Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates

David Borland[1], Vivian L. West[2], W. Ed Hammond[2]
[1]RENCI, The University of North Carolina at Chapel Hill, Chapel Hill, NC;
[2]Duke Center for Health Informatics, Duke University, Durham, NC

Abstract

We present radial coordinates, a multivariate visualization technique based on parallel coordinates. The visualization contains a number of features driven by the needs of health-related data analysis, such as integrating categorical and numeric data, and comparing user-selected subpopulations via ribbon rendering. We illustrate the utility of radial coordinates by exploring primary care trust (PCT) and practice-level data from the United Kingdom's National Health Service, using three examples: lung cancer rates among PCTs, various cancer rates among only London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Introduction

With the ever-increasing size and number of health-related datasets, new analytical tools are becoming necessary to enable enhanced understanding of the vast amount of information contained within. Visualization leverages the power of the human visual system to reveal patterns and relationships in data by mapping the data to visually salient features.

One of the challenges for visualization of health-related data is the desire to incorporate data of many types (e.g. lab results, demographics, medications, vital signs, and genomic data) from various sources. We have developed a multivariate visualization technique, radial coordinates, that enables visual analysis of a wide range of health-related datasets and handles both numeric and categorical data (Figure 1). Radial coordinates facilitates the interactive exploration of datasets to reveal patterns in the data, discover relationships between variables, and compare user-defined subpopulations. In this manner we support the pursuit of hypothesis formations that can elicit further inquiry and lead to new knowledge.

An overview of an initial radial coordinates prototype applied to query data was given previously.[1] In this paper we provide a more in-depth description of the various features of a new implementation, which includes several new features, and discuss its application to primary care trust (PCT) and practice-level data from the National Health Service (NHS) in the United Kingdom (UK). We present three examples illustrating the use of radial coordinates to explore the NHS data: lung cancer rates among PCTs, a comparison of various cancer rates among London suburb PCTs, and medical problem prevalence among over 1500 London practices.

Previous Work

Our visualization is based largely on parallel coordinates, a multivariate visualization technique which represents each dimension as a parallel axis, and each data entity as a line connecting the entity's value at each axis.[2,3] Non-parallel arrangements of axes have also been investigated.[4] Our radial coordinates arrangement differs in that the radial layout maintains a square aspect ratio even with many axes, and enables utilization of the space in the center of the radial layout. Parallel coordinates have been combined with various other visualization techniques[5-7], including direct integration of scatter plots.[8,9] In our visualization we include a scatter plot based on the first two principal components to enhance the ability to find clusters in high-dimensional data in an intuitive manner (Figure 1a). Future work will explore combinations with other techniques. We also incorporate chords representing the correlations between axes in a manner similar to Circos.[10] Extensions to parallel coordinates for incorporating categorical data include parallel sets[11] and hammock plots.[12] Both represent multiple data points as paths between axes, with the number of data points encoded as path width. Our curve spreading technique incorporates categorical and continuous data while still enabling the visualization of individual data points (Figure 2). Various techniques have been developed to combine multiple data points to enhance the understanding of large datasets[13,14] and observe clusters via edge bundling techniques.[15,16] Our ribbon rendering technique enables enhanced visualization of user-selected data points, including overlaying information of statistical data (median value and quartile ranges) of interest to the health-care community (Figure 1b). Axis ordering is an important element of parallel coordinates visualizations, as it is typically easier to notice relationships between variables with adjacent axes.[17-19] We employ a correlation-based clustering technique and also introduce dynamic reordering of categorical axis values to cluster similar values based on user-defined selections (Figures 3c, 3d).

Methods

Data

PCTs, abolished in 2013 due to NHS reorganization, were regional administrative bodies in the UK responsible for commissioning health services from providers and providing community health services. Here we investigate 26 variables measuring various health and socioeconomic factors for 147 of the 152 PCTs in England (five were removed due to missing data). Health factors include cancer rates, drug prescription rates, and factors related to diabetes prevalence and treatment. Socioeconomic factors include socioeconomic deprivation, economic output, geographic region, and local region classification (e.g. *Manufacturing Towns* and *Coastal and Countryside*) from the Office for National Statistics (ONS).

We also demonstrate our visualization with data showing the prevalence of a number of medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). There were ten SHAs in England from 2006-2013.

