Thiessen's Remarkable Polygons

Mark P. Kumler
California State University, San Bernardino

Abstract: This article presents an analysis of Alfred Thiessen's 1911 work which introduced geographers to the concept of Thiessen polygons, also known as Dirichlet regions or Voronoi polygons. The investigation reproduces Thiessen's example for precipitation data and reveals a remarkable, curious flaw. Although Thiessen described correctly the procedure for delineating the polygons, his example appears to have been mis-delineated, perhaps intentionally, such that the area-weighted average would yield precisely the same answer that would be obtained if the precipitation value was known everywhere. Possible reasons for this deception are considered, and a reconsideration of the polygons' name is recommended.
Key Words: Dirichlet regions, Voronoi polygons, Wigner-Seitz regions, pathological science.

Introduction

In the July 1911 issue of Monthly Weather Review, Alfred Thiessen presented a procedure for estimating the average for an area in which the locations with known values are unevenly distributed. The technique relies on the delineation of boundaries about each observation point to create polygons that enclose all points nearer one observation point than any other (Thiessen 1911). These polygons have come to be known as *Thiessen polygons*, and they are widely used in geographical and other spatial studies. For an exhaustive review of these polygons (over 600 references), including their definition, history, computation, statistical properties, and applications, see Okabe, Boots and Sugihara (1992). This paper will reveal a flaw in Thiessen's original presentation – a flaw that gave his polygons a rather remarkable property – and it will put forth some possible explanations for the mistake.

Background

Alfred Thiessen was a climatologist with the United States Army. He was also a Councillor and charter member of the American Meteorological Society, and he served as Observer, District Forcaster, and District Editor for the U.S. Weather Bureau in the early 1900s (Bulletin of the AMS, 1936). In Thiessen's early years as an Observer for the Weather Bureau, he contributed numerous reports on advances in a wide variety of scientific disciplines related to climatology, including "Dust in the Atmosphere", "Snow Rollers", "The Blue Color of the Sky", and "An Explanation of Wireless Telegraphy" (Thiessen, 1899a, 1899b, 1899c, and 1902, respectively). From July 1909 through 1913 the Monthly Weather Review included climatic summaries for each of twelve large weather

districts of the U.S., which corresponded roughly with the country's largest drainage basins. During these years Thiessen served as District Editor for District No. 10, Great Basin, and he authored monthly summaries of the district's climate. The summaries often included additional information, related to the interests of the individual editors, and they often served as "sounding boards" for the editors' ideas. It was in the July 1911 issue of Monthly Weather Review that Thiessen published his now-famous description of a procedure for determining the boundaries for an area-weighted average.

Thiessen's Problem, and His Perfect Solution

Thiessen presented his idea as a question about how to estimate the average rainfall over an area in which the observation points are irregularly distributed. In his hypothetical example the true precipitation was known for weather stations at the centers of each square in a four-by-four matrix of sixteen squares, and three scenarios were considered in which only six of the sixteen stations reported their precipitation values (as might be the case if there were communication problems caused by bad weather). Thiessen's example is reproduced here as Figures 1–4. Figure 1 (Case 1) is the "truth", where the precipitation is known for all sixteen stations, and Figures 2, 3, and 4 are the three possible cases.

The simple, unweighted average of the complete set of sixteen measurements was 1.875 inches (rounded to 1.88 inches in the original). The simple unweighted averages for all cases are presented in Table 1, along with the error, or "variation" of each estimate, which Thiessen reported as a percentage "too high" (above) or "too low" (below) the true value.

Thiessen highlighted the fact that the simple averages yielded errors ("variations") ranging from 24 percent too low to 15 percent too high, and then presented his proposed "solution":

> "The discordant results are due to the fact that the extent of the areas represented by the data was not considered. The amount of rain recorded at any station should represent the amount for only that region inclosed by a line midway between the station under consideration and the surrounding stations. Giving, therefore, each station its proper weight in reference to the area which it represents, we have instead of the former equation the following..." (Thiessen, 1911, p. 1083)

Thiessen proceeded to illustrate his solution by considering Case 3. He identified boundary lines between the regions – ostensibly by following the rules described above (reproduced here as Figure 5) – and he computed a new area-weighted average. Using weights proportional to

the areas of the regions he delineated, Thiessen obtained a value of 1.88 inches "which is the true average for the area" (i.e. it equaled the average of the entire set of sixteen values). Yes, it appeared somewhat remarkable that Thiessen had acquired the exact solution, with a sample of only six of sixteen observations and a novel set of boundaries and weights.

