EG UK Computer Graphics & Visual Computing (2017)
Tao Ruan Wan and Franck Vidal (Editors)

Cartographic Treemaps for Visualization of Public Healthcare Data

Chao Tong[1], Richard Roberts[1] , Robert S. Laramee[1] , Damon Berridge[2] and Daniel Thayer[2]

[1] Visual and Interactive Computing Group, Swansea University
[2] Medical School, Swansea University

Abstract

The National healthcare Service (NHS) in the UK collects a massive amount of high-dimensional, region-centric data concerning individual healthcare units throughout Great Britain. It is challenging to visually couple the large number of multivariate attributes about each region unit together with the geo-spatial location of the clinical practices for visual exploration, analysis, and comparison. We present a novel multivariate visualization we call a cartographic treemap that attempts to combine the space-filling advantages of treemaps for the display of hierarchical, multivariate data together with the relative geo-spatial location of NHS practices in the form of a modified cartogram. It offers both space filling and geospatial error metrics that provide the user with interactive control over the space-filling versus geographic error trade-off. The result is a visualization that offers users a more space efficient overview of the complex, multivariate healthcare data coupled with the relative geo-spatial location of each practice to enable and facilitate exploration, analysis, and comparison. We evaluate the two metrics and demonstrate the use of our approach on real, large high-dimensional NHS data and derive a number of multivariate observations based on healthcare in the UK as a result. We report the reaction of our software from two domain experts in health science.

1. Introduction

The United Kingdom faces massive challenges with respect to providing the best healthcare via the National Health Service (NHS). In order to provide the best service, Public Health England and the UK government collect years worth of region specific-healthcare data [NHS]. The public health profiles website [NHS] is used for publishing the latest national healthcare data in the UK. The data archive is designed to support GPs, clinical commissioning groups (CCGs), and local authorities to ensure that they provide and commission effective and appropriate healthcare services. However the size and complexity of the data creates challenges for deriving new knowledge and insight.

The NHS data includes a UK map divided into CCGs, which are groups of NHS practices. Each CCG contains the local population and high-dimensional healthcare data collected by the NHS, such as cardiovascular disease (CVD) diagnoses, indicators of respiratory health, mental health, indicators, incidents of chronic obstructive pulmonary disease (COPD), kindey disease, as well as other diagnoses.

Our goal is to develop imagery that combines UK-centric geo-spatial information with high-dimensional NHS data in a unified framework. Moreover, we believe the principles apply equally well to other multivariate data sets of this kind. A hybrid visualization we call a Cartographic Treemap combines the geo-spatial properties of cartograms with the space filling properties of treemaps, inheriting advantages of both. We provide the user interactive control over the trade off between filling the most space, like a treemap, and geo-spatial error. Currently, visualizing multi-dimensional healthcare data based on CCGs is not possible because many CCGs cover the space of only a few pixels. Many CCGs are crowded into the London region, obstructing any geo-spatial visualization without a second magnified view. We propose a cartographic treemap to integrate a modified representation of the UK based on the geo-spatial information of CCG regions combined with a modified treemap to present the multivariate NHS data. The contributions of this paper include:

(1) A new hybrid visualization, the Cartographic Treemap, combining geo-spatial information in the form of a modified cartogram with space-filling geometry for the visualization of high-dimensional data. (2) A layout algorithm for rectangular cartographic treemaps: increasing region size incrementally and avoiding overlapping regions. (3) A novel, interactive error metric and user options that trade-off screen space versus geo-spatial accuracy to facilitate user analysis. (4) The novel application of our hybrid visualization to complex, real-world NHS data from the UK. The paper by Tong et al. [TML*17] extends this work by adding time as a variate.

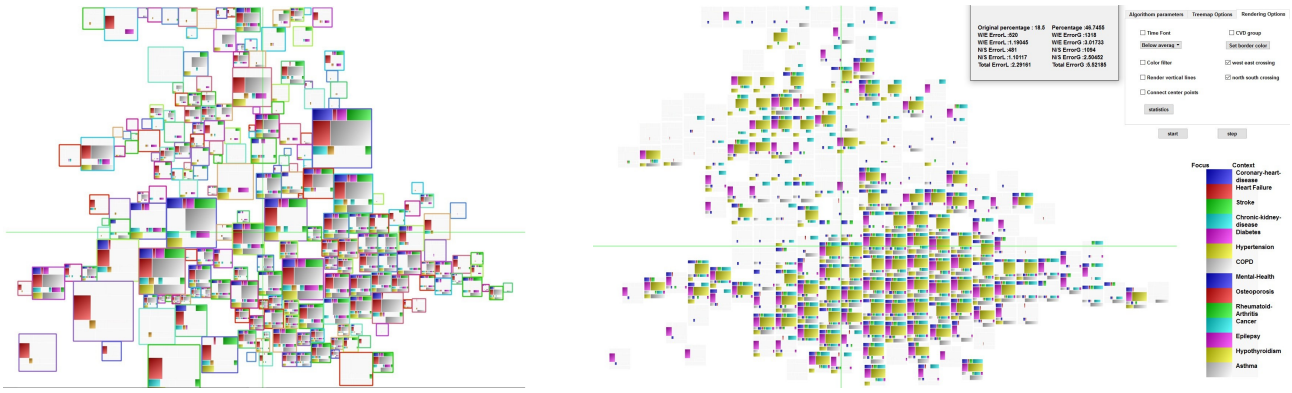

Figure 1: *This graph shows each region size proportional to its population with an added below average filter (left). The percentage of screen space occupied, s_0 = 41% and the local error, e_l = 3.5%, e_g=8.7% and uniform size output with a below average filter (right). s = 47%, e_l = 2.3%, and e_g = 5.5%. All the healthcare disorders that exhibit higher than average prevalence are filtered and shown as grey context. Note how the London region is healthier with the exceptions of diabetes and mental health. This is an observation based on multiple variates that would be difficult to make otherwise. See Figure 12 for high-resolution version.*

2. Related Work

Some very helpful survey papers provide an overview of healthcare research [KM02, RWA*13, WBH15, ML17]. However we would like to couple geo-spatial information with the healthcare data.

Geo-spatial related work falls into the areas of cartograms and spatially-ordered treemaps. we separate and review those two categories of previous paper here.

Cartographic visualization Cruz et al. [CCM15] define a cartogram as "*a technique for displaying geographic information by resizing a map's regions according to a statistical parameter in a way that still preserves the map's recognizability*". They can display geo-spatial information and another data attribute (such as population or disease prevalence) in one visualization. Tobler [Tob04] and Nursat and Kobourov [NK16] survey general cartograms. They present the development of value-by-area cartogram algorithms and performance in computer science.

Auber et al. [AHL*11] propose a layout method based on a geographic map metaphor, which facilitates the visualization and navigation of a hierarchy and preserves the order of the hierarchy's nodes.

Gastner and Newman [GN04] present a diffusion cartogram for constructing value-by-area cartograms, which provides a valuable tool for the presentation and analysis of geographic data. Keim et al. [KNP04] develop a faster algorithm for cartograms. It enables display dynamic data with cartogram visualizations. These two algorithms are categorised as contiguous area cartograms. Their performance depends on the corresponding value in each area. If the value does not correspond to the area, the cartogram may be difficult to recognize.

