Event Line View: Interactive Visual Analysis of Irregular Time-Dependent Data

Krešimir Matković[1], Alan Lež[1], Denis Gračanin[2],
Andreas Ammer[1], and Werner Purgathofer[1]

[1] VRVis Research Center in Vienna, Austria,
{Matkovic, Lez, Ammer, Purgathofer}@VRVis.at
[2] Virginia Tech, gracanin@vt.edu

Abstract. In this paper we present a novel approach to visualize irregularly occurring events. We introduce the *event line* view designed specifically for such events data (a subset of time dependent data). The *event line* view is integrated in a coordinated multiple views (CMV) system and linked with other conventional views to support interactive visual analysis. The main idea is to analyze events relative to two categorical attributes from a multidimensional multivariate dataset. Since we are interested in the categorical dimension we have also integrated and linked the *tag cloud* view in the CMV system. To the best of our knowledge this is the first integration of the *tag cloud* view in a CMV system. The *tag cloud* view can depict a ratio of the selected items versus the non-selected items.

The proposed approach is illustrated using the VAST Challenge 2008 Geo-Spatial data set that contains data on interdiction or landing of illegal immigrants in the USA. It is a multivariate multidimensional dataset with irregular events that illustrates the potential and capabilities of the proposed approach and the developed CMV system.

Key words: Interactive Visual Analysis, Coordinated Multiple Views, Events in Time, Tag Cloud

1 Introduction

Interactive visual analysis helps with data exploration and analysis in various application domains. There are numerous successful visualization approaches for various data types. Multivariate multidimensional data sets are collections of records (or rows in a table like in a spreadsheet program, e.g.) where each record contains multiple attributes. This is one of the most common data organization in information visualization.

We investigate a special case of such data sets we call *events in time*. The main characteristics is that various events (represented by single records) happen at irregular times. A good example of such a data are traffic (or any other) accidents. We do not know in advance when an accident will occur, but we are interested to find some patterns and hidden correlations in order to prevent future accidents, if possible. Another example is a publication list of a research

institution. Here we also have temporally irregular events, but some regularity might occur (larger frequency close to important conference submission deadlines, e.g.). In a business domain selling of various goods can also be considered to be an event. The same is true for events in a computer system.

Time dependent data visualization is a very active field of research [12], but we focus on a specific variant of such data sets. We present a novel view, the *event line* view, tailored specifically for the data sets containing irregular events. Since there are usually more dimensions in a data set, and we are looking for a hidden correlations, we have integrated the new view in a coordinated multiple views (CMV) system.

The correlation among the irregular events can be explored by custom grouping of the events. The events can be grouped based on the specific values of the categorical dimensions. In general, we can have a large number of categorical dimensions in a data set.

We use two categorical dimensions within a single *event line* view. In most cases, two dimensions provide sufficient grouping flexibility, e.g. considering spatial (e.g. location) and temporal (day of a week) categorical values. Linking several *event line* views within the CMV system allow us to use any number of categorical dimensions.

Once the events are selected in the *event line* view, we look for correlations among them. A correlation is indicated by the common values in other categorical dimensions. What we need is a way to show what values of a categorical dimension are the most often present in the selected events,

Categorical data is usually described with strings (or tags) that can be represented using a *tag cloud* view [2]. Therefore, we included the *tag cloud* view in the CMV system. That way, we can use the *event line* view to select events and the *tag cloud* view to identify possible correlation among the selected events. In this way a data set containing events in time can be efficiently analyzed. To the best of our knowledge we are the first to integrate the *tag cloud* view in CMV.

We explain the view design and show effectiveness of the view using the VAST Challenge 2008 Geo-spatial dataset [4]. Our entry in that challenge [11] describe the use our CMV system (ComVis tool [8]) that, in connection with some helper applications and Google Earth, allowed us to explore geo-temporal characteristics of the data set and answer the challenge questions. We build on these results in order to improve the visualization and analytical capabilities by adding new views to the CMV system

1.1 Related Work

Event data visualization and analysis tool are still not very common. Suntinger et al. [14] were among the first exploring such data sets. They have introduced a system called Event Tunnel used for business applications.