Thiessen's Flaws

A few trivial errors occur in Thiessen's example. Case 1 ("the truth") is reported as having an average of 1.88, whereas the exact value is 1.875. While it may be argued that this .005 difference is a simple effect of rounding, and not an error per se, it has repercussions in succeeding calculations. In Case 2, the average is reported as 2.17 instead of 2.166..., leading to a reported variation of 15% instead of the more accurate 16% (the exact value is 15.55...%). In Case 3, the average is reported as 1.83 instead of 1.833..., leading to a reported variation of 3% instead of the more accurate 2% (the exact value is 2.22...%). And in Case 4, a single digit appears to have been transposed in the recording of the average as 1.43 instead of the exact 1.33...; this led to the variance being mistakenly reported as 24% when in fact it is 28.88...%. All of these numbers are presented in Table 2.

The above errors would not be noteworthy if it were not for the one gross error that Thiessen made: a severe mis-delineation of the boundaries between the regions. Figure 6 depicts both Thiessen's delineation and a proper delineation for Case 3–the proper delineation having been obtained by following the procedure outlined correctly in Thiessen's own text. Note that Thiessen did not include boundary lines between regions with the same value, as they had no effect on the average, but they are included in Figure 6 to facilitate this comparison.

To further facilitate a comparison of the two delineations, several "nodes" have been identified at intersections of boundary lines with each other or with the region's perimeter, as well as at all bends along a line. The nodes are labeled with lower-case letters, and their shifts–between Thiessen's delineation and a proper delineation–are indicated with arrows in Figure 6b. Note that it is necessary to infer the locations of some of these nodes, since there is not a one-to-one relationship between physical nodes in the two networks. Nodes c, d, and e, for example, are not intersections in Thiessen's delineation, but their locations have been inferred. Similarly, nodes h and i are present in Thiessen's delineation at breaks in direction along lines, but in the proper delineation there are no such breaks. Where it has been necessary to infer the location of a node, the most generous location possible has been selected, i.e. the

node has been located such that the difference in location between the two delineations is minimized.

Consider the differences in the delineations by first examining the boundary line between the "1" and the "2" in the lower left quadrant. Thiessen has drawn a boundary vertically from "d" to "h" and then at a slight angle to "j", whereas the proper boundary would follow the perpendicular bisector between the centers at a 45° angle (illustrated in 6b). Similarly, the boundary between this same "2" and the "1" in the lower right corner should follow the perpendicular bisector, not the irregular "f" to "i" to "k" boundary drawn by Thiessen. Additional discrepancies occur between the interior "2" and the "1" in the upper-left corner, and between the interior "2" and the "4" in the upper-right.

These errors in the delineation of region boundaries are not trivial. As drawn by Thiessen, they yield a weighted average of 1.88, which is precisely the same answer he obtained from all sixteen observations. When properly drawn, an area-weighted average yields a value of 1.9375, or approximately 3.33% "too high".

A Simple Mistake, Pathological Science, or Just Plain Fraud?

Whereas the minor flaws detailed in the beginning of the previous section may all be considered simple mistakes, as they appear to be inadvertent and they do not affect significantly the conclusion, the mis-delineation, which led to polygons that yielded precisely the same answer that one would have obtained if the values at all sixteen stations were known, is a significant error and begs an explanation. Was this inadvertent, subconscious, or just plain fraudulent?

Referring again to Figures 6a and 6b, one can see that four nodes were mislocated. Perimeter nodes a, b, c, g, and k are all in their proper locations, but internal nodes d, f, and i were mis-located by one-half cell, as was perimeter node j. If one were to give Thiessen the benefit of the doubt–i.e., assume that he was just a bit careless with his numbers, as evidenced in the trivial errors–one might assume that a few such mis-locations are understandable (or at least typical), and they should be dismissed as relatively unimportant to the presentation. One might further assume that mis-locations of one-half a cell edge are reasonable errors, or at least reasonably acceptable given that one is being a bit sloppy with the geometry.

