Visualization of Spatial Data

CS&SS 569 Breakout Group Report

Nan Tang, Lauren Woyczynski, Tao Lin, Chenxi Di
Ozgur Ozkan, Summer Ai, Shihao Han

February 24, 2020

Contents

1 Introduction to Spatial Data

1.1 Types of Spatial Data

Spatial data is commonly used across many fields. To be able to apply techniques for the visualization of spatial data, you must first understand what a spatial dataset is. Spatial data comes in several common forms: areal data, raster data, and point data. Figure 1 shows pictoral example of each. Point data has GPS coordinates; this is data that is located at a specific point in space. Areal data, sometimes known as polygon data, shows a single value for a certain area. In maps, this is often some sort of political boundary, like a country, state, or county. Areal data is stored in a shapefile, a data type that has information about geometry and geographic coordinates of the shape. Shapefiles have several attributes. Each polygon within a shapefile has a geometry attribute and can also have any number of data associated. For example, a shapefile of US states might have attributes such as the state name, the population, and the median income. Raster data has one data point per pixel. Raster datasets can exist at any square resolution, such as 10m x 10m or 1mi x 1mi. Satellite data is often stored as a raster dataset. Rasters are most often used to present surfaces of data across large areas at a high spatial resolution.

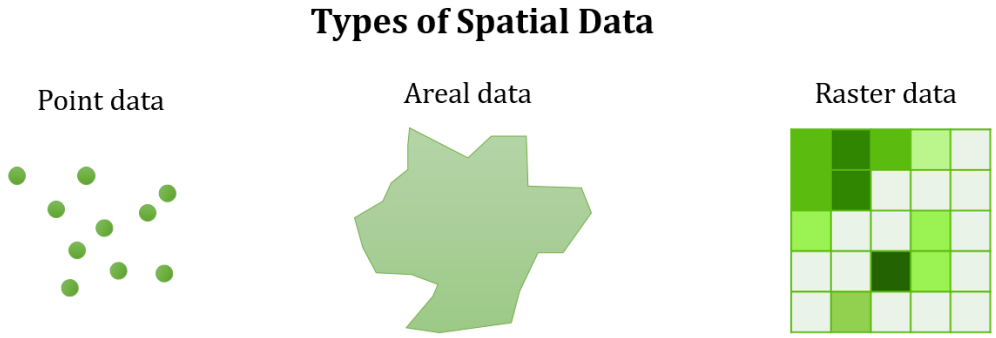

Figure 1: Examples of Spatial Data Types

1.2 Geographic Information Systems (GIS)

Geographic data is often stored in a special type of database, called a geographic information system (GIS). A GIS can be used as a framework to gather, manage, and analyze spatial data. There are several common softwares that are used to work with a GIS, the most common being ArcGIS. QGIS is a online free version of this. R also has many packages for

working with spatial data. Some popular ones include `ggmap` , `rgdal`, `rgeos`, `maptools`, and `sp`.

1.3 Common Methods to Visualize Spatial Data

Depending on the types of spatial data and the information wanted to be presented, there are different ways to visualize spatial data.

Choropleth: One of the most common methods is choropleth, in which different colors are used to represent certain distribution of a feature in different regions. It makes intuitive sense and usually works well with areal data. However, it might be misleading sometimes as big regions generally attract more attention than small regions.

Cartogram: This kind of graph is often used to avoid illusion brought by different sizes of regions. In cartograms, a thematic mapping variable, such as population, will substitute for land area or distance so that the map will distort in proportion to the mapping variable. Although it addresses some shortcomings of other types of visualization, it is important to remember that cartograms should be used with care. The audience should be able to understand the distorted graph instead of getting confused.

Heat map: Heat map works especially well with continuous variables because a spectrum of colors can be used to describe the distribution. Because heat map assigns the weight of each given data point with certain color, it makes regions with high concentrations easily stand out, which great helps with identifying geographical pattern.

Dot map: As suggested by the name, point map is suitable to point data. By using dots to indicate the presence of a variable, dot map reveals certain spatial pattern through the scatter of the data points.

