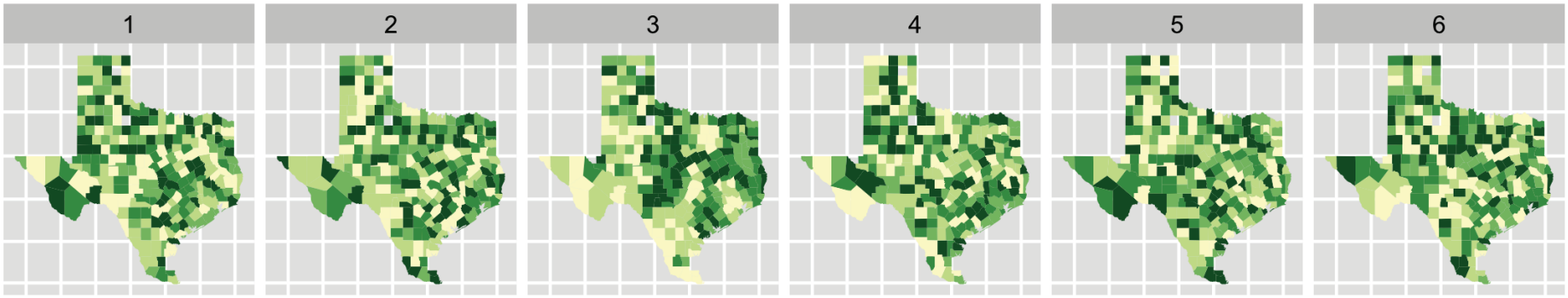


Map Line-ups: using graphical inference to study spatial structure

Roger Beecham, Jason Dykes, Aidan Slingsby, Cagatay Turkay, Jo Wood, giCentre, City University London

Graphical Inference for Infovis

Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja

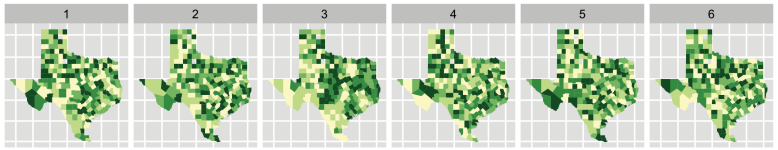


Fig. 1. One of these plots doesn't belong. These six plots show choropleth maps of cancer deaths in Texas, where darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence? If so, you've provided formal statistical evidence that deaths from cancer have spatial dependence. See Section 8 for the answer.

Abstract— How do we know if what we see is really there? When visualizing data, how do we avoid falling into the trap of apophenia where we see patterns in random noise? Traditionally, infovis has been concerned with discovering new relationships, and statistics with preventing spurious relationships from being reported. We pull these opposing poles closer with two new techniques for rigorous statistical inference of visual discoveries. The "Rorschach" helps the analyst calibrate their understanding of uncertainty and the "line-up" provides a protocol for assessing the significance of visual discoveries, protecting against the discovery of spurious structure.

Index Terms—Statistics, visual testing, permutation tests, null hypotheses, data plots.

1 Introduction

What is the role of statistics in infovis? In this paper we try and answer that question by framing the answer as a compromise between curiosity and skepticism. Infovis provides tools to uncover new relationships, tools of curiosity, and much research in infovis focuses on making the chance of finding relationships as high as possible. On the other hand, most statistical methods provide tools to check whether a relationship really exists: they are tools of skepticism. Most statistics research focuses on making sure to minimize the chance of finding a relationship that does not exist. Neither extreme is good: unfettered curiosity results in findings that disappear when others attempt to verify them, while rampant skepticism prevents anything new from being discovered.

Graphical inference bridges these two conflicting drives to provide a tool for skepticism that can be applied in a curiosity-driven context. It allows us to uncover new findings, while controlling for apophenia, the innate human ability to see pattern in noise. Graphical inference helps us answer the question "Is what we see really there?"

The supporting statistical concepts of graphical inference are developed in [1]. This paper motivates the use of these methods for infovis and shows how they can be used with common graphics to provide users with a toolkit to avoid false positives. Heuristic formulations of these methods have been in use for some time. An early precursor is [2], who evaluated new models for galaxy distribution by generating samples from those models and comparing them to the photographic plates of actual galaxies. This was a particularly impressive achievement for its time: models had to be simulated based on tables of random values and plots drawn by hand. As personal computers became available, such examples became more common.[3] compared computer generated Mondrian paintings with paintings by the true artist, [4] provides 40 pages of null plots, [5] cautions against over-interpreting random visual stimuli, and [6] recommends overlaying normal probability plots with lines generated from random samples of the data. The early visualization system Dataviewer [7] implemented some of these ideas.

The structure of our paper is as follows. Section 2 revises the basics of statistical inference and shows how they can be adapted to work visually. Section 3 describes the two protocols of graphical inference, the Rorschach and the line-up, that we have developed so far. Section 4 discusses selected visualizations in terms of their purpose and associated null distributions. The selection includes some traditional statistical graphics and popular information visualization methods. Section 5 briefly discusses the power of these graphical tests. Section 8 tells you which panel is the real one for all the graphics, and gives you some hints to help you see why. Section 7 summarizes the paper, suggests directions for further research, and briefly discusses some of the ethical implications.

