ANALYZING SEQUENTIAL DATA FROM GEOVISUALIZATION USER INTERFACES

Amy L. Griffin
School of Physical Environmental and Mathematical Sciences
University of New South Wales-ADFA
Canberra ACT 2600, Australia
a.griffin@adfa.edu.au

Abstract

User studies with geographic and other interactive information visualizations have the potential to produce datasets that contain sequences of user behaviors. One barrier to exploiting the information that these behavioral sequences contain is the difficulty of finding appropriate analytical methods for probing such datasets. This paper outlines some of the potential uses, advantages and disadvantages of two types of methodologies for analyzing sequential data: sequence alignment and sequence mining. Sequence alignment is essentially concerned with identifying regions of similarity between sequences and then grouping sequences according to a similarity score. This score is based upon the size and number of regions within the sequences that are identical. Sequence mining, on the other hand, while also concerned with finding interesting and relevant statistical patterns in sequences, focuses less on identifying groups of sequences and more on providing a wide range of sequence metrics that can be used to compare different sequences. Sequence alignment and sequence mining offer some advantages over simple visual analysis of sequences because they provide comparable quantitative measures of sequence similarity of geovisualization user behavioral patterns. A remaining challenge lies in determining which method is best used in a particular context.

Introduction

Although the last two decades have seen the development of many new methods and tools for geographic visualization, relatively few of these tools and methods are empirically evaluated, beyond simple usability testing. As a result we often don't really know whether new methods or tools are more effective or more efficient than old methods for a particular task, unless their use becomes so popular and widespread that this is the obvious conclusion. Moreover, we do not have a generalized sense for why a particular tool is more effective or more efficient at helping users to perform a particular task than another tool or method. This is particularly the case for tools that are designed to support higher-level thinking skills, such as hypothesis generation.

Perhaps one of the reasons for this relative lack of attention to why and how visualization tools work lies in the difficulty of designing experiments that will help to answer these questions, as well as the resources required, particularly the time needed,

to undertake such studies. In other cases, it may the lack of effective methods for answering a given research question. This paper reports on a trial of two methods for analyzing sequential data, sequence alignment and sequence mining, for studying the way in which geovisualization tools can support the process of hypothesis generation.

Data:

This trial of sequence alignment and mining methods uses a dataset gathered in a naturalistic experiment directed to understanding hypothesis generation when scientists used a spatially explicit simulation model of disease prevalence. The model included a number of interactive representations of the model's inputs and outputs, including linked maps and scatterplots, time series graphs, and a visual representation of the model's input parameters.

A key goal of the larger study was to try to generate a dataset that would help to uncover the ways in which interactivity facilitates visual thinking and knowledge construction (i.e., hypothesis generation). In particular, the study sought to answer two questions:

1. Do the representations a model user sees have an impact upon his or her conceptualization of the modeling problem?
2. Do the displays that the model user has seen in the past (which may have been influenced by his or her training) influence the types of representations s/he chooses for viewing the model results?

A series of simpler questions generated information that could help to answer these broader questions by documenting patterns of experts' use of the different data-display devices throughout the experiment:

- What are the patterns of use for different system components?
- What kinds of information to users attend to in the visual information display devices?
- How do participants obtain information from the system and how is this information used?
- What kinds of hypotheses are generated?

Each expert's (n = 17) participation in the experiment generated a set of behavioural sequences (or in the case of the last question, a sequence of hypotheses) that could then be analysed to answer each of these four questions. The full details of data collection and the experimental conditions can be found in Griffin (2004).

The focus of this paper is on the application of sequence alignment and mining methods to the second of these sequences – that is, on where and when experts directed their attention to particular types of data-display devices (e.g., maps, scatterplots or other

graphs) throughout the process of working with the simulation model. This sequence consisted of codes that represent different attention behaviours, identified from a transcript and video of the expert's model use session. Table 1 provides a full description of the different elements that could form a part of the sequence.

Code	**Attention behaviour**
Ia	One map, general pattern
Ib	Multiple maps, general pattern
Ic	One map, particular feature
Id	Multiple maps, particular feature
Ie	Time series graph, general trend
If	Time series graph, particular time
Ig	One scatterplot, general pattern
Ih	Multiple scatterplots, general pattern
Ii	One scatterplot, particular feature
Ij	Multiple scatterplots, particular feature
Ik	Model parameters graph, particular value

Table 1. Attention targets that could potentially appear in each attention sequence.