Visualization of event data can be considered a part of a larger topic of high-dimensional and time-dependent data visualization. Many approaches to the interactive visual analysis of time dependent data have been already introduced.

Müller and Schumann, for example, provide good overview of the visualization methods for time-dependent data [12].

The related visualization techniques can be classified into two groups based on whether or not the visual representation itself is time-dependent. Aigner et al. [1] provide an overview of visual methods for analyzing time-oriented data and discuss general aspects of time-dependent data. Multivariate data visualization techniques for multi-variate time-dependent data include the ThemeRiver [6], Spiral Graph [16] and others.

Time-dependent data often exhibit some periodic behavior. Such serial periodic data are of special interest. For example, time continues forward sequntially but includes recurring periods (weeks, months, and years) [3]. The challenge is how to simultaneously display serial and periodic attributes of the data set. Spiral arrangement of the data provides easy visual cues to both serial and periodic aspects of the data. Special visual metaphors like the Calendar View or Lexis Pencils are also used. In the Calendar View [15], van Wijk et al. use a very intuitive calendar view to display clusters of the time series data. Lexis pencils [5] display various time-dependent variables on the faces of the pencil. Pencils can depict several variables and can be positioned in space to indicate the spatial context of the data.

Extensions of traditional visualization techniques to time-dependent data include the Wormplot that provides a scatter plot for each time step. The Time Wheel leverages parallel coordinates so that the time axis is in the center and all other axes are located circularly around it. The MultiComb arranges those other axes in a star-shaped pattern.

Wegenkittl et al. [17] use extruded parallel coordinates, linking with wings, and three-dimensional parallel coordinates. Brushing the time axis to display details of the selected time frame is a very common and useful interaction technique used with static representations [8, 7]. Timebox widgets by Hochheiser et al. [7] can be used to brush both the time axis and the attribute axis of graphs.

Coordinated multiple views are also a well established methodology in data analysis. A good overview of the Coordinated and Multiple Views state of the art and what future may hold for this area is given by Roberts [13].

Tag cloud view is a standard way of visualizing text, especially for web. There are numerous papers dealing with tag clouds and improving of placement. Lohmann et al. [10] give an overview of various tag clouds algorithms. However, we are not aware of a *tag cloud* view integration into a coordinated multiple views system.

2 Data Model

Our main goal is to analyze data that contains time-based events. The events appear at irregular times, and there are more attributes describing event than just time. We will introduce the data model used first.

We explore the data sets thats contain many records, and each record contains several attributes. One record corresponds to one event. One of the attributes is

LAT	LON	Type	Pass.	USCG	Date	Death	Vessel	Day	Month	Season	Region
-80.44	28.50	Land	24	N/A	20070510	0	Go Fast	Thu.	May	Spr.	Florida1
-80.86	29.11	Land	18	N/A	20070309	0	Rustic	Fri	Mar	Win.	Florida1
-79.96	24.98	Inter	18	Stal.	20070901	0	Rustic	Sat.	Sep	Sum.	Interd
...											
-86.74	21.26	Land	10	N/A	20070704	0	Rustic	Wed.	Jul	Sum	Mexico

Dataset used in the analyis, there are 917 records in total.

Regions used to describe various landing spots.
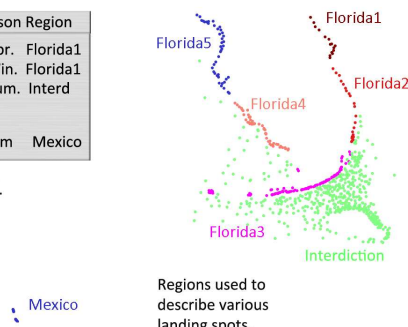
Fig. 1. The table on the left illustrate multidimensional multivariate dataset used. There are some numeric dimensions (latitude, longitude, number of passengers and number of deaths), and some categorical dimensions (week day, month, region, . . .). There is also the time dimension. One record (row in the table) represents one event in time. The implicit map on the right, created using a scatterplot and longitude and latitude shows how we have aggregated the region dimension.

time, and others can be either categorical or numerical values. Such a data set represents a typical multidimensional multivariate data set. It can be accidents data set containing time when an accident happened, information if cars, trucks, or busses were involved, were there any injuries, etc. Another example can be phone calls. There we could have time the call was started, duration, persons involved, various attribute of persons such as age, gender, etc.

