Multiple coordinated views of massive geo data using tree maps and choropleth maps

Robert Schäfer
Department of Computer Graphics, Hasso-Plattner-Institut

July 13, 2017

Abstract

Numerous visualization techniques exist for data-driven decision support systems. Two of which, namely tree maps and choropleth maps are especially useful when dealing with hierarchical and geographical data respectively. While a choropleth map is an sensible choice for geographical data, there is no obvious mapping of arbitrary hierarchical data. On the other hand, tree maps allow to visualize arbitrary, hierarchical and multidimensional data using nested rectangles. Although a tree map looks similar to a geographical map, the result may not be as intuitive and comprehensible. There is no predefined layouting for arbitrary data, let alone no well known layouting that everybody feels comfortable with. In this thesis, we propose and evaluate a coordinated multiple view to combine advantages of both visualizations. In one view, the user gets an easy access with a familiar geographical map, while in the other view the same data is displayed in a tree map.

Should we mention the combination of multiple data sets here?

1 Introduction

The human brain processes visual information better than it processes text. As a result, the most tangible data analyses usually come with some sort of data visualization. Data visualizations on a computer allow for user interactions and different levels of granularity according to the customer demands, which provides a great user experience. There are a multitude of different techniques. Coordinated multiple views take data visualizations to the next level by exploiting the respective advantages of the used visualizations as much as possible. This combination may yield a greater value to the user, but it is unclear what kind visualization techniques work best together and for which kind of data. How a user interact with coordinated multiple views and how it differs from the use of single visualizations is another question worth to investigate. In this thesis, we consider the question how a combination of a geographical visualization with

a hierarchical visualization performs, for what kind of data this combination is suitable and what interaction patterns apply.

Establish the niche, why is there further research on your topic?

Introduce the current research, what's the hypothesis, the research question?

1.1 Motivation

We create data visualizations of multi-dimensional, hierarchical and geographical data. Namely, we develop RUNDFUNK MITBESTIMMEN which is an application for German citizen to publish which public broadcasts should benefit from their broadcasting fees. The output of this application is a public user ranking and it can be used by broadcasting corporations to evaluate their program. Data visualizations guide media researchers, journalists and the general audience to draw conclusions. In this particular use case, the selection and interaction with the data may happen geographically, but the desired visualization could show e.g. changes over time or relationships within the data.

To explain this a little deeper: Public broadcasting in Germany is organized federally, a German home belongs to the jurisdiction of a public broadcasting corporation. But the produced content can be used all German citizen and it is even required to be free and available to everyone. So this means that e.g. a media researcher might want to select all users from within a certain region, but is actually interested into relationships of broadcasts that are preferred by people from that area.

So the interaction and selection of data should happen in another view than the actual data visualization. Since we deal with geographical data, we use geometry on a map for selection and interaction. For the visualization, we use a different technique, e.g. a tree map if we deal with hierarchical data.

More scenarios here

1.2 Problem statement

2.5D tree maps visualize hierarchical data on a two dimensional canvas and are particularly suitable if the proportions of the data should be emphasized. When dealing with both multidimensional and geographical data, problems arise when other features than geographical features are used for the layouting of the 2.5D tree map. As the order and placement of items depend on their specific values and hierarchy, items that should belong together according to their geographic circumstances may be scattered across the 2.5D tree map. This obstructs the comprehensibility of map and makes it especially difficult to select geographical units of items.

1.3 Research questions

When dealing with multi-dimensional data, is it helpful to have multiple views for these dimensions? What are best practices for the implementation of coor-

dinated multiple views? Is it great user experience to have one view for interaction and another view for the data visualization to create insights? Do people prefer one view for the interaction, e.g. the geographical dimension, or do they use all views for interaction and visualization alternately?

1.3.1 Hypotheses

Linkage and coupling of coordinated multiple views can be enabled through visualizations, navigations and interaction techniques. The exploration and analysis of multi-dimensional and geographical data can be supported with these coordinated multiple views.