There are many more types of maps that are regularly used in the visualizations of spatial data, such as cluster map and flow map. In discussing the potential problems of representing spatial data, the current paper used some of the maps mentioned above, each with its own advantages and disadvantages.

1.4 Common Challenges to Visualize Spatial Data

There are three common challenges in the visualization of spatial data and spatial designs. First is the problem of visualizing temporal variation. The second problem is the dimensionality problem and the visualizing multivariate data. The third challenge is the issue of uncertainty and how to deal with it.

In dealing with visualization of temporal-spatial data there are there common methods at researchers' disposal: a. Small Multiples b. Heat Maps c. Interactive Design and Animation

The two commonly embraced approaches for dealing with the issue of dimensionality and multiple variables in spatial visualization are as follows: a. Ternary maps b. Choropleths and Dot Maps

Each of these approaches have their advantages and disadvantages. Researchers have to make concessions among preattentive legibility of data, data density, and spatialiaty of the visual. These problems, solution alternatives, and their advantages are discussed below.

2 Spatial-temporal Visualization

Spatial-temporal Visualization is a technique to illustrate changes in an area over time on a map. During our discussion, we have thought of two main ways to represent time on a map: 1) the use of small multiples to create snapshots for showing changes on map with time; 2) the use of interactive map to create animation that allow readers to discover patterns among spatially and temporally scattered events.

2.1 Small Multiples Techniques

The term small multiples was popularized by Tufte (1983), these multiple static graph will illustrate change over time through a set of static map snapshot. Since they do not require a platform for animation, they are most popular in published paper and other static reports. While they are easy for readers to understand, they do need take up a lot of pages and to be scaled to be readable at very small size, especially for a large amount of time slices. One thing we have particularly discussed is how to visualize the pattern change over a long time period, for example 100 years, when we were given the data at each year. The strategy is to truncate the large period equally into some smaller intervals such as 10 years, or unevenly split it according to the characteristics of data.

Figure 2: Small Multiple showing the temporal-spatial variation of drought incidences in the US

2.2 Heat Maps

Heat maps allows incremental observation of spatial patterns. They are useful when the temporal scope at concern is wide. Heat maps allows visualization of multivariate data as well. They allows the visualization of both categorical and continuous data. However, heat maps require trading off the spatial characteristics of the data. This approach is useful especially when communicating the geographically bound nature of data is not primary concern. Figure2 is a heat map visualizing the incidence of measles across 50 states of the US over a period of 75 years.

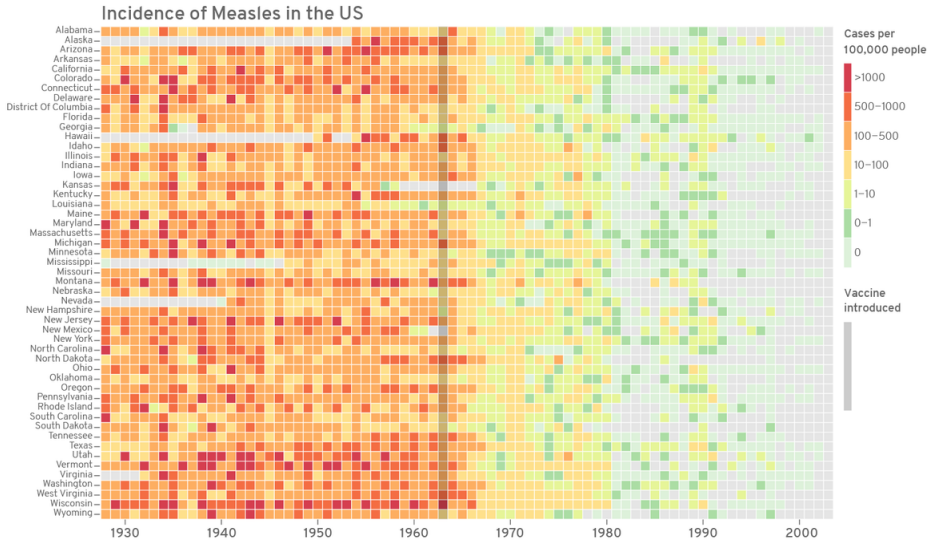

Figure 3: Heat Map showing the temporal-spatial variation of the Measles incidences across the US.