We use the data set originally issued for the VAST Challenge 2008. The data set includes the interdiction records collected by the United States Coast Guard and information from other sources about illegal landings on shore. Each record has several attributes, including longitude and latitude (either form interdiction or from the landing), date and type (interdiction or landing), type of vessel used by immigrants, name of US Coast Guard Ship in case of interdiction, number of passengers on board, and number of fatalities.

We have derived some additional data in order to easier answer questions that appear in the analysis. We have used date to compute day of week, month, and season as additional attributes. We have also grouped landing positions into seven regions, five in Florida, one in Mexico, and one for interdicted vessels.

Figure 1 illustrates the dataset on the left and shows five Florida and one Mexican regions. The seventh region represents successful interdiction spots. We have used simple scatter plot to depict latitude and longitude. The map created in this way is distorted due to the changed aspect ratio, but Florida can be easily recognized.
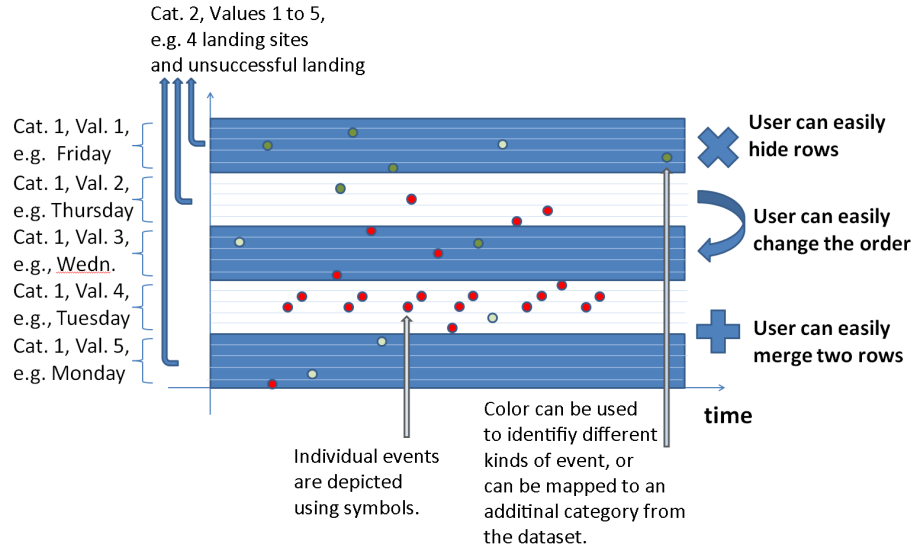
Fig. 2. The main principle behind the *event line* view. The x-axis depicts time and the y-axis represents two categorical dimensions. The y-axis is divided in strips (blue and white in the image), so that there is a strip for each value in the first category (e.g. a weekday). Each strip is than subdivided into values of another category. In our case, those values can be landing regions. Each event is now represented by a symbol in the view. There can be various symbols depending on an additional criterion. The user can merge strips, hide strips, and change strips order. The y-axis can be re-parameterized to show different dimensions.

3   The Event Line View

The events line view is tailored especially for analysis of events in time with respect to two categorical attributes. We will describe the view design first, followed by interaction and integration in a coordinated multiple views setup.

We use the x-axis as time while the y-axis is used for two categorical values. We compare the events with respect to the two categorical attributes, as this is often the case in the analysis. In the case of our data an analyst would like to see events and compare how e.g. day in week correlates with successful interdiction/landing during the time.

In order to do so we will divide the y-axis in parts. Each part represents one value for the first category. In the case of above example this would be successful interdiction in upper half of the axis and successful landing in to lower part. Now when the axis is divided, we subdivide each part in the subranges, one for each value of the second category. In our case, we will have seven subcategories, one for each day in the week. This means that number of ranges will equal to a product from number of ranges in each of the two categories, or 14 in our case: seven for landings and seven for interdiction.