Raisz [Rai34] presents the rectangular cartogram, using rectangles instead of real area shapes. Dorling [Dor11] presents the Dorling cartogram which uses circles instead of geographic area shape, similar to the modified cartogram we present. They are categorized as non-continuous area cartograms. They can display statistical information well, regardless of original shape of area, and preserve relative position. Van Kreveld and Speckmann [vKS07] present the first algorithm for rectangular cartograms. They formalize region adjacencies in order to generate processable layouts that represent the positions of the geographic regions. It converts a rectangular cartogram to a contiguous area cartogram. Our modified cartogram does not fall into the category of continuous cartograms but resembles a cross between rectangular and Dorling cartograms [NK16]. Our algorithm can be considered as a modified space-filling rectangular cartogram with the addition of a hierarchical structure and multivariate data.

Heilman et al. [HKPS04] propose a novel visualization technique for geo-spatial datasets that approximates a rectangular partition of the rectangular display area into a number of map regions preserving important geo-spatial constraints. They use elongated rectangles to fill the space whereas we use uniform rectangles to fill the space such that regions can easily be compared with one another. Their work focuses on univariate, non-hierarchical data.

Panse et al. [PSKN06] combine a cartogram-based layout (global shape) with PixelMaps (local placement), obtaining benefits of both for improved exploration of dense geo-spatial data sets. Their work also focuses on univariate, non-hierarchical data.

Slingsby et al. [SDW09] explore the effects of selecting alternative layouts in hierarchical displays that demonstrate multiple aspects of large multivariate data sets, including spatial and temporal characteristics. They demonstrate how layouts can be related, through animated transitions, to reduce the cognitive load associated with their reconfiguration whilst supporting the exploratory process. No metric for neighborhood preservation is described in this work.

Slingsby et al. [SDW10] present rectangular hierarchical cartograms for mapping socio-economic data. They present a detailed map of 1.52 million UK unit postcodes in their spatial hierarchy, sized by population and coloured by the OAC category that most closely characterises the population. However, no algorithm

for preserving geo-spatial information is provided. No metric for neighborhood preservation is described.

Alam et al. [AKV*15] present a set of seven quantitative measures (Average Cartographic Error, Maximum Cartographic Error, Adjacency Error, Angular Orientation Error, Hamming Distance, Average Aspect Ratio, Polygonal Complexity) to evaluate performance of cartograms based on the accuracy of data and its readability. They compare previous cartogram algorithms based on statistical distortion, geography distortion and algorithm complexity and evaluate their performance with respect to different properties. Nursat and Kobourov [NK16] survey cartogram research in the field of visualization and present design guidelines as well as research challenges. They state that mapping multivariate data is still a challenge in cartogram research. In general, previous cartographic visualizations focus on flat, univariate data, whereas we process hierarchical, multivariate data.

Eppstein et al. [EvKSS15] introduce a new approach to solve the association challenge for grid maps by formulating it as a point set matching problem. They present algorithms to compute such matchings and perform an experimental comparison that also includes a previous method to compute a grid map. Their work focuses on geo-spatial information and filling space. multivariate, hierarchical data is not considered.

Meulemans et al. [MDS*17] design a comprehensive suite of metrics that capture properties of the layout used to arrange the small multiples for comparison (e.g. compactness and alignment) and the preservation of the original data (e.g. distance, topology and shape). Their work focuses on geo-spatial information and neighborhood preservation. Multivariate, hierarchical data is not considered.

We note that the visualizing multivariate data is one of the top future research challenges in the latest survey by Nursat and Kobourov [NK16]. Also cartograms, in general, are not space-filling and do not necessarily make the best use of screen space.

Geo-Spatial Treemaps Mansmann et al. [MKN*07] present HistoMaps for visual analysis of computer network traffic visualization with a case study showing that a geographic treemap can be used to gain more insight into these large data sets. However the visualization is essentially univariate (one scalar per level in the hierarchy). It is also not adjacency preserving.

Wood and Dykes [WD08] provide a squarified layout algorithm that exploits the two-dimensional arrangement of treemap nodes more effectively. It is suitable for the arrangement of data with a geographic component and can be used to create tessellated cartograms for geo-visualization. They convert a geographic distribution of French provinces to a spatial treemap layout and preserve the corresponding geo-spatial relationships to some extent. However, they demonstrate that it is impossible to preserve local region adjacencies if nodes are constrained to a standard rectangle parent node. For example, a region map may only have one or two neighbors on a geographic map. We preserve geo-spatial relationships with less error by allowing gaps in screen space at the different levels of the data hierarchy.

Jern et al. [JRA09] demonstrate and reflect upon the potential synergy between information and geo-visualization. They perform this through the use of a squarified treemap dynamically linked to a choropleth map to facilitate visualization of complex hierarchical social science data. It conveys the neighborhood relationships by using a second view.

	Geo-spatial information	Neighborhood Preservation	Multi variate	Hierarchical	Space-filling
Cartograms					
Raisz, 1934					
Dorling, 1996					
Auber et al.					
Tobler, 2004					
Gastner et al., 2004					
Keim et al. , 2004					
Heilman et al. , 2004					
Panse et al.,2006					
Van et al. , 2007					
Slingsby et al., 2009					
Slingsby et al., 2010					
Alam et al. , 2015					
Eppstein et al., 2015					
Meulemans et al., 2016					
Treemaps					
Shneiderman and Johnson,1992					
Bruls et al. , 2000					
Shneiderman , 2001					
Itoh et al. , 2004					
Balzer et al, 2005					
Irnip and Shen, 2006					
Tu and Shen, 2007					
Mansmann et al. , 2007					
Wood and Dykes , 2008					
Jern et al. , 2009					
Slingsby et al., 2010	AP				
Buchin et al. , 2011	AP				
Wood et al., 2011					
Wood et al., 2011					
Duarte et al. , 2014					
Ghoniem et al. , 2015					

Figure 2: *This table shows characteristics of related work. It includes five visualization properties: geo-spatial information, neighborhood preservation, multivariate, hierarchical and space-filling. Geo-spatial information implicates whether a visualization conveys geographic information and AP in the column represents adjacency preservation only. Neighborhood preservation indicates an algorithm that features a distance metric to preserve neighborhood relationships. multivariate indicates the dimensionality of abstract data. Hierarchical indicates a type of hierarchical data and space-filling indicates how well the output visualization fills the screen. Cartographic treemaps feature all five properties.*

Slingsby et al. [SDWR10] develop an OAC (Output Area Classifier) explorer that can interactively explore and evaluate census variables. There is no inherent information preserving the geo-spatial location of regions because a synthetic grid is used to subdivide space. It is not possible to derive any information about the geography of the UK regions.

Buchin et al. [BEL*11] describe algorithms for transforming a rectangular layout without hierarchical structure, together with a clustering of the rectangles, into a spatial treemap that respects the clustering and also respects to the extent possible the adjacencies of the input layout. The work of Buchin et al. is similar to ours

with few differences. First, they do not demonstrate their layout algorithm on a full geo-spatial map, e.g. the UK. Second, the space-filling requirement results in elongated rectangles that are difficult to compare. Third, the data is univariate.

Wood et al. [WBDS11] present Ballotmaps that using hierarchical spatially arranged graphics to represent two locations (geographical areas and spatial location of their names on the ballot paper) that affect candidates at very different scales but their work does not contain any neighborhood preservation algorithm.

Wood et al. [WSD11] identify changes in travel behavior over space and time, aid station rebalancing and provide a framework for incorporating travel modeling and simulation by using flow maps. Their work focuses on univariate, non-hierarchical data.

Duarte et al. [DSF*14] propose a novel approach, called a Neighborhood Treemap (Nmap), that employs a slice-and-scale strategy where visual space is successively bisected in the horizontal or vertical directions. The bisections are scaled until one rectangle is defined per data element. Nmap achieves good space-filling visualization that couples related rectangles using a distance metric. However, the distance metric is not geo-spatial, it is also not a treemap of multivariate data nor a hierarchical visualization.