Visualization

The radial coordinates visualization, implemented using the D3 JavaScript library[20], represents each variable in a multivariate dataset by an axis, with the axes arranged radially around a circle. Each individual entity is represented by a curve that connects the value of that entity at each axis. Figure 1 gives an example applied to PCT data, with four PCT curves highlighted in red by the user.

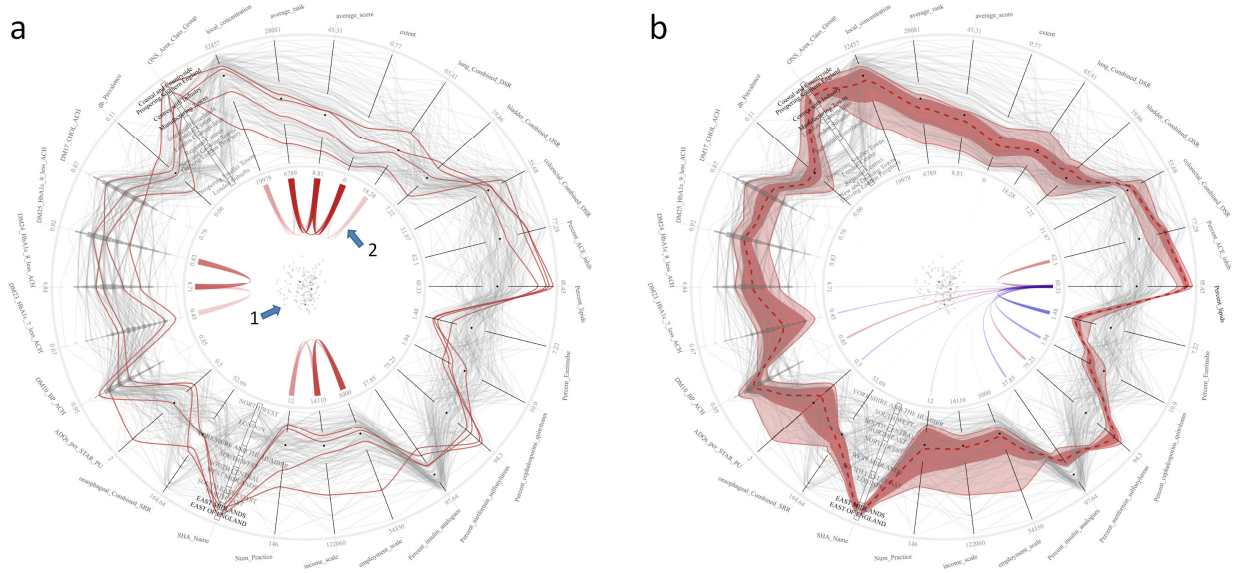

Figure 1. Radial coordinates visualizations of NHS PCT data. User-highlighted curves (red) enable the comparison of four PCTs across multiple variables (a). A linked scatterplot of the first two principal components can help show clusters in high-dimensions (a1). Chords connecting axes represent correlations (positive: red, negative: blue) above a user-defined threshold (a2). Ribbon rendering enables a simplified representation of user-defined subpopulations, displaying the data range optionally overlaid with median value and inner quartile ranges (b). Mouse over of an axis shows all correlations with that axis, regardless of user-defined threshold (b).

User selection of individual curves enables a visual comparison of how different entities relate across the various axes. A radial layout elegantly handles large numbers of axes while maintaining a square aspect ratio, also enabling the use of the center of the layout for supplemental visualizations, such as axis correlation chords and a scatterplot of the first two principal components (Figure 1a). Ribbon rendering uses a sliding window algorithm to draw the area between the innermost and outermost boundary of selected curves in a semi-transparent solid color, making it easier to see the spread of each subpopulation. An optional summary statistic overlay shows the inner quartile range and median value of each subpopulation (Figure 1b). Other visualization features include data-type dependent axis distribution visualizations and curve spreading for categorical and discrete data (Figure 2).

