Geo-Coordinated Parallel Coordinates (GCPC): A Case Study of Environmental Data Analysis

Maha El Meseery and Orland Hoeber*

Department of Computer Science, University of Regina,
Regina, Canada
{elmeseem,orland.hoeber}@uregina.ca

Abstract. Knowledge discovery in scientific and research datasets is an extremely challenging problem due to the high dimensionality, heterogeneity, and complex relationships within the data. When these datasets also includes temporal and geospatial components, the challenges in analyzing the data become even more difficult. A number of visualization approaches have been developed and studied to support the exploration and analysis among such datasets, including parallel coordinate plots, dimensional subsetting, geovisualization, and multiple coordinated views. In this research, we combine and enhance these approaches in a system called Geo-Coordinated Parallel Coordinates (GCPC), with the goal of supporting interactive exploration, analytical reasoning, and knowledge discovery.

1 Introduction

With advances in data collection and storage technology, the volume and complexity of scientific and research datasets are becoming increasingly overwhelming. Analyzing and understanding these datasets is an essential step in hypothesis development and scientific discovery. Discovering new and unexpected knowledge requires making sense of large amounts of high-dimensional and interrelated data. The need to derive insights from data collected for a particular domain or problem is driving researchers to design, develop, and study new tools and techniques to support data analysis and knowledge discovery.

Knowledge discovery is the process of identifying and understanding new meaningful patterns and trends contained within datasets [14]. It is a complex process that requires multiple iterations of data processing and transformation, hypotheses generation, and finally interpretation and reasoning about what has been discovered [14]. Such a process is an extremely challenging problem when the data is high-dimensional and heterogeneous, contains complex relationships among attributes, and has important temporal and spatial aspects.

Modern knowledge discovery systems utilize automated data analysis methods based on research from various fields including data mining, statistics, artificial intelligence, and machine learning. Even though these automated methods

* Corresponding author.

may be used to identify previously unknown aspects of the data, they provide researchers with few explanations about how or why the knowledge has been acquired, and provide little aid to the researchers in interpreting what has been discovered. Exploratory data analysis takes a different approach, with the aim of keeping humans involved in the discovery process . This often includes iterative investigation of the data, with the support of automated data processing, leading to the understanding of the patterns and the acquisition of new knowledge.

Visual analytics is an emerging approach that is increasingly being employed to support exploratory analysis of data [7, 23]. By combining information visualization, data processing, data mining, and interactive interfaces, analysts are able to explore, analyze, reason, and make sense of highly complex data [23]. Merging multiple visual and interactive representations helps analysts to generate hypotheses, identify new lines of inquiry, understand patterns, and derive new insight from what is being shown.

Our goal in this research is to develop a method to support the exploration and understanding of complex patterns and trends within high dimensional, heterogeneous, and geotemporal data. Geo-Coordinated Parallel Coordinates (GCPC) takes a visual analytics approach to the problem domain, using multiple coordinated views to simultaneously show, filter, and examine the data using parallel coordinates, micro-visualizations of the statistical features of the data, geovisualization, and investigative scatter plotting. Interactive and automated features support the knowledge discovery process, as well as the necessary task of hiding the complexity of the data to reveal the patterns.

The remainder of this paper is organized as follows. Section 2 provides a review of the key literature that has informed this research, including overviews of high dimensional data visualization, geovisual analytics, and multiple coordinated views. Section 3 outlines the key features of GCPC, followed by a case study in Section 4 that illustrates how these features can be used for data exploration and analysis activities, leading to new knowledge generation. The paper concludes with Section 5, which outlines the key contributions of this work, the limitations of the approach, and future work.

2 Literature Review

The use of visual analytics to support exploration, reasoning, and knowledge discovery within high dimensional geotemporal data is an active research domain with applications in many scientific fields [13, 22]. The following literature review focuses on three main topics that are relevant to our work: high dimensional data visualization, geotemporal data visualization, and multiple coordinated views.

2.1 High Dimensional Data Visualization

While a multitude of approaches have been developed over the years to visualize high dimensional data, each has its limitations [30]. Dimensional reduction methods use computational techniques such as principle component analysis [2] or multidimensional scaling [30] to transform the data to a lower dimensional space while preserving the relative proximity between data points [11]. The end result is a visual representation of the data in a coordinate space that has no obvious correlation to the actual dimensions, introducing complexity while exploring the data [11]. Dimensional subsetting methods use algorithmic techniques or user preferences to select a small subset of the dimensions to visualize. The success of such an approach is dependant on choosing which dimensions contain the most relevant and useful information [30]. This approach can be extended by displaying several dimensionally subsetted views of the data in a small multiples configuration. However, as more views are added to explore the relationships between the different dimensions, it becomes increasingly difficult to detect and interpret patterns within the data [20]. Instead of using simple shapes to represent each data point within these plots, glyphs can be used to encode additional dimensions of the data with shape, size, colour, orientation, or other visual attributes. However, there is a limit to the number of dimensions that can be visualized using glyphs before they become incomprehensible [12].

