Experiments on the asymptotic behavior for some geometric data conflation algorithms

Carlos López-Vázquez
Technical University of Madrid, Spain. (*)

Data Fusion, Data Harmonization and Conflation are equivalent terms which denote the process of merging dataset A and dataset B to produce dataset C which hopefully has better properties than the original ones. Each term is popular within specific communities (Remote Sensing, Computer Science, Cartography, etc.) without a clear preeminence. This paper is devoted to Geometric Conflation, defined as the process of modifying coordinates of objects of (for example) dataset B in order to fit as much as possible those also available in dataset A, believed to be more accurate.

Recent proposals splits the process in three steps. The first one identifies all (or most) corresponding objects that appear both in A and B. Depending on the goal, usually such objects are only "well defined points" (like cross-roads) but the set might also include polylines (like roads) and/or polygons (like parcels). Second step is to report about the differences found among such homologue objects. Third step is to estimate a mathematical transformation which moves objects in A onto corresponding objects in B, exactly or approximately. Common practice considers only steps one and three, but there exist applications which requires just step one and two.

This paper is devoted to step three. We assume that corresponding objects (as many as possible) have already being identified, and our problem is to find the best mathematical transformation to apply to all objects in B in order to use them together with those of A.

The sought transformation should minimize some metric of discrepancy between objects in A and the transformed coordinates of its corresponding objects in B. The metric should consider a significant number of well distributed objects in order to be meaningful. We have followed cartography standards based upon RMSE of coordinates.

The transformation function should be estimated using part of the available corresponding objects (M out of N objects), and best practice suggests to left aside another part just to test the goodness of fit over N-M objects. We wish/speculate that this residual a) should diminish when the number of M increases and b) should not improve significantly if M>M0 (M0 large) thus showing an asymptotic behavior. Note that there exist well documented counterexamples in the numerical analysis literature, like the Runge's phenomenon for high order polynomials.

In the paper we describe the experiment designed to test our speculations. It requires a large number N of corresponding objects, a statistically sound procedure to perform the experiment, and a suitable function subspace where to look for the transformations. The large N cannot be achieved using just points, so we expand the set using *pseudo-homologue* points derived from polylines and polygons. We test up to M~10.000, being N~11.000.

Our findings indicate that fitting figures improves with the number of homologue points, but the process does not show an asymptotic behavior. We speculate that the reason might be that we are looking for the mathematical transform in a not-too-rich function subspace. Most methods in the geospatial interpolation literature prescribe the shape of the function (inverse distance weighting, polynomial functions (splines), Kriging, etc.) in a closed form, leaving almost no room to better fit the data. We tested them as well as slightly variations which satisfy other conditions (like being a conformal transformation) with no significant differences. Future research paths are also discussed.

(*) Also ORT University, LatinGEO Lab, Montevideo, URUGUAY