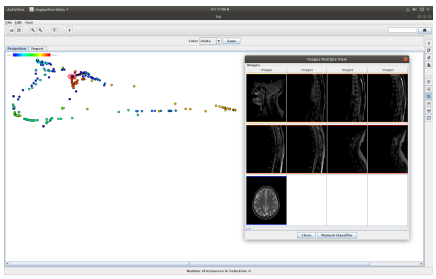

Figure 4.35: Second projection

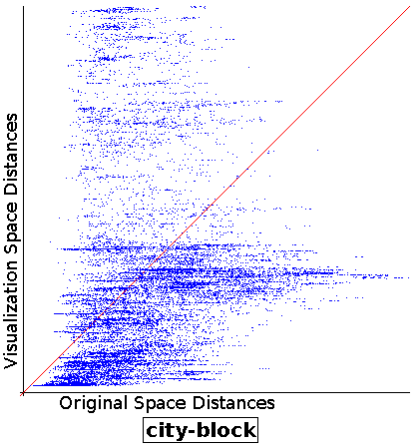

Figure 4.36: Stress Curve

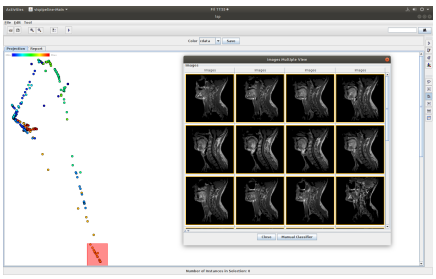

Figure 4.37: Third projection

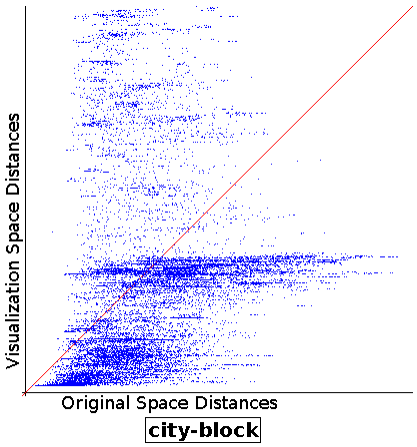

Figure 4.38: Stress Curve

4.3.2 t-SNE

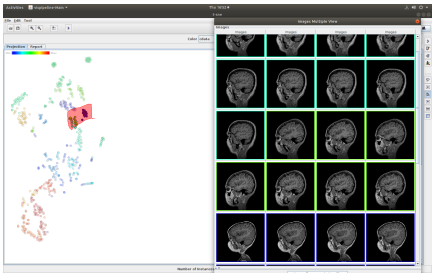

Figure 4.39: First projection

Figure 4.41: Second projection

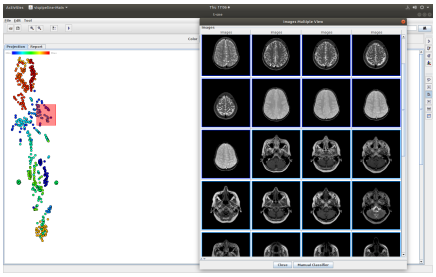

Figure 4.43: Third projection

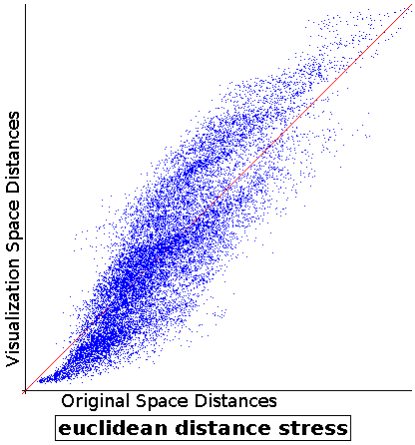

Figure 4.40: Stress Curve

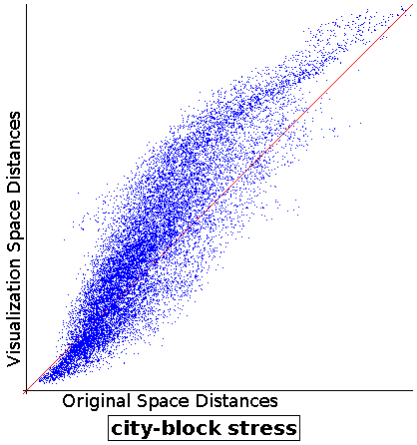