- *Hadley Wickham is an Assistant Professor of Statistics at Rice University, Email: hadley@rice.edu.*
- *Dianne Cook is a Full Professor of Statistics at Iowa State University.*
- *Heike Hofmann is an Associate Professor of Statistics at Iowa State University.*
- *Andreas Buja is the Liem Sioe Liong/First Pacific Company Professor of Statistics in The Wharton School at the University of Pennsylvania.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

2 What is inference and why do we need it?

The goal of many statistical methods is to perform inference, to draw conclusions about the population that the data sample came from. This is why statistics is useful: we don't want our conclusions to apply only to a convenient sample of undergraduates, but to a large fraction of humanity. There are two components to statistical inference: testing (is there a difference?) and estimation (how big is the difference?). In this paper we focus on testing. For graphics, we want to address the question *"Is what we see really there?"* More precisely, is what we see in a plot of the sample an accurate reflection of the entire population? The rest of this section shows how to answer this question by providing a short refresher of statistical hypothesis testing, and describes how testing can be adapted to work visually instead of numerically.

Hypothesis testing is perhaps best understood with an analogy to

3 Protocols of graphical inference

This section introduces two new rigorous protocols for graphical inference: the "Rorschach" and the "line-up". The Rorschach is a calibrator, helping the analyst become accustomed to the vagaries of random data, while the line-up provides a simple inferential process to produce a valid p-value for a data plot. We describe the protocols and show examples of how they can be used, and refer the reader to [1] for more detail.

3.1 Rorschach

The Rorschach protocol is named after the Rorschach test, in which subjects interpret abstract ink blots. The purpose is similar: readers are asked to report what they see in null plots. We use this protocol to calibrate our vision to the natural variability in plots in which the data is generated from scenarios consistent with the null hypothesis. Our intuition about variability is often bad, and this protocol allows us to reduce our sensitivity to structure due purely to random variability.

Figure 4 illustrates the Rorschach protocol. These nine histograms summarize the accuracy at which 500 participants perform nine tasks. What do you see? Does it look like the distribution of accuracies is the same for all of the tasks? How many of the histograms show an interesting pattern? Take a moment to study these plots before you continue reading.

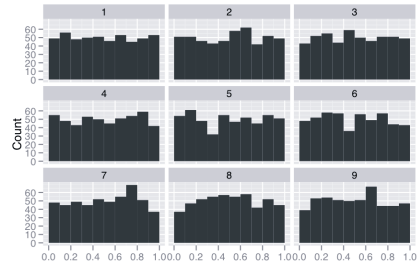


Fig. 4. Nine histograms summarizing the accuracy at which 500 participants perform nine tasks. What do you see?

It is easy to tell stories about this data: in task 7 accuracy peaks around 70% and drops off; in task 5, few people are 20-30% accurate; in task 9, many people are 60-70% accurate. But these stories are all misleading. It may come as a surprise, but these results are all simulations from a uniform distribution, that is, the distribution of accuracy for all tasks is uniform between 0 and 1. When we display a histogram of uniform noise, our expectation is that it should be flat. We do not expect it to be perfectly flat (because we know it should be a little different every time), but our intuition substantially underestimates the true variability in heights from one bar to the next. It is fairly simple to work out the expected variability algebraically (using a normal approximation): with 10 observations per bin, the bins will have a standard error of 30%, with 100 observations 19% and 1000, observations 6%. However, working through the math does not give the visceral effect of seeing plots of null data.

To perform the Rorschach protocol an administrator produces null plots, shows them to the analyst, and asks them what they see. To keep the analyst on their toes and avoid the complacency that may arise if they know all plots are null plots [8] the administrator might slip in a plot of the real data. For similar reasons, airport x-ray scanners randomly insert pictures of bags containing guns, knives or bombs. Typically, the administrator and participant will be different people, and neither should know what the real data looks like (a double-blinded scenario). However, with careful handling, it is possible to self-administer such a test, particularly with appropriate software support, as described in Section 6.

Even when not administrated in a rigorous manner, this protocol is still useful as a self-teaching tool to help learn which random features we might spuriously identify. It is particularly useful when teaching data analysis, as an important characteristic of a good analyst is their ability to discriminate signal from noise.

3.2 Line-up

The SJS convicts based on difference between the accused and a set of known innocents. Traditionally the similarity is measured numerically, and the set of known innocents are described by a probability distribution. The line-up protocol adapts this to work visually: an impartial observer is used to measure similarity with a small set of innocents.

The line-up protocol works like a police line-up: the suspect (test statistic plot) is hidden in a set of decoys. If the observer, who has not seen the suspect, can pick it out as being noticeably different, there is evidence that it is not innocent. Note that the converse does not apply in the SJS: failing to pick the suspect out does not provide evidence they are innocent. This is related to the convoluted phraseology of statistics: we "fail to reject the null" rather than "accepting the alternative".

The basic protocol of the line up is simple:

- Generate $n-1$ decoys (null data sets).
- Make plots of the decoys, and randomly position a plot of the true data.
- Show to an impartial observer. Can they spot the real data?

In practice, we would typically set $n = 19$, so that if the accused is innocent, the probability of picking the accused by chance is $1/20 = 0.05$, the traditional boundary for statistical significance. Comparing 20 plots is also reasonably feasible for a human observer. (The use of smaller numbers of n in this paper is purely for brevity.) More plots would yield a smaller p-value, but this needs to be weighed against increased viewer fatigue. Another way of generating more precise p-values is to use a jury instead of a judge. If we recruit K jurors and k of them spot the real data, then the combined p-value is $P(X \leq k)$, where X has a binomial distribution $B(K, p = 1/20)$. It can be as small as 0.05^K if all jurors spot the real data ($k = K$).

