Ying-Chi Lin*, Anika Groß, and Toralf Kirsten

Integration and visualization of spatial data in LIFE

https://doi.org/10.1515/itit-2016-0025
Received May 9, 2016; revised July 20, 2016; accepted March 23, 2017

Abstract: It is usually a challenging task to integrate and analyze huge amounts of heterogeneous data in large medical research projects. Often meaningful new insights can be achieved by visualizing medical data on geographical maps. For instance in epidemiological studies, data is often explored on a spatial dimension. LIFE is a large epidemiological study, managed by the LIFE Research Center for Civilization Diseases at Leipzig University. The study investigates the health-related states of the local population, e.g. by looking at the role of lifestyle factors on major civilization diseases. To allow for an effective data exploration, the development of sophisticated data analysis and spatial visualization techniques is necessary. Here, we present the interactive web application *LIFE Spatial Data Visualization System* (LIFE-SDVS) that adds a geographical facet to the data integration and analysis workflow of the LIFE research project.

Keywords: Spatial data visualization, medical research data, LIFE.

ACM CCS: Applied computing → Life and medical sciences → Health informatics

1 Introduction and background

Large medical research projects often encounter challenges in data integration and analysis processes due to a huge amount and variety of data. In particular, it can be very challenging to find meaningful information in massive data volumes. Therefore, data visualization plays an important role in the data analysis pipeline, often leading to greater insights in less time. One sub-discipline of medical research is *medical or health geography* where spatial data is used to support the study of diseases and health care. It has been shown in the past that geographical information adds an indispensable value to medical research. One of the most famous examples is the research of the English physician John Snow who plotted the clusters of cholera cases around public well pumps to find the most likely source of the outbreak during the epidemic of 1854 [1]. Also nowadays, it has been shown to be very valuable to link medical and geographical data as demonstrated by applications of geographic information systems (GIS) in the health sector [2, 3].

In epidemiological studies, it is quite common to map data and analysis results onto geographic maps in order to explore disease patterns in the spatial dimension. The general goal of epidemiological studies and health surveys is to determine the healthy state of a population in a geographic region of interest and to recognize imbalances in health or disease developments. Typical results of such studies are prevalence values for specific diseases to quantitatively describe the portion of the affected population. Often, such prevalences differ between various geographic regions, e.g. the prevalence of bronchial asthma in Scotland is nearly three times higher than that in Germany (18.4% vs. 6.2%) [4]. To analyse the underlying causes for the discovered effects, such as an increasing rate of allergies, environmental factors like industry density, traffic or non-domestic plants in the region are often considered. Geographical maps have been used to identify specific locations where changes in health policy are necessary to improve the quality of care or to increase the accessibility of services (e.g., [5–7]). Moreover, health atlas platforms are practical examples that associate collected health data on geographic maps. Examples include the *U.S. Cancer Atlas* [8, 9], the *Community Health Map* [10], or the *Global Health Atlas* [11] of the World Health Organization (WHO). Typically, such platforms need to integrate different types of data from various authorities, in heterogeneous formats, with possibly overlapping or contradictory content, impeding the data analysis process.

LIFE is a large epidemiological study conducted and managed by the LIFE Research Center for Civilization

***Corresponding author: Ying-Chi Lin,** Universität Leipzig, Institut für Informatik, Augustusplatz 10, 04109 Leipzig, Germany, e-mail: lin@informatik.uni-leipzig.de
Anika Groß: Universität Leipzig, Institut für Informatik, Augustusplatz 10, 04109 Leipzig, Germany
Toralf Kirsten: Universität Leipzig, LIFE Research Center for Civilization Diseases, Philipp-Rosenthal-Straße 27, 04103 Leipzig, Germany

Diseases at Leipzig University. The goal of LIFE is to investigate the healthy state of the local population. Currently, more than 25,000 participants have attended different sub-studies. For instance, the *LIFE Adult* study investigates prevalences, genetic predispositions, early onset markers, and the role of lifestyle factors on major civilization diseases, such as metabolic diseases, depression, or sleep disorders [12]. Between 2011 and 2014, approx. 10,000 randomly selected participants aged 18–79 from Leipzig completed the baseline examination. Another sub-study, the *LIFE Child* study, is examined on a long-term basis and investigates how metabolic, environmental and genetic factors affect the health from fetal life to adulthood [13]. Follow-ups are carried out annually over a period of ten years.

Since the beginning of LIFE in 2011, a large amount of heterogeneous data has been collected by means of questionnaires, structured interviews, physical examinations, and biospecimen collection. Most of the assessment data and analysis results are integrated into a comprehensive research database. So far, there are more than 800 different assessments (i.e., investigations) and 39,000 items (i.e., questions and measurements). To facilitate the data retrieval process, the 'LIFE Investigation Ontology' (LIO) is used to semantically describe and classify the entities and their relationships in the research database [14, 15]. The amount of collected data is still increasing with the ongoing experiments and follow up studies. To allow for an effective data exploration, the support of sophisticated data analyses and visualization techniques are necessary.