Methodology

Sequence alignment is a method that was developed in the field of bioinformatics for the purpose of analyzing DNA sequences. Since its development, the methodology has been adapted for use with social science and geovisualization sequences (e.g., Wilson 1998; Shoval and Isaacson 2007; Fabrikant et al. 2008), which differ from DNA sequences in the number of different elements they may contain. Sequence alignment is essentially concerned with identifying regions of similarity between sequences and then grouping sequences according to a similarity score. This score is based upon the size and number of regions within the sequences that are identical. In the context of this research problem, we might think of these common patterns as similar information display viewing patterns. One such behavior might be looking at the general pattern of a map to get an overview and then later focusing in on particular aspects of that pattern. For the purposes of this research, I have used a package called ClustalTXY, the latest version of ClustalG for sequence alignment (Wilson et al. 1998).

Sequence mining, on the other hand, while also concerned with finding interesting and relevant statistical patterns in sequences, focuses less on identifying groups of sequences and more on providing a wide range of sequence metrics that can be used to compare different sequences. In this research, I have used the TraMineR package, which has been developed as an add-on to R, the open-source statistical package (Gabhadino et al 2009).

Results

Before discovering sequence analysis methods, I originally undertook a simple visual analysis of attention patterns to identify groups of individuals with broadly similar visual attention behaviours. In this analysis I identified three groups: focused attention, attention-switching and mixed focus and attention-switching. These groups differed principally on the frequency with which they switched between one visual attention target (as coded using the elements in Table 1) and another. 'Focused attention' users switched attention targets at a low frequency, 'attention-switching' users switched targets at a high frequency, and 'mixed focus and attention-switching' users employed both behaviours at some point in their interactions with the simulation model (Figure 1).

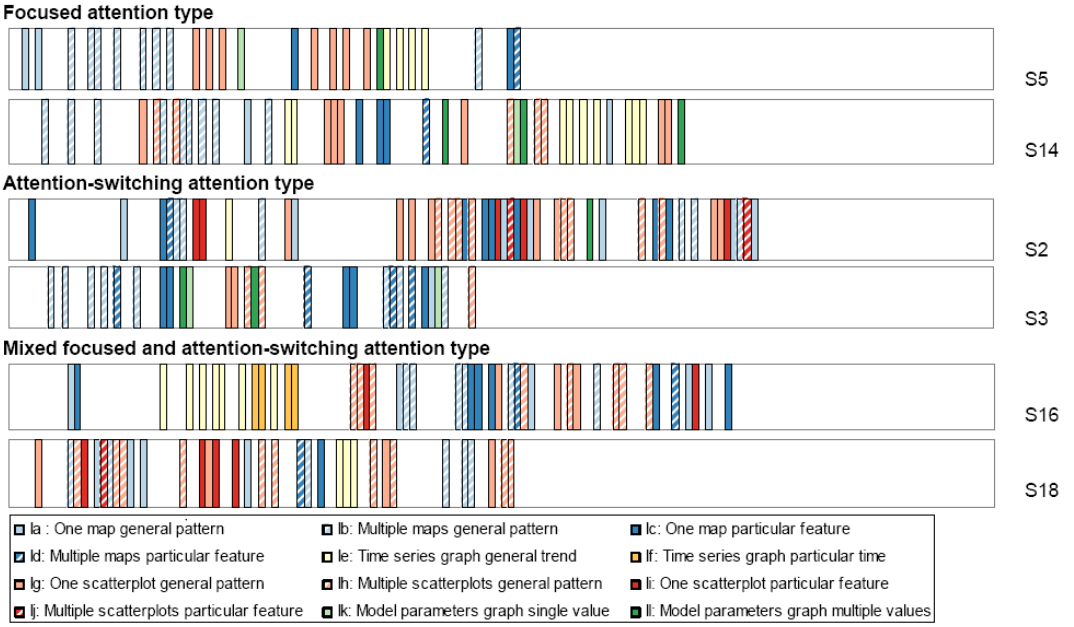

Figure 1. Visual attention behaviours identified through visual analysis and comparison of sequences.

The sequence alignment analysis also identified three groups (Figure 2). The left half of the figure, when compared with the cluster-tree at the right, shows that groups are comprised of individuals with similar attention patterns. For example, users S4, S8 and S15 relied more heavily on scatterplots than other representations and did not tend to flip from one device to another at high frequencies, while S18 (at the top), who also relied heavily on scatterplots and is in a different cluster, (on its own) does exhibit rapid attention-flipping. Comparing the visual analysis groups with the sequence alignment groups, we can see that there is some level of group membership similarity. However, sequence alignment takes into account both the frequency of attention targets and the order in which the user's attention was targeted at a particular aspect of a visual display rather than just the order of instances of visual attention.