Ghoniem et al. [GCB*15] present a weighted maps algorithm, which is a novel spatially dependent treemap. They present a quantitative evaluation of results and analyze of a number of metrics that are used to assess the quality of the resulting layouts. The work of Ghoniem et al. is similar to ours with some important differences. They place emphasis on evaluating adjacency relationships between nodes rather than geo-spatial positions. Requiring 100% space-filling results in higher geo-spatial error and elongated nodes. Also the data is not multivariate.

Treemaps: Geo-spatial information versus adjacency preservation: In general, the treemap layout algorithms attempt to reflect geo-spatial information implicitly through adjacency relationships between the nodes. As shown by Ghoniem et al. [GCB*15], this leads to high geo-spatial error, e.g. in the 40%-50%. It also leads to elongated rectangles which may be difficult to compare. It may be difficult to recognize the correspondence to the original geo-spatial map when looking at a treemap. In contrast, our algorithm emphasizes geo-spatial preservation with less emphasis on adjacency relationships. We give the user new interactive control over the amount of error and allow spaces and gaps to reduce geo-spatial error.

The work we present here differs from previous work in that it attempts to combine the space-filling, hierarchical characteristics of ordered space-filling treemaps together with the geo-spatial information conveyed by a cartogram. Table 2 compares the current work with the work presented here. No previous algorithm combines all five properties. Cartographic Treemaps convey geo-spatial information. They feature an error-driven distance metric between nodes and visualize multivariate hierarchical data. They also give the user interactive control over how much screen space is used.

3. NHS Data Description

The NHS data includes a UK map divided into CCGs, groups of NHS practices. A standard map of the UK only covers about 18 % of screen space due to its awkward shape. Each CCG contains various categories of disease in prevalence value. Prevalence is the proportion of a population who have a specific medical diagnosis in a given time period, typically an illness, a condition, or a risk factor such as depression or smoking. Prevalence is a derived metric of the local population of each region. Prevalence is usually expressed as a percentage.

Typically this data is displayed using line charts, bar charts, and pie charts. The map provided by public health England is a standard UK map with 209 CCG regions. The boundaries of CCG regions vary and are difficult for presenting high-dimensional data. The CCGs coupled directly to the geography do not make efficient use of space. The UK map itself only occupies 18% of screen space. For visualization purposes the CCG regions in London for example, crowd together and hamper our ability to visualize multi-dimensional data clearly. This will be true in the capital region of most countries and other densely populated areas. Other healthcare data, for example, the population distribution data is typically visualized using a single line chart showing the percentage of age groups distributed from 0-4 to 85+. Standard graphs show no connection with other health data attributes such as geo-spatial location and clinical diagnoses. This challenging data set is the inspiration behind cartographic treemaps. See the supplementary PDF for a description of the health disorders.

4. Cartographic Treemaps

This section describes the cartographic treemaps construction algorithm and interactive error control, starting with an overview. The processing begins with reading the UK geo-spatial information and high-dimensional healthcare data. The algorithm is as follows:

(1) Compute region center points: We use the QGIS [QGI] tool to calculate the center points of each CCG region. The center points are the starting positions of the rectangular region nodes. (2) Update node size: We start with a unit square to represent each CCG region as a node in the cartographic treemap and increase the size of each node according to the user's chosen space-filling target or error constraint. (3) Update cartographic layout: During the region growing process, one region may shift adjacent neighboring regions to remove overlap and preserve relative position. When all regions reach their maximum size or the user-specified geo-spatial error is reached, the cartogram layout stops. We use the fast overlap removal algorithm [DMS06, DMS07] incrementally for this process.

(4) Treemap node layout: After the cartographic node layout is completed, an ordered squarified treemap layout is used to present the multivariate healthcare data in each CCG region, the lowest (finest) level in the treemap hierarchy. (5) Interactive user options: For further exploration, analysis and region comparison, several user options are designed to present the results focusing on different user requirements, such as modifying algorithm parameters, region selection for detail, modifying the color legend, and exploring the hierarchy.

4.1. Updating Node Size

After calculating the center point of each CCG node, we initialize CCG nodes as unit squares on the cartographic treemap. The al-

gorithm increases the size of each node to make the most efficient use of space. It terminates when the user-specified geo-spatial error or a target screen space percentage is reached. The algorithm can also increase the size of each node based on any property of the region (or proportion to a fixed maximum size region), e.g. the local population of the CCG like a traditional cartogram. Because we gradually increase the size of each CCG region node, the relative geo-spatial position between nodes is preserved. After the area of each square is increased by a small amount (1 pixel by default) some adjacent nodes may overlap. We then update the position of each node in the tree by running the fast node overlap removal algorithm [DMS06, DMS07] described in the next section. We provide an animation to present the incremental processing from 1 pixel to maximum size. Slingsby et al. demonstrate the benefit of animation in this context [SDW09].

4.2. Updating Region Node Position

We use the fast node overlap removal algorithm presented by Dwyer et al. [DMS06, DMS07] for removing overlap between neighboring region nodes. With this algorithm, the overlap is reduced in the quickest, most effective way. That means if a node, n, overlaps with its northern neighbor, n_n, running this algorithm shifts n south or its neighbor n_n north, the most effective way to remove overlap. By constraining the overlap to a small area, the relative position of adjacent nodes is preserved. If we increase all nodes to their maximum size before running the overlap removal algorithm, relative geo-spatial position of region nodes is not preserved as well. The reason for this is when a node (n) is much smaller than its neighbor (n_n), it may lie completely inside its neighbor after its size has expanded to its maximum. In this case, it is faster to reduce overlap without preserving relative position.

The fast node overlap removal algorithm has two phases. In the first phase a number of constraints are applied that derive the separation distance between nodes. In the second phase, the solution is searched based on location as close as possible to the original node positions [DMS06]. To address relative geo-spatial position preservation, we run the fast node overlap removal algorithm incrementally. In each pass, we increase the size of nodes by 1 unit and run the fast node overlap removal algorithm. In this way, the algorithm removes overlap and preserves relative position. The process is repeated until all nodes have reached their maximum size or a user specified error threshold is reached. (Some examples are shown in supplementary file.) We can also animate the region growing process in order to increase the legibility of the visualization. Please see the accompanying video for a demonstration. Observing the evolution of each region provides benefit [SDW09].

4.3. A Neighborhood Preservation Error Metric

We introduce a novel neighborhood preservation error metric that objectively quantifies how closely the relative geo-spatial positions of the resulting nodes correspond to their original positions. In other words, a west neighbor n_a should remain west of a given node after the layout is updated. Likewise for the east, north, and south directions. We consider an error when the relative geo-spatial position of the region center points cross. We use global error, e_g, to record any two center points crossing while we use local error, e_l, to record center points crossing when the distance between two center region points is less than a user specific threshold in Euclidean space, e.g. 20% of screen space.

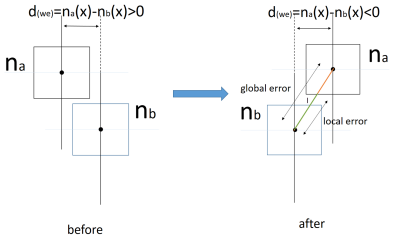

Figure 3: *The illustration of global and local error for neighborhood preservation. The error distance is decoupled into x(west-east) and y(north-south) components. The x components is illustrated here.*

As shown in Figure 3, we focus on the relative position of the center points of regions n_a and n_b. After looping through the layout algorithm, an error is counted if the longitudinal line of n_b crosses the longitudinal (along y) line of n_a. i.e. the longitudinal distance $d_{(we)}$=$n_a(x)$ -$n_b(x) > 0$ initially and $d_{(we)}$=$n_a(x)$ -$n_b(x) < 0$ after updating the node positions. That means the relative longitudinal positions of n_a and n_b are not preserved, thus we count this case as one error, similarly for the north-south orientation/position. If the total distance between the centers of n_a and n_b is less than a user specific distance, we consider this error as local error, e_l. We consider the worst-case scenario or maximum geo-spatial error when the whole map is flipped both latitudinally and longitudinally, similar to the worst case of bubble sort $O(n^2)$. Figure 4 shows an actual depiction of this error.

