University College Cork

CS3205 Lab Report 1

HDI Trends and Multidimensional Projections

Jack O'Connor

February 26, 2022

Contents

1 Introduction **1**
- 1.1 List of Acronyms . . . 1
- 1.2 The Report . . . 1
- 1.3 The Role of Visualisation . . . 1

2 Data Description **2**
- 2.1 Task 1 Datasets . . . 2
- 2.2 Task 2 Datasets . . . 2

3 Task 1 **3**
- 3.1 Free Exploration . . . 3
  - 3.1.1 HDI Histogram . . . 3
  - 3.1.2 HDI Global Heatmap . . . 3
  - 3.1.3 Average HDI Trend . . . 3
  - 3.1.4 Min/Max HDI . . . 3
  - 3.1.5 Male/Female Mean Years Education . . . 4
- 3.2 Specific Observations . . . 4
  - 3.2.1 Spurious Values . . . 4
  - 3.2.2 Correlated Attributes . . . 4
  - 3.2.3 Uncorrelated Attributes . . . 5
  - 3.2.4 Hypotheses for Previous Patterns . . . 5
  - 3.2.5 Usefulness of All Visualisations . . . 5
  - 3.2.6 Alternative Visualisations . . . 5

4 Task 2 **7**
- 4.1 Corel Projections Comparison . . . 7
  - 4.1.1 LSP . . . 7
  - 4.1.2 t-SNE . . . 8
  - 4.1.3 PCA . . . 9
- 4.2 CBR Projections Comparison . . . 10
  - 4.2.1 LSP . . . 10
  - 4.2.2 t-SNE . . . 11
  - 4.2.3 ProjClus . . . 12
- 4.3 Medical Images Projections Comparison . . . 13
  - 4.3.1 LSP . . . 13
  - 4.3.2 t-SNE . . . 14
  - 4.3.3 PCA . . . 15
- 4.4 Headlines Projections Comparison . . . 16
  - 4.4.1 LSP . . . 16
  - 4.4.2 t-SNE . . . 17
  - 4.4.3 ProjClus . . . 18
- 4.5 Projections of HDR . . . 19
  - 4.5.1 LSP . . . 19
  - 4.5.2 t-SNE . . . 20
  - 4.5.3 PCA . . . 21

5 Conclusions **21**

6 References **21**

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

All **datasets** used are linked here.
All **scripts** used are linked here. (Scripts use relative filepaths based on project structure found in the GitHub repo.)

2.1 Task 1 Datasets

Two source datasets were used to complete task 1, on which data transformations were performed:

1. **CS3205_2020_statistical_annex_all.xlsx** is an Excel spreadsheet containing survey data collected by the Human Development Centre from their HDR 2020 report.

2. **hdi.csv**, a csv file containing the Human Development Index (HDI) score assigned to each country in the world from the years 1990 to 2019 provided to all students taking CS3205 on Canvas.

CS3205_2020_statistical_annex_all.xlsx contains all of the original sheets supplied in the HDR, unmodified. When selecting which attributes to include in the task 1 visualisations I found it most expedient to simply copy and paste the columns from the existing sheets to a new sheet called ReformattedData. This was possible due to the relatively small scall of the number of attributes and records in the dataset.

I then was able to manually add the HDI Label column by selecting groups of cells between each of the highlighted HDI Level rows which were in the original sheets (possible since countries are ordered by HDI rank). Finally I removed the highlighted HDI Level rows since they were only for the benefit of human observers and do not actually contain any attribute information.

hdi.csv was used to create two additional datasets. **hdi_with_levels.csv** is identical to hdi.csv except for each YEAR_HDI attribute of the dataset a corresponding YEAR_HDI_LEVEL attribute following the same scheme as in the HDR is added using the script **add_hdi_levels.R**.

hdi_pivoted.csv was created using **pivot_hdi_levels.R** and it performs a pivot transformation such that instead of having a YEAR column for each year in the period, a YEAR column and HDI column pair is used. This drastically reduces the number of repetitive attributes in the dataset at the expense of introducing repeat values for COUNTRY_NAME AND COUNTRY_CODE to make up for the increased number of rows in the dataset. This transformation was especially useful for aggregating data by year and country, since it cut down on the number of attributes which had to be included in any visualisation. A HDI_Label column was also added to this dataset.

2.2 Task 2 Datasets

A total of five primary multidimensional data sets and two auxiliary stopwords datasets for CBR and Headlines textual datasets were used to create projections in VisPipeline. **hdr.data** was the only dataset which was not readily provided on Canvas.

1. **Corel** is a collection of images with 10 distinct classes: African tribes, beaches, buildings, buses, dinosaurs, elephants, flowers, food, horses, mountains.

2. **CBR** is a collection of paper abstracts and references spanning four main topics: case-based reasoning (CBR), inductive logic programming (ILP), information retrieval (IR), sonification (SON) and six intruders.

3. **Medical** is a collection of X-Rays of the human body mainly of the skull and spine from different angles.

4. **Headlines** is a collection of articles from AP, BBC, Reuters and CNN which were all published during the same time period in the mid 2000's.

5. **HDR** contains all of the same attributes as were included in ReformattedData but excludes HDI and HDI_Rank.

To create the hdr.data file it was enough to create a new spreadsheet in Excel, remove the requisite columns, move Country column to the beginning and Label column to the end, export the new sheet to csv using a semi-colon as the delimiter and finally remove the Country and Label headers in the exported cvs while adding the necessary .data read-in file parameters to the top of the file.