1.4 Objectives

No section without text

Basic coordinated multiple view The existing VISUAL ANALYTICS PLATFORM is supplemented with a basic coordinated multiple view system. Modules, interfaces and functionalities of the coordinated multiple view system are designed, coordinated and prototypically implemented. In particular, a method for arranging multiple coordinated multiple view widgets in a coordinated multiple view layout and storing coordinated multiple view layouts is developed.

Interaction with the coordinated multiple view We develop coupling mechanisms and interaction mechanisms between 2D maps (for map-based representation) and 2.5D tree maps (for abstract information representation). The functionality includes zoom per object or selection with a bounding box. This creates a powerful selection mechanism, which can be used to select the data from the map-based representation in the 2.5D tree map.

Demonstration and evaluation Coordinated multiple view layouts and suitable views are implemented and tested for the selected test data. Based on the test data sets, the coordinated multiple view implementation is examined and evaluated for design criteria[16], general usability aspects[14] and usage for typical visual analytics tasks.

How elaborate is it to conduct a study for the usability of the coordinated multiple view? Who has done that before?

1.5 Methodological approach

A literature research is used to gather knowledge about the state of the art with respect to coordinated multiple views. Existing concepts and implementations available on the internet are examined and reused if possible. Interviews of people from the target group are conducted and the define the requirements for the application. A minimal viable product is developed to further validate the

user requirements. Also, common user behaviour is observed during user tests. The prototype is continuously developed to allow for further experimentation.

Scenarios While implementing more features, we have two scenarios with different kind of data and different user requirements:

- In our work on the project RUNDFUNK MITBESTIMMEN we design and prototype visualizations for media researchers. We fully control the database and the database schema as well as on the user facing application on top of it. User requirements are tied to journalists, media researchers, broadcasting corporations. The data has geographical, hierarchical, temporal and correlated characteristics.
- The VISUAL ANALYTICS PLATFORM for administrative data is used as a more general purpose application. There is no single database schema but combatibility with many sources or services. User requirements are potentially unknown and part of the resarch. The usage focuses especially on geographical and hierarchical data.

For the scope of this master thesis we therefore compare implemented visualization and views. How do both approaches differ in development speed, value for the customer? What considerations need to be done regarding the database schema?

What is the area(s) of research, in which this thesis can be placed into?

2 Structure of the work

In section 3 we will give an overview on coordinated multiple views and focus on visual analytics as well as massive, geographical data. Existing concepts and implementations of coordinated multiple views are examined and summarized in an overview.

3 State of the art

Data visualizations are a key part in data-driven decision support systems[9][13]. Few [5] mentions sense-making (also called data analysis) and communication as some of the most important purposes of data visualization. Statistical information is abstract and in data visualization "we must find a way to give form to that which has none."[5]

Visualizations are an obvious choice for managers who demand a quick overview on performance data. In fact Kusinitz [8] explains that the human brain processes visual information 60,000 times faster than text and visual content makes up even 93% of all human communication. Data visualizations are essential here, as managers often do not have the resources to do an in depth analysis with the numbers only. We can expect to see these technologies more in more in business applications. McAfee and Brynjolfsson [12] from the MIT Center of Digital Business showed that organizations driven most by data-based decision making had 4% higher productivity rates and 6% higher profits. However, little research has been done regarding the performance of coordinated multiple views in the field of decision making. There might be a great potential. Back in 1997 Mayer [11] conducted eight studies to compare the effect of using multimedia on university students. The studies showed that when using combined visual and verbal explanations the generation of creative problem solutions increased by an average of more than 50%.

So the application of combined data visualization techniques in decision making seems to be a promising strategy. Nevertheless is is unclear, which visualization techniques are the most suitable to be used in combination. If we know what kind of data we are dealing with, what are the best suited visualization techniques? Let's say we have multidimensional data, is there an order in how people access these multiple dimensions? How do these visualizations perform and what are best practices to be considered for their implementation?

