An Address-Based Precinct Definition

Max Fan

October 14, 2020

Contents

1 Abstract

Over the past few decades, several processes for quantifying gerrymandering have been proposed. When applied to real-world situations, these processes frequently rely on unreliable data from states, counties, and a variety of other jurisdictions. In some cases, precinct lines are traced by hand over scanned copied of paper maps. In many instances, identical precincts are difficult to distinguish algorithmically, making it hard to calculate district stability. Unreliable and unstable data sources pose a significant obstacle in producing reliable and accurate gerrymandering research, requiring extremely labor-intensive data vetting. A faster and more accurate method of cleaning and representing precinct lines would reduce the amount of labor required to conduct gerrymandering research and make it more feasible to run multi-state and nationwide experiments. This paper proposes a simpler and more accurate method of representing districts: defining precincts as the set of addresses contained within. Additionally, the information-theoretic equivalence of such a representation of such a will be shown.

2 Introduction and Motivation

The current standard of representing voting precincts is a geo-spatial one. Precincts are represented as polygons and manipulated through various geometrical techniques. These representations are saved in a geo-spatial "shapefile", allowing for easy analysis and sharing. This format is the most visually intuitive representation of voter precincts: representing them as shapes.

One prominent metric used in gerrymandering research is the compactness metric. Intuitively, a district with a greater perimeter to area ratio is more likely to be gerrymandered than a district with a lower ratio. This is the essence of the various compactness measurement techniques proposed in the literature.

For example, a commonly proposed compactness method is exactly the method described previously: the ratio of the perimeter to the area. Another commonly used, alternative metric is the cut edges metric. The cut edges metric is generated from the dual graph of the polygon-based representation of the precincts in a district. The dual graph is generated by converting every precinct to a node, with every adjacent face of two precincts as the edges. Then, the cut edges metric is the number of edges in the dual graph that are **cut** by a districting plan.

Some compactness metrics, particularly the perimeter-area metric, can vary from year to year or even within the shapefile itself. This is particularly the case with states and counties that have inconsistent resolutions across precincts and years. If the shorelines in one precinct are extremely fine, whereas the shorelines in another are extremely coarse, the perimeter-area metric can be disproportionately inaccurate in reliably detecting gerrymandering. However, the cut edges metric does not suffer from this particular flaw for shorelines.

Several other data quality issues exist for cut edges. For example, sometimes inter-precinct boundaries can sometimes be extremely craggily, for example, if a river separates two different precincts. In such a case, many edges would be created for the same adjacent face. Without human, geographical context, it is difficult to distinguish from the legitimately many-edged precincts, such as a gerrymandered precinct, from the naturally many-edged precincts, such as those delineated by rivers or lakes. In fact, making such a determination, if possible, would be so labor intensive that these data quality issues are typically ignored due to the immense difficulty in determining which polygon edges are worthy of inclusion in the dual graph. In short, any process to clean up the quality issues in a dual graph based upon real-world precinct measurements is fraught with many subjective, arbitrary decisions. As the goal of creating objective metrics to quantify gerrymandering is to remove subjectivity in the analysis, a different approach is required.

3 An Address-based Representation

Non-important or trivial changes in the polygon-based precinct definition are difficult to computationally distinguish. Previous work relied upon giving error tolerances for the boarders to account for varying degrees of boarder and coastline resolution from year to year. However, by defining each precinct as the set of addresses contained within, which I call an address-based representation, only changes in inhabited areas are modeled. In this definition, a precinct is determined not by a noise-intolerant polygon boundary, but as the set of addresses contained within the precinct. In this representation, many illegal polygon-based precinct representations are unrepresentable in the address-based representation, following loosely from the general principle of making illegal states unrepresentable.

In addition, each address point more closely correlates to each individual in precincts, allowing for a more granular representation of the most important part of voter precincts: the voters themselves.

3.1 Advantages

3.1.1 Data quality and Sources

When evaluating the merits of a new representation method, it is essential to take into consideration the practicalities of potential data sources, particularly, the quantity, quality, and provenance of available data sources.

The quantity of data for the address-based precinct definition is greater than that of the polygon-based one. In many states, the boarders of precincts are ill-defined or stored in a non-digital format. Address, on the other hand, do not suffer from a quantity issues. The United States Department of Transportation maintains a National Address Database, which attempts to keep track of every address in America. Additionally, the postal office, the judiciary, and spam mailers all appear to have ready, high-quality national address lists. In a court case or in a legislative redistricting plan, these supplementary sources would all be available for redistricting use.

Additionally, the quality of data for the address-based precinct definition is greater than that of the polygon-based one. States themselves already maintain high-quality voter registration lists for the purposes of voter identification and address-checking at the poll booth on Election Day. Even if the voter registration lists, which are the ground truth of all eligible voters, are unavaiable, generic address lists are a good approximation of voters. However, for the polygon-based representation, some states do not maintain digital copies of the election precincts and force researchers to manually trace over scanned, paper maps in order to conduct research. In some cases, the maps can also be confusing and require calling each individual county to confirm precinct boundaries This process is error-prone and adds noisy to the boundary of voter precincts.