We consider the worst-case scenario when the center of every region node n crosses every other region node, $n-1$. We adopt the result that $n+(n-1)+(n-2)+...+1 = n(n+1)/2$. In our case n is 209, however node n cannot cross itself. Thus we use $n(n-1)/2$ as our worst case result. The worst-case number of crossings in our application is 21736. And all error can be expressed as a percentage of this total.

We do not claim that this is the best distance metric in all of the literature. Ghoniem et al. [GCB*15] and Nusrat and Kobourov [NK16] provide a comprehensive review and comparison of distance and error metrics for cartograms and spatial treemaps. In fact many of those could be substituted here. Our contribution is that this error metric is interactive as the user controls the level of error. For the first time the user controls the trade-off between filled screen space and relative error of geo-spatial position.

4.4. Ordered Treemap Algorithm

After the size and position of each CCG region node is computed, a treemap node layout algorithm is used to visualize the non-spatial, multivariate health indicator data within each CCG. We require this data is layed out consistently for each CCG region node to facilitate comparison between CCGs. Ordered treemap algorithms create rectangles in a visual order that match the input order of the data.

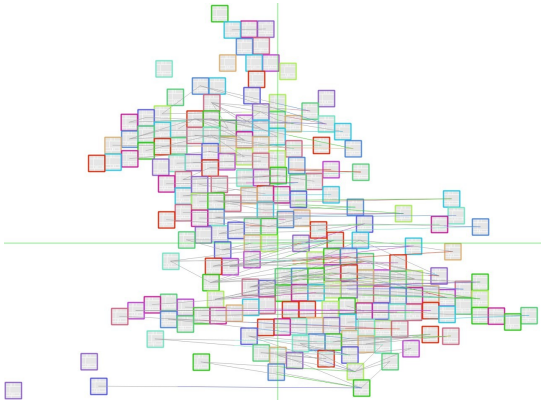

Figure 4: *Visualization of errors: This figure shows error crossing edges in north and south orientation. The screen space-filling percentage, s, is 20% and e_l is 0.9%, and e_g is 1.8%.*

Bederson et al. [BSW02] present two algorithms to display ordered treemaps: A Pivot treemap and the Strip treemap algorithm. Compared to the Pivot treemap algorithm, the Strip treemap results in a lower rectangular aspect ratio. This version is more squarified with a higher readability score. So we choose the Strip treemap algorithm to present data inside each individual CCG node.

4.5. Interactive User Options

For further exploration and analysis, several user options are available to explore and present the results focusing on different requirements such as filling the maximum space, specifying the local or global error, animating the node layout algorithm, modifying layout parameters, region selection for detail, modifying the color legend, and exploring the hierarchy.

Geo-spatial Error and CCG Region Node Size As our goal is to combine the geo-spatial properties of cartograms with the space filling properties of treemaps, the first user controlled parameter setting is the maximum geo-spatial error of the CCG regions. All CCG region sizes are uniform by default in order to facilitate comparison between regions. However, their size can also be proportional to the maximum sized region. The size of each CCG region can be mapped to the size of its local population or any health data indicator like a traditional cartogram. So we enable the user to set the maximum size of the region with the largest population and the other regions are adjusted relative to the maximum. As in Figure 5.

Node Size Increment and Animation In the cartographic treemap layout algorithm, the region size grows incrementally. As discussed in section 4.2, immediately increasing the node size to its maximum does not preserve the geo-spatial relationship between regions as well, while iterative increments take more time to generate the final result. So we provide a user option to explore an ideal size of area increase in a single layout algorithm pass. The increment size is set between 1 and 10 pixels. The layout takes more time when the increment size is small, but the accuracy of geo-spatial neighborhood relationships is increased. There is a trade-off between processing speed and accuracy of the geo-spatial relationship between nodes. A user option of animating the region node layout process is provided so the user can observe the correspondence between the original node position and the final visualization. Slingsby et al. [SDW09] demonstrate the value of animation. The multi-pass layout algorithm is shown gradually from initial to final layout.

Uniform Size Regions A cartographic treemap node for a single region represents the prevalence of various health disorders. As the size of each CCG region may be uniform or represent its population, the size of bottom level rectangles represents the proportion of the population with a particular health disorder in the respective region. We can get an overview of the prevalence of various diseases in CCG regions. As in Figure 6. However, as the population sometimes varies greatly among CCG regions, the size of bottom level rectangles may not be directly compared with other CCG region nodes. For example, a large population of heart failure in Oxfordshire CCG may not indicate heart failure there is relatively prevalent. The prevalence of heart failure in Oxfordshire is 0.51 which is lower than the average of 0.73. In order to facilitate direct comparison of health disorders across CCG region nodes, we provide a user option to generate uniform size region-level nodes set to true by default. In this way, the size of rectangles at the bottom level of the treemap hierarchy can be compared directly. As in Figure 7.

Difference Cartographic Treemap and Focus+Context To make the healthcare visualization clearer, we introduce a user option: a difference cartographic treemap. The size of each rectangle at the bottom level of the healthcare treemap does not represent the absolute prevalence value of each health disorder. Instead, it represents the difference from the average UK value. Using this option, we can emphasize how the prevalence of a specific health condition differs from the national average level and understand the conditions in a particular region. As in Figure 8. We also use a focus+context visualization incorporating a focus+context color map. The user may choose to focus on above average or below average values. Focus attributes are then rendered in color while context rectangles are rendered in grey-scale. As in Figures 10 and 1.

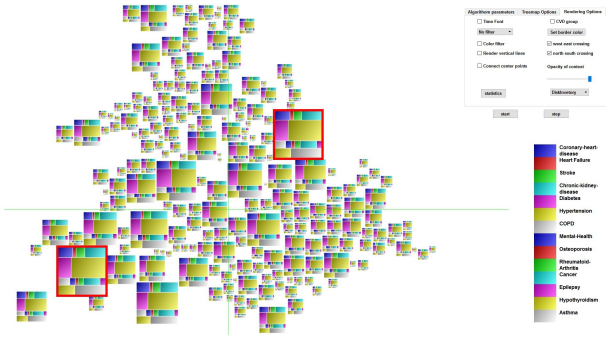

Figure 5: *Nodes proportional to CCG size. The screen space-filling percentage, s=36% and e_l=2.4%, e_g = 4.5. The two red outlines show the two biggest region nodes on the map: Cambridgeshire Peterborough and North East & West Devon. This is unexpected since we hypothesized the largest regions to be in London or Birmingham. This example uses color map from the Disk Inventory X tool [dis]. See Figure 13 for high-resolution version.*

