



Unlocking the Mysteries of the Bounding Box

Persistent URL for citation: <http://purl.oclc.org/coordinates/b1.pdf>

Douglas R. Caldwell

Douglas R. Caldwell (e-mail: Douglas.R.Caldwell@erdc.usace.army.mil) is employed as a cartographer and geospatial analyst at the US Army Engineer Research & Development Center, Topographic Engineering Center, Research Division, Information Generation and Management Branch, 7701 Telegraph Road, Alexandria, VA 22315.

Date of Publication: 08/29/05

Abstract

Few geospatial data representations are more basic than the bounding box; a rectangle surrounding a geographic feature or dataset. Bounding boxes are a key component of geospatial metadata and lie at the heart of many computational geometry algorithms as well as spatial indexing systems. Despite their ubiquity and common use, bounding boxes are more complicated than they first appear. The phrase that ‘spatial is special’ applies to this humble representation as well as to more sophisticated geospatial representations. This paper explores the nuances of correctly understanding, using, and interpreting bounding boxes.

Keywords: Bounding box, Minimum Bounding Rectangle (MBR), metadata, map projection, geographic information systems, GIS

Introduction

The bounding box, also known as the Minimum Bounding Rectangle (MBR) [\[1\]](#) or envelope [\[2\]](#) is “A rectangle, oriented to the x and y axes, which bounds a geographic feature or a geographic dataset. It is specified by two coordinates: xmin, ymin and xmax, ymax.” [\[3\]](#) (See Figure 1)