Like the Rorschach, we want the experiment to be double-blind - neither the person showing the plots or the person seeing them should know which is the true plot. The protocol can be self-administered, provided that it is the first time you've seen the data. After a first viewing of the data, a test might still be useful, but it will not be inferentially valid because you are likely to have learned some of the features of the data set and are more likely to recognize it. To maintain inferential validity once you have seen the data, you need to recruit an independent observer.

The following section shows some examples of the line-up in use, with some discussion of how to identify the appropriate null hypothesis for a specific type of plot and figure out a method of generating samples from the appropriate null distribution.

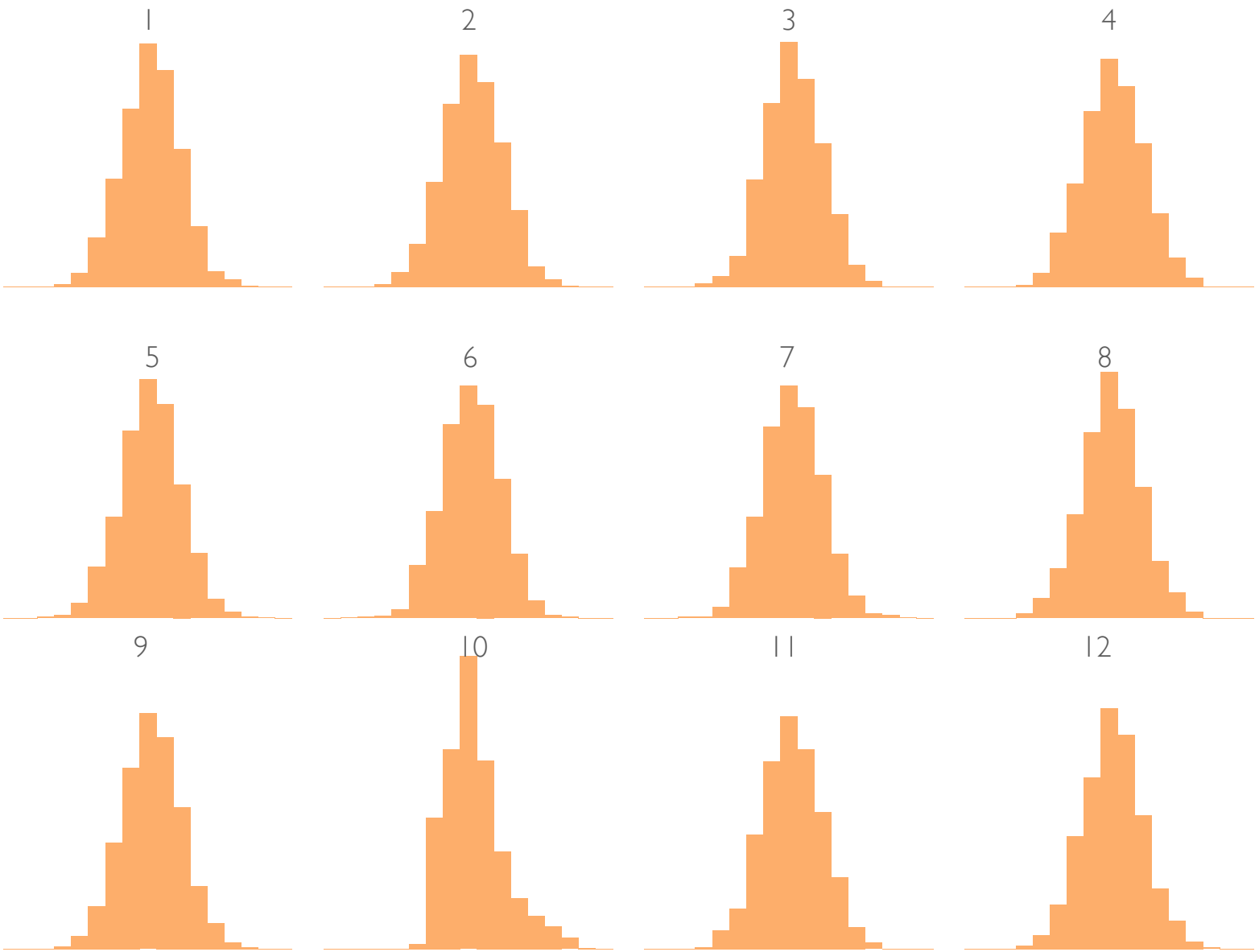
4 Examples

To use the line-up protocol, we need to:

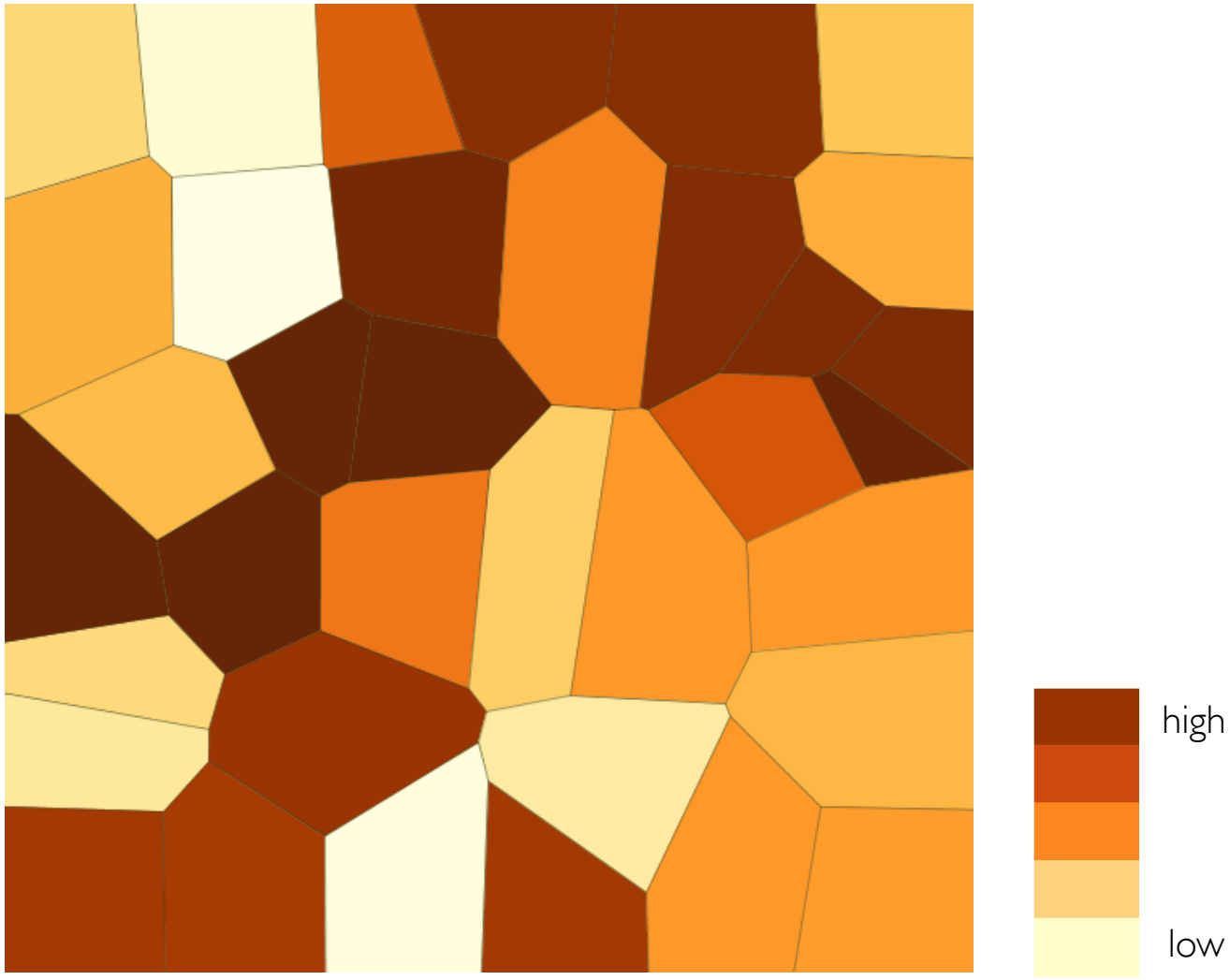
- Identify the question the plot is trying to answer.
- Characterize the null-hypothesis (the position of the defense).
- Figure out how to generate null datasets.

This section shows how to approach each of these tasks, and then demonstrates the process in more detail for two examples. Section 4.1 shows a line-up of a tag cloud used to explore the frequency distribution of words in Darwin's "Origin of Species" and Section 4.2 shows a line-up of a scatterplot used to explore the spatial distribution of three point throws in basketball.