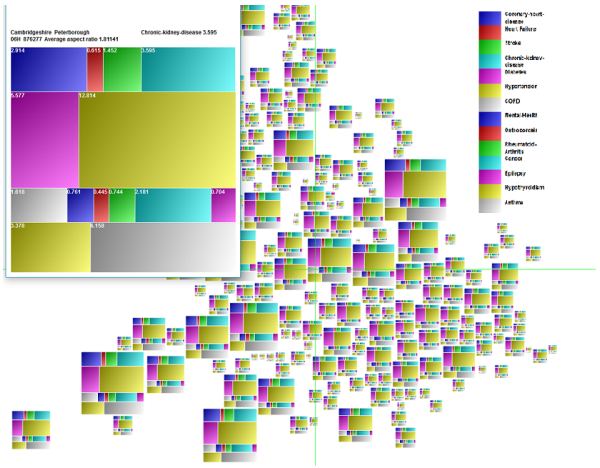

Figure 6: *This visualization shows the output of cartographic treemap with region size proportional to population, and with a details-on-demand window for one region node. s=30%, e_l=2.4% and e_g= 5.1%. The first three rectangles in each region node represent three CVD health disorders. Note the prevalence of hypertension and diabetes is very widespread the UK. This type of multivariate observation display itself clearly with this type of visualization. See Figure 14 for high-resolution version.*

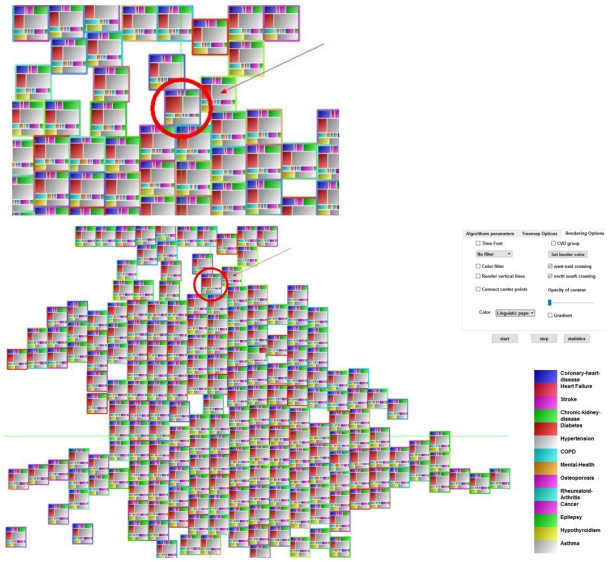

Figure 7: *This graph shows the output of cartographic treemap with uniform size region nodes. s=50% and e_l=2.4%, and e_g=5.8%. The region with the red circle (Bradford City) contains the largest purple rectangle which indicates the highest relative prevalence of diabetes in the UK. This example uses a published color-map from Setlur and Stone [SS16].*

Area Groups We introduce area groups to classify CCG regions into 27 area groups in the treemap hierarchy based on area code. This option creates a more space-filling cartogram and another hierarchy level in the treemap. It facilitates comparison of CCG regions healthcare data within their own CCG groups and enables exploration and analysis. As in Figure 9. It also results in a more space-filling layout with greater resemblance to a traditional treemap.

Details-on-Demand and on-mouse-over: For the finest (lowest) level of data detail in CCG regions or area groups, a details-on-demand feature is implemented. By hovering the mouse over or clicking on any region, a new window opens with a higher resolution treemap, providing the CCG code, CCG name and value of each health diagnosis category. As in Figure 6. To improve the appearance, we also add user options for various color maps and color gradient styles (See Figure in supplementary file). The color maps come from different sources; one is from the disk inventory X tool [dis], the second one is from ColorBrewer [Col], the third one is from Telea [Tel14], the fourth is from QGIS [QGI], and the last one is from Setlur and Stone's paper [SS16]. As in Figure 7.

5. Results and Discussion

In this section, we present the results of our interactive metrics and derive a number of observations based on cartographic treemaps.

Accompanying Demonstration Video URL

https://vimeo.com/199637583

The images here are lower resolution due to space limitations. The video and supplementary PDF contain higher resolution imagery.

Evaluation of Space and Error Metrics To evaluate the performance of our algorithm, we measure the percentage of filled screen-space, s, versus the local and global geo-spatial error. As the original map is narrow, the space filled with respect to the screen is 18.5% and by using our algorithm the percentage of filled screen can reach up to 70%. The relationship between error and screen space filled is shown in Figure 11.

Based on the algorithm described in section 4.3, the local and global error is shown in Table 1 and Figure 3. It shows the connection between e_l, e_g and s. It presents percent space filled along with local and global percentage and frequency of center-axis crossings.

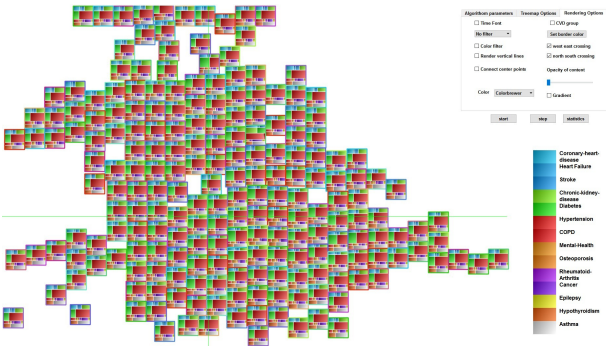

Figure 8: *This graph shows the cartographic treemap using average difference maps. s=50%, e_l=2.4%, and e_g=5.8%. The larger a bottom level rectangle is, the more it deviates from the UK average. This example uses a well-known color map from color-brewer [Col].*

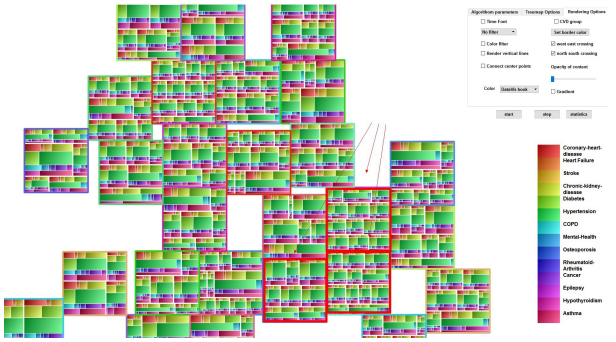

Figure 9: *This graph shows the cartographic treemap with 27 area groups. s= 70% and e_g = 5.2% . The regions in red highlights are London areas. This example uses Telea's color map [Tel14].*

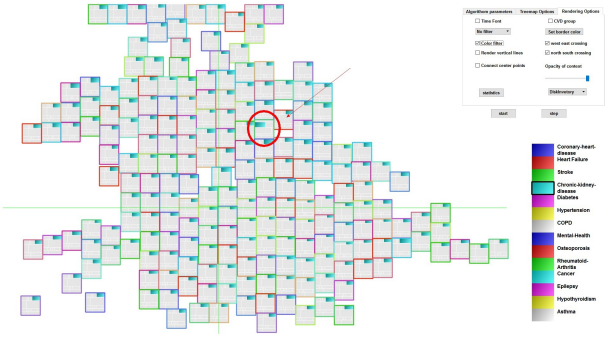

Figure 10: *A focus+context cartographic treemap visualization with uniform size regions. s=50%, e_l=2.4%, and e_g=5.8%. The data is mapped to two color scales: one for the focus data and the other for context. All the healthcare prevalence categories are shown as context except for user selected data attributes. The red circle shows the relatively largest rectangle in the map that represents the highest prevalence of Chronic-kidney-disease disorder in the UK (Nottingham North And East).*

We can see that e_l increases linearly with s occupied while e_g increases more rapidly. We can achieve 65% screen-space occupancy with only 1-4% error.

Performance and Observation The algorithm requires less than a second to run (85ms-1000ms). The computer used to run this algorithm is an MSI desktop with Intel 3.4GHz CPU, 8GB RAM, GeForce GTX 770 graphic card and Windows 10. We slow it down for purposes of animation and user observation.