Figure 4.42: Stress Curve

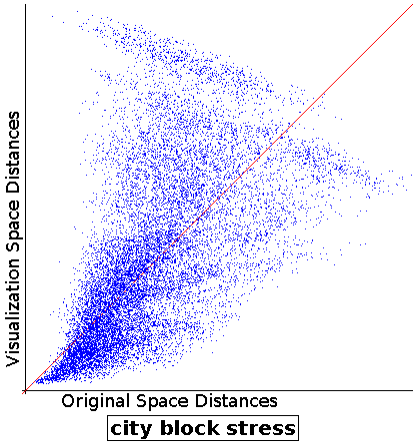

Figure 4.44: Stress Curve

4.3.3 PCA

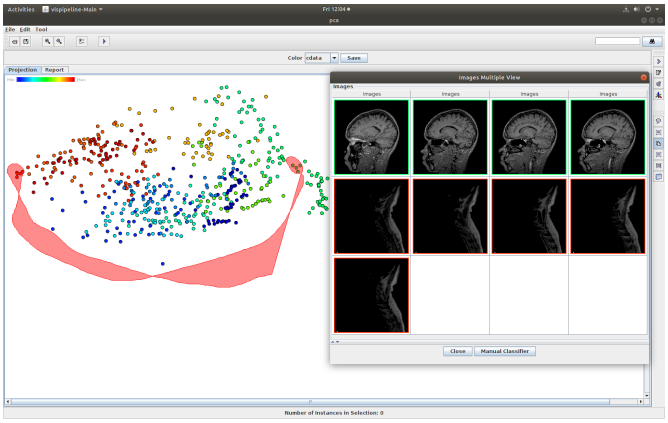

Figure 4.45: First projection

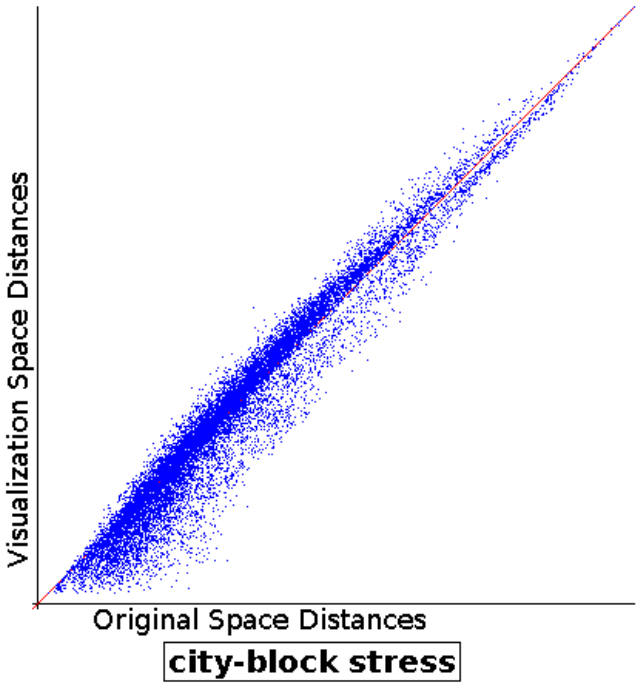

Figure 4.46: Stress Curve

4.4 News Headlines Projections Comparison

4.4.1 LSP

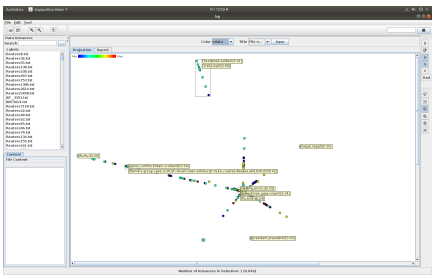

Figure 4.47: First projection

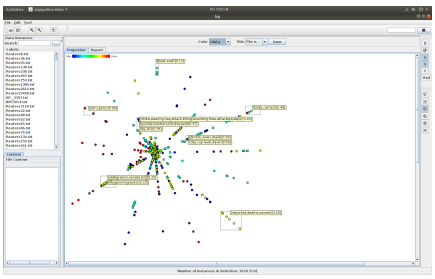

Figure 4.49: Second projection

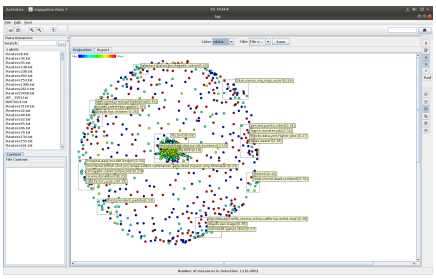

Figure 4.51: Third projection

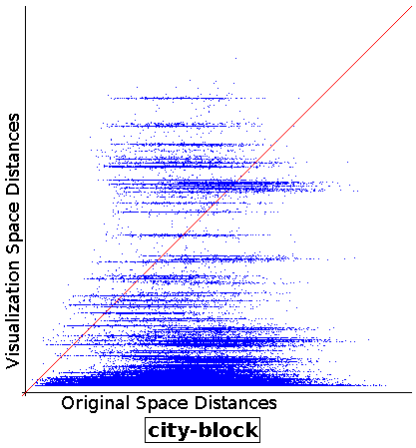

