

Tomo Lennox, VP of R&D
DataMap, Inc.
7525 Mitchell Road, Suite 300
Eden Prairie MN 55344-1958
E-mail: ~~TFL@Datamap.mn.org~~

THE NEED FOR COORDINATE QUALITY METRICS IN PREDICTING ACCURACY IN GIS QUERIES

A coordinate pair defines the center of some object (example: a ZIP-code, fire district, address) in a GIS database. The coordinates alone provide little information that can be used to determine the quality of placement. The user must assume that the accuracy is the same for each coordinate and trust that accuracy is sufficient. This paper calls for a new standard in which additional metrics would be included with each coordinate pair. These additional metrics allow the measurement of positional error, and in the context of a specific geographic query, allow a confidence metric to be calculated.

CASE ONE: COORDINATES ONLY

Coordinate pairs are the bricks from which a GIS system is built. Mathematically, the coordinate pair specifies a precise longitude and latitude that together specify an infinitely small point on the surface of a sphere. In real life, no point can be measured without error. No floating point number can be expressed in a computer without round-off errors. The objects being described by the coordinates are rarely infinitely small. Nor is the earth a perfect sphere, but this error is insignificant compared to the others.

Say, for example, one is given the specification for the location of some houses shown in Table 1.

Coord.	Longitude	Latitude
1	-87.294834	29.956023
2	-87.2	30
3	-87.1145	30.23851
4	-87.1145	30.23851
5	-87.2566	29.8234

Table 1

What is the accuracy of this data? What is the maximum distance these points could be from the front door of the houses they represent? What is the average error on all points? With only coordinates, there are very few clues one can use to determine accuracy.

The first method available is to count the number of digits after the decimal place in each coordinate. A number with only four digits of precision may be less accurate than one with six digits of precision. Unfortunately, this approach has two flaws. First, some points will naturally fall on round numbers. Because computers do not usually follow the engineering convention of keeping significant digits, 30.2000000 will be displayed as 30.2, and therefore, look imprecise. The other flaw is that rounding and calculation errors often create more digits. These extra digits do not indicate extra precision.

The second method available if you have a large number of points is to analyze them for "stacking". It is easy to create coordinates if accuracy is not a concern. All one has to do is take an atlas and find the coordinates for the center of each state (the state centroid). For each item needing coordinates, use the coordinates of the center of the state. The result of this method is that every house in the state has the exact same coordinates. In other words, the points are all "stacked". The technique is perfectly logical, as long as accuracy of ± 200 miles does not create a problem. Though few data suppliers would represent state centroids as accurate data, county centroids, city centroids or ZIP centroids are common.

There are also two problems with attempting to measure data quality by analyzing stacking. First, if the data supplier wants to make their data look better, dithering (adding small random numbers to the data) hides the stacking while slightly decreasing the quality. Second, stacked points are sometimes real. Two addresses in the same apartment building should, probably, have the same coordinates. Likewise, two houses across the street from each other on a street without an

odd/even numbering system may both be placed in the center of the street. This will stack the points, but represents an error that could be less than 20 meters.

In summary, all the techniques for measuring data quality using only coordinate values have serious limitations.

CASE TWO: QUALITY GRADES

When different sources or techniques are used to calculate the coordinates for a point, some data suppliers pass on indications of these to the user. Though this indication could be expressed in any form, report card style letter grades are common. In simplified form, data marked with an "A" has good quality and the data marked with a "D" has lower quality. Table 2 illustrates how the data in the previous section would be presented with a grade for each coordinate.

Coord.	Longitude	Latitude	Grade
1	-87.294834	29.956023	D
2	-87.2	30	A
3	-87.1145	30.23851	A
4	-87.1145	30.23851	A
5	-87.2566	29.8234	B

Table 2

The grade provides some real information on quality. One could purchase a list and request only grade "B" or better coordinates. Alternatively, one could pay less for grade "D" coordinates than for grade "C" coordinates. If the project using the data had the resources to verify and correct some of the coordinates, correcting 25% of the data with the worst grades could produce more improvement than verifying a random 25% sample.

The problem is defining "good" in units that can be used when processing the data. How are the grades assigned? How can the user tell if grade B is accurate enough for a planned use? Can the grade be used to calculate which coordinates are off by more than 3 kilometers?

CASE THREE: RESOLUTION

Resolution measures the potential error in the location of a coordinate in units of distance. The resolution often expresses the same information as the grade does, but since it is expressed as a distance, it is more useful in calculations. In the example shown in table 3, the resolution is expressed in \pm kilometers from the given coordinate.

