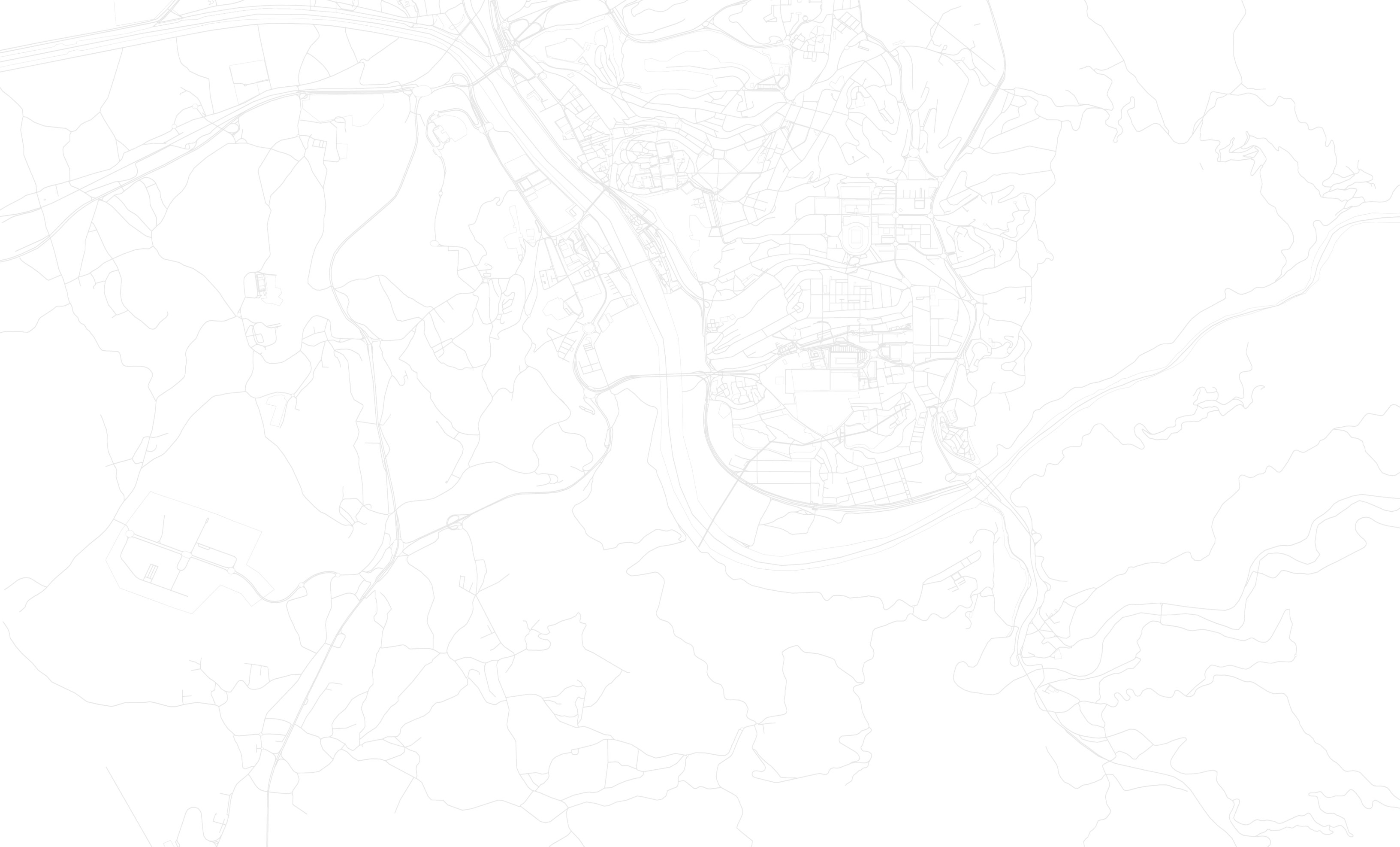
Visual Tools for the Study of Urban Mobility

DESIGN STRATEGIES FOR VISUALIZATION AND ANALYSIS OF RAW AND SEMANTIC DATA TO UNDERSTAND URBAN MOBILITY

Masters dissertation in Design and Multimédia
Faculty of Sciences and Technology
University of Coimbra

Evgheni Polisciuc
evgheni@student.dei.uc.pt

Supervisors:
Ana Alves // Artur Rebelo // Penousal Machado



Visual Tools for the Study of Urban Mobility

DESIGN STRATEGIES FOR VISUALIZATION AND ANALYSIS OF RAW AND SEMANTIC DATA TO UNDERSTAND URBAN MOBILITY

Masters dissertation in Design and Multimédia
Faculty of Sciences and Technology
University of Coimbra

Evgheni Polisciuc
evgheni@student.dei.uc.pt

Supervisors:
Ana Alves // Artur Rebelo // Penousal Machado

Abstract

In the modern world, the use of mobile devices and web services is rapidly becoming part of our everyday life. By using most of services on smart devices, a user leaves *digital footprints*—a precise data in terms of spatial and temporal location. Semantic information about places helps understand the reason why person has visited that place. Collection and further analysis of information of *when*, *where* and *why* a person has performed activities can reveal patterns in land use and urban mobility.

This dissertation explores diverse visualization techniques suited for representing time-spatial information, using static and interactive applications. Through the use of maps along with the narrative the user may reveal knowledge from a set of raw data. For that reason the dissertation starts by looking at theory and existing visual methods related to the representation of spatial and temporal data, as well as visualization of meanings of places. Then it describes principles and methods of the design process.

The first part of the dissertation presents visual artifacts that help understand urban mobility, through the visualization of the data related to the use of public transport. The developed visualization model, along with the map, reveals patterns of abnormally high use of buses in urban area. Visual exaggeration, which caused by the use of the blob technique, makes it efficient in general views, while in zoomed views the representation transforms in a rigorous visualization.

The second part of the dissertation covers visualization methods to represent semantic information. It starts with the exploration of visual methods that represent a organic shape of clusters of POIs. Then it describes the different approaches to use of typography to represent POI and category names. The combination of presented techniques, plotted on the map, serves for displaying the city from the perspective of the user. The final visualization helps to understand how urban area is used by its residents.

Keywords

Information visualization, Information aesthetic, Interactivity, Urban mobility, Land use, Semantic enrichment.

Acknowledgements

Aos pais, por me terem apoiado quando foi tão necessário.

À irmã, pelas longas conversas sobre tudo e mais alguma coisa.

Aos orientadores, pelos mapas, pela paciência e pela partilha da experiência.

Aos membros do laboratório, pela possibilidade da realização da dissertação e por terem aberto novos caminhos na vida profissional.

Aos amigos de outros países, pelas pausas de café.

Aos colegas do curso, pelas palavras certas.

Aos amigos, pelas entrelinhas e almoços.

À minha esposa, por ser a razão de eu acordar nos dias difíceis.

Obrigado.

Посвящается моей семье, Андрею, Елене, Алине и Олесе.

Contents

Chapter 1
Introduction

For centuries analysis of statistic information has relayed on information visualization techniques. All sorts of design strategies in information design evolved during last 500 years and have been well documented. However, these techniques fail when complex dynamic information is considered. Since such research areas as urban mobility and, very late, semantic enrichment of places have only been recently affirmed, this dissertation focuses mainly on developing appropriate visualization techniques, in order to help the study of land use and urban mobility.

The traditional way of land use analysis relies on census surveys. This methodology has limitation in terms of spatial and temporal scale. However, with the advent and the wide deployment of pervasive computing devices (e.g. cell phones, GPS devices, smart cards and digital cameras) this limitation can be reduced. Collecting and analyzing information of how people use urban space can be done dynamically and in more precise way. Understanding the urban dynamics are one of the central pillar of urban planning and managing.

By using most of services on smart devices, a user leaves digital footprints. This is a precise data in terms of spatial location (***where***) and temporal location (***when***), and in general can be captured without human intervention. The information about human activity (***what***) if not explicitly introduced by human may be inferred by other ways. One of which is about to retrieve the information about visited place. This place, denominated Point of Interest (POI), offers a range of services and has special utility. Such information is not always available hence it is necessary to ***Enrich Semantically*** the information about place where the person was, in order to understand what he was doing there. Collecting information of how crowds uses urban space become a very important task on creation of the image of the city from the perspective of a user, since places are often associated by meaning (e.g. the user's relationship with place or its physical properties).

Most of smart devices integrate contextual processing. However, it is difficult to enable context-awareness without semantic information. While semantic information has been available for years, the Internet, in most cases, abandons such information. In a recent work were presented various perspectives on semantic enrichment of place and extraction of such information from the Internet. Furthermore, there are studies about machine learning techniques to automatically classify POIs within a standard taxonomy. POIs should be classified within the same standard in order to make land use analysis reliable.

The urban dynamics can also be studied through the analysis of mobility patterns. Perhaps analysing ticketing information will show how crowd uses urban transportation system. This data consist in geospatial and temporal information, in other words where and when the user catches or leaves a bus/metro. From this point we will call this data as raw data (e.g. counts of passengers, standard deviations of norm, timestamps, etc.), although these data have been roughly processed but in comparison with the high-level semantic data we will assume it as raw data. Through the visualization of urban mobility we can find patterns in everyday usage of buses or metro and consequently some deviations in these patterns can be detected becouse the routine in people's live is not constant. There is always confounding factors that break systematic usage of public transport. One of these factors can be an event occurring in the city. Crossing event information with raw data will allow us to understand the reason of these deviations.

Motivation 1.1

Since the area of urban planning from semantic perspective is a young field, there is small number of suitable information visualization techniques. In cartography there are a lot of well-documented methodologies to visualize geospatial data. However, due to the subjective nature and complexity of semantic information these techniques fail. Yet, in order to understand time-spatial patterns of people's behavior, there should be custom visualization methods suited for representing such information.

Due to the change in information visualization aesthetic, large amount of data and its dynamic nature, there is a need of creation of aesthetically appealing visual artifacts that conveys only the essence of complex data, and visualization that are capable to adopt to changes in data. From point of view of contemporary design, it remains only those graphic elements that clearly convey message. All the "pyrotechnics", defined as *chartjunk* by Edward Tufte [7], used in information visualization nowadays are considered as a bad design practice. Therefore, from these observations we proposed a research to develop methodologies and visual artifacts that will clearly show urban dynamic through the visualization of semantic and raw data.

Placement 1.2

Present dissertation was placed in the research project designated as CROWDS*. The aim of this project basically is to understand urban dynamics through analysis of semantically enriched points of interest and events. The project is developed under the *AmILab* laboratory, a research laboratory dedicated mainly to Ambient Intelligence, Pervasive Computing and Ubiquitous Computing. The *AmILab* is part of Center of Information and Systems of University of Coimbra, in particularly is part of Cognitive Media System group.

***Supported by the Fundação para a Ciência e a Tecnologia of the Ministério da Educação e Ciência and Fundo Social Europeu, PTDC/EIA- EIA/115014/2009.**

1.3 Scope

Since this work is a part of one research project, which comprises several members, it is important to identify its boundaries. The dissertation includes:

1. Investigate a theory in visualization of information field.
2. Present existing visualization techniques related to the representation of temporal and geospatial data.
3. Analysis of given information and development of a set of suited visual artifacts and methodologies, which will facilitate the analysis of urban space use, more precisely development of visualizations for urban mobility and land use study.
4. Further analysis and validation of developed techniques.
5. Definition of necessary conditions for implementation and establishment of design requirements.
6. Implementation of graphical user interface that will facilitate exploration of data sets.

All the data treatment (e.g. acquiring, parsing, mining and filtering of data) is not a part of this dissertation. Yet, all the visualization techniques that are not suited to visualize given information (e.g. not related to time and space data) will not be covered in this dissertation.

1.4 Questions & Goals

— *What are design requirements and visualization techniques needed to visualize urban dynamics?* Information visualization is a subjective discipline however there is a number of variables and requirements that must be defined in order to have consistency of further work. Yet, there are a lot of existing visualization techniques and theories that should be used in order to develop suited visual artifacts.

In order to answer this question, the following goals must be met:

1. Research the state of the art in information visualization and cartography fields.
2. Analyze data and develop prototypes that meet the structure of given data.
3. Develop visual artifacts to visualize land use and public transport use.
4. Refine and validate developed methodologies on another case of study.
5. Implement a graphical user interface in order to explore the data.

Dissertation Summary 1.5

Chapter 2, state of the art, is divided in two sections: theory and visualization techniques. The first one covers a basic theory about analyzing information, graphical system properties and cognitive functions while perceiving graphics, as well as methodology of design process. The second section contains an introduction to cartography and a review of existing visualization techniques related to the temporal and geospatial information. Finally, basic interaction techniques are discussed.

In chapter 3, we will study two different approaches to visualize urban mobility and land use related to this thesis work. The first describes a project with developed techniques to visualize relationship between cellphones-time use and events/POIs. Second presents techniques to visualize the use of urban space by its residents through the visualization of data from web sources.

Chapter 4, urban mobility, covers whole process of creation of visualizations that represent raw data. In the beginning the data, tools, implementation and design requirements are presented. Then, the basic visualization is described and the series of improvements of efficiency of graphics are detailed. Finally, we will describe additional graphic and application of developed model. All the presented work is discussed at the end of chapter.

Chapter 5, land use, is divided in two sections: visualization of clusters of POIs and visualization of automated classification of POIs. In the beginning of this chapter we will introduce the data, used tools and established design requirements. The first section starts with basic representation. Existing and our algorithm detailed then. Next we will present a method to smooth corners of a polygon. Further improvement in visualization and interaction finally described. The second section starts with basic representation and experiments with color and size. Small multiples are then presented. Finally we will cover approaches with uniform distribution of points and predominant error. All the presented work is discussed at the end.

Chapter 6 includes the conclusions.

“Visualizing the data is just like any other type of communication: success is defined by your audience's ability to pick up on, and be excited about, your insight.”

— Ben Fry

Chapter 2
State of The Art

This chapter is divided in two sections: theory and visualization techniques. The first one reviews the of methodologies of information analysis and graphical system properties defined by Jacques Bertin, as well as Cleveland's studies about our cognitive task in graphic perception. The second section is divided in three sub-sections: diagrams, maps and interaction. In the diagrams part we will discuss simple, yet efficient, visualization techniques, mostly related to visualization of time-based data and to storytelling. In the maps section we will study cartography and techniques for visualizing space related information. Finally we will present methods that enables data exploration by users known as zooming and panning.

2.1 Theory

Analysis and graphic representation of semantic information and row data requires consideration of existing classification schemes in the field of information visualization. The understanding of general theory can guide the development process in an efficient way, discarding unwanted repeating processes. Thought, there are a lot of proven and well documented studies in the field of information visualization, in current dissertation we will rely upon Jacques Bertin's theory of graphical symbols and modes of graphics representation. This is a comprehensive theory that could be useful in earlier stages of the project, when analyzing information and choosing graphical elements to represent that information, and in further result analysis. Yet, we will consider Cleveland's theory on cognitive tasks in perceiving graphics, since the process of encoding the information into a visual form requires the knowledge about people's perception, i.e., the way user reads the visual information. Finally, the methodology of design process proposed by Ben Fry will be analyzed and discussed.

2.1.1 Bertin's Semiologie Graphique

Jacques Bertin, a french cartographer and theorist, in a comprehensive way theorized graphic representation in the field of data visualization. It appears to be similar to how Mendeleev arranged chemical elements, Bertin did for graphics. He organized visual and perceptual elements according to the features and relations in data and graphical systems. Before his book *Semiologie Graphique* (*Semiology of Graphics*) [1], edited for the first time in 1967, visual elements were used as the best data representation each person chose. Bertin's study of graphics as a sign-system introduced new directions in their application. The data visualization process begun to have more control over information analysis and its representation. Until nowadays his study is considered as one of the main theory in the data visualization field, although the context suffered a lot of changes: the appearance of computers that perform fast calculations, which allow animations; high-resolution monitors that are capable to display accurate, vivid and colorful images; and mobile devices expanded the world of information visualization with many new possibilities of graphic representation. In order to correctly understand Bertin's theory we need to consider his assumptions: printable on white paper; visible at a glance; reading distance of a book or atlas; normal and constant lighting; readily available graphic means.

In the world of information its communication is possible due to combination of two known concepts: **content** and **container**. Content is the information to be transmitted and container are properties of the graphic system. In data visualization design or translation of data into visual representation process is necessary to separate the content from the container.

Information Analysis (content)

First of all the information must be analyzed and categorized. Bertin defined three stages in information analysis. Let us consider the following example: "*On July 10 the overall load of the road segment X is equals to 60%; on July 11 it is equals to 80%*".

In given information the **invariant** part is the road segment X. This is the central notion common in each phrase. The data that varies here is the day and the segment load amount. This two variables are considered as **components**[*] in information analysis theory. So, the information translated from the above phrase is composed of one invariant — "*the road segment X*" and two components — "*network segment load amount in percents*" and "*the date*". According to Bertin, this is the first stage of information analysis.

The second stage consists on the identification of **elements** (or categories, or classes) and the **length** of given components. The notion of elements stands for identifiable parts of a component. For example categories *bus*, *car* and *motorbike* are elements of component *vehicle*. The length is the number of elements that we are able to identify. The component *vehicle* has the length of three.

The third step in information analysis is the identification of the relationship between components or elements, which defines organization levels. Correct identification of **organization level** of a given component a posteriori reflects on the effectiveness of the visualization model. There are three levels of organization (indeed there are four levels, but two of them belong to the same level).

The first one is the **qualitative** level or nominal—concepts of simple differentiation, i.e. this is similar to that (*similarity*) defines **associative** level of organization or i.e. this is different from that (*differentiation*) defines **selective** level of organization. For the correct transcription to graphical representation these two levels are considered as separated notions.

***In graphic sign systems components are also called as visual variables or variables. We will talk about them in following section.**

The **ordered** level defines another organizational level. Components that belong to this level contain elements that have single and universal order—a temporal order (e.g. age, generation, geologic era); an order of sensory discrimination (e.g. black-grey-white, small-medium-large).

The last level of organization is the **quantitative** or interval-ratio level, contains countable units (this is quarter or triple than that). A series of numbers could be quantitative when they specify the variation in distance among the categories.

Looking with more attention at these four levels of organization we can notice the inclusiveness of them. Thus, all quantitative series can be considered as ordered; all the categories of an ordered series can be considered as differentiated; all the categories of qualitative series can be considered as similar (Fig. 1). This inclusiveness enables us to identify the perceptual approaches and to choose appropriate representation for components of equal level.

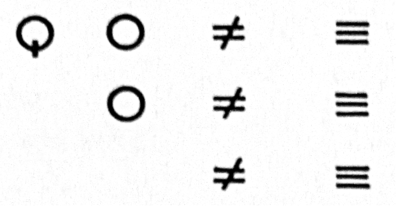


Fig. 1 Inclusiveness of levels of organization (from left to right): a quantitative series; a component whose categories arc equidistant and inscribed in a single, universally acknowledged order; a qualitative component whose categories are defined and equidistant; a qualitative component whose differential characteristic can be desegregated.

Graphic System Properties (container)

Having the information analyzed how do we graphically represent its elements? In our verbal language, in order to convey a message we encode it in universal system that comprises phonemes, letters and combination rules. Note that the phonemes and the letters by themselves are meaningless. Bertin's defines possible graphical language that can be understood by anyone, moreover nowadays due to technological advances we have more graphical elements at our disposal.

So, there are visual units that designer has at his disposal, defined by Bertin as **marks**—something that has to have the power to reflect light in order to be visible. In the context of digital media it would be the pixels of a display. Designers encode information by variation of marks. They can vary in position on paper, in how we place them (implantation) and by their visual characteristic (retinal variables). Variation of a mark in position can be expressed by **two planar dimensions**. When fixed at given point on the plane it can vary in **size**, **texture**, **value**, **color**, **orientation** and **shape**. Thus, the graphic system has eight components that are called visual variables—two planar and six retinal variables. On the plane mark can be assigned as point, line or area. Berin termed these types of significations as **implantations** (*classes of representation*).

The plane and classes of representation

In graphic design theory there are three primitive figures of plane geometry—the point, the line and the area*. These graphic elements are the "*building blocks*" for creating images, diagrams, typography, patterns, among many others complex designs. In information visualization theory implantations define classes of representation, as they have different visual characteristics. Bertin defines different features when distinguishing classes of representation: the length of retinal variables varies with the selected class of representation; the representation of quantities varies according to the selected implantation; difference in classes of representation are selective; in a single image, the same concept cannot be represented by different classes of representation.

Point is the first class of representation. Represents a position in space or on a plane. Geometrically speaking, a point is a pair of x and y coordinates that has no theoretical length or area. It has no mass at all. However, graphically it takes a form of a dot to be visible. Marks that are assigned to a point can vary in *all visual variables*.

Line is the following element. Basically, a line can be described by an infinite series of points. A line is the connection between two points and theoretically has length and position but no area. If a line appears at the edges of two areas it creates boundary between them. Also lines can represent routes, connections, paths, etc. Marks that indicate line can vary in *size (width), value, texture, color, orientation of its components and shape of detail.*

Area is the last implantation element. As we stated earlier areas in Bertin's theory have slightly different meaning and application than in graphic design. Areas are those graphical elements that have measurable size on the plane. They can vary in *position*, but marks applied to areas *cannot vary in size, orientation or shape* without changing their meaning, however they can vary in *value, texture and color.*

So, as we said earlier the plane is continous and is limited only by the limits of perception. According to Bertin [1], the plane offers the longest visual variable that is capable to encode the longest components of information. Since the plane has no breaks in continuity it is very difficult to distinguish perceptible meaningful parts and those which do not represent anything. When encoding information on the plane designer relies only on the conventions adopted within the *signifying space* (the part of the image that convey meaningful information). Therefore, in signifying space, lack of signs means absence of phenomena.

***In graphic design field the third graphical element is a plane and it is defined as a flat surface that has height and width. Basically, plane is the path of a moving line and when delimited by closed lines it becomes a shape or "a bounded plane". A plane can be opaque or transparent, textured or not, solid or perforated, etc. [2]. According to Bertin, plane is basically, a flat continuous surface that serves as holder for graphic elements and its divisibility is only limited by the threshold of perception and areas are something on the plane that have a measurable size. Thus, a mark assigned to an area transcribed to the definition of Ellen Lupton is a plane. For convenience we will use the term of "area" and not "the plane" in the context of information visualization. Further in this chapter we will see how Bertin defines the plane in more detail.**

In many cases absense of phenomena can be confused with missing data, as in a signifying space any sign means something. Consequentially, information must be applied to the entire area of signifying space. The convention that we rely on is the invariable of information and if changes it transforms the meaningful structure of the distribution. Bertin also mentions the "*inset*" that is basically a supplementary image that delimits a part of meaningful area simplifying and clarifying representation of that part. He also stated that the frame delimits signifying space, but not necessary the phenomenon. It means that whenever we delimit a part of meaningful area there is a presumption of an extension of a phenomenon.

In information visualization theory the utilization of two planar dimensions is called "*imposition*", according to Bertin [1]. The way how designer establishes the correspondence on the plane defines different groups of graphic representation. There are four groups predefined by Bertin: diagrams, networks, maps and symbols.

In the first group, the construction is a **diagram** when the correspondence is established between all elements of one component and all elements of another component. For example, a line graph which represents the trends of *stock X* over the time. Here we have one component of price and another of date. The correspondence expressed on the plane results in a line graph and we call this kind of construction a diagram.

In the second group of representation, **networks**, the correlation is created among all the divisions of the same component. For example, a circular graph which represents the connection between friends. So, here we have a component of different friends and an invariant is the relationship between different individuals. The construction is following: we uniformly place points on a circumference which represents each person; we draw lines from each individual to every individual he is friend of.

The third group is maps. The representation is called **map** when the correspondence of variables is established among all elements of the same component and arranged according to geographic order. In other words, this kind of representation is a geographic map. For example the representation of roads on the plane is a map.

The fourth group is **symbols**. In this type of construction the correspondence is created not between individual elements of a component, but between a single element on the plane and the user. This group of graphic representation involves semiotic studies of symbolism and since it is out of scope of this dissertation we will not consider this topic.

Diagrams and networks may have various impositions, can be expressed on the plane in different ways. This defines the types of impositions. Components can be inscribed on the plane according to their arrangement or according to their construction: **rectilinear**, **circular**, **orthogonal** and **polar** (Fig. 2). Despite Bertin have defined these groups and types of imposition as universal, nowadays there are more of them. In the section of visualization technique of current chapter we will discuss more on those that are in the scope of current dissertation.

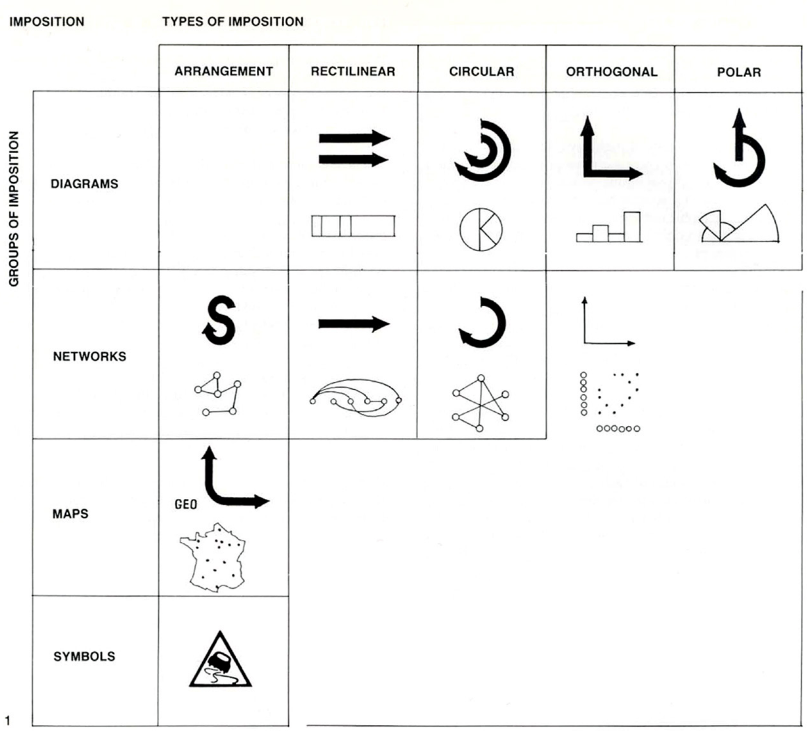


Fig. 2 Types and groups of imposition.

Visual variables and the levels of organization

As mentioned earlier in this chapter there are eight visual variables: position *x* and *y*, *size*, *texture*, *value*, *color*, *orientation* and *shape*. Position is the planar variable and the other six are retinal variables. With the introduction of third component or second component in cartography we must resort to the retinal variables. In this section we will describe modes of how to *encode/represent* information using these visual variables by establishing the relationship with levels of organization of information.

Starting with the position, we can say that position it is the variation in space along *x* and *y* axis. Position possesses all four levels of organization: *selective*, *associative*, *quantitative* and *order*. By position of any implantation element we can encode any quantitative information. Taking in account inclusiveness of levels of organization, we can say that other types of data can be encoded in the similar way. For example, the information relative to the born dates of great artists, we could encode this information by placing dots along *x* axis, where the position would represent dates. Thus, we can say that this variable is quantitative, as it is countable and it specifies the variation in distance, also we can say that it is ordered, as we can say if an artist is older than other, and for nominal level, we can say that it is similar to or different from other in terms of born date. So the plane is at once selective, associative, ordered and quantitative. As a rule in cartography, this visual variable is reserved to encode geo information.

Speaking of retinal variables we can say that *any* of them can be used to represent any component. However, each variable is not suited to every component. In order to solve this problem Bertin introduces the notion of level of organization. Relying on perceptual characteristics of each retinal variable, he defines various levels of organization for each of them.

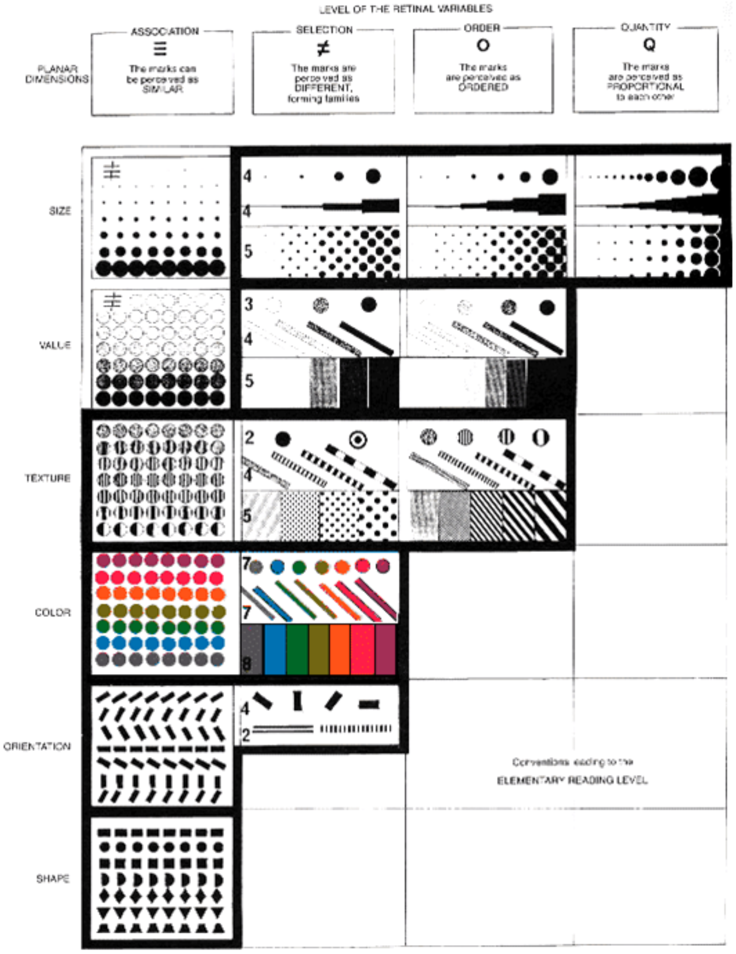
Associative perception is useful when designer is looking for equalization of variation and group correspondences within the same group. So, from Bertin's observation he defines *shape*, *orientation*, *color* and *texture* as *associative*, whether *value* and *size* as *dissociative*. Basically dissociative variables are those which create differentiation in uniformity of the area.

Selective perception is useful to identify elements of similar category forming so called "*families*". For all three implantation *shape* is not the best choice for selective perception, nor is orientation when represented by area.

Ordered perception is utilized when comparing several orders. For example gray perceived as intermediate between white and black. It is obvious that shapes, colors and orientations are not ordered. What is immediately perceived as ordered are *texture*, *value* and *size*.

Quantitative perception is useful when one seek to define numerically the ratio between two signs. So, only variations in *size* can correctly represent quantitative values. Another perceptual characteristic of visual variables is the length. Length defines across how many changes the distinctions are still perceptible. All above declarations of levels of organization are summarized in Fig. 3.

Fig. 3 Levels of organization of retinal variables. Associative—utilization of all variables (size and value are dissociative); Selective—best represented by size, value, texture, color and orientation (except orientation when represented by area); Ordered—size, value and texture; Quantitative—size only.