Figure 4.48: Stress Curve

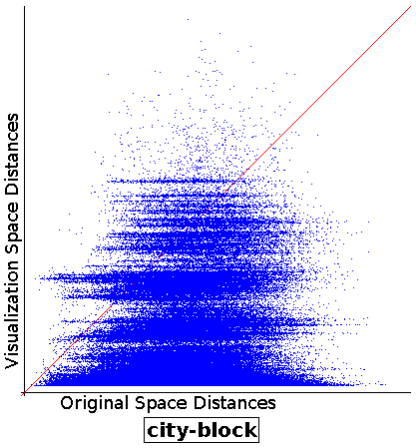

Figure 4.50: Stress Curve

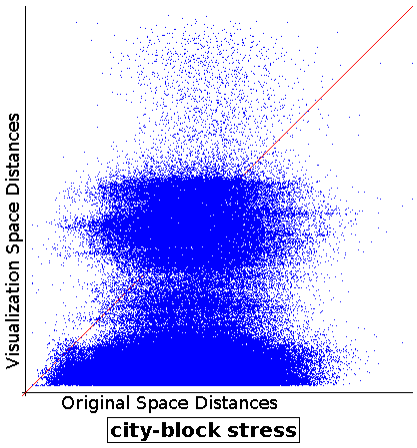

Figure 4.52: Stress Curve

4.4.2 t-SNE

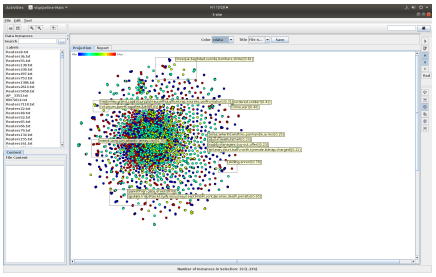

Figure 4.53: First projection

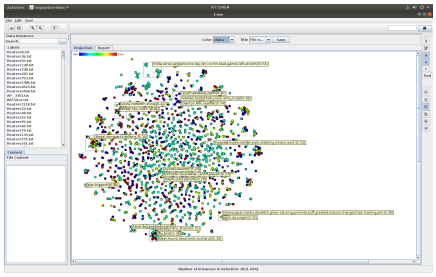

Figure 4.55: Second projection

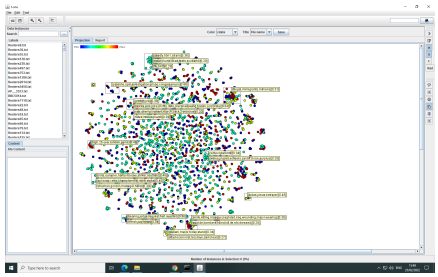

Figure 4.57: Third projection

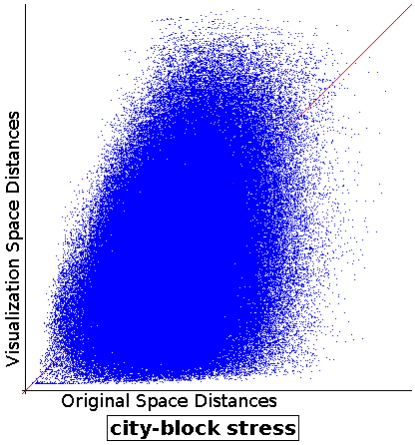

Figure 4.54: Stress Curve

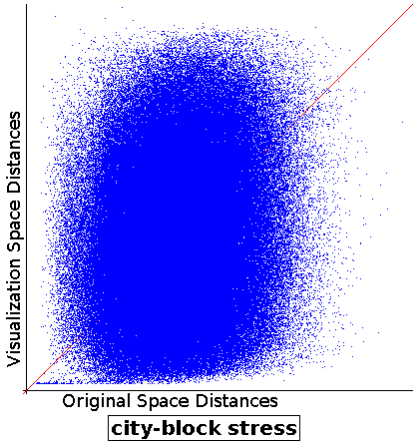

Figure 4.56: Stress Curve

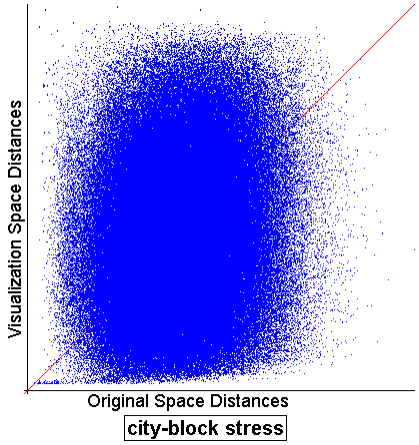