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

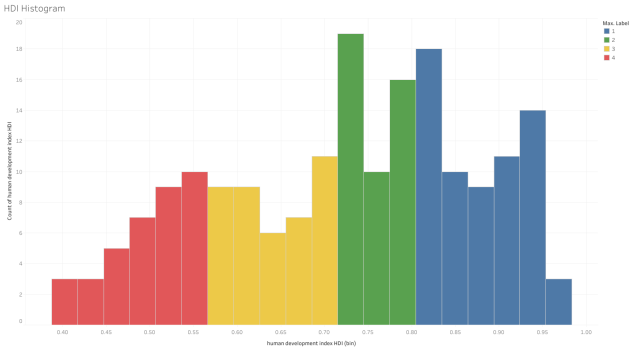

Figure 3.1: HDI Distribution with HDI Level as Colour

3.1.2 HDI Global Heatmap

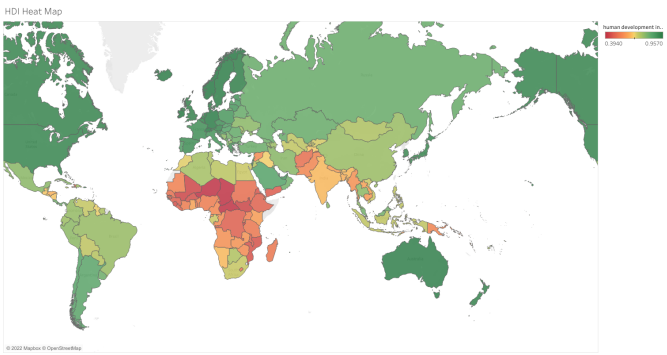

Figure 3.2: Darker greens imply higher HDI, darker reds lower HDI

3.1.3 Average HDI Trend

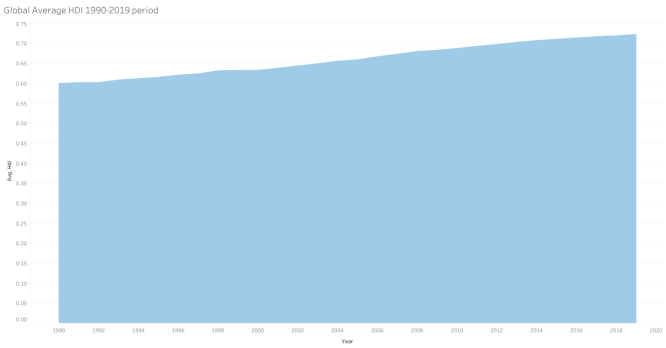

Figure 3.3: Global average HDI strictly rises over time

3.1.4 Min/Max HDI

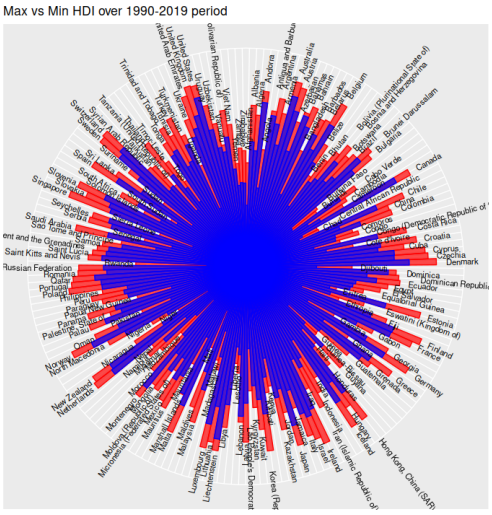

Figure 3.4: Max HDI over period in red, Min HDI over period in blue

3.1.5 Male/Female Mean Years Education

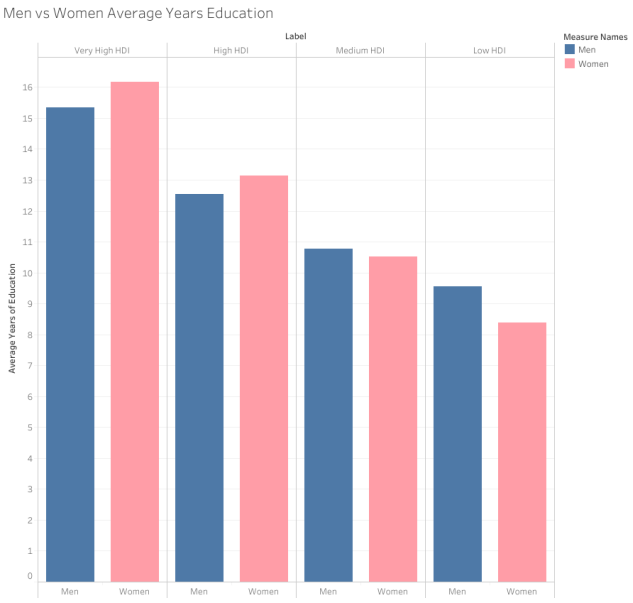

Figure 3.5: Male mean years education represented by blue, female pink

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: Distribution of future to current population ratio by country, colour represents HDI Level

Mean vs Expected Years Schooling

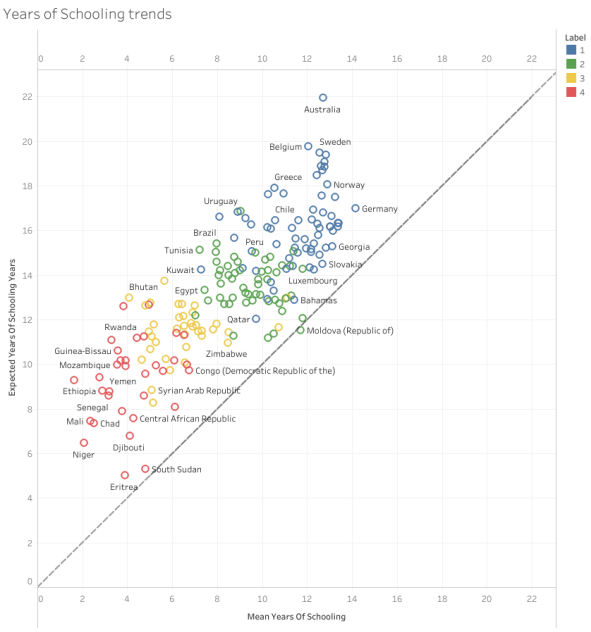

Figure 3.7: Amount of education is expected to rise, colour represents HDI Level

3.2.2 Correlated Attributes

Fertility Rate vs Adolescent Birth Rate

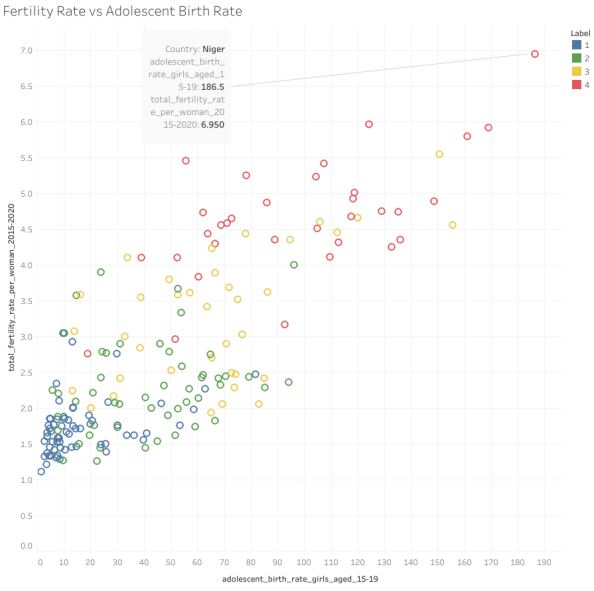

Figure 3.8: Fertility rate highly correlates with adolescent birth rate, colour represents HDI Level

HDI vs Gross National Income (GNI)

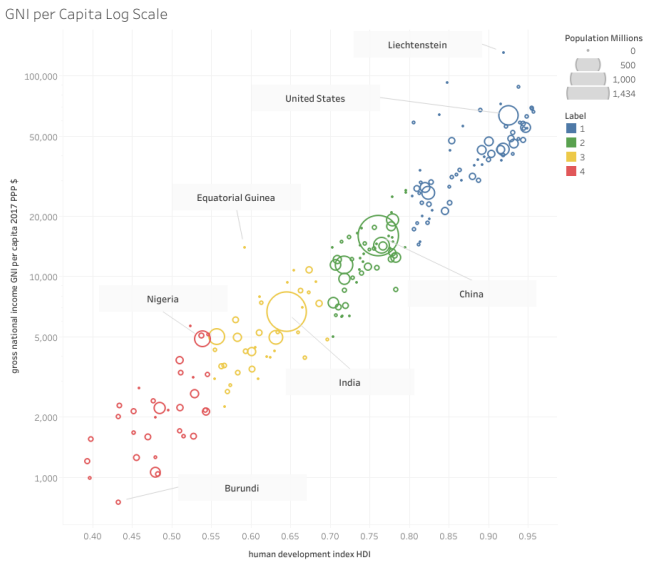

Figure 3.9: GNI accounts almost totally for HDI, colour represents HDI Level, size represents total population

3.2.3 Uncorrelated Attributes

Women Share of Seats in Government vs Average Years Education

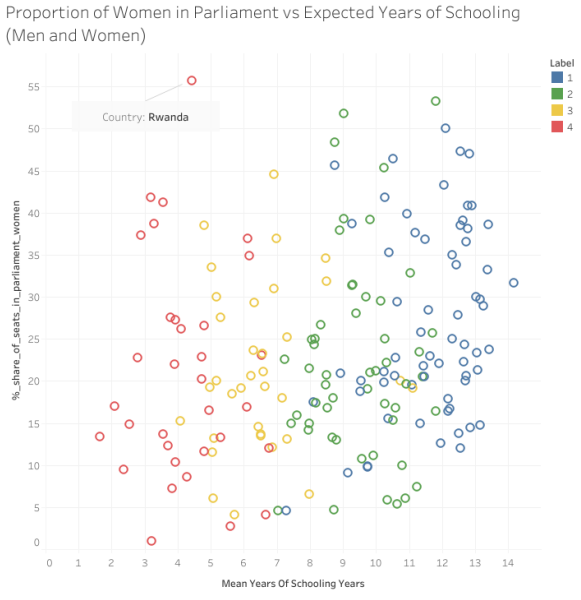

Figure 3.10: Average years of schooling plotted against percentage of women in government, colour represents HDI Level

Labour Force Participation vs GDP

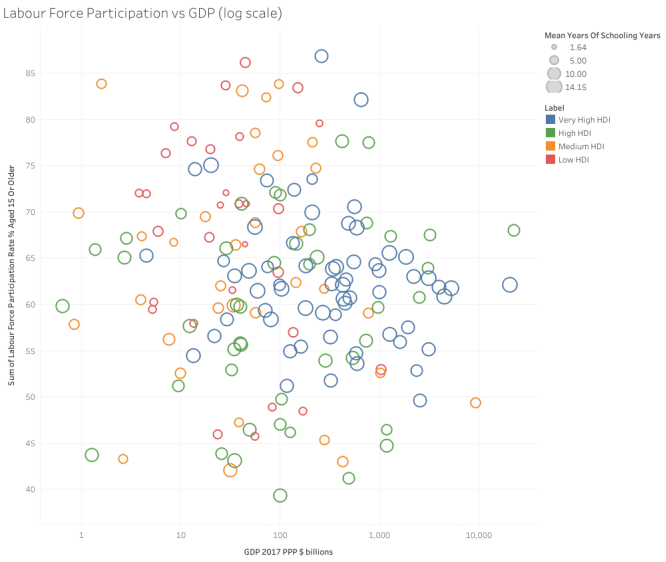

Figure 3.11: Labour force participation plotted against GDP, colour represents HDI Level, size represents mean years schooling