At this early stage of data analysis in LIFE, the mapping of results into geographic maps has been realized only occasionally and manually. However, geospatial data visualization and analyses will be frequently applied to LIFE data. First, this will be useful to gain new insights when interpreting data related to geographical information. Second, LIFE results should be presented in a comprehensible way for both, researchers and interested users such as the inhabitants of Leipzig. In order to assist human visual perception, it is especially important to provide well structured maps with respect to map labeling. Basic techniques tend to produce confusing results, such as putting a district's name or small bar charts into the wrong district on a city map. To overcome manual spatial data visualization, we designed the interactive web application *LIFE Spatial Data Visualization System* (LIFE-SDVS), a prototype that adds a geographical facet to the data integration and analysis workflow of the large medical research project. LIFE-SDVS is very flexible in that it allows visual analysis of various kinds of LIFE data. Throughout this paper, we make the following contributions:

- We describe the system architecture of LIFE-SDVS as well as the statistical data processing and map visualization workflow. (Section 2)
- We provide algorithms for a good label positioning and boundary labeling on maps. (Section 3)
- To show the functionalities of LIFE-SDVS, we present two selected use cases on 1) the body mass index and 2) hand grip strength. (Section 4)

2 Architecture and workflow

Due to the large variety of data, it is important to provide a system which let the users choose what data they want to investigate. Moreover, an efficient and user friendly interface is needed for the many researchers who might not all have programming skills. The following sections describe the architecture and workflow we designed to achieve these requirements.

2.1 Architecture

LIFE-SDVS has a three layer architecture that consists of data source, functionalities and presentation (Figure 1). The data source layer includes the LIFE Research Database (RDB) and the LIFE Metadata Repository (MDR). The functionalities layer comprises two main units: data processing and map generation functions including various map labeling options. The presentation layer integrates the outputs, including statistics and maps, and displays them on the web interface. The system is implemented in R [16] and the web application uses the Shiny framework.

2.2 Workflow

The workflow starts from obtaining data from the LIFE database that is followed by data aggregation processes.

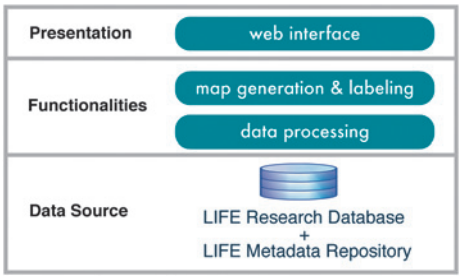


Figure 1: Three-layered architecture of LIFE Spatial Data Visualization System.

The aggregated data are integrated into maps or listed in form of a table. Finally, the web interface displays the maps and tables. Furthermore, the web interface provides interactive functionalities allowing users to visualize specified data and to produce customizable maps.

The LIFE assessment data are aggregated by geographical regions and the resulting statistics are presented on their associated regions on the map. The spatial visualization of statistics focuses on two data types: continuous data and categorical data. The continuous data are visualized as so-called *choropleth style maps*, i.e. the regions of Leipzig are shaded based on statistics aggregated from continuous data, to reveal the patterns of epidemiological study results. The aggregated categorical data are visualized as pie charts or bar charts within each region on the map. For these different ways of visualization, different data aggregation methods and plotting functions are used for each data type. In the following, we describe these workflow steps in more detail including the different processing methods for continuous and categorical data.

LIFE Research Database and data access

The RDB is the central data hub in LIFE. It integrates all relevant data that have been recorded during the study. The MDR collects the metadata describing the data in RDB and, hence, allows to create queries over the integrated data in RDB. All study data are associated with participant pseudonyms. Similarly, identification data, names, and addresses are associated with the participant pseudonym but managed in a separate system (and database) due to data privacy requirements. Address data allow us to find the area codes of the administrative regions of Leipzig. From data privacy point of view, this mapping (participant pseudonym – area code) is feasible as long as the exact living place (i.e., address) is not recognizable. Both, the area mapping and the scientific data (including gender and age at time of study participation) of RDB, are the basis for the visual analysis in LIFE-SDVS.

To enable the flexibility for exploring the great variety of LIFE assessment results, LIFE-SDVS is designed to have direct access to both, the LIFE Research Database and the Metadata Repository. On the web interface, the users can specify which data they want to explore. This request is transformed into an SQL query to obtain data from the database. Each row of the raw data set retrieved from the database contains information for one participant. A row in a continuous data set includes (1) the region where the participant lives, (2) gender, (3) age and (4) the *value* of the assessment result of the participant. Categorical data sets differ in the last element (4) that contains the information to which *category* the participant's assessment result belongs.

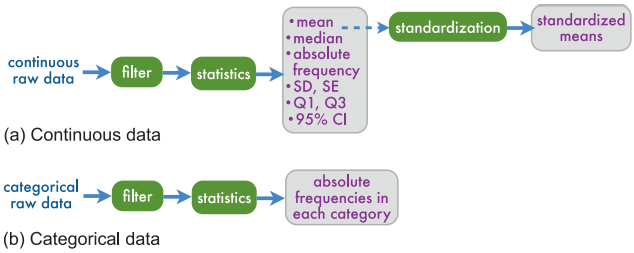


Figure 2: Data processing pipelines in LIFE-SDVS. Main data processing components are in green boxes and the items in the gray boxes are the outputs. SD: standard deviation, SE: standard error, 1Q and 3Q: first and third quantile, CI: confidence interval.

Data processing

The data processing phase aims to aggregate the raw data obtained from LIFE Research Database into statistics for each region. Figure 2 illustrates the data processing pipelines for continuous and categorical data types. Both pipelines contain the components *filter* and *statistics*. An optional *standardization* step can also be applied to the means of continuous data.