Table 1: *Neighborhood Preservation Metric*

s	e_l	local error frequency	e_g	global error frequency
10%	0.4	164	0.7	293
20%	0.7	308	1.5	667
35%	0.9	409	2.5	1073
57%	1.1	476	3.1	1369
66%	1.2	524	3.6	1593

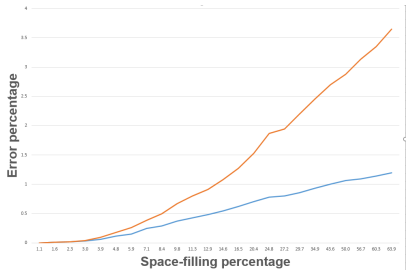

Figure 11: *This figure shows the relationship between percentage of both local and global error versus the amount of filled space. The red line shows the global error while the blue line indicates the local error.*

Based on the cartographic treemap visualization, several observations can be derived from the public health data. Several of these observations would be very difficult without the cartographic treemap.

(1) From the region node layout in Figure 9, we can see that the London area contains the most CCG group regions (32 in total) and the largest population. (2) The individual CCG regions with largest population are Cambridgeshire Peterborough and North East & West Devon. As in Figure 5. This is not what we would expect but rather the largest populations in a London CCG. (3) Hypertension is most prevalent health disorder with the largest proportion throughout the UK while the second largest health disorder is diabetes. As in Figure 6. This is clear from an overview cartographic treemap.

(4) Three kinds of CVD related disorders (Coronary-heart-disease, Heart Failure, Stroke) are prevalent throughout the UK, and coronary heart disease is the most common disorder in the CVD disorder group (a multivariate observation). As in Figure 6. (5) From the uniform size nodes, the regions with a significantly higher prevalence of health disorder can easily be observed. Bradford City has the relatively highest diabetes in the UK. As in Figure 7. We can also find the highest relative chronic kidney disease disorder prevalence in the Nottingham North & East CCG. As in Figure 10 and the highest relative mental health disorder prevalence is found in Islington. (6) Compared to the average value across all health disorders, regions in London are generally better than the average in most health categories with the exceptions of mental health and diabetes. As in Figure 1. This is another multivariate observation found using the cartographic treemaps capabilities.

(7) The North regions are higher than average in most health disorders, such as, Cumbria and Northumberland. The values are higher than the average for a range of health disorders. For example, diabetes is more prevalent in Northern regions than Southern regions. This is shown in Figure 1. Cartographic Treemaps facilitate these kind of multivariate observations.

6. Health Science Domain Expert Feedback

Domain expert 1: "Data analysts are often required to analyse complex sets of spatial, multivariate, longitudinal, and event history public health data in order to answer research questions as part

of major studies such as CORTEX, ELAStiC and the Carmarthenshire Housing Project. Cartographic treemaps facilitate the recognition of patterns within the data such as geographical clustering and temporal trends, as well as the identification of salient features including outliers and extreme values, thereby helping to complement machine learning and data mining techniques and to inform statistical modelling. This visualization will make a major contribution towards helping data analysts to achieve their research objectives. Therefore, we are delighted that this new technique will be utilised by data analysts in the Farr Institute @ CIPHER within Swansea University Medical School. We are confident that the cartographic treemaps will provide data analysts with the opportunity to gain additional deeper insights into their complex public healthcare data."

Domain expert 2: "Some of the biggest challenges of working with linked population health datasets relate to the sheer volume of the data: the scale is daunting in terms of the population sizes, and dimensionality. There are thousands of potentially interesting facts stored in various data sources. The depth and breadth of the data make it hard to see the big picture of what is going on in a population, as well as to sort through the noise to identify what information is relevant. These challenges are multiplied if the data is to be used directly in a clinical setting by people who are not expert analysts. Something that is necessary to derive maximum benefit from available data resources. Visualization is a key technology to help users, both academic and clinical, make sense of the data. The cartographic treemap approach described here addresses our challenges by allowing a number of related variables to be presented simultaneously. Geography is often an important dimension in health research and service planning, and this technique allows data to be organized geospatially while transcending some of the limitations of traditional map-based visualizations. The ability to see geography, population sizes, and several health measures at the same time will help users get a much more accurate, at-a-glance understanding of the data and the population it represents. It has potential to aid research, particular in the hypothesis-generation phase; and it could be quite beneficial in the healthcare sector, supporting activities such as service planning."

7. Conclusion

This paper presents a novel hybrid visualization, the Cartographic Treemap, combining geo-spatial information, a novel interactive neighborhood preservation metric, and space-efficient geometry for the interactive visualization of geo-spatial, and high-dimensional data. It combines the advantages of both cartograms and treemaps. We implement and demonstrate this visualization with a real-world high-dimensional healthcare data collected by NHS to support clinical commissioning groups (CCGs) and the healthcare service providers. Several interactive user options are available to explore and present the results focusing on different user requirements for further exploration, analysis and comparison. Also, we present several multivariate observations based on the cartographic treemap visualization and report feedback from two domain experts in health science. Future work includes investigating more optional color maps for high-dimensional data and a more in-depth user feedback study.

8. Acknowledgments

Thanks to Liam McNabb, Dylan Rees, Dave Greten and Sean Walton for proofreading this paper.

References

[AHL*11] AUBER D., HUET C., LAMBERT A., SALLABERRY A., SAULNIER A., RENOUST B.: Geographical Treemaps. *Technical Report* (2011). 2

[AKV*15] ALAM M., KOBOUROV S. G., VEERAMONI S., ET AL.: Quantitative Measures for Cartogram Generation Techniques. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 351–360. 3

[BEL*11] BUCHIN K., EPPSTEIN D., LÖFFLER M., NÖLLENBURG M., SILVEIRA R. I.: Adjacency-Preserving Spatial Treemaps. In *Algorithms and Data Structures*. Springer, 2011, pp. 159–170. 3

[BSW02] BEDERSON B. B., SHNEIDERMAN B., WATTENBERG M.: Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. *ACM Transactions on Graphics (TOG) 21*, 4 (2002), 833–854. 6

[CCM15] CRUZ P., CRUZ A., MACHADO P.: Contiguous Animated Edge-Based Cartograms for Traffic Visualization. *IEEE Computer Graphics and Applications 35*, 5 (2015), 76–83. 2

[Col] ColorBrewer. http://colorbrewer2.org/. 7

[dis] Disk Inventory X. http://www.derlien.com/. 6, 7, 13

[DMS06] DWYER T., MARRIOTT K., STUCKEY P. J.: Fast Node Overlap Removal. In *Graph Drawing* (2006), Springer, pp. 153–164. 4, 5

[DMS07] DWYER T., MARRIOTT K., STUCKEY P. J.: Fast Node Overlap Removal Correction. In *Graph Drawing* (2007), Springer, pp. 446–447. 4, 5

[Dor11] DORLING D.: Area Cartograms: Their Use and Creation. *The Map Reader: Theories of Mapping Practice and Cartographic Representation* (2011), 252–260. 2

[DSF*14] DUARTE F. S., SIKANSI F., FATORE F. M., FADEL S. G., PAULOVICH F. V.: Nmap: A Novel Neighborhood Preservation Space-filling Algorithm. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2063–2071. 4

[EvKSS15] EPPSTEIN D., VAN KREVELD M., SPECKMANN B., STAALS F.: Improved Grid Map Layout by Point Set Matching. *International Journal of Computational Geometry & Applications 25*, 02 (2015), 101–122. 3