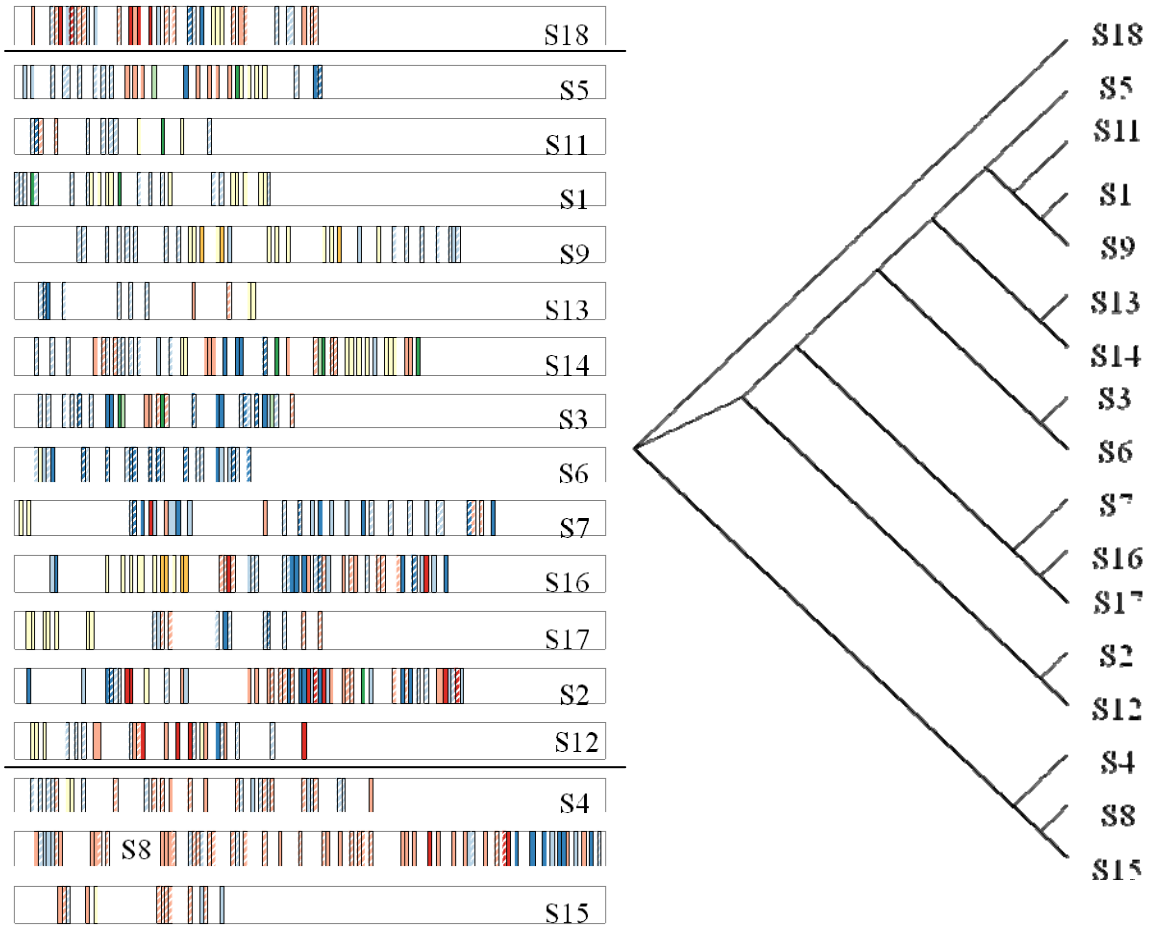

Figure 2. Groupings produced by the sequence alignment method.

Sequence alignment has the potential to provide a quantitative measure of similarity of behavioral patterns of geovisualization users. By explicitly accounting for both the frequency of a behavior and when in the sequence the behavior occurs, it can provide a more nuanced understanding of the strategies that geovisualization users take to accomplishing tasks. While the interpretation of cluster groupings becomes more difficult as the coding scheme contains a larger number of different types of actions, this method can be very useful for comparing relatively simple sequences.

Sequence mining provides us with a number of insightful metrics about when particular types of visual attention occurred. For example, from the state distribution plot (Figure 3), which looks at the frequency of when types of attention occurred at different points within the sequences, it is possible to see that Ib (Multiple maps, general pattern) occurs more frequently towards the beginning of the model use session, while Ig (One scatterplot, general pattern) is most frequent in the middle of model use sessions,

suggesting the users perhaps relied on map sequences to understand the general pattern of dynamics predicted by the model and then used a scatterplot to investigate a specific relationship that may have been noticed among the general patterns seen in the maps.