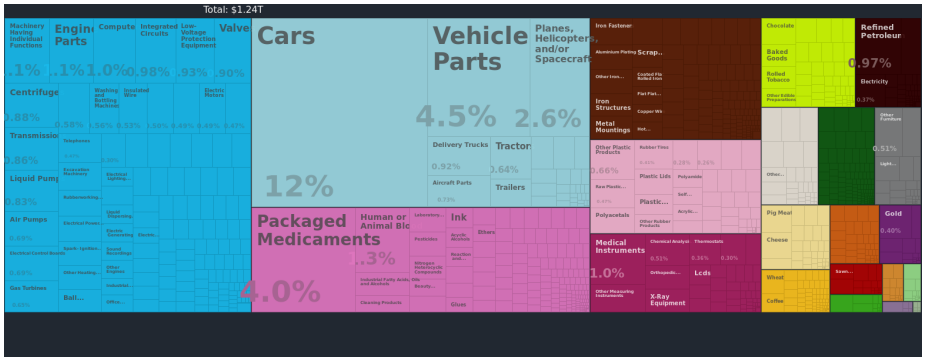
3.1 Data related challenges

Add independent section and explain tasks like data cleansing or data reliability regarding different sources here

3.2 Visualization of hierarchical data

Due to the hierarchical nature of the of our use cases, we focus on the visualization of hierarchical and geographical data The visualization of hierarchical data has a long tradition. The traditional representation of a tree is a rooted, directed graph with the root node at the top. An everyday use case is a directory tree example of a file system, e.g. in file browsers or command line utilities like `tree` in UNIX. As Shneiderman [15] mentions, this visualization becomes increasingly large when displaying more than one node and soon exceeds the entire screen size. Johnson and Shneiderman [7] proposes the tree map visual-

ization technique, in which each node is a rectangle whose area is proportional to some attribute, thus making 100% use of the available screen size. As we can see in figure 1 large boxes are labeled with generic tems like “cars” and “medicaments” and include smaller boxes with more specific meanings. We apply the same rules to ordinary maps. The world can be divided into continents, which can be divided into countries, which can be divided into provinces and so on. The difference is that there is no predefined algorithm for layouting, which brings up one of the major disadvantages of tree maps: As the order of placement depends on the respective features of the nodes, small changes in the input data can lead to a large change in the layout of the resulting tree map.



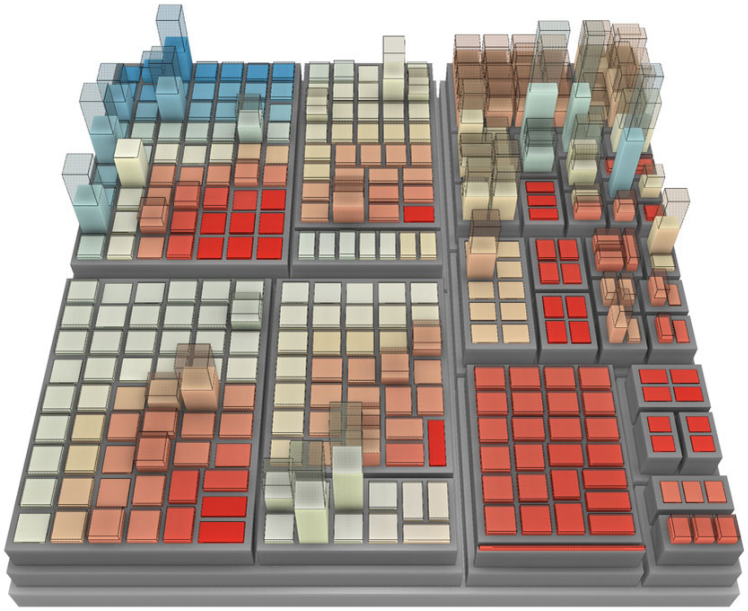
Observatory of Economic Complexity[6]

Figure 1: German exports visualized as a tree map

3.2.1 3D tree maps and 2.5D tree maps

In 2004, Bladh, Carr, and Scholl [1] transfer the concept of tree maps from two dimensional into three dimensional space. The introduce StepTree, which is a three dimensional tree map to display a directory layout. It “differs from Treemap in that it employs three dimensions by stacking each subdirectory on top of its parent directory.”[1] 3D tree maps are superior to 2D tree maps for tasks with a pronounced topological challenge. User perform significantly better in interpreting the hierarchical structure. However, 3D visualizations also introduce some disadvantages as superimposition of objects and a complex view point navigation.