2.3 Interactive Graph Techniques

Map animations could be created with a variety of different software such as Shiny. Besides providing a way to dynamically represent the spatio-temporal data, the interactive graph would also offer sliders for readers to view the snapshot of a particular point in time that they want to know about by sliding the slider. However, even with the interactivity, readers may also experience cognitive difficulties such as change blindness.

For example, when visualizing the spread of infectious disease (2019-ncov - the novel coronavirus disease initially discovered in Chinese city of Wuhan) through time and space, the dynamic graphs have the capability to recognize and track the changes in the complex system. The current static dashboard built by Center For Systems Science and Engineering at John Hopkins University did great job on identifying specific number of confirmed and recovery cases (the left side chart) and other attributes in different location. However, when presenting the change, especially when the infectious number dramatically increase as time moves forward, the red bubble illustrating the magnitude of confirmed cases in one province will essentially overlap with the other one in the nearby provinces, thus, creating difficulties to identify specific patterns. Figure2 is the dashboard visualizing the disease spread on Feb 3. Figure3 is the dashboard visualizing the disease spread on Feb 23 as the confirmed cases exploded.

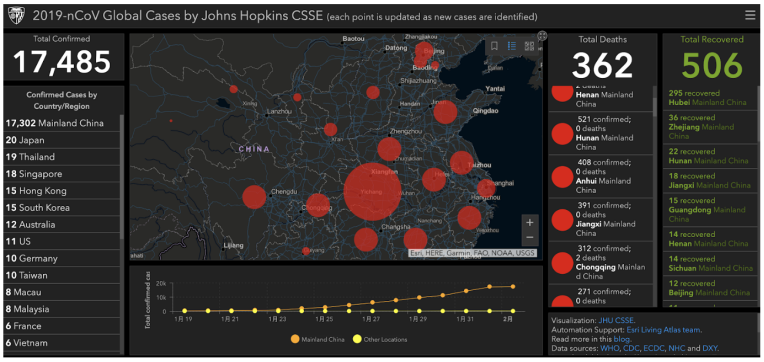

Figure 4: Date: Feb 3, 2019-ncov dashboard, Center For Systems Science and Engineering at JHU

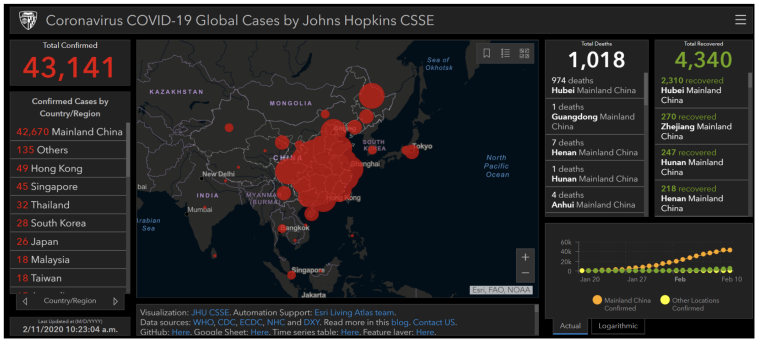

Figure 5: Date: Feb 23, 2019-ncov dashboard, Center For Systems Science and Engineering at JHU

Therefore, to better visualize the time and space variation simultaneously, we thought that an interactive dashboard including a dynamic choropleth map, a time slider which allows viewer to see a snapshot of a particular point, a time series trend graph and some appropriate labels or measures. However, another question was raised during the discussion: since the data is strongly skewed, the choice of color scale becomes especially important. Usually, we employ equally spaced color cutoffs, but this will inevitably blind all the changes across time and space. One solution raised during discussion is to choose the color bins based on Quantiles so that the variation would be highlighted.