For both data types, the user can utilize the *filter* function to select particular data sub groups. All assessment data stored in the LIFE database have three attributes in common: (1) gender, (2) age, and (3) absolute frequency (i.e. number of participants). Hence, these three attributes are chosen for the filter function. The gender filter specifies if the data of only male, only female or of both genders shall be aggregated. Currently, LIFE-SDVS focuses on the visualization on data of the *LIFE Adult* study, hence the age filter contains three age groups: (18,40], (40,60], and (60,80+]. Sometimes, only a limited number of participants attended an assessment in a specific region. This leads to small sample sizes such that statistics are not representative. Therefore, the absolute frequency filter enables the users to set a threshold on minimum absolute frequency. In the visualization of continuous data, regions with a sample size below this threshold obtain a specific color to indicate there is too little or no data available in the region. Similarly, the pie charts or bar charts are absent in such regions.

The selected sub groups of assessment data are further aggregated into various *statistics* for each region (see Figure 2a). For continuous data sets, we compute the mean, standard deviation, standard error and 95% confidence interval. The three quantiles (the first, third quantiles and the median) are also provided since they are less susceptible to long-tailed distributions and outliers, compared to

the mean. The means of continuous data can further be standardized. *Standardization* in epidemiological study is a statistical adjustment technique used in instances where the statistics in focus is influenced by some other underlying factors, such as the age or gender structure of the sampled populations [17]. We apply *direct age standardization* which is the most predominant technique among the many proposed standardization methods, as described in Ahmad et al. [18]. For categorical data sets, the absolute frequencies in a region are aggregated according to categories of an assessment measurement.

Map generation and labeling
In this workflow step, the outputs from the data processing phase are visualized on maps. The core features of map generation in LIFE-SDVS are:

1. generating choropleth maps for continuous data,
2. generating maps with bar charts or pie charts placed in each region for categorical data,
3. finding a good labeling position for each region,
4. realizing boundary labeling.

The system enriches the administrative regions of Leipzig with the results obtained by the LIFE assessments. The spatial data of administrative regions are provided by the Leipzig city government and are available in two spatial resolutions: in 63 districts (Ortsteil) or in 10 boroughs (Stadtbezirk). These administrative regions consist of sets of points where each set defines a border of a region and the points are stored as coordinates.

To color each region, we use the assessment means, medians or absolute frequencies obtained from continuous data sets. A single plotting function based on the R plotting system `ggplot2` enables the automatic generation of these choropleth maps. Many map features are available to improve the aesthetic presentation of the maps such as label size, font and area filling colors. Users can assign either an internal label, an external label (i.e. boundary labeling) or no label to each single region on the map.

For categorical data, a pie chart or bar chart showing the aggregated absolute frequencies of each category of an assessment item (e.g. body mass index) is placed in each region on the map. The categories of an assessment item are stored as numeric codes in RDB. The corresponding text denotations of these categories are stored in MDR. The codes and their corresponding text denotation are extracted as a metadata view from the MDR. The legends of the maps are then obtained from this view.

To show this textual and non-textual information on the map, LIFE-SDVS needs to provide different labeling options. We can use short internal labels such as the district IDs to denote all regions in the limited space. Furthermore, it is also desirable to show statistical graphics in each region of Leipzig. To find the visually sound positions for placing text or graphics, a novel algorithm called *Label Positioning Algorithm* (LPA) is proposed (see Section 3.1). Often, it is necessary to use long label texts that do not fit into the regions. In such cases, boundary labeling (see Section 3.2) can be applied to place the labels outside of the map area (for an example see Figure 6 in Section 4). To the extent of the authors' knowledge, LIFE-SDVS is the first R package containing the boundary labeling function for maps.

Interactive web application
Users need to have basic knowledge of R to make direct use of the map generating functions. Moreover, it is a laborious job to manually input the argument values whenever changes on the maps are necessary. We therefore built an interactive web application to create a user friendly data exploration and map generation tool. With just a few clicks and settings, the LIFE scientists are able to produce customizable maps in a short time. The resulting maps can also be downloaded for further usage in scientific publications or reports.

3 Map labeling in LIFE-SDVS

Since the Leipzig map in LIFE-SDVS comprises solely of administrative regions, the corresponding labeling problems belong to the field of area-feature labeling (for a classification of labeling problems see [19]). However, conventional area-feature labeling usually only considers text labels. By contrast, we wish to enrich each region on the map not only with short text, but also with graphics and external labels. Furthermore, we also allow partial overlaps of the labels so that label sizes do not need to be scaled down. By doing so, labels could slightly exceed the border of their regions. However, we observed that as long as the midpoints of the labels are well-located, the association of labels and their corresponding regions can still be easily identified.