With these generous assumptions, let us consider Thiessen's errors. Four internal nodes were misplaced by one-half cell, and one perimeter node

was misplaced by one-half cell. Each of the internal nodes could have been mislocated by a half-cell in four different directions, and the perimeter node could have been mislocated be a half-cell in either of two directions. There are thus 4^4 possible combinations of mislocations of the four internal nodes, plus two possible mislocations of the perimeter node, for a total of 4^4*2, or 512, possible ways to make this number of these kinds of errors. If one considers the possibility of stumbling on the correct locations for some of these nodes (thus factoring in the possibility of making fewer errors), the number swells to 1874 (5^4 * 3 - 1, the 1 being the correct solution) possible ways to mislocate some or all of these five nodes. Of this very large number of possible ways to make these kinds of errors, it is, in this author's opinion, truly remarkable that Thiessen stumbled on one mis-delineation that generates a weighted average of exactly 1.88 inches.

Given that it seems quite unlikely that Thiessen stumbled on such a perfectly erroneous solution by chance, we must consider whether it was an intentional deception or a case of pathological science. The term "pathological science" is credited to Nobel-laureate Irving Langmuir (1881-1957). Langmuir spent many years pursuing Nobel-caliber research in chemistry; he also had a hobby of investigating science that he termed "pathological". Langmuir characterized these cases as "... where there is no dishonesty involved, but where people are tricked into false results by a lack of understanding about what human beings can do to themselves in the way of being led astray by subjective effects, wishful thinking, or threshold interactions" (Langmuir 1953). Langmuir investigated and revealed several such cases in particle physics, isotopic chemistry, and parapsychology (extrasensory perception). More recently the term has been applied to several cases in the field of nuclear chemistry, the most notable being the much-publicized "discovery" of cold fusion by Pons and Fleischmann (reviewed in Rousseau 1992, and Cromer 1993).

Langmuir listed six symptoms that he considered characteristic of pathological science. Two that might apply to Thiessen are "claims of great accuracy" and "the effect is of a magnitude that remains close to the limit of detectability" (Langmuir 1953). There is no question that Thiessen's polygons were highly accurate: their weighted average led to a perfect estimate. But what was the magnitude of the effect? How much improvement was effected by using Thiessen's approach instead of the simple unweighted average?

Table 2 presents both the simple unweighted average and the (properly-delineated) Thiessen-weighted average for each of Thiessen's three cases. Note that in Cases 2 and 4 the Thiessen-weighted average yields a significantly better estimate than the simple average does, but in Case

3–the case that Thiessen unwittingly selected to illustrate his point–the simple unweighted average is in fact more accurate. The simple unweighted average yields an estimate 2.22...% "too low", while a properly-delineated Thiessen-weighted average yields an estimate 3.33...% "too high". Not only is the magnitude of the effect rather slight, it is in the wrong direction; if Thiessen had not mis-delineated his polygon boundaries, his areal-weighting would have increased the magnitude of the error of the estimate by a full 50%!

In trying to assess whether the magnitude was near the limit of detectability, one must consider both the apparent effect (based on the mis-delineation) and the real effect. The apparent effect was dramatic–an elimination of all error, whereas the real effect is equally dramatic–the estimate does not improve at all but in fact becomes worse. Considering both possibilities, it seems inappropriate to characterize the effect as "near the limit of detectability". Thiessen's example then meets only one of the six symptoms that Langmuir offered for pathological science, and it seems unfair to characterize Thiessen as a self-deluded pathological scientist.

The one remaining possible explanation, however, is even less charitable: that Thiessen's mis-delineation was simply fraudulent. It could be that Thiessen, hoping to promote his idea, may have started with a proper delineation and then manipulated the boundaries until he obtained a weighted average that matched precisely the true average. This is no simple feat–as the reader might discover by trying to adjust boundaries from figure 6b to achieve an average of exactly 1.875–but it could be done.