3 Visualizing Spatial Categorical Data

In descriptive epidemiology, there is often a desire to display data by as many facets as possible, so that readers can look for trends across dimensions all in one figure. Of course, the more dimensions that are displayed, the more complex a figure is. Sometimes, however, data has multiple facets of importance to the reader that are hard to display in one figure. One such example is spatial data that is categorical or proportional in nature. Maps are often displayed best as chloropleths or cartograms. Neither of these are particularly conducive to categorical data. One such dataset in presented in Wiens et al's 2018 paper, "Global variation in bacterial strains that cause tuberculosis: a systematic review and meta-analysis". Here, the authors extract data from 206 global studies on the distribution of TB strain lineages by country. The strain distributions vary spatially, as well as the total sample size in each country. In Figure 4 below, the authors use pie charts on a world map, where there is a pie for each country with data and a color in the pie for each possible Lineage (8 possible). The diameter of the pie corresponds to the sample size in that country.

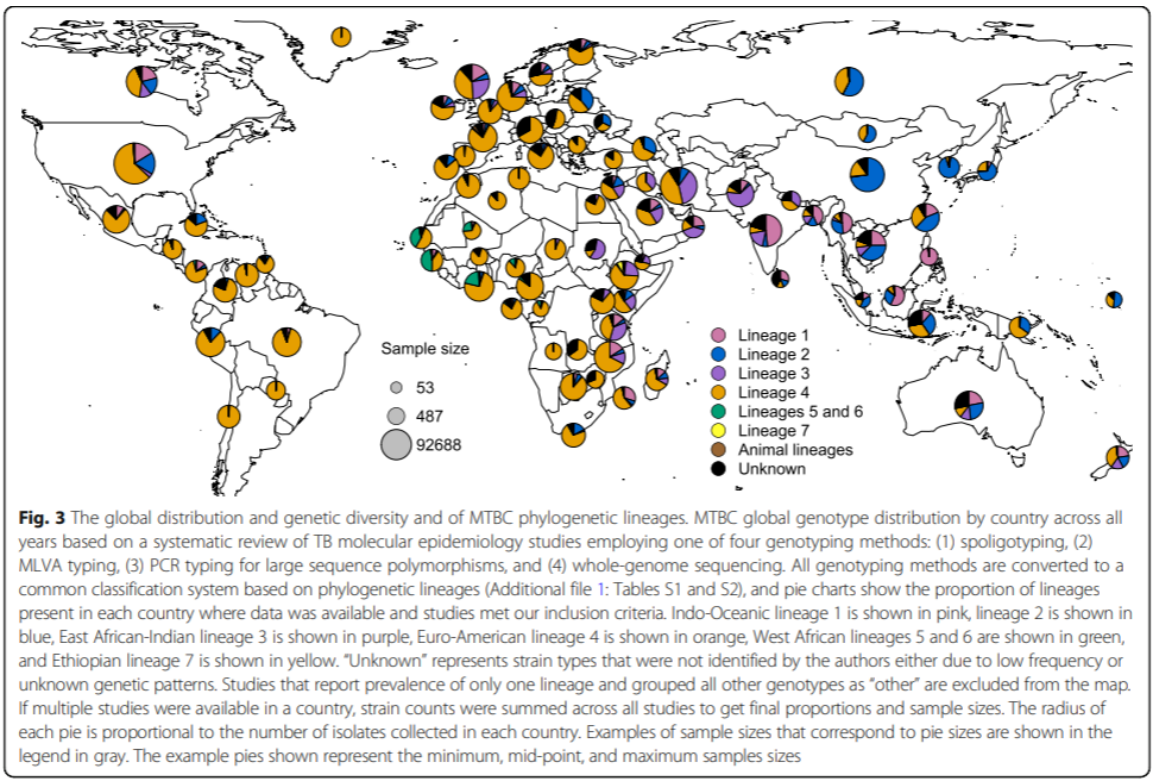

Fig. 3 The global distribution and genetic diversity and of MTBC phylogenetic lineages. MTBC global genotype distribution by country across all years based on a systematic review of TB molecular epidemiology studies employing one of four genotyping methods: (1) spoligotyping, (2) MLVA typing, (3) PCR typing for large sequence polymorphisms, and (4) whole-genome sequencing. All genotyping methods are converted to a common classification system based on phylogenetic lineages (Additional file 1: Tables S1 and S2), and pie charts show the proportion of lineages present in each country where data was available and studies met our inclusion criteria. Indo-Oceanic lineage 1 is shown in pink, lineage 2 is shown in blue, East African-Indian lineage 3 is shown in purple, Euro-American lineage 4 is shown in orange, West African lineages 5 and 6 are shown in green, and Ethiopian lineage 7 is shown in yellow. "Unknown" represents strain types that were not identified by the authors either due to low frequency or unknown genetic patterns. Studies that report prevalence of only one lineage and grouped all other genotypes as "other" are excluded from the map. If multiple studies were available in a country, strain counts were summed across all studies to get final proportions and sample sizes. The radius of each pie is proportional to the number of isolates collected in each country. Examples of sample sizes that correspond to pie sizes are shown in the legend in gray. The example pies shown represent the minimum, mid-point, and maximum samples sizes