A fundamentally different approach to the problem is the use of parallel coordinate plots, where data are represented within a structure that maps each of the dimensions to a parallel axis [21]. Individual data points are represented using polylines, intersecting each axis at the appropriate location for the value on the specific dimension. This approach is very flexible and scalable with respect to the dimensionality of the data; adding a new dimension can be achieved by adding a new parallel axis and extending the polylines to their appropriate values [17]. The primary value of this approach over other high dimensional data visualization techniques is that all aspects of the data are shown, and the relationship between pairs of attributes can be investigated by interactively placing their axes beside one another.

However, there are also a number of important limitations. When multiple data points intersect an axis at the same location, ambiguity is introduced. This can be addressed by interactively highlighting the data points, or replacing the polylines with curves or density functions, both of which make it easier to perceive and follow the data points through the parallel coordinate structure [17]. When a large number of data points are shown using parallel coordinates, overplotting may occur, resulting in visual clutter. Some have explored clustering and outlier detection algorithms in order to reduce the amount of data that is shown, and to highlight those data points that are different from the norm [16, 32]. Since overplotting may also make it difficult to grasp the distribution of the data, adding statistical information to the parallel axes may be useful [18], but may also further contribute to the complexity of the display. When a particular dimension of the data represents discrete qualitative values, the ambiguity and overplotting problems become even more acute. Parallel sets provide an alternative for visualizing such categorical data, using ribbons between the coordinates to represent the frequency of each category [24]. However, combining this with traditional parallel coordinates when the data includes a combination of qualitative and quantitative data is not feasible.

2.2 Geotemporal Data Visualization

Geotemporal data visualization is a challenging problem due to the complexity of representing different scales and relations between the geospatial and temporal aspects of the data [7]. Simply mapping the data with traditional GIS tools, using different layers for each temporal range, limits the ability to dynamically analyze the features of the data. Several approaches have been proposed to support interactive visual analysis of geotemporal data. One of the earliest methods is the space-time cube, where location is represented in the map dimensions, and time represented in the third dimension [5]. Another straightforward method is using two coordinated visualizations: one to represent the temporal aspect of the data and the second to represent the geospatial aspect of the data. In this approach, a thematic dot map [22] or a choropleth map [16] may be used to visualize the location of the data; the temporal representation can then be used to interactively filter what is shown in the map. Others have studied methods for directly visualizing the temporal aspect of the data overlaid on the map (e.g., using a glyph to representing an aggregation of monthly temporal data [4]). However, by using visual attributes to represent the temporal aspect of the data, these attributes cannot be used to represent other multivariate aspects. Pre-processing the data to calculate changes over specified timeframes, and then visualizing these differences, can allow interesting features to be identified that would be difficult to discern otherwise [19]. Such an approach can be beneficial when the goal is to analyze how the data are changing over space and time.

2.3 Multiple Coordinated Views

Considering data that consists of both high dimensional attributes and geotemporal aspects, exploring and analyzing this type of data becomes an extremely challenging problem. A single visualization method is not adequate to support exploration, comparison, analysis, and knowledge discovery across the different aspects of such complex data. A common approach is to provide multiple visualizations of the data, which are linked together such that interactive manipulation in one (e.g., zooming, filtering, and focusing) results in a corresponding change in all others [27]. Views of the data that are customized to the specific meanings of the attributes have been used in this manner, such as the combination of a scatter plot matrix, time series visualizations, and word clouds [13, 16]. From a geospatial analysis perspective, the coordinated combination of parallel coordinates and geovisualization approaches have been studied for many years [3, 15]. However, there remains a shortage of geovisual analytics systems that support interactive analysis of high dimensional geotemporal data [6]. Furthermore, linking these approaches within an integrated analysis of the qualitative non-spatial attributes remains an open problem that we wish to address in this research.

3 Geo-Coordinated Parallel Coordinates (GCPC)

Discovering knowledge and testing hypothesis in environmental studies is a challenging problem due to the complexity of environmental data. Such data often consists of multiple heterogeneous factors with complex interrelations and important spatial and temporal aspects. Motivated by the challenges of exploring among such data, we have developed a geovisual analytics system to support analysis and reasoning about high dimensional heterogeneous geotemporal data. Geo-Coordinated Parallel Coordinates (GCPC) has been designed to enable the interactive analysis activities described in Keim's visual analytics mantra: "analyze first; show the important; zoom, filter, and analyze further; details on demand" [23].