Coord.	Longitude	Latitude	Resolution
1	-87.294834	29.956023	1.0 lan
2	-87.2	30	.041 lan
3	-87.1145	30.23851	.052 lan
4	-87.1145	30.23851	.052 lan
5	-87.2566	29.8234	.3855 lan

Table 3

The resolution metric is a much more reliable way of measuring accuracy than counting decimal digits, as described above. Coordinate 1 in the table has more decimal digits than coordinate 2, but the resolution shows that coordinate 2 is over twenty times more accurate. The fewer number of decimal digits may be the coincidental chance of the calculation resulting in a round number.

Resolutions in point data

Inherent in the definition of resolution is that a coordinate has a statistical error, and that the resolution is a

measurement of the probability of a true mathematical point being within a circle. For example, the supplier of coordinate 1 in the table may be claiming, "There is a 99% probability of this point being within one kilometer of the given coordinate." On the other hand, the supplier could provide the same table and claim that there is only a 50% chance of the coordinate being within one kilometer.

A resolution is more useful than a grade in cases where the coordinate represents a point or very small area (such as a house). Resolution can be used to answer questions like, "What percentage of the file contains coordinates likely to be off by more than 500 meters?" After verifying a random sample of resolutions, statistical analysis of the resolutions in a data file provides an excellent mechanism for measuring overall quality.

Resolutions in centroid data

When the coordinate represents the centroid of a region, the resolution is often increased to factor in the radius (distance from the centroid to the farthest border) of the area. For example, a circular region 2 kilometers in radius that has been perfectly placed may be reported to have a resolution of 2 kilometers. This would be interpreted as, "Any point in this region is within 2 kilometers of the centroid." If the same centroid were placed with a one kilometer uncertainty, the reported resolution would have to be 3 kilometers. Note that this system gives large areas big resolutions regardless of their placement accuracy.

When the resolution is used to indicate the maximum distance between the centroid and any point in the region, it does not separately indicate the radius from the placement error. In this case a second metric, such as grade, must be used to indicate the quality of the placement of the centroid. It would also be possible to use a resolution to express the error in centroid placement and a radius to express the size of the region. Adding these will indicate the probable farthest point from the centroid.

CASE FOUR: CONFIDENCE

When coordinates are used to determine in what region a point is included, the resolution alone is not very useful in measuring accuracy. In Table 4, the same coordinates shown in the other tables have been compared to a set of regions (such as fire districts, sales territories, earthquake zones) and the region enclosing each coordinate has been displayed.

Coord.	Region	Grade	Resolution	Confidence
1	X	D	1.0 km	99%
2	Z	A	.041 km	51%
3	Y	A	.052 km	99%
4	Y	A	.052 km	99%
5	Z	B	.3855 km	82%

Table 4

The confidence is expressed, here, as a percentage. It can also be displayed as 1 to 10 index, or as "High", "Medium", or "Low". Whatever the unit, a high value shows that the assignment of a point to a region is likely to be correct. A confidence of 50% would indicate that, due to large resolution errors, closeness to a boundary, or both, the coordinates could be equally likely to be assigned to a different region. The question the user of this data would want to ask is, "How sure can I be that each coordinate has been assigned to the correct region?" Neither "The grade is C." nor "The resolution is 1 km." answers the question very well. "The confidence is 99%", is a very useful answer.

Confidence is calculated by intersecting the resolution area surrounding the coordinate with the region assigned to the coordinate. This is illustrated in Figure 1. For the purposes of illustration, assume that the definitions of the regions are mathematically perfect, but assume that the coordinates in the table have the realistic accuracies described above. Since the coordinates each have some error, the assignment of regions must also have some error. In this example, the resolution areas surrounding each coordinate are circles. Other shapes may be used to better represent the resolution errors around other types of data.

