ISSN: 2473-4901
v9 n5912

Spatial Join in Spatial Data Analytics

Peter Y. Wu
wu@rmu.edu
Department of Computer Information Systems
Robert Morris University
6001 University Blvd, Moon, PA 15108

Abstract

In data analytics, when the data sets are spatially related, it often calls for certain special skill sets to process the geometric and topological relationships involved, such as those in the Geographic Information System. Spatial join is one such operation. We analyze the operation and focus on how it operates on two tactically different approaches of the point-in-polygon test. The two approaches are basically one, from the perspective of the aggregation of data. We describe the two approaches in details. In this paper, we explain the operation and introduce it for spatial data analytics. We explore three applications in data analytics where the spatial join is practicable: in political re-districting, in assessing pollution credits, and in distribution logistics.

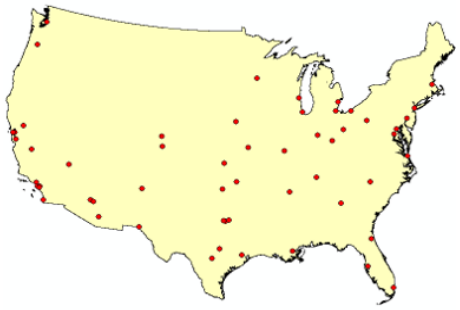
Keywords: Spatial Join, Spatial Data Analytics, GIS, Geographic Information System.

1. INTRODUCTION

Data analytics has a long history but did not come into prominence until the past decade (Albright, Winston 2017; Sedkaoui 2018; Sharda, Delen, Turban 2020). The approach to data analytics has made a lot of progress, but the key is in the availability of data and the immense increase in processing power. An important component of artificial intelligence today is based on data analytics of a huge pool of data gathered. Yet in the proliferation of data analytics, it often calls for certain special skill sets to process data sets that are spatially related. Spatial data analytics needs a combination of skills in data analytics with that of the geographic information systems (Zhou, Su, Pei, et al, 2020; Wu, Igoche 2020). The GIS operates on the geometric and topological relationships, combined with data analytics working on massive amounts of data. One such operation is the spatial join which relates data sets by their spatial relationship. The generic spatial join focuses on the point-in-polygon test to relate spatial data. While there are two approaches to spatial join, we analyze the operation to resolve that they are basically one and the same. Then we proceed to describe various applications in which spatial join is fundamental. These applications include political re-districting, calculating pollution credits, and distribution logistics.

2. SPATIAL JOIN

Spatial join is the operation to connect multiple data sets based on their spatial relationship. Suppose we have a data set of the major cities of the United States and the population in each of the cities, as illustrated in Figure 1.



Cities	
Name	**Population**
Seattle	593,350
Phoenix	1,502,129
Oakland	411,240
…	…

Figure 1. Cities of the US

ISSN: 2473-4901
v9 n5912

We also have a data set of the states of the United States, as illustrated in Figure 2.

States	
Name	**Population**
Washington	5,894,121
Arizona	6,392,017
California	33,871,648
…	…

Figure 2. States of the US

The two data sets have no apparent relationships except when we show them in the same map. Figure 3 puts the two on the same map, showing which cities are in which states. Spatial join makes use of the location information to resolve that, and allows us to pursue further analysis.

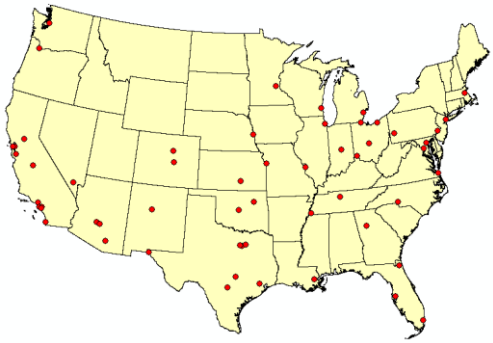


Figure 3. Cities and States together

The core of the spatial join is the point-in-polygon test to sort out which polygons contain which points. There are different approaches to speed up the processing (Wu 2005), but we are not concerned about the efficiency of the operation here. We reason that generically, the two data sets in the spatial join will have one set representing points, such as the cities, and the other set representing polygons, such as the states, in our example.

3. TWO APPROACHES AND AGGREGATION

The output of the spatial join depends on what we are interested in. If we cast the polygon data set over the point data set, that is, the states over the cities, we will have in our result a modified point data set. We will have for each city, the state it is in, along with all other attributes of the state, such as the state population, as illustrated in Figure 4, and we may also calculate the percentage of the city's population in the state. The crucial issue is that each point can only be in one polygon: each city can only be in one state.