3.2.4 Hypotheses for Previous Patterns

Health Expenditure ordered by Life Expectancy

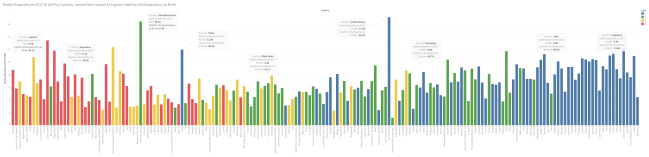

Figure 3.12: Health expenditure as a fraction of GDP, from left to right in order of increasing average life expectancy

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

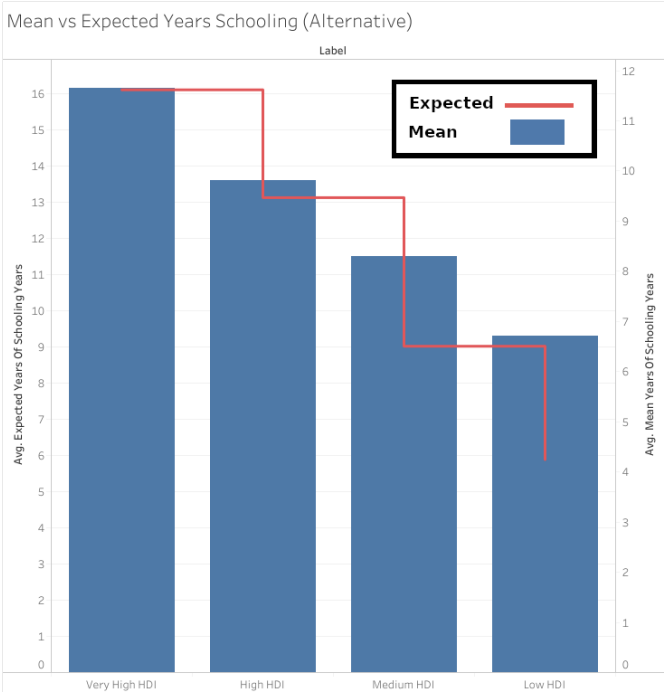

Figure 3.14: Lose individual countries

4 Task 2

Results for task 2.

4.1 Corel Projections Comparison

4.1.1 LSP

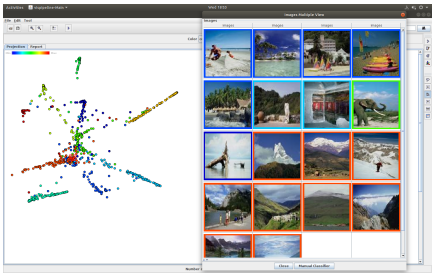

Figure 4.1: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

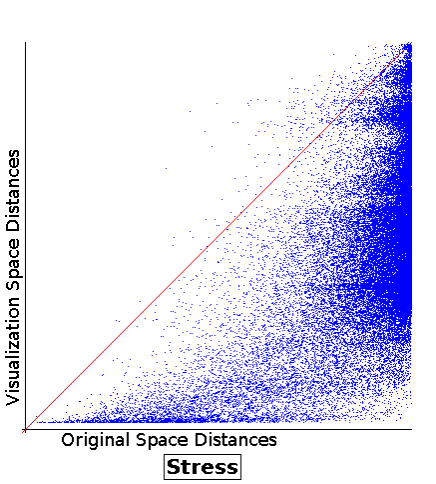

Figure 4.2: Cosine Stress Curve

Silhouette Coefficient: 0.5155

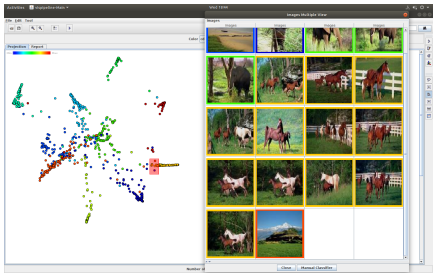

Figure 4.3: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 12, No: Neighbours: 10, Dissimilarity: Cosine

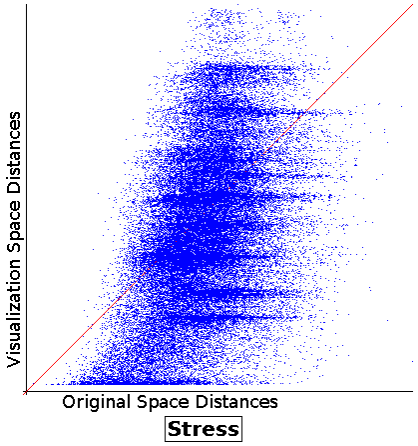

Figure 4.4: Euclidean Stress Curve

Silhouette Coefficient: 0.4635

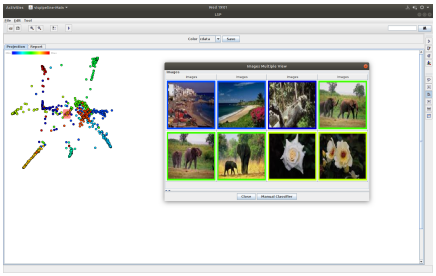

Figure 4.5: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 15, Dissimilarity: Euclidean

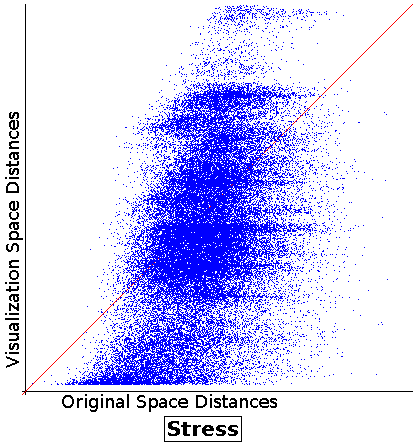

Figure 4.6: Euclidean Stress Curve

Silhouette Coefficient: 0.4762

4.1.2 t-SNE

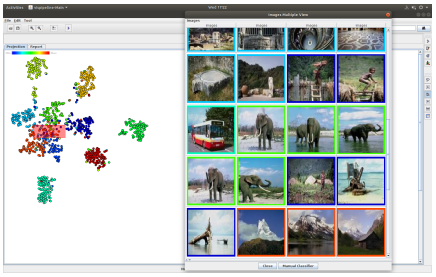

Figure 4.7: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

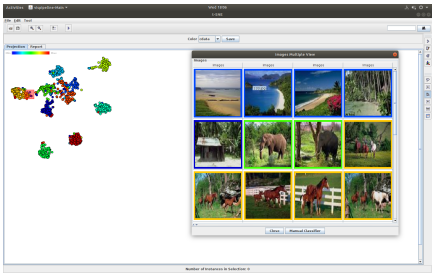

Figure 4.9: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Cosine

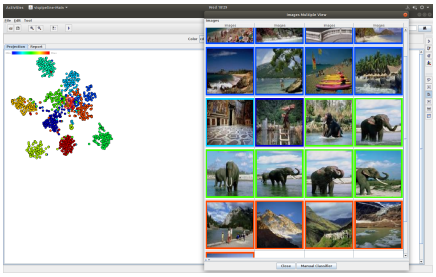

Figure 4.11: Initial Dimensions: 15, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1000, Dissimilarity: Euclidean