The events themselves are depicted using a point (or any other symbol) in the corresponding range strip and time position. Figure 2 illustrates the main idea. The user can choose the categories at any time and change the parametrization of the view.

Since the number of ranges on the y-axis can be very large and make the view less useful, we propose several mechanisms to cope with this issue. One option is to hide less interesting ranges. That saves space, but at the same time reduces information available. Another possibility is to merge several ranges into one. This can be done for the ranges that are less interesting, or in the case when the screen space is limited. When the ranges are merged, the points or symbols used to depict events have different shape or color depending on their original ranges.

The shape and color of the symbols can always be customized dependent on range or any additional attribute from the data set. We currently support circles and capital letters as symbols. Letters are convenient when ranges are merged and we want to distinguish between original events. Of course, any other symbol could be used, but it is challenging to design a large number of easily distinguishable small symbols.

In case of a long time span, or a coarse time scale (like for example events described with month (or day) only, it is possible that more than one event share the same position in screen. We propose two techniques here, one is symbol size increase based on the number of events represented by a single symbol, and the other one is dispersion of the symbols in a close neighborhood. The overlapping symbols can also occur when several ranges are merged. The same technique is used to overcome the problem in this case, too. Additionally, a zoom of the time axis is useful for the very long time spans.

The order of ranges on y-axis can enhance the comparison possibilities significantly. The patterns in neighboring ranges can be spotted more easily. Our events line view makes it possible to rearrange ranges by moving a specific range up or down.

The events line view is integrated in a coordinated multiple views system. Whenever a subset of data is brushed (selected) in any other view, corresponding events in the *event line* view are highlighted. The view supports brushing too, the user can brush certain events in the view, and corresponding items in all other views will be highlighted. There can be several *event line* views with different parameterizations active simultaneously.

4   The Tag Cloud View in Coordinated Multiple Views

The *event line* view is used to analyze various events in respect to two categorical values. Categorical values themselves can be visualized in numerous ways. Parallel sets [9] have been developed especially for categorical values. A histogram, stacked bars, a pie chart or even a scatter plot can be used as well.

The *tag cloud* view is a standard view for displaying tags or keywords. It is widely used on the web and represents a well known and very efficient way

Fig. 3. The *tag cloud* view integrated in the CMV system. The top left *tag cloud* view shows months as contained in our data set. The user selected only the interdiction region (in a histogram view not showed here) and the selection is shown with context in the lower row (horizontal and vertical), and without the context in the top right *tag cloud* view.

of visualizing frequency of appearance of single words or short phrases. Since categorical values are often represented by words, and we focus on comparison of events in respect to categorical data, we decided to include the *tag cloud* view in the CMV system.

There are many ways how to optimize layout of a *tag cloud* view [10]. Our goal is not to improve the *tag cloud* view itself, but to include it in the CMV system and use it to depict categorical data. To the best of our knowledge, this is the first attempt to use the *tag cloud* view as a part of the CMV system.

If the user brushes a subset of the data in any view, the selection will be highlighted in the *tag cloud* view as well. In order to show the ratio of the selected and unselected items for each category value (or a word in the *tag cloud* view), we show the original *tag cloud* view in gray (or any user selected context color), and highlight a part of the word depending on the selection using focus color (red in our case). The selection can be depicted horizontal or vertically.

Figure 3 illustrates the idea. The *tag cloud* view shows distribution of months in our dataset. Just as expected, most of the events occurred during warm, summer months.

The top left *tag cloud* view shows the overall distribution. We have brushed interdictions now (selection was done using a simple histogram not shown in Figure 3). The bottom row shows two versions (horizontal and vertical) of the selection.