Conclusion

Thiessen's polygons were truly remarkable. Although Thiessen described correctly the procedure for their construction–a procedure that is widely used today in many geographical studies–he illustrated his idea with a perfectly erroneous delineation. In his example, properly-constructed Thiessen polygons yield a weighted average that is 3.33% greater than the truth, yet Thiessen's mis-delineated polygons yield the exact truth. Whether this was inadvertent, a case of pathological science, or mere fraud we shall never know. But it certainly appears unlikely that it was inadvertent.

These polygons – or regions that encompass all points nearer an observation point than any other point – have been "discovered" independently by many individuals in a variety of fields. In mathematics they are known as *Dirichlet regions* or *Voronoi polygons*, after mathematicians who

introduced them in 1850 and 1908, respectively (Dirichlet 1850, Voronoi 1908). In physics they are known as *Wigner-Seitz regions* (Wigner and Seitz, 1933) or "domains of the atom" (Frank and Kasper, 1958). And in ecology they were introduced relatively recently as *"plant polygons"* (Mead, 1966).

Given the errors that Thiessen made in his original presentation, I recommend that henceforth these polygons be known no longer as *Thiessen polygons*, but rather as *Dirichlet regions*, after the mathematician who first introduced them to the scientific community in 1850.

References

Boots, B.N. 1986. Voronoi (Thiessen) Polygons. Concepts and Techniques in Modern Geography (CATMOG) 45, (Norwich, UK: Geo Books), 51 pages.

Bulletin of the American Meteorological Society, January 1936, 17(1), p. 18.

Cromer, A. 1993. Pathological Science: An Update. *Skeptical Inquirer,* 17(4):400–407, Summer.

Dirichlet, G.L. 1850. Über die reduction der positeven quadratischen formen mit drei unbestimmten ganzen zahlen. *Journal für die Reine und Angewandte Mathematik,* 40:209–227.

Frank, F.C. and J.S. Kasper 1958. Complex Alloy Structures Regarded as Sphere Packings. I. Definitions and Basic Principles. *Acta Crystallographica,* 11:184–190.

Horton, R.E. 1917. Rational study of rainfall data makes possible better estimates of water yield. *Engineering News-Record,* 79:211–213. [Oka]

Langmuir, I. 1953. Pathological Science. Transcribed and edited by R.N. Hall from a microgroove disk found among Langmuir's papers in the Library of Congress of a colloquium given at the Knolls Research Laboratory, December 13 1953. *Physics Today.* 42: 36–48, October 1989.

Mead, R. 1966. A Relationship Between Individual Plant Spacing and Yield. *Annals of Botany,* N.S., 30:301–309.

Okabe, A., B. Boots, and K. Sugihara 1992. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. (England: Wiley), 521 pages.

Rousseau, Denis L. 1992. Case Studies in Pathological Science: How the loss of objectivity led to false conclusions in studies of polywater, infinite dilution, and cold fusion. *American Scientist,* 80(1): 54–63, January.

Thiessen, A. H. 1899a. The Dust in the Atmosphere. *Monthly Weather Review,* 27(2): 63, February.

_____ 1899b. Snow Rollers. *Monthly Weather Review,* 27(3): 100, March.

_____ 1899c. The Blue Color of the Sky. *Monthly Weather Review,* 27(3): 113–114, March.

_____ 1902. An Explanation of Wireless Telegraphy. *Monthly Weather Review,* 30(12): 570–576, December.

_____ 1911. Precipitation Averages for Large Areas. *Monthly Weather Review,* 39(7): 1082–1084, July.

Voronoi, G. 1908. Nouvelles applications des paramètres continus à la théorie des formes quadratiques, deuxième memoire, recherches sur les paralleloèdres primitifs. *Journal für die Riene und Angewandte Mathematik,* 134: 198–287.

Wigner, E. and F. Seitz 1933. On the Constitution of Metallic Sodium. *Physical Review,* 43:804–810.

Table 1

Averages and variations from true amount for Thiessen's cases.

	Number of stations	Average ppn, in inches	Variation from true amount
Case 1	16	1.88	0.
Case 2	6	2.17	15 percent too high.
Case 3	6	1.83	3 percent too low.
Case 4	6	1.43	24 percent too low.