Limberger et al. [10] introduce the concept of a 2.5D tree map which is a constrained 3D tree map. A 2.5D tree map has all items attached to the ground. For the rest of this thesis, we will refer to this type of tree map. We can see an example of a 2.5D tree map in figure 2



© Hasso-Plattner-Institut[4]

Figure 2: Example of a 2.5D tree map

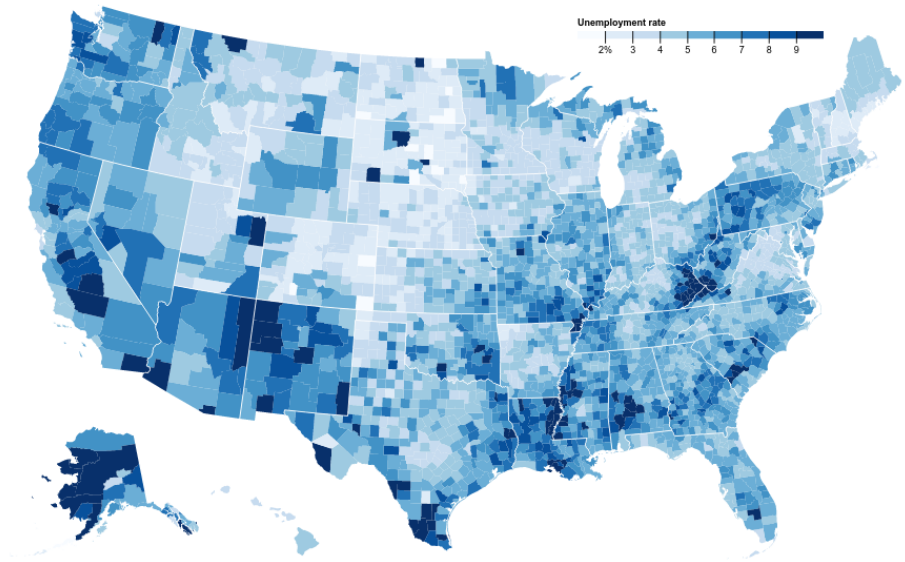
3.3 Choropleth map

A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map. A popular use case is the display of population density or per-capita income. We can see an example of a choropleth map in figure 3. Choropleth maps are extremely popular and so the audience is likely to understand them. They are very helpful when data is attached to enumeration unites like counties, provinces and countries.

How do tree maps relate to geographical data?

3.4 Coordinated multiple views

According to Roberts [14] coordinated multiple views is just "a specific exploratory visualization technique that enables users to explore their data". Coordinated multiple views are characterized by the fact, that they show multiple views side-by-side. Most multiple coordinated views also provide some kind of brushing technique. "The technique of brushing is the principle approach, where elements are selected (and highlighted) in one display, concurrently the same information in any other linked display is also highlighted."[14] We can see an example in figure 4. It displays an on-time performance of airlines, visual-



GPLv3 Mike Bostock[2]

Figure 3: Unemployment rate in the USA

ized with the "Crossfilter" javascript library. The user can set the borders of an interval with the mouse in each of the views. The visualization takes the most recent 80 flights from the database that match all given filters. All visualizations are then updated in real time. As we can see in the example in figure 4 there seems to be a correlation of a long delay with a later time of the day.

4 Concept

Since we deal with real world problems, we aim to evaluate the developed tools on real users and existing data.

4.1 Data sets

For our uses cases, we have two different data sets: One data set consists of a user ranking of public broadcasting in Germany, i.e. entities, mostly TV and radio broadcasts, are liked or disliked by people. This data is public data and can be used by media researchers of broadcasting corporations but also targets media journalists and the general audience. The other set, called `RISO`, consists of statistical data from various German administrations and is used by the authorities for urban planning and policy strategies.