LEVEL OF THE RETINAL VARIABLES
ASSOCIATION
≡
The marks can be perceived as SIMILAR
SELECTION
≠
The marks are perceived as DIFFERENT, forming families
ORDER
O
The marks are perceived as ORDERED
QUANTITY
Q
The marks are perceived as PROPORTIONAL to each other
PLANAR DIMENSIONS
SIZE
VALUE
TEXTURE
COLOR
ORIENTATION
SHAPE
4
4
5
3
4
5
2
4
5
7
7
8
4
2
Conventions leading to the
ELEMENTARY READING LEVEL

2.1.2 On Graphical Perception

One of the main goals of data visualization is to convey the information efficiently through graphical representation in order to answer questions about given data. In early 20s the Bauhaus school performed studies of forms in terms of basic geometric elements based on their beliefs that this language would be universal. Today, for example software designers, uses this idea in order to organize visual elements in convenient way for the user [2]. So, what about the information visualization field? How to convey information in efficient and universal way? The graph construction, the choice of graphic elements that represent the data must be based the knowledge about how people perceive graphics. In the previous subchapter we talked about information analysis and properties of graphical system. Correct information encoding or graph construction is crucial in further visual decoding or graphical perception. "*...a graph is a failure if the visual decoding fails.*" [3]. William S. Cleveland did a rigorous study about our cognitive tasks in graphical perception, about how do people decode visual information. The information encoding process must be based on that kind of knowledge. In this subchapter we will talk about the model for analyzing the information and graphical representation developed by Cleveland.

The Cleveland's model

In the very first stage Cleveland makes taxonomic definition of information. There could be either **quantitative**, either **categorical** values. Quantitative values are usually numerical values (e.g. percentages, absolute values). In other hand, information that describes different concepts could be defined as categorical (similar to Bertin's qualitative or nominal information). For example, "Country" is the categorical type of information as it could describe values like "Portugal", "Italy", etc. There are cases when the information, depending on context, could be defined either as quantitative, either as categorical. For example, time values in many applications are quantitative values (e.g. evolution of something along given time interval). In other cases information can contain categorical values (e.g. comparison of something in different years). In the second example, time information simply indicates different seasons.

Quantitative and categorical values can be described as **scale** or **physical** information. When quantitative values are in units of the data it is scale information (ex. GDP values like 12.800 euros per capita for the year 2012). For categorical information it is a name (ex. "Portugal", "Italy"). The physical information is the description of the quantitative and categorical information on a graph (ex. considering that the information about GDP is plotted on bar graph and countries are represented with different colors, so one of the physical information would be colors).

So, now when taxonomy is defined and given information is classified accordingly how do we encode it to perform in an efficient way? First we need to understand how the information is read. In other words how the information is visually decoded. Basically, the *scale* information is read by **table look-up**. Considering a simple bar graph, first we look at the peak of a bar that interests us and then we scan to the left where we can find the corresponding numeric value. Table look-ups are mainly focused on individual values, usually those that interesting to the viewer. Cleveland defines as **pattern perception** visual decoding of the *physical* information. In the case of a bar graph we visually filter and then group bars of the same color and perceive them as an unique pattern. So, the pattern perception is basically the detection and the assemblage of geometric elements to reveal patterns.

The process of table look-up as pattern perception can varies from slower sequential operations to focused attention depending on user's goal and complexity of information. However, there are three sequential visual operations that user performs for each information decoding process. For the pattern perception they are **detection**, **assembly** and **estimation**. In a first stage, detection consists in visual recognition of a geometric elements that encode physical information (ex. detection of all bars with the same color). Later, assembly detects graphical elements that are visually grouped into one pattern. Finally, estimation operation consists in evaluating relative magnitudes of two or more quantitative values. The process of estimation is divided in three progressive levels:

1. **Discrimination**—visual determination if *a* equals or not to *b*.
2. **Ranking**—visual comparison of values if *a* is bigger or smaller than *b*.
3. **Ratioing**—the evaluation of the ratio between *a* and *b*.

The effectiveness and efficiency of described operations is determined by speed and accuracy of perception.

In the Cleveland's model the visual decoding by table look-up consists in scanning, interpolation and matching operations. On the example of the bar graph we scan along vertical line left and right perpendicularly to fix interesting points and then interpolate them by estimating the distance between fixed points and a baseline. After that, we convert physical interpolation to an interpolation in data units by looking at the tick marks on the left — matching operation.

So, in the presented model Cleveland proposes the taxonomy of information that is useful in correct visual encoding of data and further graphical perception taking in account people's cognitive aspects. He distinguishes the information in two classes — quantitative information and categorical information. Each kind of data has its scale values and physical values. The visual decoding process of each type of data consists in three sequential steps however they proceed in different ways. In pattern perception these phases are detection, assembly and estimation. In other hand table look-up occurs by scanning, interpolation and matching. Knowledge of how we decode visual information determines what graphical elements to use and what composition is more efficient.

2.1.3 Methodology of Design Process

Information visualization involves application of methods from diverse fields of computer sciences, statistics, graphic design, etc. Meaningful solution requires insights from all these fields, even more regarding to complexity of given information. The problem identified by Ben Fry [4] is that usually all of these parts are isolated from each other. Computer scientists can learn how to visualize their data using prefabricated visualization software. There are tons of software-based tools for interacting with and representing various kinds of data, however they undervalue the aesthetic principles of visual design. Graphic designers, in other hand, can learn basic insights in statistics and are capable to map the data to a visual form. However, in contrast with data mining field, graphic design have no methodologies for leading with big amounts of data. The methods themselves are not new, but their isolation impede them from being used together. Ben Fry proposes a methodology for integrating all of these parts in single information design process[4]. We will use it as a guide* within this dissertation.

***In the second part of this subchapter we will discuss the universality of this method and describe our vision of Ben Fry's methodology adopted to be used by designers. In practical part of this dissertation we will use our methodology based on the Ben Fry's methodology and experience gained during the development of practical work.**

The model

Any project of understanding data begins with questions and data sets. Correct identification of questions will create a starting point to the answer. Then the followed path is established with seven steps proposed by Ben Fry. In order to correctly understand each of the stages we need to clarify the difference between data, information and knowledge. According to Aaron Marcus the data is a “simple perceptual or conceptual input” [5], i.e. it is raw input of numbers and pieces of concepts without any kind of treatment. Information is significant patterns of data resulted from application of diverse statistical, data mining and other methods over data. Knowledge is the significant patterns of relevant information resulted from analysis the information. So, in order to move up from data-to-knowledge chain we will use Ben Fry’s model of information design process.

Acquire—the very first step is the acquisition of data. This step involves collection of data from any available source (e.g. hard disc, internet, large data bases, etc.). Yet, we must concern how the user will access to the data. If he is downloading over internet we must ensure that the time required does not reflects on user experience.

Parse—second step is parsing data. This consists in changing of data’s initial structure/format to the one that is intended for the project. The importance of parsing could be undervalued. Reader can ask the question, “*Why not to start working from existing structure?*”. For example in some cases we need the data to be loaded partially, moreover when working with large amounts of information.

Filter—filtering the parsed data is the third step. This step consists in removing parts that are not interested to us. For example if the question we want to answer is “*What is Portugal’s GDP in comparison to other Europe countries?*” we obviously will exclude all countries except Europeans. This step reduces loading and computational time.

Mine—this step could be one of the most sophisticated steps, as it involves math, statistic, and data mining. Computer scientists and statisticians can tell that this is the core in data understanding, as it reveals patterns and extracts meaningful parts of data. Basically it transforms raw data in something meaningful that can be analyzed, in information. However, there are problems that can be understood through the visualization only, namely when working with geo data. In some cases patterns can be revealed only when the information is represented on the map.

Represent—fifth step consists in representation. In this step we choose a basic visual model, and here is where the information takes its first visual form. We can use representation methodologies proposed by Bertin, or simply use one of prefabricated tools/techniques (some of which we will be described later) as this step consist only in rough, but quick representation.

Refine—next step is the refinement step. In this step we improve basic representation by applying graphic design methods in order to draw user's attention to the meaningful parts of representation, to clarify the graph, to augment the readability, as well as to make the representation aesthetically appealing.

Interact—in the last step designer adds interaction, enabling the user explore the information in order to uncover knowledge. Due to continuous changing nature and complexity of the data, this step is as important as other ones. With increases in computational power and storage capacity, as well as in data collection, the data is no more static. It has values that change over the time. So the modern software-based tools must be prepared to these adjustments. Consequentially, the interaction makes it possible to track and analyze the data.

Obviously, as in any design process, this model is not static and is not linear. In some stages of the process we can return to previous steps and improve them without affecting other steps. Ben Fry explains the iteration between each stage [4]. The interaction with other stages starts in representation. For example in the first basic representation we can conclude that the data must be filtered differently in order to answer the given question. Or if there is not enough data to answer this question we must acquire more data. In the interaction step we always can improve visualization by returning to refinement step or if after exploring the data we say that the information still need to take more treatment in order to show something, we can go back to the mining. All of the declarations above are summarized and schematized in the Fig. 4.

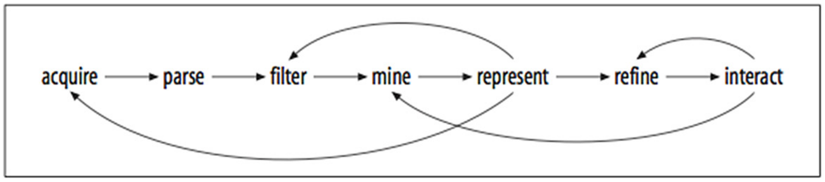


Fig. 4 The seven stages and interactions between them.

As we said before, in this part of the current subchapter we will present our adopted model for design process. We based on methodology of design process described earlier in this sub-chapter and in our experience gained along the development of our practical part* of current dissertation.

For us it was obvious that as designers we will focus on the three last steps: represent, refine and interact. As process progressed we were interacting with other members of the team in order to achieve desired results. As in any information visualization projects we started with question: "*What and when bus stops are more/less used and what is the conditional for that?*". The data set was earlier acquired, parsed, filtered and mined by other members of the team. When arrived to us it has the following structure: the first file has values of stop id, stop name, latitude and longitude; the second file contained the information about the date, time, stop id and standard deviation. Both files were in .csv format.

First step was to represent the information. We used different visualization techniques such as bubble graph or representing the information with triangles and lines (further we will explain those in more detail). However, the result was not that we desired. So, we relied on bubble graph and focused on its improvement. At this point we realized that starting right away with the representation is not the best solution, as there are solid theories on how to correctly choose graphic elements and types of construction in order to achieve efficient results. We proceeded to use Bertin's theory—starting with information analysis (definition of the content) and choosing the correct way to represent it (elaboration of the basic representation). Consequentially, this stage was divided in two sub-stages: 1. understanding given information; 2. create mapping from data to visual representation. Later we found this methodology more efficient in terms of time and visual solution.

In the next step we focused on improvement of basic visualization model. According to Ben Fry this was the refinement step. So, basically we were improving the visual aspect of bubble graph and the performance of the resulting visualization model. This step consisted in presenting the visual artifacts on the digital support (or in other cases it could be the physical support).

As in Ben Fry's model, such as in our model, the interaction step makes all the sense in any project where the goal is to explore the data and reveal any knowledge. In our particular case we added buttons that allows the user to navigate through the time. Yet, we implemented the zoom functionality, so the user can explore the data in more detail.

***The practical part of this dissertation consists in four exercises. However, we will focus only on the first project, since this one started with use of Ben Fry's model and lead us to adjust the methodology. The following three projects were guided by the same model.**

The last step in our project was to ensure the credibility of our visualization model. Using Cleveland's methodology we verified the usefulness of the representation, the way it is presented and its interaction methods. Although this step is not required when using existing visualization techniques, since the usefulness of those might have been proven earlier.

Summarizing all said above, Fig. 5, our methodology is oriented to designers who works in teams with members from other fields such as computers science, statistics, or others, and consists in:

1. Understand the structures of related information.
2. Create mapping from data to visual representation.
3. Present the improved visual representation on the digital or physical support.
4. Provide methods of interacting.
5. Verify the usefulness of the representation.

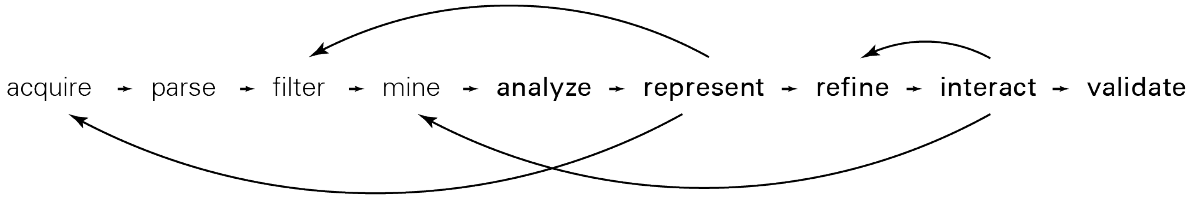


Fig. 5 Complete pipeline of methodology.

Visualization Techniques 2.2

In the field of data visualization there are tons of visualization techniques. Major part is proven and well documented and almost all of existing visualization methods is software-based, enclosed in libraries with provided API or applications with user interface, and most part is open-source. While the complete review of existing visualization techniques and application of them would worth a book, the following section will describe diagrams constructions, precisely time-series and small multiples, and maps, in particular subjective and thematic maps. Finally some basic interaction methods will be analyzed.

Diagrams 2.2.1

According to Bertin, diagram is a construction on the plane when the correspondence is established among all elements of one component and all elements of another component [1]. Statistical data in many cases are quantitative results of observation of phenomena over continuous interval (e.g. the traffic density over time in road system). In other words, when we want to represent the correspondences among two components and one of them is quantitative. Time-series graphs are well known and widely used diagrams that represent continuous quantities. The time in most cases is represented by *x-axis*, regardless where it starts counting, time is always continuous, unless it is divided in time ranges, or if represents seasons. The time varying features are represented by another variable that is the function of the first. What makes visualizations of such of data interesting is the possibility to observe curve variation and correlation between several categories over the time, as well as tell a story. In this section we will discuss about simple and multiple line charts, stacked area charts and small multiples.

Simple Line Charts

A basic representation of continuous quantities considered as a simple line chart. This type of charts projects two components—one varying feature as a function of another among *x* and *y* axis. Usually *y* is a function of *x* value. Perhaps this convention is a fruit of human's perception of horizontal and vertical planes in almost all cultures. If we think about how we write, regardless direction, the path will still be in a horizontal plane. From that observation we deduce that for us a horizontal plane means something continuous in opposite of vertical plane that would correspond to variation of a value along of *x-axis*.

Simple line charts are advantageous for basic representation of non-complex data, the data that posses less than three components. In other words simple line chart technique efficiently represent quantitative values observed over continues interval, Fig. 6. The line charts are widely used diagram. According to Edward Tufte, 75 percent of 4000 graphics published in 15 of the world's newspapers from 1974 to 1980 where line charts or *time-series* by definition of author [7] . The most significant information that can be derived from these charts is the changing in curve over time, as well as individual values. Likewise, the maximum and minimum values are obvious on such graphs.

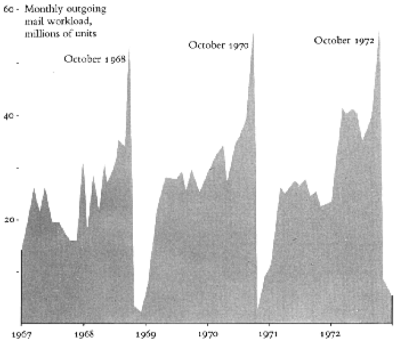


Fig. 6 "Mail, House of Representatives", (Extract from p.37 in Tufte, E. 1983. *The Visual Display of Quantitative Information*. Graphic Press, Cheshire Connecticut.). This graph describes how the Representatives of US House use their privilege in order to send free mail newsletter to promote their campaigns. We can see that the activity peak occurs every year right before the Christmas.

There are cases when we need a small line graphics that accompanying the description of a data, in order to give a context and rapid overview of past occurrences. Edward Tufte, in his *Beautiful Evidence*, introduced the sparklines—"*Intense, Simple, Word-Sized Graphics*" [6]. These small graphics, illustrated on Fig. 7, similar in construction, but different in use to the simple line graph, usually embedded in a full context of words, numbers, images. As in typography, letterforms are designed to be read in any varieties of sizes, without losing the legibility at small sizes, the sparklines can be efficient even at size of a letter [7]. Moreover, words present quick visual recognition, regarding to the overall shape and letter-by-letter detail. Same with sparklines—they present overall shape and aggregated pattern of local detail.

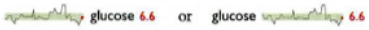


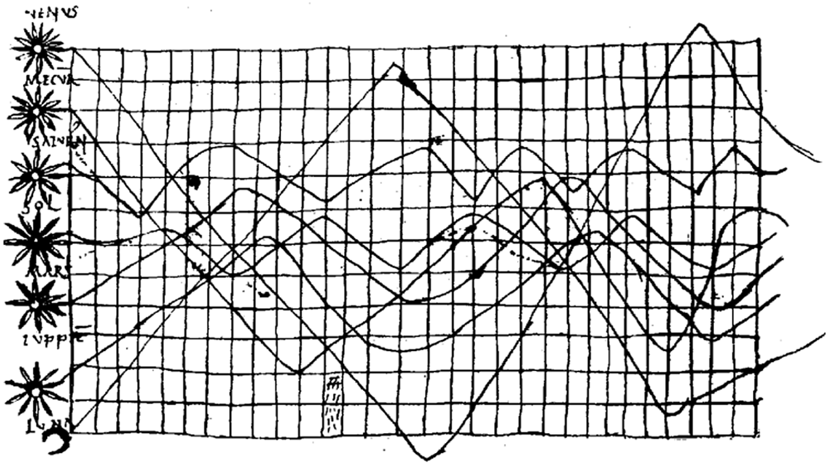
Fig. 7 The use of sparklines to add helpful context. These graphs are showing the history of patient's level of glucose (gray background is a normal range of glucose). Scan from the book "Beautiful Evidence" of Edward Tufte[6]

Consequentially, these small graphics cannot exceed the height of a text line, when presented in text. The length may vary regarding to the aspect ratio. Basically, the length should be greater than the height and, according to Cleveland[8], in time-series best is an aspect ratio that comes from average hill-slopes of 45º. Therefore, sparklines should be embedded in the text and tables accordingly to the definition above, in order to provide a helpful context for interpreting.

Multiple Line Charts

So, line chart is useful to represent the behavior of individual elements of information. However, how do we compare variations of several elements over time interval, i.e. growth of market *A* and *B*? In this case we would use multiple line charts. Basically these types of diagrams are similar to simple line chart, with the only difference that here we have each line representing one category (Fig. 8). This example is a 10th century graph that illustrates values changing and represents planetary movement over time. Here, instead of drawing one graph next to the other, author plotted all graphs on a single coordinate system, what let the user directly compare variation of several variables over time.

Fig. 8 Planetary movement, 950, (H. Gray Frunkhouser, "A Note on a Tenth Century Graph", Osiris, 1 (January 1936), pp. 260-262.). One of mysterious graphs in history, since the next similar diagram appears 800 years later. According to Frankhouser, the graph does not correspond to actual planetary path, since the astronomical content is confused.



While the planetary movements chart may be considered as time-series graph, the first correct definition of time-series line graph technique was made by William Playfair—scottish engineer, considered a father of modern graphical methods for statistical visualizations [9]. In 1786 he published his "*Commercial and Political Atlas*", which contains 44 statistical charts. In each graph Playfair uses one line to represent imports to and another to show exports from England along the time. The space between the curves is a balance of trade with England—against, if negative, painted in red color, or favor, if positive, painted in green color (Fig. 9). So, basically he used two simple line chars incorporated in single system within the same dimension (expanse in million of pounds). Comparing each other reveals the area between the two curves, which is the difference between imports and exports and to make it more obvious it is reinforced with colors.

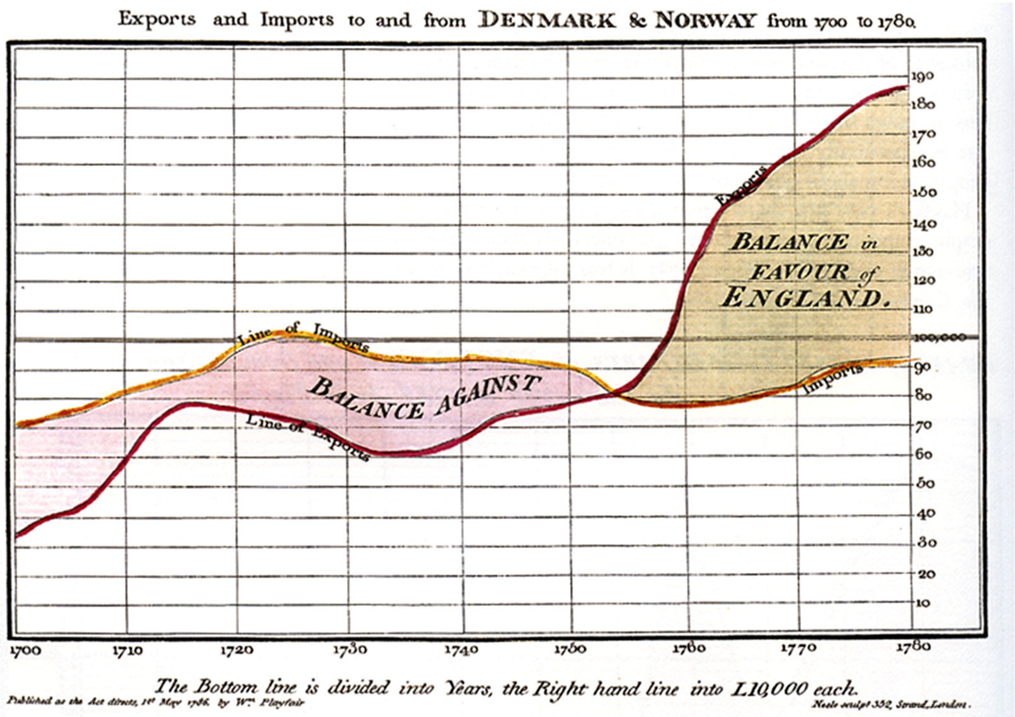


Fig. 9 Playfair's time-series line graph, 1786. One of the graphs of series "Commercial and Political Atlas" representing the trade balance between England and Denmark and Norway — the space between curves painted in red (before 1765) and in green thereafter.

However, according to Cleveland [8], this kind of representations is inaccurate for pattern perception. Fig. 10 demonstrates that the visual difference between lines decreases from left to right, however, focusing on circles we can see that the difference is constant. Thus, in situations when values are rapidly increasing or decreasing the further conclusions might be inaccurate. For the multiple line graph the solution would be to represent the difference separately using all the same line graph.

Stacked Area Charts

Stacked area charts displays cumulated totals of values, usually in percentages, over the time. This technique is useful when representing different nominal elements (or categorical) within the same component that changes over time. In other words, stacked area graph shows trends over time among related elements, where each element is represented by a line and a filled area bellow the line. The representation of following element is plotted on top of the previous one, and so on.

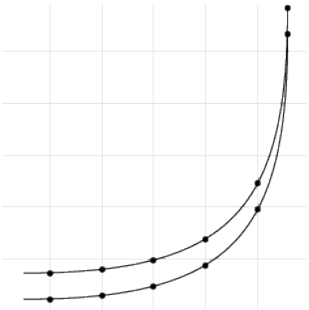


Fig. 10 Multiple line graph that represents two elements. The difference among values is constant, but the visual impression is that the difference decreases.

Consider the graphic on Fig. 11, presented by Edward Tufte[10]. This graph illustrates marketing trends in the development of pop/rock music, covering the time period from 1955 to 1978. Each "*stream*" is a time-series that tracks the length of time, along with an estimation of its share of total record sale, for popular performers or music styles. Stacking the charts of each category allow us to compare the influence of multiple artists for the same time period and trends of music style.

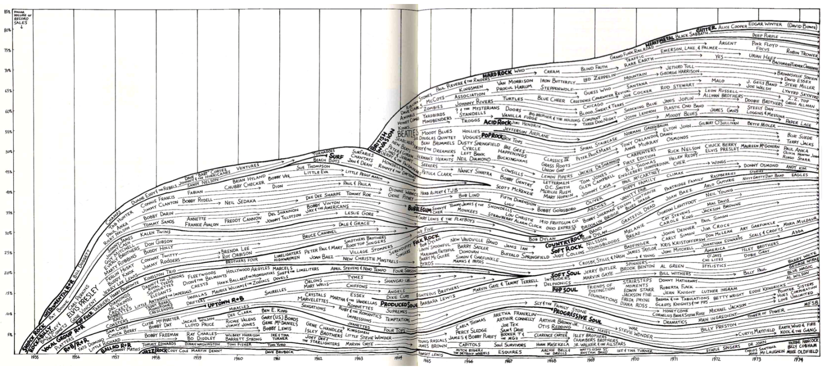


Fig. 11 This graphic is a reproduction by Edward Tufte of the originally designed by Steve Chappel and Reebe Garofalo in Rock 'N' Roll is Here to Pay: The History and Politics of the Music Industry in 1977. More than 700 artists and 30 styles of music are mapped in current graph.

More recent example of use of stacked area graph, showed onFig. 12. This interactive visualization, created by Moritz Stefaner[11], represents the use of tags in social networks and its frequency over the time. Each tag is represented by a strip, stacked on top of each other, that varies along the *x-axis* according to the frequency of appearance of the tag. So, this diagram allow us to see the frequency of using tags from different sources over the time and through the interaction we can highlight interested elements and access to more detailed information.

Stacked area graphs can fall in the same issue as multiple line graph—abrupt changes in line slope might lead to incorrect interpretation of a graph. From the above observations we can say that the correct application of such elementary techniques as multiple line chart can reveal non-evident information in data. Yet, stacked area graphs are capable to display trends in time-series and convey it in comprehensive and meaningful way.

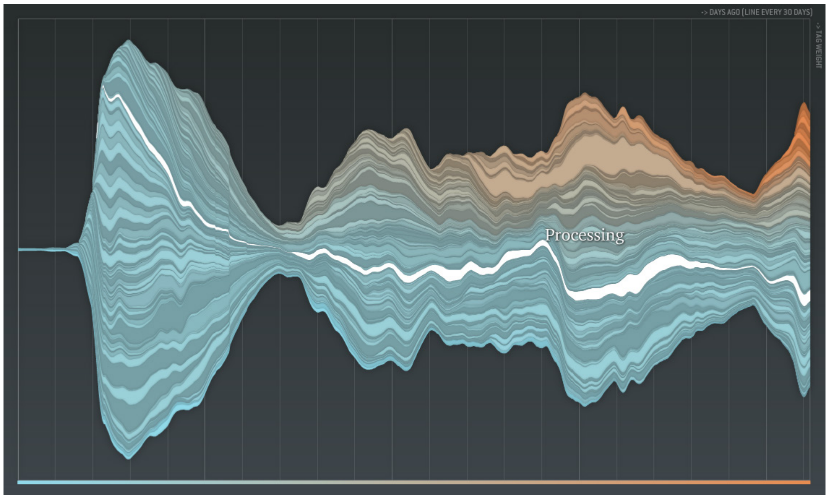


Fig. 12 This diagram, created by Moritz Stefaner in 2007, visualizes the tag structure. Using stacked area technique, he represents the use of tag in social networks and its frequency over the time. Tags are stacked from bottom to top in order of their appearance. Colder colors mean an earlier appearance and warmer colors means more recent tags

Small Multiples

“Inevitably comparative; Deftly multivariate; Shrunken, high-density graphics; Usually based on a large data matrix; drawn almost entirely with data-ink; efficient in interpretation; Often narrative in content, showing shifts in the relationship between variables as the index variable changes”

— Edward Tufte

Defined by Edward Tufte [12], small multiple enables visualization to answer the question: “*Compared to what?*”. By direct visual comparison of changes among elements of multivariate, aka complex data, the data that have three or more components. Essentially, small multiple is a series of displays with the same properties/structure of graphical system arranged in a grid within the scope of the eyespan. In other words, each graph of a small multiple series must have same graphical composition, differing only in the data they represent.

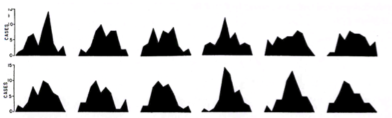


Fig. 13 This small multiple shows the effects of sampling errors. Each graph is based on a sample of 50 random normal deviates. Edmond A. Murphy, 1964 [14]

Small multiples can vary from very simple, Fig. 13, to more sophisticated representations, which makes them to expand from bounds of diagram construction. For example lets consider the visualization of taxis’ traces in Vienna presented on Fig. 15. This small multiple representation focuses on the understanding of patterns of taxis’ movement. Each graph illustrates the change of the gestalt with one hour interval, starting at midnight of 25 of July 2011 and ending at the midnight of the following day. The frames are arranged in four by six grid and the consistency is remained through all the frames, so that the user can focus only on shifts in the data.

Small multiple is one of the most efficient visualization technique for multivariate data. “for a wide range of problems in data presentation, small multiples are the best design solution.” — Tufte [12]. Mental cost of the perception is low, since the decode and comprehension process occurs only once for one of the frames. Then the consistency among all the other frames enables the user to focus only on the changes in data and not in graphical composition. For example the differences in a typeface’s wights within one superfamily can be understood by using small multiple, Fig. 14. This simple representation clearly illustrates changes among different font weights and serves as visual insight for the classification of fonts.

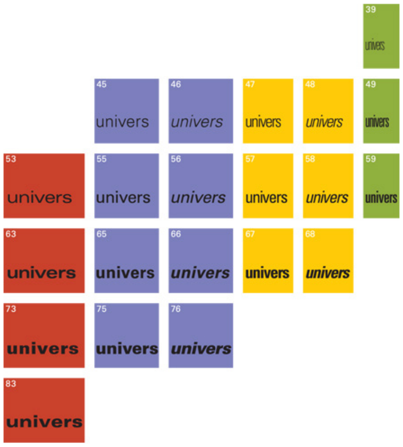


Fig. 14 Univers type family, designed by Adrian Frutiger in 1957 [15].