Cities – after Spatial Join			
Name	**Population**	**State.Name**	**State.Population**
Seattle	593,350	Washington	5,894,121
Portland	551,302	Oregon	3,421,399
Sacramento	462,910	California	33,871,648
Oakland	411,240	California	33,871,648
…	…	…	…

Figure 4. Cities after Spatial Join

If we cast the point data set over the polygon data set, that is, the cities over the states, we will have a more complicated result. The crucial issue is that a polygon may contain multiple points or no point at all. In our case of the states versus the cities, we will have in our result a modified data set for the states, included for each state, the various aggregations of the cities in that state depending on what we want for the application. We will certainly have the number of cities in each state, but we may also have to the total population of the cities in the state, as illustrated in Figure 5. From there, we may also have the percentage of city population in the state.

States – after Spatial Join				
Name	**Population**	**count**	**City.pop**	**City.pop %**
Washington	5,894,121	1	593,350	10.07%
Arizona	3,421,399	3	2,492,038	72.84%
California	33,871,648	10	9,484,420	28.00%
…	…	…	…	…

Figure 5. States after Spatial Join

The two approaches depend on what we want to get as the result, but they are not fundamentally different. If we focus on the spatial join result of the cities in Figure 4, we can apply aggregation by the state to obtain the spatial join result of the states in Figure 5, as we would apply "GROUP BY" to the data table and draw up the appropriate aggregate functions as needed to the attributes as shown.

ISSN: 2473-4901
v9 n5912

4. POLITICAL RE-DISTRICTING

Every 10 years, the United States runs the census and according to the results, the state governments may re-draw the map of political districts according to the new data. The census produces demographics data of where people groups are located. As we re-draw the political districts, we can apply spatial join casting the demographics point data set over the districts. Every district newly drawn is a polygon against which we apply spatial join on the demographics map layer, to identify the percentage of each kind of the population group such as ethnicity, gender, age, income level, educational level, and the like.

Since it can be done quickly, it allows us to easily verify the general representation in each district, to meet the requirements of re-districting. That, however, also leads to the GIS becoming an aid to the practice of gerrymandering (Crane and Grove 2018; Forest 2018; Wu, DePlato and Combs 2020). Politicians can carefully re-draw to the districts to include or exclude voters they want, as illustrated in Figure 6. While the debate on how to identify gerrymandering to prevent it, there are also discussions about how to stop the practice using the GIS (Wu and Igoche 2022).

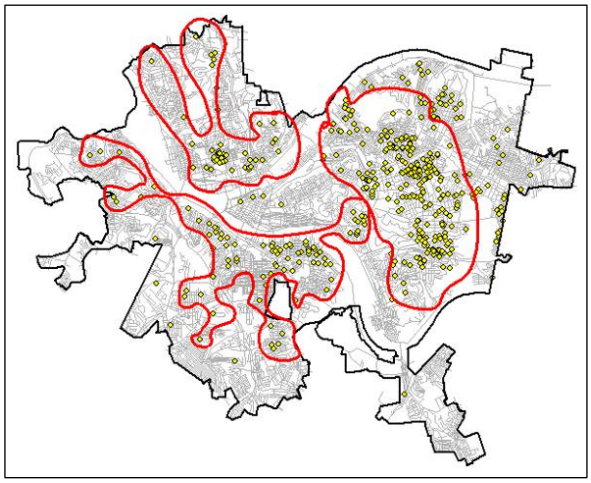


Figure 6. Re-districting by Choosing Voters

5. POLLUTION CREDITS

Consider a certain manufacturing company with several plants in operation. The locations of these plants are in different states. Environmental protection in each state requires the plants to purchase certain pollution credits for the exhaust fumes produced each year. Hence, we need to estimate the amount of exhaust, but for those within each state, we may choose to shift the credits from one plant to another. Figure 7 shows the case of multiple plants in the Michigan, Ohio and Pennsylvania region. Treating each state as a region, we can apply spatial join to the map of the production plants, identifying the plants in each state. Then we can apply the sum aggregate function to sum up the exhaust fume capacities as well as the pollution credits. That will allow adjustments periodically as appropriate.

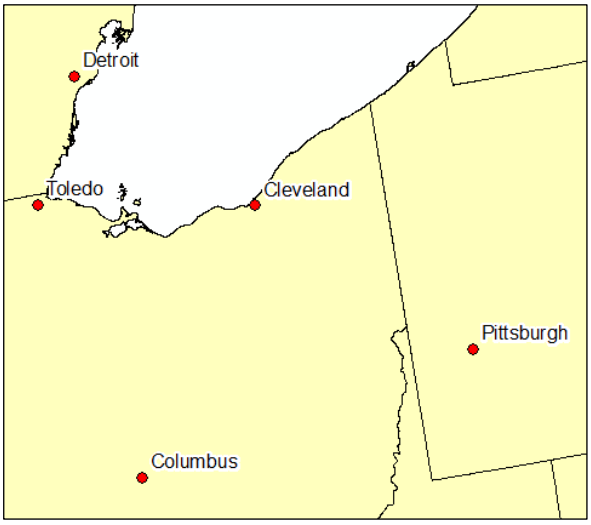


Figure 7. Pollution Credits in the Tri-States

6. DISTRIBUTION LOGISTICS

During the Covid pandemic, we had to ship vaccine to supply different areas. But there were also different priorities set up for different groups of people to be vaccinated, and we needed to estimate the amount of vaccine to ship to each distribution center. Based on the locations of the distribution centers, we built Voronoi diagrams (Aurenhammer, Klein and Lee 2013; Yamada 2017).