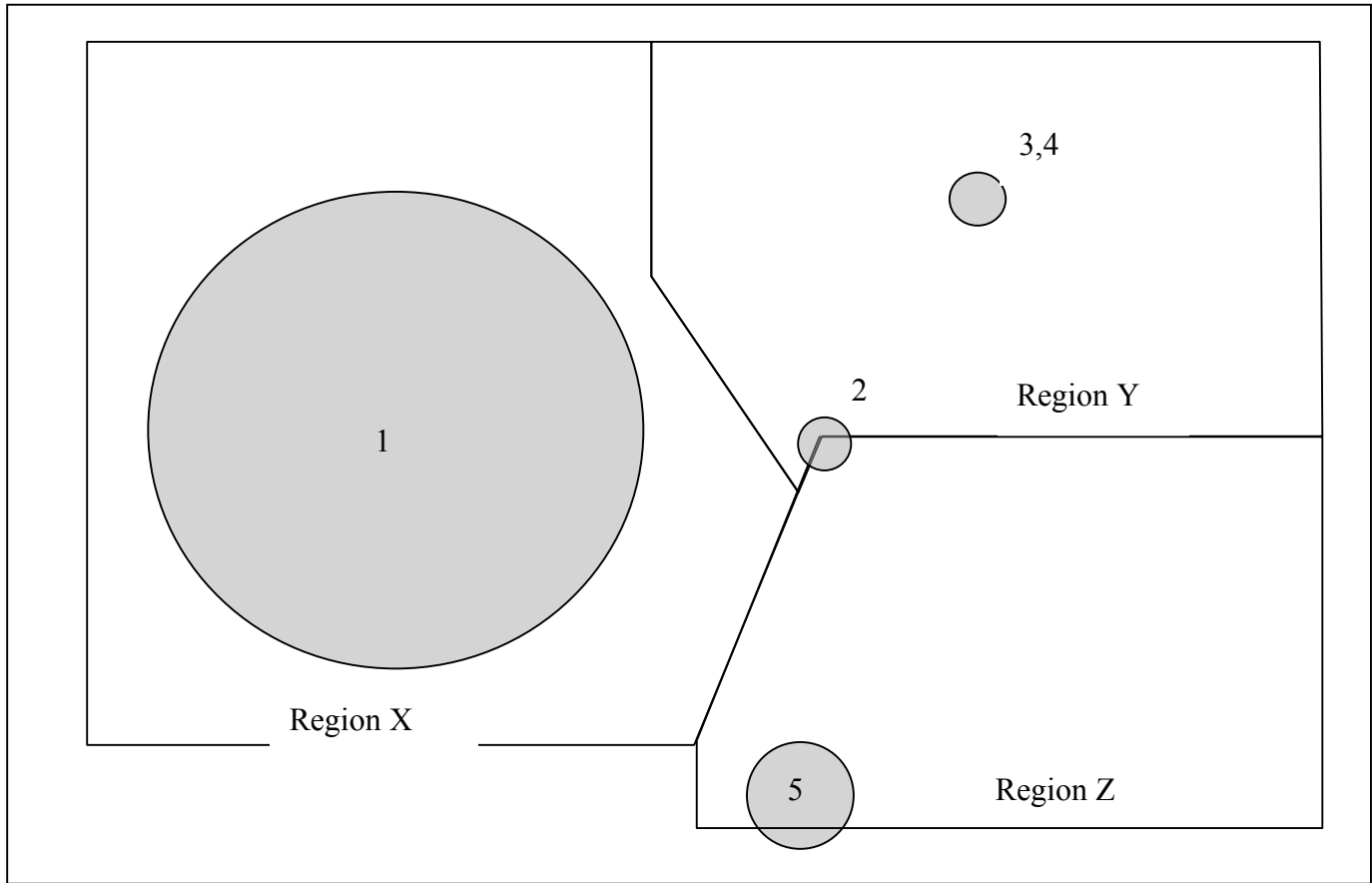


Figure 1

Coordinate 1 has a 99% confidence because all of the resolution area (the area that has a 99% chance of containing the true location of the coordinate) is completely within the region. Coordinate 5 has an 82% confidence because 82% of the circle defining the resolution area is within the region. Note that the confidence is not directly related to the grade or resolution.

Some coordinates with large resolution errors have high confidences. Rather, the low confidences are associated with coordinates whose resolution areas are large compared to their proximity to the border of the region.

SUMMARY: THE NEED FOR METRICS

In the preceding examples, the same data has been presented with different metrics. Counting decimal places would lead us to believe that coordinate 1 was the most accurate and coordinate 2 was the least accurate. The resolution shows that the exact opposite is true. The stacking in coordinates 3 and 4 could be an indicator of low quality, but again the resolution shows this not be the case. Resolution works well in determining the quality of positioning points, but in the assignment of coordinates to regions, it is a poor indicator. Neither the resolution nor the grade of coordinates 1 and 2 predicts the quality of region assignment.

The measurement of quality requires the proper metric for the task. To determine the quality of a coordinate file, each coordinate should have a resolution. If resolution is not available, grade is the next best choice. To determine the quality of a centroid file, each centroid should have two metrics, presumably a grade and a resolution. To determine the quality of coordinates assigned to regions, each assignment should have a confidence.

Without these metrics, all coordinate data is of "trust me" quality.