The user can optionally choose to show brushed values only, as depicted in the top right *tag cloud* view. In this case only words that belong to the selection are shown (only focus, without context), and resized accordingly. The context is not shown, but comparison within selection is easier. The user can easily switch between the two modes in order to gain better insight into data.
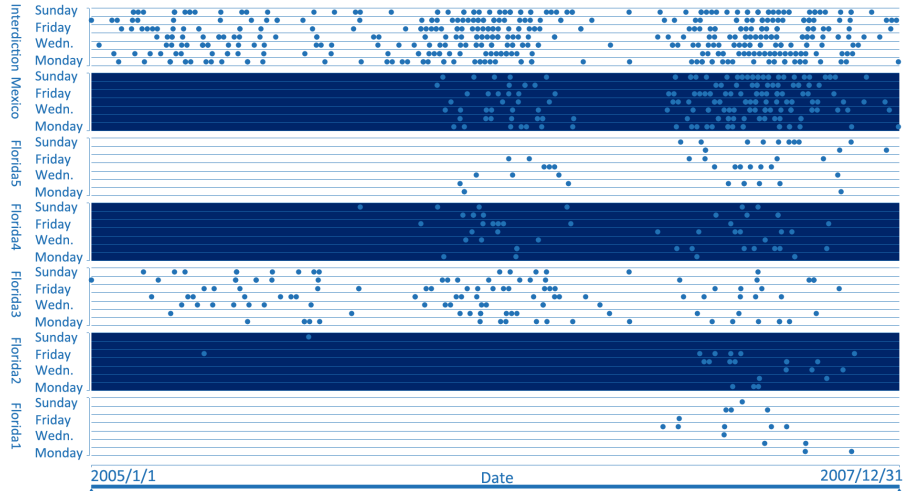
Fig. 4. The *event line* view for years 2005, 2006, and 2007. Some interesting things can be noticed. First of all, the increase in the number of events is accompanied by the increase in the number of landing spots over time. We can see that some regions are active during all three years (*Florida3*), and some became active in 2007 (*Florida1* and *Florida2*). Also, looking at week days it can be noticed that *Florida2* is almost never active during weekends and that *Florida5* has a big gap on *Monday*.

5   Illustrative Example

In this section we provide examples to illustrate usefulness of the proposed approach. We will start with the complete data set which contains events from years 2005, 2006, and 2007. The data set used is described in Section 2.

We were interested in the events with respect to the regions and days of the week. We wanted to see if there are some days when there are especially many or few events, if there is some rules like going to the same region on certain days, etc. We use the *event line* view with *y*-axis showing regions (there are seven strips: *Florida1* to *Florida5*, *Mexico*, and *Interception*) and days of a week (seven lines per strip, one for each day). Figure 4 shows the view.

From this view many interesting things can be observed. First thing is the visible increase of the number of events during time, and increase of number of landing spots. This could be due to increased activity of the immigrants or due to US Coast Guard becoming more effective (for example it could mean that the fleet became bigger). Other things are, for example, *Florida3* is active during all three years, *Florida4*, *Florida5* and *Mexico* became active in 2006 while *Florida1* and *Florida2* mostly became active in 2007. Looking at the week days we can notice that *Florida2* is almost never active during weekends and that *Florida5* has a big gap on *Monday*.

Since there are most of events in 2007, we will focus on this year only from now on. We want to analyze events in respect to a region and boat type now. We
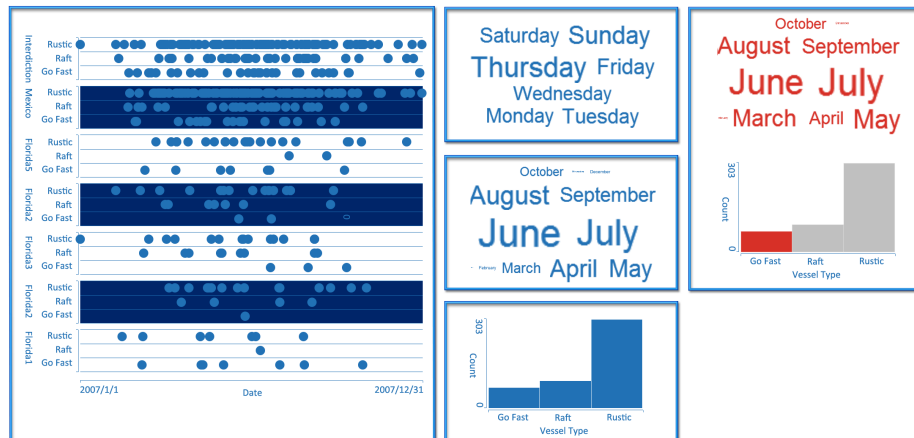
Fig. 5. Events in 2007. It is very obvious that there are no trends and patterns in boat type usage, but also that the number of boats of some type are not the same. While *Go Fast* and *Raft* types are very similar in numbers, *Rustic* type is used in much larger numbers. Very important thing is that there is much much less immigrant activity during fall and winter which is probably due to bad weather. This is also confirmed in the *tag cloud* view with months where *January*, *February*, *November*, and *December* (winter months) are barely visible. The *tag cloud* view with week days shows us that *Thursday* and *Sunday* are the most active days and there is a growing trend towards the end of the week. The histogram confirms that *Rustic* boat type is used most often. At the right we can see that *Go Fast* boat type is rarely used during winter (with the selected brush only option inthe *tag cloud* view, the winter months are barely visible or not visible at all).