Table 2

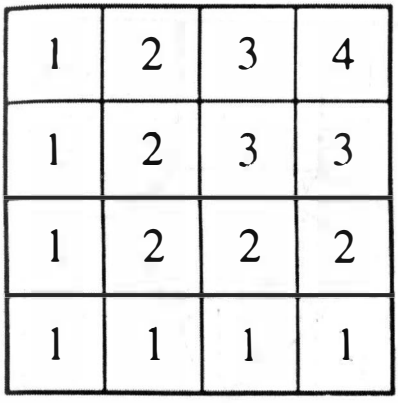
Averages and variations from true amount for Thiessen's cases. The leftmost four fields in this table are verbatim from Thiessen's original article.

	Number of stations	Average ppn, in inches	Variation from true amount	Exact Average	Exact "Variation"
Case 1	16	1.88	0.	1.875	0.0
Case 2	6	2.17	15 percent too high.	2.166...	15.55...% high
Case 3	6	1.83	3 percent too low.	1.833...	2.22...% low
Case 4	6	1.43	24 percent too low.	1.333...	28.88...% low

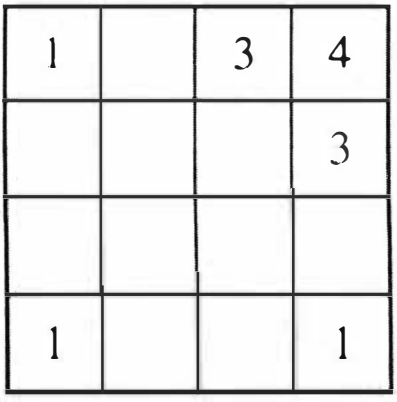
Table 3

Simple unweighted and Thiessen-polygon-weighted averages for Cases 2, 3, and 4.

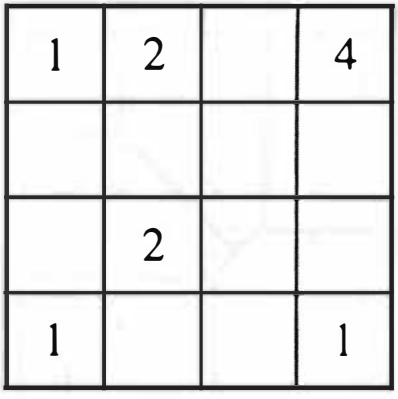
	Number of stations	Simple unweighted average	Error or "Variation"	Thiessen-weighted average	Error or "Variation"
Case 1	16	1.875	0.	1.875	0.0
Case 2	6	2.166...	+15.55...%	1.825	-2.66...%
Case 3	6	1.833...	-2.22...%	1.9375	+3.33...%
Case 4	6	1.333...	-28.88...%	1.7265625	-7.9166...%



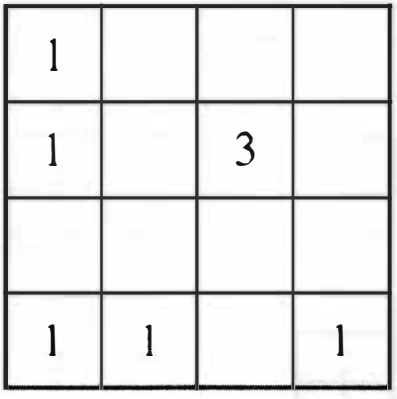
Case 1, Fig. 1



Case 2, Fig. 2



Case 3, Fig. 3



Case 4, Fig. 4

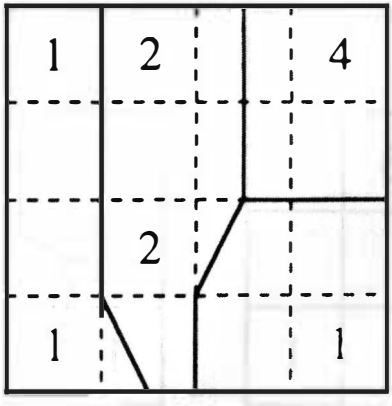


Fig. 5

Figures 1-5. These figures are re-drawings of those in Thiessen's original article. The figures and cases are numbered here as in the original, for ease of comparison.