Both data sets share some characteristics. The administrative data connects

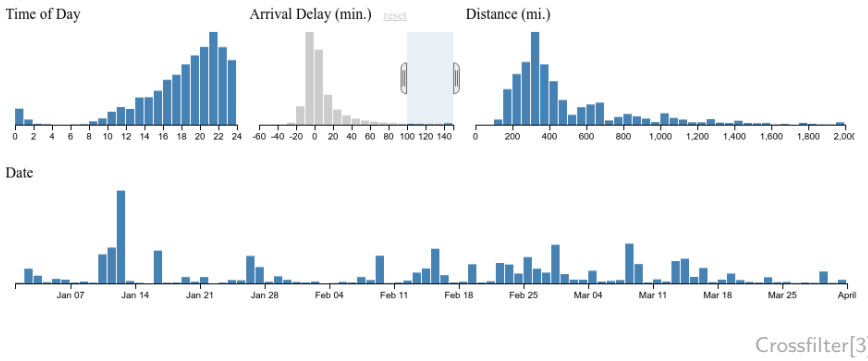


Figure 4: Airline on-time performance: Correlation of time of day with arrival delay. Most recent flight with a delay of more than 100 minutes selected.

certain features with certain regions of Germany. As Germany is a federal state, larger regions consist of many other smaller regions.

The second one consists of user data that was collected through a web application called Rundfunk MITBESTIMMEN.

4.1.1 RISO

The `RISO` data base is used in by local authorities to get insights about governmental KPIs to assist local and regional decision making. It is a relational database and a part of the data base schema is shown as an ER diagram in figure 5.

The largest table is called `data` with approximately 10,466,600 records, which holds all values along with the survey date.

Features This data is connected to a feature table through a foreign key called `id_mm`. In the feature table we can find the description for every referenced feature, e.g. population density, working population in agriculture, education spending. The `RISO` system groups all features in a 4-level hierarchy:

1. `katalog_daten_1_kategorien`
2. `katalog_daten_2_themenfelder`
3. `katalog_daten_3_themen`
4. `katalog_daten_4_merkmale`

The actual features table is the last one in the list. At the lowest level within the hierarchy, this is the largest table with 1234 records.

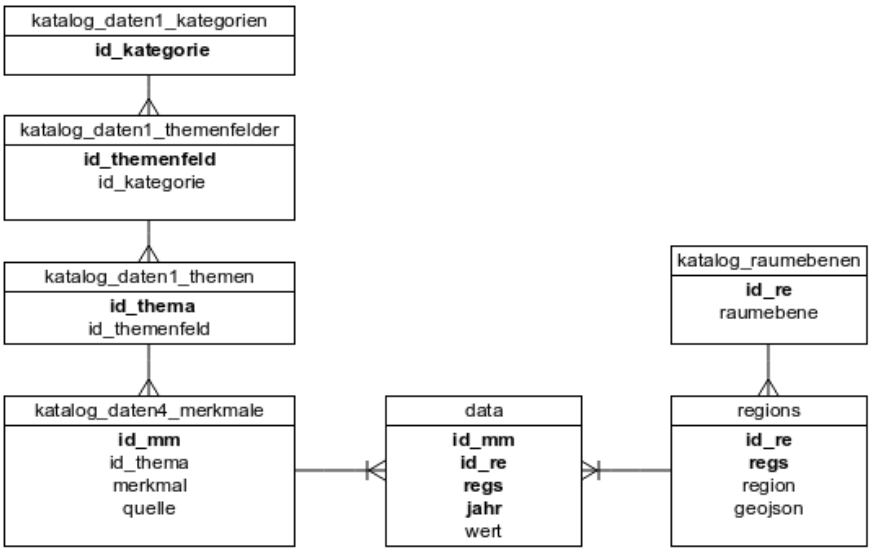


Figure 5: Part of the `RISO` database schema. Primary keys are set in bold.