The core of the system is comprised of two tightly coordinated features: a parallel coordinate plot and a geovisualization. These two views allow the system to represent the high dimensional, heterogeneous, temporal, and geospatial aspects of the data simultaneously. An optional scatterplot view allows analysts to interactively investigate correlations between pairs of factors. To further support exploration among the data, these visual components are linked through coordinated interactions: filtering, zooming, and highlighting the data to focus on interesting features in one view results in similar actions in the other views. Selections for visual encoding (e.g., colour, size) are replicated across all views, reinforcing the interpretation of the coordination across the views. A screenshot of the system is provided in Fig. 1.

Since environmental data analytics is seldom performed on just a single data set, the system was designed with data flexibility in mind. In order to load an arbitrary dataset, an automatic pre-processing step converts raw data into a tabular format with a single geospatial location per data point. The data types

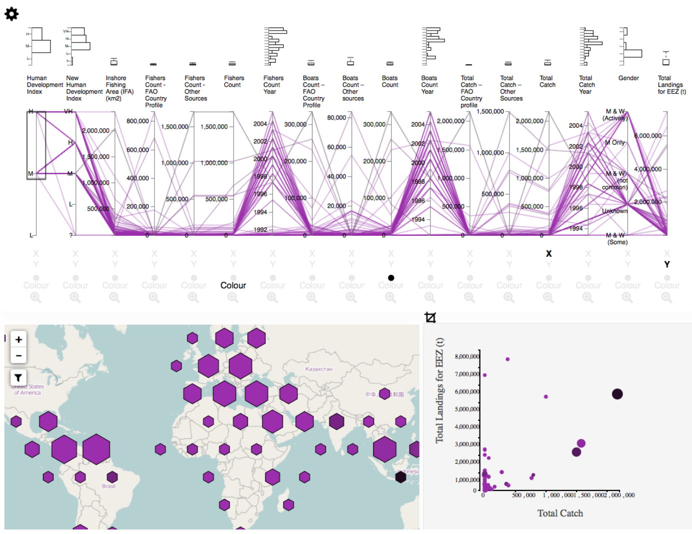

Fig. 1. The main view of the GCPC consists of parallel coordinates, micro-visualizations of the statistical properties of each dimension, interactive controls for configuring the visualizations, a geovisualization of the geospatial distribution of the data, and the investigative scatterplot. Here, the data is filtered for specific values on the first dimension and coloured based on the values in the sixth dimension.

are automatically detected, allowing appropriate visual encodings within the GCPC interface.

The software was developed within a web-based interface, using the Data Driven Documents (D3) library [8] as the core. Existing parallel coordinates [10] and geovisualization [1] plugins were used and extended to add the additional visual and interactive features. In the remainder of this section, the specific features and design considerations of GCPC will be explained.

3.1 Parallel Coordinate Plot

The parallel coordinate plot in GCPC allows multiple interactions to support exploration within the high dimensional data. Since the direct relationship between a pair of dimensions can only be seen if the dimensions are placed adjacent to one another, the system supports interactive reordering of the coordinates. Dimensions of the data that are not relevant for the current analysis activity can be interactively hidden. Investigating interesting subsets of the data is supported by filtering the data using brushing operations on the coordinates. Such filtering is immediately applied throughout the system to support further investigation in the other views. Detailed analysis is also supported by allowing the user to zoom-in on the brushed data, resulting in a rescaling of the data displayed on the coordinate. This can allow for the study of a subset of data that is tightly clustered within a narrow range of values.

The analyst may choose to colour the data according to the value on a chosen dimension, which is also reflected in the other visualizations. In order to ensure proper interpretation of the colour encoding, the encoding scheme is different for each data type. Quantitative data is encoded with a continuous, perceptually ordered scale; ordinal data is encoded with a discrete perceptually ordered scale; and qualitative data is encoded with perpetually distinct colour scales. Colour scales were chosen with an awareness of colour theory and the human interpretation of colour [31], using ColorBrewer [9] as the starting point for the specific scale selections. Since a given dataset may contain multiple different temporal aspects, it was decided to not provide a single timeline to filter the data, but instead to include these within the parallel coordinates. This allows the temporal features to be studied, filtered, and manipulated in the same way as other aspects of the data.

Statistical Descriptors One of the criticisms of using parallel coordinate plots is the difficulty in identifying the distribution of the data on a given dimension when there are a large number of data points. The compact nature of the parallel coordinates may result in overplotting and visual clutter, making it difficult to identify the precise data points going through a given value on a specific dimension. To address this limitation, micro-visualizations have been added to each parallel coordinate to illustrate the statistical properties of the data. While others have explored similar solutions by overlaying the statistical descriptors on the coordinates [18], GCPC provides these on top of each coordinate, allowing the information to be observed as needed, without interfering with the interpretation of the data shown in parallel coordinate plots.