A common practice of placing labels on choropleth maps is to use the centroids ([20] and in R, the `map.text` function in package `maps` and the `gCentroid` function in the package `rgeos`). The application of centroids as labeling points on the Leipzig map shows several violations of good labeling positions (see Figure 3). The label of district `53` locates outside of its area (purple color). Many of the centroid points are too close to the border, such as districts `10`, `51`, `72`, `81`, `93` (blue color). Some centroid points (e.g.,

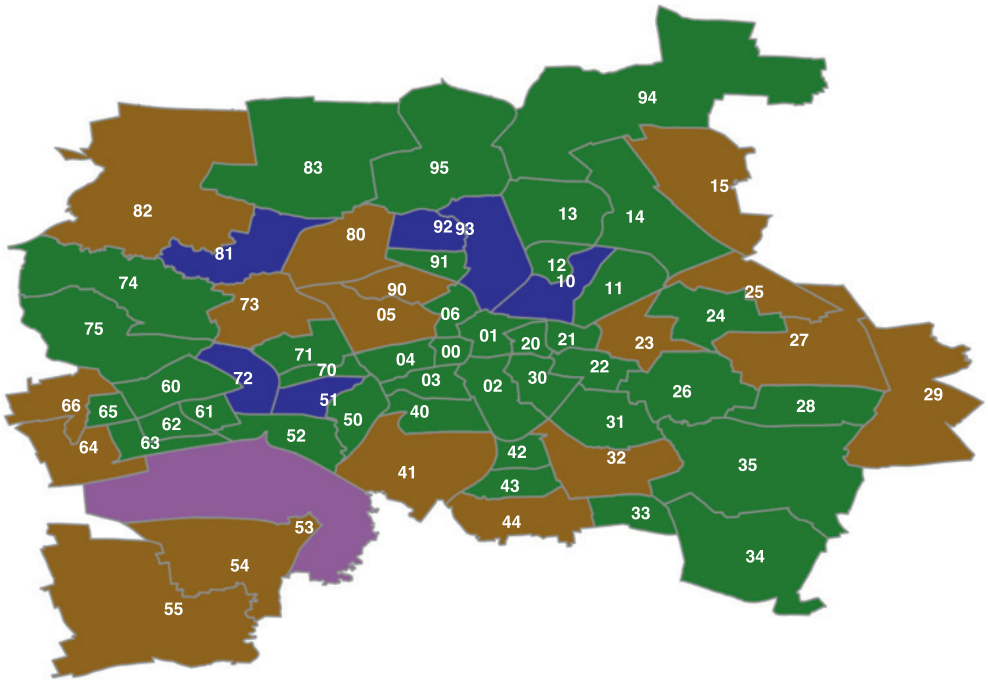


Figure 3: Map of Leipzig with district IDs positioned at the centroid points of each region. The districts are colored based on the positions of their centroid points: (purple) the centroid point locates outside of its area, (blue) the centroid point is too close to the border, (brown) the centroid point does not situate close to the middle of its area.

districts 15, 41, 55, 82) do not situate close to the middle of the associated regions (brown color). Therefore, we developed a fast heuristic to locate good labeling positions.

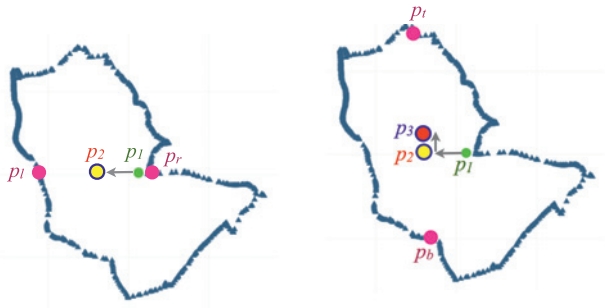
3.1 Label Positioning Algorithm

Labeling quality depends on many factors and reflects human visual perception and experience [21, 22]. Thus, there is not a single exact definition of an optimal label placement of a region, such that a heuristic solution is needed. We propose the *Label Positioning Algorithm* (LPA) to find good labeling positions in LIFE-SDVS. LPA is a generic algorithm, i.e. its application is not limited to the map of Leipzig. The algorithm first generates five candidate positions for each region, and then selects the most suitable labeling position out of the candidate set.

Candidate generation

The border of a region comprises many border points given by x- and y-coordinate pairs. The first labeling position candidate of a region is the *middle point* p_1 whose x-coordinate is calculated with $(x_{max} - x_{min})/2$ where x_{max} and x_{min} are the maximum and minimum values of x-coordinates among all border points of the region. The y-coordinates of p_1 is obtained analogically. The green point in Figure 4a shows the middle point of a district in Leipzig.

Further vertical and horizontal position adjustments generate the other four labeling position candidates. As



(a) a horizontal adjustment (b) a horizontal and then a vertical adjustment

Figure 4: (a) A horizontal position adjustment and (b) a horizontal followed by a vertical position adjustment of LPA.

seen in the example in Figure 4a, the middle point might locate close to the region border. To move the labeling position away from the border, one can shift the labeling position *horizontally* to the left. This is done by taking the right and left points of the region border which have the same y-coordinate as p_1 (i.e. p_r and p_l in Figure 4a) and compute again the middle of the x-coordinates of these two points, i.e.$(x_{p_r} - x_{p_l})/2$. The new position is the second labeling position candidate p_2. The third labeling position candidate p_3 is generated by taking a further *vertical* shift as illustrated in Figure 4b. That is, the y-coordinate of p_3 is the middle of the y-coordinates of the following two border points: the border point on top p_t and the border point on the bottom p_b. To conclude, the position of p_2 is generated by taking a horizontal adjustment from p_1 and a further vertical adjustment from p_2 generates p_3.

Analogically, the fourth candidate p_4 is obtained by taking a vertical adjustment from p_1 and a further horizontal adjustment from p_4 generates the fifth candidate p_5.