Regions On the other side, the geographical data is stored in the `regions` table. The geometry data for each region is stored in the `geojson` column and as the name suggests, the data type is a `geojson`. The foreign keys that connect the tables `data` and `regions` are called `id_re` and `regs`. Unlike the feature table, the regions are grouped through the `id_re` that indicates the hierarchy level. So the values of the `id_re` column denominate the level of the hierarchy. E.g. a region with a `id_re` of 1 is a federal state of Germany, a region with id 13 is a constituency. A textual description for the hierarchy level can be found in the `katalog_raumebenen` table in column `raumebene`. Both column `id_re` and `regs` belong to the primary key of the regions table, so there will never be two regions on the same hierarchy level with the same `regs` id.

Characteristics As we can see, the schema of the `RISO` database follows a rather denormalized approach. The schema does not make a lot of assumptions regarding the input data. It allows to add data of arbitrary size, features and completeness as long as there is some kind of numerical data associated with some kind of geographical unit. This approach is suitable for a data base that incorporates data from different sources, as it is the case with the `RISO` data base.

4.1.2 Rundfunk MITBESTIMMEN

Unlike the `RISO` database, the data base of Rundfunk MITBESTIMMEN is used as persistency layer. For that reason the data base schema follows the

requirements of a productive web application.

As outline in section 1.1 RUNDFUNK MITBESTIMMEN is an evaluation platform for public broadcasting in Germany. First, users vote on broadcasts, i.e. they decide if they want to support broadcast or if they do not want to support. As a next step, user can then make a priorisation by distributing a virtual, monthly budget among the chosen broadcasts.

Figure 6 shows the data base schema of the application. A `user` is connected to `broadcast` through a `selection`. If the `user` supports some a broadcast, the `response` on the given `selection` will be 'positive'. If the `user` does not wish to support a `broadcast`, the `response` will be 'neutral'. The `user` can allocate virtual money to supported broadcasts. The money will be stored in the column `amount` of the `selection`. The sum of all amounts for one user will never exceed the virtual budget of 17.50€.

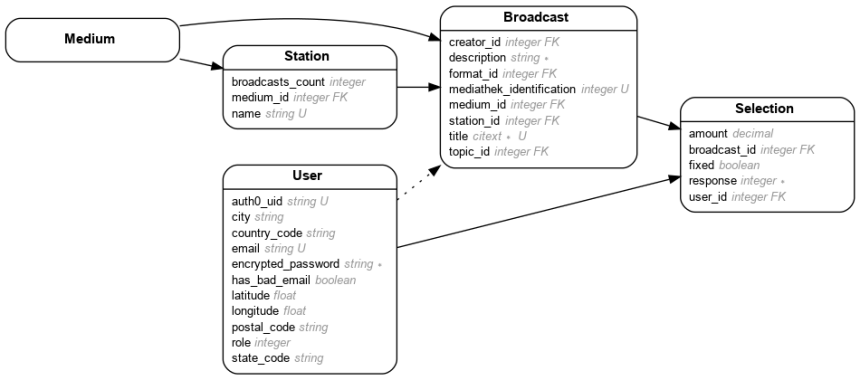


Figure 6: Database schema of the RUNDFUNK MITBESTIMMEN app

Features We have both numerical as well as nominal features. A numerical feature could be the number of supporters from an area in Germany. A nominal feature could be a list of the most supported broadcasts from an area in Germany. Numerical and nominal features can be combined, so we could request for every region, a distribution of the desired expenditure for radio, TV, online and other broadcasts.

Regions RUNDFUNK MITBESTIMMEN stores the geometry for each region in `geojson` files. This file holds a `FeatureCollection`. Every `Feature` is a region, the identifier is stored as a property. We merge the geometry data with features for every request. To be precise: We get all the user data, group it by the identifier `state_code` and merge it with the geometry in the `geojson`.

Characteristics The data base schema is a result of the specific requirements of a persistency layer. Changes in the source code may require a migration of the data base schema.

However, we can ask a lot of questions already with common data base queries or standard data analysis tools:

1. How does the actual support of a broadcast compare to the average support of a broadcast?
2. What are the most popular broadcasts in Berlin?
3. What is the desired ratio of genres of supported broadcasts? How important is education compared to sport?
4. How does the support of a broadcast change over time?
5. According to the user ranking, which broadcasts are similar to each other?