The format of these statistical descriptors depends on the type of data they describe. Quantitative data is visualized using Tukey box plots [29], providing a compact representation of the median, quartile, and fifth/ninety-fifth percentiles. For qualitative and ordinal data, such measures are meaningless; instead histograms of the distributions are provided. Both formats allow the analyst to quickly observe and interpret the different types of data, providing an overview of the features of the dataset. Any filtering of the data automatically results in a recalculation of the statistical properties of what remains, and an update in these micro-visualizations.

Outlier Detection A second criticism of parallel coordinate plots is the difficulty in identifying outliers, due to the significant visual weight that is given to the dominant pattern within the data. As a result, it is difficult to visually isolate data points that are different from the norm. In some cases, such outliers may be uninteresting, and there may be a desire to remove these to reduce the additional visual clutter they cause. In other cases, the outliers may be important for the analysis at hand, and there may be a desire to highlight them. In order to support outlier analysis, GCPC includes an automatic outlier detection algorithm.

The approach employed is designed specifically for high dimensional data, based on the comparison of angles between multi-dimensional vectors [25]. This is based on the observation that for an anomalous data point, the angle to other pairs of data points in the collection will be small because of its distance from the other data. Conversely, for data points that are not anomalous, they will be surrounded by other points, resulting in large angles to other pairs of data points. The statistical variance of the angle is computed for each point to all other pairwise points in the dataset, which is then used as an outlier score to rank the data points. A data point is labeled as outlier if the score is lower than an empirically set threshold. Although not entirely accurate, qualitative and ordinal data are mapped and normalized to numerical values in this process.

More specifically, the angle based outlier detection score ($ABOD$) of point $\boldsymbol{A}$ is computed as:

$$ABOD(\boldsymbol{A}) = VAR_{\boldsymbol{B},\boldsymbol{C}\epsilon D}\left(\frac{\langle \bar{AB}, \bar{AC}\rangle}{\| \bar{AB} \|^2 \| \bar{AC} \|^2}\right)$$

where $\boldsymbol{B}$, and $\boldsymbol{C}$ represent all pairs of data points in the dataset D, and VAR is the statistical variance over these data. What is being calculated are the angles between point $\boldsymbol{A}$ and all pairs $\boldsymbol{B}$ and $\boldsymbol{C}$, noramlized by the length of the vectors $\| \bar{AB} \|, \| \bar{AC} \|$, which gives more weight to the score if points $\boldsymbol{B}, \boldsymbol{C}$ are nearer to point $\boldsymbol{A}$.

Because this algorithm must compare each data point to all other pairs, it is computationally expensive ($O(n^3)$). While classifying the data using an algorithm such as k-nearest neighbours can speed up the approach [25], for our purposes it is not necessary to calculate these outliers in real-time. Instead, the outlier ranking scores can be calculated during the pre-processing step and stored as part of the data, but keeping the cut-off threshold for outlier classification as an interactive parameter. The analyst may then choose how sensitive to make the outlier detection, and whether to use this to filter out the outliers or highlight them for detailed investigation.

3.2 Geovisualization

The purpose of the geovisualization is to allow the analyst to observe and interpret the spatial distribution of the data. This is an essential part of GCPC, allowing for the exploration among the relations between multiple factors and the geospatial aspects of the data. GCPC contains two main modes of displaying geospatial features on the map: a dot map that represents each point as a circle at the appropriate location, and hexagonal binning that represents aggregated spatial data on a hexagonal grid. While the process for producing the dot map is straightforward, there is some complexity in the creation of the hexagonal bin map. A grid of hexagonal polygons are layered over the map, and the data is aggregated based on which bin it falls into [26]. The default is to simply count the number of data points in each bin, but more complex aggregation such as total or average calculations are also possible. The size of the hexagons are then used to encode the data aggregated within the bins.

Settings below the parallel coordinate plot allow for the manipulation of two visual variables within the geovisualization: colour and size. When the dot map is shown, the size and colour of the dots are encoded based on the dimensions of the data chosen for these values. When the data is aggregated in the hexagonal bins, this colour and size encoding cannot be used directly. Instead, the size of the hexagon continues to be calculated as normal, but the colour is determined by the average value for quantitative data or the most frequent value for qualitative data.