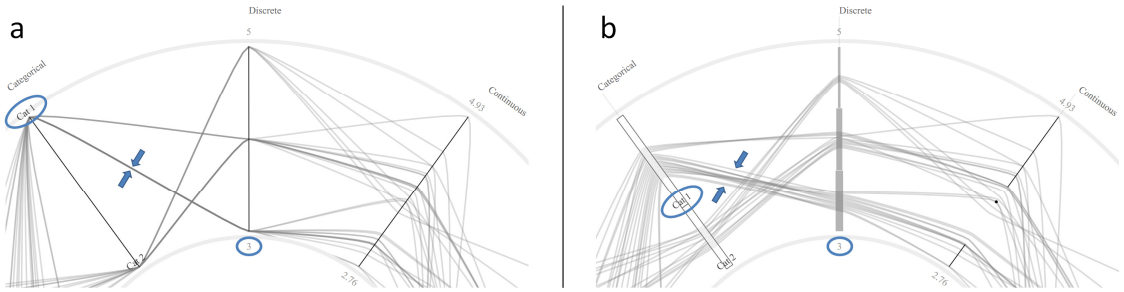

Figure 2. A sample data set without (a) and with (b) data-type dependent axis distribution visualizations and curve spreading. Axis distribution visualizations represent categorical axes as a stacked bar chart, discrete numeric axes as a histogram, and continuous numeric axes as a quartile plot[21], enabling rapid evaluation of the data type and overall distribution of the data for each axis. Curve spreading for categorical and discrete axes enables improved visualization of individual curves and clusters of curves, such as the number of data points with a Categorical value of Cat 1 and a Discrete value of three (highlighted in blue).

Results

Lung Cancer Prevalence

In Figure 3 the user has clicked on the lung cancer rate axis (*lung_Combined_DSR*), causing PCTs in the upper quartile of lung cancer rate to be automatically colored red, and the lower quartile blue. High and low lung cancer rates can now be compared across all dimensions in the data (Figure 3a). In the upper portion of the visualization it is apparent that PCTs with high and low lung cancer rates also tend to have high and low values for *extent*, *average_score*, *average_rank*, and *local_concentration* (also indicated by the correlation chords connecting these axes), which represent measures of social deprivation (poverty rate, socioeconomic status, etc.)

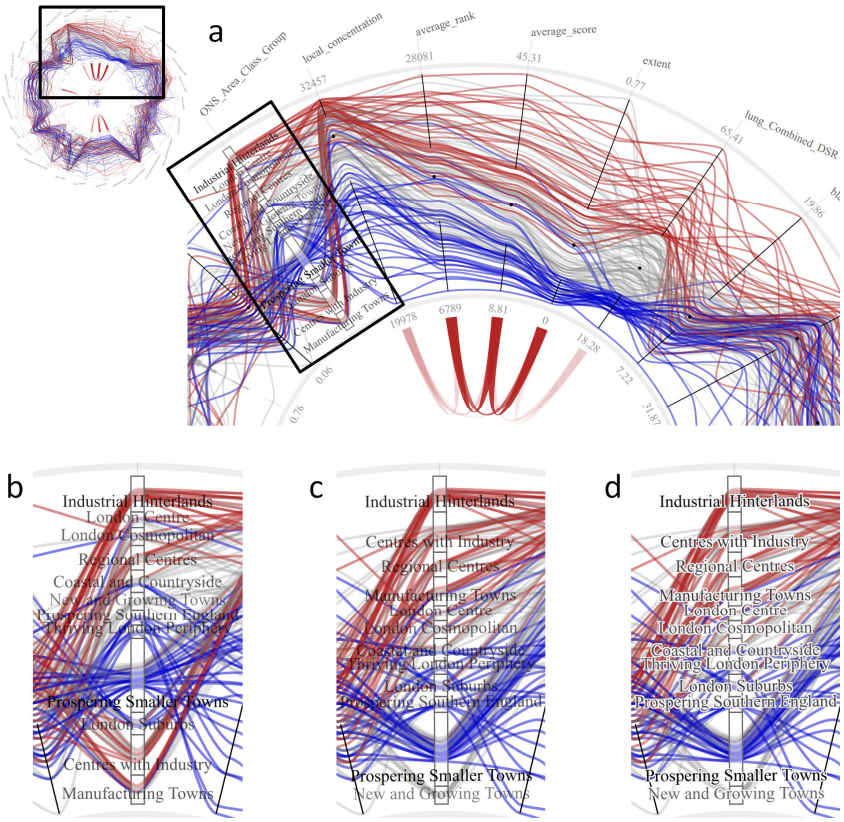

Figure 3. Visualization of lung cancer rates (red = upper quartile, blue = lower quartile) in 147 primary care trusts (PCTs) in the UK. High and low lung cancer rates tend to cluster based on regional classification (b), made clearer with automatic categorical axis reordering to cluster similar regions (c, d).