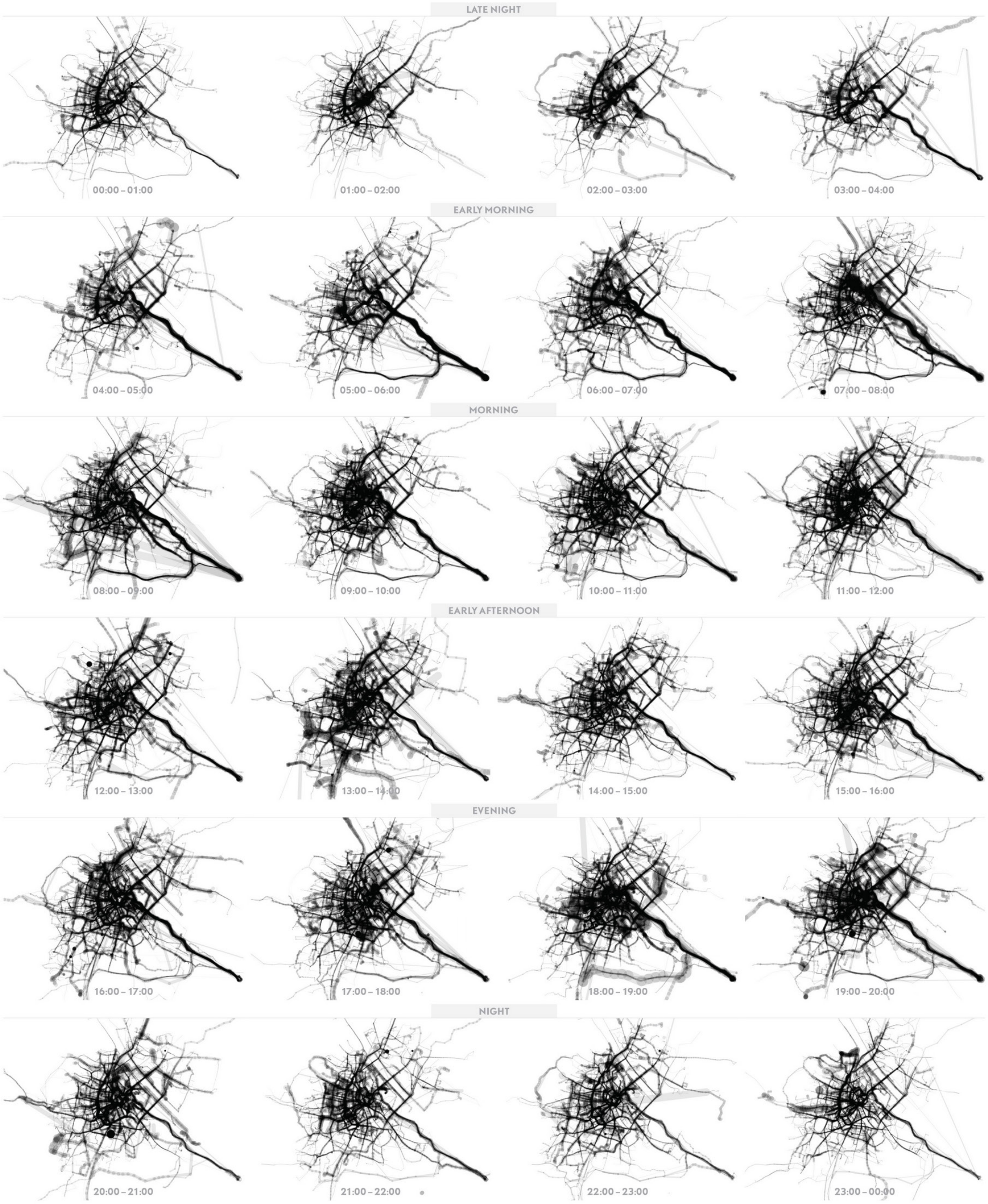


Fig. 15 Sense of Patterns — Twenty Four hours of Taxis. This small multiple shows 24 hours of taxis’ traces in Vienna, 2011 [13].

2.2.2 Maps

In this subchapter we will review a part of basic theory of cartography—geographic coordinate system and map projection. In the second part we will discuss subjectivity in cartography, precisely anamorphic, tube and typographic map, and diverse application in information visualization. In the third part, topographic and contour maps are reviewed. Finally we will talk about statistical maps and how they where used in the past and in modern statistical analysis.

Cartography

While the entire subject of cartography and mapmaking worth a book, the following section will describe the minimum of information the reader needs to understand what is a map and how it is displayed. In theory chapter we saw that the map is a visual representation of established correlation among all elements of the same component arranged in geographic order. In other words a map is a visual representation of relationship between objects, regions and other elements of earth's surface and/or geo-referenced data. Maps provide the unique ability to visually display physical location/spacial information. Sometimes the information about location is called geodetic or geographic coordinates. In maps this information is described by **latitude** and **longitude**, an absolute precise location on earth's surface. The latitude is frequently considered the *y* coordinate and the longitude the *x* coordinate and measured in **degrees**. The latitude starts with zero degrees at the equator and increases to 90 degrees at either pole. The longitude is used to measure distances east and west from prime meridian, which passes trough Greenwich, England. The longitude starts with zero degrees at prime meridian and increases east and west until opposite side of the world, which called the International Date Line, that gives 180 degrees.

Geodetic information transferred from a spherical surface to a flat plane is called a **map projection**. It might be visualized as the process of projecting each point on a globe onto a simple geometric shape such as a **cylinder**, **cone** or **plane** (aka azimuthal) by **normal**, **transversal** or **oblique** projection types [16], Fig. 16. Obviously the transformation from the three-dimensional ellipsoid to the two-dimensional plane is not possible without any forms of distortions. This distortions affects shapes, distances and direction. Each available projection methods result in different distortions and determines which map projection is suitable or not for a certain purpose. We will focus only on normal cylindrical projection, in particular on **Mercator** projection.

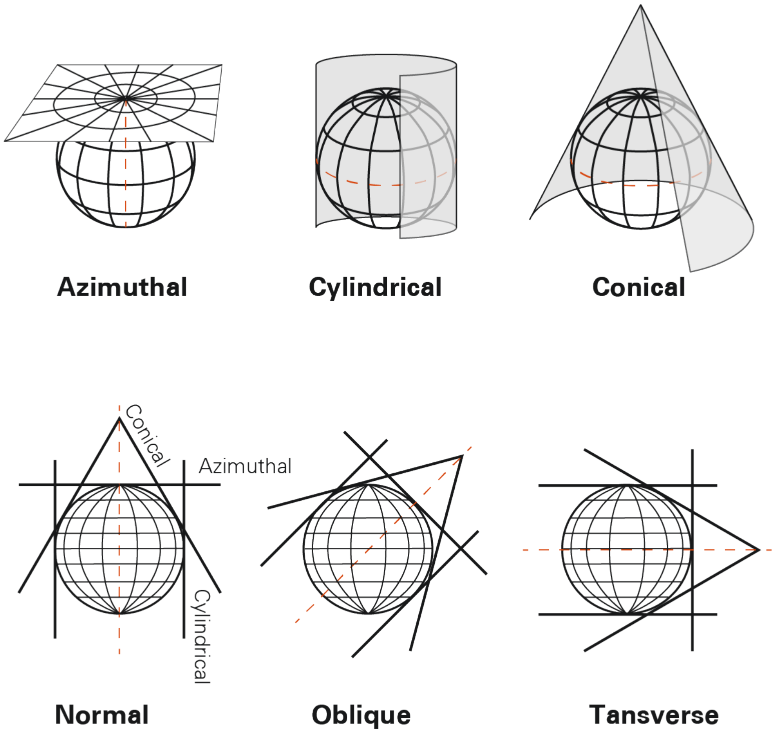


Fig. 16 Normal, transverse and oblique projection of a globe onto a cylinder, cone and plane surface [16].

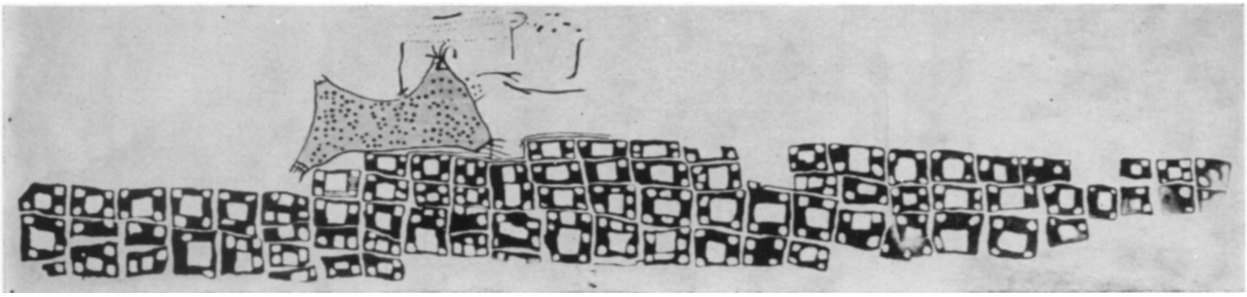
Before starting with the Mercator projection we would like to introduce few terms used in modern cartography. So, a world standard used in cartography for calculation of position, distance, etc., is called The World Geodetic System (WGS). It comprises a standard coordinate frame for the earth (e.g. data related to the shape of the ellipsoid, data on gravity of the earth, etc.). The most recent revision of this system was made in 1984 and is referred to as WGS84. In many cases WGS84 is referenced as description of ellipsoid with the major radius equals to 6378.137 km. The WGS84 is also known as EPSG:4326, which is the identifier in European Petroleum Survey Group (EPSG). There is one particular spherical mercator projection coordinate system that was popularized by two major web map services—Google maps and OpenStreetMaps referenced as EPSG:3857.

So, the Mercator projection is cylindrical map projection, which distorts the size and shape of large objects, as they move away from the equator. It is because of the projection process—meridians are mapped to equally spaced vertical lines and parallels (circles of latitude) are mapped to horizontal lines. Consequentially, each circumference is stretched to equal length and small elements located closer to the poles become bigger than they are. For example, Alaska takes as much space as Brazil, although Brazil's area is nearly five times bigger than of Alaska. Although the Mercator projection is not suited to general reference world maps due to its distortion of land area, many major online street mapping services still use this projection's variations for their map images (e.g. Google Maps, OpenStreetMaps, Bing Maps).

Subjective Maps

In early years, due to absence of accurate and objective geographic data, maps where produced based on subjective knowledge. Fig. x shows the ancient map of a settlement Catal Hoyuk in Turkey, drawn according to the knowledge of the habitants about the approximate location of their houses and vulcan located near the settlement. Therefore, **subjective maps** are defined as description of a spatial structure regarding to the users point of view, hence reflecting how the city/urban space is perceived [17].

Fig. 17 The Earliest known map, Catal Hoyuk, Turkey, 6200 BC [18].



Anamorphic maps and variations

Anamorphic maps and its variations are kind of subjective maps —a map in which non-geographic, as a rule a quantitative, variable (e.g. population, travel time, etc.) are represented with land areas or distances—geometric distortion of the map. In the example of the Fig. 18, a deformation of spatial size shows the travel time from Paris to various places in France over 200 years. So, the locations to which the travel time is short where placed closer to Paris, while further in travel time locations appear further on the map.

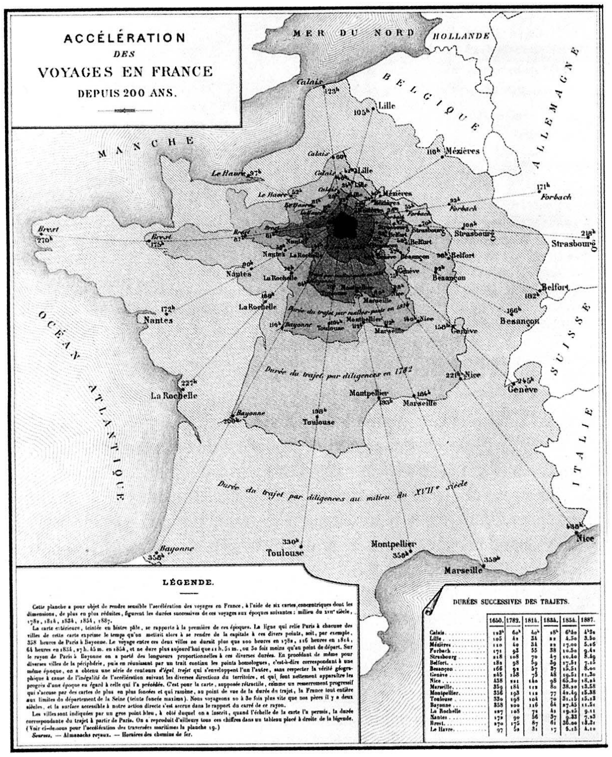


Fig. 18 Using a deformation of spacial size to show travel times from Paris to various places in France over 200 years, Émile Cheysson, France, 1888[18]

The concept of distorted map can be applied to many cases of visualization of geographic related information. The *UrbanCyclr* application from the Kitchen Budapest team, distorts the city of Budapest in order to convey the information regarding to biking patterns[19]. Here, all the distortions of the map reveals higher biking activities in the respective area at given time, Fig. 19. This visualization uses slightly different technique rather than anamorphic maps. In order to visualize geographic and time-based data, they relayed on some variations of the fisheye distortion. This led to areas with more importance appear closer to the user and in contrast the areas with less importance loose focus becoming smaller.

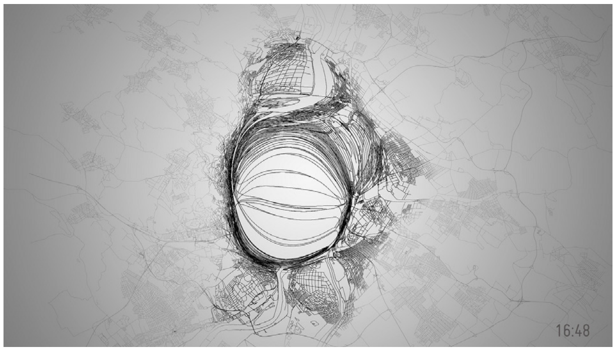


Fig. 19 Visualization of biking traffic pattern at 16:48. Application is created by Kitchen Budapest team in 2011 [19].

Schematic maps

A schematic map is particularity of schematic diagrams, which is the representation of a systems using abstract, graphic elements. In schematic maps, as a rule, these systems are road or subway networks. Since, the first schematic map of a subway, designed by Henry Beck in 1931, Fig. 20, was London's underground network, commonly known as the Tube, we will designate this design concept as **Tube Map**.

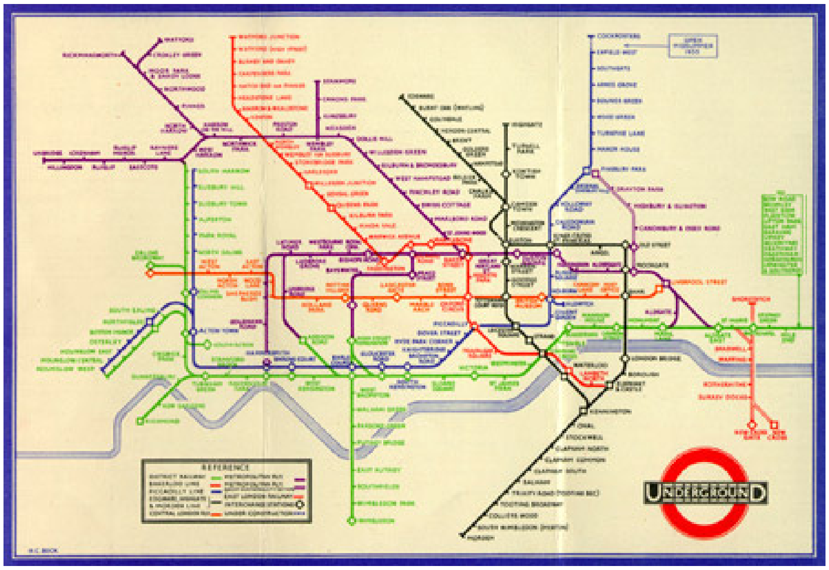


Fig. 20 Re-design of the London's underground network, using schematic representation, by Henry Beck in 1931.

The omission of all detail that not relevant to the information to be conveyed, the representation of relative position of stations and their connective relations with each other with abstract graphical elements are the main characteristics of tube maps. For example the Beck's map consist in representation of stations and straight line segments connecting them. Lines are only vertical, horizontal or on 45 degree diagonal. Stations are differentiated between ordinary stations, marked with tick marks, and interchange stations, marked with diamonds sign. The omission of cutter makes tube map easy to read, even in great distances.

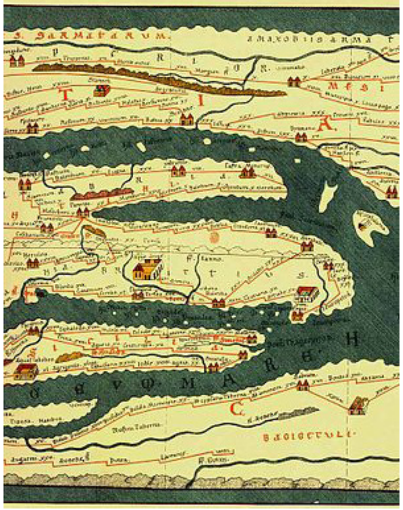


Fig. 21 Part of Tabula Peuntigeriana (Peuntinger Map), 366-335 BC [18].

A map with combination of schematic elements with realistic elements is known as semi-schematic maps. This creates the compromise between purely abstract and realistic representations. Fig. 21 shows a map with these characteristics. In this example locations, orientations and distances of roads and geographic features are represented as accurate as possible, while the map have no similarities with reality. The symbolism of these kind of maps may have stronger impact in information communication.

Let us consider the example of famous Buckminster Fuller's "*Dymaxion Map*" [20]. This is the flat map of the entire surface of the Earth which represents our planet as one island in one ocean. Yet, this map does not represent any visual obvious distortion of the relative shapes and sizes of the land area, Fig. 22. While it is the most precise representation of the earth, why it is not so popular as other maps? There are many reasons for that—regarding to political views, regarding to stereotypes, regarding to humans knowledge and culture, among any others. It is simply uncomfortable for us to orient in this kind of map. Thus, the subjective maps sometimes are more efficient than objective, since the communication must rely on how people may perceive this information.

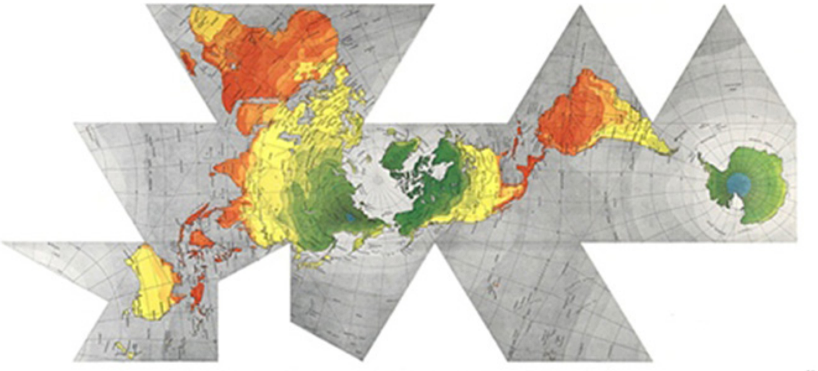


Fig. 22 Dymaxion map designed by Backminster Fuller in 1943

Typographic maps

First of all we would like to briefly introduce the concept of typography. Since the whole area of typography is a hugely complex and specialized discipline, we will introduce only essential concepts. So, basically typography is the craft of transforming human language into a visual independent forms, the craft of arranging type in order to make language visible [21].

The arrangement of type consist in following:

- Choosing a **typeface** and a **weight**.
- Choosing a point **size**.
- Adjusting space between pairs of letters, aka **kerning**.
- Adjusting spaces between groups of letters, aka **tracking**.
- Choosing a line **length**.
- Adjusting line spacing, aka **leading**;
- … among others.

All the existing typefaces are **classified** according to the system established in nineteenth century [22]. So, the classification is divided in three categories: **humanist**, **transitional** and **modern** typefaces. Then typefaces are classified as **serif** or **sans** serif. Yet, there is one more classification category called **egyptian**, aka slab serif.

Another relevant concept in typography is the organization of typefaces into **families** that include *roman*, *italic* and *bold* variations of a typeface, sometimes called as typographic weights [22]. There are modern typographic families that are composed with more than three weights (e.g. ultra light, light, regular, semibold, black, etc.). For example, as we saw earlier Univers type family possesses twenty one variations. Some families have both serif and sans serif versions.

So, typographic kind of maps may be seen as "*artistic*" representation of semantic information, rather than an accurate mapping of geographic data. A semantic information is the description of the relationship of the place and its meaning and it depends of many human, political, social, historical factors, thus, these kind of maps may be considered as the most subjective of all maps.

In typographic maps the semantic information is presented trough the use typography, hence the name. For example, the most part of maps by Paula Scher are purely typographical painted maps of the world, its continents, countries, islands, etc, Fig. 24. Like the maps of Axis Maps, Fig. 23, the information about location/space is represented by typography, moreover words takes the form of the object the represent (e.g. streets, parks, rivers, etc.).

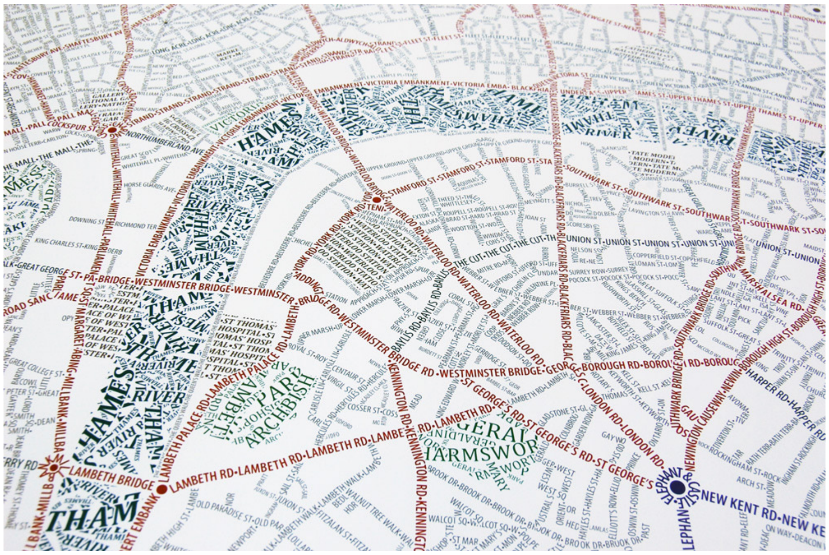


Fig. 23 A detail of manually crafted map of Boston produced by Axis Maps team. Their maps uses nothing but description of places (streets, water channels, parks, etc.) [24].

The maps on the second example were composed using software-based tools and more accurate typography, yet, the graphical elements are placed geographically precisely. Though, these maps possess information overload and present alternate versions of reality [23], a good typographical hierarchy makes them to be readable and efficient in conveying subjective and not precise information. Elements with more significant meaning appear in bigger sizes or in caps and elements that are less important appear smaller. So, we may say that in order to convey subjective information the choice of graphical elements depends not only on rigorous, almost mathematical study, but on the way it communicates best, regarding to design/artistic methods of expressing information.

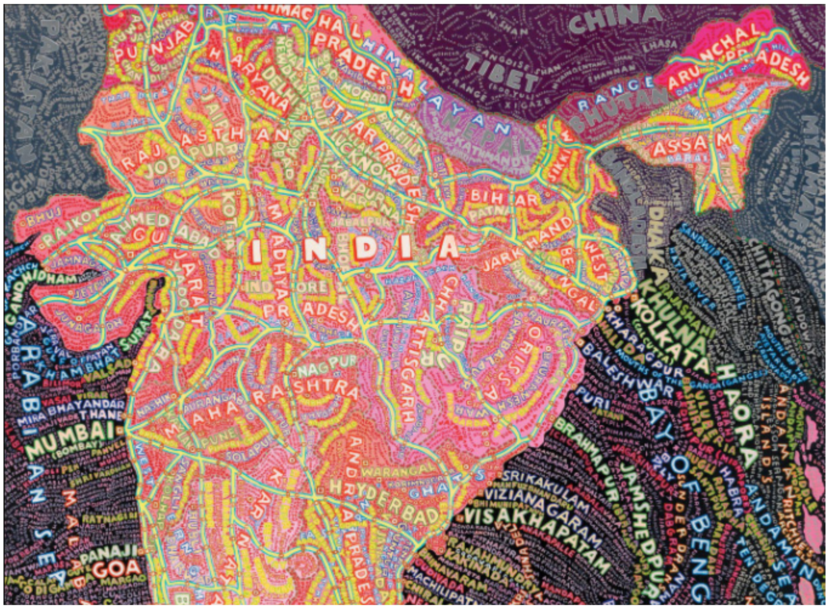


Fig. 24 A map painted by Paula Scher, which presents personal reaction against information overload and the constant stream of news of skewed reality [23].

Topographic and Contour Maps

A topographic map is a type of map that displays the shape and elevation of the earth's surface or simply relief. In modern cartography and in first representation of topographic maps, contour lines are used to represent these information, Fig. 25. These lines are a type of isoline, which are the lines of equal values. In topographic maps these lines connect points of equal elevation, points on pseudo *z-axis*. Yet, in charts and maps the isoline can be used to represent multivariate data—contour maps, where the data is not related to the description of the earth's surface, which makes them a statistical map that we will talk further in this chapter.

The technique of a contour map is used in information visualization when the first two variables are coordinates in geographic space and the third one is a non-geographic variable to be displayed on the map. On the example shown on Fig. 26. the population densities are represented using the contour lines. In this visualization each line marks the threshold of 50 population density units. The intervals of 200 units are represented with thicker line and labeled with the number, that repeats along the corresponding line.

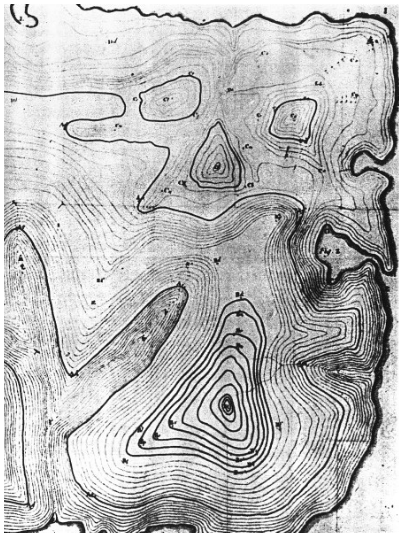


Fig. 25 This map is considered as first topographical map, produced by Marcellin du Carla-Boniface in 1782, France [18]

Statistical Maps

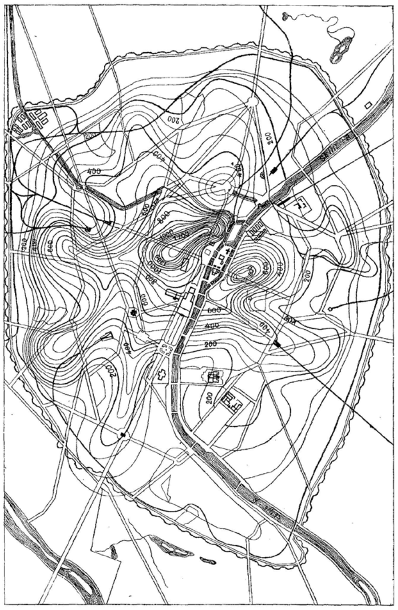


Fig. 26 Statistical use of a contour map introduced by Louis-Léger Vauthier in 1874 [18]. This map represents population densities by use of contour lines.

Sometimes referred to as a thematic or a data map [7], a statistical map displays quantitative information regarding to locations, distances, areas, etc. As a rule the base map containing geographic information onto which the information is represent is deemphasized with only enough information included to orient the viewer. Here the balance is found by using design strategies to represent the map—quantity of *meaningful information* vs *visual clutter*. In some cases the base map representation boils down to only road information using thin lines, so it is clear and visually pleasing. The graphical choice depends on the information and how we want to encode it. In the theory section we discussed diverse properties of graphical system defined by Jacque Bertin. Using his theory in combination with Cleveland's "*graphical perception*" designer can plot any kind of information over a map.

One of the first statistical maps was a chart displaying trade winds and monsoons observable in the seas between and near the tropics represented on a geographic map Fig. 27. This data map was produced by Edmond Halley in 1686, where the sharp edge of strokes are pointing out from where the wind continually comes.

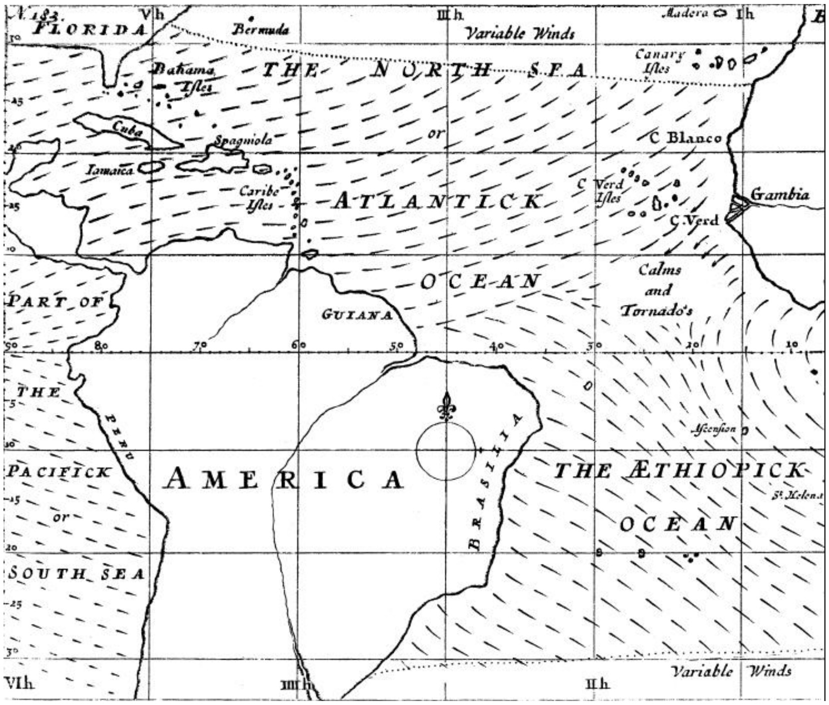


Fig. 27 One of the first statistical maps. A historical account of the trade winds and monsoons observable in the seas between and near tropics represented on map. Edmond Halley, 1686.

According to Tufte[7], one of the most worthy use of a data map was the famous dot map of Dr. John Snow, Fig. 28. In this visualization deaths are represented by dots and eleven water pumps are represented by crosses. The observation led Snow to discover that cholera occurred in the areas near the Broad Street water pump. The reader can think that the link between the pump and disease might have been revealed by analysis and simple deduction/calculation. However, the cost would be the hard work and many spent hours, while here the graphical representation of data is far more efficient than calculation.

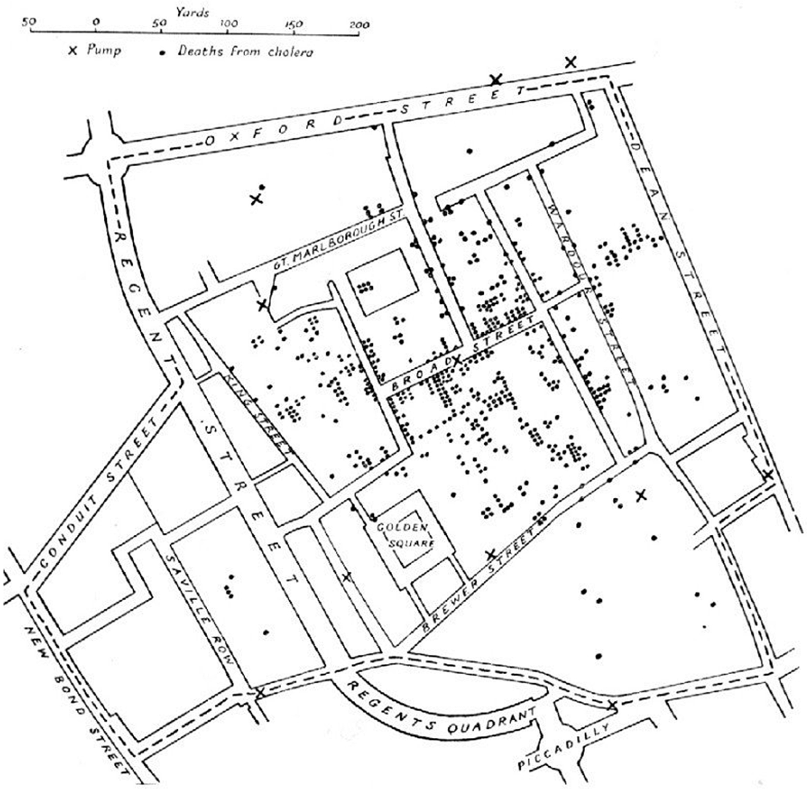


Fig. 28 Use of a data map to display epidemiological information produced by Dr. John Snow, 1854.