Finally, by concatenating and deduplicating several different address lists, a more accurate address list can be obtained. In this manner, the Nation Address Database maintained by the Department of Transportation can be combined with state-provided data to result in a list more accurate than either input list.

In all three aspects, the address-based representation proves promising.

3.1.2 Equivalence and Voronoi Diagram

In addition to having great data sources and excellent theoretical properties, the address-based representation is convertible to an equivalent, traditional, polygon-based representation using the Voronoi diagram. The address-based representation is convertible to a traditional, polygon-based representation using the Voronoi diagram. The Voronoi diagram is composed of Voronoi cells, with each cell representing an address. More formally, each cell is defined as the set of points in the plane such that the closest address to each point is the same. By having an equivalent, visualizable representation, the address-based representation allows for previous work on polygon-based precinct representations to be applied to an address-based precinct representation, without the inherent data quality issues.

Each cell in the Voronoi diagram of the address-based representation is defined as the set of points nearest to each particular address. This creates a diagram containing convex cells where cell size roughly correlates to population density.

By representing population in the model itself, more populated geographical regions are forced to have smaller voter districts. However, with increasing granularity comes larger and more computationally intensive models. The address-based representation, unaggregated, is several orders of magnitude larger than any previously examined representation.

The Voronoi diagram is useful for several reasons. Firstly, it allows for the usage of cut edges and other metrics developed on the polygon-based precinct definition to be applied. The dual graph is generated by converting each cell in the Voronoi diagram to a node and each two adjacent cells to contain an edge. This is also known as a Delaunay triangulation. Since much previous work has been done on the dual graph of the polygon-based metric, the ability to convert to a Delaunay triangulation is an important consideration.

Secondly, it allows the address-based representation to be easily visualized in a manner easy to understand by the public and courts. This is an important consideration when considering new gerrymandering and redistricting techniques.

Thirdly, as a result of each cell in the Voronoi diagram being convex, strange boarders and shapes are less likely to be generated by the Monte Carlo Markov Chain methods.

3.1.3 Monads

A monad in computer science is typically used to help with code abstraction. It is the functional programming equivalent of writing classes and utilizing Factory design patterns. More formally, a monad M is defined as something that implements the “bind” and the “return” functions.

```
return :: a → M a
```

```
bind :: M a → ( a → M b ) → M b
```

The address-based representation and the address-based representations of precincts are equivalent. More formally, there is a monad from the address-based representation to an equivalent polygon-based representation utilizing the Voronoi diagram, which I call the AddressVoronoi monad.

The return function is defined as computing the Voronoi diagram. The bind function is defined as taking in the generated Voronoi diagram and applying any polygon-based function to the Voronoi diagram and returning an altered Voronoi diagram. Additionally, a function called "just" that returns back the address-based representation is implemented by returning the address within each cell of the Voronoi diagram.

The existence of the AddressVoronoi monad allows any function defined on the polygon-based representation to be applied to the address-based representation enabling previous work to be utilized.

4 Experimental design (TBD)

To examine the usefulness and practical implications of an address-based precinct definition, the address-based representation of precincts in Massachusetts in 2002 and 2016 were utilized. Massachusetts was selected due to its poor data management practices and the apparent precinct boarder changes reflected in the data, which contradict the legally stated precinct boarders.

The first use case examined was the calculation of precinct stability. Precinct stability is defined as the percentage of precincts that have the same geographical boarders from year to year. Previously, with polygon-based precinct definitions, precinct stability in Massachusetts was difficult to reliably determine due to, in many instances, uninhabited regions such as lakes and forests changing from one precinct to another year over year.

The second use case examined was the differences in running a Monte Carlo Markov Chain on an address-based representation. In particular, the focus was on determining if population was better represented in the address-based representation when compared to the original polygon-based representation.

The third use of the address-based representation was to test the robustness of the current redistricting code base. As the address-based Voronoi diagram is several orders of magnitude greater than any previously examined representation, examining the performance of the codebase under such extreme conditions was useful.

5 Results (TBD)

6 Conclusion

7 Future work

8 Scratch (ignore this part)

Firstly, while some states do not keep electronic shapefiles, all states have some way of mapping each voter to the precinct they vote in. Therefore, the address-precinct mapping must exist, in some form, at a very high level of accuracy.

NB: Ignore everything below this line.

Secondly, the address-based representation is a significantly better representation of

Secondly, given the large, reasonably accurate national address databases, both governmental and proprietary, an accurate address-based representation can be easily recovered from a noisy shapefile. Since typically the noisy regions of a polygon are uninhabited, they are not represented in an address-based scheme. This allows an address-based representation to be used for cleaning up noisy polygon-based shapefiles as only the important parts of the precincts – namely the addresses contained within the precinct – are preserved.

Thirdly, the address-based representation and the traditional, visual polygon-based representation are easily convertible. Thirdly, the voronoi diagram of the addresses can be used to convert back to a polygon-based visual representation of the precincts.

Thirdly, not only is there a conversion from the polygon-based shapefiles to an address-based representation of precincts, but there is also a lossless conversion from an address-based representation of precincts to a polygon representation. The voronoi diag

Finally,