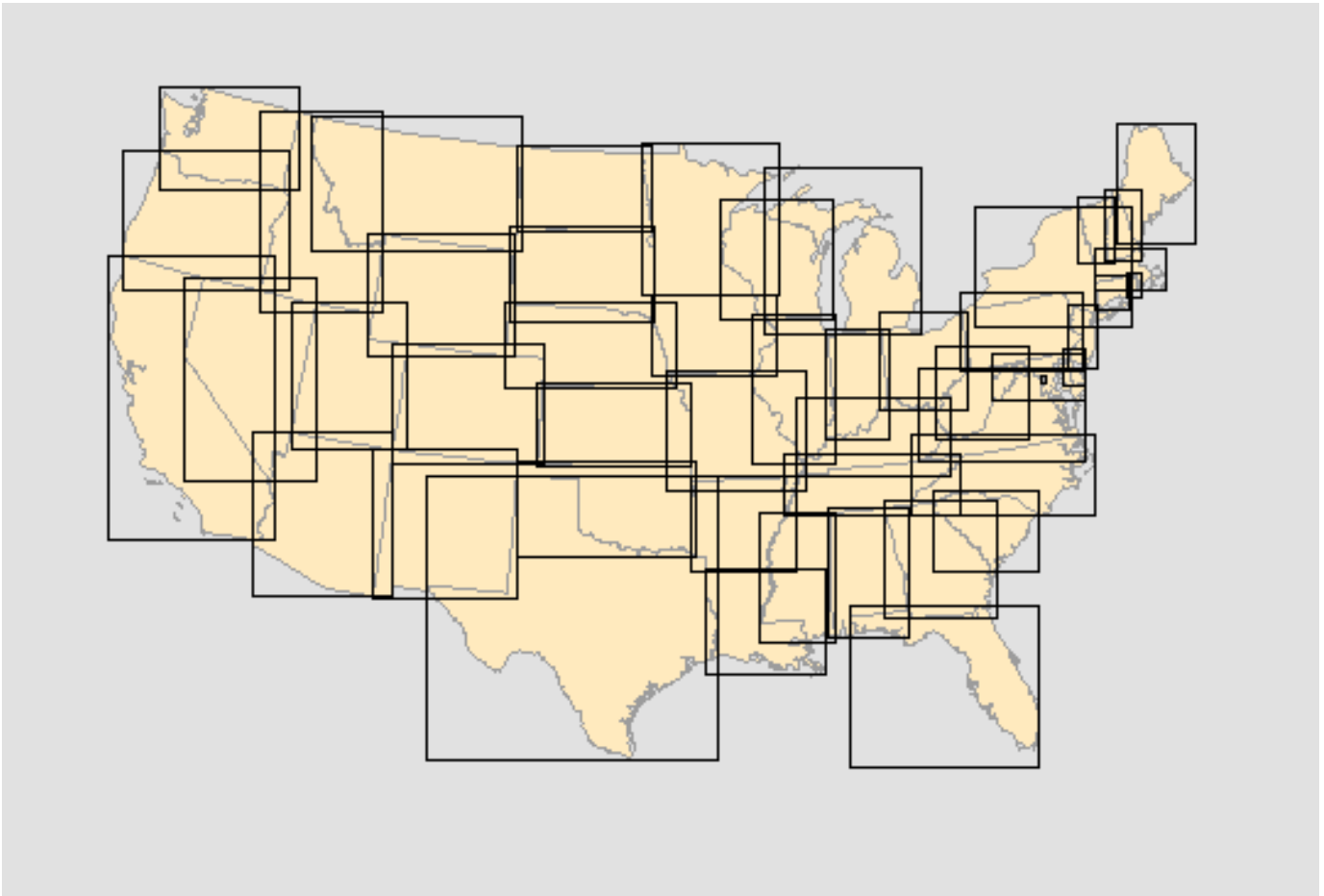


Figure 1. State Bounding Boxes for the Continental United States. The map is drawn in an Albers Equal Area Projection and the bounding boxes were generated from the projected data.

Bounding boxes are one of a number of bounding container shapes. Additionally, there has been work done utilizing other types of spatial footprints, such as the bounding diamond, the minimum bounding parallelogram, the convex hull, the bounding circle or bounding ball, and the bounding ellipse. [4] In many situations, these alternatives may more tightly bound a geospatial feature, yet despite these advantages, “The bounding box is the computationally simplest of all linear bounding containers, and the one most frequently used in many applications.” [5]

Bounding boxes lie at the heart of many computational geometry applications, such as ray tracing, collision avoidance, and hidden object detection. [6] Spatial indexing schemes, such as R-trees, use minimum bounding rectangles to subdivide space. [7] These applications typically hide the implementation of bounding boxes from the user.

Bounding boxes are visible to geospatial data users in metadata, where they are one of a number of methods for describing the extent of a dataset. Other methods include a textual description or name, geocode, point, or other polygon. [8] Bounding boxes specify the extent or limits of their associated features. They also serve as an approximation for the areal coverage of the feature. The major metadata standards all include the concept of a bounding box.

The Content Standard for Digital Geospatial Metadata, from the Federal Geographic Data Committee

(FGDC), defines a bounding box under the Spatial Domain heading as "bounding coordinates":

Bounding Coordinates - the limits of coverage of a data set expressed by latitude and longitude values in the order western-most, eastern-most, northern-most, and southern-most. For data sets that include a complete band of latitude around the earth, the West Bounding Coordinate shall be assigned the value -180.0, and the East Bounding Coordinate shall be assigned the value 180.0 [\[9\]](#)

ISO 19115:2003(E), Geographic information – Metadata defines bounding boxes under the EX_Extent entity, which records the spatial and temporal extents of the data. This element falls under the identification information (MD_Identification) and is mandatory under certain conditions. EX_GeographicBoundingBox is defined in terms of geographic coordinates (longitude and latitude) that bound the limit of the dataset extent. [\[10\]](#)

The Dublin Core Metadata Initiative (DCMI) defines identifiers for a place in the Coverage element, which can include the DCMI Bounding Box:

The DCMI Box encoding scheme is a method for identifying a region of space using its geographic limits. Components of the value correspond to the bounding coordinates in north, south, east and west directions, plus optionally up and down, and also allow the coordinate system and units to be specified, and a name if desired.