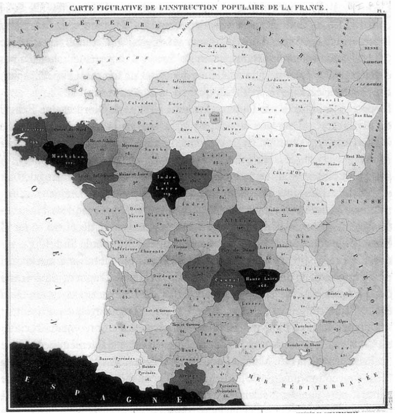


Fig. 29 The distribution and intensity of illiteracy in France, represented with choropleth map with shadings from black to white by Charles Dupin, 1826.

One of the modern statistical maps is Charles Dupin's choropleth map. Sometimes referred to as a shaded map, a choropleth map displays area data by means of shading, color and sometimes texture. The areas might be countries, states, territories, zip codes, trading areas, etc. In Dupin's map he uses shading from black to white to represent the distribution and density of illiteracy in France, Fig. 29.

In the example of the Fig. 30, Charles Minard shows quantity as well as direction. This map show exports of French wine, where the width of a stroke is proportional to the amount of wine. This kind of maps define so called flow maps. So, flow maps say little or nothing about the path, but include the information of what is flowing (moving, migrates, etc.), what direction the flow is moving and how much is being transferred. They are useful to represent exports, however in the Minard's example this could be done by using other technique, say a country by country matrix for two—way flows, or by a simple table. For example in Harness' transportation flow map, presented in Fig. 31, the data density is higher, so the use of another technique would not be that efficient.

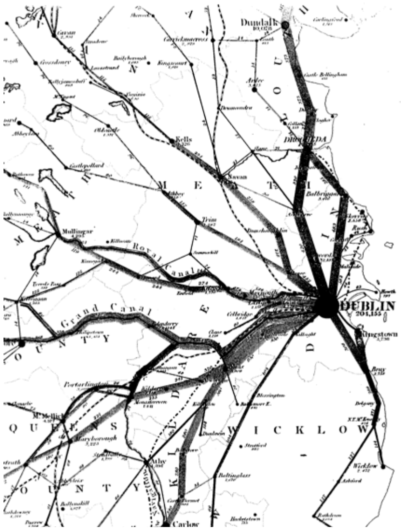


Fig. 31 This flow map shows transportation by means of shaded lines, where the widths are proportional to amount of passengers. Henry Drury Harness, 1837.

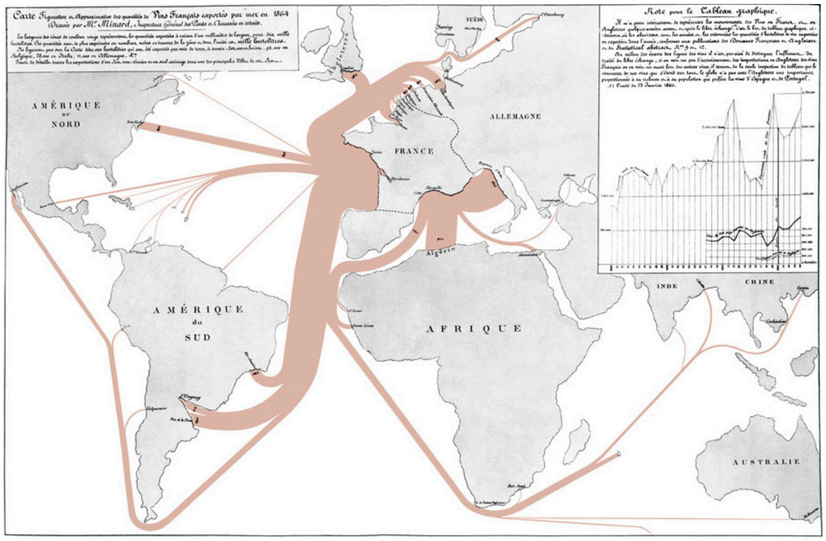

Carte Figurative et Approximative des quantités de Vins Français exportés par mer en 1864
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Note pour le Tableau graphique.
AMÉRIQUE DU NORD
FRANCE
ALLEMAGNE
SUÈDE
Algérie
AMÉRIQUE du SUD
AFRIQUE
INDE
CHINE
AUSTRALIE

Fig. 30 The flow map of exports of French wine, accompanied by a time-series graph, produced by Charles Joseph Minard, 1864

An effective enhancement of the explanatory power of maps representation is to add a time-series description to the design of graphic. So, the data are moving over two/three-dimensional space, as well as over time. On of the excellent example is the classic chart of Napoleon's march on Moscow produced by Charles Joseph Minard in 1869, Fig. 32. This graphic portrays a sequence of losses of Napoleon's army. The width of a line represents the size of the army in each place of the map and color distinguishes invasion (brown) and retreat (black) of the army. The path from Moscow is linked to a temperature scale and dates at the bottom of the graph. This flow map is considered as one of the best statistical graphics ever made[7], since it possess six variables efficiently presented in a single image—the size of the army; location on a two-dimensional space; direction of movement; temperature and dates.

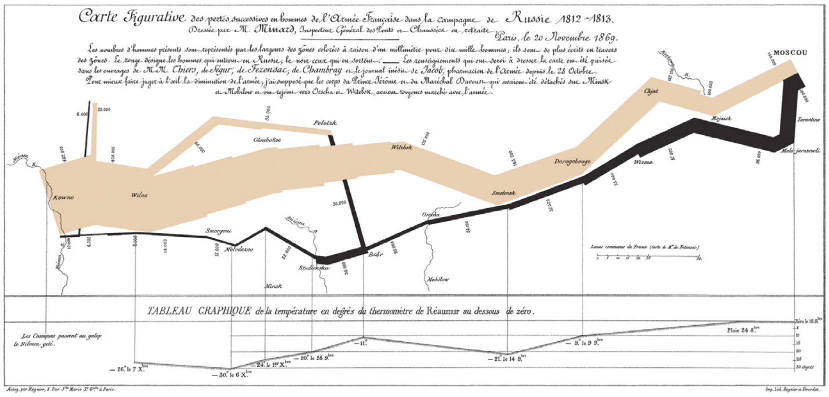

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wizma
Orscha
Bobr
Studienska
Minsk
Molodezno
Smorgoni
Mohilow
Niemen R.
Berezina R.
422.000
400.000
22.000
6.000
60.000
33.000
175.000
145.000
100.000
30.000
10.000
4.000
12.000
14.000
28.000
37.000
24.000
50.000
55.000
87.000
96.000
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niemen gelé.
— 26° le 7 Xbre
— 30° le 6 Xbre
— 24° le 1er Xbre
— 20° le 28 9bre
— 11°
— 21° le 14 9bre
— 9° le 9 9bre
Pluie 24 8bre
Zéro le 18 8bre
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Fig. 32 Flow map of Napoleon's march on Moscow produced by Charles Joseph Minard in 1869. This map represents six variables in one single image: size of the army; location; movement direction; temperature; dates.

In modern analysis of flow data the variation of this visualization technique is many times used in visualization of origin-destination(OD) information. Having for instance OD matrices for subway stops, resulted from aggregation of amount of passengers' check-ins/outs per stop in intervals of time, we can visualize mobility patterns and understand which stops are overused in certain hours or days [25]. The Fig. 33 shows an application of this technique. The OD information is represented by connection of two stops by an arc, and density is represented by color and thickness of a line.

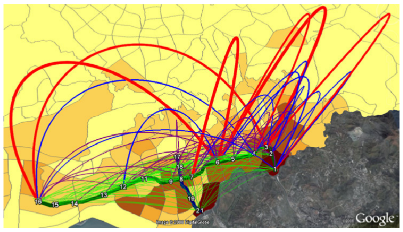


Fig. 33 Visualization of mobility pattern, regarding to the amount of traveling people in subway. This screen shows connections in one of weekdays. This work is a part of a project Urban Mobility Landscape [25].

In urban mobility study the flow can be visualized using animation or line traces of geo-referenced data. In the example of the Fig. 34 authors visualized the flow of taxis in Lisbon by representing each taxi's movement over the time on the map of Lisbon. The traces of each vehicle were colored according to the velocity and intensity of movement—green represent faster traffic and red represents slow traffic. Thought, the flow map visualize the flow of something from one place to another, the variation of this technique result in representing patterns of slow/fast traffic in the city.

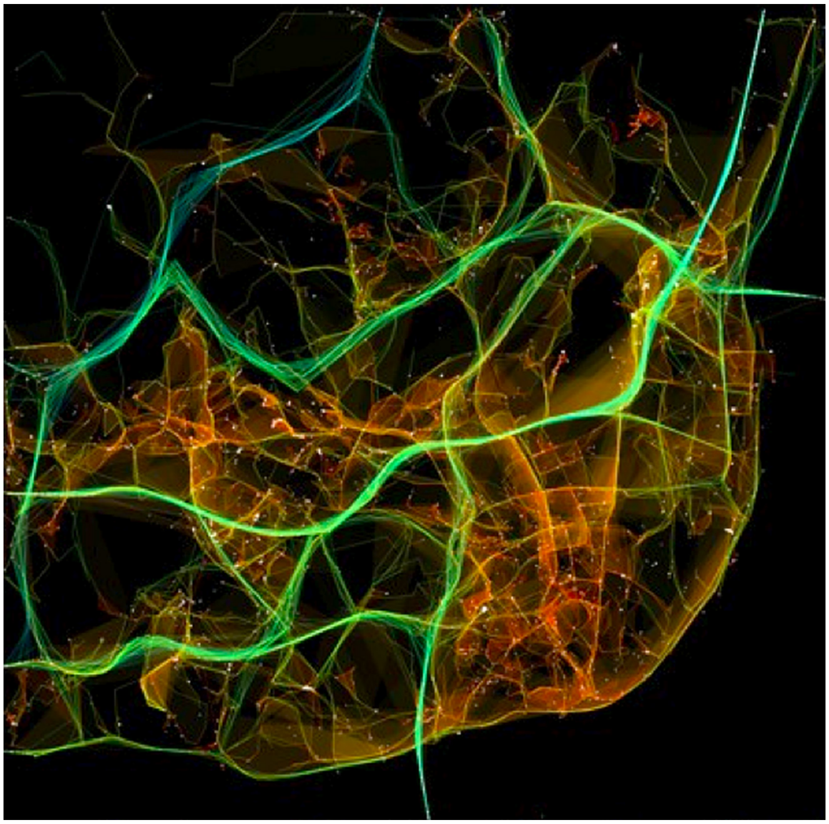


Fig. 34 Snapshot of the visualization Lisbon's slow traffic produced by Pedro Cruz in 2010 [26].

Finally, we would like to review a technique used in statistical maps called heatmaps. Based on map everlying images, heatmaps allow us to assign a quantitative value to every pixel of that image according to each point on the map. Then, according to the value, each pixel changes its opacity and/or color value. Usually in heatmaps color value varies from yellow to deep red. In the example of the Fig. 35 heatmap technique is used to visualize the use of a Microsoft's Live Search Maps, one of online mapping systems [27]. It represents the number of downloads of imagery to the user by coloring corresponding to geographical location pixel–locations from which were made many downloads represented with opaque reddish pixels and location with less downloads are represented with almost transparent orange pixels.

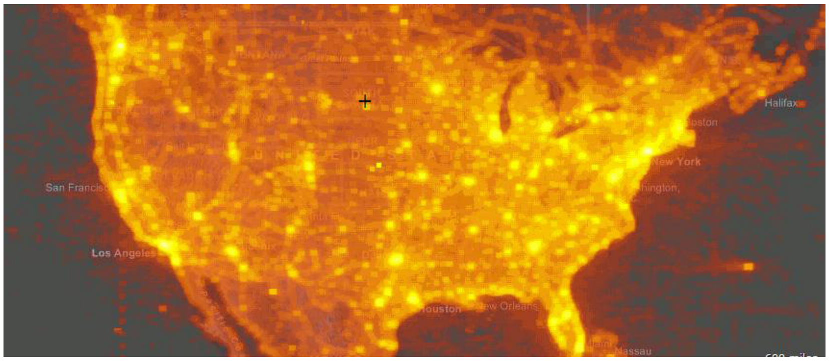


Fig. 35 Screenshot of the application Hotmap, that uses heatmap technique to visualize Microsoft's Live Search Map imagery, 2007

In urban mobility studies this technique can be used to represent information related to travel-times. For example in the mySociety project travel-times is represented as follows: fixed an origin and start time; choose a location of interest; collected and calculated the travel-time from defined point to every point on the map (selected region) [28]. Then the image is drawn over the map, where each pixel is colored according to journey time. In this representation warm colors indicate short journeys and cool ones longer journeys, Fig. 36.

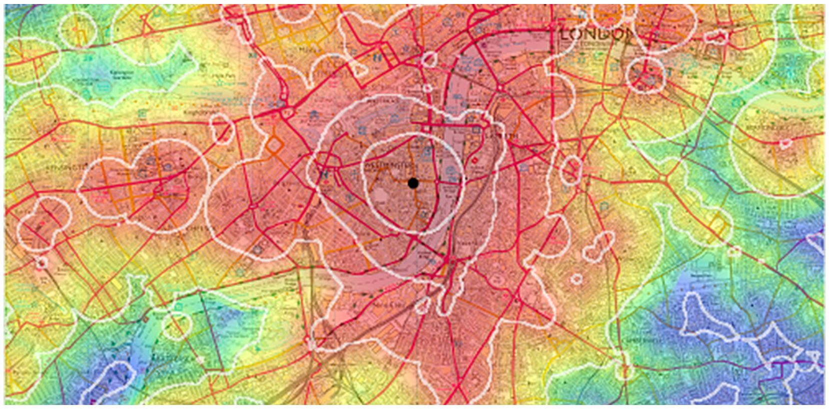


Fig. 36 mySociety visualization—based on heatmap visualization technique each pixel are colored according to the travel-time value—worm color indicates short journey and cool ones longer journey.

Interaction 2.2.3

Navigation and interaction are essential part in exploration of big data sets, moreover when the information is presented on maps or in complex networks. The interaction between the user and the computer-based visualization is made by graphical user interface or GUI. The goal of GUI design is to make the user's experience efficient, productive and, yes, pleasing. In this subchapter we will present some of the basic interaction techniques as zooming and panning. Yet, we will discuss some of existing integrate systems consisted in different graphs that enable the analysis of the same data from different points of view.

Zooming and panning

In large and detailed data sets we are forced to make a compromise between resolution and clarity. Moreover, it is true for maps where the density in many time are so huge that the visual clutter is dominates the meaningful information. Therefore we need an interaction method that will allow us to examine interested areas in closer view, without making any changes in data structure. In designer's dispose there are simple zoom and panning and local zoom technique.

Simple zoom is one of the most common user interaction methods. Its use can be found mostly in every interactive map application (e.g. google maps), as well as in many other interactive visualization tools. So, basically by zooming user can create the display of the data in higher resolution. In other words, user can lay out the visualization with as much space as he needs to identify all interested details. In almost all modern applications the zoom functions are implemented with GUI elements such as a slider or/and with mouse's scrolling wheel.

Simple zoom often works in combination with panning. This interaction technique allows the user to move the data space around the display. This is useful when the representation does not fit in one single display and expands beyond the visual space. By using a set of the directional buttons or by "*dragging*" the representation, the user can bring these hidden parts to the view.

Local zoom is an advanced zooming technique that enlarges only a selected area of the display and acts like a magnifying glass and serves as the alternative to the zooming-and-panning methods. Its implementation can vary from simple uniform augmentation of given area in form of a circle or a square, or more sophisticated like *fish-eye* lens technique. The last is usually implemented by using non-linear equations. Fig. 37 illustrates the behavior of a variation of fish-eye technique. Since the typical fish-eye zoom does not provide large levels of zoom (e.g. from the city overview to the detailed representation of the streets), Pedro came to his approach using a customized lens equation [29]. The resulted implementation behaves in more interesting and pleasing way, while exploring areas with large concentrations of points.

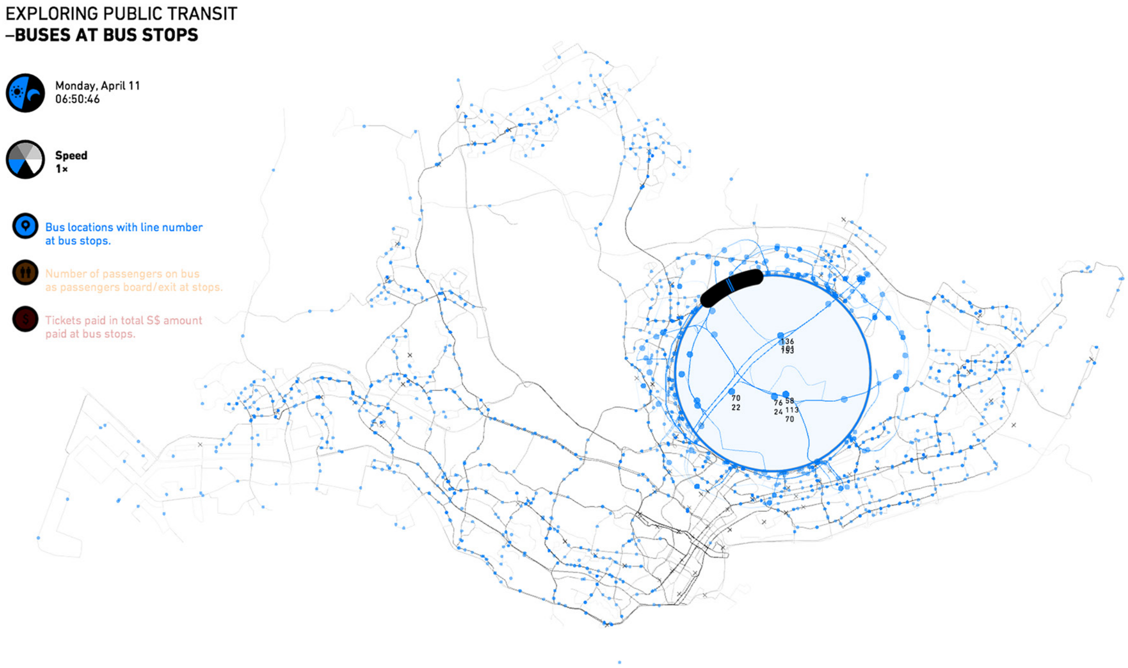

EXPLORING PUBLIC TRANSIT
–BUSES AT BUS STOPS
Monday, April 11
06:50:46
Speed
1×
Bus locations with line number at bus stops.
Number of passengers on bus as passengers board/exit at stops.
Tickets paid in total S$ amount paid at bus stops.
136
103
70
22
76
58
24
113
70

Fig. 37 A screenshot of the visualization of public transport in Singapore—Data Lenses, a part of LIVE Singapore project, produced by Pedro Cruz, 2012

Chapter 3
Related Work

In this chapter we will discuss projects related to our dissertation. In order to prevent possible problems in the realization of practical work and to study in detail used visualization techniques we will focus on the Real Time Rome and Pulse of the City projects. We will study their design process in order to find strongest and weakest sides of used techniques and graphical choices. These projects are similar to our study of urban mobility in terms of questions to be answered. Although the goals are similar, our approaches are different. The sources and processing of the data are distinct. Yet, we will review the Livehood project, which consisted in collection of Foursquare check-ins and further visualization of filtered clusters in order to understand how people use the urban area.

3.1 Real Time Rome

Real Time Rome is a project that focuses on developing of a real-time monitoring system that collects, processes and visualizes the data acquired from telecommunication networks and public transportation system. More precisely, it uses the location of cell phones, public busses and taxis of the city of Rome, Italy in 2006 [30]. The goal of this project is to understand the patterns of daily routine in the city. Their real-time applications help understand the correlation between densities of people and public transport, the correlation between different social groups and the space. In order to reveal the relationship among fixed and dynamic urban elements the diverse mobility information is overlaid on the map of Rome. This project has been presented in the 10th International Architecture Exhibition in Venice.

The collection of the data has proceeded as following: three servers were providing different location data; then it was stored in MySQL database and processed with Java applications in the central server; finally the information was visualized in real-time (with time intervals of 15 minutes) using Google maps as a background. The raw data consisted of location of buses and taxis; traffic noise; amount of telecommunication traffic; density of tourists, pedestrians and and vehicles (including also the speed of these). The location data were provided by GPS devices. The traffic noise data was provided by wireless sensors installed in Rome. The amount of telecommunication traffic was provided by local Telecom Italia Company and was collected and aggregated in 15 minutes intervals. The densities were obtained by using more sophisticated techniques. Basically, using the approximate distances and strengths from cell phone to GSM base stations they calculated and aggregated the location data in matrices with corresponded elements to 250 by 250 meter squares [30].

Real Time Rome uses six different visualizations to present real-time information described above. Each visualization application provides answers to different questions.

1. “*Where in Rome are people converging over the course of a day?*” [30]. They call this visualization tool **Pulse**. Basically, this software visualizes the intensity of mobile calls in the Rome at a given time. In order to represent this information, the Rome urban area was divided by 40x40 meter squares and the telecommunication traffic intensity values were assigned to each of these squares. The correlation was established by using the exponential distribution function. The software shows the real-time data, divided in intervals of 15 minutes, on the left side. On the right side of the screen user can find representation of the data related to yesterday’s values Fig. 38.

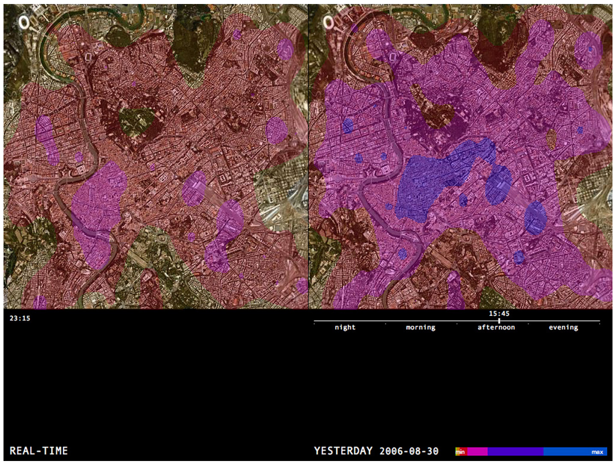


Fig. 38 Sreenshot of the application Pulse. On the left side — visualization of the current day. On the right side — visualization of the previous day.

2. “*How do people occupy and move through certain areas of the city during special events?*” [30]. This visualization tool called **Gathering**. Basically, it shows pre-recorded movements of mobile phone users during the special events. Visualization is based on 3D approach, where the urban area was divided in squares and each one represents the traffic intensity. Plotted over Google maps, this visualization uses 3rd dimension and color to represent intensity values. In terms of interaction, the user can find a verbal description of location and time, as well as description of actions that happened in a given moment as time progresses. Yet, software presents a timeline where the user can find the numerical and graphical representation of current visualization time, Fig. 39.

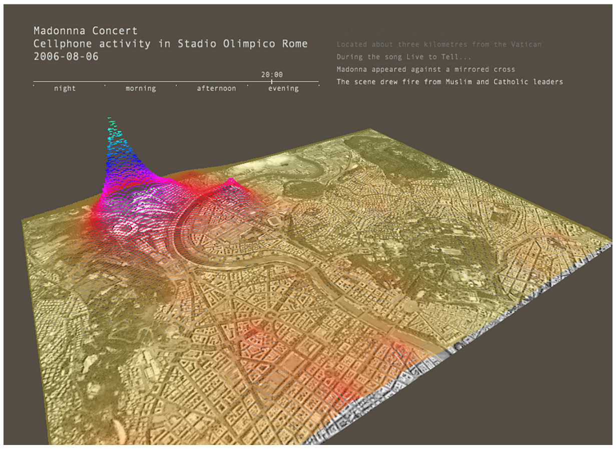

Madonnna Concert
Cellphone activity in Stadio Olimpico Rome
2006-08-06
20:00
night
morning
afternoon
evening
Located about three kilometres from the Vatican
During the song Live to Tell...
Madonna appeared against a mirrored cross
The scene drew fire from Muslim and Catholic leaders

Fig. 39 Visualizing Madonna's concert in Rome on August of 2006. During the event were recorded great spikes of mobile phone use during particularly emotional moments.

3. "*Which landmarks in Rome attract more people?*" [30]. **Icons** software represents the density of mobile phone users at different historic points in Rome. In order to visualize this information, they used a weighted mean for each attraction, based on traffic intensity associated with the square corresponded to that attraction point. This information was represented by using a bar on top of each attraction point, which height was mapped to the relative intensity values. At the bottom user can find a line graph that represents a week-long data of the most popular (green line) and the least popular (red line) place. The visualization is updated every 15 minutes, when the new data is available, Fig. 40.

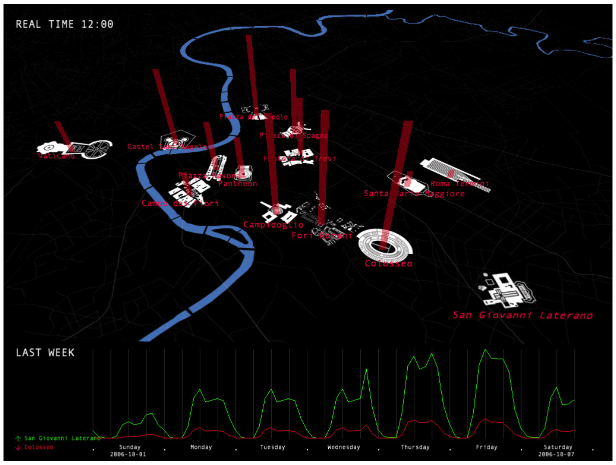


Fig. 40 Screenshot of the software, Icons, that represents the density of people at the attraction points in Rome.

4. "*Where is the concentration of foreigners in Rome?*" [30]. **Visitors**, the software that shows the concentration of the foreigners in Rome, using the same technique as for Gathering. In this case, investigators focused on areas around the *Stazione Termini* (Terminal Station) in Rome. Basically, it highlights the tourists speaking on mobile phone with a 24 hour loop, using the 3rd dimension and color to represent the intensity data, Fig. 41.

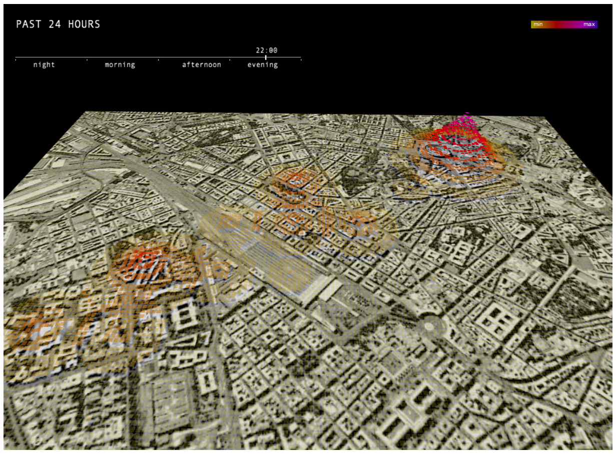


Fig. 41 A screenshot of the visualization application Visitors.

5. “*Is public transportation where the people are? How do the movement patterns of buses and taxis and pedestrians overlap in the Stazione Termini neighborhood of Rome?*” [30]. They call this application **Connectivity**. This visualization shows the movement of buses and taxis indicated by yellow points and the relative densities of people represented by red areas. This software estimates the trajectory of bus’s and taxi’s movements, based on previous locations represented by yellow tails on the map. Using a color (combination of saturation and opacity) the software represents pedestrian densities with intervals of 5 minutes, Fig. 42.

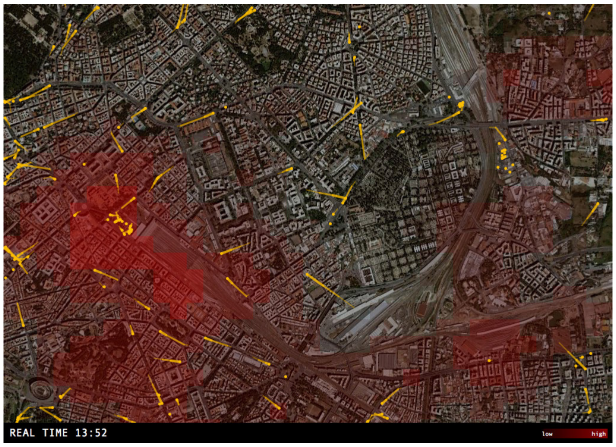


Fig. 42 Snapshot of the visualization application Connectivity.

6. “*Where is traffic moving?*” [30]. The software **Flow** visualizes the movement of the people traveling in vehicles. It focuses on the area of the *Stazioni Termini* and the *Grande Roccordo Anulare* in Rome. The urban area was divided in squares, where the red ones represent slow traffic and the green one represent quick traffic. The arrows represent the dominant direction of movement, Fig. 43.

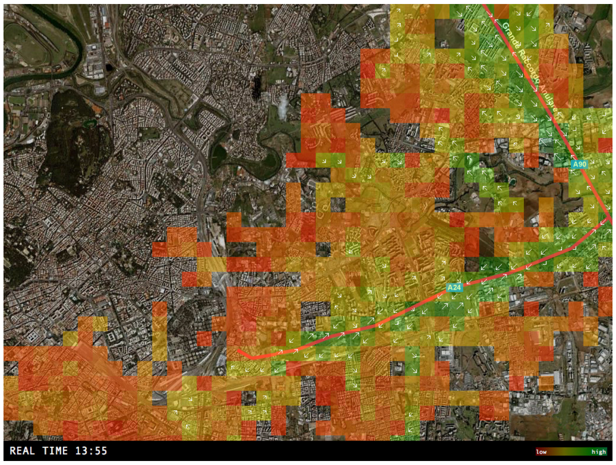


Fig. 43 Snapshot of the visualization of Flow software. Red color represents slow traffic and green color represents quick traffic.

The project Real Time Rome used five different visual approaches in order to represent diverse types of information that help to understand the relationship between people and urban space/transportation. Most representations were laid-out over the Google maps. When we first analyzed this work, we ask ourselves "*why did they use satellite imagery for the map?*" The combination of "*artificial*" graphical elements and the natural map may cause aesthetic conflicts, as they speak in different visual languages. Since the researchers use a map only for geographical reference, the visual aspect of the map could be customized to combine with other graphical elements in terms of color, tone of voice, style of lines, etc. Yet, the map could present the semantic information as well. Since the final visual artifacts are presented to the public it would be easier orient in space for the people that had no contact with Rome.