[GCB*15] GHONIEM M., CORNIL M., BROEKSEMA B., STEFAS M., OTJACQUES B.: Weighted Maps: Treemap Visualization of Geolocated Quantitative Data. In *IS&T/SPIE Electronic Imaging* (2015), International Society for Optics and Photonics, pp. 93970G–93970G. 4, 5

[GN04] GASTNER M. T., NEWMAN M. E.: Diffusion-based Method for Producing Density-equalizing Maps. *Proceedings of the National Academy of Sciences of the United States of America 101*, 20 (2004), 7499–7504. 2

[HKPS04] HEILMANN R., KEIM D. A., PANSE C., SIPS M.: Recmap: Rectangular Map Approximations. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004.* (2004), IEEE, pp. 33–40. 2

[JRA09] JERN M., ROGSTADIUS J., ASTROM T.: Treemaps and Choropleth Maps Applied to Regional Hierarchical Statistical Data. In *IEEE Information Visualisation 2009 Conference* (2009), IEEE, pp. 403–410. 3

[KM02] KOSARA R., MIKSCH S.: Visualization Methods for Data Analysis and Planning in Medical Applications. *International Journal of Medical Informatics 68*, 1 (2002), 141–153. 2

[KNP04] KEIM D. A., NORTH S. C., PANSE C.: Cartodraw: A Fast Algorithm for Generating Contiguous Cartograms. *IEEE Transactions on Visualization and Computer Graphics 10*, 1 (2004), 95–110. 2

[MDS*17] MEULEMANS W., DYKES J., SLINGSBY A., TURKAY C., WOOD J.: Small Multiples with Gaps. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 381–390. 3

[MKN*07] MANSMANN F., KEIM D. A., NORTH S. C., REXROAD B., SHELEHEDA D.: Visual analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1105–1112. 3

[ML17] MCNABB L., LARAMEE R. S.: Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization. *Computer Graphics Forum 36*, 3 (2017). 2

[NHS] NHS:. http://fingertips.phe.org.uk/profile/general-practice. 1

[NK16] NUSRAT S., KOBOUROV S.: The State of the Art in Cartograms. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 619–642. 2, 3, 5

[PSKN06] PANSE C., SIPS M., KEIM D., NORTH S.: Visualization of Geo-spatial Point Sets via Global Shape Transformation and Local Pixel Placement. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 749–756. 2

[QGI] QGIS:. http://www.qgis.org/en/site/. 4, 7

[Rai34] RAISZ E.: The Rectangular Statistical Cartogram. *Geographical Review* (1934), 292–296. 2

[RWA*13] RIND A., WANG T. D., AIGNER W., MIKSCH S., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B., ET AL.: Interactive Information Visualization to Explore and Query Electronic Health Records. *Foundations and Trends in Human–Computer Interaction 5*, 3 (2013), 207–298. 2

[SDW09] SLINGSBY A., DYKES J., WOOD J.: Configuring Hierarchical Layouts to Address Research Questions. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 977–984. 2, 5, 6

[SDW10] SLINGSBY A., DYKES J., WOOD J.: Rectangular Hierarchical Cartograms for Socio-economic Data. *Journal of Maps 6*, 1 (2010), 330–345. 2

[SDWR10] SLINGSBY A., DYKES J., WOOD J., RADBURN R.: OAC Explorer: Interactive Exploration and Comparison of Multivariate Socioeconomic Population Characteristics. *proceedings of GIS Research UK* (2010), 167–174. 3

[SS16] SETLUR V., STONE M. C.: A Linguistic Approach to Categorical Color Assignment for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 698–707. 7

[Tel14] TELEA A. C.: *Data Visualization: Principles and Practice*. CRC Press, 2014. 7, 8

[TML*17] TONG C., MCNABB L., LARAMEE R. S., LYONS J., WALTERS A., BERRIDGE D., THAYER D.: Time-oriented Cartographic Treemap for Visualization of Publice Health Care Data. *Proceedings of the Conference on Computer Graphics and Visual Computing (CGVC), 2017* (2017). 1

[Tob04] TOBLER W.: Thirty Five Years of Computer Cartograms. *ANNALS of the Association of American Geographers 94*, 1 (2004), 58–73. 2

[vKS07] VAN KREVELD M., SPECKMANN B.: On Rectangular Cartograms. *Computational Geometry 37*, 3 (2007), 175–187. 2

[WBDS11] WOOD J., BADAWOOD D., DYKES J., SLINGSBY A.: BallotMaps: Detecting Name Bias in Alphabetically Ordered Ballot Papers. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2384–2391. 4

[WBH15] WEST V. L., BORLAND D., HAMMOND W. E.: Innovative Information Visualization of Electronic Health Record Data: A Systematic Review. *Journal of the American Medical Informatics Association 22*, 2 (2015), 330–339. 2

[WD08] WOOD J., DYKES J.: Spatially Ordered Treemaps. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1348–1355. 3

[WSD11] WOOD J., SLINGSBY A., DYKES J.: Visualizing the Dynamics of London's Bicycle-hire Scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization 46*, 4 (2011), 239–251. 4