Figure 6: Global Distribution of Tuberculosis Strain

This figure has several flaws that make this approach to displaying these facets ineffective. Firstly, each pie is too small to see the distinct colors besides the piece that is taking up the largest portion. Therefore, the pie does not serve the purpose of showing the distribution in this example. Additionally, using diameter to indicate sample size is notably ineffective for pie charts, since readers are often bad at comparing areas, especially with such small pies. However, this is a very interesting and useful data set with definitively spatial patterns. A question remains about whether this data needs to be visualized on a map, or whether the spatial components of this dataset can be displayed through other visualization methods.

For datasets with two to three categories, binary and ternary plots can be used. I attach Figure 5 and 6 here as examples. Figure 5 shows both uncertainty and child mortality, with one axis for each. This approach shows both the indicator and the associated uncertainty by having one dimension of color represent the width of the uncertainty level for that unit. These 16 distinct colors are mapped to second-level administrative units in low and middle income countries. Figure 6 shows educational attainment in three bins in India and Nigeria. Panels a, b, d, and e use the package `tricolore` to create these ternary plots, that show the distribution of points within three categories, as well as create a ternary color key that is

then mapped onto administrative units. Both of these approaches take some effort from the reader to orient, but do a good job of displaying a large complex dataset and simultaneously being compact and delivering a lot of information. Both Figure 5 and 6's approaches are chloropleths, but display several dimensions (2 and 3, respectively), instead of the single dimension of data that color traditionally represents.

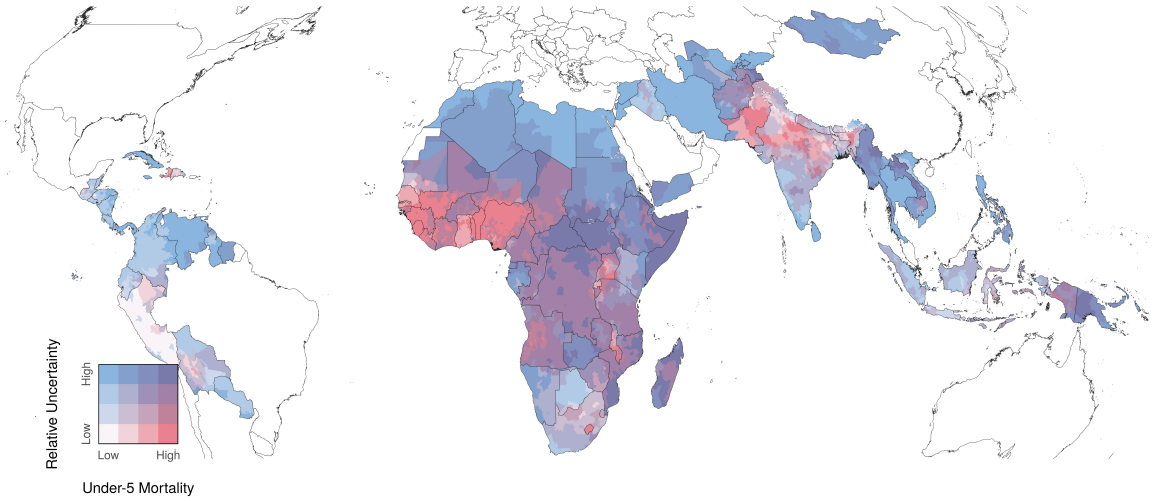

Figure 7: District-level Under-5 Mortality and Associated Uncertainty in Low- and Middle-Income Countries