We identify a place by considering the minimal rectangular box which fully encloses the place, whose faces are aligned parallel with the axes of an identified Cartesian coordinate system. [\[11\]](#)

Despite their ubiquity, simplicity and common use, bounding boxes may not be well understood by many users. The phrase, "spatial is special," [\[12\]](#) applies to this humble representation. This paper explores the nuances of correctly using, interpreting, and understanding bounding boxes through the examination of four problems: the Content Quandary, Global Gotchas, Projection Problems, and Approximation Assessment.

Content Quandary

The bounding box is supposed to represent "the limits of coverage of a data set," but the meaning of limits is not clearly specified in the Content Standard for Digital Geospatial Metadata. The ISO 19115 standard and DCMI standard are similarly unclear. The bounding box can be interpreted as either the extent of the data collection area or the extent of the data records in the data set. These are both useful, but present significantly different views of the bounds of the data set.

In the example in Figure 2, we have a map of the imaginary Snagglehuff distribution in the continental United States. There are twenty Snagglehuffs, all located in the southeastern United States. The bounding box for the data collection area covers the continental United States, while the bounding box for the actual data in the data set covers a smaller area in the southeastern corner of the country.

Snagglehuffs in the Continental United States



Legend

- Snagglehuffs
- Data Collection Area Bounding Box
- Data Bounding Box

Figure 2. Comparison of Data Bounding Box with Data Collection Area Bounding Box.

Users interested in determining whether Snagglehuffs are found in their area of interest would prefer the smaller bounding box (the red box in Figure 2), which shows the extent of the data. However, this presents a partial picture of Snagglehuff distribution and does not accurately represent the data collection area, which was the continental United States (the blue box in Figure 2). When presented with a bounding box showing the extent of the data in the Snagglehuff dataset, a user interested in learning about Snagglehuffs west of the Mississippi River might believe that there is no information about Snagglehuffs outside of the southeastern United States. While there may be no Snagglehuffs located in other parts of the country, it is important to recognize that the dataset was collected for the entire continental United States.

The distinction between the spatial extent of the data and the spatial extent of the data collection area is rarely clearly identified and stated. One exception to this is the Biological Data Profile of the FGDC. To go along with Bounding Coordinates specifying the extent of the data in the dataset, the profile specifies a mandatory Description of Geographic Extent element to provide a text description of the extent of the study and/or data set. [13] Techniques for automatically calculating metadata, similar to those used in ESRI's ArcGIS software, calculate the spatial extent of the data records. These would have to be manually

overridden if the extent of the dataset does not match the extent of the data in the dataset.

Unless the extent of the data in the dataset happens to exactly match the extent of data collection area, the two will differ. This was the case in the Snagglehuff distribution map. It is important for users to understand the content of their bounding boxes, as the information lends itself to answering different types of questions.

Global Gotchas

While there is no beginning or end on a globe, digital spatial data sets have an artificially defined beginning and end. Longitude extends from 180 degrees west (-180) to 180 degrees east (+180) of Greenwich, United Kingdom, and latitude extends from 90 degrees south (-90) to 90 degrees north (+90) of the Equator. This artificial segmentation of geographic coordinates results in a 'Global Gotcha' for bounding boxes of features spanning the 180-degree meridian.

Consider the imaginary country of Boxtopia, which has a southwest corner at (170, 40) and a northeast corner at (-170, 50). The width of this box is 20 degrees. In a spatial database, Boxtopia would be represented by two rectangles, one which has a southwest corner at (170, 40) and a northeast corner at (180, 50) and a second that has a southwest corner at (-180, 40) and a northeast corner at (-170, 50). This split is required because the longitude must be between -180 and 180 degrees.

Since the bounding box is a single feature that extends from the minimum to the maximum longitude and latitude values, the bounding box for Boxtopia has a southwest corner at (-180,40) and a northeast corner (180,50). This bounding box has a width of 360 degrees, rather than 20 degrees, exaggerating by eighteen times the width of the Boxtopia.