are interested if migrants are using specific boats for specific routes, for example, or if there is any other pattern in the data. As we have multiple views we will use the *tag cloud* view to depict weekdays so to see if there are any preferable days in the week when the events occur, and another *tag cloud* view to depict months. We have also included a histogram showing frequency of usage for each boat type (Figure 5).

We were first interested if there any trends or patterns considering boat types being used, which could then be compared to trends in immigrants activities to see is there a specific boat type which is more effective than the others. We can see that *Rustic* vessel is used most often and *Go Fast* vessel the rarest.

It is interesting to see the distribution of events with specific boat and landing spots. *Go Fast* succeeded only once in reaching *Florida2*, e.g., and only twice in reaching *Florida4*. It reached *Mexico* quite often. The usage of *Go Fast* is not frequent at the end of the year. There were very few successful landings and very few successful interdictions in the last months. In order to confirm this finding we have brushed *Go Fast* vessels in the histogram and select option show brush only in the *tag cloud* view with months.

Fig. 6. Immigrant successfulness. The left view shows unsuccessful (context) as gray and successful (brushed) as red. *March*, *May* and *June* were most successful months for the immigrants. Right view shows same thing but with brush only option selected (so the context is not shown).

Figure 5 shows the *tag cloud* view and the brushed histogram on the far left. Note that *January*, *February*, *November*, and *December* (winter months) are barely visible in this *tag cloud* view.

The next task could be identifying months when the immigrants are most successful (most landings). We can visualize this using the *tag cloud* view of months with all successful landings selected (Figure 6). There are two views, one with context and one without. The red part represents successful landings. Immigrants are most successful during *March*, *May* and *June*.

Last thing we are interested in is how often did the events happen day after day in some region, which could also lead us to patterns. These cases can be identified in the *event line* view as dots in line with a slight slope to the right when looking from *Monday* to *Sunday* (Figure 7).

Florida4 has none, *Florida1*, *Florida2*, *Florida3* and *Florida5* have few, *Mexico* has many of those cases and *Interdiction* has few examples when there were events in every day of the week.

6   Summary and Conclusions

In this paper we have presented a new way of visualizing irregularly occurring events. We have introduced a new type of view, the *event line* view, and integrated it into a coordinated multiple views (CMV) system. The *event line* view supports events analysis based on two categorical attributes. Since we are mostly dealing with categorical data we also added the *tag cloud* view which is well suited for categorical data visualization. We also provided illustrative examples of usefulness of this new view type.