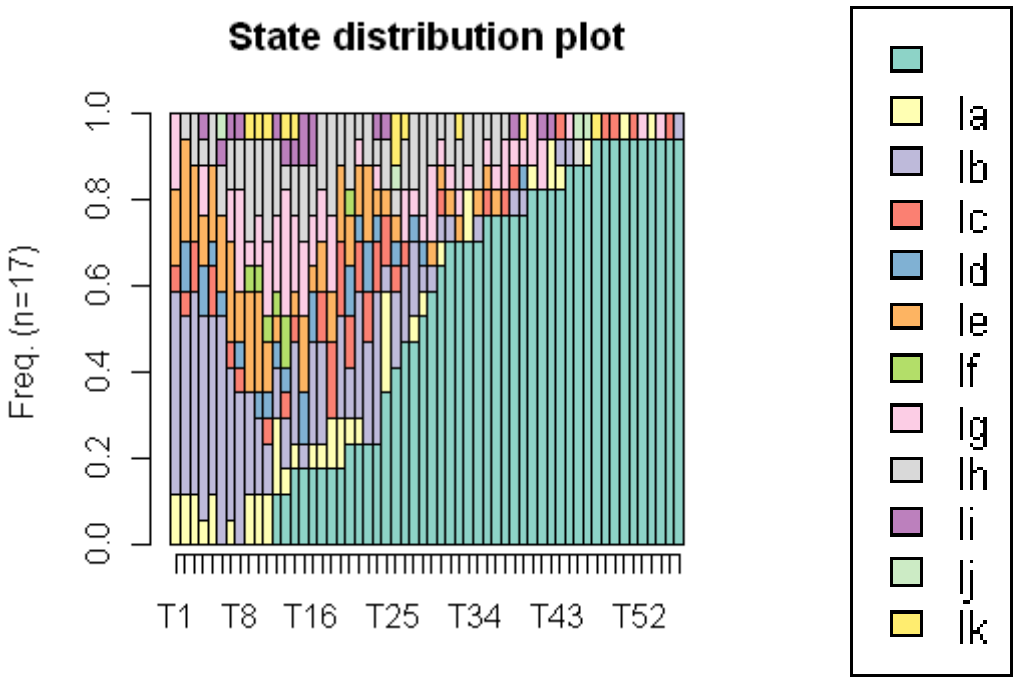

Figure 3. State distribution plot of visual attention targets.

Table 2 shows commonly occurring transitions between attention targets. These results suggest that there is a high degree of autocorrelation between attention targets. The two most common transitions in attention targets were not transitions at all – they were repeated instances of attention being directed to the same targets (Ie to Ie and Ib to Ib). The remaining frequently occurring transitions tended to be transitions between different aspects of a single display-device (e.g., maps) rather than between display-devices (maps and scatterplots).

From	**To**	**Percent of transitions from the first target**
Ie	Ie	63%
Ib	Ib	46%
Id	Ib	40%
Ih	Ib	37%
Ig	Ia	36%
If	Ie	29%
Ia	Ib	20%

Table 2. Most common attention target transitions. The percentage quoted is the percentage of transitions from the first target to any other target. That is, 63% of all transitions from Ie were to Ie. Hence, the percentages do not add to 100.

Finally, sequence mining also offers the ability to group similar sequences into groups. Furthermore, it generates both state distribution plots for each group as well as mean frequency diagrams for each group (Figure 4). This method placed fourteen sequences in group one, three sequences in group two (S2, S12 and S16), and one sequence in group three (S8). From these diagrams, it is apparent that the sequences in group one were much shorter than those in either groups two or three (the large teal bar), and generally lower amounts of time spent on each attention target, while group three is distinguished from the first two groups by its heavy reliance on scatterplots and low amount of attention to time series graphs.

