Identifying spatial patterns on choropleth maps: A comparison between humans and deep learning models

1 Study Objectives

Maps are an effective means of presenting and communicating information on space (Crampton, 2001). Humans have long developed map reading skills (Presson, 1982; Gilhooly et al., 1988) to comprehend information from the map naturally. Among many things, spatial patterns such as how the phenomenon represented on the map concentrates or spread on space are especially suitable to be depicted on the map and therefore for humans to understand (Slocum et al., 2009; Kimerling et al., 2016). Because maps are artifacts made by humans and, more importantly, *for* humans to read, reading maps has been challenging for traditional computer algorithms, if not impossible. In recent years, however, dramatic progress has been made in computer vision. Thanks to the advances in deep neural networks (Goodfellow et al., 2016), methods in computer vision have been successfully applied to problems in various domains where images need to be recognized or classified (Szeliski, 2022). For example, in facial recognition, deep learning models are applied on a face photograph database designed for studying the problem of unconstrained face recognition and have achieved an accuracy that approaches or is even beyond humans (Wang and Deng, 2021). Our own preliminary research (Li and Xiao, 2019) has also shown that deep learning models can be used to successfully identify maps and map elements such as areal symbols, titles, and legends. While we can agree that artificial intelligence may not have reached a point that machines can read maps as humans do, would it be possible that computers can recognize some contents of maps utilizing the cutting-edge algorithms in computer vision and then do some analysis and therefore can recognize spatial patterns presented on the map?

The overall goal of this study is to explore how well artificial intelligence algorithms can compare with humans in identifying spatial patterns in a specific type of map, choropleth maps. Popular in a wide range of applications, choropleth maps are used to display the spatial distribution of quantitative data across the enumeration units such as counties and states in a region. A common approach to making this type of map is to group the data into a number of classes and then each class is assigned a color (see Figure 1 for an example). In other words, each unit is rendered using areal symbol (i.e., the shape of the unit filled with a color). These symbols hold the key to understand the spatial pattern of the data. In the example illustrated in Figure 1 that will be used in this study, one should recognize a unit with a certain value on the map tends to associate with units with a similar value because similar colors tend to be close to each other. One of the major goal of representing the data using a choropleth map is to allow readers to recognize such patterns. In the past five years, we have been developing deep learning to identify symbols from choropleth maps and the results can be used to analyze spatial patterns on the map using existing algorithms (Li and Xiao, 2019; Li, 2021). It is possible to assemble computer programs that can be used to automatically identify the spatial patterns on the map, and we are interested how such computer programs can compare or compete with humans.

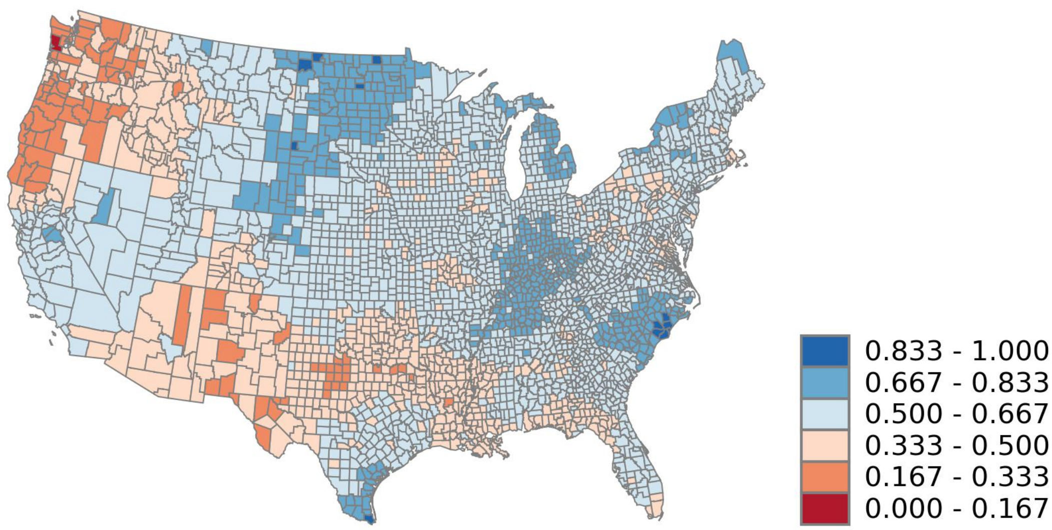

Figure 1: An example map that will be used in this study.

In order to conduct the comparative study, it is critical to understand how humans perform on recognizing spatial patterns on the map when the exact same map will also be processed by computer programs. On the computer side, deep learning-based object detection models in computer vision will be used to detect the mapping area of a choropleth map (LeCun et al., 2015), and then algorithms in traditional computer vision (Szeliski, 2022), data mining (Han et al., 2001), and spatial analysis (Xiao, 2016) will be applied to extract information from the mapping area and analyze its spatial patterns. On the human side, we plan to conduct surveys that use the same set of maps to ask our subjects if they can recognize the patterns presented on the maps. In this protocol, we detail the design and procedures of the surveys. We do not need and will not record personal identifiable information in this study.