The red and blue curves also form clusters on the *ONS_Area_Class_Group* axis, a local region categorization from the ONS. Investigating this axis (Figures 3b-d) shows that *Industrial Hinterlands*, *Centres with Industry*, *Regional Centers*, and *Manufacturing Towns* all have high lung cancer rates, whereas *Prospering Smaller Towns*, *Prospering Southern England*, *London Suburbs*, and *Thriving London Periphery* all have low lung cancer rates. The discovery of such relationships via exploring the data visually drives the formation of causal hypotheses (e.g. pollution levels or smoking prevalence), which can be investigated further.

London Suburb Comparison

In Figure 4a a single PCT, Harrow, was seen to have the lowest lung, bladder, and colorectal cancer rates compared to all other PCTs, and has been highlighted in red. Harrow is classified as a *London Suburb*, so in Figure 4b the user has highlighted the other London suburbs in blue for comparison, made easier in Figure 4c via ribbon rendering. Harrow is shown to have a much higher value for the *oesophageal_Combined_SRR* axis, and thus a much higher esophageal cancer rate, than the other London suburbs, which are almost all in the lower quartile. This visualization raises the question of why Harrow has such a disparity in the rates of different cancers.

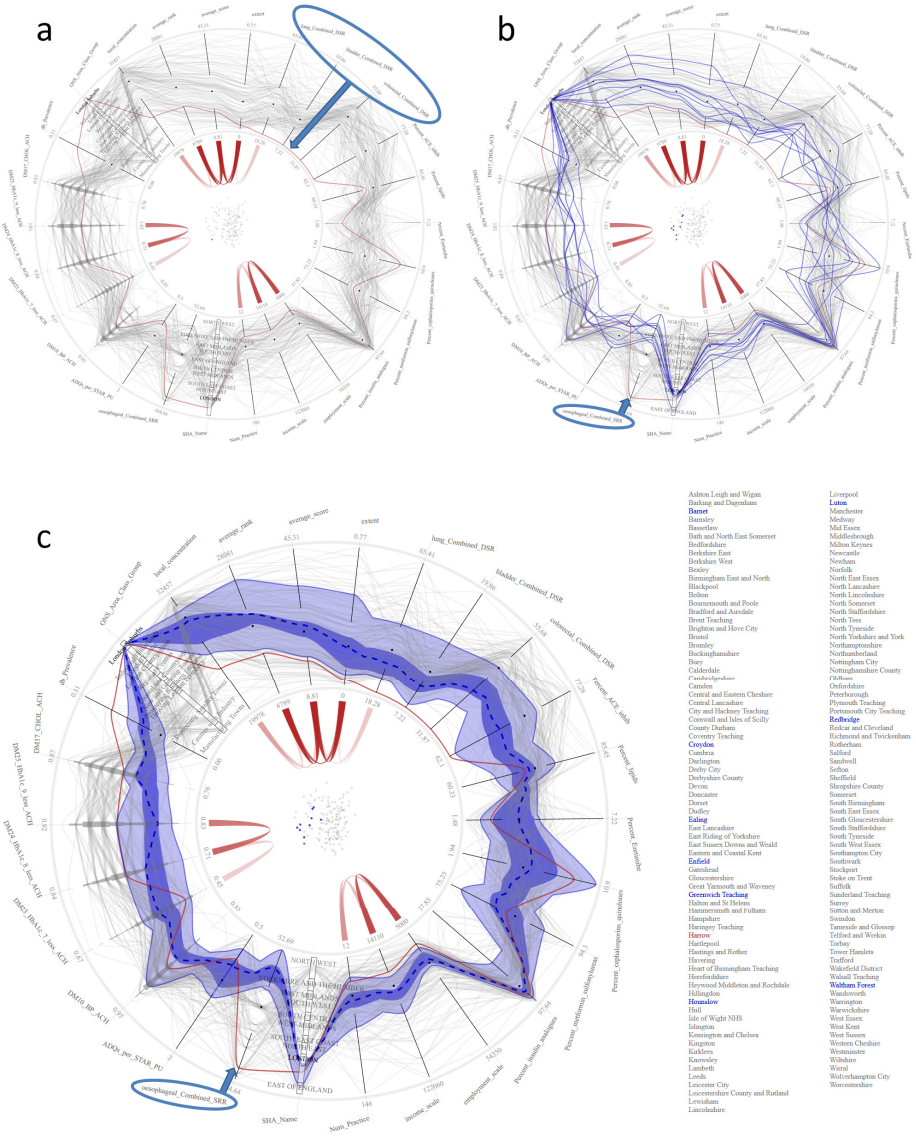

Figure 4. The Harrow PCT (red) has the lowest lung, bladder, and colorectal cancer rates (circled) among all 147 PCTs in the NHS dataset (a). Comparing Harrow to the other London suburbs (blue) reveals that Harrow has a much higher esophageal cancer rate (circled) than the other suburbs (b). Ribbon rendering makes it easier to visually compare Harrow with the other London suburbs (c).