The Global Gotchas can be clearly seen in Figure 3, where Russia, the United States, Kiribati, Fiji, New Zealand, and Antarctica span the 180-degree meridian. The bounding box for Antarctica is a good approximation, as the continent spans the globe. In the other cases, the bounding box is a poor approximation of the feature extent. There are also problems when the bounding box does not actually cross the 180-degree meridian, but contains parts, which straddle the line. These bounding boxes are smaller than 360 degrees, but clearly much larger than they should be. Searches based on these extents may return large numbers of irrelevant results.

Unfortunately, no simple and elegant solution exists to solving the Global Gotchas. Multipart bounding boxes are a possible alternative, but they add complexity to the database and search process, defeating the simplicity of the bounding box.

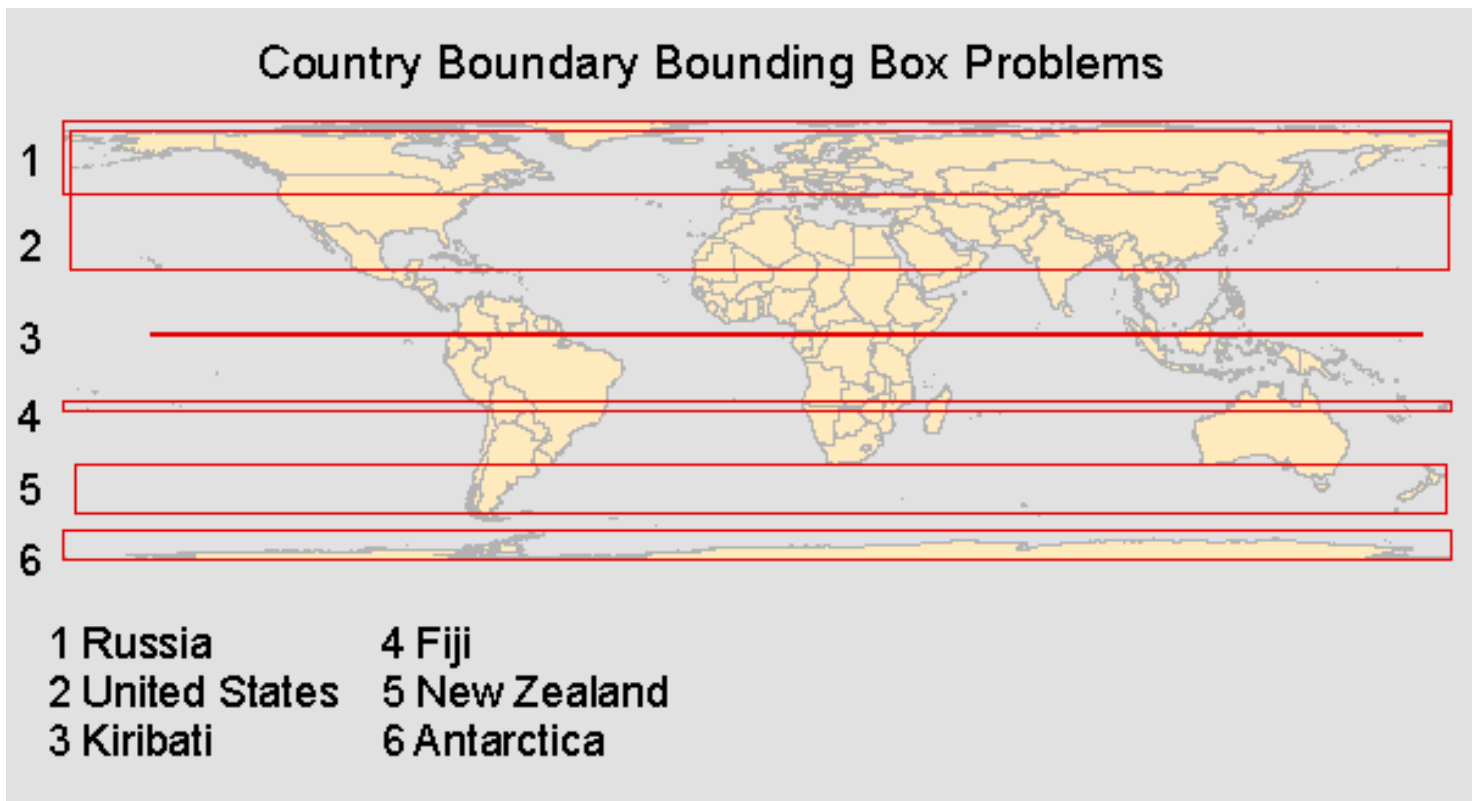


Figure 3. Bounding Boxes for Countries and Continents Which Straddle 180 Degrees.