In the case of Pulse visualization it is not obvious which screen represents what data. Yet, regarding the layout, there is an unused whitespace, that create uncomfortable visual gap between elements—the screens and the labels. Also, it is not clear why the screen on the right possess the timeline and one on the left not and why the time indicators display different time. If both of the screens are not synchronized, why it is not indicated somehow? The absence of such of information may lead to incorrect conclusions, while comparing one with other.

As for the visualization model, in our opinion, the color variable was the right choice in order to represent the intensity data, with few exceptions, assuming that the information was classified as ordered (divided in ranges that qualify the intensity) and not quantitative. The changes in hue can provide efficient visual detection and assembly, by visual delimiting the boarders of shapes. However, for the estimation of ranking, the use of hue is not that efficient in pattern perception [8]. Thus, this model does not present immediate perception of ordered information, but works fine for detecting patterns of distribution of mobile phone users. The application Connectivity, in contrast, represents quantitative information using opacity.

In other hand, the flow visualization presents correct use of color for pattern perception. As in Pulse, the hue changes allow user to detect and visually group graphical elements in patterns. Moreover, this model provides correct discrimination of elements, while estimating the values interpolating the color from red to green. Yet, the lengths of arrows increase efficiency of the visual estimation. Also, the use of arrows provides the direction of flow, when presented.

As we mentioned earlier three of the visualization tools use 3D approach. The Icons application lacks of any kind of physical information (e.g. a reference grid, tick marks) making it impossible to visually estimate quantitative values, in this case densities of people that use mobile phone at attraction points. The use of 3D distorts representation, so the interpolation operation is inefficient and inaccurate. For example, let us compare bar height for *Colosseo*, *Fori Romani* and *Piazza del Polo*, considering the point of view represented on Fig. 40. As the Colosseo and Fori Romani are located near each other, by visually estimating the height of each bar, we can say that they are more or less the same. Then, the Piazza del polo visually looks smaller, however, taking in account the distortion caused by 3D projection, we can say that it is taller. Anyway the conclusion will be inaccurate. Thus, the table-lookup process is impossible in this visualization and it leads to subjective and inaccurate conclusions.

Use of 3D in data visualization, requires the observation from all possible points of view, in order to make more accurate conclusions. The problem is that the distortion caused by 3D projection, makes graphical elements appear bigger/smaller than they are. This leads to an inaccurate estimation of their size. Other issue is the overlapping of higher peaks with lower ones. This quantity of visual clutter is not contributing to significant communication. In case of Pulse visualization, the division of the surface by square tiles and elevation of them, making the shapes semi-transparent, as well as use of color, helps the representation to be more efficient. So, the user decodes the information estimating not only the hight of peaks, but the color hue.

3.2 Pulse of the City

This project was created for the President Obama's 100th day in office. The visualizations of mobile phone digital footprints during the inauguration period answers the questions: "Who was in Washington, D.C. for President Obama's Inauguration Day? When did they arrive, where did they go, and how long did they stay?". The data was provided by large telecom operators and aggregated by time intervals and physical location. The information was embodied in two visualizations—The City and The World, we will focus only on the first one [31].

The first visualization, The City, answers the first question, by representing the phone call activity in Washington D.C. Fig. 44. Based on the same technique used in Pulse software, this visualization improved aesthetically and in efficiency through the addition of two new features such as a meaningful timeline, represented by a bar chart, and another bar chart that summarizes the phone activity from 50 states and 138 foreign countries. The timeline represents the trend of call activity during the week of the inauguration. The beginning of important moments is marked on the timeline by a vertical line and description of that moment.

At this time the map uses simple and clear visual language. The map is simplified to the representation of only the road network, using a thin white line. This reduction increased the visual integration with other graphical elements. The information is represented by using 3D surface, which is divided by square tiles of 150x150 meters. The call activity in each area is assigned to corresponding tile and the value is encoded by *z* position and the color of each tile (from yellow to red according to low and high activity values).

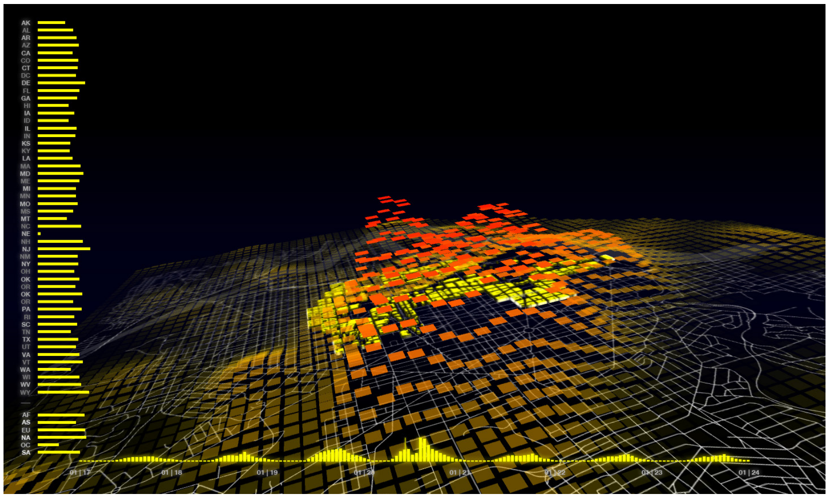


Fig. 44 Snapshot of the application The City. Washington, D.C.

Livehoods 3.3

The Livehoods project consists in a clustering model for mapping a city regarding on the collective behaviors of its residents and further visualization on the map [32]. Based on Foursquare check-ins they shaped the dynamics, structure and character of a city using a computational model, since it requires hundreds of hours of observation and interviews for city planners and researchers to learn about the city. Given geospatial social data from hundreds of thousands people they developed an algorithm that represents distinct areas of the city regarding to the activity patterns. The resulting aggregated clusters of check-ins represent so called mental map of the city, the vision of urban space from users perspective.

The data for their clustering model came from Foursquare online service, and consisted in 18 million check-ins in 2011. Since Foursquare check-ins are by default not publicly visible these 18 millions of check-ins were collected from Twitter public timeline and then where aligned with venue information from the Foursquare API. The data consisted of the user ID, the time, the latitude and longitude, the name of the venue and the category of the place. With further filtering their data consisted of 42787 check-ins from 3840 users at 5349 venues from Pittsburg metropolitan area [32].

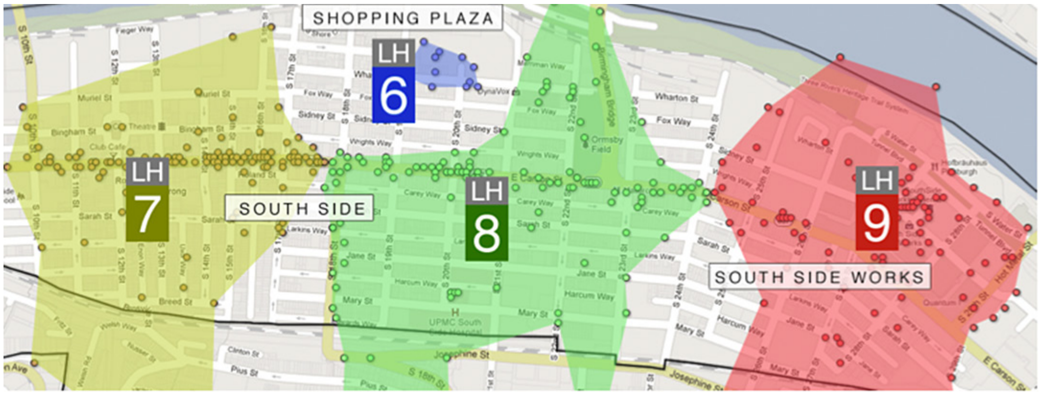
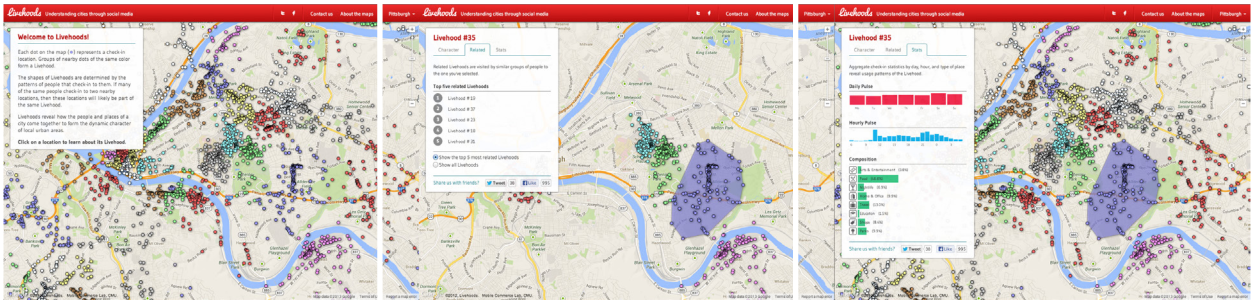


Fig. 45 Snapshot of the web application Livehoods. The municipal borders are represented with black lines and Livehoods are represented with colored shapes.

The output from their algorithm, named Livehood clusters, was visualized in an interactive web application, allowing user to explore various check-ins and venue statistics for each Livehood and compare the structure of different Livehoods. In Fig. 45 we can see some of Livehood clusters represented with different colors and shapes defined by their clustering model. Yet, check-in locations are represented with points of the same color of the corresponding cluster. The underlying map is provided by Google Maps and uses its default theme.

The interactive web application provides some useful functionalities to explore the data [33]. When the user opens the visualization he is presented with all check-ins displayed on Google Maps and is asked to choose one of them to learn about corresponding cluster. When a point is clicked the side panel, with information regarding to the cluster appears. In this panel we can find three tabs: character, related and stats. In the character tab the user can find the popular check-in locations and unique types of places found in selected cluster. In the second tab the user can find related clusters visited by similar groups of people to the selected cluster. When this tab is opened all the points are muted except the points of selected and related clusters. In the stats tab the user can find detailed statistics of aggregated check-ins by day, hour and type of place, Fig. 46.



Welcome to Livehoods!

Each dot on the map (●) represents a check-in location. Groups of nearby dots of the same color form a Livehood.

The shapes of Livehoods are determined by the patterns of people that check-in to them. If many of the same people check-in to two nearby locations, then these locations will likely be part of the same Livehood.

Livehoods reveal how the people and places of a city come together to form the dynamic character of local urban areas.

Click on a location to learn about its Livehood.

Livehood #35

Related Livehoods are visited by similar groups of people to the one you've selected.

Top five related Livehoods

1 Livehood #19
2 Livehood #37
3 Livehood #23
4 Livehood #18
5 Livehood #31

● Show the top 5 most related Livehoods
○ Show all Livehoods

Livehood #35

Aggregate check-in statistics by day, hour, and type of place reveal usage patterns of the Livehood.

Composition

Arts & Entertainment	(3.8%)
Food	(46.6%)
Nightlife	(6.9%)
Home & Office	(9.9%)
Travel	(13.0%)
Education	(1.5%)
Shops	(8.4%)
Parks	(9.9%)

Fig. 46 Livehood interactive web application. On the left image all check-ins are presented. On the central image it is presented selected and related clusters. On the right image additional statistical information is presented.

Discussion 3.4

By presenting above projects we discussed diverse visualization techniques for studying crowds' behavior and relationship with the urban space. Using the relative information of users' mobile phone calls (e.g. location, call time, amount of telecommunication traffic) and public transport location we can apply these techniques in order to answer the questions about the urban use and mobility. By visualizing intensities of phone calls we can understand the patterns of crowds' moviments over along the day. These gestalts are created by mapping ranges of intensity with different color hues. While visualizing the intensity of phone calls during special events we can find the most emotional moments. The team of Real Time Rome used visualization technique based on 3rd dimension. The values are represented with changing of *Z* position of square tiles they are assigned to, as well as with the change of color. By comparing densities of mobile phone users with location and speed of public transport (e.g. buses and taxis) we can understand if there is enough accessibility to the public transportation in places of large concentration of people. Finally, by filtering phone users that move abnormally quickly and concluding that these people are moving in vehicles, we can visualize this information, using color hues and arrows, in order to understand where and how the traffic moves. In other hand Livehoods used the data collected online, which previously was introduced by users. Their clustering model enables people who wants to learn about the city, explore the information through the interactive web application. Thus allowing a interactive learning about the city. So, this visualization relies completely on web resources in order to reveal crowds behavior and understand the relationship between people and the space they use.

Chapter 4

Urban Mobility

In this chapter we will discuss the whole process of visualization of our data. Starting with the description of the data and tools we used. Then we will cover our first approaches to visualize the data and how the conclusions of one influence the improvements of the following representation. Our visualization model will be then discussed in detail. Finally we will present the used interaction techniques to explore the data set.

Data

According to the pipeline of data visualization process described in the methodology section, we started with representation, or to be more precise with information analysis that was previously collected, filtered and treated by other members of the team. To us the data comes in groups of two files. One contained the data related to the location, name and ID for bus stops in Coimbra and London, and bus stops and metro stations of Singapore cities. The second file contained counts, means and standard deviations of people aggregated by stop, arrivals/departures and time intervals, which will vary through different stages of the project (e.g. on the 13 of April of 2012 at 10 o'clock eleven persons cutched the bus on stop *R. Ant. Neves da Costa*, which resulted in standard deviation of 2.2).

The identifiable parts of given information vary in each case, however the common variables are the latitude and longitude, count, mean and standard deviation. The data from Singapore have one more component which is the stop ID that possesses two categories of bus and metro stops. Yet, the data from Singapore is grouped by arrivals and departures, which is particular of the data from Singapore. Data sets from other cities have only arrivals. Later we will differentiate bus from metro stops by representing one group at each time, since the crowd behavior and, consequentially, the scale of data varies when using buses or metro transport. As well as arrivals and departures will be visualized separately due to amount of visual clutter. So, the construction of a graphic will be a statistical map, since the data is geo referenced. The chosen graphical elements will be discussed later in this chapter.

Tools and Implementation

In the main framework of our project we used Processing [34]. This programming language, based on Java, provides all possible tools for working with graphical elements, starting with drawing simple lines and circles, through to the shader based 3D animations. In combination with the Processing we used Java libraries such as Java Map Projection Library, which provides diverse types of map projection [35], and custom libraries for loading the data from CSV files, although newer versions of Processing provides the API for that. The later improvements of our visualization model were implemented in OpenGL Shading Language (GLSL) using the API provided by Processing framework. All the map data comes from OpenStreetMaps [36] and filtered using Open Source QGIS [37] application. Then maps were exported to the *.svg* format, since Processing provides useful tools to work with vector graphics, such as loading graphics and changing their properties with a "*one code line*". The final retouch of maps was made in Adobe Illustrator program, while filtering and ordering of data were made with Python scripts.

Design requirements

In order to guide the design of our visualization in a specific way, we established the following design requirements, that defines the boundaries for the project. The visualization should create a digital layer of urban space. Using a simple and clear visual language it should establish a strong relationship between a crowd and the urban space. Thus, the representation should reflect the geographic nature of data. For that reason the information should be visualized on a map. Yet, the visualization should be interactive, allowing the user to explore the data set in space and time. Moreover, it should be understood by the users with no analytical background, so there should be a good balance between aesthetics and functionality, without visual overload. Finally, the visualization should run in real time, so that the user could explore the data.

4.1 First Approach

This was the first step in data visualization process—the basic representation of given data. By directly representing count values by bubbles, more precisely the counts of each bus stop and metro station in given time interval were mapped to the radius of representing circle, Fig. 47.

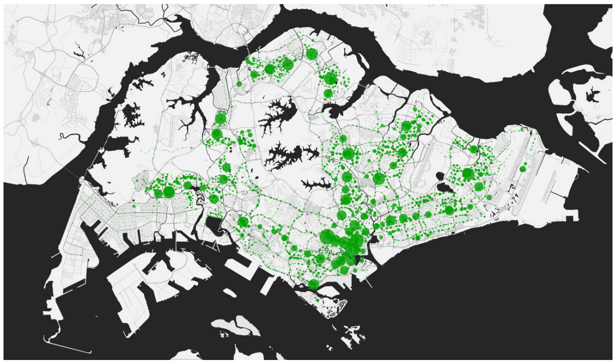


Fig. 47 Simple mapping of counts per bus stop and metro stations in Singapore by green circles overlaied over a map.

Then we proceeded to represent each stop by two overlapping circles, one for arrival counts, painted in green color, and another for departure counts, painted with black color. The used colors were applied with 50% of transparency, so both could be visible. The data was aggregated in time intervals of half hour (Fig. 48).

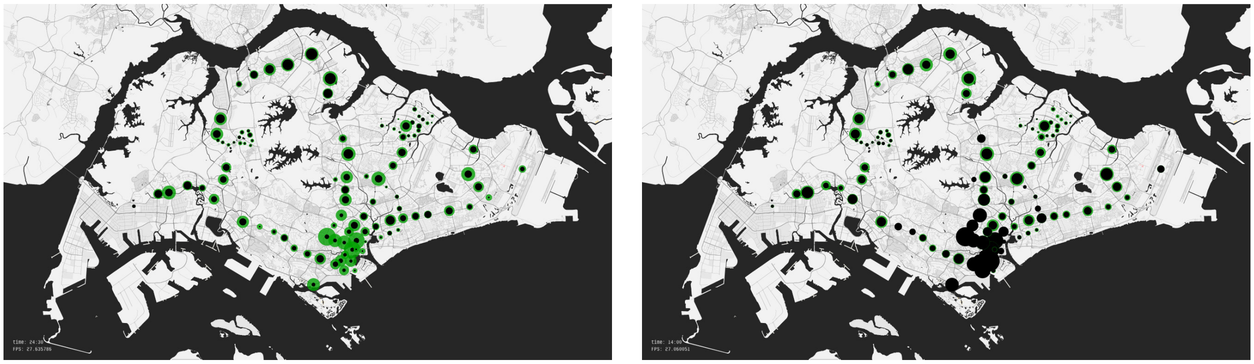


Fig. 48 Simple mapping of counts per metro stations in Singapore area. Green represents arrivals and black represents departures. Image on the left displays the city at 24.30h and image on the left displays the city at 14.00h

Due to the high density of data the amount of visual clutter made the representation inaccurate. Yet, the perception of the representation of counts by radius of circles is not precise, since the numeric and visual scales are different (Fig. 49). So, we concluded that this basic representation, which was done in short time period, is not efficient and we need to find another approach to represent our data.

Perpendicular Vectors 4.2

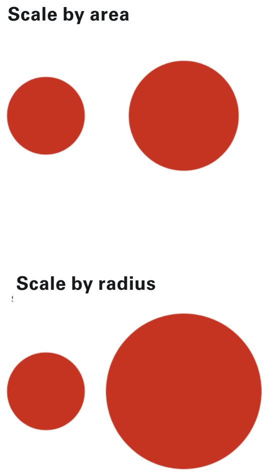
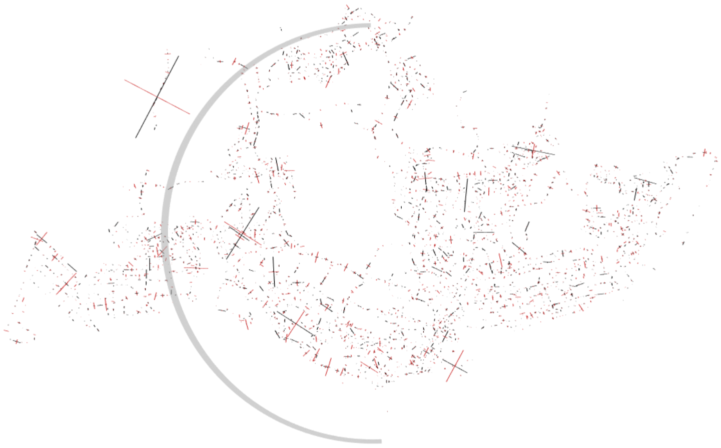


Fig. 49 Comparison of scale by radius and by area. The area and the radius of the small circles are equal. Now lets scale them by a factor of two. The left image illustrates the scale by area and the right by radius. By scaling by area the perceived amount of information is exactly two times, however the circle scaled by radius appears more than two times bigger. The equation for calculating radius from given area is sqrt(A/PI) where A is the area of the circle.

Due to the high density of data we needed graphical elements that were capable to represent tradeoff between the occupied space and amount of conveyed information. On the dot map of disease, John Snow [7] represents deaths from cholera and water pump sites with dots and crosses respectively. By using elementary graphical elements this visualization efficiently conveys the information. The perpendicular arrangement of dots in relation to streets, that represent deaths, makes this visualization precise, aesthetically clear and comprehensive.

With this observation we proceeded to represent our data in a similar way. The challenge was to find parallel vectors to streets closest to given stops. For the basic representation it would be time consuming to develop a sophisticated mechanism in order to find this information, so we manually indicated the corresponding streets to the bus stops. As a metro stations does not belong to a unique street it was problematic for us to identify the directional vectors, though the visually closest street was indicated. The calculation of the parallel vectors was trivial—each street's directional vector was duplicated, normalized and rotated 90° CW, so the perpendicular vector points outside the street. On the Fig. 50 these vectors are represented with red lines. Having parallel vectors we could represent our data with marks associated to the line implantation. So, the representation below shows counts for arrivals and departures in red and black colors respectively. The count values are mapped to the length of the lines having cross point location equal to the bus/metro stops location.

Fig. 50 Counts are represented with black and red lines for arrivals and departures respectively. The red lines are perpendicular to the corresponding streets.



4.3 From General to Particular

Although the first basic visualization was not enough efficient to our research, we were be able to conclude that representing simple counts would not tell as enough about patterns of crowds behavior. From what we know the behavior of crowds repeats in weekly loop. So, this can be the key to estimate deviations in the public transport use. First of all we needed to distinguish weekdays from weekends since the behavior variations are distinct. To be more precise we calculate deviations aggregated by the same weekdays (e.g. all the Mondays in data set). So, we went back to the mining stage and by using the ***c*m/d*** equation, where ***c*** is the count of arrivals/departures for each stop, ***m*** is the mean for each time interval and ***d*** is the standard deviation, we calculated anomalies, i.e. unpredicted use of public transports. As a result we obtained values that revealed unusually high or low passenger counts of departures/arrivals along the course of the day. This created another issue—the output of the equation could be either positive either negative. So, how to graphically distinguish these two concepts?

First of all, we wanted to improve aesthetic aspect of representation and its efficiency. This small experience was rallied on triangle graphical elements, due to its pointing nature. Irregular triangles conduct user's attention to the direction they pointing at. Yet, the size of an area implantation is the most efficient to represent quantitative values. From that observation we concluded that the triangle's pointing direction could be used to represent positive or negative numbers and size could represent the value. Yet, the arrivals and departures are represented one at each time, as well as metro or bus stops. Another important aspect of analysis of such information is that we need to pay more attention to high deviations. These ones are more crucial in adjustment/management of public transport system. In order to emphasize these anomalies we used transparency of the triangles. So, anomalies with great values are represented with opaque triangles, while small ones are represented with almost invisible triangle, Fig. 51.

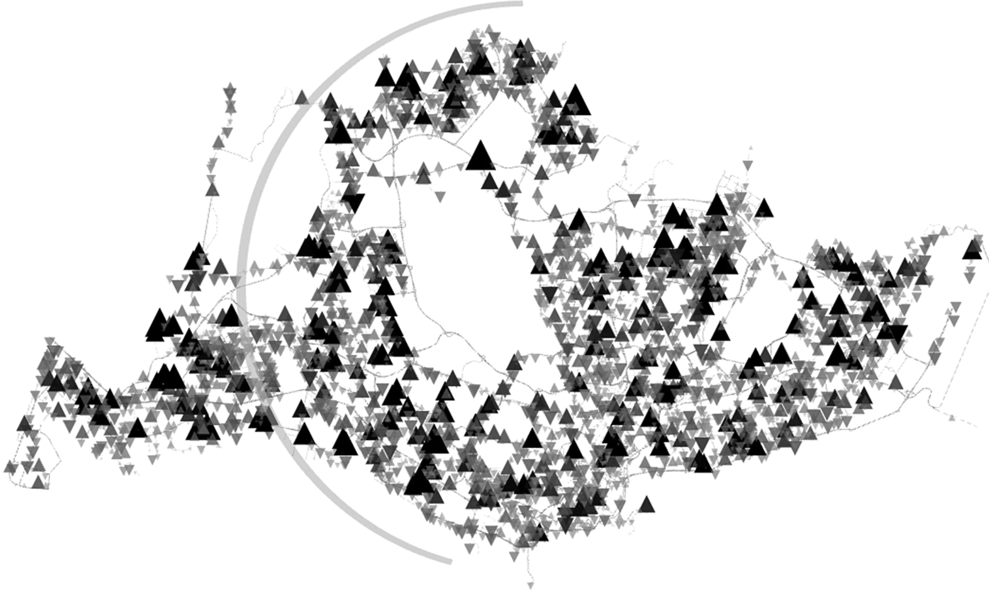


Fig. 51 Snapshot of visualization of Singapore public transport use. The deviations are represented with size and transparency of triangles. Positive and negative values are represented with the triangles pointing up and down respectively.

At this point we realized that representation of the high dense data in one single image was not efficient. So, we decided to go back to the basic representation, based on the parallel lines model, where each small diagram represents the data of corresponding stop. In this approach we straightened the orientation of the diagrams, so that the horizontal and vertical lines form a cross, which serves as a local Cartesian coordinate system. The horizontal lines represent departures and the vertical represent arrivals. Positive and negative values correspond to the sign of *x* and *y* axis. The relative origin of each diagram specify the physical location of corresponding stop (Fig. 52).

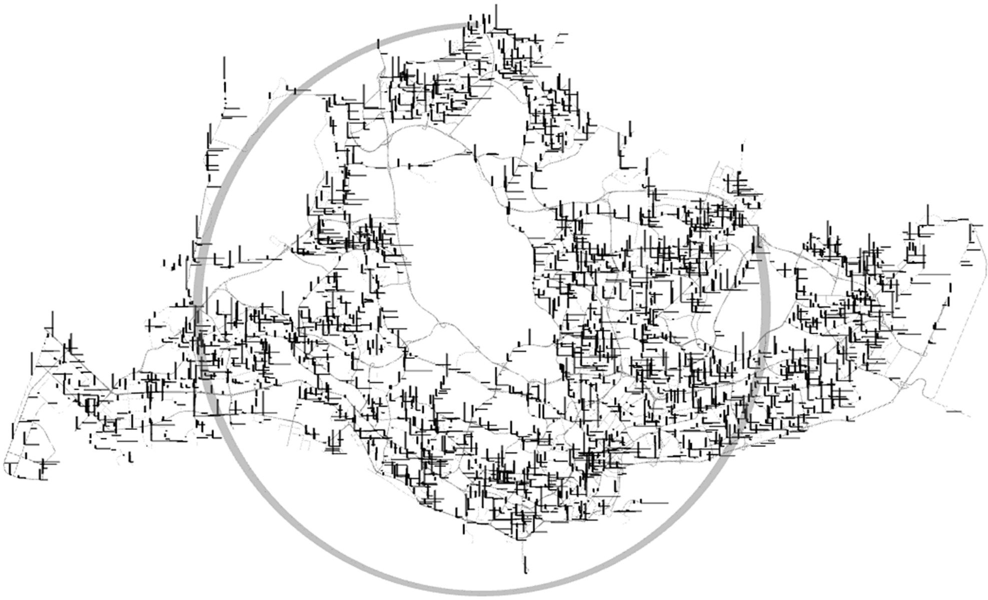


Fig. 52 Visualization of anomalies in public transport use in Singapore. Displayed are only bus stops and anomalies for arrivals and departures.

As the amount of visual clutter did not reduce and the visualization was still confusing, due to the high density of stops, the visualization had to be expanded with interaction capabilities. We added the ability to zoom the image, which allowed the user to explore the representation in detail. Simple zooming and panning method partially resolved the problem. Since the scale was applied to the whole image, map and diagrams, the visualization still presented a lot of overlapped lines. So, we decided to retain the initial scale of diagrams, so that the ratio of line sizes was equal, Fig. 53.

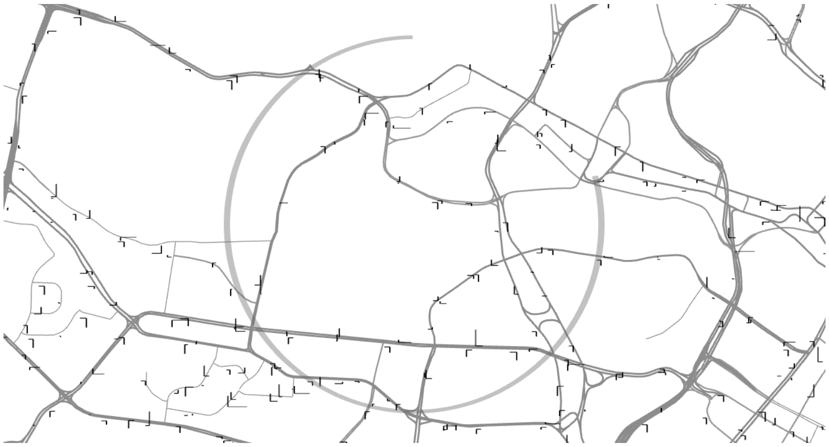


Fig. 53 Zoomed visualization retains the correct relationship between stops in terms of anomaly values. Lengths of lines are scaled down proportionally to the zoom scale factor in order to reduce overlaps.

We noted that the zooming technique has not only exploratory aspects, but it allows the user to detect graphical elements and group them in one visual pattern in general view, and accurately estimate their relative sizes by zooming to the interested part. With this observation we divided pattern recognition in two stages: analyzing the general view representation and zoomed part. So, why not to represent the information using the appropriate for each perceptional stage, visualization models? Could it be one visualization model that transforms in each stage? Or it could be different models with the similar properties that morph from one to another?

Introducing Subjectivity 4.4

So, in the previous section put the hypothesis that the zoom technique could be used in order to transform the same visualization model or morph between different representations. Obviously, the development of the visualization model that transforms, retaining the same visual aspect, could be done with less cost, since the morph among two or more representation requires establishing of common characteristic, finding the same visual language and implementing each of them.

So, what are the properties of simple transformation model? First of all we needed to define what kind of question each view may answer. What do we want to see in each view? The general view must have strong detection and assembly characteristics, so that the user can detect and assemble visual patterns in one perceptional instant. Like in *UrbanCyclr* visualization the user can quickly detect areas with high biking traffic. In our opinion the graphical representation at this level can be subjective, since this kind of representation in some cases have better communicative abilities. In other hand, the zoomed view must have efficient visual estimation characteristics. As the user zooms the image in, he wants to see it as rigorous as possible. At this view the essential is the accurate perceptual estimation of graphical elements.

So, our approach was based on the *bubble* representation and on the *Metaball*, aka blob, technique [38]. It is important to describe the metaball nature, since it distorts representation of input data. Basically, *Metaball* defines a function, in our case in a two-dimensional space, which takes as input the *x* and *y* coordinate of a point, and outputs a floating point value: deciding on threshold for the output the value become 1 if above the threshold and 0 if below [38]. The calculated points define the so called *isosurface*. Most common function to define isosurface is magnetic field equation which is $\boldsymbol{1/i^2}$ where $\boldsymbol{i}$ is the force of influence representing the distance from the charge point to the calculated pixel. One property of this function is that the output never comes to 0. That means that every pixel of the iso-surface will be influenced by every charge point. On the Fig. 54 red circles represent inputs and black circles represent the output from equation. As we can see the small circles are influenced by the big one, so their radiuses become bigger than they were initially.