Figure 4.58: Stress Curve

4.4.3 ProjClus

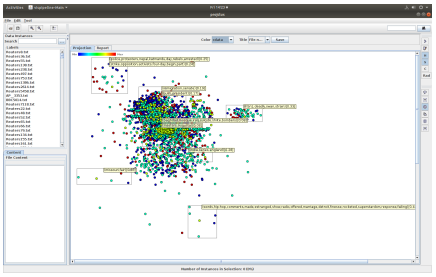

Figure 4.59: First projection

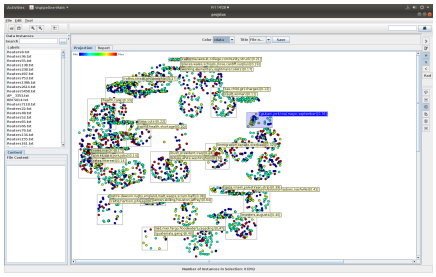

Figure 4.61: Second projection

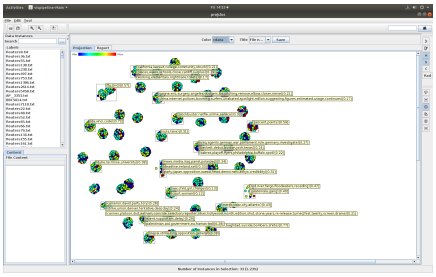

Figure 4.63: Third projection

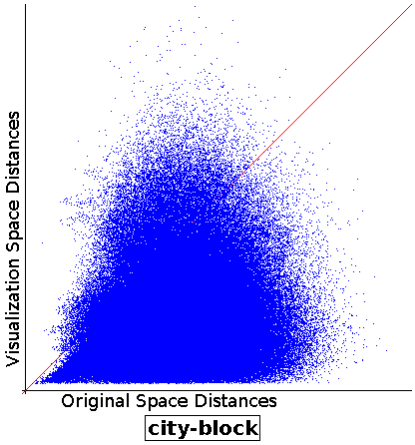

Figure 4.60: Stress Curve

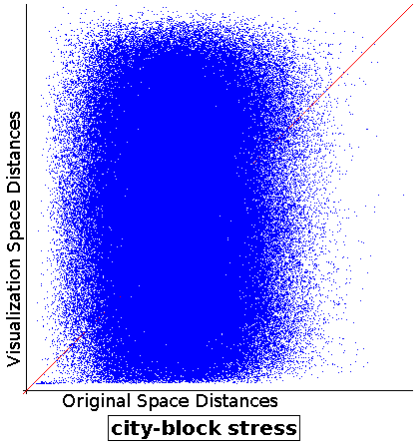

Figure 4.62: Stress Curve

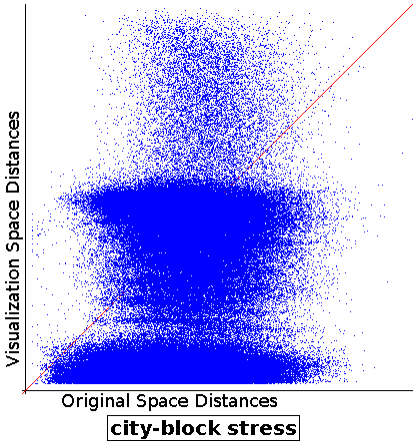

Figure 4.64: Stress Curve

4.5 Projections of HDR

4.5.1 LSP

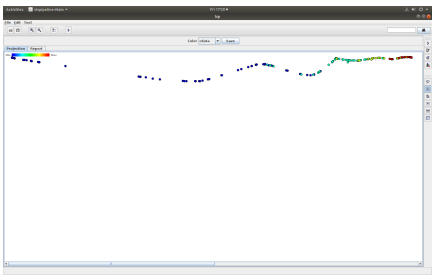