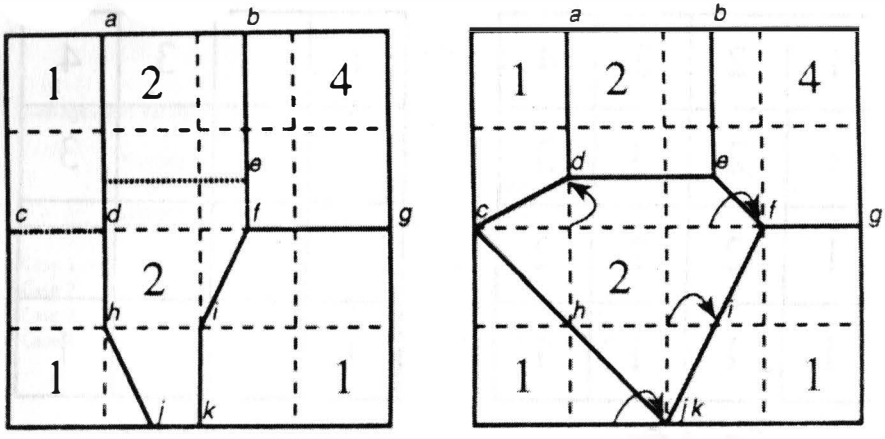
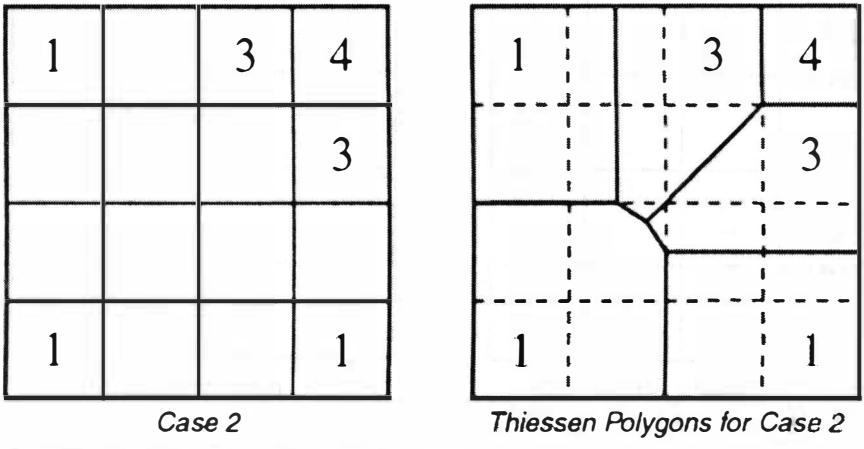


Figure 6. Thiessen's incorrect delineations (a) and the proper delineation (b)



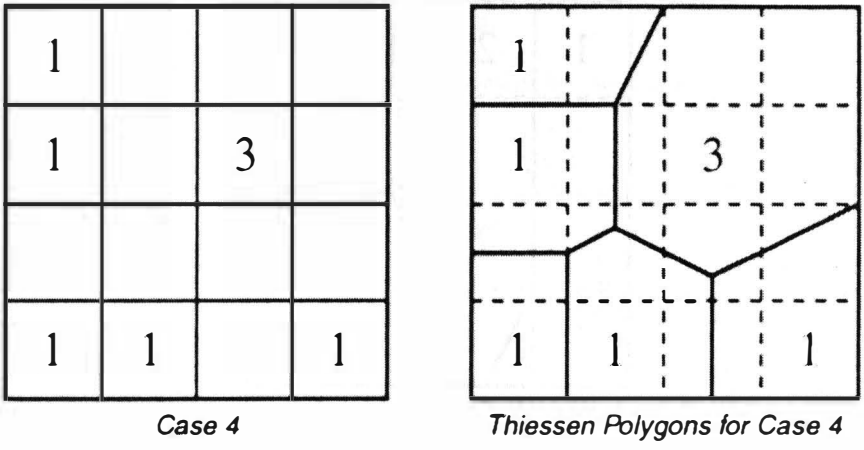


Figure 7. Thiessen's Cases 2 and 4, and their corresponding Thiessen polygonizations.