Overall, these alternative approaches fix some, but not all issues presented with the original figure. Effectively displaying categorical spatial data is a complex task that may not be able to be compacted to a single figure. Overall,the use of multi-dimensional color scales and small multiples offer promise in achieving this. Descriptive epidemiology offers the public health field large datasets ripe with information that is disaggregated across many dimensions. These types of data have wide appeal, and thus being able to display many dimensions of the data at once can be a very powerful tool for display of results. As mentioned in previous sections, these large datasets are often best served by interactive web displays that allow the use to toggle between several settings, where single facets can be focused on and compared one at a time. This approach offers clarity in each display, flexibility to the user's specific interests, and ability for design and animation that is not possible in print.

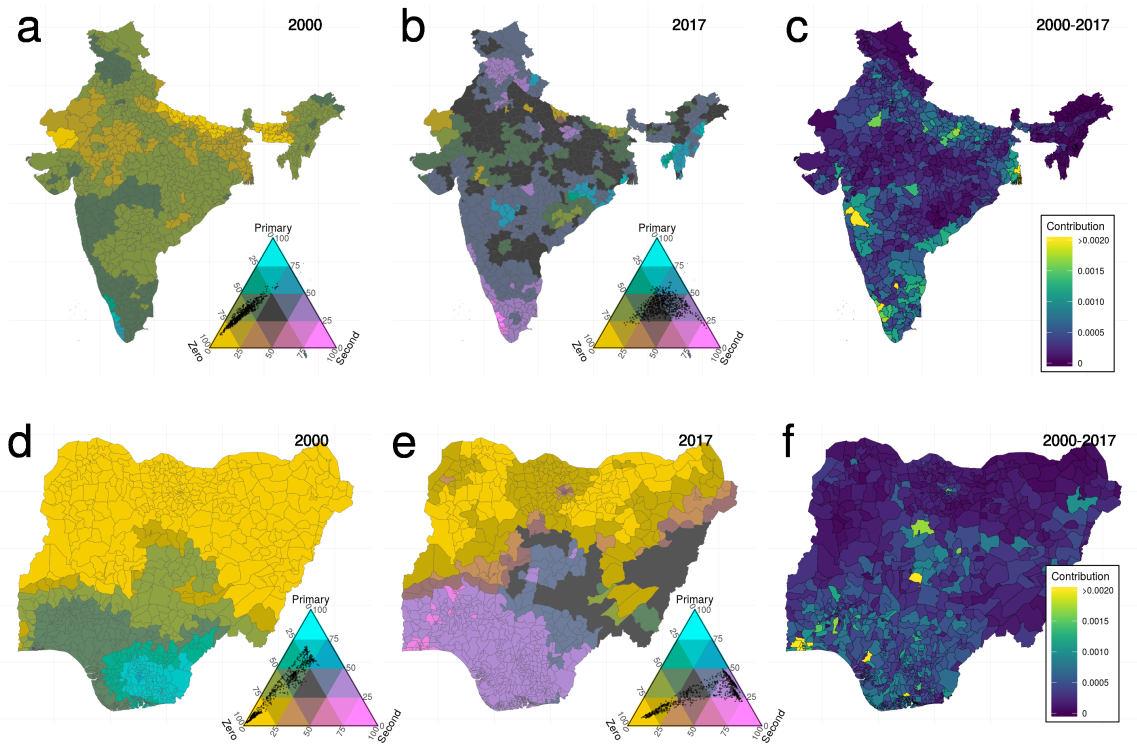

Figure 8: Distribution of Educational Attainment in Nigeria and India

4 Mapping Uncertainty

4.1 Spatial Data Uncertainty

Data uncertainty can be broadly divided into two categories, measurement error, and variation of data. When a map is created from GIS software, information loss is unavoidable. Measurement error directly lead to the difference between real geographic phenomenon and user's understanding. Therefore, informing viewer with accuracy and completeness of the spatial data that have been applied on visualization is important. Variation on event data is another cause of uncertainty. Mostly in research of environment and prediction, people tend to use mean value as representation of the event, for example, the mean daily temperature. It is obvious that temperature varies within the day, hence displaying uncertainty is a critical step to improve completeness of data communication.