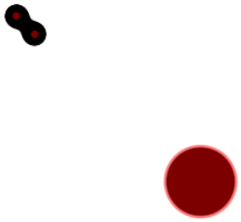


Fig. 54 Comparison of metaballs (black) to bubbles (red) with equal input radiuses. The figure illustrates that the small circles changed their radiuses with the influence of a big one. The sum of all influence forces doubles the radius of each small circle.

The visual output from the equation of magnetic field results in blobs with equal radiuses. That is not what we wanted. So we transformed the equation to the $\boldsymbol{r/d^2}$ where $\boldsymbol{r}$ is the calculated radius of an imaginary circle and $\boldsymbol{d}$ is the distance between charge point position and current pixel location, Fig. 55. When $\boldsymbol{d}$ is minor or equal to $\boldsymbol{r}$ the output will be 1 and grater. Consequentially, if we set the threshold of 1 and give only one charge point, then the visual output will be one circle with the radius equal to input value.

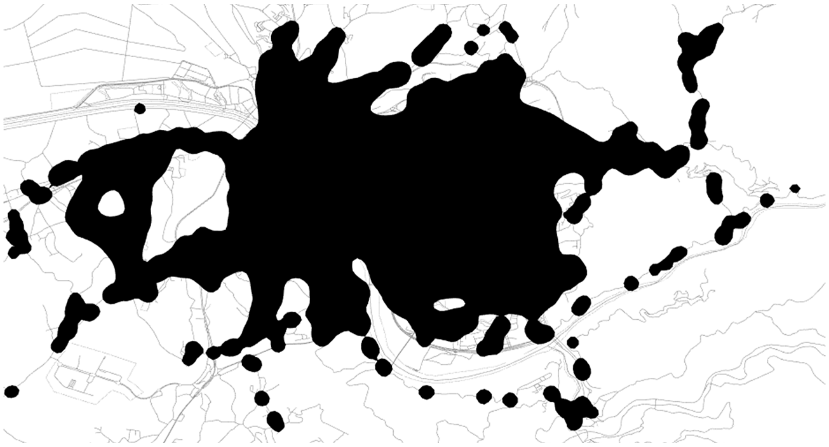
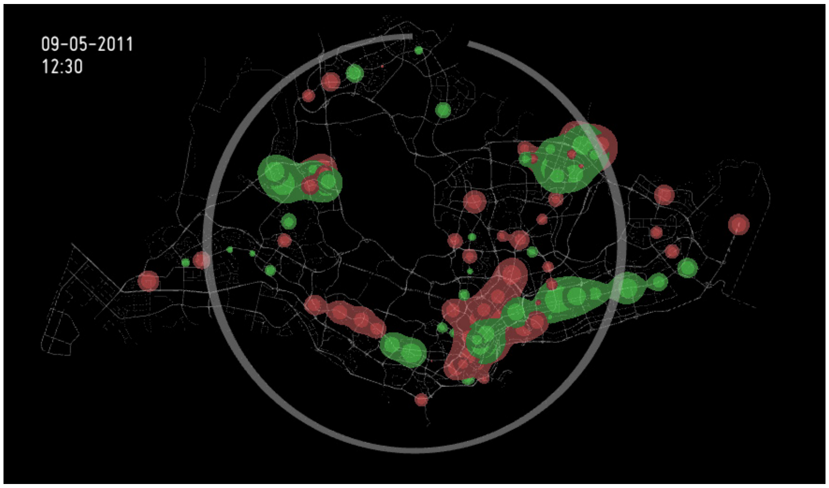


Fig. 55 Visual output from the equation plotted on the map with 500 stops of Coimbra city.

In Processing the zoom can be done by changing the scale of an image. Internally the scale is done by applying a *transformation matrix* over each point (e.g. vertices of a shape). So, let the scale be equals to 3 and let the point *P* have coordinates *P(100, 100)*. The result of multiplying vector *P* with the transformation matrix will be P_t*(300, 300)*. So, it triples the distance from the origin. So, as the influence of point depends on the distance between them and by scaling the image we change the distances, the combining of the zoom technique with metaball result in what we were seeking for: in general view we have subjective representation and by zooming the visualization becomes rigorous with bubbles' areas equal to the input values. The application of the described technique is illustrated on the Fig. 56.



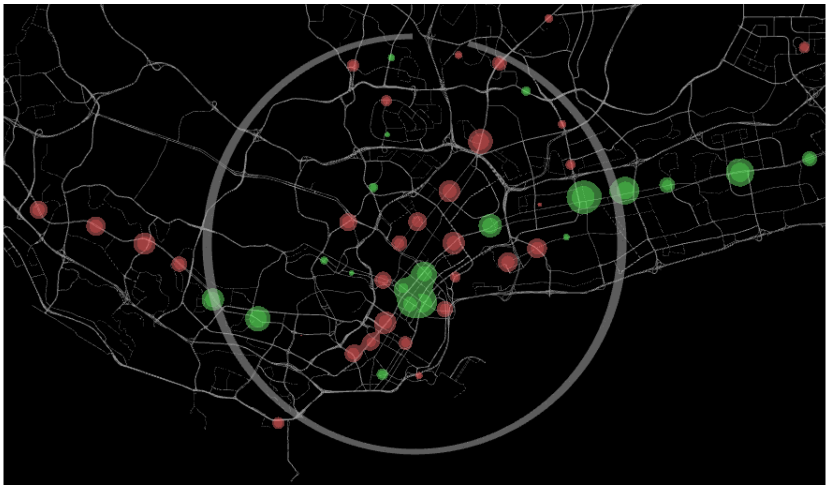


Fig. 56 Applying metaball technique to our data. Negative values are represented by red color and positive are green colored. In general view, image on top, we have distortions in areas of major density. As we zoom in, image in bottom, the radiuses become closer to the input values.

So, visual artifacts were produced based on the metaball technique, where deviation counts were represented with the size of blob's area, and positive and negative values were represented by colors. According to the Cleveland [8] color is one of the most efficient visual variables to encode categorical values. So, a positive deviation, i.e. an abnormally high number of passengers, was represented in red, while a negative deviation was represented in green. In order to improve performance the data was pushed to video card and the output was calculated using fragment and vertex shaders written in GLSL. We generate an isosurface where the position of each charge point is the GPS position of each stop transformed to the screen coordinates, and the force is the absolute value of the deviation. In this approach we used two-band threshold. Forces superior to 1 are represented with an alpha component of 0.8, forces inferior to 1 and superior to 0.5 are represented with an alpha component of 0.5. As it can be observed, forces superior to 1 tend to result in circles since the influence of other charge points tends to be negligible by comparison. This contrasts with the areas of transparent green and red generated by forces in the [0.5,1] interval, which assume a more organic nature filling in the gaps among strong forces and highlighting areas where anomalies are occurring. When we zoom in, the exaggerations caused by the representation of these small forces are reduced and the visualization becomes more rigorous.

Finally, we experimented the representation of the data by other graphical elements such as hexagons. The idea was based in calculating the color of each cell of a hexagonal grid using the mettaball algorithm, instead of calculating each pixel of the image. First we implemented the configurable grid, which cells could vary in size, Fig. 57. Then we applied the mettabal algorithm in the similar way we did for pixels, Fig. 58. Since the amount of cells is smaller than pixels the algorithm performed effectively, due to low number of calculation operations. So, this representation was implemented without the help of hardware acceleration. The resulted visual artifacts were aesthetically interesting, however the precision of the visualization decreased, due to the not meaningful extra parts added by hexagons. We could not afford that amount of visual distortion, so the further implementation relied on previously discussed model.

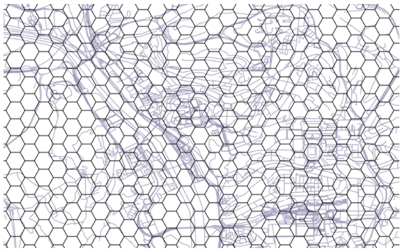


Fig. 57 Configurable hexagonal grid.

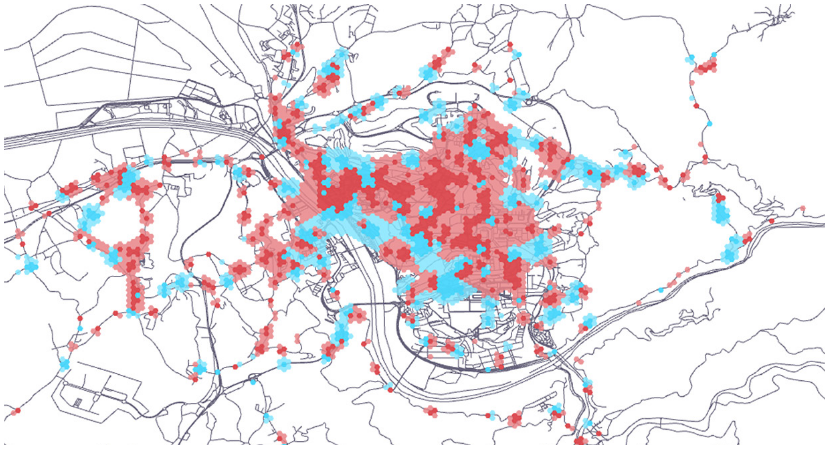


Fig. 58 The render of hexagonal Mettaballs. Red represents positive deviation and blue represents negative deviation.

Illuminated Map 4.5

This model was based on the same technique described above. The difference is that the output of the function is applied directly to the shape vertices of the map, Fig. 59. So, the road map was retrieved from OpenStreetMap and unnecessary objects were filtered. Then, the *Metaball* technique was applied to determine the color for each vertex of the map. Further interpolation of colors along the line was calculated in fragment shader. Although, this approach is less computationally expensive, in our opinion, neither the visual output is clear nor informative.

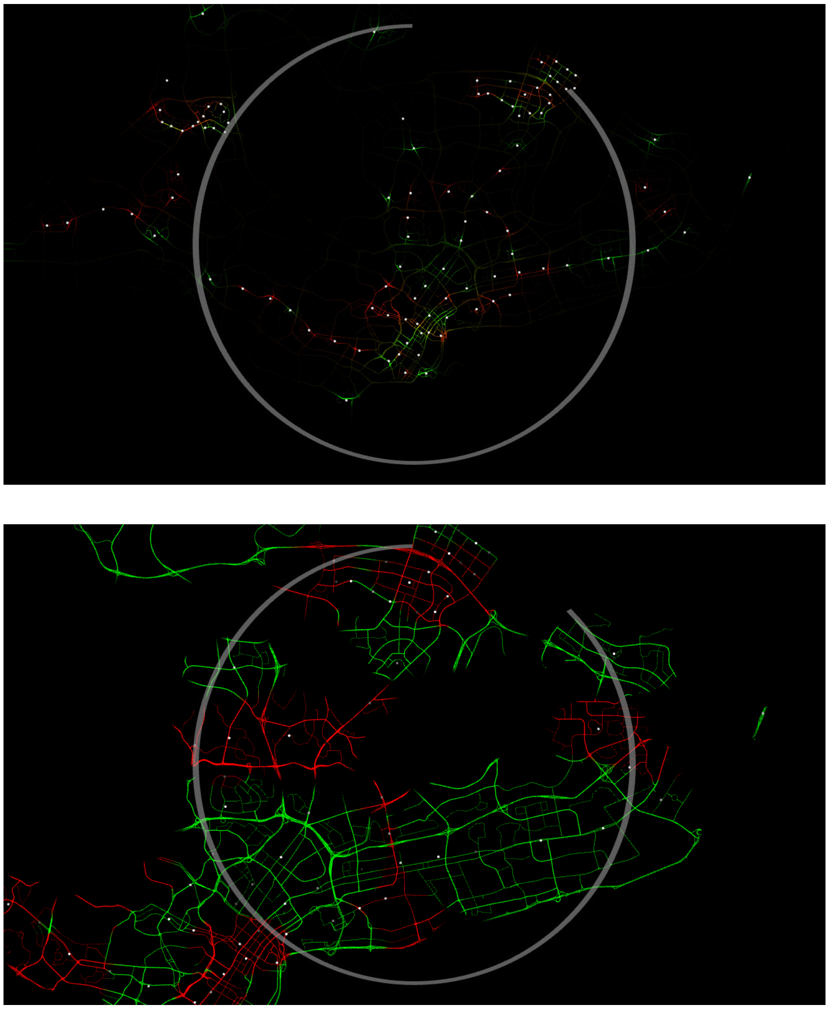


Fig. 59 Visualizations of the urban mobility created by applying the Metaball technique to colorize the vertices of the map with global view, the image on top, and zoomed view, the image in bottom

4.6 Improvement and Interaction

As we discussed earlier this two stages, refine and interact, consist in visual improvement of graphics and expanding the application with interaction methods. Yet, we wanted to validate our model on other case studies. So, the refinement and interaction stage were based on the data set artificially generated for Coimbra and another data set provided by London's TfL.

So, in order to improve visual impact of graphical elements we introduced some changes in the color palette. The colors of the bubbles retained untouched. Pure red and green, more precisely ***rgb(255, 0, 0)*** and ***rgb(0, 255, 0)*** in the interval [0, 255] respectively. We changed the background color from pure black to dark blue—***rgb(23, 24, 25)***. When the bubble's colors are mixed with the background they become softer, Fig. 60.

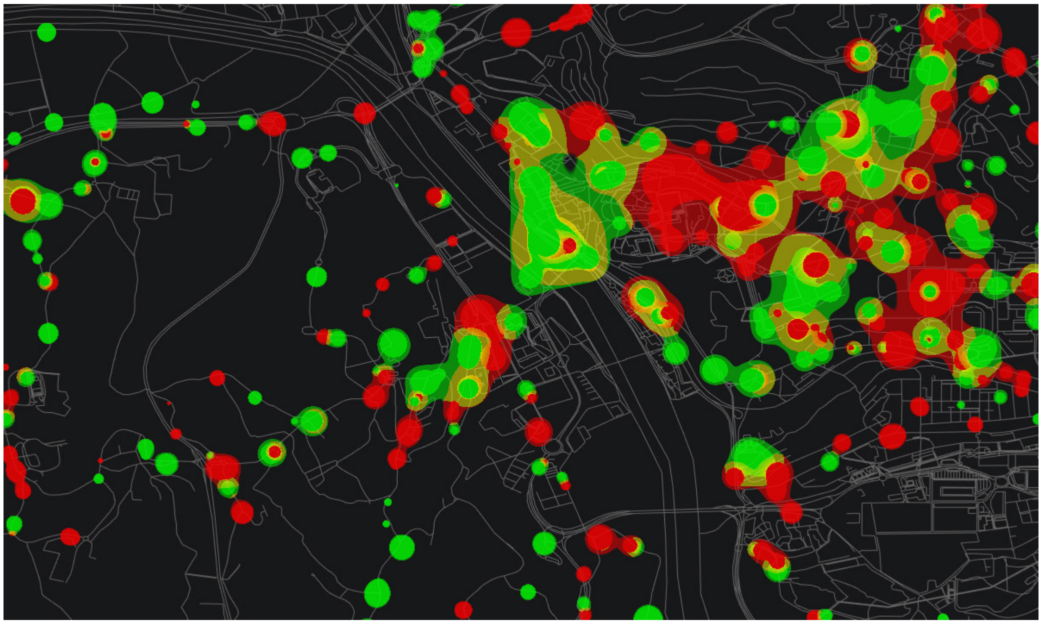


Fig. 60 Improved color palette resulted in soft, vivid blobs. The visualization is slightly zoomed.

In other hand the vector map changes its stroke width and color while zooming. The thickness of a stroke is proportionally inverse to the scale factor (*1/s*, with *s* as scale factor), constraining the width between 1px in general view and 0.01px in zoomed view. The color varies in its brightness between 40 in general view and 70 in zoomed view, with the brightness interval of [0, 100] and the variation function of *70/s*, where *s* is the scale factor. So, the map looks like the view from the airplane at night—slightly visible illuminated streets from great heights and more distinguishable from closer perspectives.

In the second model the idea of changing stroke's thickness was adopted with exaggeration. The Fig. 61 illustrates the exaggeration of segments' widths. The representation comprises two vector maps: the background map with the behavior similar described above; the overlaid vector map with applied vertex shader. The color of vertices is calculated as follows:

Due to the technical limitations, restricted amount of accepted data by shader, the model is limited to display 500 stops.

- All the vertices of vector map are sent to the graphic card, as well as position and the data of each stop.
- The vertex shader program written in GLSL takes all this information and calculates the color for every vertex using the metaball algorithm we described before.
- All the calculated data is pushed forward throughout the graphical pipeline and renders to the screen.

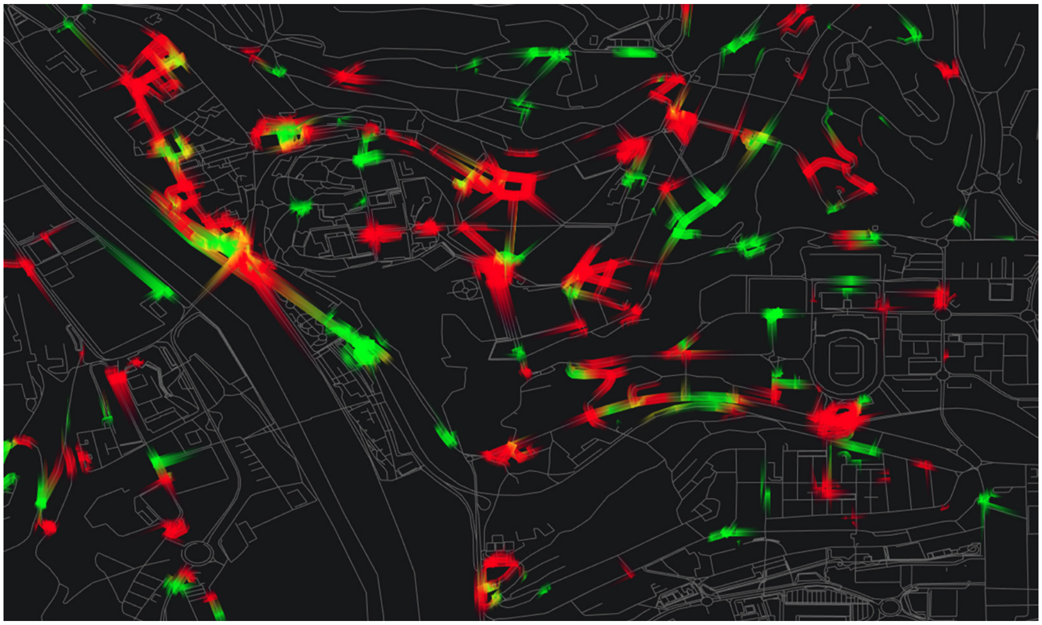


Fig. 61 The detail of zoomed view. The strokes of vector maps are exaggerated.

Before talking about interaction aspect of the application we would like to cover the changes in the metaball equation. So, while validating our model with TfL data we noticed that the amount of exaggeration was too high. The Fig. 62 illustrates the subjective parts of the representation. In order to change the exaggeration we added one variable that controls the slope of the equation. So, the equation become as follow: $(r/d^2)^b$, where ***r*** is the radius of imaginary circle, ***d*** is the distance from calculated pixel/vertex to the charge point, and ***b*** is the variable that controls the slope of function curve. As the *b* increase the resulted blobs tend to the shape of circle with radius equal to the *r*. While the low *b* values makes the blobs to be more expanded. In the example of Fig. 62 we used *b* equals to 3.0.

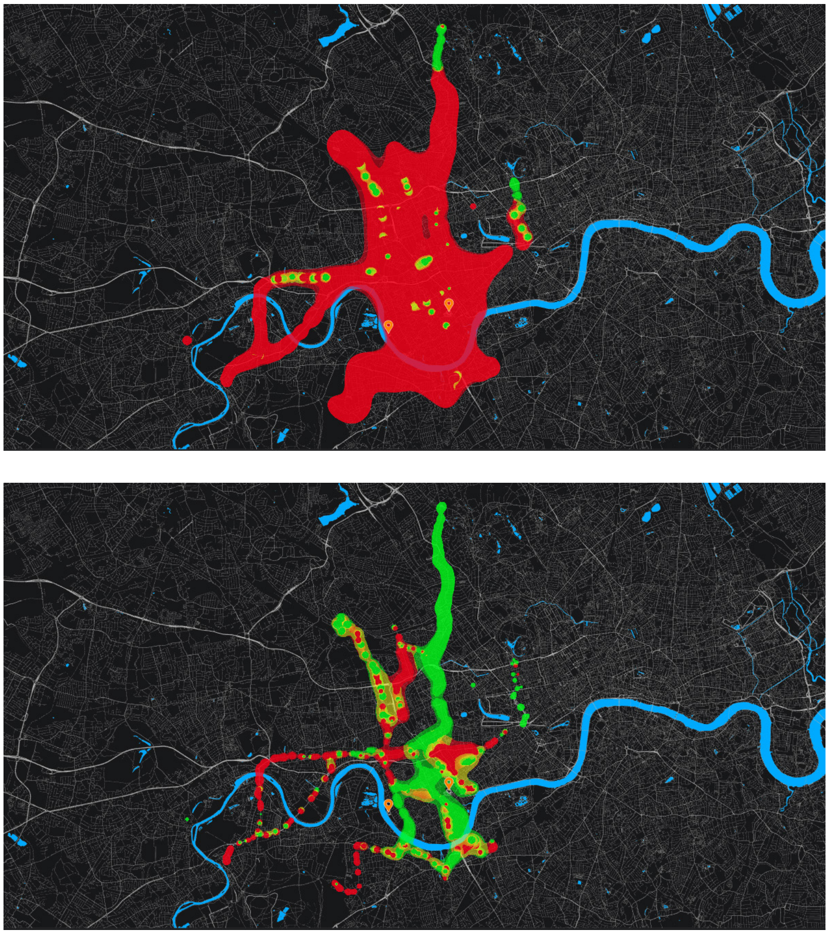


Fig. 62 Visualization of TfL data—standard deviations from normal, with controlled slope of metaball function. Image on the top shows a network load at 9.00h and image in bottom shows the deviations at 21.00h. Eight bus routs were represented.

In order to explore the data set we needed a mechanism to navigate in time. We added the panel with the timeline and additional information, located at the bottom of the application window. On left part of the panel we added labels that explain the meaning of colors and the source of data set. On the top of the same part the user can find the current date (e.g. year, month and day), as well as change the date using two buttons located at the right side. The navigation in time along the day can be done by clicking on timeline located at the right side of the panel. The selected time interval is marked by a semi-transparent line with a numerical indicator, Fig. 63.

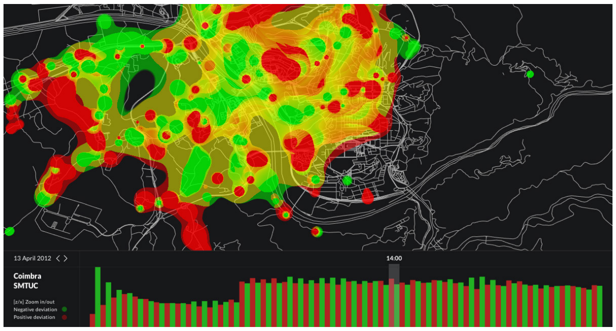


Fig. 63 Basic interaction for navigating in time. The time line is located at the bottom of the screen.

According to Tufte [7], an efficient enhancement of the exploratory power of the map representation is to add time-series description. Usually the time is represented by a line graph, since it is a continuous quantity. However, in our data set the time was divided in intervals, so we considered it as a discrete, individually separated, quantity. Thus, the graphical representation was based on a bar graph technique. Each bar in a graph represents an absolute sum of all deviations for given time interval. Yet, we aggregated the data in positive and negative deviations and represented them with red and green colors respectively. The ratio was determined from maximum and minimum peaks found in whole data set. In the first approach we displayed the absolute values, since we wanted to compare areas with abnormally high or low number of passengers, instead of comparing of the notions of positive and negative values. The two such distinct colors are efficient in the assemblage of visual patterns, however in cases of higher frequencies (e.g. when the time is divided in 15 minutes time intervals) the representation become less efficient, Fig. 64.

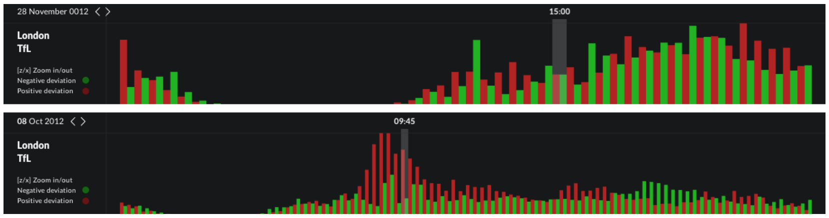


Fig. 64 Comparison of timeline, represented with bar graph, divided in 1 hour intervals, image on top, and 15 minutes intervals, image in bottom

So, we decided to project the green bars to the bottom, starting from the zero line, Fig. 65. This resolved the visual assembly problem. The graph presented all the aspects of efficient representation—easy recognition of graphical elements, visual grouping of detected elements and estimation of elevations among different elements. Moreover, the total height of a bar, red plus green bar, represents the sum of all deviations in the network. Yet, the model presents efficient table look-up aspects. For that purpose we added the reference grid in back, so that the user can visually scan to the left of the graph, estimate the distance from the origin line and match the fixed high with labels.

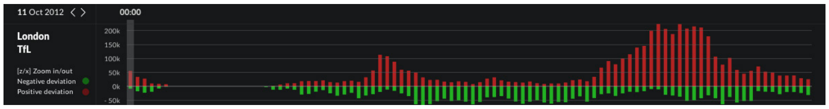


Fig. 65 Positive and negative deviations are represented with red and green bars respectively, with the reference grid in the back

In the following experiment we considered the time as a continuous quantity. In order to represent this time-series that was divided in two groups, positive and negative deviations, we relayed on the stacked line technique (similar to stacked area with except that the areas between lines was not painted). As we have seen this technique is capable to convey cumulated elements of information. The Fig. 66 illustrates the comparison between bar graph and stacked line graph that we plotted. Therefore, the graph has presented efficient representation of totals and communication of trends along the time, the abrupt changes in the data made it inaccurate, due to the rapid changes in line slope.

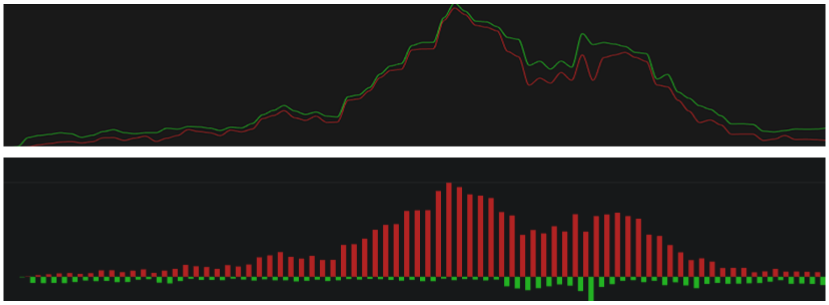


Fig. 66 Comparison of a bar graph, image at the bottom, and a stacked line graph, image on the top.

In the Fig. 67 the representation of the data for eight routes in London's bus network were arranged in a 3x4 matrix. Each display shows visual patterns in general view for given time interval, starting at 14.15h and ending 17.00h. This small multiple shows that the bus network is highly used between 14.15h and 16.00h. Then, by observing the reduction in meaningful area and appearance of green blobs, we conclude that the state of the network starts to fall to normal and low use than it was expected.

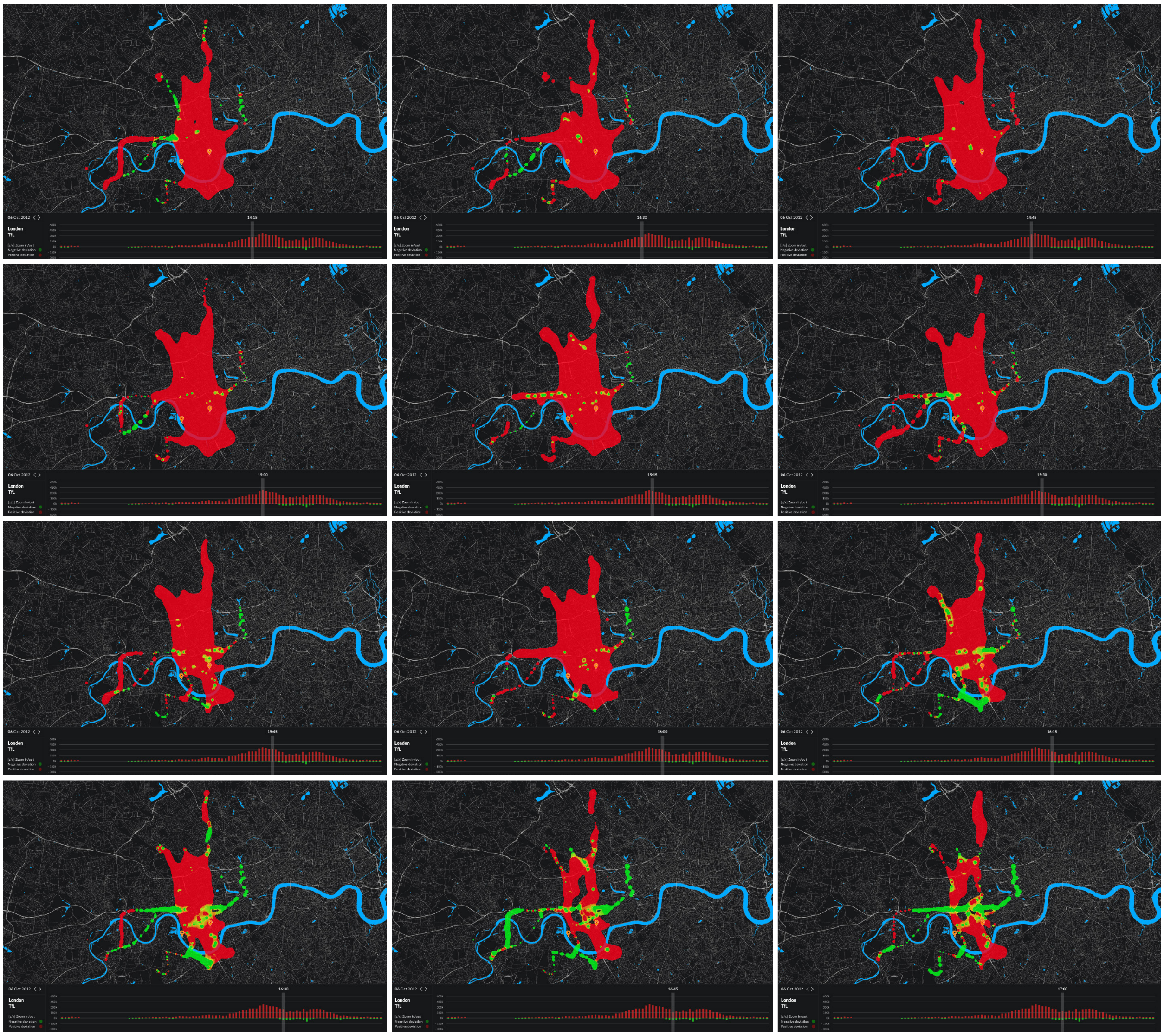


Fig. 67 The small multiple visualization of peak hours of bus network in London starting at 14.15h with time interval of 15 minutes.