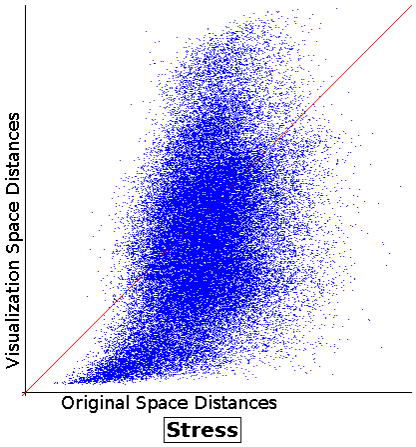

Figure 4.8: Euclidean Stress Curve

Silhouette Coefficient: 0.4693

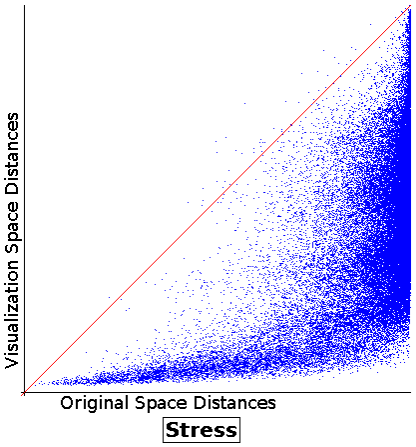

Figure 4.10: Cosine Stress Curve

Silhouette Coefficient: 0.4961

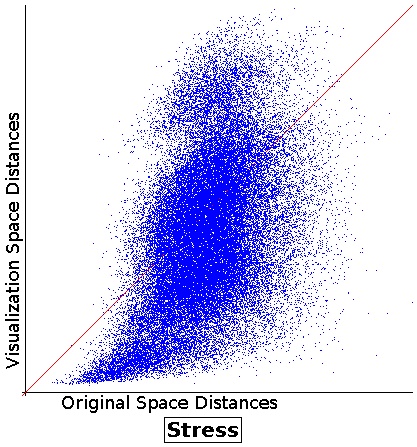

Figure 4.12: Euclidean Stress Curve

Silhouette Coefficient: 0.4484

4.1.3 PCA

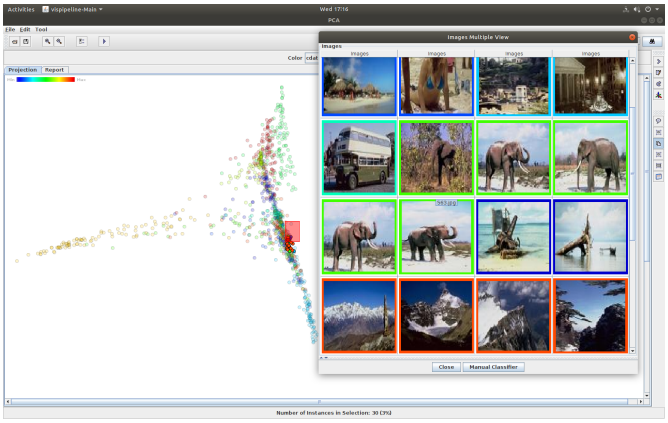

Figure 4.13: No parameters

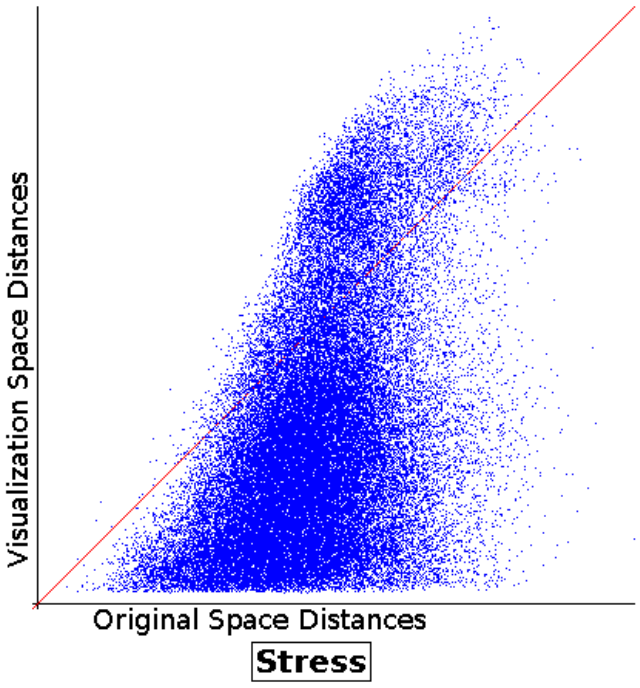

Figure 4.14: Stress Curve

Silhouette Coefficient: 0.5201

4.2 CBR Projections Comparison

4.2.1 LSP

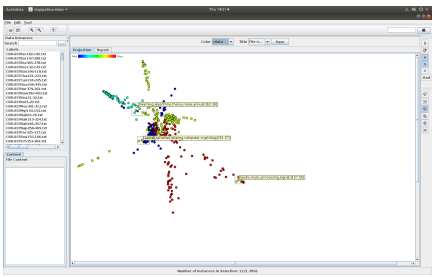

Figure 4.15: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

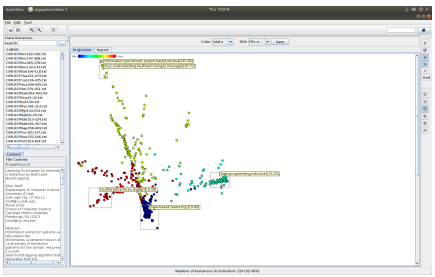

Figure 4.17: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Cosine

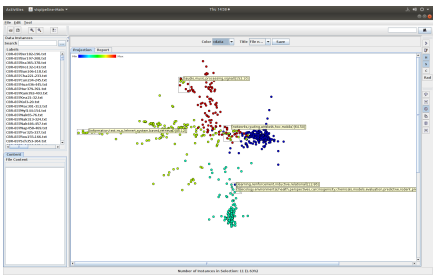

Figure 4.19: No. Iterations: 80, Fraction of Delta: 8.0, No. Control Points: 20, No: Neighbours: 12, Dissimilarity: Cosine

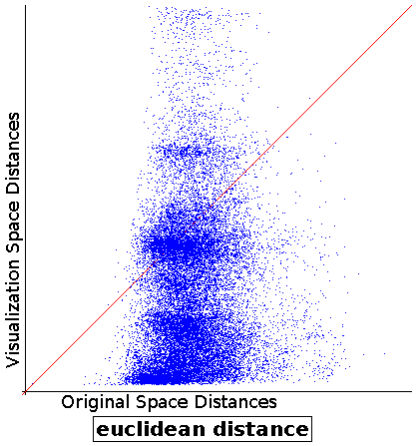

Figure 4.16: Stress Curve

Silhouette Coefficient: 0.3550

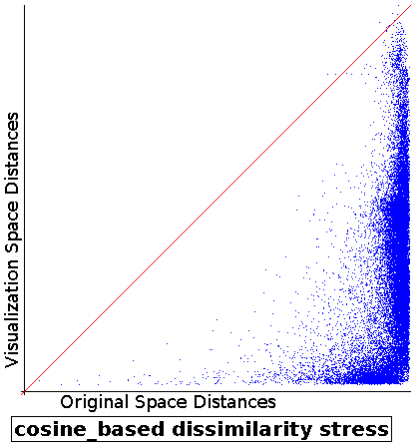

Figure 4.18: Stress Curve

Silhouette Coefficient: 0.5100

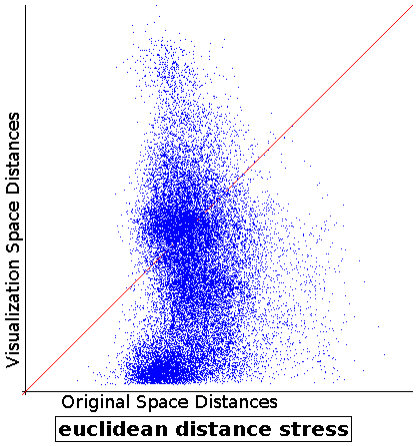

Figure 4.20: Stress Curve

Silhouette Coefficient: 0.5116

4.2.2 t-SNE

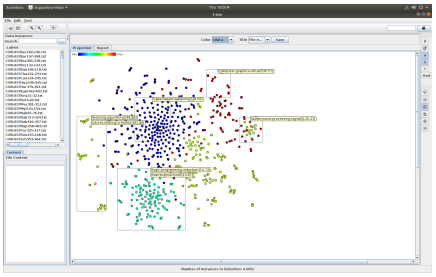

Figure 4.21: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

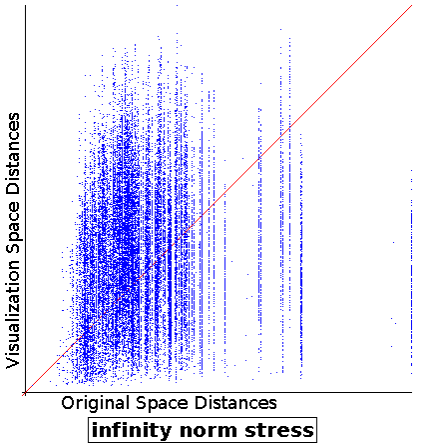

Figure 4.22: Stress Curve

Silhouette Coefficient: 0.2731

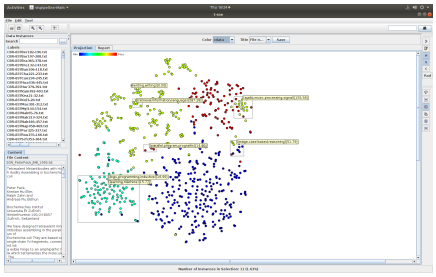

Figure 4.23: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1500, Dissimilarity: Cosine