Projection Problem

The projection problem occurs when projecting a bounding box from its original projection/datum to a new projection/datum. It is quite possible to discover that the bounding box no longer 'bounds' the feature, making the extent information invalid. Instead, the bounding box may intersect the feature (See Figure 4).

Continental United States Bounding Box Projected from Geographic to Albers Equal Area

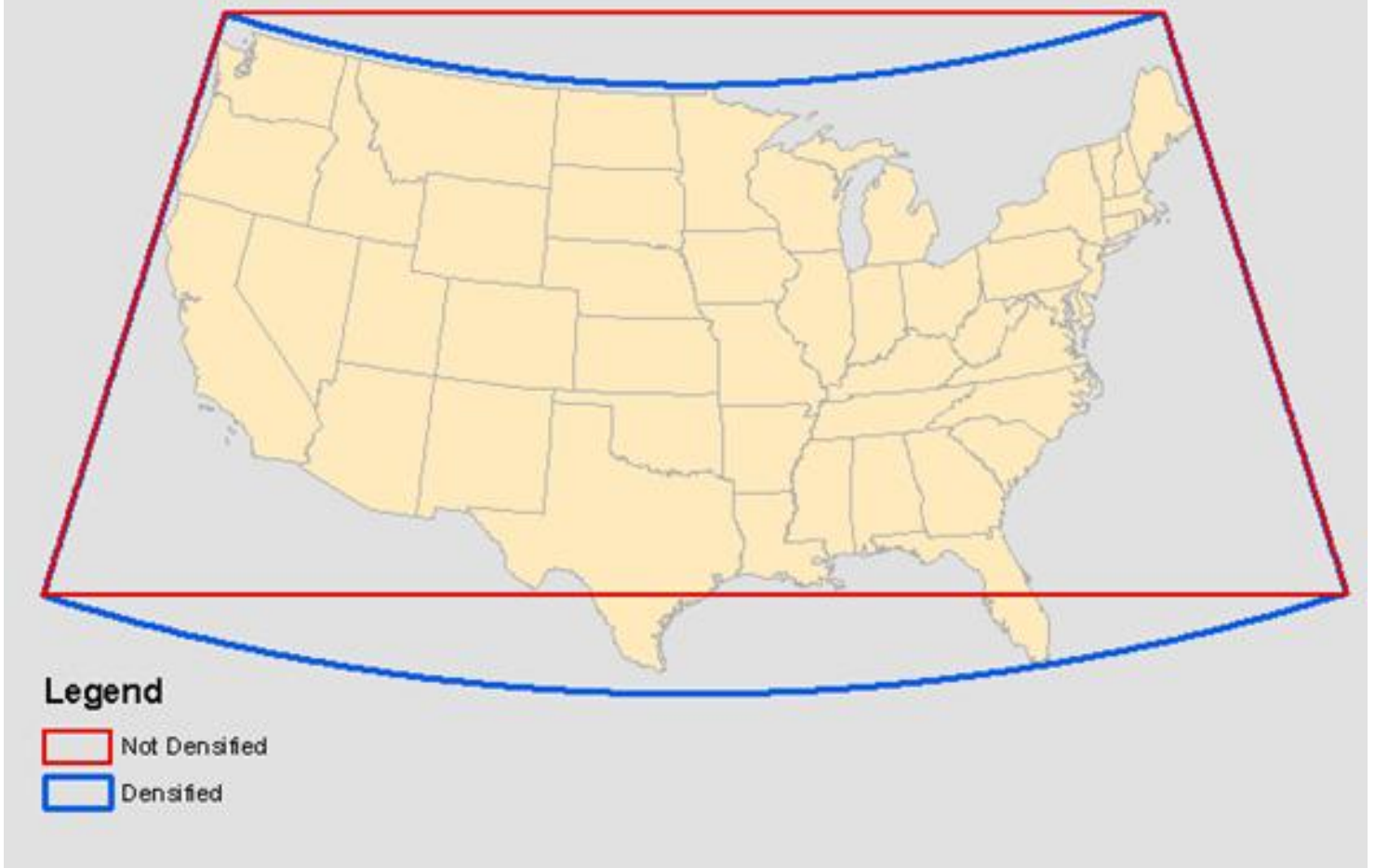


Figure 4. Bounding Boxes for a Dataset with Geographic Coordinates Projected to Albers Equal Area. The red box shows the effects of projecting the box defined only by the four corner coordinates. The blue box has been densified with additional vertices.

The Projection Problem is caused by the fact that a bounding box, defined by corner coordinates, is an undersampled representation of the true bounding box. It only provides information at the corner locations, but no information along the lines connecting the corners. When a bounding box defined by the four corners is projected, the lines connecting the corners remain straight lines. This means that queries against this box may have an incorrect extent and risk missing areas that should be included, a significant problem. In addition, areas outside the original bounding box may be included in the search. This is less of a problem, as the areal coverage of a bounding box is already understood to be an approximation greater than the area of the feature. In the example in Figure 4, the bounding box for the United States misses the southern tips of Florida and Texas.

There are two potential solutions to the Projection Problem. The first solution is to add vertices to