4.2 Strategies to Represent Uncertainty

Whether adding uncertainty of data into the graph as additional information depends on kinds of the problem to solve. Generally, when working on model of prediction or using mean

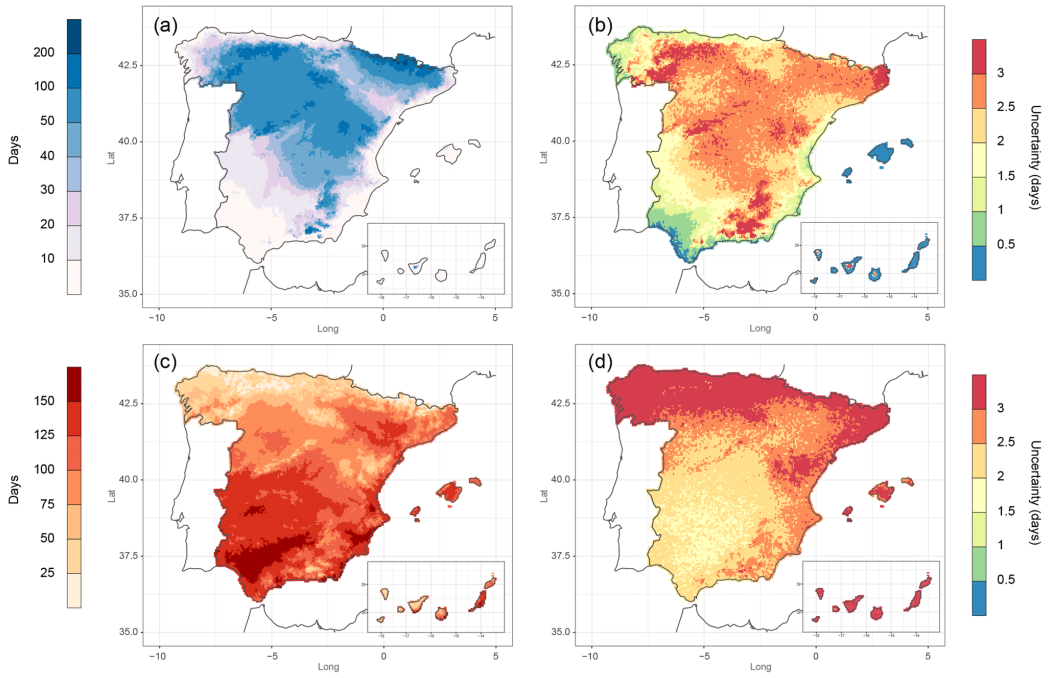

Figure 9: Uncertainty of Annual Frost and Summer Day

value as data point, visualizing uncertainty could generate new perspective and understanding.

In general, methods to display uncertainty are categorized as intrinsic and extrinsic mapping. Common intrinsic strategies to display levels of uncertainty is referring to the uncertainty visualization cube, where each corner represents one way of distinction. They are distinctive color value (fade out and fade in), texture (coarser or denser), transparency and fuzziness. These methods are not universally acceptable to all problems, in cases of more than two dimension of data, the addition layer that represents uncertainty might confuse viewer. In contrast, extrinsic strategy is the preferable choice for most of the cases. As the name means, extrinsic strategy utilizes adjacent views (paired maps) or small multiples which prevents extra dimension of data from cluttering with original views. Figure 7 is an example of mean annual number of frost day around Spain. Righthand side displays mean values of frost and summer days, while righthand side are their corresponding uncertainty. In this case, uncertainty are range of the data.