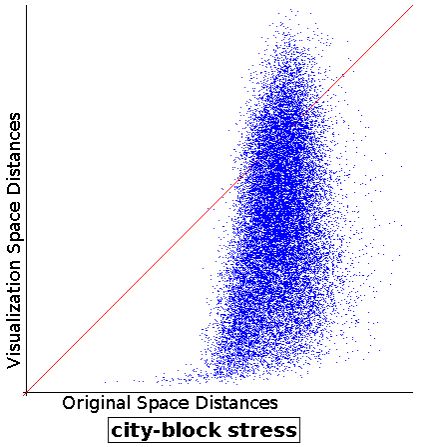

Figure 4.24: Stress Curve

Silhouette Coefficient: 0.3481

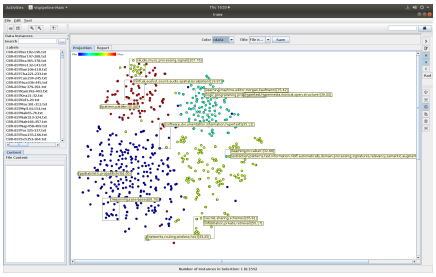

Figure 4.25: Initial Dimensions: 500, Target Dimension: 2, Perplexity: 100, Max No. Iterations: 2000, Dissimilarity: Cosine

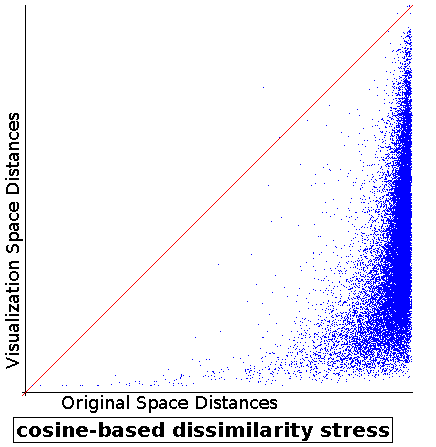

Figure 4.26: Stress Curve

Silhouette Coefficient: 0.2612

4.2.3 ProjClus

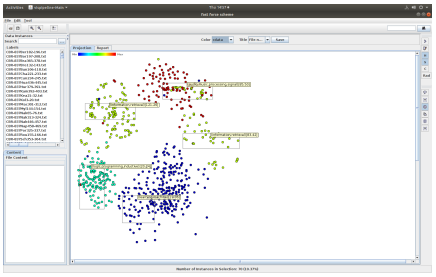

Figure 4.27: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 4.5, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

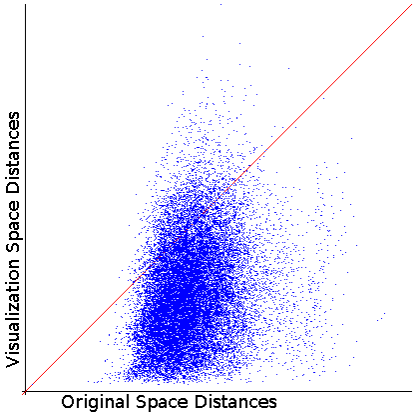

Figure 4.28: Stress Curve

Silhouette Coefficient: 0.3777

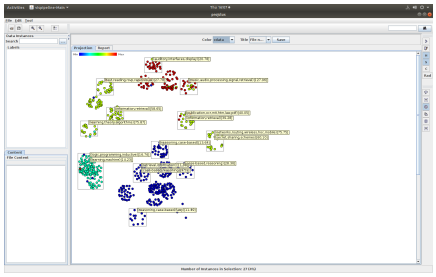

Figure 4.29: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 9.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

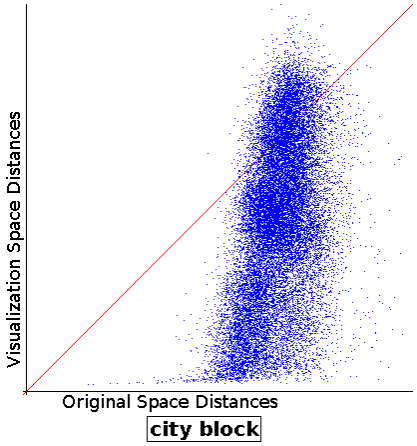

Figure 4.30: Stress Curve

Silhouette Coefficient: 0.4225

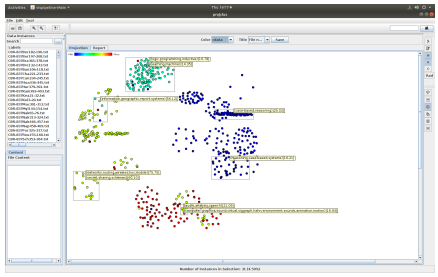

Figure 4.31: No. Iterations: 200, Fraction of Delta: 15.0, Cluster Factor: 7.0, Type of Projection: Fastmap, Dissimilarity: Cosine

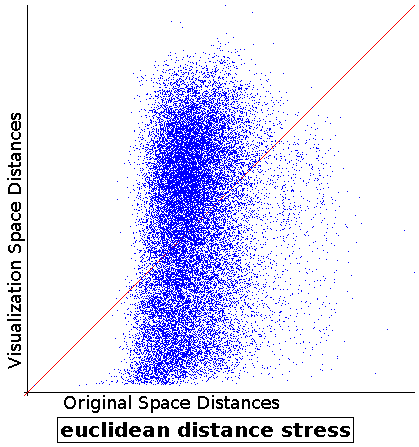

Figure 4.32: Stress Curve

Silhouette Coefficient: 0.4637

4.3 Medical Images Projections Comparison

4.3.1 LSP

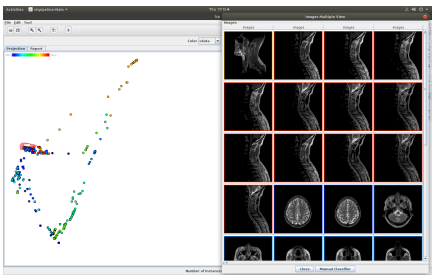

Figure 4.33: No. Iterations: 100, Fraction of Delta: 4.0, No. Control Points: 12, No: Neighbours: 80, Dissimilarity: Cosine

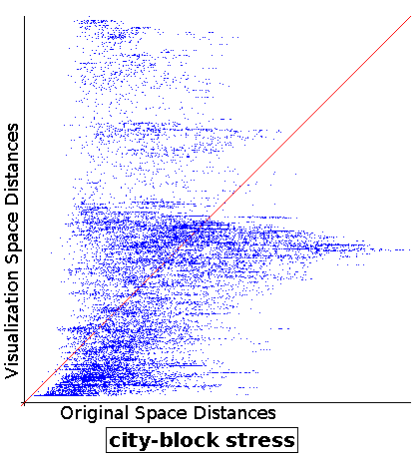

Figure 4.34: Stress Curve

Silhouette Coefficient: 0.0375

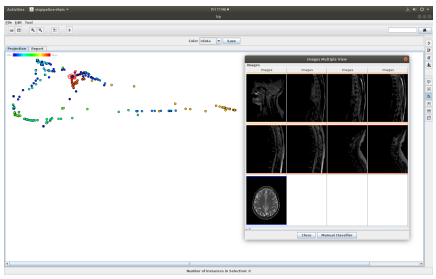

Figure 4.35: No. Iterations: 100, Fraction of Delta: 8.0, No. Control Points: 12, No: Neighbours: 6, Dissimilarity: Cosine

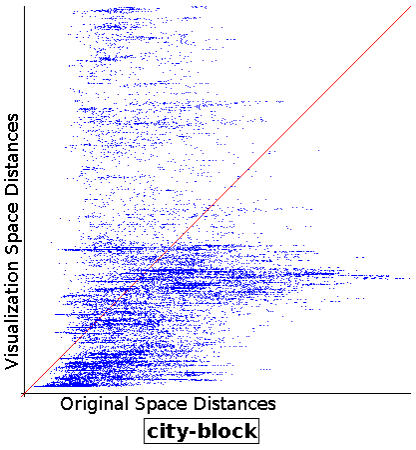

Figure 4.36: Stress Curve

Silhouette Coefficient: -0.01736

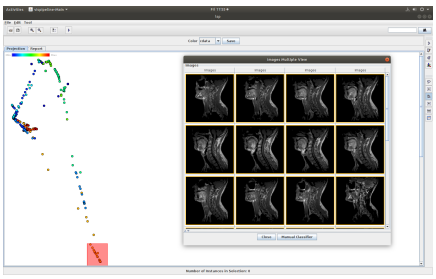

Figure 4.37: No. Iterations: 100, Fraction of Delta: 2.0, No. Control Points: 14, No: Neighbours: 6, Dissimilarity: Cosine