The normal pan and zoom operations on the map allow the analyst to view more closely the geospatial relationships among the data. In order to further understand and explore among the data, the analyst may activate a geographical filter. The system allows the user to draw polygons to create arbitrary shapes that overlap a region of interest. The filter will remove all data points outside of the drawn shape, both within the geovisualization and also from the other visual representations. Coordinated highlighting allows the analyst to select specific data points within the map in order to isolate their attributes in the other dimensions using the other visual representations. These features enable the co-exploration of the data within both the high-dimensional elements and the geospatial elements.

3.3 Investigative Scatterplot and Correlation Analysis

Investigating the correlation between different factors and dimension is essential to understanding the complex relations within the data. While the order of the dimensions in the parallel coordinates may be manipulated to observe the pattern of the relationship, an analyst may wish to investigate such relationships in more detail and in a more fluid and interactive way. Selecting dimensions of the data to plot on the x- and y-axis using the controls under the parallel coordinates results in the creation of a scatterplot of the data. This enables a direct and intuitive analysis of the correlation between the selected attributes. Any selections of the colour and size encoding will also be present in this scatterplot, allowing for the interactive visualization of four dimensions of the data.

Because the analyst can easily change the dimension of the data to use for the axes, correlations can quickly be investigated and examined. Following the same coordinated interaction within the other views of the data, brushing over a region of this scatterplot will filter the other views, and selecting individual points will highlight their counterparts within the parallel coordinates and the geovisualization.

3.4 Data Inspection

During the exploration of the data, it is important to maintain the ability to drill down to the raw data in order to allow the analyst to inspect the actual values. This inspection may be used by the analyst to confirm what has been shown visually. When an individual data point is selected in any of the other views, a details window is populated with the complete set of data for this point. In addition, as the analyst filters the data using the parallel coordinates, geovisualization, and investigative scatterplot, they may wish to extract this specific subset of the data for detailed inspection and export into other software. A table view of the data supports this process, which only shows the data that matches the current filter settings.

4 Case Study

To demonstrate the features and utility of GCPC in the analysis of environmental data, we describe below three exploration scenarios of a dataset from the fisheries domain. The dataset was provided by the Too Big To Ignore (TBTI) research project, whose goal is to document and study the impact and importance of small scale fisheries around the world [28]. It consists of 127 data points over nineteen dimensions that include quantitative, qualitative, ordinal, temporal, and geospatial attributes that describe the small-scale fishing industry around the world. While the size of this dataset is relatively small, its high dimensionality, heterogeneity, and geotemporal attributes made it difficult to analyze using traditional means.

Initial Exploration The analyst in this case study is an environmental researcher trying to explore and compare the impact of small scale fisheries across the broad range of attributes collected. Loading the dataset into GCPC will automatically identify the type of each dimension and calculate initial statistical distributions. As shown in Fig. 2, the system will default to showing all of the data in the parallel coordinates plot, and the locations at which this data was collected in the geovisualization. From this overview of the entire dataset, global patterns in the data may be observed, including the distribution of the data over the dimensions, the correlation between adjacent coordinates, and the geodistribution of the data. This initial assessment of the data can then be used as the basis for confirming what is known (e.g., the extent of small-scale fishing in Central America), and developing and evaluating new hypotheses about the data.

One aspect of the data that can be readily observed from the overview is that it is highly irregular, with a small number of extreme values that extend the range of some coordinates (e.g., *Inshore Fishing Area* (third parameter), and *Fishers Count* (sixth parameter)). This pattern in the data causes the remaining data points to be clustered at the other end of the scale, making it difficult to discern their pattern. There are two mechanisms built into GCPC that can address this problem: using the automatic outlier detection to hide these data points that are substantially different than the norm, or using the interactive focusing, filtering, and zooming features on the coordinates of interest. Supposing that the analyst wants to retain interactive control over the analysis process, the first step in exploring these coordinates is to inspect the extreme values. By clicking on each, the researcher will observe that they correspond to countries with large fishing regions (i.e., Canada, Indonesia, and Australia). The data on these dimensions can be filtered easily, by interactively dragging a bounding box over the coordinates. Clicking on the zoom icon will cause the selected range to fill the available space for the coordinate in question. The results of this filtering and zooming operation can be seen in Fig. 3.

Knowing that there are outliers in the data, the analyst may wish to have the system automatically find these so that they can be evaluated, and then perhaps hidden when conducting future analyses of the data. Fig. 4 shows these

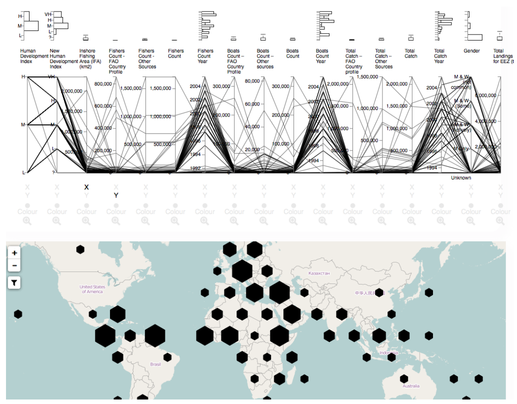

Fig. 2. The default view of the data loaded in GCPC.