4.7 Discussion

In this chapter we presented the visualization model that enables the user to explore the relationship between the crowd and urban space through the visualization of a public transport use. We started with introducing the input data that consisted in the data sets from Coimbra, Singapore and London transport networks. The data for Coimbra was artificially generated, while we awaited the data from other official sources. So, each data set was comprised of one csv file with the stops description, and another *csv* file with the data about the use of public transports. Further experiments allowed us to understand that the deviation from the normal use of urban transports, aggregated by the same weekdays, may reveal anomalies in the routine behavior of crowd.

Then we introduced the tools we used to work with graphics. The Processing framework allowed us to calculate and presents all the graphical elements to the screen. With the QGIS we filtered unnecessary objects from the road map, previously retrieved from OpenStreetMaps in shape file format. Then the map was exported to the .svg format, since the Processing provides useful tools to work with vector graphics. For example, while running the application changes the color and thickness of strokes of a vector map. The further implementation of a graph was proceeded by using the fragment and vertex shaders, programmed in GLSL language. This enables the visualization to run in real time.

The established design requirements defined the boundaries for our project. In order to reflect the geographic nature of the data the visualization should be plotted over a map. This would create a digital layer that reflects the relationship between crowd and urban space. The visual language should be simple and clear without overloading the representation. The visualization should be interactive in order to explore the time-spatial data.

Our first attempt to represent data was based on simple representation using a bubble graph. Each stop was represented as a circle, which radius corresponded to the counts of passengers. Later we demonstrated that the use of area lead to more accurate conclusions. While the bubble graph could be one of the solutions for representing our data, we observed a high amount of overlapping circles, since the stops' locations appeared to close to each other. So, we proceeded to experiment with other graphical elements such as lines.

First of all we experimented with simple graphical elements. Based on the dot map of disease we proceeded to indicate the closest streets to the stops vectors and represented the counts with lines perpendicular to those vectors. The idea and visual results appeared interesting to us, though the high amount of visual clutter made the visualization inefficient, yet, it did not work for metro station, since the metro stations belong to various streets. However, the implemented system could serve as a “skeleton” for further exploration. We could somehow embody the resulted diagrams to a dot representation, or represent with symbolic marks, which could vary according to the amount a passenger counts.

In the next step we went back to the mining stage and calculated new values that represented the anomalies in public transport use. These consisted in abnormally high or low passenger counts that were expected. So, this component was comprised of two categorical elements—positive and negative values. In this experiment we encoded our data with irregular triangles. Due to their pointing nature we encoded the positive values with the triangles pointing up and negative values with the triangles pointing down. The size of those represented the corresponding absolute value. Yet, the opacity varied according to the abnormally high amounts—these appeared totally opaque, while the low values were represented with almost transparent triangles. This quick experiment led us to the conclusion that our data should be represented using more sophisticated mechanisms.

Further experimentation led us to the conclusion that the zoom technique can be used to transform the representation from one state to another or to morph from one visualization model to another. In detail we discussed a transformation of the same model, which in each state efficiently represent the data. However, we did not cover the transformation with morphing, which must be studied separately, since it involves the development of methodology and various visualization models that share same characteristics.

So, we introduced a metaball based visualization model. The developed representation presents strong pattern perception characteristics at general views and accurate estimation of information at zoomed views. Yet, we discussed the subjectivity added by the model at general view. Due to the nature of the metaball equation extra pixels are added to the visualization, which creates a blobby aspect of representation. Since the calculation is computationally expensive, we relayed on hardware acceleration by using GPU and shaders to produce visual artifacts. This work was presented at SIGGRAPH conference [46]. Finally, we developed an application that uses the developed visualization model with supplementary graphic. This represented the data regarding to the time. Using a bar graph we visualized summed positive and negative deviations aggregated by time intervals.

Future work will focus mainly on exploring the transformation of various visualization models from one to another using morphing techniques. This will enable the visualization of geo-referenced data to be as efficient as possible, since different visualization models that suites best for each view will be used. Yet, the developed application will be enhanced with supplementary graphics, such as sparklines, that will display the time based information per stop along with its description.

Chapter 5
Land Use

This chapter starts with the description of given data, used tools, implementation and established design requirements. Then, the first section describes the examples of visualization of clusters of semantically enriched points of interest. The second section presents diverse visual techniques which illustrates the quality of automated classification of points of interest.

Data

This project was based on the semantic information of places of urban area. Like in urban mobility project the data was previously acquired, filtered and mined by other members of the team. So, basically the row data consisted in so called digital footprints, the precise data in terms of spatial and temporal location [39]. In the earlier work, in order to understand people's activity, the visited places, aka points of interest (POIs), were semantically enriched, due to the lack of information about these places. Since, places are often associated by its meaning (e.g. the user's relationship with place or its physical properties), this information creates an images of the city from user's perspective. So, the semantic information was retrieved from diverse web sources and grouped in clusters of POIs using clustering methods developed by members of CROWDS project [40, 41].

So, when we started to work with the data, it has the following format: cluster ID, its centroid's latitude and longitude, data related to the belonging POIs (ID, latitude, longitude, name, source), name of a cluster and its relevance. Yes, all clusters were weighted by mechanisms developed by researchers of CROWDS project [39]. So, the information consisted mostly in nominal components, except the weight of clusters, which is quantitative variable, and geographical location of its centroids and its constituent POIs.

The second part of the project was based on the automatic classification of POIs. Basically, the visualization should compare the census data and the output from automated classifier in NAICS* taxonomy standard [45]. The data consisted in ID of POIs, its name, latitude and longitude, the belonging category, true NAICS and predicted NAICS.

***North American Industry Classification System (NAICS) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. (ref. U. C. Bureau. North american industry classification system (naics): Introduction, February 2010. http://www.census.gov/eos/www/naics/.)**

Tools, implementation and design requirements

In this project we used already introduced tools, such as Processing 2.0b8 and its libraries, QGIS and shape files retrieved from OpenStreetMaps. The vector map of Boston area was saved in svg format, due to the reasons described in previous chapter. The data was saved in tables of csv file. All of experiments were implemented in Processing and Java language (e.g. custom libraries for working with geometry), as well as custom library with integration of Java Map Projection library, that provided simplified API for projecting GPS coordinates to the screen. Basically, it takes geographical bounding box (latitude and longitude of top-left and bottom-right corners) and window's width and height as arguments, and then it outputs the X and Y in windows coordinate system of projected latitude and longitude of a point. Regarding to the design requirements, they are similar to the urban mobility project: simple and clear visual language; run in real-time; displayed on a map; interactive. Moreover, all of the represented nominal data must be as readable as possible, in order to efficiently convey the information.

The design requirements for the second part of the project are slightly different. The visual artifacts are intended to be printable in A4 white sheet of paper, since they served as explanatory graphics. So, they should be static and should use simple visual language, in order to be readable and understandable at small sizes. The produced visuals should be clear and should convey only essential information in order to avoid visual clutter. They should present a good tradeoff between accuracy of visualization and quantity of transmitted information.

Visualization of Clusters of POIs 5.1

This section covers all the process of development visualization of semantic information. First, basic representation is presented. Then the detailed process and the algorithm to find a shape of clusters of POIs is described. Following subsection presents a method to smooth corners of a polygon. Finally, a typographic weight as a visual variable and visual improvements are presented and discussed.

Basic representation 5.1.1

In order to understand the distribution of points of interest in space and within the corresponding clusters, we plotted them on the map. Yet, we wanted to visualize the distribution of the most relevant areas on the map. So, first we represented the names of categories, where the center of the word corresponded to the projected centroid's location and the size of the word corresponded to the relevance of the category. The size was linearly mapped to the weights of clusters. So, we observed that, due to the high density of data the resulted in huge amount of overlapping words, which made the representation looked more like chaotic gray blot, Fig. 68.

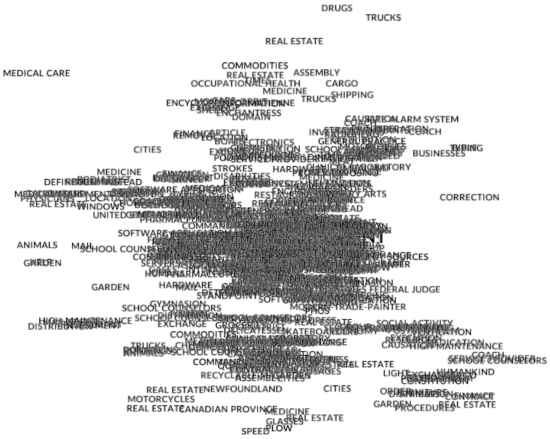


Fig. 68 Basic representation of names of categories, where the size of words represents its relevance.

In order to resolve that issue we proceeded to implement a mechanism that detected the overlapped words and regarding to the relevance of category displayed the one with biggest weight, Fig. 69. This approach resulted in more clear basic representation, which led us to discover that there was one category with the highest relevance that made other words to appear with more or less equal sizes—linear scale problem. Moreover, due to the high amount of words, more precisely 751 clusters' names, the complexity, which is approximately 564001 intersection calculation operations per frame, made the visualization perform slowly (6.5 vs. 60 fps, with and without intersection detection correspondingly).

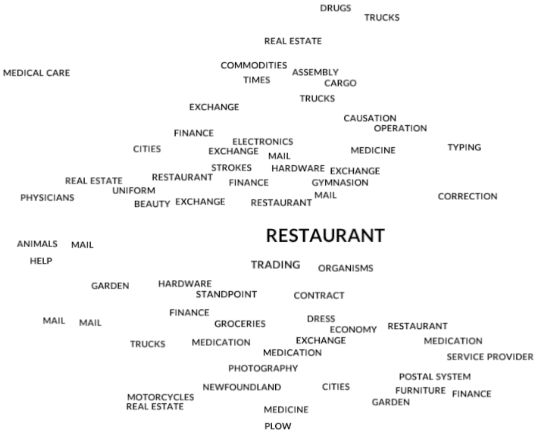


Fig. 69 Basic representation of names of categories, with intersection detection mechanism.

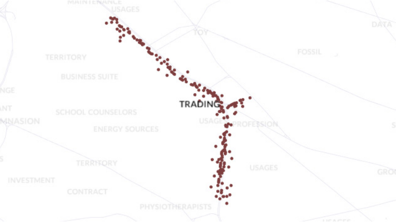


Fig. 70 POIs of the Trading category.

Next we represented points of interest on the map, where each POI was encoded with a small circle and a color, which distinguished points belonging to different groups, Fig. 71. We observed that each cluster have interesting and organic shape. For example, on the Fig. 70 POIs of the Trading category forms a geo-referenced shape that uniquely identifies that particular cluster. Like countries or continents of the world have its own unique geographic shape to which are associated diverse meanings and symbolisms. We found this idea particularly interesting and proceeded with the implementation of mechanisms that would allow us to calculate shapes of clusters and use them for further experiments.

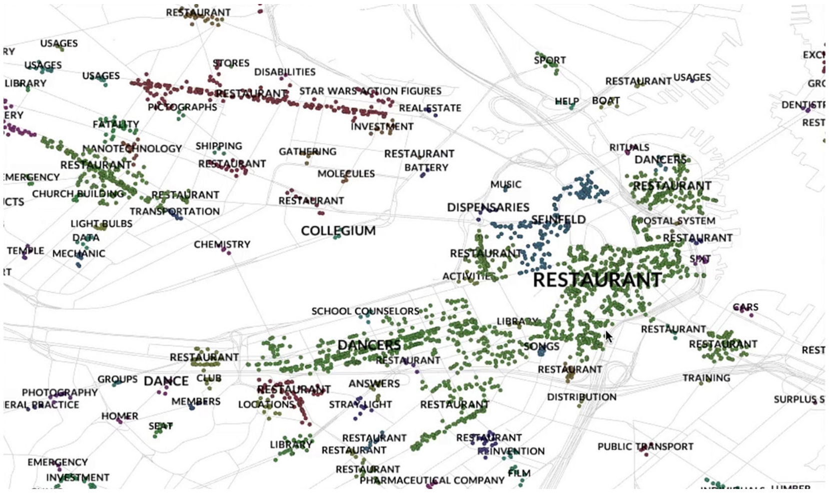


Fig. 71 Basic representation of names of categories, as well as POIs. Points of equal color belong to the same group.

Shape of a cluster 5.1.2

In order to find so called “*hull*” of a set of points in a real vector space we relayed on convex hull technique. So, basically, the convex hull can be visualized using a rubber band analogy, Fig. 72. This is the minimal convex set of points containing in a whole set. We calculated the convex hull using the Mesh Processing library [42] as follows: we randomly generated a set of points x in a window space; the library received the x as an input; then outputs a set of points that defined the convex hull. The visual output is illustrated on Fig. 73.

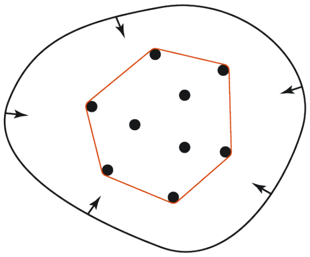


Fig. 72 Convex hull, red line, or convex envelop of a set of points as a rubber band analogy.

The result was not that we were looking for, since it did not reflect an organic shape of a set of points. So, we proceeded to use a non-convex, aka concave, hull algorithm, more precisely an alpha-shape with help of JTS Java library [43]. Basically, alpha-shape calculates a polygon which embraces all the homogeneously distributed points, but with less area compared to the convex hull. The figure x visualizes a concave hull with the red line and convex hull with the blue line (diagram). So, we applied the alpha-shape algorithm to calculate a concave hull o randomly generated set of points, Fig. 74, image on the left. The visual output was much better comparing to the Fig. 73. However, when applied to a non-homogeneous distribution of points or to the sets with low density of points, the output may consist in holed polygon, images on the center and on the right of Fig. 74.

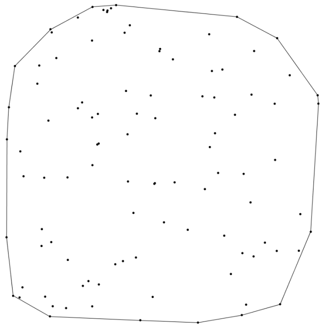


Fig. 73 Visual output of the convex hull algorithm applied for a randomly generated set of points.

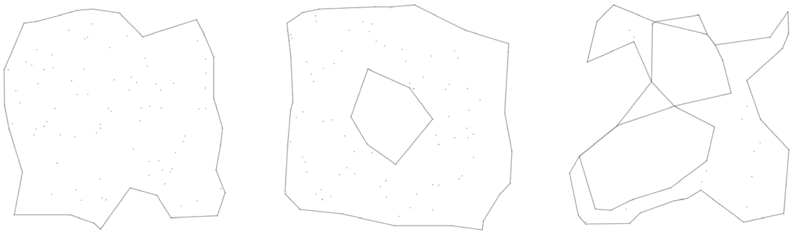


Fig. 74 The output from alpha-shape algorithm applied to randomly generated homogeneously distributed points, image on the left, 100 points with resulted holed polygon, image in the center, and 50 points with resulted holed polygon, image on the right.

The next experiment was based on already described metaball algorithm in combination with blob detection. So, again we randomly generated points on the screen and calculated an isosurface. Then using blob detection mechanisms provided by blobDetection processing library [44] we calculated the hull polygon with different threshold values. On the Fig. 75 we can see the visual output using different thresholds. The resulted hulls were what we expected—nice, soft edged and organic shapes. However, in order to calculate a single polygon the threshold must be controlled manually, since the high values of threshold results in multi-polygonal shape. Due to the high amount of clusters of POIs it would be time consuming and inefficient.

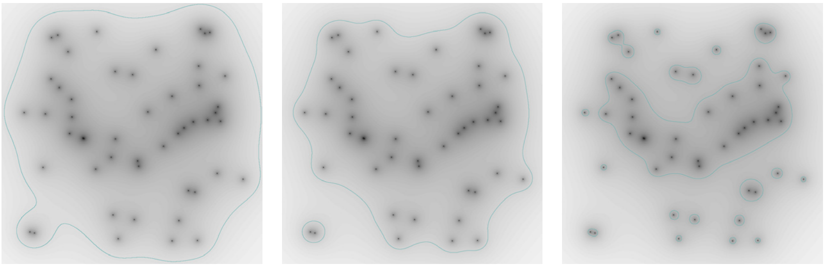


Fig. 75 The visual output from the combination isosurface and blob detection techniques. Thresholds are 0.15, 0.18 and 0.28 in the order corresponding to the order of images.

What we needed was a totally automated algorithm that receives a set of points and calculates a shape of every cluster of POIs in a set. For that reason we proceeded to develop our algorithm, based on simulation of rubber band, or if visualized in three-dimensional space a plastic bag with vacuum inside which makes the bag to embrace all the fixed points until it is totally stretched.

So, let the set of points *X* to be our input, then the algorithm is as follows:

1. Defined an empty array list that will contain the points that define a hull, say *L*.
2. Calculated a convex hull that defines the initial set of points that are appended to *L*.
3. Each edge between two points in *L* is divided by half at the center.
4. It defines a triangle with *A*, *B* and *C* corners, where *A* is the starting point, *B* is the ending point and *C* is the central point.
5. For every iteration (the number of total iterations is defined by the user):

- ▹ *C* is pushed forward perpendicularly to the segment *AB*.
- ▹ The pushing force *F* varies accordingly to the length of the edge. i.e. short edges have low *F*.
- ▹ If one of the points, say *P*, in *X* is inside the triangle, then the segment *AB* is deleted and two new edges, *AP* and *PB*, are appended to the *L*.
- ▹ The process for new defined edges starts from stage 3. or stops if their lengths are equal or smaller than was defined by the user.

Detection of point *P* in triangle *ABC* is done by calculating the cross products of vectors *AP* and *AB*, *BP* and *BC*, *CP* and *CA*. If all values minor than zero, then the point is inside the triangle. If there is more than one point inside the triangle, it is considered the clossest point to the central point *C*. The algorithm is summarized in the Fig. 76.

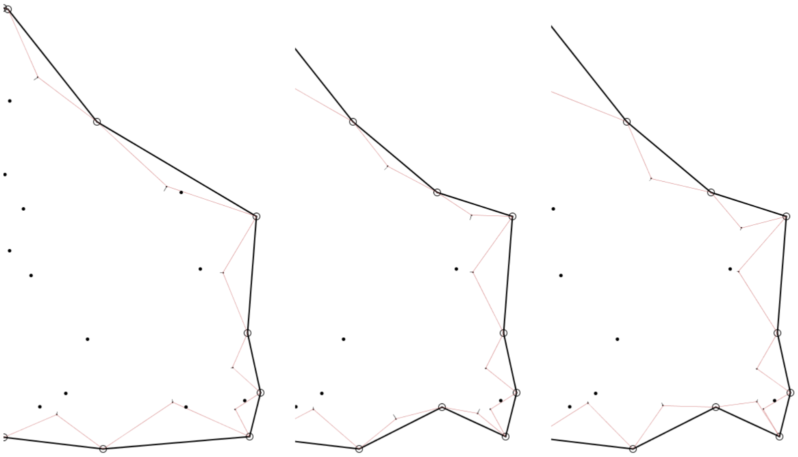


Fig. 76 The illustration of the rubber hull algorithm, progressing from left to right.

So, the visual result was exactly what we were looking for. In the Fig. 77 are represented two shapes of clusters are represented from our data set. We observed that even complex shapes are correctly defined, Fig. 77, image on the right.

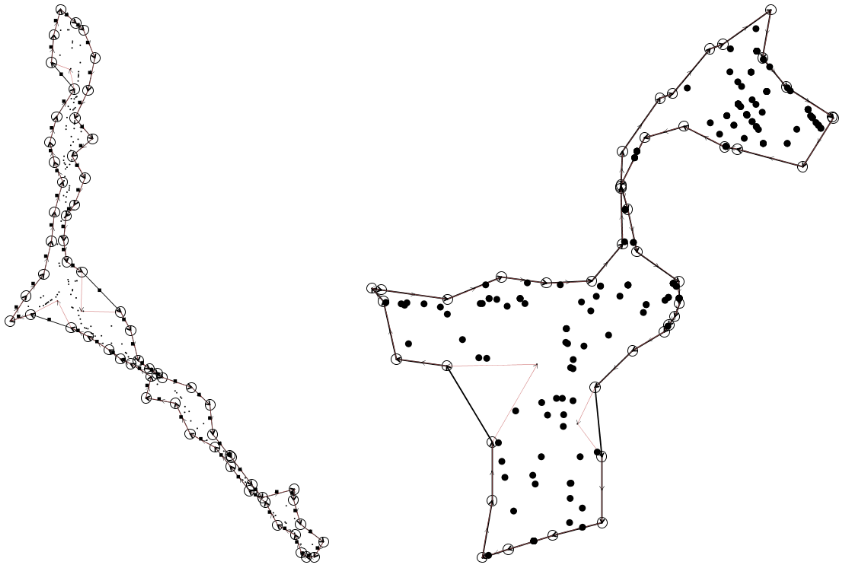


Fig. 77 Calculated shape of clusters of "Trading", image on the left, and "Seinfeed" categories, the image on the right. Circumferences indicate points that compose a hull, with the arrows inside that indicate the order of points.

The next step consisted in smoothing corners of a shape to be similar to the metaball based results. Our first idea on how to round corners is illustrated in Fig. 78. However, when we implemented the algorithm we noted that in cases when acute angles was too closed the polygon was calculated incorrectly. So, we proceeded to calculate acute angles using bezier equation. The result, illustrated on the Fig. 79, shows the smooth bezier lines that describe acute angles. Yet, the representation becomes more organic comparing to the purely geometric construction.

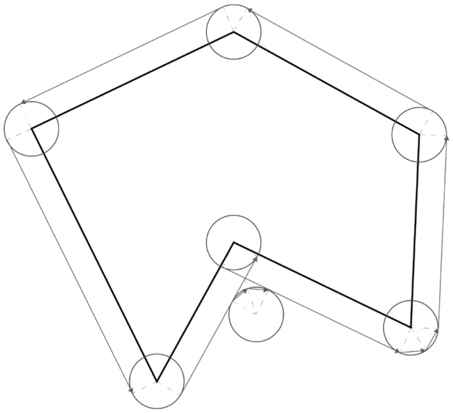


Fig. 78 Schematic representation of round corners construction.

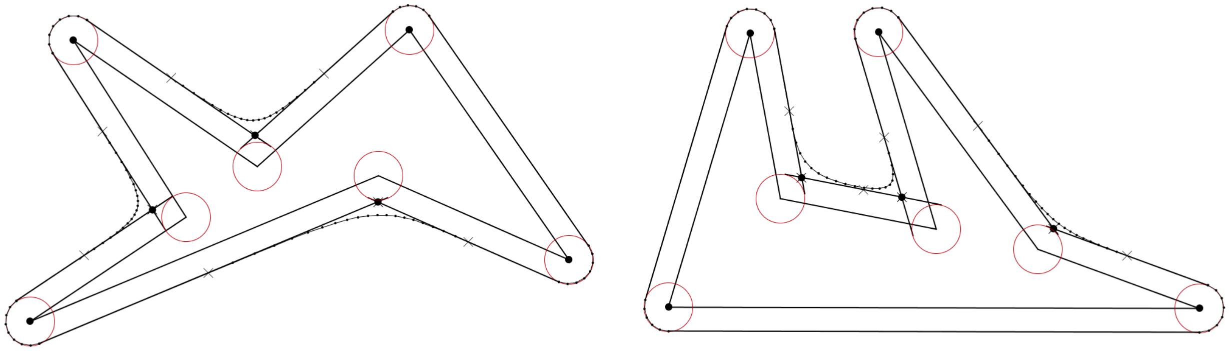


Fig. 79 Construction of round corners with bezier equation for acute angles. Image on the right demonstrates that even in extreme cases the algorithm performs efficienly.

While combining these two methods we pre-calculated the shapes of our clusters. Some of them are displayed on the Fig. 80. Then, we plotted these on a map with names of categories, Fig. 82.

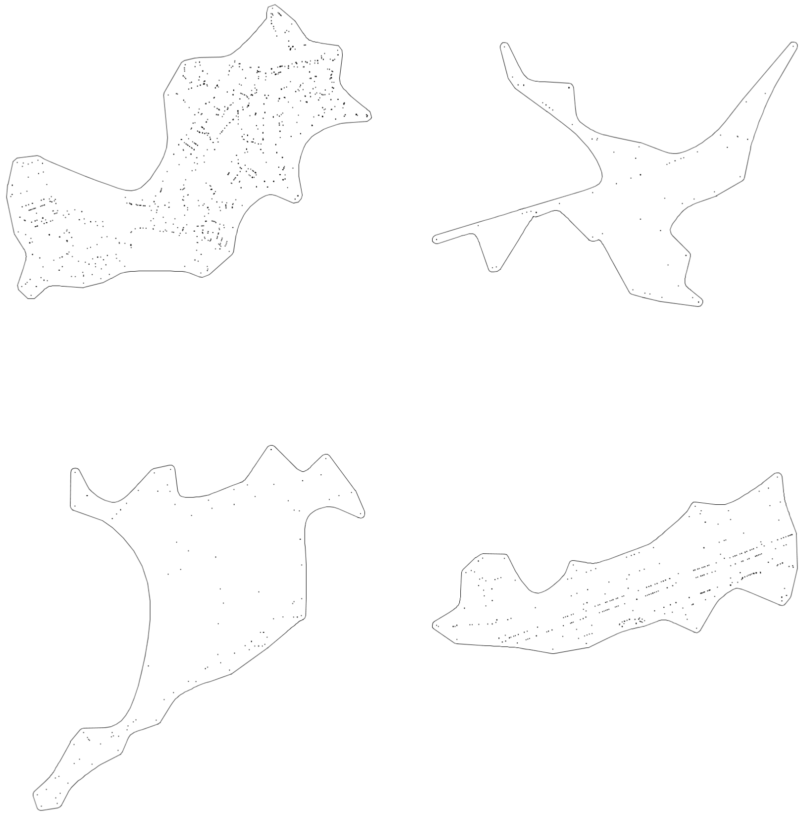


Fig. 80 Some of the examples of resulted geometries.

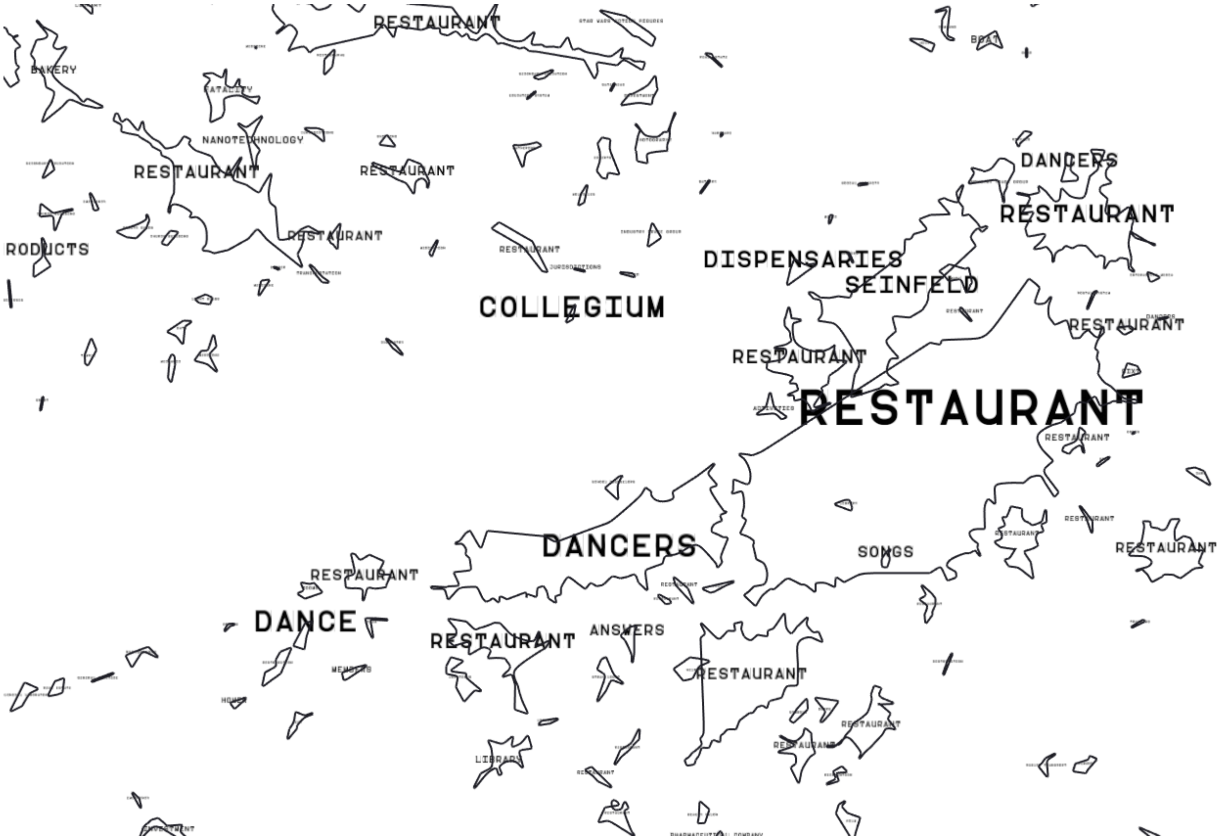


Fig. 82 Visualization of shapes of clusters and names of corresponding category on the map.

5.1.3 More Semantic Information

So, a shape of clusters and POIs represented with points on their own did not tell us enough about meaning of places. For that reason we decided to add more semantic information to our visualization—names of POIs. Having defined shapes we decided to use them as containers for graphical elements. This approach consisted in placing words on contour of a shape. So, basically we put every character on the path perpendicularly to the corresponding segment. When a letter appear on a joint of two segments it uses a weighted angle depending on the percentage of occupied space on each of segments, Fig. 83 demonstrates the construction of a graphic and Fig. 84 illustrates the representation of one cluster.