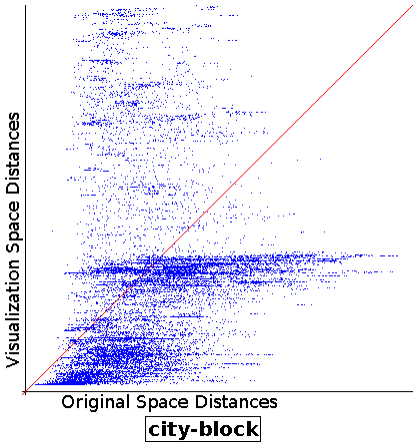

Figure 4.38: Stress Curve

Silhouette Coefficient: -0.0921

4.3.2 t-SNE

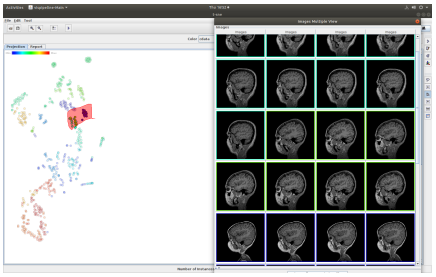

Figure 4.39: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

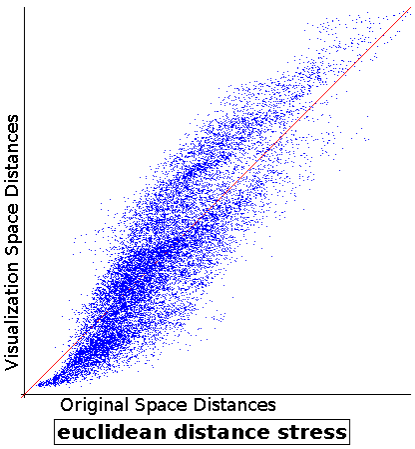

Figure 4.40: Stress Curve

Silhouette Coefficient: 0.1451

Figure 4.41: Initial Dimensions: 4, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1500, Dissimilarity: Euclidean

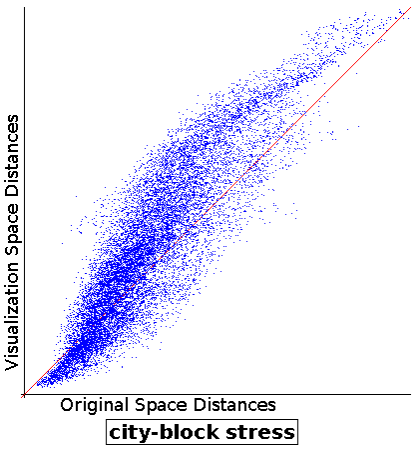

Figure 4.42: Stress Curve

Silhouette Coefficient: 0.1089

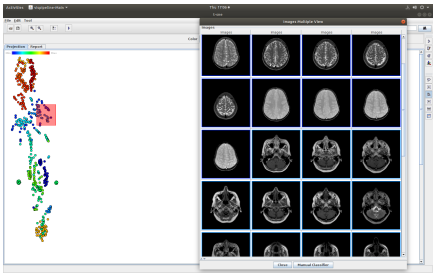

Figure 4.43: Initial Dimensions: 256, Target Dimension: 2, Perplexity: 50, Max No. Iterations: 1500, Dissimilarity: Cosine

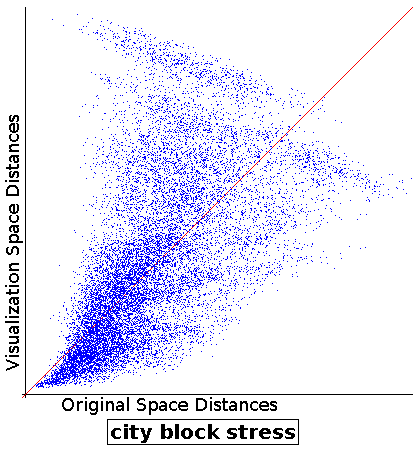

Figure 4.44: Stress Curve

Silhouette Coefficient: 0.0301

4.3.3 PCA

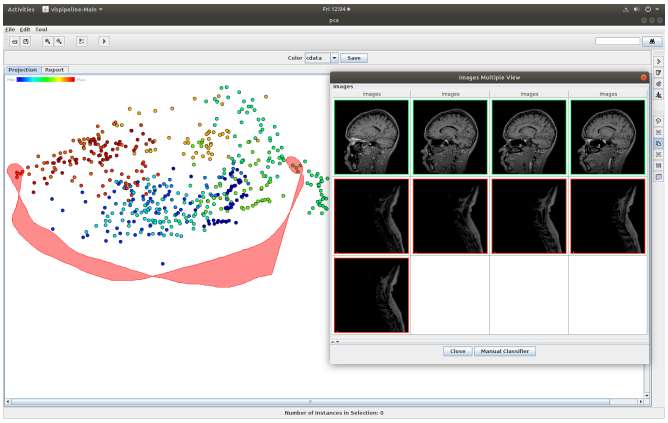

Figure 4.45: No parameters

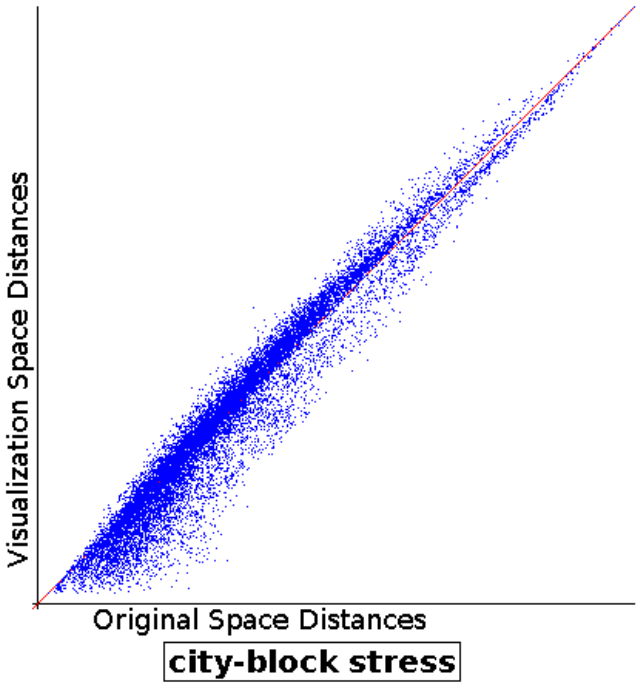

Figure 4.46: Stress Curve

Silhouette Coefficient: 0.3501

4.4 Headlines Projections Comparison

4.4.1 LSP

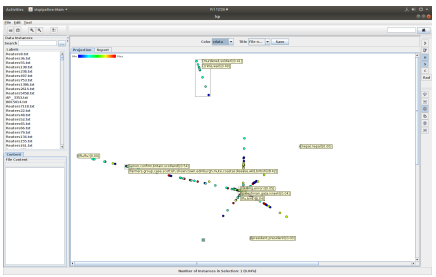

Figure 4.47: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 10, No: Neighbours: 10, Dissimilarity: Euclidean

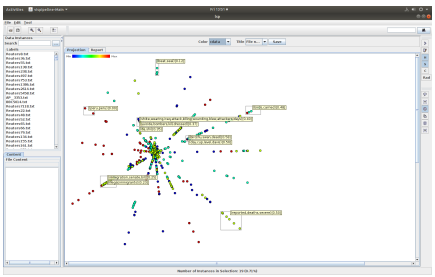

Figure 4.49: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 20, No: Neighbours: 10, Dissimilarity: Cosine

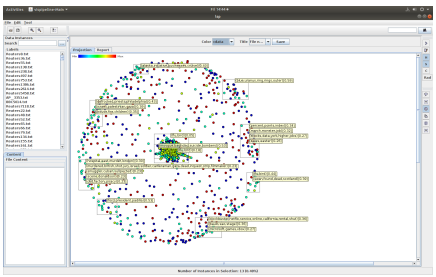

Figure 4.51: No. Iterations: 100, Fraction of Delta: 35.0, No. Control Points: 600, No: Neighbours: 200, Dissimilarity: Cosine