the bounding box before projecting it. This ‘densification’ supports a more accurate representation of the projected boundary and should be done whenever using the bounding box to determine the extent of the data. This solution is appropriate when the original data are no longer available. If the data are available, a second solution is to generate a new bounding box after reprojecting the data.

Approximation Assessment

The Approximation Assessment, or measure of how well bounding boxes approximate the coverage of a feature, is the final issue. This is especially important for applications involving bounding boxes used to estimate the area of a feature, as poor approximations will lead to larger numbers of non-relevant results. Approximation effectiveness is analyzed using the Bounding Box Factor, which is the ratio of the area of the bounding box to the area of the feature. The Bounding Box Factor ranges from 1, where the bounding box and the feature are identical, to infinity, where the bounding box is infinitely larger than the feature.

In order to better understand the range of values for the Bounding Box Factor, tests were run on a three different datasets, representing political and natural features at multiple levels of aggregation. These included datasets for Census Tracts, Ecoregions and Hydrologic Units. Because the Bounding Box Factor can change for different projections of the same feature or dataset, all the data was projected to an Albers Equal Area projection to allow the direct comparison of areal measurements.

Census Tracts

The Census Tract dataset is the 2004 Edition of the U.S. Census Tracts produced by Geographic Data Technology for ESRI and distributed on the ESRI Data and Maps CD-ROM (See Figure 5). It covers all 50 states and the District of Columbia. Data were analyzed at four levels: Census Tract component parts, Census Tracts, Counties, and States. Some Census Tracts are multipart features, so they were broken into their component parts to analyze data at its atomic level.

