Chart Families

Ashish Anand

1 Objective

I am making this rolling[1] tutorial discussing different chart types. In visualization task, we face two primary questions:

- what is the primary objective of the visualization: whether we want to explore data distribution; whether we want to show hierarchical relations or temporal pattern etc.
- what are the different possible charts one can use for the identified objective and which chart type among all possible options will be the most appropriate choice for conveying the message in a comprehensive manner.

Aim of the tutorial is to help you in answering the two questions based on the organization of chart families as discussed in [1].

2 Chart Families

Different charts offer distinct ways to represent and visualize data. In [1], Kirk has categorized the charts into **five** main families based on their primary objective. The five families are -

1. **Categorical:** For comparison of categories and distributions of quantitative values
2. **Hierarchical:** Revealing *part-to-whole* relationships and hierarchies
3. **Relational:** Exploring correlations and connections
4. **Temporal:** Plotting trends and intervals over time
5. **Spatial:** Mapping spatial patterns through overlays and distortions

Kirk has given a five letter mnemonic **CHRTS** representing these chart families [1].

Now, we will discuss representative charts from each of the families. It is noteworthy to mention that -

[1] I will keep updating this tutorial

- some charts are likely to fall into more than one families, however, it was kept in one which suits the primary objective of the chart most. It does not mean that the particular chart cannot or should not be used in other scenario. We will highlight this aspect while discussing such charts.

- you can add additional *presentation* aspects such as "making it interactive", "adding texts and legends" etc.

Since, during the course of assignments, we already have explored several chart types belonging to **categorical** family, I will start the tutorial with discussion of **temporal** family.

2.1 Temporal Family

Charts from this family are primarily used to show trends, events/actions or activities at a set of time-intervals. Let us look at some of the examples.

Line Chart

A line chart displays how quantitative values have changed over time for different categorical items. A line chart is drawn around a **continuous temporal x-axis** and **quantitative y-axis** with values plotted at relevant coordinates. Connecting lines join up adjacent and related categorical items to form slopes which are then extended along the full timescale.

Fig 1 shows the the cumulative runs scored in Test matches by English Test batsmen between 1947 and 2018.

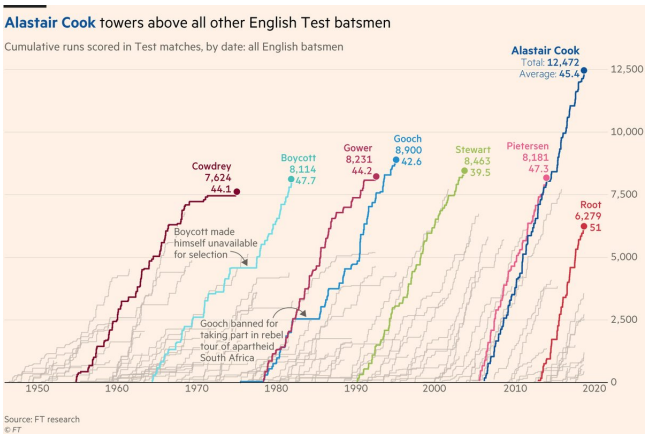

Figure 1: Runs scored by English Test batsmen between 1947 and 2018. Source: Andy Kirk, Chapter 6 [1]

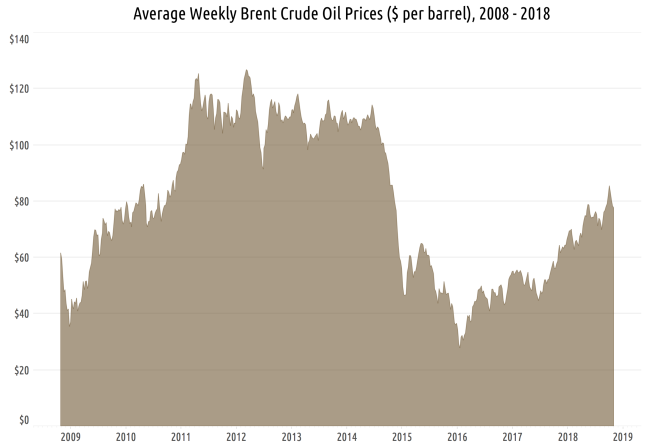

Figure 2: Crude oil price for Brent (US$ per barrel) 2008-2018. Source: Andy Kirk, Chapter 6 [1]