According to the 2011 Census[22] Harrow is very diverse, with 63.8% of its population from the Black and Minority Ethnic communities, including the highest concentration of Sri Lankan Tamils and Gujarati Hindus in the UK and Ireland. India is known to have relatively low cancer rates in general, but some of the highest rates for oral and esophageal cancers in the world[23], which may help explain this phenomenon. Although further analysis is necessary, this example shows the utility of radial coordinates and ribbon rendering to compare subpopulations.

Practice-Level Data

Figures 8a and 8b show the prevalence of various medical problems (e.g. diabetes, dementia, and obesity) in the 1504 practices in the London strategic health authority (SHA). Figure 8a highlights in red two practices that appear to be outliers in the PCA scatterplot. Ribbon rendering makes apparent that they have the two highest prevalences for 12 of the 21 medical problems represented in the data. Figure 8b applies ribbon rendering to the remaining 1502 practices, making it easier to compare maximum and minimum values of medial problem rates for the two subpopulations.

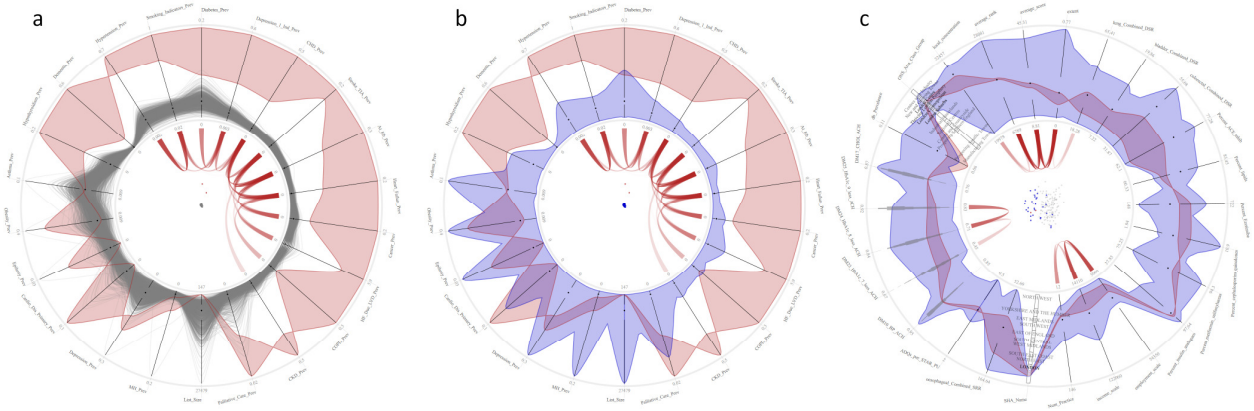

Figure 8. Two out of the 1504 practices in the London SHA, highlighted in red, have the two highest prevalences for 12 of the 21 medical problems represented in the NHS practice-level data (a and b). Comparing the PCTs containing these practices (red) to all other London PCTs (blue) does not reveal any major differences (c).

The two practices highlighted in red are Royal Hospital Chelsea in the Kensington and Chelsea PCT, and Nightingale House in the Wandsworth PCT. Because these two practices stood out so dramatically in the practice-level data, the user performed a PCT-level comparison of all London PCTs (Figure 8c). Interestingly, the Kensington and Chelsea and the Nightingale House PCTs (red) do not appear very different when compared to the other London PCTs (blue). Further research determined that Royal Hospital Chelsea is a retirement and nursing home for British soldiers and Nightingale House is a nursing home for the Jewish community that specializes in dementia, which may explain the high prevalence of problems such as dementia, hypertension, stroke, heart failure, and cancer in these two practices.