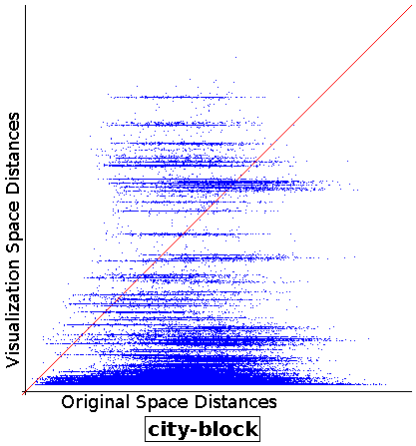

Figure 4.48: Stress Curve

Silhouette Coefficient: -0.1885

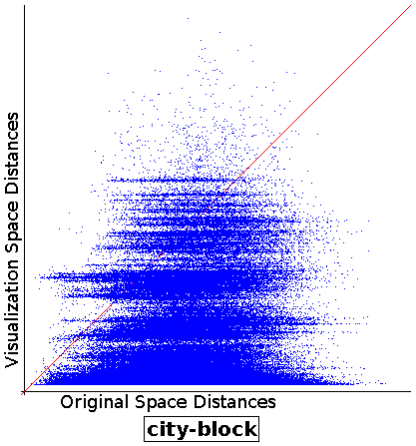

Figure 4.50: Stress Curve

Silhouette Coefficient: -0.0756

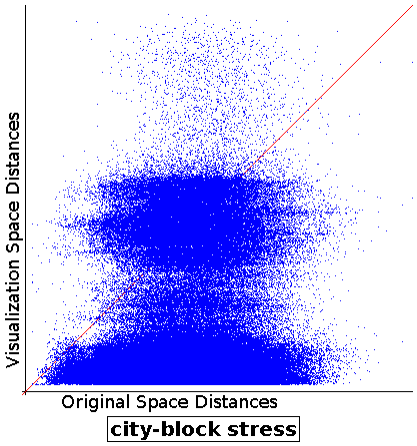

Figure 4.52: Stress Curve

Silhouette Coefficient: -0.1269

4.4.2 t-SNE

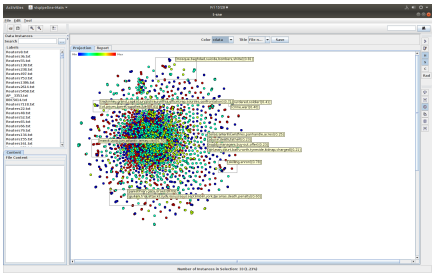

Figure 4.53: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

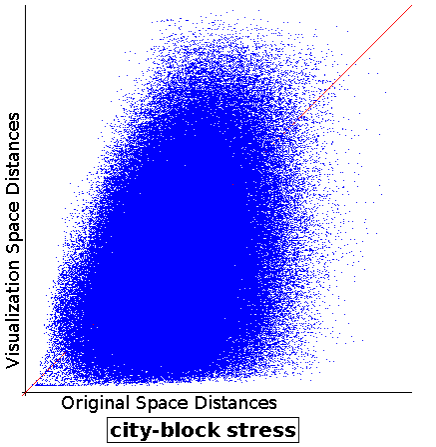

Figure 4.54: Stress Curve

Silhouette Coefficient: -0.0714

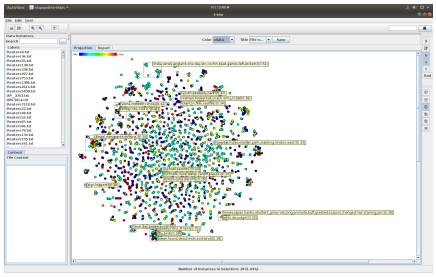

Figure 4.55: Initial Dimensions: 60, Target Dimension: 2, Perplexity: 40, Max No. Iterations: 5000, Dissimilarity: Cosine

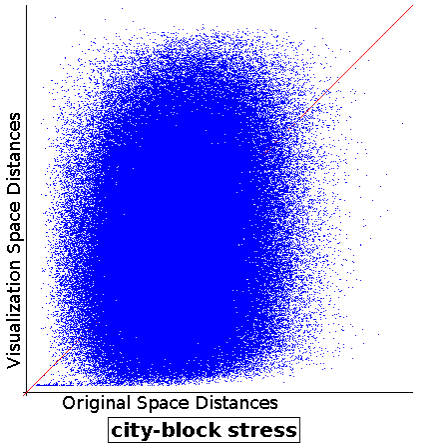

Figure 4.56: Stress Curve

Silhouette Coefficient: -0.0275

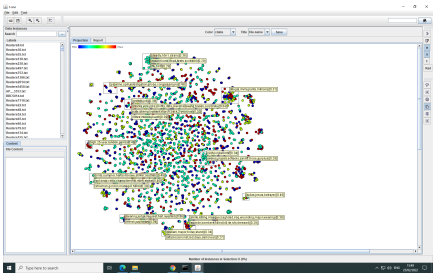

Figure 4.57: Initial Dimensions: 240, Target Dimension: 2, Perplexity: 80, Max No. Iterations: 750, Dissimilarity: Cosine

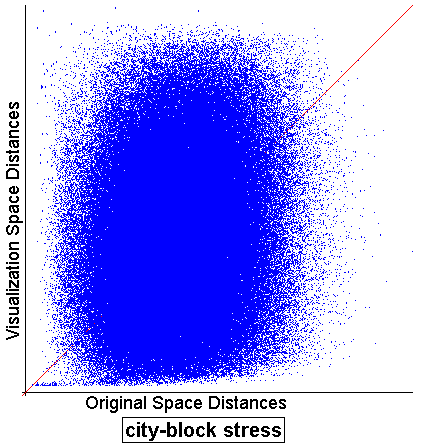

Figure 4.58: Stress Curve

Silhouette Coefficient: -0.0197

4.4.3 ProjClus

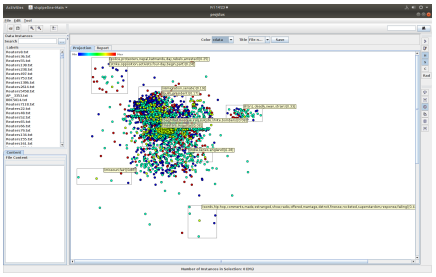

Figure 4.59: No. Iterations: 50, Fraction of Delta: 8.0, Cluster Factor: 4.5, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

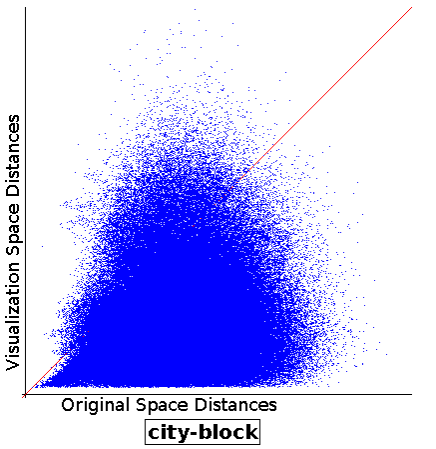

Figure 4.60: Stress Curve

Silhouette Coefficient: -0.1216

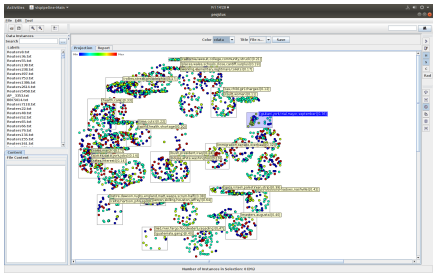

Figure 4.61: No. Iterations: 100, Fraction of Delta: 8.0, Cluster Factor: 8.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

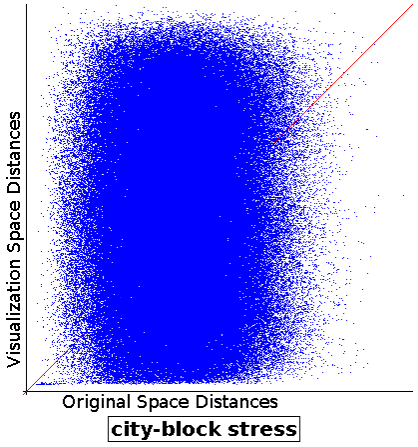

Figure 4.62: Stress Curve

Silhouette Coefficient: -0.04515

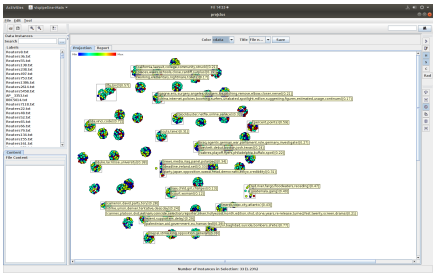

Figure 4.63: No. Iterations: 75, Fraction of Delta: 40.0, Cluster Factor: 20.0, Type of Projection: Nearest Neighbor Projection, Dissimilarity: Cosine

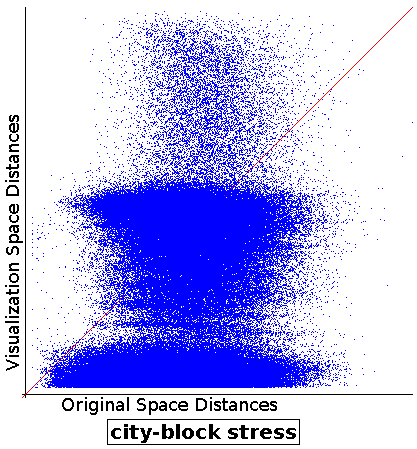

Figure 4.64: Stress Curve

Silhouette Coefficient: -0.0195

4.5 Projections of HDR

4.5.1 LSP

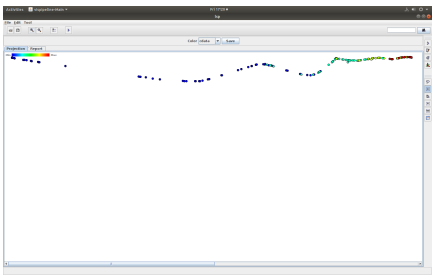

Figure 4.65: No. Iterations: 50, Fraction of Delta: 8.0, No. Control Points: 18, No: Neighbours: 10, Dissimilarity: Euclidean