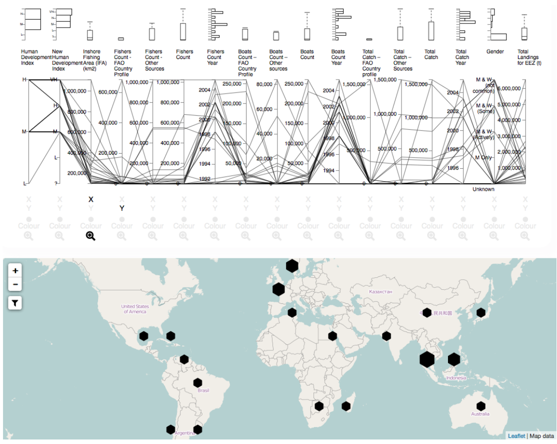

Fig. 3. Zooming the scale on the *Inshore Fishing Area* dimension from over 1,000,000 to 100,000 reveals the pattern at the lower values on this dimension.

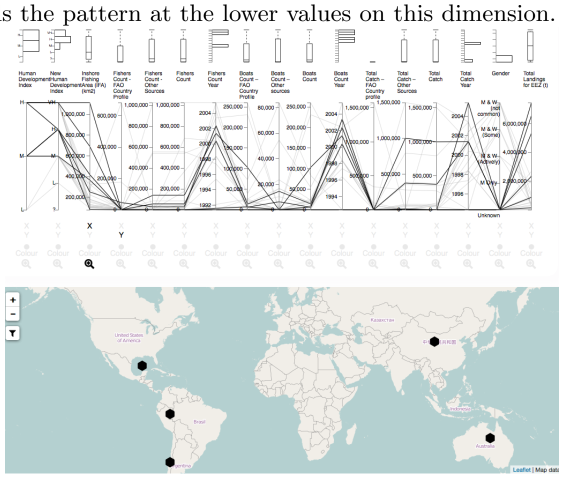

Fig. 4. Highlighting the outliers data allows the analyst to easily inspect these, and the subsequently hide them.

anomalies, and dims the remaining data so that the overall pattern can still be observed. The algorithm identified countries that are considerably different than the normal pattern of the data across multiple dimensions (e.g., Chile and Mexico). Another set of outliers detected were data points with multiple missing information, which render them substantially different than normal data (e.g., Australia). Isolating these anomalies from the rest of the data would be tedious and cognitively taxing had it been done manually.

Analysis of Attribute Relationships After this initial observation and exploration, the researcher may be interested in the development and testing of an hypothesis that relates *Total Catch*, *Boats Count*, *Fishers Count*, and *Total Landings*. A first step in such an examination is to re-order the coordinates such that they are adjacent to one another. Doing so, allows the analyst to observe direct or inverse correlations easily. More complex relationships can be observed using the investigative scatterplot, mapping these attributes to the x-axis, y-axis, colour, and size options (see Fig. 5). Since the colour and size parameters are also represented on the map, the locations where the *Total Landings* and *Fishers Count* are large can be observed. This analysis shows a pattern of the correlation between these parameters, as well as the instances of data points that are counter to the pattern (e.g., Japan, with low *Fishers Count*, but high *Boats Count* and *Total Catch*).

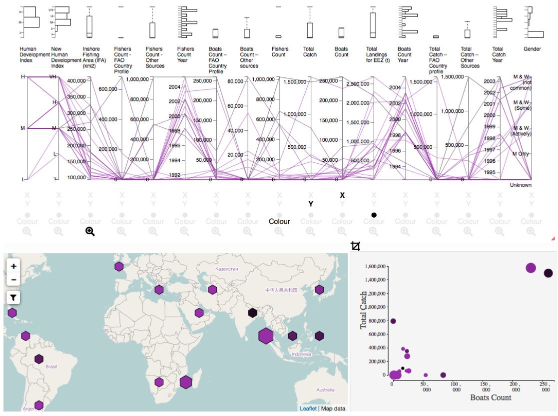

Fig. 5. Investigating the relations between *Total Catch*, *Boats Count*, *Fishers Count*, and *Total Landings* using the investigative scatter plot and visual encodings.