Best candidate selection

The selection of the *best* position is based on the idea that a good labeling position shall keep good distance from any point of the region border so that labels or graphics do not overlap with the border excessively. Therefore, LPA chooses the candidate that locates furthest away from any region border point as the labeling position. This is realized by first calculating the Euclidean distances of each candidate to every region border point. Then, the minimum Euclidean distances of each candidate to the border points (i.e., the shortest distances) are stored. Finally, among the five candidates the one with the largest 'shortest distance' is selected as best labeling position.

The application of LPA on Leipzig districts gives good results (for examples see Section 4). Moreover, each of the five different candidate position types has been selected for some of the regions indicating that all of them are beneficial to find the best labeling positions.

3.2 Boundary labeling

The boundary (external) labels are designed to be placed on the left, right and bottom margins around the map in three corresponding labeling groups: `right_group`, `left_group` and `bottom_group`. Users can assign regions to each of these labeling groups. The position of each labeling group is defined with three parameters. For example, the position of the `right_group` is defined by: (1) the y-coordinate for the label on the top (2) the y-coordinate for the label on the bottom and (3) the x-coordinate for the left alignment of all the labels in the `right_group`. The positions for the other two labeling groups are defined analogously with the labels in the `left_group` aligned on the right and those in the `bottom_group` aligned on the top.

The line segments that connect each external label to their associated regions are called *leaders*. Straight lines are chosen as the leader type for maps since they are simple, often used by professional graphic designers and show good performance with respect to user readability [23]. The labeling positions found by LPA are used as starting points for the leaders.

One important aim is to avoid intersecting leaders. The order of the boundary labels within each labeling group determines if the leaders of the boundary labels intersect with each other. The Label Alignment Problem (LAP) is to determine which region's label shall be assigned to which boundary labeling position. Given a set of internal labeling positions P_{in} and a set of boundary labeling positions P_{bo}. The aim is to find an bijective assignment of the elements in P_{in} to the elements in P_{bo} so that no line intersection occurs. The *Label Alignment Algorithm* (LAA) is designed to solve LAP.

Label Alignment Algorithm

Taking boundary labels on the right margin as an example, one can observe that a line-line intersection occurs when the leader of the boundary label above has larger slope than the leader of the boundary label beneath. Thus, LAA solves the intersection problem by re-ordering the boundary labels based on their line segment slopes. The concept of LAA is similar to the idea of Bekos et al. [24] though they do not explicitly refer to the utilization of slopes. In LIFE-SDVS, LAA is embedded in the R plotting function. As a consequence, when the users adjust boundary labeling parameters on the web interface, the generated map reacts dynamically to these changes and the boundary label positions are re-aligned to the corresponding regions interactively.

4 Use cases and analysis

Two use cases are selected to demonstrate the spatial visualization using LIFE-SDVS: body mass index (BMI) for the categorical data type and hand grip strength for the continuous data type. We use different types of maps as well as various features including different labeling types to show the LIFE-SDVS functionalities. In both use cases, only districts with absolute frequencies for more than 30 participants (called *valid district* in the following text) are considered.

Use case I: body mass index

People who are obese have higher risk of many diseases and health problems such as high blood pressure, sleep apnea, type 2 diabetes and stroke [25–28]. Hence, it is important for the LIFE study to investigate the prevalence of obesity in Leipzig and how it is related to geographical regions within the city. The body mass index (BMI) can be used for the assessment of overweight and obesity [29]. It is computed by an individual's weight in kilograms divided by the square of height in meters (i.e. in units kg/m^2). WHO classifies the BMI values into four standard categories and this classification is used to assign the LIFE participants.

Figure 5 shows the absolute frequencies in each BMI category of an example data set plotted as pie charts in each valid district. On top of the map we show the cor-

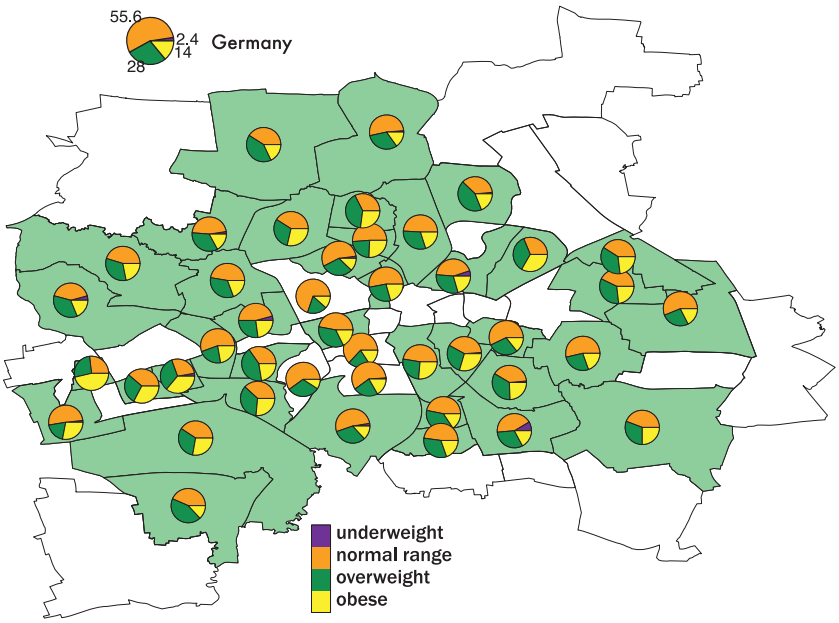


Figure 5: An example of categorical data visualization in LIFE-SDVS using BMI in Leipzig districts. The proportions of each BMI category are plotted as pie chart in districts with at least 30 participants. Districts are colored in green, if the sum of proportions for *overweight* and *obese* exceeds the corresponding sum for the German reference data.