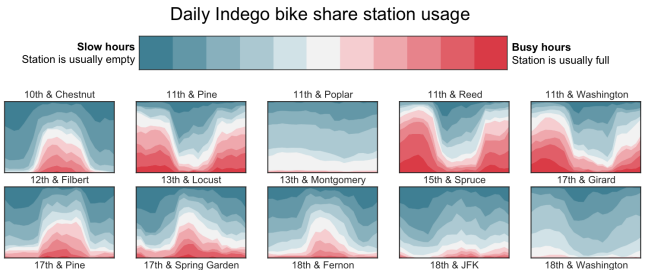

Figure 3: Source: Andy Kirk, Chapter 6 [1]

Area Chart

Area chart is an extension of line chart, with the area between the x-axis (representing time) and the line (quantitative measure).

Fig. 2 shows changes in the average weekly price (US$ per barrel) of Brent crude oil between 2008 and 2018.

Stacked Area Chart

Stacked area chart is an area chart displaying how quantitative values have changed over time for multiple categorical items. The idea is very similar to stacked bar chart and allows visualization of evolution of both total and contribution of component sub-categories. However, if there are too many subcategories, then one can think of using **Area Chart** with **faceting**.

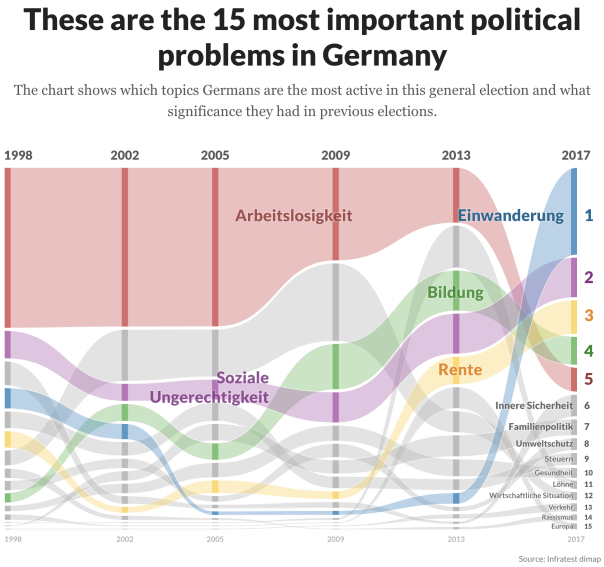

Figure 4: Changes in rank of the most politically important topics for Germans between 1998 and 2017. Source: Andy Kirk, Chapter 6 [1]

Bump Chart

Bump chart is an alternative to Line chart, when we want to show how quantitative values, in terms of *ranking measurement*, have changed over time for different categorical items. In this case, we keep rank on the Y-axis and X-axis represents continuous time scale.

Bump chart usually uses variation in the size (width) of each line to represent a quantitative measure, usually the absolute value informing the ranking measurement (Fig. 4).

Slope Graph

Slope Graph is used to show trend of change of quantitative values during the two time points for different category items. The graph consists of two parallel quantitative axes with a common value range. A line is plotted connecting the two axes together with the vertical position on axes representing the respective quantitative values. These connecting lines form slopes that indicate the upward, downward, or stable trend between the two temporal axes.

Fig. 5 shows an example of slope graph comparing the changes in the share of power sources across US states between 2004 and 2014.

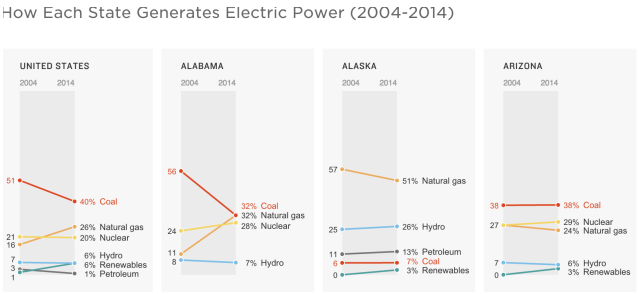

Figure 5: How US states generates power from different sources: Coal, gas, nuclear, hydro? Source: Andy Kirk, Chapter 6 [1]

Stream Graph

Stream Graphs display the changes in data over time of different category items. It uses flowing, organic shapes that somewhat resemble a river-like stream. The axis that a Stream Graph flows parallel to, is used for the timescale.