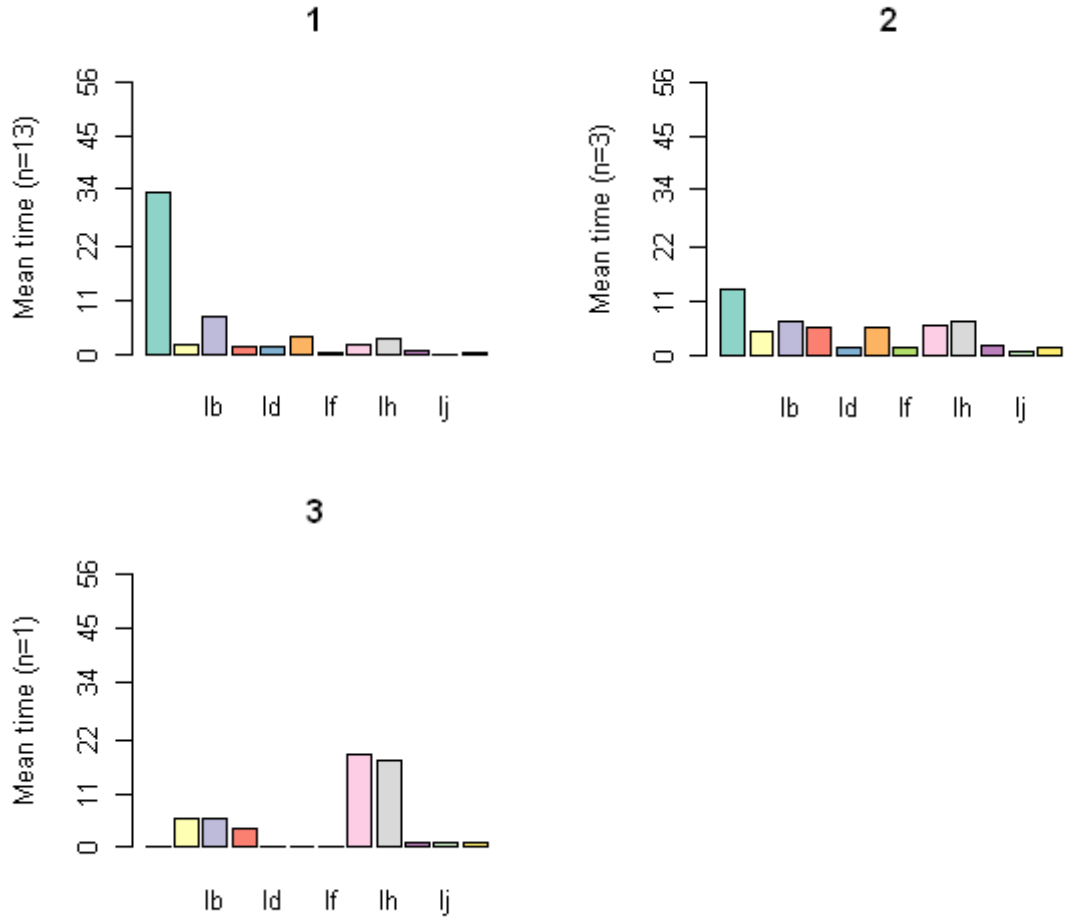

Figure 4. Mean frequencies of occurrences of attention directed to different targets in the three groups generated by the sequence mining analysis.

Conclusions

Sequence alignment and sequence mining offer some advantages over simple visual analysis of sequences because they provide comparable quantitative measures of sequence similarity of geovisualization user behavioral patterns. None of these methods, however, allows for the comparison of how multiple sequences covary with each other (e.g., for example how visual attention might covary with characteristics of hypotheses that have been generated by using a model). Furthermore, as the number of sequence states (i.e., visual attention targets in the example used here) grows larger, the interpretation of cluster grouping becomes more difficult. Nevertheless, these methods work relatively well for simple sequences and offers an improvement on simple visual inspection of sequences. A remaining challenge lies in determining which method is best used in a particular context.

References

FABRIKANT, S.I., REBICH-HESPANHA, S., ANDRIENKO, N., ANDRIENKO, G. AND D.R. MONTELLO. (2008). "A Novel Method to Measure Inference Affordance in Static Small-Multiple Map Displays Representing Dynamic Processes." The Cartographic Journal, 45(3): 201-15.

GABADINHO, A., RITSCHARD, G., STUDER, M., AND N. S. MÜLLER. (2009). Mining sequence data in R with the TraMineR package: A user's guide for version 1.21. University of Geneva, 2009. http://mephisto.unige.ch/pub/TraMineR/Doc/1.2/TraMineR-1.2-Users-Guide.pdf. Last accessed 29 July 2009.

GRIFFIN, A. L. (2004). Understanding how scientists use data-display devices for interactive visual computing with geographical models. Unpublished PhD thesis, Department of Geography, The Pennsylvania State University.

SHOVAL, N. AND M. ISAACSON. (2007). "Sequence Alignment as a Method for Human Activity Analysis in Space and Time." Annals of the Association of American Geographers, 97(2): 282-97.

WILSON, C. (1998). "Activity pattern analysis by means of sequence alignment methods." Environment and Planning A, 30: 1017-38.

WILSON, C., HARVEY, A., AND J. THOMPSON. (1999). ClustalG: Software for analysis of activities and sequential events. http://www.ssc.uwo.ca/sociology/longitudinal/wilson.pdf. Last accessed 29 July 2009.