responding proportions for the German population based on data from *Gesundheitsberichterstattung des Bundes (GBE)* [30] as reference value. If the sum of the proportions in the categories *overweight* and *obese* of a valid district is higher than that of German data, the district is colored in green. The example map demonstrates that the majority of the valid districts have higher proportions of *overweight* and *obese* cases than the German reference data. Furthermore, the valid districts which are not colored in green locate close to the center of the city. Therefore, by utilizing this type of map, the districts with especially high proportion of overweight or obese can be easily identified so that further investigation or intervention can be taken. For instance, we plan to investigate the relationships between overweight/obesity and lifestyle factors or sociodemographic clusters in Leipzig (e.g., income, employment rate).

Use case II: hand grip strength

Low hand grip strength in healthy adults across all ages predicts increased risk of premature mortality, disability in later life and risk of post-surgery complications or longer hospitalization (e.g., [31–33]). Hand grip strength is measured by squeezing a dynamometer and the values are taken in kilogram. Due to its ease of measurement and high predictive power on short-term mortality, hand grip strength is considered as an important biomarker.

Figure 6 illustrates two different maps produced using LIFE-SDVS to present aggregated continuous data such as the median of hand grip strength. Different coloring palettes can be chosen to fill the regions. The regions colored in gray are the regions with a sample size below the given threshold. We applied all three labeling options, i.e. internal labeling, boundary labeling and no labeling, to demonstrate the labeling features of LIFE-SDVS. In Figure 6a, the district IDs are used as internal labels to denote the regions. The labels have been placed using the best labeling positions found by LPA. Alternatively, customizable boundary labels can be assigned to highlight specific results. For instance in Figure 6b, only the 15 districts with the lowest handgrip strength values have been labeled. For better readability and clarity, the users can also choose to remove all labels for the remaining regions.

5 Conclusion and future work

The spatial visualization tool LIFE-SDVS is built to cope with the large amount and variety of data in the LIFE-project. LIFE-SDVS is an efficient data exploratory platform and a map generation tool enabling sophisticated data analysis and spatial visualization. LIFE-SDVS has direct access to the LIFE database ensuring the flexibility of accessing various assessment results based on current data. We proposed two algorithms, LPA and LAA, to optimize the label positioning and boundary labeling problems. These algorithms are also applicable to other maps composed of different regions. Hence, though LIFE-SDVS is a prototype based on LIFE data, other health-related studies can profit from its functions. In the future, we plan to release the implemented R package covering the optimized labeling functions for maps.

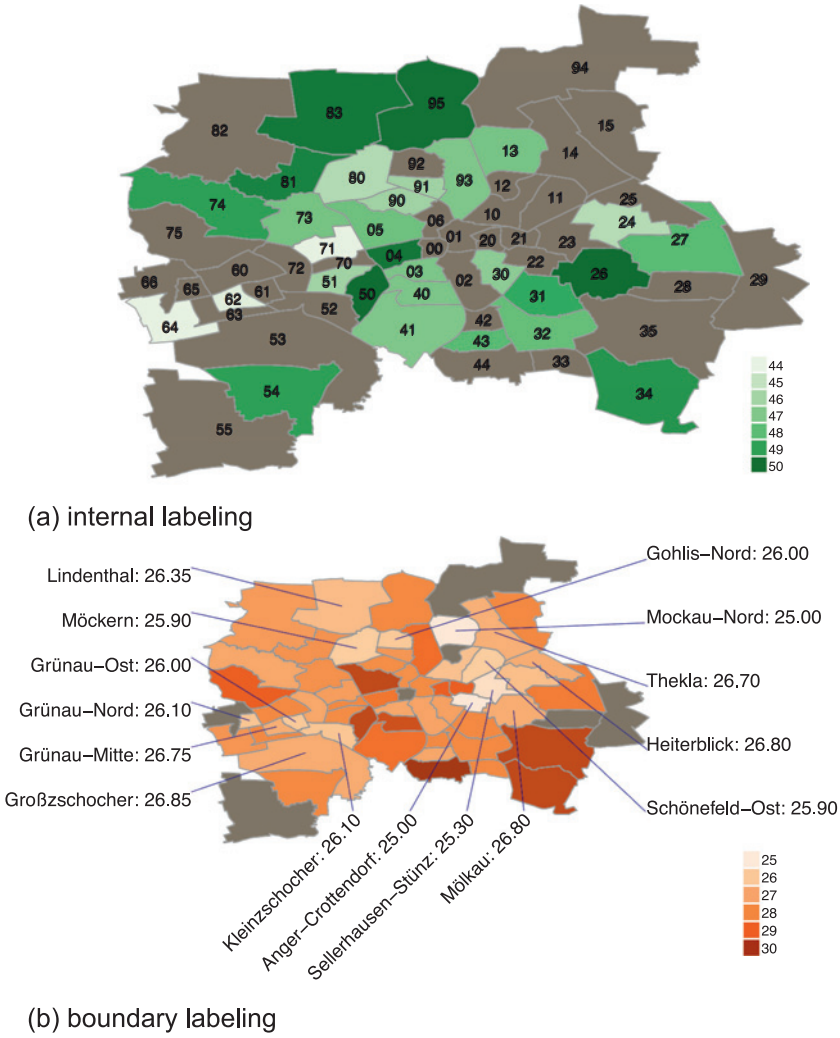


Figure 6: Examples of visualizing continuous data in Leipzig districts. Subfigure (a) uses internal labeling (b) uses boundary labeling. The districts are colored based on the aggregated hand grip strength data (in kg). The boundary labels in (b) show the district names with the aggregated values.