In a Stream Graph, the size of each individual stream shape is proportional to the values in each category. Often more than one attributes of the category item are shown. The area occupied is filled with an attribute of color to represent another categorical attribute. For example, in the Fig. **??** changes in the total domestic gross takings (in US$) and the longevity of all movies released between 1986 and 2008 are shown. Height of the stream indicates weekly box office revenues and width shows longevity at the box office. The area (and associated color) indicates total domestic gross. Instead of plotting values against a fixed,

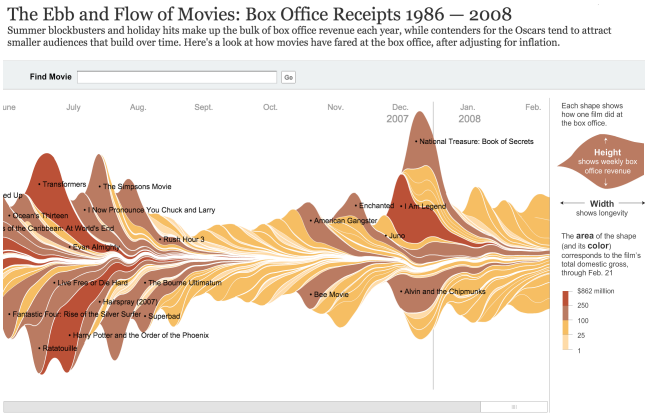

Figure 6: Changes in the total domestic gross and the longevity of all movies released between 1986 and 2008. Source: Andy Kirk, Chapter 6 [1]

straight axis, a Stream Graph has values displaced around a varying central baseline.

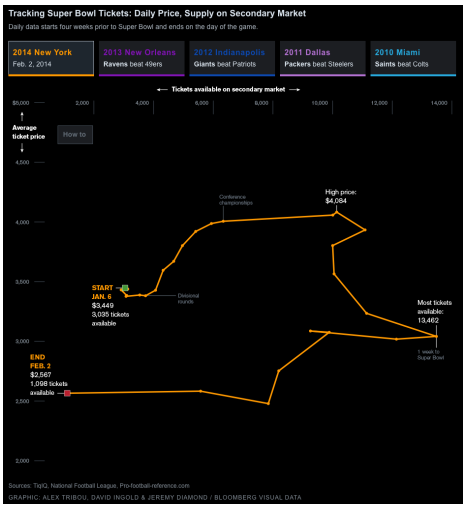

Figure 7: Changes in the daily price and availability of Super Bowl tickets on the secondary market. Source: Andy Kirk, Chapter 6 [1]

Category items may be plotted up and down the implied y-axis, in order to optimise the layout and not to reflect any notion of positive or negative values.

Connected Scatter Plot

Connected scatter plot is an extension of joined version of *line chart* and *scatter plot*. It has two variations. In one of the variation, *line chart* and *scatter plot* are combined together to show the relation between two quantitative variables.

The second variation is used to compare two quantitative variables over a time series. Thus, as an input this variation requires three variables. The chart is formed of two quantitative x- and y-axes, and with the values represented by point marks at the respective coordinates, one for each measurement over time. The individual points are then connected using lines joining each consecutive time points to form a sequence of change.

Fig. 7 displays changes in the daily price (variable 1) and availability of Super Bowl tickets (variable 2) on the secondary market four weeks prior to the event across five Super Bowl finals.

Gantt Chart

Gantt chart is used to visualize sub-tasks/sub-events of an event or a project over time. Typically, x-axis is used to represent the continuous time intervals and sub-tasks/sub-events on the y-axis. Corresponding to each item on the y-axis, size (length) of a horizontal bar indicates the duration of the activity. Horizontal bars corresponding to the different activities may have different start

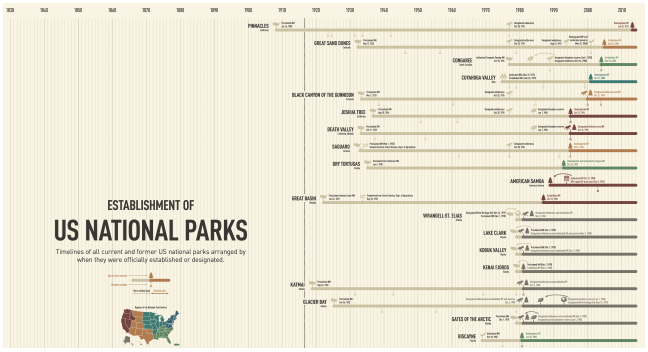

Figure 8: Establishment of the U.S. National Parks. Source: Andy Kirk, Chapter 6 [1]

and end time points. It can be further extended to visualize dependencies among several sub-tasks.