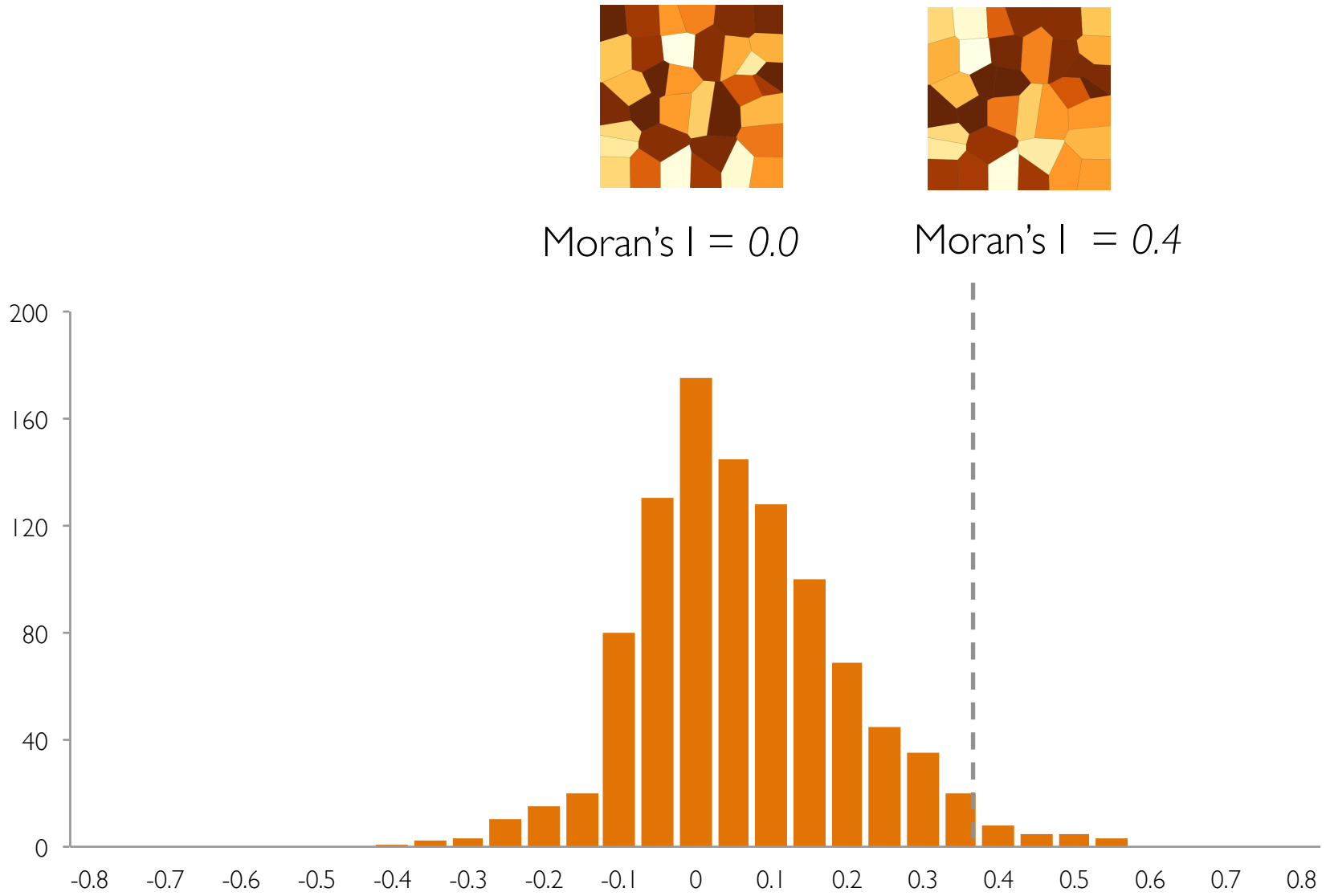
Line-ups

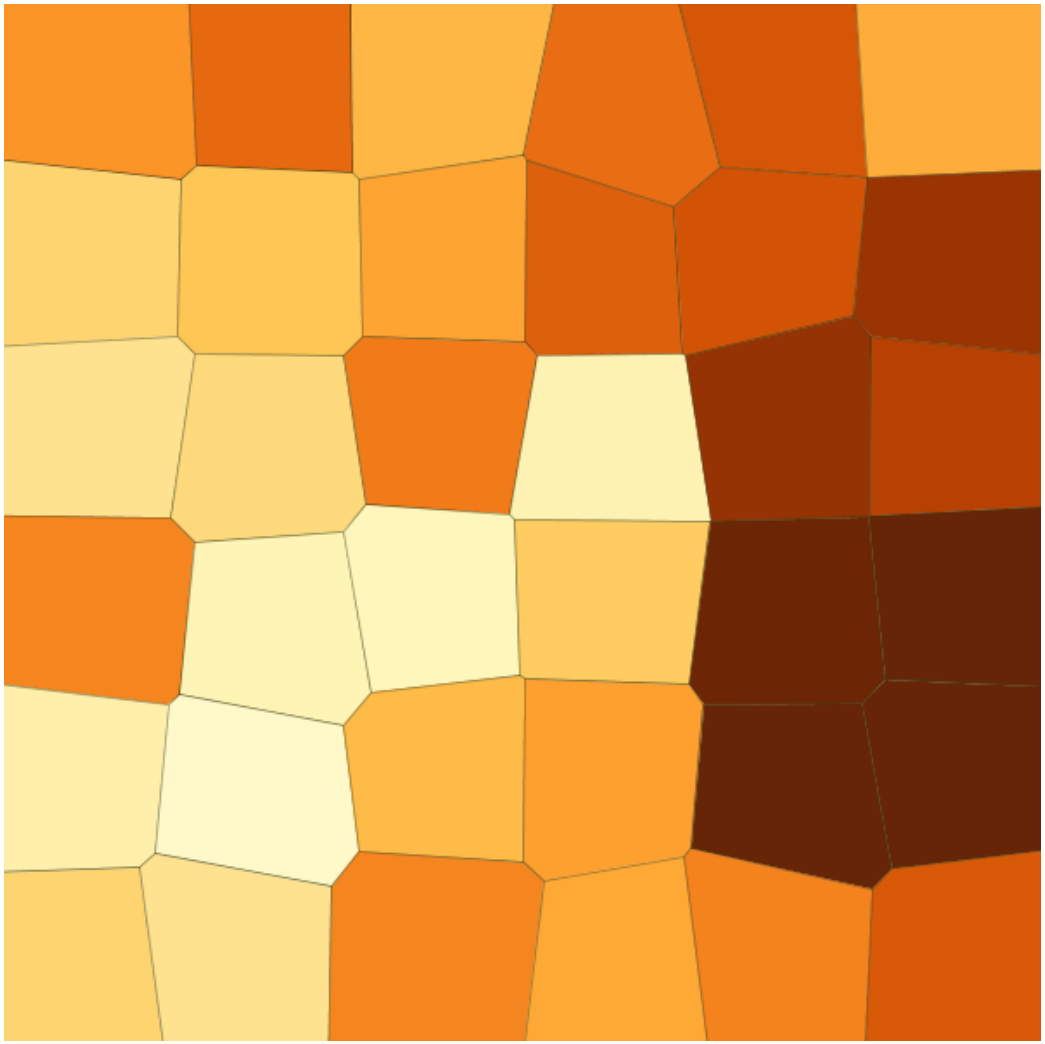

1
2
3
4
5
6
7
8
9
10
11
12

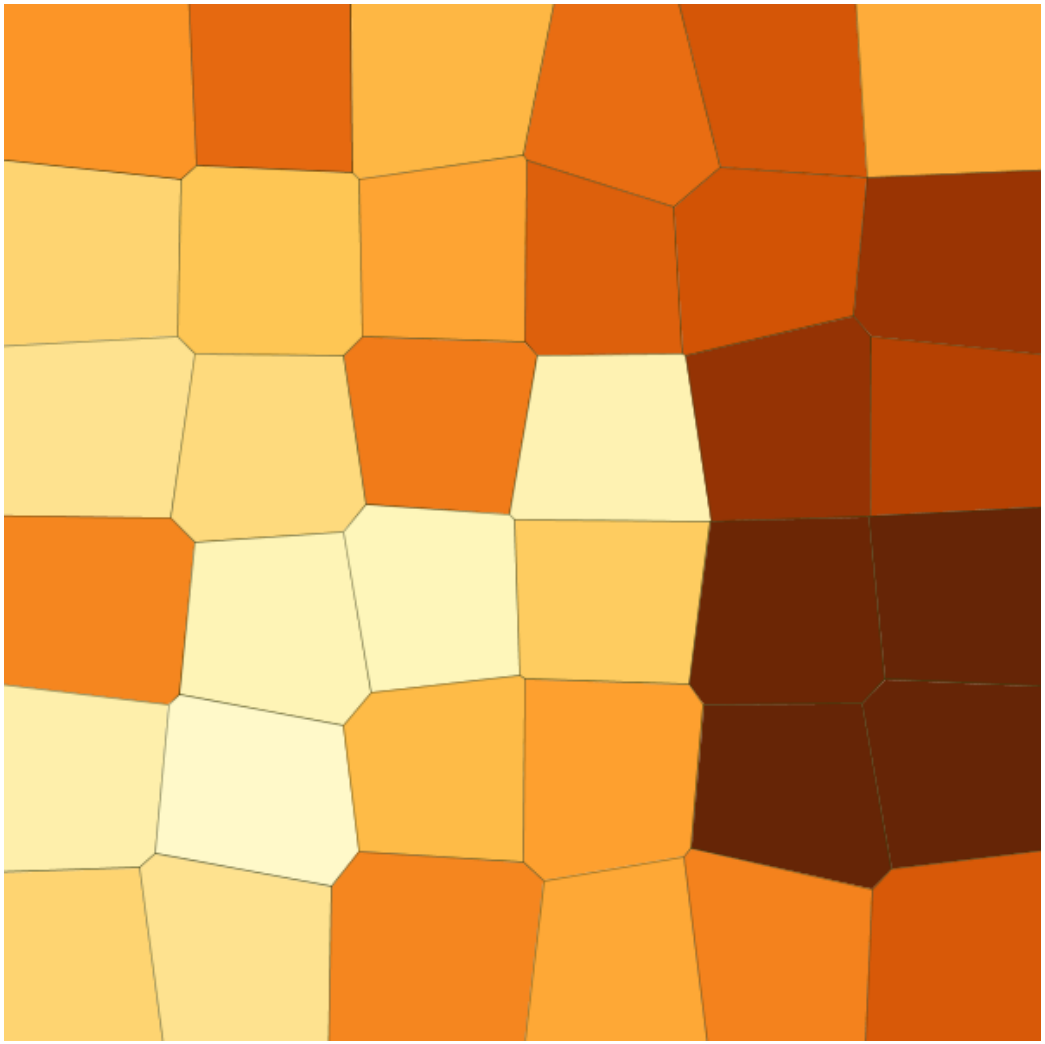
Map Line-ups


high
low

Moran's I = *0.4*

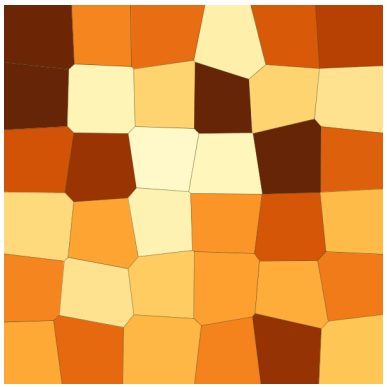

Moran's I = 0.0
Moran's I = 0.4
200
160
120
80
40
0
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8