4.2 Implementation

An example of aggregated user data merged with geometry data can be seen in listing 1.


```
{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "geometry": {
        "type": "MultiPolygon",
        "coordinates": []
      },
      "properties": {
        "NAME_1": "Baden-Württemberg",
        "state_code": "BW",
        "user_count_total": "34",
        "user_count_normalized": "0.10149253731343283"
      },
      "id": 0
    },
    {
      "type": "Feature",
      "geometry": {
        "type": "Polygon",
        "coordinates": []
      },
      "properties": {
        "NAME_1": "Bayern",
        "state_code": "BY",
        "user_count_total": "36",
        "user_count_normalized": "0.10746268656716418"
      },
      "id": 1
    }
  ]
```

```
33 }
```

Listing 1: Geojson example

We can use this data as input for our common VISUAL ANALYTICS PLATFORM, e.g. figure 7 shows the user distribution of RUNDFUNK MITBESTIMMEN.

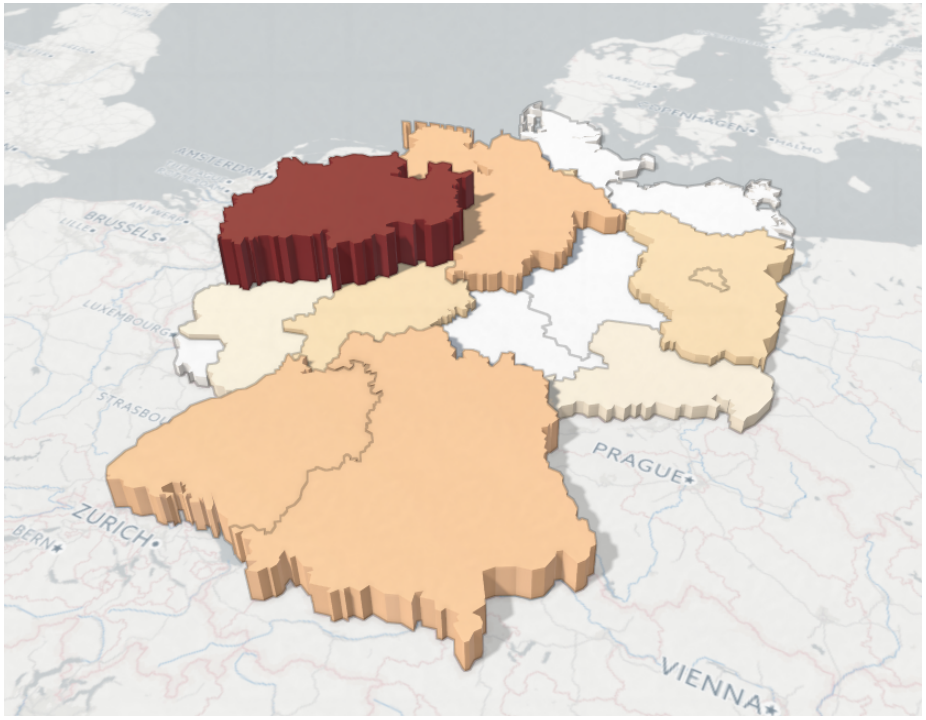


Figure 7: User distribution of RUNDFUNK MITBESTIMMEN across German federal states

References

[1] Thomas Bladh, David A. Carr, and Jeremiah Scholl. "Extending Tree-Maps to Three Dimensions: A Comparative Study". In: *Computer Human Interaction: 6th Asia Pacific Conference, APCHI 2004, Rotorua, New Zealand, June 29-July 2, 2004. Proceedings*. Ed. by Masood Masoodian, Steve Jones, and Bill Rogers. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 50–59. ISBN: 978-3-540-27795-8. DOI: 10.1007/978-3-540-27795-8_6. URL: http://dx.doi.org/10.1007/978-3-540-27795-8_6.

[2] Mike Bostock. *Choropleth*. July 9, 2017. URL: https://bl.ocks.org/mbostock/4060606 (visited on 07/10/2017).

[3] Mike Bostock. *Crossfilter*. July 2017. URL: http://square.github.io/crossfilter/ (visited on 07/10/2017).