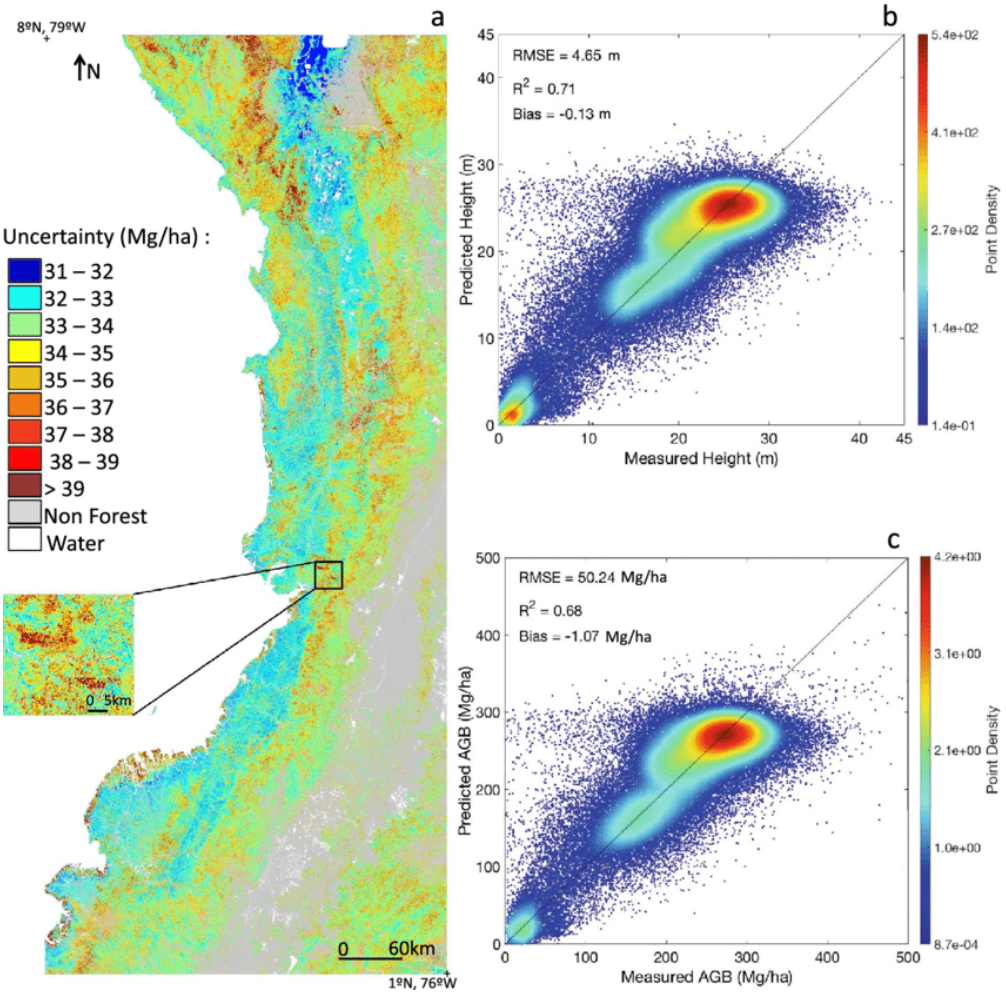

Figure 10: Uncertainty of AGB Map

Using interactive system for large scale of map enables user to check the variation and bias for spacial data. Figure 8 shows the example of displaying prediction model of Above Ground Biomass (AGB) map takes into account error of random foresting model and the lidar model.

5 Conclusion

In many cases, the use of maps may restrict effective visualization the data at hand due primarily to the structural limitations discussed above. In this sense, it is important to ask ourselves if our data is really spatial. Even if our data is collected or categorized through spatial units we may not really need to use a map to visually communicate it. In most cases the geographically bound nature of data could still be retained even when we choose not to use maps.