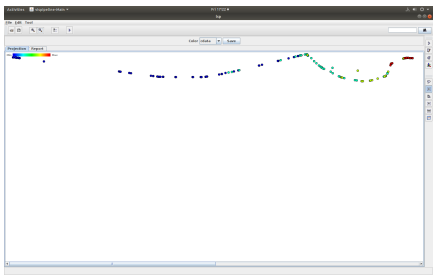

Figure 4.67: No. Iterations: 50, Fraction of Delta: 4.0, No. Control Points: 6, No: Neighbours: 8, Dissimilarity: Euclidean

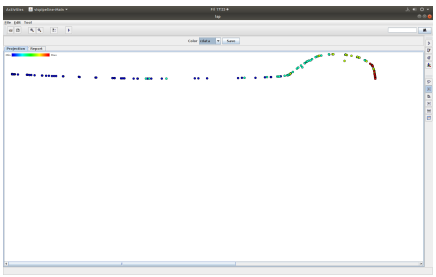

Figure 4.69: No. Iterations: 75, Fraction of Delta: 6.0, No. Control Points: 4, No: Neighbours: 15, Dissimilarity: Euclidean

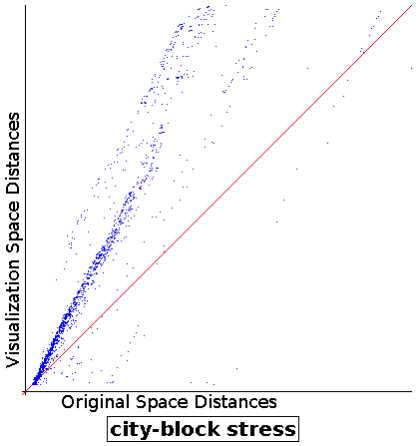

Figure 4.66: Stress Curve

Silhouette Coefficient: 0.2117

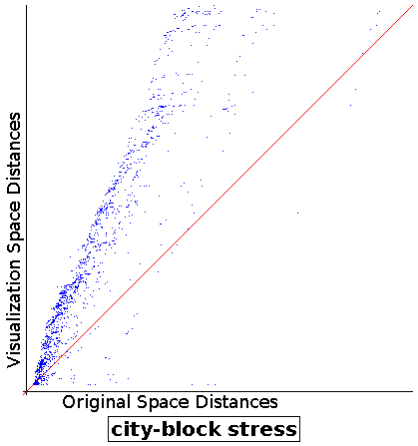

Figure 4.68: Stress Curve

Silhouette Coefficient: 0.2338

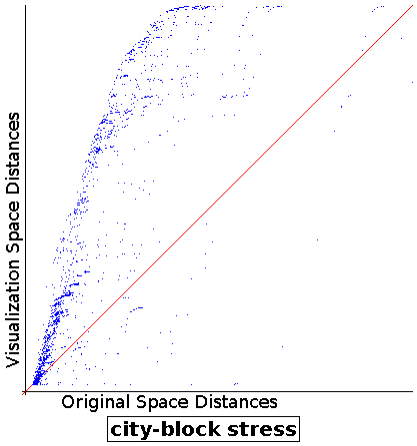

Figure 4.70: Stress Curve

Silhouette Coefficient: 0.2779

4.5.2 t-SNE

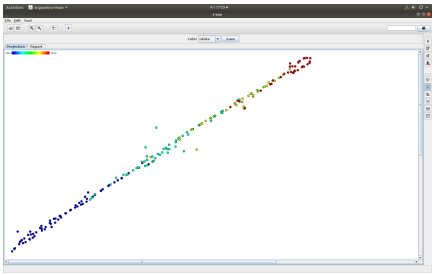

Figure 4.71: Initial Dimensions: 30, Target Dimension: 2, Perplexity: 30, Max No. Iterations: 1000, Dissimilarity: Euclidean

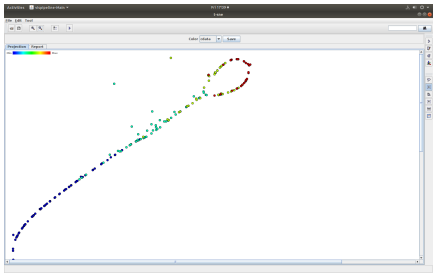

Figure 4.73: Initial Dimensions: 24, Target Dimension: 2, Perplexity: 100, Max No. Iterations: 1000, Dissimilarity: Euclidean

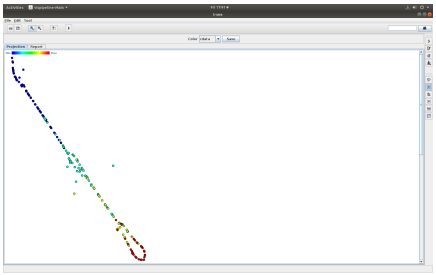

Figure 4.75: Initial Dimensions: 6, Target Dimension: 2, Perplexity: 60, Max No. Iterations: 1000, Dissimilarity: Infinity Norm

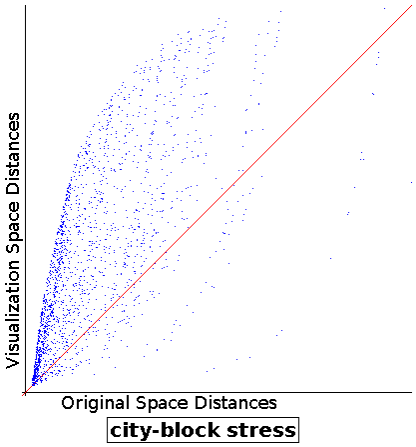

Figure 4.72: Stress Curve

Silhouette Coefficient: 0.3648

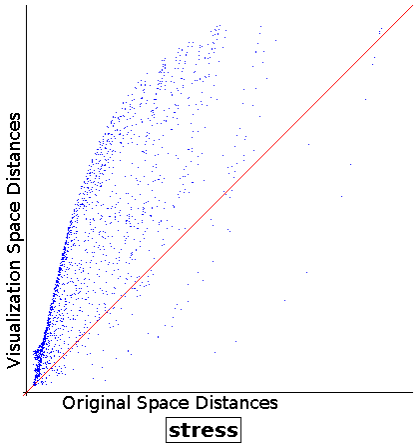

Figure 4.74: Stress Curve

Silhouette Coefficient: 0.3267

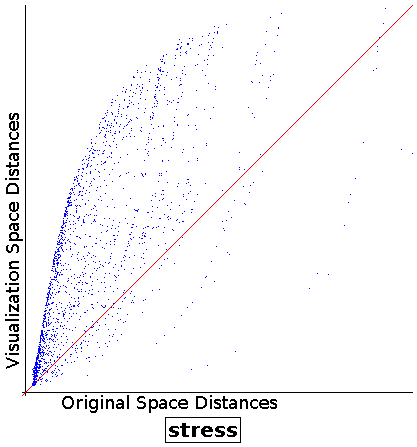

Figure 4.76: Stress Curve

Silhouette Coefficient: 0.3427

4.5.3 PCA

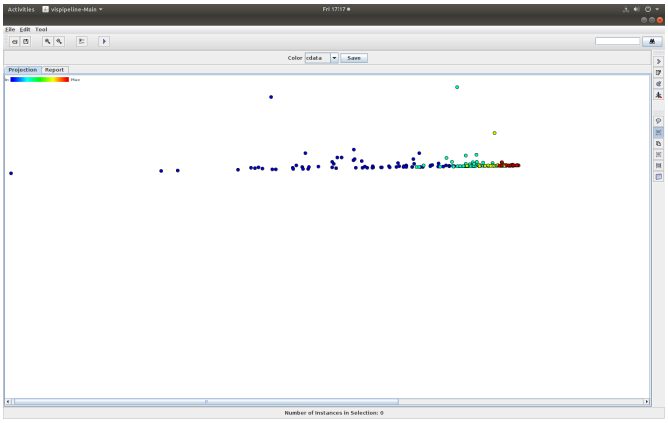

Figure 4.77: No parameters

Figure 4.78: City-Block Stress Curve

Silhouette Coefficient: 0.1985

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