Figure 4.65: First projection

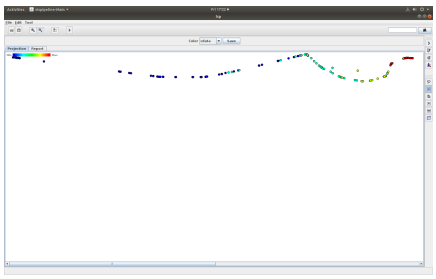

Figure 4.67: Second projection

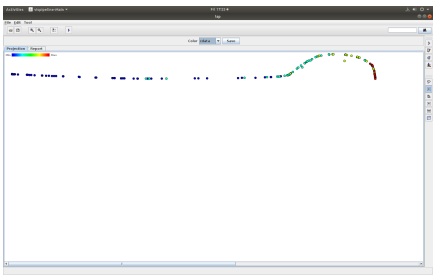

Figure 4.69: Third projection

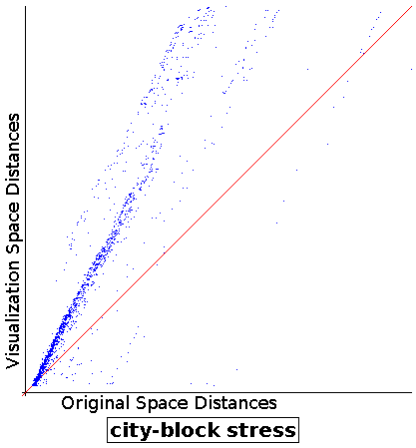

Figure 4.66: Stress Curve

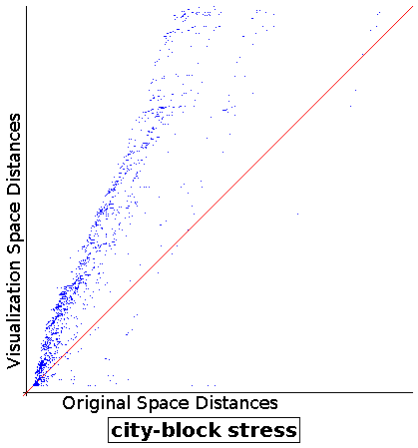

Figure 4.68: Stress Curve

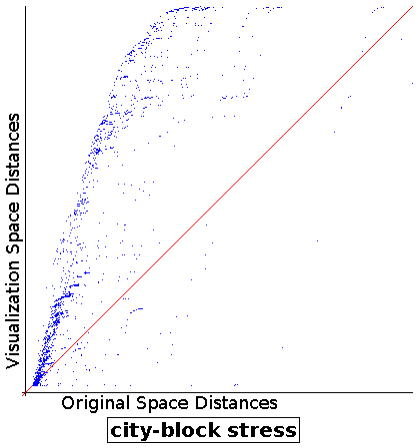

Figure 4.70: Stress Curve

4.5.2 t-SNE

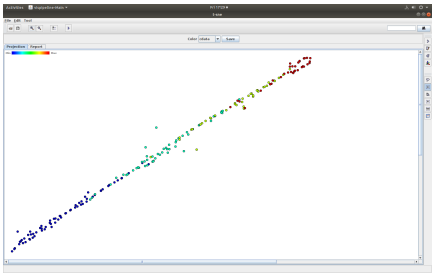

Figure 4.71: First projection

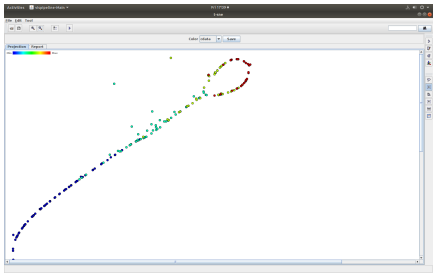

Figure 4.73: Second projection

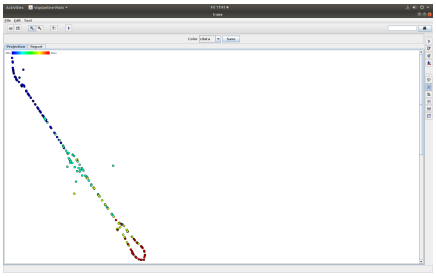