Analysis of Spatial Relationships An important step in analyzing data such as this is to make comparisons across different geographical regions. Suppose the researcher wishes to study the gender distribution of small-scale fishers between Europe and Africa. The map can be zoomed to these regions independently, and then free-form shapes can be drawn around the areas of interest. Doing so filters the data shown in the parallel coordinates, which may be further filtered, perhaps in order to focus on the gender distribution in the most recent data. Screenshots showing these two analyses are provided in Fig. 6. From this, we can readily observe that in France, Italy, and Greece, it is not common for both men and women to fish. However, in sub-Saharan Africa, recent data shows that both men and women are actively involved in the small-scale fisheries.

5 Conclusion and Future Work

In this paper, we presented Geo-Coordinated Parallel Coordinates, a visual analytics system designed to support the exploration and analysis of high dimensional, heterogeneous, geotemporal data. The main contribution of the system

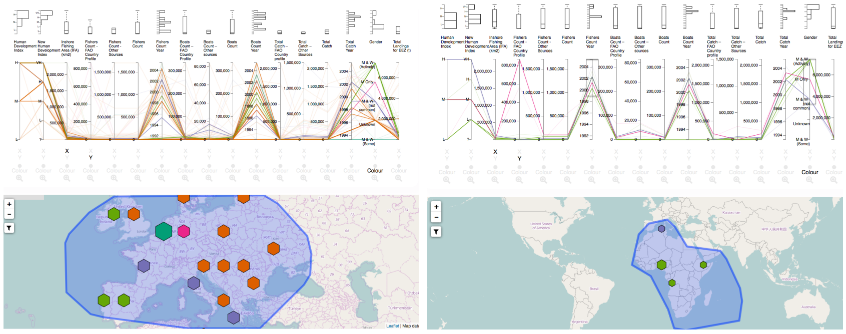

Fig. 6. Zooming the map and filtering the data to specific geographic regions allows for the isolation of this data within the parallel coordinates, enabling the comparison of parameters such as the differences on the *Gender* dimension.

are: (1) the integration and coordination of multiple visualization and interaction techniques; (2) the micro-visualizations of the statistical information for each dimension added to the parallel coordinates; (3) the use of automatic outlier detection to allow highlighting and filtering of outlier data; and (4) the flexible design that allows arbitrary high-dimensional data to be loaded into the system.

Even though this paper demonstrated through a case study the benefits of GCPC in facilitating analytic reasoning through interactive exploration of the data, there are some limitations. The current implementation assumes the geospatial data are point data, and does not support data that represents geospatial regions. Complex data types such as hierarchical data and missing data cannot be represented within the parallel coordinate structure. GCPC does not currently support the analysis of complex temporal patterns such as temporal ranges and movement data. Even though we endeavoured to address some of the fundamental limitations of parallel coordinates in GCPC, it remains difficult to detect patterns over more than four dimensions of the data.

Future work may include addressing some of these limitations. Supporting different geospatial features (e.g., regions, trajectories) will allow the system to more readily support the analysis of such data. Modifying the angle-based outlier detection approach to more accurately detect differences in qualitative and ordinal data will improve the anomaly detection. Clustering the data may address the overplotting issues within the parallel coordinates, allowing high-level patterns within the data to be identified. Adding a separate view for temporal data (e.g., a timeline), will enable users to analyze different types of temporal data and identify complex geotemporal patterns and trends. Since this approach is highly interactive, we are currently in the planning stage for an empirical evaluation with expert data analysts, which will provide evidence of the value and usefulness of the approach for real-world data analysis activities.

6 Acknowledgements

The authors wish to thank the Too Big To Ignore (TBTI) project for their support and dataset they provided. This work was supported in part by grant from Social Sciences and Humanities Research Council (SSHRC) held by the second author.