9. Appendix

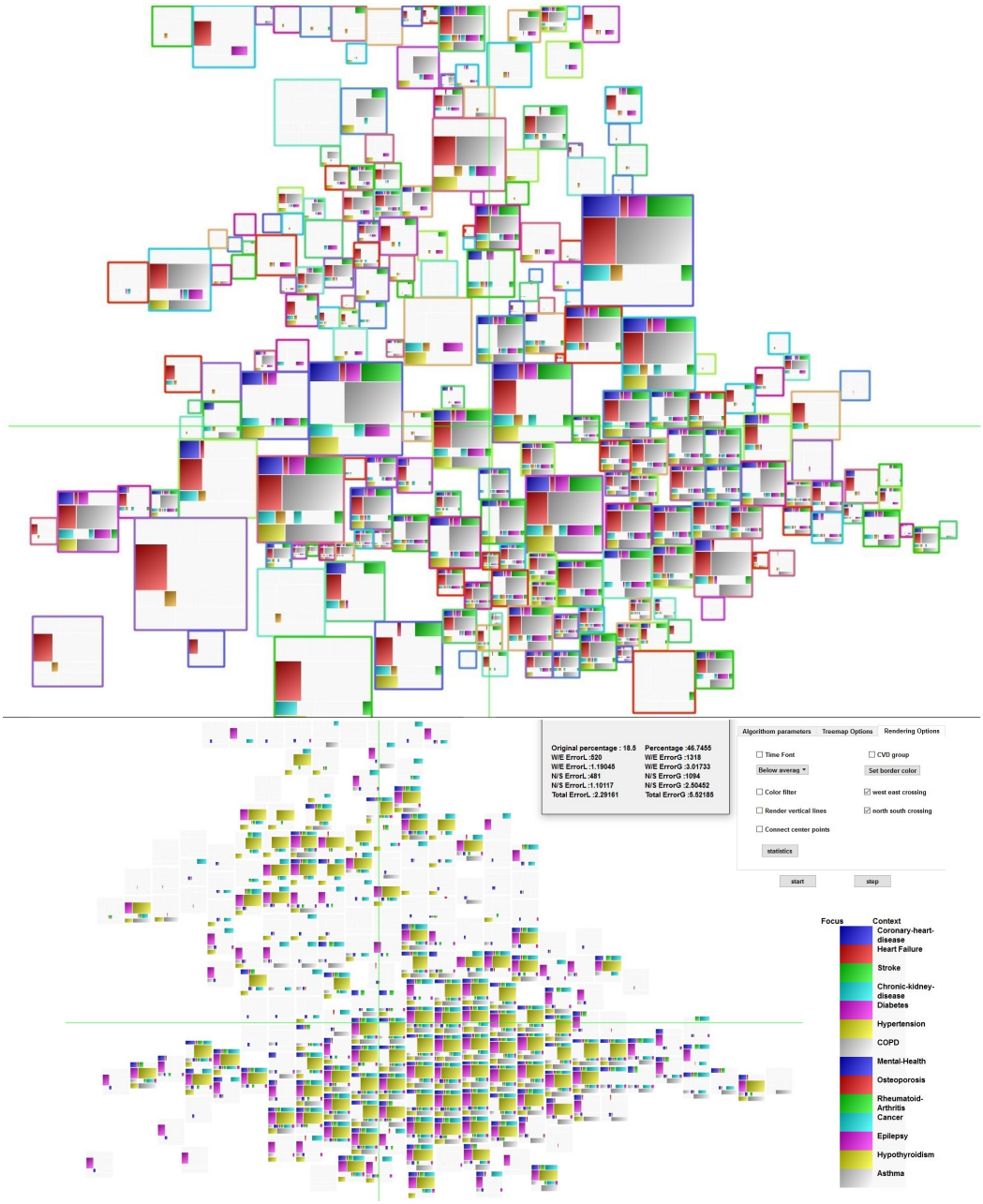

Figure 12: *A high-resolution version for Figure 1. This graph shows each region size proportional to its population with an added below average filter (left). The percentage of screen space occupied, s_0 = 41% and the local error, e_l = 3.5%, e_g=8.7% and uniform size output with a below average filter (right). s = 47%, e_l = 2.3%, and e_g = 5.5%. All the healthcare disorders that exhibit higher than average prevalence are filtered and shown as grey context. Note how the London region is healthier with the exceptions of diabetes and mental health. This is an observation based on multiple variates that would be difficult to make otherwise.*

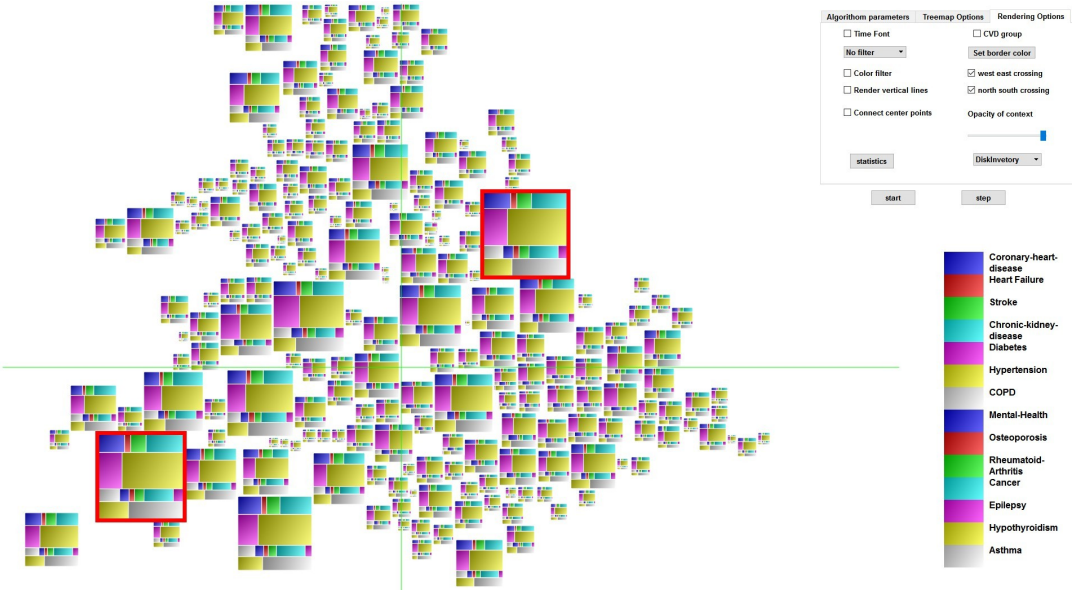

Figure 13: *A high-resolution version for Figure 5. Nodes proportional to CCG size. The screen space-filling percentage, s=36% and e_l=2.4%, e_g = 4.5. The two red outlines show the two biggest region nodes on the map: Cambridgeshire Peterborough and North East & West Devon. This is unexpected since we hypothesized the largest regions to be in London or Birmingham. This example uses color map from the Disk Inventory X tool [dis].*

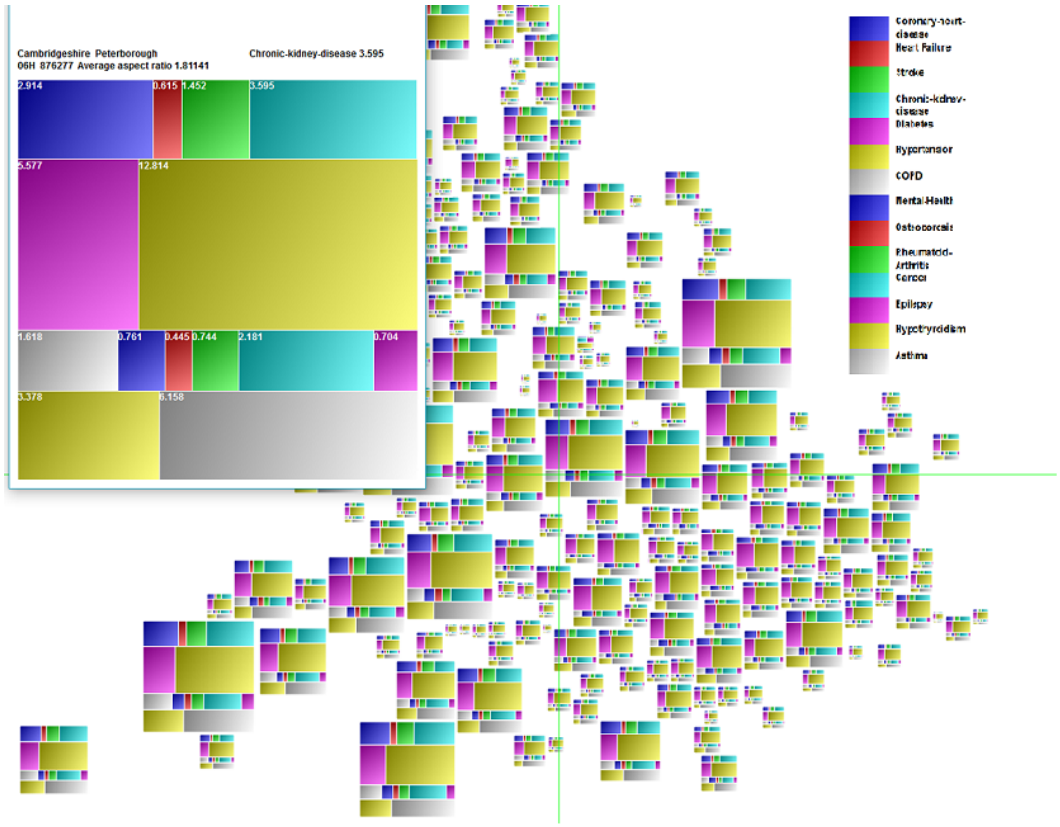

Figure 14: *A high-resolution version for Figure 6. This visualization shows the output of cartographic treemap with region size proportional to population, and with a details-on-demand window for one region node. s=30%, e_l=2.4% and e_g= 5.1%. The first three rectangles in each region node represent three CVD health disorders. Note the prevalence of hypertension and diabetes is very widespread the UK. This type of multivariate observation display itself clearly with this type of visualization.*