Conclusion

We have presented radial coordinates, a multivariate visualization technique based on parallel coordinates that incorporates features, such as per-axis population distribution visualizations based on data type (continuous, discrete, and categorical), direct visualization of correlations between variables, curve spreading for discrete and categorical data, visualization of summary statistics for user-selected subpopulations via ribbon rendering, and automatic reordering of categorical values based on user selection, driven by the needs of health-related data visualization.

We have applied radial coordinates to data from the UK's NHS at both the PCT and individual practice levels. Visualization of lung cancer rates among PCTs discovered possible relationships among lung cancer rate, socioeconomic factors, and regional classification. A comparison of London suburb PCTs revealed a potentially interesting PCT with a much higher esophageal cancer rate than other similar PCTs. Visualizing medical problem prevalence among over 1500 London practices showed two practices that have much higher rates of many medical problems. These examples illustrate the utility of the combination of visualization techniques embodied in our radial coordinates tool, and underline the need for further research in the use of visualization to aid in the analysis of complicated health-related datasets.

Acknowledgments

NHS and other UK data were made available courtesy of the BT Health Cloud. This work is supported by the US Army Medical Research and Materiel Command (USAMRMC) under Grant No. W81XWH-13-1-0061. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation.

References

1. West V, Borland D, Hammond WE. Visualization of EHR and Health Related Data for Information Discovery. In Proceedings of the 2013 AMIA Workshop on Visual Analytics in Healthcare. November 2013.
2. Gannet H. General summary, showing the rank of states, by ratios. 1880.
3. Inselberg A. The plane with parallel coordinates. Visual Computer. 1985;1(4):69-91.
4. Tominiski C, Schumann H. An event-based approach to visualization. In Proceedings of the Eighth International Conference on Information Visualization (IV'04). July 2004;101-107.
5. Rodrigues Jr. JF, Traina AJM, Traina Jr. C. Frequency plot and relevance plot to enhance visual data exploration. In Proceedings of the XVI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'03). 2003;117-124.
6. Edsall RM. The parallel coordinate plot in action: Design and use for geographic visualization. Computational Statistics and Data Analysis. 2003;43(4):605-619.
7. Siirtola H. Combining parallel coordinates with the reorderable matrix. In Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization. July 2003;63-74.
8. Holten D, van Wijk JJ. Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum. 2010;29(3):793-802.
9. Harter JM, Wu X, Alabi OS, Phadke M, Pinto L, Dougherty D, Petersen H, Bass S, Taylor II RM. Increasing the perceptual salience of relationships in parallel coordinate plots. In Proceedings of SPIE Visualization and Data Analysis 2012. January 2012.
10. Krzywinksi M, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Research. September 2009;19(9):1639-1645.
11. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. July/August 2006;12(4):558-568.
12. Schonlau M. Visualizing categorical data arising in the health sciences using hammock plots. In Proceedings of the Section on Statistical Graphics, American Statistical Association. 2003.
13. Fua YH, Ward MRE. Hierarchical parallel coordinates for exploration of large datasets. In Proceedings of the Conference on Visualization '99: Celebrating Ten Years. 1999;43-50.
14. Heinrich J, Weiskopf D. Continuous parallel coordinates. IEEE Transactions on Visualization and Computer Graphics. 2009;15(6):1531-1538.
15. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. Computer Graphics Forum. May 2008;27(3):1047-1054.
16. Heinrich J, Luo Y, Kirkpatrick AE, Zhange H, Weiskopf D. Evaluation of a bundling technique for parallel coordinates. In Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications. 2012;594–602.
17. Ankerst M, Berchtold S, Keim DA. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In Proceedings of the IEEE Symposium on Information Visualization. 1998;52-60.
18. Peng W, Ward MO, Rundensteiner EA. Clutter reduction in multi-dimensional data visualization using dimension reordering. In Proceedings of the IEEE Symposium on Information Visualization. 2004;89-96.
19. Seo J, Shneiderman B. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In Proceedings of the IEEE Symposium on Information Visualization. 2004;65-72.
20. Bostock M, Ogievetsky V, Heer J. D3: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis). 2011;17(12).
21. Tufte ER. The Visual Display of Quantitative Information. 2nd ed. Cheshire, CN:Graphics Press. 2001.
22. Office for National Statistics. 2011 Census: Ethnic group, local authorities in England and Wales. 2012.
23. Sinha R, Anderson DE, McDonals SS, Greenwald P. Cancer risk and diet in India. Journal of Postgraduate Medicine. July-September 2003;49(3).