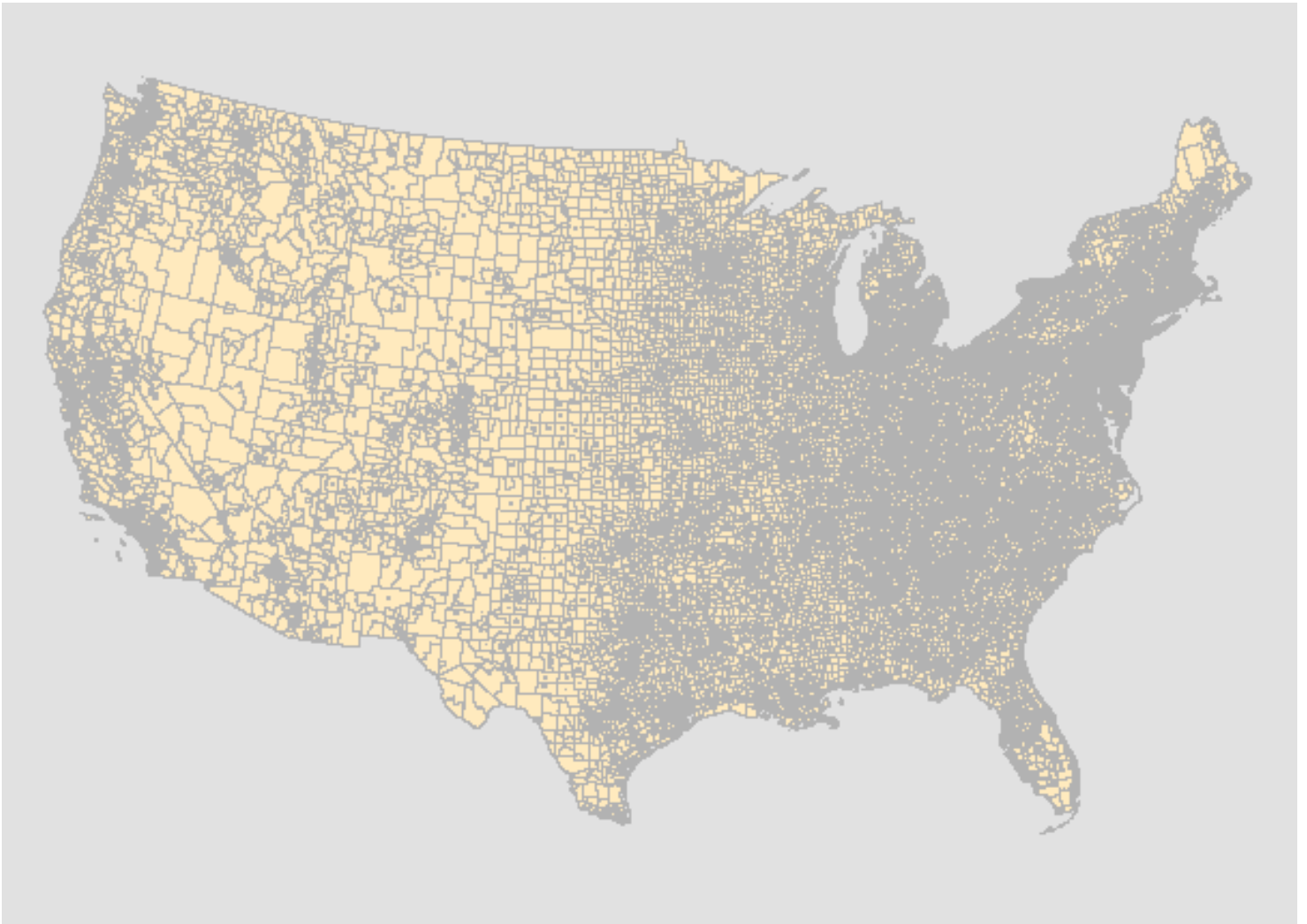


Figure 5. Map of Continental US Census Tracts (2004) – Tracts.

Census Tracts (2004)		Bounding Box Factor			
Geography Level	Feature Count	Minimum	Maximum	Mean	Standard Deviation
Tracts (Component)	66906	1.003001	43.402738	1.870394	0.772512
Tracts (Multipart)	65344	1.003001	3158.839174	1.938390	13.185470
Counties	3141	1.004461	42.077043	1.609442	0.897699
States	51	1.083702	11.852433	2.085811	1.533791

Table 1. Bounding Box Statistics for Census Tract Data

The Census Tract Bounding Box Factor data are reported in Table 1. The data has the lowest overall mean values of all datasets for the Bounding Box Factor. This is not unexpected, as Census

Tracts are designed by humans using guidelines that place an emphasis on compactness. [14] The minimum Bounding Box Factors are close to 1, meaning that the bounding box very closely approximates the shape of the feature. This dataset has the largest maximum value for the Bounding Box Factor. This occurs at the multipart, unpopulated Census Tract with a FIPS Code of 09009000000. This Census Tract, located in New Haven County, Connecticut, has two parts separated by more than 20 miles, with a total area of 0.049 square miles and a bounding box area of 156.212 square miles (See Figure 6).