2 Research design

We will conduct surveys that ask participants questions about spatial patterns on a series of choropleth maps. For a given map, we focus on two kinds of spatial patterns that can be observed from the map by asking each participants two questions. The first question asks about whether the symbols are concentrated:

Is the phenomenon represented by the visual symbols (colors) concentrated in some area?

- Yes
- No

And the second question asks about whether values on the map are spatially associated:

Based on your observation, the values on the map tend to occur...
- Near the values that are similar to them.
- Near the values that are different from them.
- Without clear associate with similar or different values.

While there are different kinds of spatial patterns, these two types (concentration and association) illustrate meaningful and typical information that can be observed on a choropleth map. In addition, we can develop computer programs to answer exactly the same questions. This will provide a common basis for our comparative study on spatial pattern recognition using choropleth maps between humans and computer algorithms.

We will design and generate all the maps for the survey. Each map will have only two visible elements: the mapped area where the spatial units are rendered, and a legend. In other words, the maps used for our surveys will be different from the ”normal” maps that would have a title and other elements. Because our study will focus on how humans and computer programs can use symbols (colors, in this specific case) to recognize spatial patterns on choropleth maps, the way the maps are designed will make sure the symbols are the focus during the survey. Our computer program will be built specifically to recognize and analyze the mapped area and legend. Our maps will vary in the following aspects:

- **Base map.** We use two sets of spatial units to provide different levels of cognition challenges. The first is based on the 88 counties in Ohio, and the second are the 3,109 counties for the conterminous United States.
- **Data.** The spatial data used in the maps presents different spatial patterns for our subjects to recognize. We synthesize spatial data for the two base maps with four types of spatial autocorrelation: negative spatial autocorrelation, no spatial autocorrelation, low positive spatial autocorrelation, and high positive spatial autocorrelation.
- **Number of classes.** In general, a choropleth map with a high number of classes may provide more detailed information while making it difficult for us to read. On the other hand, it is reasonable to assume that computers would not have a problem with the number of classes, but identifying the colors on the map may be more of a challenge for the computer. The cartography literature generally suggests a relatively small number of classes between 4 and 8 (Dent, 1996). We will use three different numbers of classes: 4, 6, and 8 to render each map.
- **Color scheme.** The set of colors used for the classes on a map is called a color scheme. Color schemes vary based on what kind of colors are used and how they differ. We will use two color schemes: a sequential color scheme where color range from light blue to dark blue, and a red-blue diverging color scheme where dark blue and dark red on both ends progress to light blue and light red, respectively, toward the middle of the scheme (see, for example, Figure 1). These are all colorblind safe schemes as suggested by ColorBrewer[1], the defacto color standards for cartographers and beyond (Harrower and Brewer, 2003).

In sum, each map will be defined by its base map (Ohio or the U.S.), color scheme (sequential or diverging), number of classes (4, 6, or 8), and spatial pattern (negative, random,

[1]https://colorbrewer2.org/

small positive, or large positive associations). We will make two copies of each combination (the data are randomly generated). Therefore, we have a total of 96 maps, or 2 (base maps) x 2 (color scheme types) x 3 (numbers of colors) x 4 (spatial patterns) x 2. In the survey, we will allocate 16 map images with two questions for each to one participant. It will take one participant about one minute for questions attached to one map. They will spend about 20 minutes on the survey totally. The answers from our subjects will be collected using Qualtrics survey reports, which will then be analyzed (see 5.2).

3 Sample

We plan to recruit approximately 60 adult (age 18-65) participants from Ohio state students, faculty, and staff members to answer the questions in our survey. For the student recruits, the survey will be distributed to two undergraduate classes in the Department of Geography: Design and Implementation of GIS (GEOG 5223) and GeoVisualization (GEOG 5201). Students are not required to take this survey. Those who do take the survey will receive bonus participation points. In addition, we will use online bulletin boards, email mailing lists, in-class announcements, and personal contacts to reach out to potential participants.

4 Measurement and Instrumentation

We will use Qualtrics to conduct our surveys. Each participant will need a computer or a mobile device (phone or tablet). All experiments will be done online through using a web browser. The survey can start at any time or at any place without log-in requirements. Data will be saved on the server of Qualtrics, which can be retrieved for further analysis. All the data collected here is anonymous, and no personal identifiable information will be recorded. The surveys are conducted on the platform of Qualtrics, and the choropleth map images are generated using the Python programming language.

For each question, the participants will be asked to select one from the choices. The answers will be saved in a survey report of the Qualtrics platform. Then, the accuracy rates for the questions of maps will be calculated. The accuracy rates will be compared with rates by computers.