At this early development stage, some features can still be improved. For instance, users currently need to assign regions to boundary labeling groups manually. This process can be automated. Further, we plan to provide some rule-based assignment methods for the boundary labels such as labeling only regions with an average above a given threshold. We will extend LIFE-SDVS by several additional functionalities. For instance, more types of statistical results might be displayed on maps. Moreover, the comparison of results (e.g. different gender, age groups etc.) could be facilitated by placing two or more maps on one page. We further plan to add temporal aspects to visualize developments over time which is essential for long-term studies such as the *LIFE Child* study.

Acknowledgement: This publication is supported by LIFE and LHA. LIFE (Leipzig Research Center for Civilization Diseases) is funded by means of the European Union, by the European Regional Development Fund (ERFD) and by means of the Free State of Saxony within the framework of the excellence initiative. LHA (Leipzig Health Atlas) is funded by German Federal Ministry of Education and Research (grant 031L0026, "i:DSem – Integrative Datensemantik in der Systemmedizin")

References

1. J. Snow. On the mode of communication of cholera. John Churchill, 1855.
2. E. C. Fradelos, I. V. Papathanasiou, D. Mitsi, K. Tsaras, C. F. Kleisiaris, and L. Kourkouta. *Health based geographic information systems (GIS) and their applications*. Acta Informatica Medica, 22(6):402, 2014.
3. G. Rushton. *Public health, GIS, and spatial analytic tools*. Annual Review of Public Health, 24(1):43–56, 2003.
4. M. Masoli, D. Fabian, S. Holt, and R. Beasley. *The global burden of asthma: executive summary of the GINA Dissemination Committee report*. Allergy, 59(5):469–478, 2004.

5. S. L. McLafferty. *GIS and health care*. Annual Review of Public Health, 24(1):25–42, 2003.
6. C. E. Noon and C. T. Hankins. *Spatial data visualization in healthcare: supporting a facility location decision via GIS-based market analysis*. Proceedings of the 34th Annual Hawaii International Conference on System Sciences, pages 10 pp.–, 2001.
7. N. Ray and S. Ebener. *AccessMod 3.0: computing geographic coverage and accessibility to health care services using anisotropic movement of patients*. International Journal of Health Geographics, 7(1):1, 2008.
8. W. J. Blot, J. M. Harrington, A. Toledo, R. Hoover, C. W. Heath Jr, and J. F. Fraumeni Jr. *Lung cancer after employment in shipyards during World War II*. New England Journal of Medicine, 299(12):620–624, 1978.
9. D. M. Winn, W. J. Blot, C. M. Shy, L. W. Pickle, A. Toledo, and J. F. Fraumeni Jr. *Snuff dipping and oral cancer among women in the southern United States*. New England Journal of Medicine, 304(13):745–749, 1981.
10. A. Sopan, A. S.-I. Noh, S. Karol, P. Rosenfeld, G. Lee, and B. Shneiderman. *Community health map: a geospatial and multivariate data visualization tool for public health datasets*. Government Information Quarterly, 29(2):223–234, 2012.
11. *Global Health Atlas of World Health Organization*. http://gamapserver.who.int/mapLibrary/default.aspx, 2016.
12. M. Loeffler, C. Engel, P. Ahnert, D. Alfermann, K. Arelin, R. Baber, F. Beutner, H. Binder, E. Brähler, R. Burkhardt, U. Ceglarek, C. Enzenbach, M. Fuchs, H. Glaesmer, F. Girlich, A. Hagendorff, M. Häntzsch, U. Hegerl, S. Henger, T. Hensch, A. Hinz, V. Holzendorf, D. Husser, A. Kersting, A. Kiel, T. Kirsten, J. Kratzsch, K. Krohn, T. Luck, S. Melzer, J. Netto, M. Nüchter, M. Raschpichler, F. G. Rauscher, S. G. Riedel-Heller, C. Sander, M. Scholz, P. Schönknecht, M. L. Schroeter, J.-C. Simon, R. Speer, J. Stäker, R. Stein, Y. Stöbel-Richter, M. Stumvoll, A. Tarnok, A. Teren, D. Teupser, F. S. Then, A. Tönjes, R. Treudler, A. Villringer, A. Weissgerber, P. Wiedemann, S. Zachariae, K. Wirkner, and J. Thiery. *The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany*. BMC Public Health, 15(1):1–14, 2015.
13. M. Quante, M. Hesse, M. Döhnert, M. Fuchs, C. Hirsch, E. Sergeyev, N. Casprzig, M. Geserick, S. Naumann, C. Koch, M. A. Sabin, A. Hiemisch, A. Körner, and W. Kiess. *The LIFE Child Study: a life course approach to disease and health*. BMC Public Health, 12(1):1–14, 2012.
14. T. Kirsten, A. Kiel, et al. *Ontology-based registration of entities for data integration in large biomedical research projects*. In GI-Workshop – Informationsintegration in Service-Architekturen, pages 711–720, 2010.
15. A. Uciteli and T. Kirsten. *Ontology-based retrieval of scientific data in LIFE*. In BTW Workshops, pages 109–114, 2015.
16. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
17. D. Schneider and D. E. Lilienfeld. Lilienfeld's Foundations of Epidemiology. Oxford University Press, 4th edition, 2015.
18. O. B. Ahmad, C. Boschi-Pinto, A. D. Lopez, C. J. Murray, R. Lozano, M. Inoue, et al. Age standardization of rates: a new WHO standard, 2001.
19. E. Imhof. Die Anordnung der Namen in der Karte. International Yearbook of Cartography, 2:93–129, 1962.
20. D. Dörschlag, I. Petzold, and L. Plümer. *Placing objects automatically in areas of maps. In Proc. 23rd International Cartographic Conference (ICC'03), Durban, South Africa*. Citeseer, 2003.
21. J. Christensen, J. Marks, and S. Shieber. *An empirical study of algorithms for point-feature label placement*. ACM Transactions on Graphics (TOG), 14(3):203–232, 1995.
22. K. G. Kakoulis and I. G. Tollis. *Labeling algorithms*. In Handbook on Graph Drawing and Visualization, pages 489–515. Chapman and Hall/CRC, 2013.
23. L. Barth, A. Gemsa, B. Niedermann, and M. Nöllenburg. *On the readability of boundary labeling*. In Graph Drawing and Network Visualization, pages 515–527. Springer, 2015.
24. M. A. Bekos, M. Kaufmann, A. Symvonis, and A. Wolff. *Boundary labeling: models and efficient algorithms for rectangular maps*. In Graph Drawing, pages 49–59. Springer, 2004.
25. K. Bhaskaran, I. Douglas, H. Forbes, I. dos Santos-Silva, D. A. Leon, and L. Smeeth. *Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults*. The Lancet, 384(9945):755–765, 2014.
26. S. Kasen, P. Cohen, H. Chen, and A. Must. *Obesity and psychopathology in women: a three decade prospective study*. International Journal of Obesity, 32(3):558–566, 2008.
27. F. S. Luppino, L. M. de Wit, P. F. Bouvy, T. Stijnen, P. Cuijpers, B. W. Penninx, and F. G. Zitman. *Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies*. Archives of General Psychiatry, 67(3):220–229, 2010.
28. National Heart, Lung, and Blood Institute (NHLBI). Managing overweight and obesity in adults: systematic evidence review from the obesity expert panel, 2013.
29. J. S. Garrow and J. Webster. *Quetelet's index (W/H2) as a measure of fatness*. International Journal of Obesity, 9(2):147–153, 1984.
30. *Gesundheitsberichterstattung des Bundes*. http://www.gbe-bund.de/, 2013.
31. R. Cooper, D. Kuh, R. Hardy, et al. *Objectively measured physical capability levels and mortality: systematic review and meta-analysis*. BMJ, 341:c4467, 2010.
32. C. R. Gale, C. N. Martyn, C. Cooper, and A. A. Sayer. *Grip strength, body composition, and mortality*. International Journal of Epidemiology, 36(1):228–235, 2007.
33. C. Martin-Ruiz and T. von Zglinicki. *A life course approach to biomarkers of ageing*. A Life Course Approach to Healthy Ageing, page 177, 2013.