The *event line* view has been proven very useful in exploring data sets that motivated the development of the view. When combined with the *tag cloud* view and with the other well known views (histogram, scatter plot, parallel coordinates, etc.) in a coordinated multiple views (CMV) system, it provides
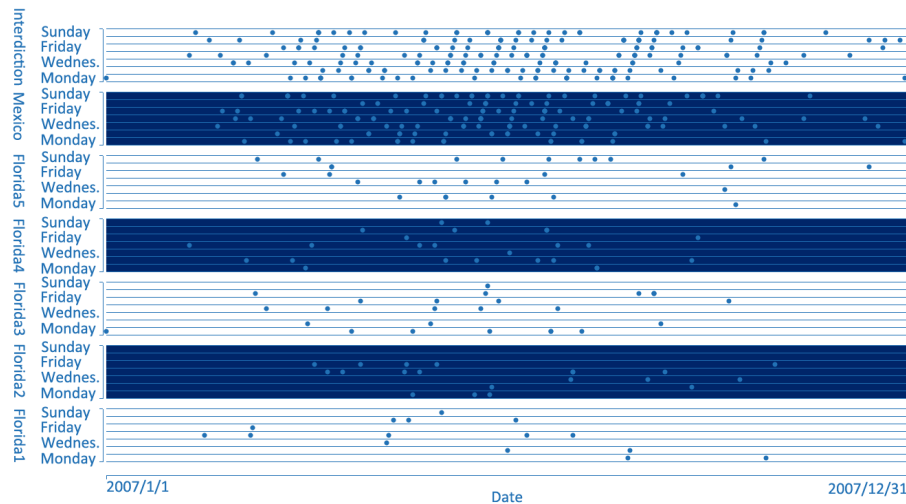
Fig. 7. Searching for events happening day after day in particular region. This can be identified in events line as dots in line with a slight slope to the right when looking from *Monday* to *Sunday*. We can see that one region (*Florida4*) has none of such cases, and one has many (*Mexico*). *Interdiction* has some examples where there were events in every day of the week. Again, there are no visible patterns and the number of these cases is proportional to the number of events in a region overall.

new analytical capabilities. These capabilities may enable new discoveries which were previously harder (or less intuitive) to do or even impossible.

7   Acknowledgments

References

1. W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.
2. P. Bausch, J. Bumgardner, and C. Fake. *Flickr Hacks: Tips & Tools for Sharing Photos Online (Hacks)*. O'Reilly Media, Inc., 2006.
3. J. V. Carlis and J. A. Konstan. Interactive visualization of serial periodic data. In *UIST '98: Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, pages 29–38. ACM Press, 1998.

4. VAST 2008 challenge. http://www.cs.umd.edu/hcil/VASTchallenge08.

5. B. Francis and J. Pritchard. Visualisation of historical events using Lexis pencils. Advisory Group on Computer Graphics, 1997.

6. S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large documents collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

7. H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

8. Z. Konyha, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual analysis of families of function graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1373–1385, 2006.

9. R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

10. S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In T. Gross, J. Gulliksen, P. Kotz, L. Oestreicher, P. A. Palanque, R. O. Prates, and M. Winckler, editors, *Proceedings of INTERACT (1)*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer, 2009.

11. R. Miklin, T. Lipić, Z. Konyha, M. Berić, W. Freiler, K. Matković, and D. Gračanin. Migrant boat mini challenge award: Simple and effective integrated display. In D. Ebert and T. Ertl, editors, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '08)*, pages 203–204, 19–24 Oct. 2008.

12. W. Müller and H. Schumann. Visualization for modeling and simulation: Visualization methods for time-dependent data - an overview. In *WSC '03: Proceedings of the 35th Conference on Winter Simulation*, volume 1, pages 737–745. Winter Simulation Conference, 2003.

13. J. C. Roberts. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In G. Andrienko, J. C. Roberts, and C. Weaver, editors, *Proc. of the 5th International Conference on Coordinated & Multiple Views in Exploratory Visualization*. IEEE CS Press, 2007.

14. M. Suntinger, H. Obweger, J. Schiefer, and M. E. Gröller. Event tunnel: Exploring event-driven business processes. *IEEE Computer Graphics and Applications*, 28(5):46–55, Sept. 2008.

15. J. J. van Wijk and E. R. van Selow. Cluster and calendar based visualization of time series data. In *INFOVIS '99: Proceedings of the 1999 IEEE Symposium on Information Visualization*, page 4, Washington, DC, USA, 1999. IEEE Computer Society.

16. M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In *Proc. of the IEEE Symp. on Information Visualization*, pages 7–13, 2001.

17. R. Wegenkittl, H. Löffelmann, and E. Gröller. Visualizing the behavior of higher dimensional dynamical systems. In *Proceedings of the IEEE Visualization (VIS '97)*, pages 119–125, 1997.