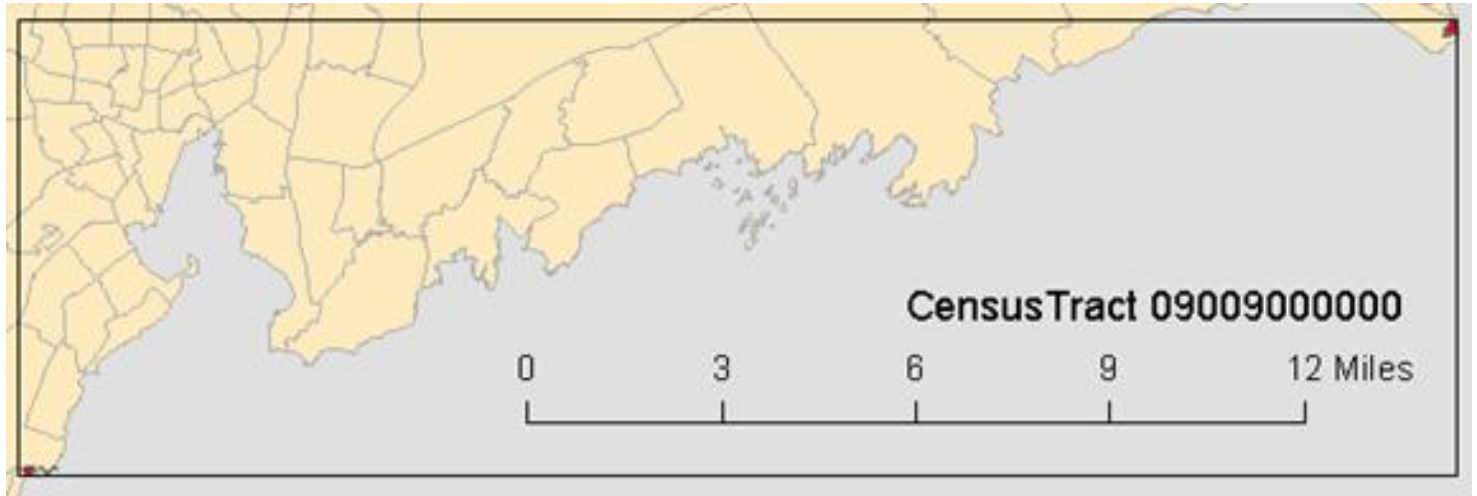


Figure 6. Multipart Census Tract 09009000000 Shown With Bounding Box.

Although they are difficult to see, the two small features shown in red at the southwest and northeast corners of the bounding box are the components of the Census Tract.

Ecoregions

The Ecoregions dataset is the USDA Forest Service dataset for ‘Ecoregions and Subregions of the United States, Puerto Rico, and the U.S. Virgin Islands,’ published in 2004 (See Figure 7). [15] According to the metadata accompanying the dataset:

This data set shows ecoregions, which are ecosystems of regional extent, in the United States, Puerto Rico, and the U.S. Virgin Islands. Four levels of detail are included to show a hierarchy of ecosystems. The largest ecosystems are domains, which are groups of related climates and are differentiated based on precipitation and temperature. Divisions represent the climates within domains and are differentiated based on precipitation levels and patterns as well as temperature. Divisions are subdivided into provinces, which are differentiated based on vegetation or other natural land covers. The finest level of detail is described by subregions, called sections, which are subdivisions of provinces based on terrain features.

The dataset covers all 50 states and the District of Columbia. Some Sections are multipart features, so they were broken into their component parts similar to the Census Tract data.