2.2 Spatial Family

This section describes some of the common charts used to visualize spatial data.

Choropleth Map

Choropleth maps are used to display a data variable associated with distinct, and defined geographical areas or spatial regions. Each region is defined by a polyglonal shape and all regions together define a target landscape or geographical area. For example, individual states representing regions and together they represent a country. State-wise *Population density* or *per capita income* can then be visualized by using a **choropleth map**. Colour attribute is used to represent quantitative measurements or categorical variables for all regions.

Depending on the aim of the visualization, a specific color progression should be used. For example, *single-hue progression* is commonly used to represent magnitude. The dark shade of the chosen color fade to a very light or white shade of relatively the same hue.The darkest hue is typically mapped to the greatest number in the data and the smallest number is represented by the lightest hue.

Two important things to remember while choosing choropleth map for visualization:

1. Geographical area is well defined, such as countries, states, counties etc.

2. The target measurement/variable is directly associated with and continuously relevant across the spatial region. In other words, the raw data should not be used and it must be normalized to reveal rates, aggregate statistics or ratios. Let us understand this with an example. If we want to

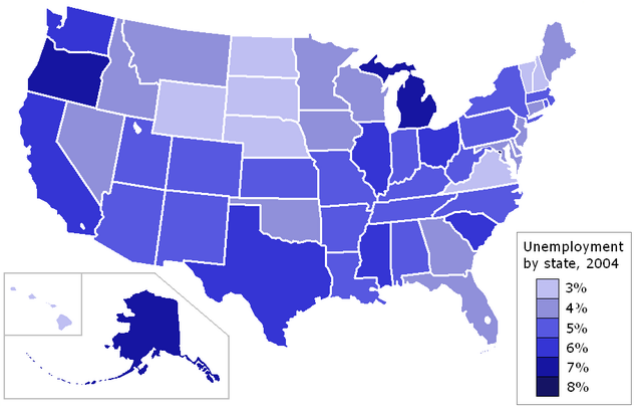

Figure 9: Unemployment in US in 2004. Source: https://tinyurl.com/ycjujzyt

visualize population distribution across different states of a country, then instead of using population, one should map population density of each state to have a meaningful visualization. If we use population then the states in bigger in area are likely to have more population than smaller states and visualization does not reflect the true picture.

Proportional Symbol Map

Similar to *choropleth map*, **proportional symbol map** is also used to represent quantitative variable associated with a priory well defined geographical areas. **Proportional symbol map** uses proportionally sized shapes (circles, bars etc.) positioned with the centre mid-point over a given location coordinate. One can add colour attribute to represent further categorical distinction.

Proportional symbol map is useful when raw data needs to be visualized or when raw data cannot be transformed into ratio or proportion (a requirement for **choropleth map**). Fig 10 shows the fund raised for Hillary Clinton across the USA.

Prism Map

Prism Map generates a 3d view corresponding to a quantitative variable over geographical regions. The quantitative values are represented proportionally in the form of lines or bars. Attributes of colour are also used to emphasize certain characteristics such as large values etc.

Fig 11 shows the population of trees for each 180 square km of land across the globe.

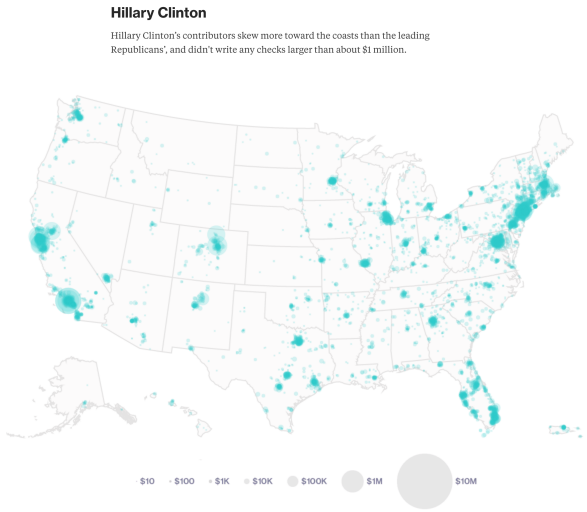

Figure 10: Mapping fund amounts raised across the USA for Hillary Clinton during the first half of 2015.

