Arbitrarily re-composable hierarchies: modern geospatial science needs normal-form data structures

Mark Padgham, Michael Sumner and Angela Li

Modern GIS standards generally represent spatial data as nested lists, whether in accordance with the Simple Features (SF) standard of the Open Geospatial Consortium, or in `geojson` format. Most commonly used geometric libraries are based on one or both of these two standards. We argue that (1) the agreed representations in modern GIS geometry effectively restrict ongoing development of GIS as a whole, and (2) the enforced representation of geometry as nested lists as a central form is inefficient.

Simple Features

SF does not address what "non-simple" features are or might be, yet clearly these include important application domains such as GPS data, transport networks, point clouds, computer aided design, virtual and/or augmented reality, and 3D games. Each of these significant arenas have their own standards which are difficult to reconcile or unite without risking fragmentation and inefficiency.

SF and nested-list representations are limited because:

- Shapes are not represented as topological primitives and so internal boundaries are precluded.
- Shapes are represented as paths so only planar polygonal shapes are possible.
- Shapes may exist in XY[Z[M]] geometry, but this is not extensible, with no capacity to store data against component geometry elements.
- Shapes have no persistent naming of features or their components.
- There is no capacity for internal topology of shapes or within collections (no vertex-, edge-, or path-sharing).

These limitations mean that SF cannot fully represent every-day data forms from tracked objects, transport, Lidar, 3D models, statistical graphics, topological spatial maps, TopoJSON, CAD drawings, meshes or

triangulations. Translations between geospatial forms and the grammars of data science can be disjointed, relying on localized implementations that are lossy or inefficient, require third party workflows, or involve unnecessary tasks.

GIS applications generally diverge from common standards in different ways but none currently provide a normal-form model. There is no standard way to normalize data by detecting and removing redundancy (topology), or to densify data (a common necessity in planning domains). There is no standard way to extend the types although complex forms are well established in other domains.

Arbitrarily re-composable hierarchies

The common "well-known" formats of encoding geometry (WKB/WKT for binary/text) represent (pre-)aggregated data, yet the input levels of aggregation are often not directly relevant to desired or desirable levels of aggregation for analysis. A key stage in many GIS analyses is thus an initial disaggregation to some kind of atomic form followed by re-aggregation.

We propose a common form for spatial data that is inherently disaggregated, that allows for maximally-efficient on-demand re-aggregation (arbitrarily re-composable hierarchies), and that covers the complexity of geometric and topological types widely used in data science and modelling. We provide tools in R for more general representations of spatial primitives and the intermediate forms required for translation and analytical tasks. These forms are conceptually independent of R itself and are readily implemented with standard tabular data structures.

There is not one single normal form that should always be used. There is one universal form that every other model may be expressed in, but also other forms that are better suited or more efficient for certain domains. We show that conversion between these forms is more straightforward and extensible than from SF or related types, but is also readily translated to and from standard types. The most important forms we have identified are "universal" (edges and nodes), "2D primitives" (triangles), "arcs" (shared boundaries), and "paths" (normalized forms of SF types).

Further details are available at https://github.com/hypertidy/silicate.