Given a set of points in the 2D plane, the Voronoi diagram subdivides the plan into regions closest to each point. An algorithm to construct it in optimal time was first published in 1975 (Shamos & Hoey), and it is generally available in the GIS.

Making use of the Voronoi diagram, we obtained a polygon map of the regions in the area, each region represents the region each distribution center was responsible for. We then applied spatial join with the polygon map of the Voronoi diagram on the demographics map of the whole area, as illustrated in Figure 8**.** We could sum up the populations of the various demographic groups in each region. That gave us a reasonable estimate of the amount of vaccine needed in each distribution center at different times as the vaccination priority changes in time.

ISSN: 2473-4901
v9 n5912

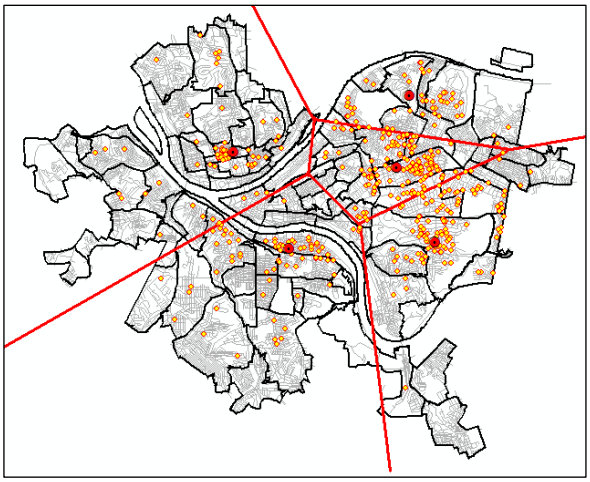


Figure 8. Using Voronoi Diagram to determine distribution logistics

7. SUMMARY

Spatial join is a fundamental operation in GIS work for spatial data analytics, linking the processing of two data sets which are spatially related. We explained the generic form of spatial join in two approaches, based on the point-in-polygon test. The two approaches however are one and the same, just a matter of convenience about what we may want in the result. We then proceeded to show how spatial join is useful in a few spatial data analytics cases: in political re-districting, in calculating pollution credits, and in distribution logistics.

8. REFERENCES

Albright, S.C., Winston, W.L. (2017). Business Analytics: Data Analysis & Decision Making, (6th edition). Cengage.

Aurenhammer, F., Klein, R. and Lee, D-T. (2013). Voronoi Diagrams and Delaunay Triangulations. World Scientific.

Crane, N.J. and Grove, K. (2018). An Active Role for Political Geography in Our Current Conjuncture. *Geography Compass* 12(11). Wiley Online Library.

Forest, B. (2018). Electoral Geography: From Mapping Votes To Representing Power. *Geography Compass* 12(1). Wiley Online Library.

Sedkaoui, S. (2018). Data Analytics and Big Data. Wiley.

Shamos, M.I. and Hoey, D. (1975). Closest-point Problems. *16th Annual Symposium on Foundations of Computer Science*.

Sharda, R., Delen, D. and Turban, E. (2020). Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support, (11th edition), Pearson.

Wu, P.Y. (2005). A Distributed Approach To Fast Polygon Overlay. *6th Annual Central Appalachian Geo-Spatial Conference*, California University of Pennsylvania, Southpointe, PA, August 2005.

Wu, P.Y., DePlato, J. and Combs, A. (2020). Geographic Information System For and Against Gerrymandering. *Journal of Information Systems Applied Research 13(3), pp.4-10*. ISSN: 1946-1836, November 2020.

Wu, P.Y., Igoche, D.A. (2020). Analytics, Spatial Data Analytics for the Pandemic. *Proceedings of the Conference on Information Systems Applied Research* (CONISAR), 13(5389), ISSN:2167-1508, November 2020.

Wu, P.Y., Igoche, D.A. (2022). GIS for Democracy: Toward A Solution Against Gerrymandering. *Journal of Information Systems Applied Research 15(2), pp.47-53. ISSN 1945-1836*, July 2022.

Yamada, I. (2017) *Thiessen Polygons*. The International Encyclopedia of Geography. Richardson, et al, eds. John Wiley & Sons.

Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B. and Song, C. (2020) COVID-19: Challenges to GIS with big data. Geography and Sustainability.