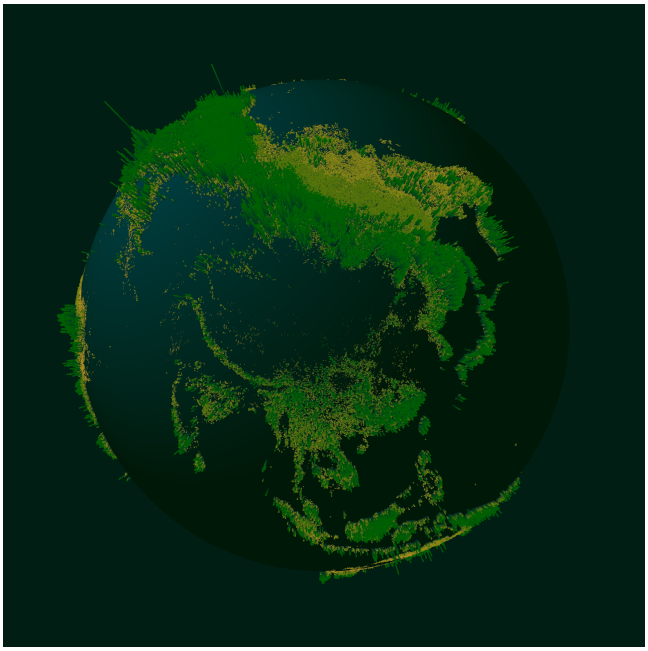

Figure 11: Trees across the globe. Source: Andy Kirk, Chapter 6 [1]

Dot Map

Dot Map is used to display spatial distributions or densities of data representing a large number of single objects. As the name suggests, equally sized dots

are used in the visualization. There are two types of *Dot Maps*: **one-to-one**, where one dot or point represents a single count or object,e.g. 1 dot = 1 tree; **one-to-many**, where one dot or point represents a particular unit of the object, e.g. 1 dot = 100 people.

Dot Maps are good in revealing spatial patterns when the points cluster over geographical regions in the map. Fig 12 shows each USA resident on their location during the 2010 Census across different ethnicity.

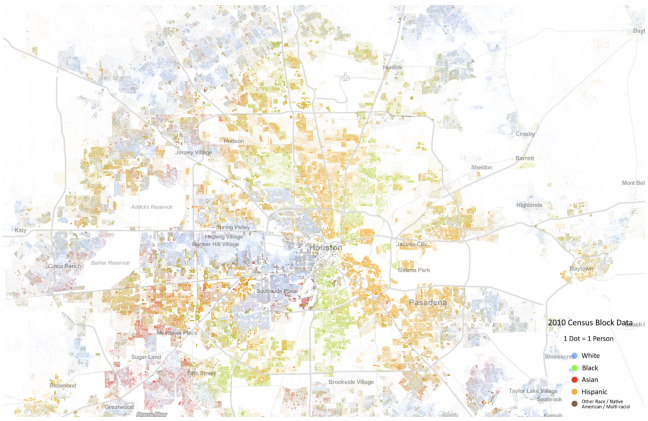

Figure 12: The Racial Dot Map. Image Copyright, 2013, Weldon Cooper Center for Public Service, Rector and Visitors of the University of Virginia (Dustin A. Cable, creator) Source: https://demographics.coopercenter.org/racial-dot-map

Isarithmic Map

Isarithmic Map displays distinct spatial surfaces on a map sharing the same quantitative classification. In contrast to the *Choropleth Map*, spatial boundary is not defined by geopolitical boundaries, rather it is regions sharing a certain quantitative value or interval scale. These regions are formed by connecting points of similar measurement. The relevant quantitative value are colour coded and regions are encoded with the respective colours. Such map is good to show continuous phenomena such as temperature, pressure or precipitation etc. https://wiki.gis.com/wiki/index.php/Isarithmic_map is a good resource to know about *isarithmic map* and its variants.

Fig 13 shows temperature values across the US. It uses temperature data recorded at individual weather stations across the country and then interpolation is done to obtain estimated values for other locations [2]. Boundaries are drawn upon the map to represent zones in which the temperature is assumed to be the same and then color, value, or saturation is added to enhance the map [2].