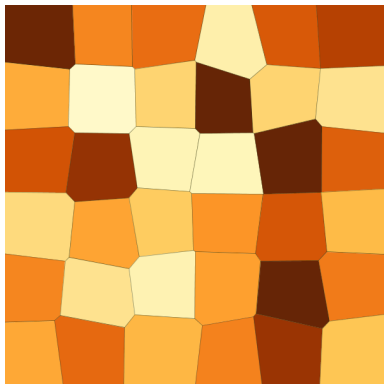


Moran's I *0.9*

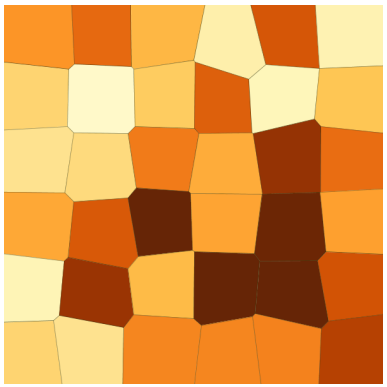
1



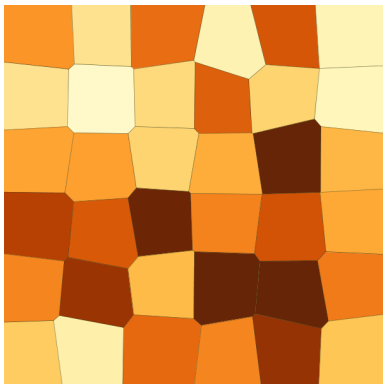
2



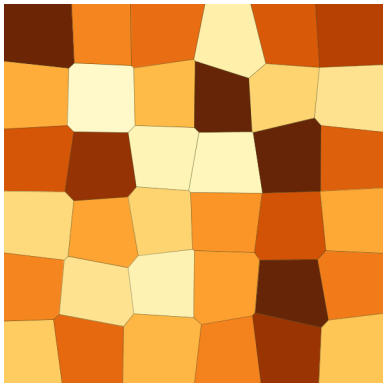
3



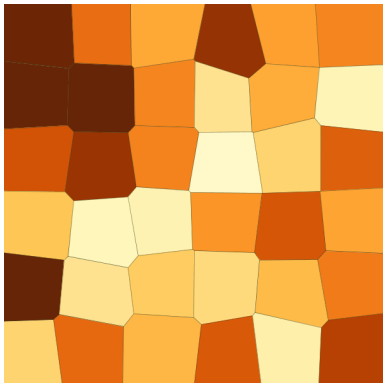
4



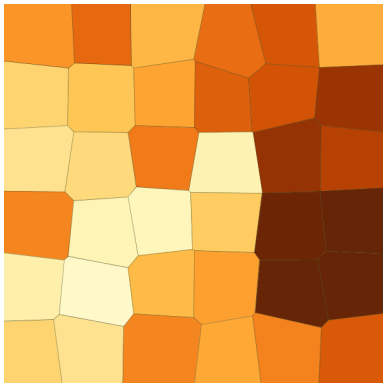
5



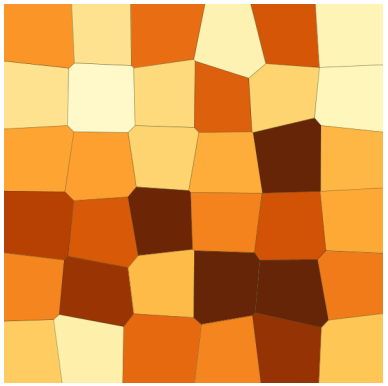
6

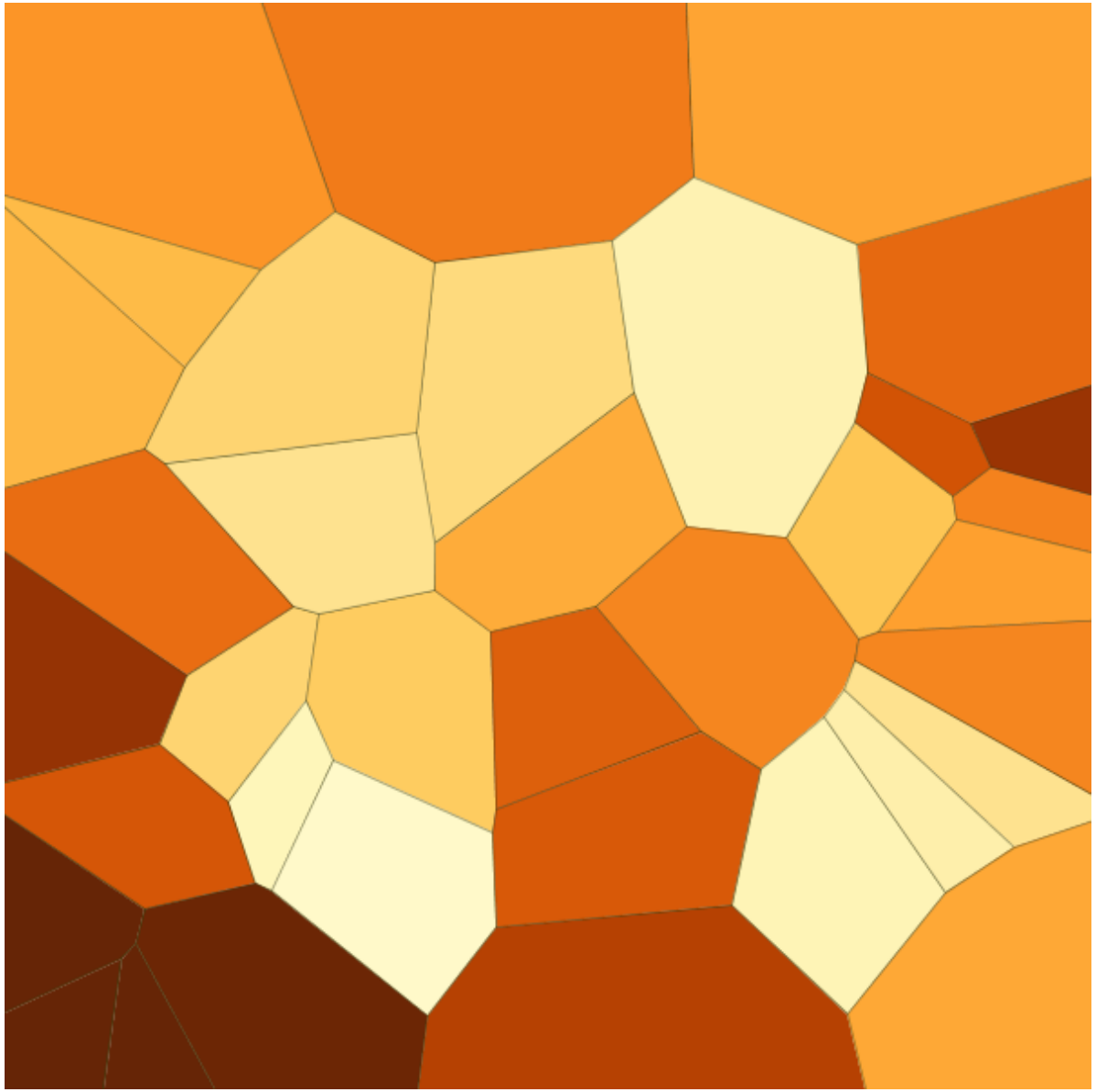


7