[4] Jürgen Döllner. *Visualization Techniques for Big Data*. July 2017. URL: https://hpi.de/doellner/masterarbeiten/visual-software-analytics.html (visited on 07/11/2017).

[5] Stephen Few. "Data visualization for human perception". In: *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* (2013).

[6] Macro connection group. *The Observatory of Economic Complexity*. July 2017. URL: http://atlas.media.mit.edu/en/profile/country/deu/ (visited on 07/03/2017).

[7] Brian Johnson and Ben Shneiderman. "Tree-Maps: A Space-filling Approach to the Visualization of Hierarchical Information Structures". In: *Proceedings of the 2Nd Conference on Visualization '91*. VIS '91. San Diego, California: IEEE Computer Society Press, 1991, pp. 284–291. ISBN: 0-8186-2245-8. URL: http://dl.acm.org/citation.cfm?id=949607.949654.

[8] Sam Kusinitz. *12 Reasons to Integrate Visual Content Into Your Marketing Campaigns*. July 18, 2014. URL: https://blog.hubspot.com/marketing/visual-content-marketing-infographic (visited on 07/08/2017).

[9] Nada Lavrač et al. "Data mining and visualization for decision support and modeling of public health-care resources". In: *Journal of Biomedical Informatics* 40.4 (2007). Public Health Informatics, pp. 438–447. ISSN: 1532-0464. DOI: http://dx.doi.org/10.1016/j.jbi.2006.10.003. URL: http://www.sciencedirect.com/science/article/pii/S153204640600116X.

[10] Daniel Limberger et al. "Dynamic 2.5D Treemaps Using Declarative 3D on the Web". In: *Proceedings of the 21st International Conference on Web3D Technology*. Web3D '16. Anaheim, California: ACM, 2016, pp. 33–36. ISBN: 978-1-4503-4428-9. DOI: 10.1145/2945292.2945313. URL: http://doi.acm.org/10.1145/2945292.2945313.

[11] Richard E. Mayer. "Multimedia learning: Are we asking the right questions?" In: *Educational Psychologist* 32.1 (1997), pp. 1–19. DOI: 10.1207/s15326985ep3201_1. eprint: http://dx.doi.org/10.1207/s15326985ep3201_1. URL: http://dx.doi.org/10.1207/s15326985ep3201_1.

[12] Andrew McAfee and Erik Brynjolfsson. *Big Data: The Management Revolution*. Oct. 2012. URL: https://hbr.org/2012/10/big-data-the-management-revolution (visited on 07/08/2017).

[13] Thiago Poleto, Victor Diogho Heuer de Carvalho, and Ana Paula Cabral Seixas Costa. "The Roles of Big Data in the Decision-Support Process: An Empirical Investigation". In: *Decision Support Systems V – Big Data Analytics for Decision Making: First International Conference, ICDSST 2015, Belgrade, Serbia, May 27-29, 2015, Proceedings*. Ed. by Boris Delibašić et al. Cham: Springer International Publishing, 2015, pp. 10–21. ISBN: 978-3-319-18533-0. DOI: 10.1007/978-3-319-18533-0_2. URL: http://dx.doi.org/10.1007/978-3-319-18533-0_2.

[14] J. C. Roberts. "State of the Art: Coordinated Multiple Views in Exploratory Visualization". In: *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*. July 2007, pp. 61–71. DOI: 10.1109/CMV.2007.20.

[15] Ben Shneiderman. "Tree Visualization with Tree-maps: 2-d Space-filling Approach". In: *ACM Trans. Graph.* 11.1 (Jan. 1992), pp. 92–99. ISSN: 0730-0301. DOI: 10.1145/102377.115768. URL: http://doi.acm.org/10.1145/102377.115768.

[16] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. "Guidelines for Using Multiple Views in Information Visualization". In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '00. Palermo, Italy: ACM, 2000, pp. 110–119. ISBN: 1-58113-252-2. DOI: 10.1145/345513.345271. URL: http://doi.acm.org/10.1145/345513.345271.