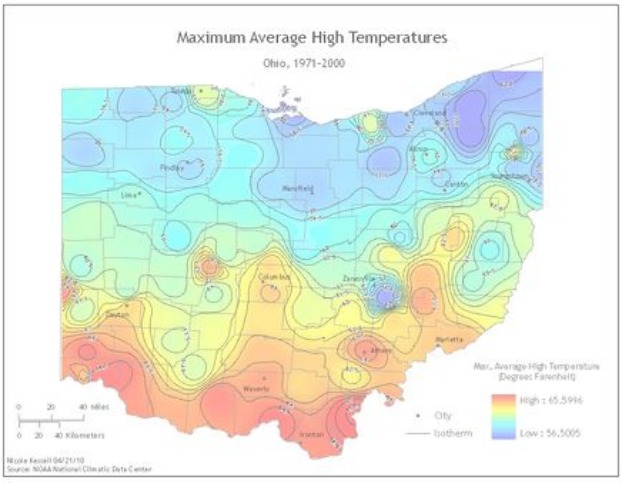

Figure 13: Caption. Source: https://tinyurl.com/y7h3puc2

Flow Map

In contrast to the other spatial maps, *Flow Map* displays characteristics of movement from one spatial region (location) to another, or connections between spatial regions. For example, *flow maps* are used to show community travel patterns (very useful for Covid-19 spread pattern analysis), movement of goods transportation, traffic volume etc. Such maps can show both qualitative and quantitative data.

Flow maps use lines to show the movement of objects (e.g. people, goods) between locations. The quantity of flow can then be represented by the width of the line. Additionally, flow map generally displays characteristics of origin and destination (positions on a map), route, directions (using arrow or tapered line width) and categorical classification (colour).

An example of the flow map is shown in the Fig 14. This figure shows the average number of vehicles using Hong Kong's main network of roads during 2011.

Fig 15 shows another example of flow map displaying transmission of diseases from one geographical region to another. You can find similar examples of Covid-19 transmission at many websites.

Relevant R Packages

A few relevant R resource for spatial maps and data analysis:

1. Resources at https://r-spatial.org. A good resource for analyzing spatial and spatial-temporal data using R.

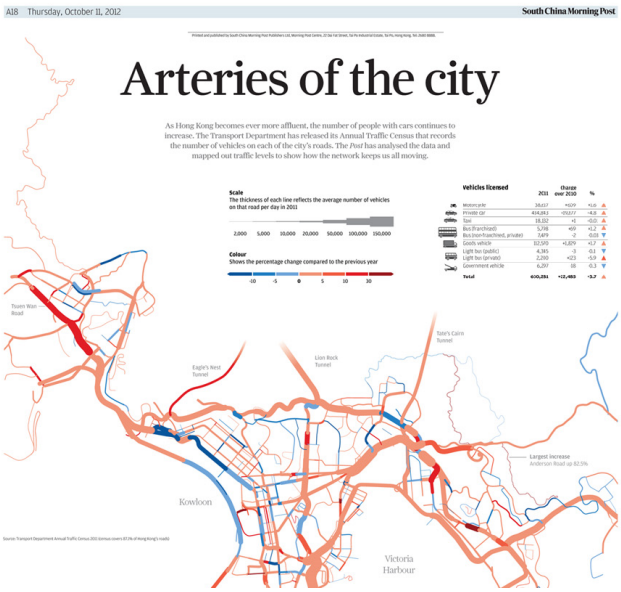

Figure 14: Average number of vehicles on the Hong Kong main network of roads. Source: Andy Kirk, Chapter 6 [1]

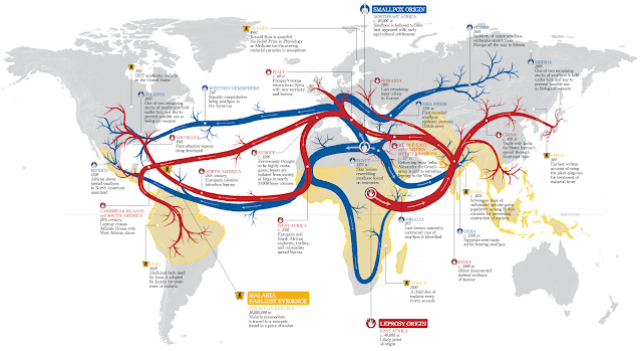

Figure 15: Displays origin and transmission pathways of Malaria, Small Pox, and Leprosy. Source: https://tinyurl.com/ya8nbsd3

2. Geocomputation with R[2]. R Lovelace, J Nowosad, J Muenchow.

References

[1] Andy Kirk. *Data Visualisation: A Handbook for Data Driven Design*. Sage, London, 2019.

[2]https://geocompr.robinlovelace.net/index.html

[2] GIS Website. *http://wiki.gis.com/wiki/index.php/Isarithmic_map*.