References

1. Agafonkin, V.: Leaflet. http://leafletjs.com/ (2010), accessed: December 2014
2. Aggarwal, C.C.: Data Mining, The Textbook. Springer International Publishing, Switzerland (2015)
3. Andrienko, G., Andrienko, N.: Exploring spatial data with dominant attribute map and parallel coordinates. Computers, Environment and Urban Systems 25(1), 5–15 (2001)
4. Andrienko, G., Andrienko, N.: Spatio-temporal aggregation for visual analysis of movements. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology. pp. 51–58 (2008)
5. Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S.I., Jern, M., Kraak, M.J., Schumann, H., Tominski, C.: Space, time and visual analytics. International Journal of Geographical Information Science 24(10), 1577–1600 (2010)
6. Andrienko, G., Andrienko, N., Keim, D., MacEachren, A.M., Wrobel, S.: Challenging problems of geospatial visual analytics. Journal of Visual Languages & Computing 22(4), 251–256 (2011)
7. Andrienko, N., Andrienko, G., Gatalsky, P.: Exploratory spatio-temporal visualization: An analytical review. Journal of Visual Languages & Computing 14(6), 503–541 (2003)
8. Bostock, M., Ogievetsky, V., Heer, J.: D3 data-driven documents. IEEE Transactions on Visualization and Computer Graphics 17(12), 2301–2309 (2011)
9. Brewer, C.A., Hatchard, G.W., Harrower, M.A.: Colorbrewer in print: A catalog of color schemes for maps. Cartography and Geographic Information Science 30(1), 5–32 (2003)
10. Chang, K.: Parallel coordinates toolkit. https://syntagmatic.github.io/parallel-coordinates/ (2012), accessed: December 2014
11. Choi, J.Y., Bae, S.H., Qiu, X., Fox, G.: High performance dimension reduction and visualization for large high-dimensional data analysis. In: Proceedings of the IEEE/ACM International Conference on Cluster, Cloud and Grid Computing. pp. 331–340 (2010)
12. Chung, D.H., Legg, P., Parry, M.L., Bown, R., Griffiths, I.W., Laramee, R.S., Chen, M.: Glyph sorting: Interactive visualization for multi-dimensional data. Information Visualization 14(1), 76–90 (2015)
13. Dörk, M., Carpendale, S., Collins, C., Williamson, C.: VisGets: coordinated visualizations for web-based information exploration and discovery. IEEE Transactions on Visualization and Computer Graphics 14(6), 1205–12 (2008)
14. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. Advances in Knowledge Discovery and Data Mining, AI Magazine 17(3), 37–54 (1996)

15. Gahegan, M., Takatsuka, M., Wheeler, M., Hardisty, F.: Introducing Geo VISTA Studio: An integrated suite of visualization and computational methods for exploration and knowledge construction in geography. Computers, Environment and Urban Systems 26(4), 267–292 (2002)
16. Guo, D., Chen, J., MacEachren, A.M., Liao, K.: A visualization system for space-time and multivariate patterns (VIS-STAMP). IEEE Transactions on Visualization and Computer Graphics 12(6), 1461–1474 (2006)
17. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. In: Proceedings of Eurographics - State of the Art Reports. pp. 95–116 (2013)
18. Ho, Q.V., Lundblad, P., Åström, T., Jern, M.: A web-enabled visualization toolkit for geovisual analytics. Information Visualization 11(1), 22–42 (2012)
19. Hoeber, O., Wilson, G., Harding, S., Enguehard, R., Devillers, R.: Exploring geo-temporal differences using GTdiff. In: Proceedings of the IEEE Pacific Visualization Symposium. pp. 139–146 (2011)
20. Im, J.F., McGuffin, M.J., Leung, R.: GPLOM: the generalized plot matrix for visualizing multidimensional multivariate data. IEEE Transactions on Visualization and Computer Graphics 19(12), 2606–2614 (2013)
21. Inselberg, R.: The plane with parallel coordinates. The Visual Computer 1(2), 69–91 (1985)
22. Jänicke, S., Heine, C., Scheuermann, G.: GeoTemCo: Comparative visualization of geospatial-temporal data with clutter removal based on dynamic delaunay triangulations. In: Proceedings of the International Conference on Information Visualization Theory and Applications. pp. 160–175 (2013)
23. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In: Kerren, A., Stasko, J.T., Fekete, J.D., North, C. (eds.) Information Visualization: Human-Centered Issues and Perspectives, pp. 154–175. Springer-Verlag (2008)
24. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics 12(4), 558–68 (2006)
25. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 444–452 (2008)
26. Ramakrishna, A., Chang, Y., Maheswaran, R.: An interactive web based spatio-temporal visualization system. In: Proceedings of the International Symposium on Advances in Visual Computing. pp. 673–680 (2013)
27. Roberts, J.: State of the art: Coordinated & multiple views in exploratory visualization. In: Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization. pp. 61–71 (2007)
28. TBTI: Report on enhancing stewardship in small-scale fisheries through ecosystem approaches and other means. Tech. rep., Too Big to Ignore Latin America and the Caribbean Joint Workshop with Working Group 4, Curitiba, Brazil. TBTI, Canada (Aug 2013)
29. Tufte, E.R.: The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT, USA, 2nd edn. (2001)
30. Ward, M., Grinstein, G., Keim, D.: Interactive Data Visualization: Foundations, Techniques, and Applications. A. K. Peters, Ltd., Natick, MA, USA (2010)
31. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann Publishers Inc., Waltham, MA, USA, 3rd edn. (2013)
32. Zhou, H., Yuan, X., Qu, H., Cui, W., Chen, B.: Visual clustering in parallel coordinates. Computer Graphics Forum 27(3), 1047–1054 (2008)