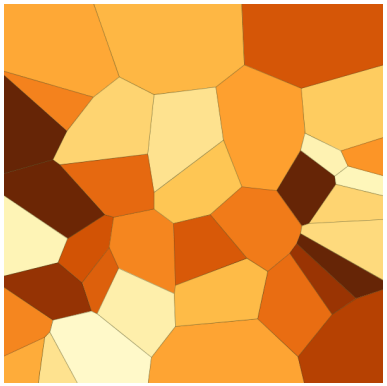


8

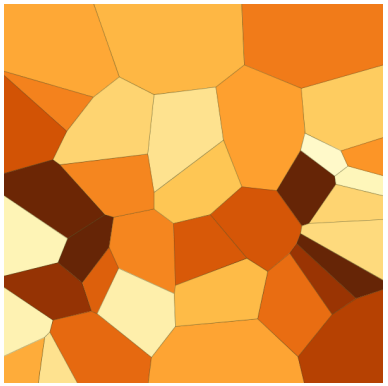




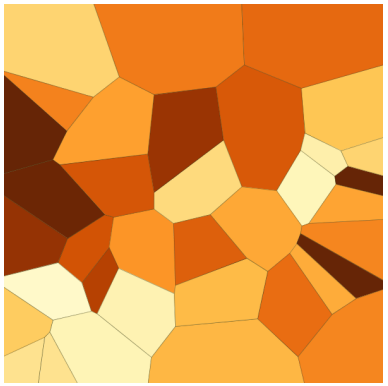
1



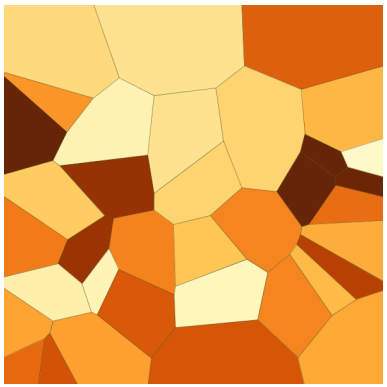
2



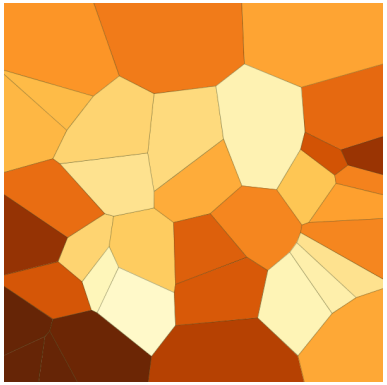
3



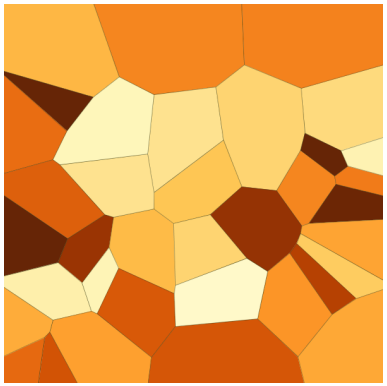
4



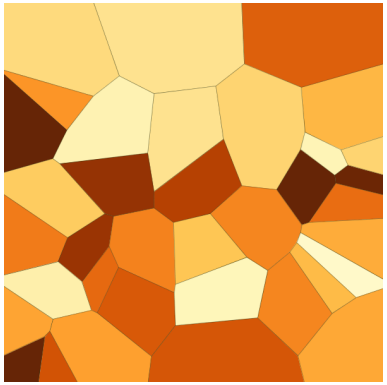
5



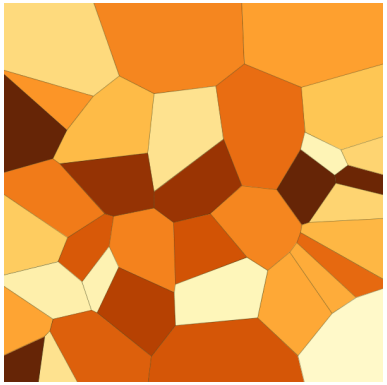
6



7



8



empirical tests

questions