Figure 4.75: Third projection

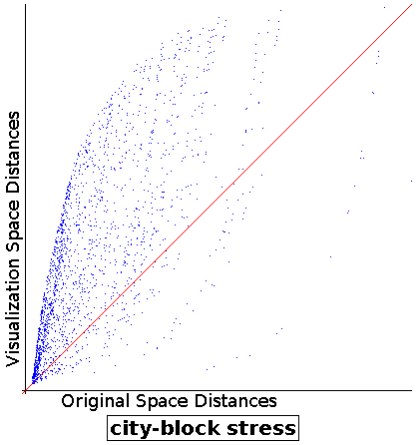

Figure 4.72: Stress Curve

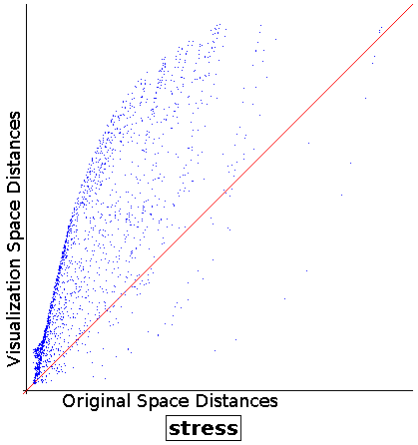

Figure 4.74: Stress Curve

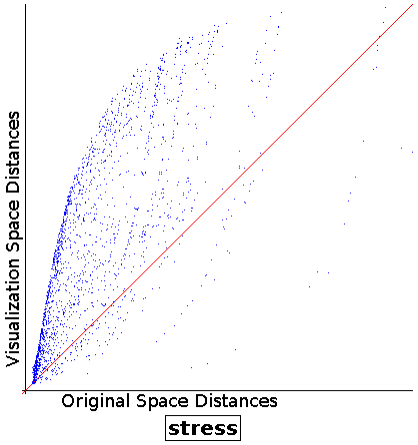

Figure 4.76: Stress Curve

4.5.3 PCA

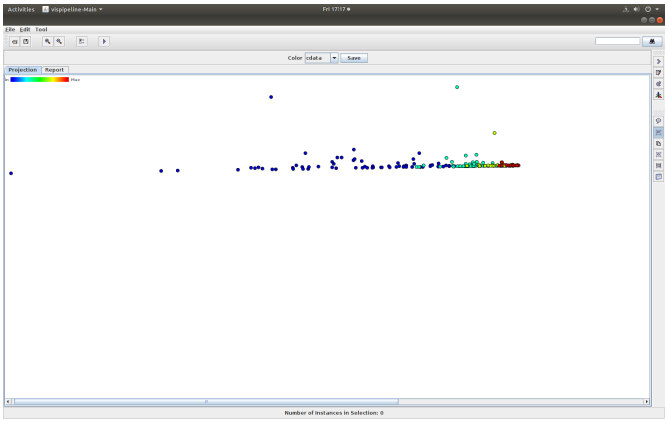

Figure 4.77: First projection

Figure 4.78: Stress Curve

5 Conclusions

Conclusions from overall assignment, emphasis on data sets, exercises and using visualisations.

6 References

HDR 2020 dataset, Human Development Data Center, https://hdr.undp.org/en/2020-report