5 Procedures

Our survey contains four sections: informed consent form, instruction, background information, the two question, and a random completion code generated at the end for the undergraduate students from the two geography classes.

- The informed consent form explains the the goal of the study, procedures and tasks, risks and benefits, confidentiality, and participant rights. Our participants will have a basic understanding of this project, then they can decide to continue the survey or not.

- The instruction section introduces what the participants will see and what they should do during the survey. Some examples of maps with the two questions are included in the instruction. After this section, the participant should know what to do during the survey.
- In the section of background information, the participant will be asked to answer their background in geography and cartography, which will be used for experiment control purposes. We will also ask if the participant is taking the survey for bonus credit for their class.
- In the question section, the participant will be presented a series of choropleth maps with different spatial patterns and be asked the two questions about spatial pattern for each map. They should read the maps carefully and answer the questions based on what they read and learn from the mapping areas. Each participant will be asked to view 16 maps.
- At last, a random completion code will be presented to students in the two geography classes (GEOG 5223 and GEOG 5201) mentioned above. The completion code will be used as the proof of completion for bonus points. The code is randomly pulled from a pool of codes that are prepared beforehand. The codes do not contain any personal information and are not associated with the participants in any manner. In other words, given a code we cannot trace back to the survey and questions answered by the participant, or vice versa. In this way, personal information of the participants will not be identified.

5.1 Internal validity

The surveys are designed to measure the ability of participants in identifying spatial patterns. The two questions about spatial patterns of the phenomenon presented on a map are non-trivial. They represent the reason why maps are useful: to learn how things distribute on space. These questions require participants to read the symbols on the map carefully to recognize the patterns.

The experiment will not discriminate against any population and will be open to all on a voluntary basis. We use colorblind-safe color schemes as suggested by extensive research behind colorbrewer (Harrower and Brewer, 2003). This experiment utilizes a forced choice response in which the correct choices in the experiments are predefined by the researchers. We will also make slides with instructions to introduce the experiments and how the Qualtrics survey will work. By using the instructions, we expect the participants will be familiar with the general process of the experiments and then concentrate on the experiments when they start.

5.2 Data handling analysis

Data would be collected by asking participants to read the maps and answer the questions. Personal identifiable information such as name and personal email will not be collected. The random completion code is generated by an online random code generator[2] and this code does not contain any personal information. Thus, subjects are anonymous during the whole research process and subject confidentiality would therefore be ensured throughout the whole experiment.

[2] https://www.randomcodegenerator.com/en/home

The average accuracy rate of the responses different choropleth maps will be used to reveal the impacts of the map design variables (e.g., base maps and color schemes) on the reading performance of our subjects. We will also develop statistical models to understand the relationship between the human reading performance on choropleth maps and the map design variables. We will run our computer programs to answer the exact same questions. As a final step of this study, we will compare the human responses with those from the computers. The results of this comparison will shed light on how close today's computer based algorithms can compete with humans in reading spatial patterns on choropleth maps.

References

Crampton, J. W. (2001). Maps as social constructions: power, communication and visualization. *Progress in human Geography 25*(2), 235–252.

Dent, B. D. (1996). *Cartography: Thematic Map Design* (5th ed.). Boston, MA: WCB/McGraw-Hill.

Gilhooly, K. J., M. Wood, P. R. Kinnear, and C. Green (1988). Skill in map reading and memory for maps. *The Quarterly Journal of Experimental Psychology Section A 40*(1), 87–107.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Han, J., M. Kamber, and A. K. H. Tung (2001). Spatial clustering methods in data mining: a survey. In H. Miller and J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*, pp. 188–217. London: Taylor and Francis.

Harrower, M. and C. A. Brewer (2003). Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal 40*(1), 27–37.

Kimerling, A. J., P. C. Muehrcke, J. O. Muehrcke, and P. Muehrcke (2016). *Map use: reading, analysis, interpretation*. ESRI Press Academic.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature 521*(7553), 436–444.

Li, J. (2021). Cartographic Recognition to Understand Choropleth Maps Semantically: a Case Study for State-level Choropleth Maps of the United States. Master's thesis, Department of Computer Science and Engineering, Ohio State University, Columbus, OH.

Li, J. and N. Xiao (2019). Using machine learning methods to identify and classify the regions and projections of online maps. In *Proceedings of GeoComputation 2019*. The University of Auckland.

Presson, C. C. (1982). The development of map-reading skills. *Child Development 53*(1), 196–199.

Slocum, T. A., R. B. McMaster, F. C. Kessler, and H. H. Howard (2009). *Thematic cartography and geovisualization* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Szeliski, R. (2022). *Computer Vision: Algorithms and Applications* (2nd ed.). London: Springer.

Wang, M. and W. Deng (2021). Deep Face Recognition: A Survey. *Neurocomputing 429*, 215–244.

Xiao, N. (2016). *GIS Algorithms*. London: SAGE Publications.