Bionotes

Dr. Ying-Chi Lin
Universität Leipzig, Institut für Informatik,
Augustusplatz 10, 04109 Leipzig, Germany
lin@informatik.uni-leipzig.de

Dr. Ying-Chi Lin is a PhD student in Computer Science at Leipzig University. She received her first master degree in Entomology from National Chung-Hsin University, Taiwan and a PhD degree in Environmental Sciences from University of East Anglia, U.K. She has worked as a consultant for a content management software company. She is currently doing research on ontology-based annotation on biomedical forms and the influence of ontology evolution on annotations at the Database Group in Computer Science at Leipzig University and for LIFE.

Dr. Anika Groß
Universität Leipzig, Institut für Informatik,
Augustusplatz 10, 04109 Leipzig, Germany
gross@informatik.uni-leipzig.de

Dr. Anika Groß is a Postdoc in the Database Group at Leipzig University. She studied bioinformatics at Martin-Luther-Universität Halle-Wittenberg and received her PhD in Computer Science at Leipzig University in 2014. As a Postdoc she is teaching current topics such as data integration, NoSQL databases and cloud data management. Her research interests include large scale data integration, in particular the mapping and evolution of ontologies and annotated data sets, in the biomedical and other application domains.

Dr. Toralf Kirsten
Universität Leipzig, LIFE Research Center
for Civilization Diseases,
Philipp-Rosenthal-Straße 27,
04103 Leipzig, Germany
toralf.kirsten@life.uni-leipzig.de

Dr. Toralf Kirsten received a PhD in Computer Science (2007) from the Leipzig University. He was a member of the *Database and Data Integration* group (head: Prof. Erhard Rahm) from 2002–2010 at the Interdisciplinary Center for Bioinformatics, Leipzig University. In 2010, he moved to the LIFE Research Center for Civilization Diseases, University of Leipzig, where he heads the *Database and Software Development* group. From 2011 to 2012, he was external lecturer for database systems at the Telecom University of Applied Sciences in Leipzig and from 2012–2014 visiting professor for database systems at the Leipzig University of Applied Sciences, both in part-time. He has strong experiences in data integration, schema and ontology matching with applications in the life science domain.