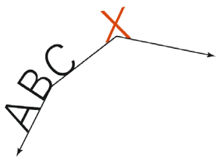
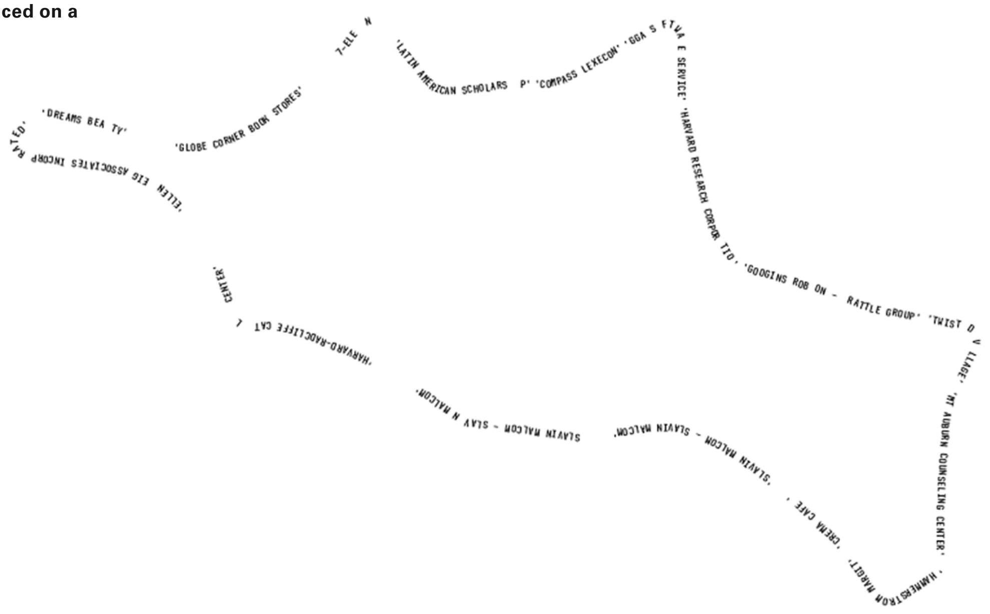


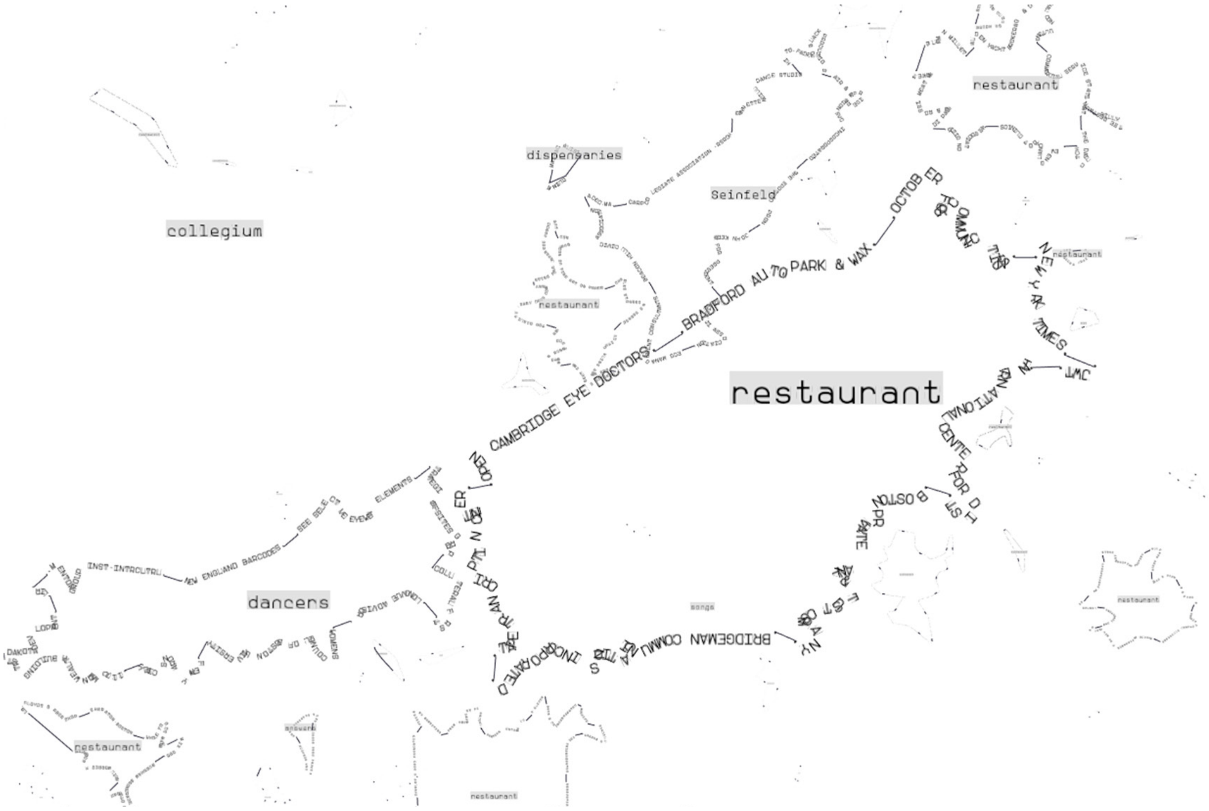
Fig. 83 Illustration of construction of words that organically accompanying a corrstponding path.

Fig. 84 Names of POIs placed on a polygon's guiding path.


'DREAMS BEA TY'
'GLOBE CORNER BOOK STORES'
7-ELE N
'LATIN AMERICAN SCHOLARS P' 'COMPASS LEXECON' 'GGA S ETVA E SERVICE' 'HARVARD RESEARCH CORPOR TIO' 'GOOGINS ROB ON - RATTLE GROUP' 'TWIST O V LLAGE' 'MT AUBURN COUNSELING CENTER'
'HAMMERSTROM HARGIT' 'CREM CAFE'
'SLAVIN MALCOM - SLAVIN MALCOM' 'SLAVIN MALCOM - SLAV N MALCOM'
'HARVARD-RADCLIFFE CAT L CENTER'
'ELLEN EIG ASSOCIATES INCORP RATED'

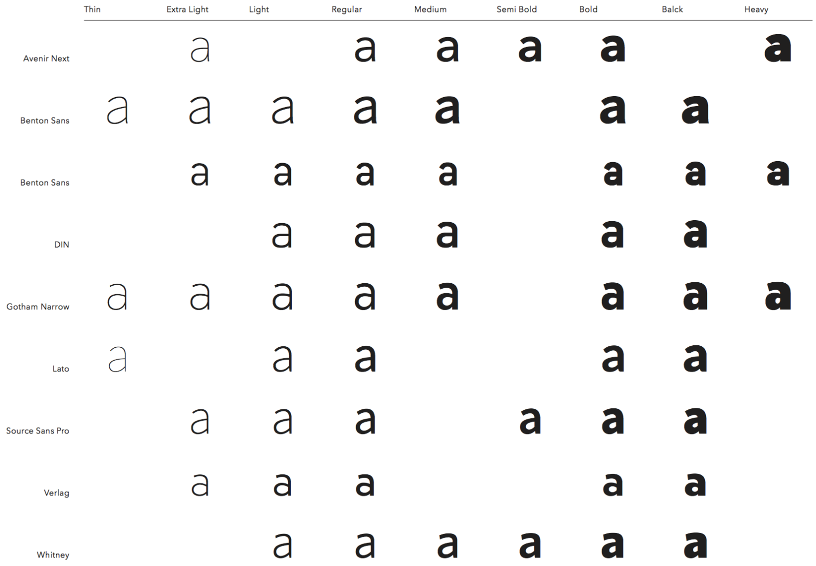
Example on the Fig. 85 is the resulted visualization as the combination of all methods described above: names of categories with different sizes; organic geo-referenced shapes of clusters; names of POIs placed on a contour of corresponding shapes. Yet, the size of POI names also vary according to the relevance of a category.

Fig. 85 Representation of POIs names with varying, accordingly to the relevance of category, size.



5.1.4 Improving representation

With all techniques described above we have significantly reduced the amount of visual clutter, namely at closer views. However, in general view words with large sizes continued to occupy huge amount of space. For that reason we proceeded to use other visual variable to represent names of categories. Within the typographic system there are different font weights—starting with thin (or it variations like hair) and ending with extra black or similar. Since the beginning we relayed on the sans serif typefaces, since it presents better legibility on the screen. So, we chose type families that presented high variation of weights. The Fig. 86 shows the chosen typefaces arranged in a table. The one that presented most variety, eight weights, was Gotham Narrow. In other words we had a visual variable with the length of eight.



	Thin	Extra Light	Light	Regular	Medium	Semi Bold	Bold	Balck	Heavy
Avenir Next		a		a	a	a	a		a
Benton Sans	a	a	a	a	a		a	a	
Benton Sans		a	a	a	a		a	a	a
DIN			a	a	a		a	a	
Gotham Narrow	a	a	a	a	a		a	a	a
Lato	a		a	a			a	a	
Source Sans Pro		a	a	a		a	a	a	
Verlag		a	a	a			a	a	
Whitney			a	a	a	a	a	a	

Fig. 86 Sans serif typeface families arranged in table by weight.

In order to use this visual variable we needed to transform our data from quantitative to categorical. This was done by dividing whole gamut of values in eight ranges. Then every range was encoded with typeface weight, i.e. names of categories with low relevance were represented with Gotham Narrow Thin, Fig. 87.

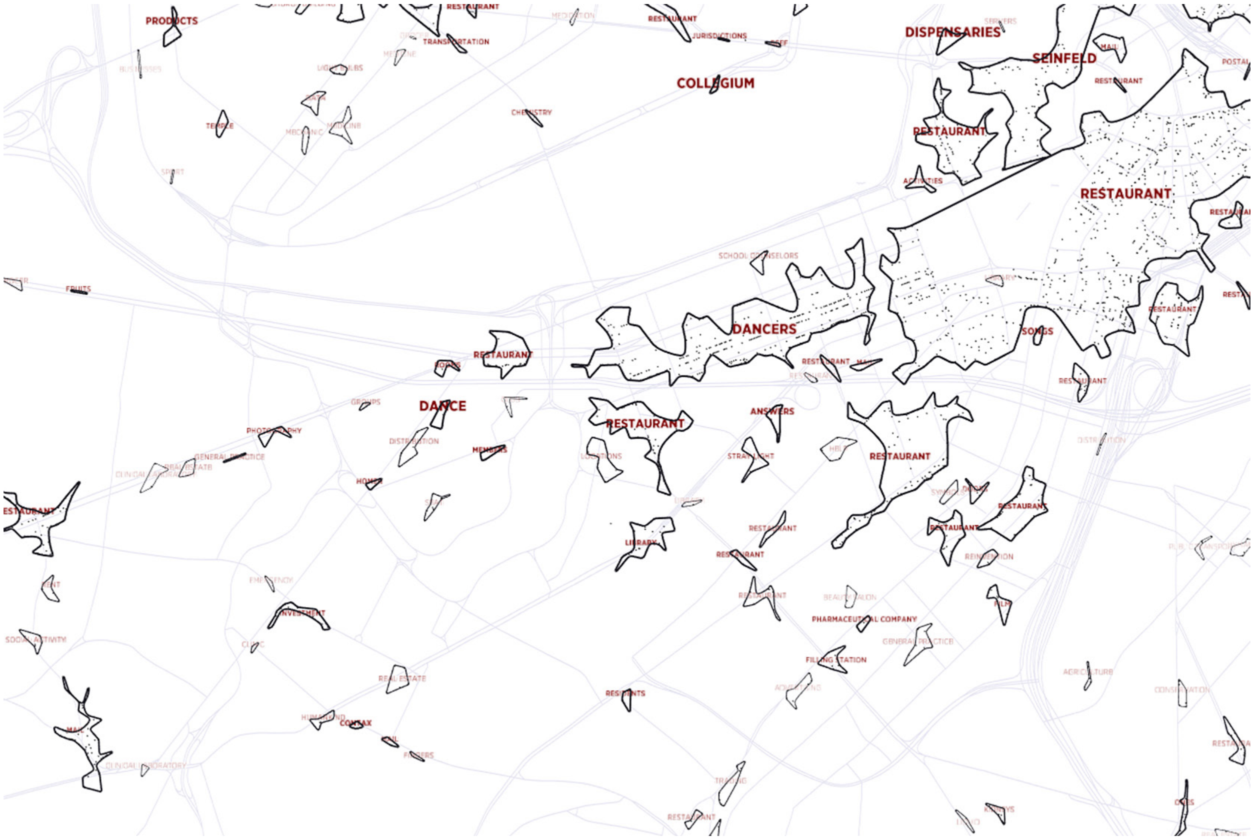


Fig. 87 Category names have equal sizes with varying weights of the typeface, according to the relevance of corresponding category.

Another small improvement in efficiency of visualization was made by adding a color to represent different categories, since there are cases with more than one shape that belong to the same category. The names of POIs have no longer represented by size. A constant minimal size of 10pt was defined, so that the names could be readable, Fig. 88.

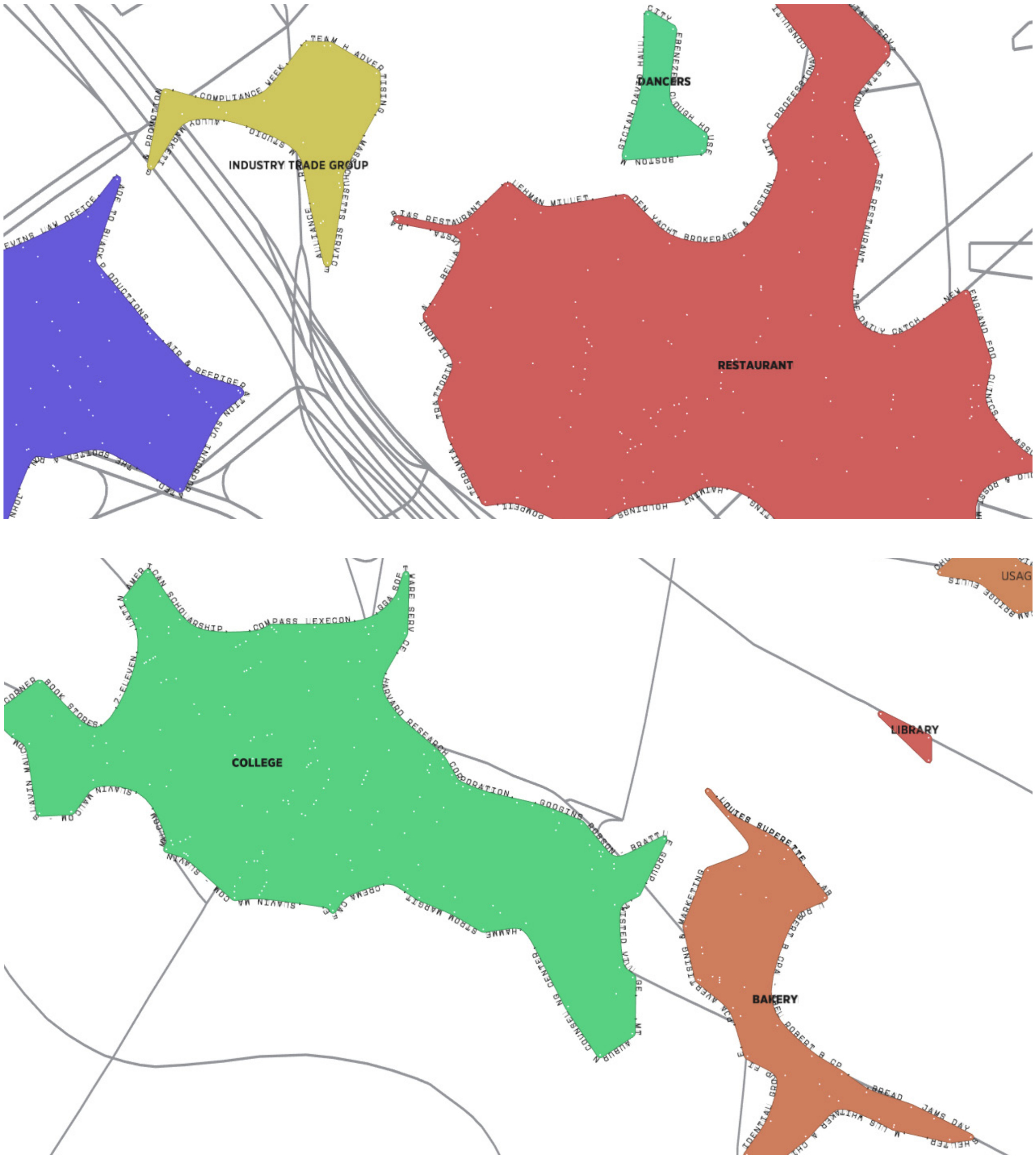


Fig. 88 Representation of names of categories, shapes of clusters and POIs' titles. Clusters of the same category are represented with equal color.

Visualization of automated classification of POIs 5.2

In land use analysis, which is a central pillar in urban planning, if the POIs do not share a common taxonomy, than the whole analysis is unreliable, since POI classification from one source may differ with one from other. In this project we explored visualization techniques to understand and identify possible ways to improve the performance of the developed models.

So, in this project we pretended to develop visual techniques that allow researchers to compare the data from census surveys (e.g. InfoUSA, Yahoo!) and the classification data that was produced by their classifier.

On the other hand, we had estimated employment data (size and location), densities at the Block level, retrieved from InfoUSA and Yahoo![4]. This data is the application of the classified POIs to the urban modeling task. So, the next step of the project was a representation of the difference between two data sets (InfoUSA vs Yahoo!). At the final phase we compared these two visualizations, classification accuracy and estimated employment densities, in order to understand if one influenced the other.

Basic representation 5.2.1

We started by visualizing the errors of the learned POI classifier. Due to the hierarchical structure of the NAICS taxonomy, there are six levels of error, considering a level 0 as no error. Therefore, we define the error level of a POI, say P, as the first index, from left to right, where the predicted NAICS code is different from the true one. So, for example, if the classifier mislabels a POI from NAICS 921130 as 921120, we say it made a level 2 error. On the other hand, if the classifier mislabels it as 238320, we say it made a level 6 error. Correctly classified POIs are said to have an error level of 0 [45]. In order to understand the density and distribution of points in space we started with basic representation, direct mapping of values. The POIs were represented with points of different color hues, ranged from 0 to 150, depending on error level, Fig. 89.

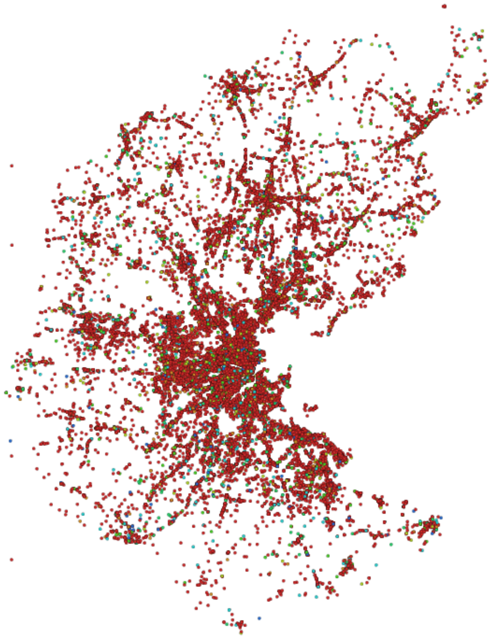


Fig. 89 Direct mapping of POIs and its error values. Color ranges between red and green. Most red points represent POIs with no error in classification and most green ones represent POIs with maximum error.

5.2.2 Color and size

The first analysis showed that there were almost no errors in classification, which in comparison to data set was incorrect. In order to better understand the error distribution we added one more visual variable to represent the error — area of circle, Fig. 90.

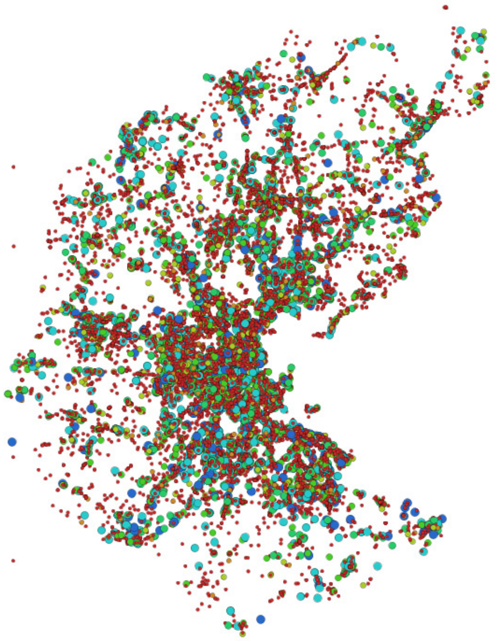


Fig. 90 Visualization with two visual variables for representation of classification error. Distribution ranges from red small to big bluish-green circles.

This visualization model still was not that we were expected. We needed to find a way to augment visual accuracy. With this in mind we reanalyzed given information in more detail and concluded that the error variable was comprised by ordinal values, rather than quantitative data. From this observation we ordered layers of POIs representation aggregated by error level, i.e. POIs with no error in classification was placed on the bottom layer. As the result become clearer we decided to remove the forth variable—area of a circle. Furthermore, we inverted the colors, due to the color symbolism, i.e. green means no error and red means bad classification, Fig. 91.

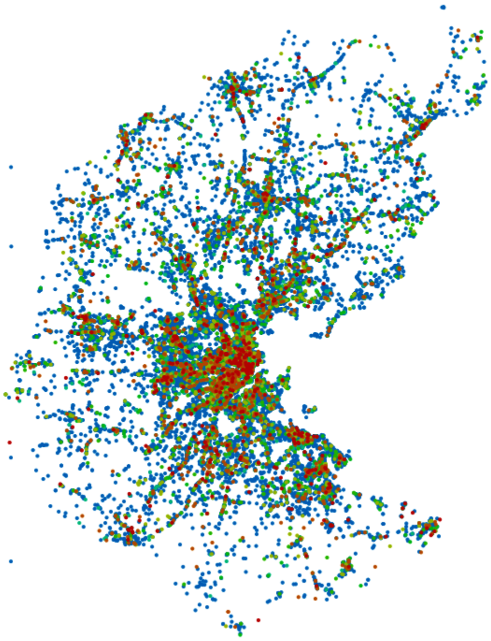


Fig. 91 Visualization with ordered layers of POIs grouped by error value. The colors ranges from bluish-green to red points representing POIs with most and no error corresponding.

Finally, we experimented with shades of gray. POIs with low classification error were represented with light gray, almost white color, and POIs with high classification error were represented with black color. Yet, we lowered the opacity to 50%, in order to bottommost layer be visible, Fig. 93. That improved the aesthetic aspect and efficiency of the representation.

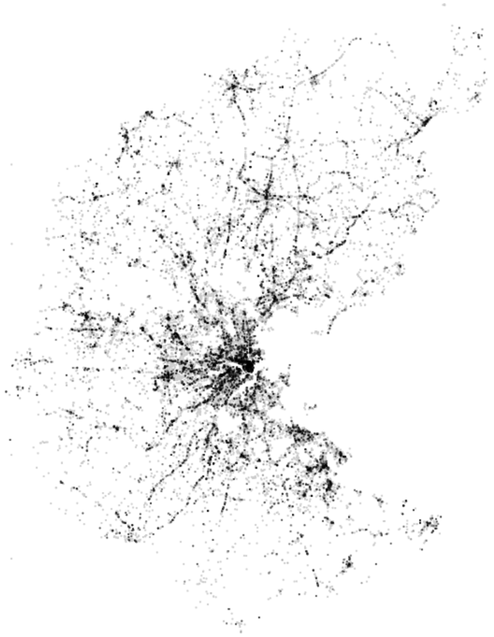


Fig. 93 Use of gray shades, instead of using hue values.

Further improvement of colors made this visualization model in this case the most efficient, Fig. 92. Here the colors varied from dark blue, through the light green, until the orange. Finally, a label was added to the visualization.

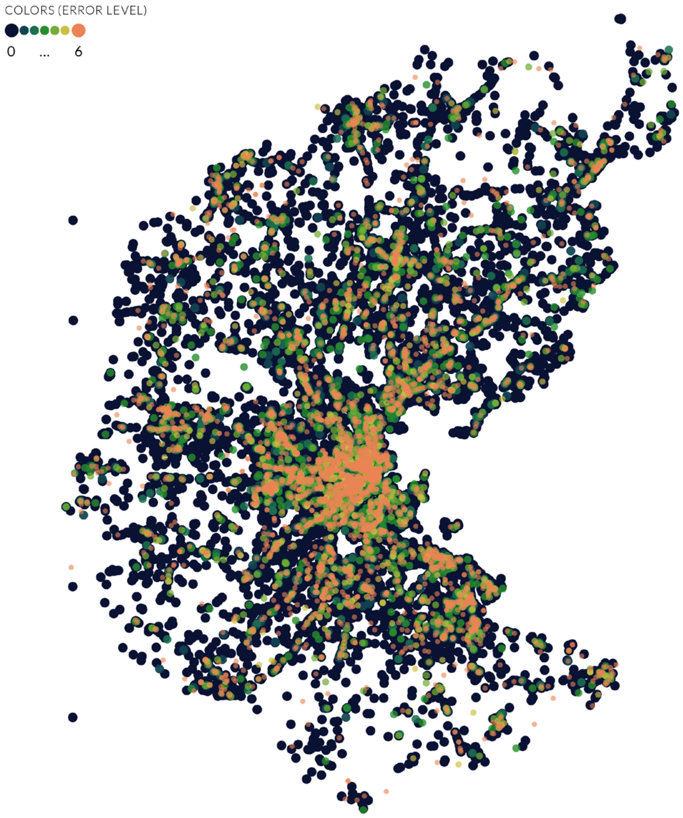


Fig. 92 Improved visualization model with the physical label added.

Small multiple 5.2.3

The above methods led us to the idea of using small multiples. As we have seen earlier, this technique is usually used to compare various versions of data sets with the same structure in order to visualize shifts in in data [7]. For that reason, we separated layers and composed them in sequenced order— first figure represented POIs with no errors in classification and the last represents POIs with error of level 6, Fig. 94. The analysis showed us that this representation could be more efficient by representing the percentage, or absolute quantities, of POIs with different error level. Yet, we observed was that the error was all most homogeneously distributed throughout the urban area of Boston.

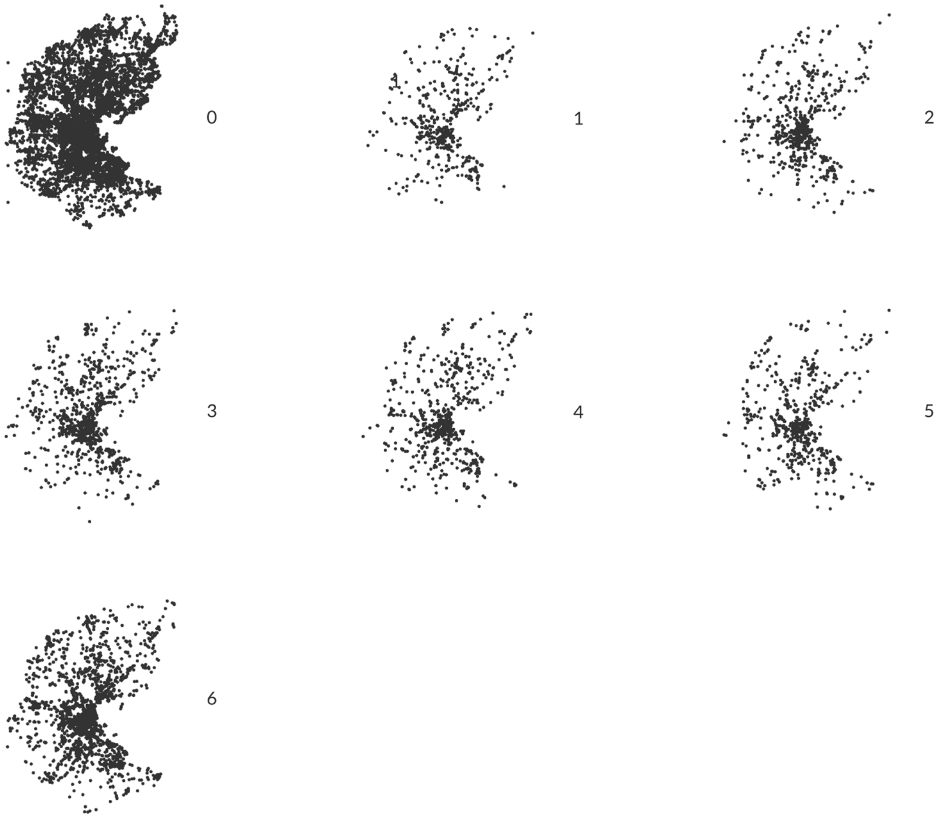


Fig. 94 Use of small multiples technique in order to find shift in error distribution in geo space. The visualization reads from left to right and from top to bottom.

5.2.3 Uniform distribution

From the observation above, we proceeded with another approach. The idea was to normalize density of POIs and represent only the error spread. The process was divided in three steps.

The first step consisted in the construction of a modular grid, with configurable size, which reflects on visual impact, as we will see on later examples.

Next we calculated the predominant error level for each grid cell, by aggregating the POIs contained in each cell. This is made as follow: for each cell we calculated the histogram of POIs error; then we normalized it; then we sum up all the normalized values.

For example, let us consider histogram array _{10, 5, 4, 6, 4, 2}_, the result of normalization will be _{10/31, 5/31, 4/31, 6/31, 4/31, 2/31} = {0.32, 0.16, 0.13, 0.19, 0.13, 0.06}_, summation will result in _0*0.32 + 1*0.16 + 2*0.13 + 3*0.19 + 4*0.13 + 5*0.06 = 1.81_. This is the predominant error for given cell.

The last step consisted in the representation of normalized values. In the first approach we relied on painting each cell with shades of gray, so that the representation did not take in count the density of POIs, Fig. 95. With this method the user could easily identify the zones with big errors in classification. However, the graph did not tell how critical it was. For example consider the zone A with only one POI that was classified with error of level 6 and zone B with 100 POIs with predominant classification error of level 4. This meant that zone B has a lot of more bad classified POIs than zone A. However, it was impossible to understand that through the visualization of normalized values.

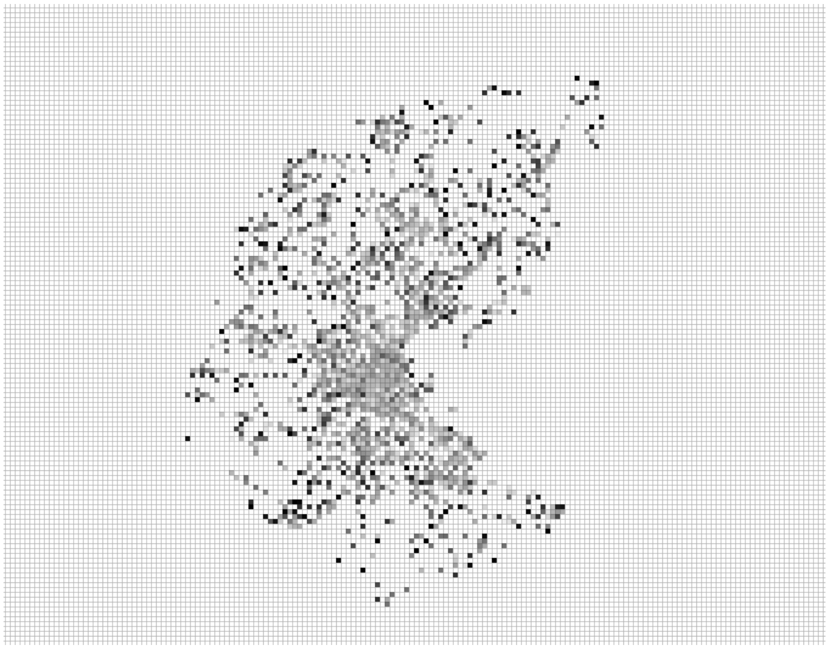


Fig. 95 Visualization of predominant error distribution in Boston, MA.

Yet, in terms of visual perception we found this a little contrasting representation. Aesthetic aspect was improved by use of circles. So, each circle was centered within a corresponding cell and its area was defined by predominant error in that cell, Fig. 96.

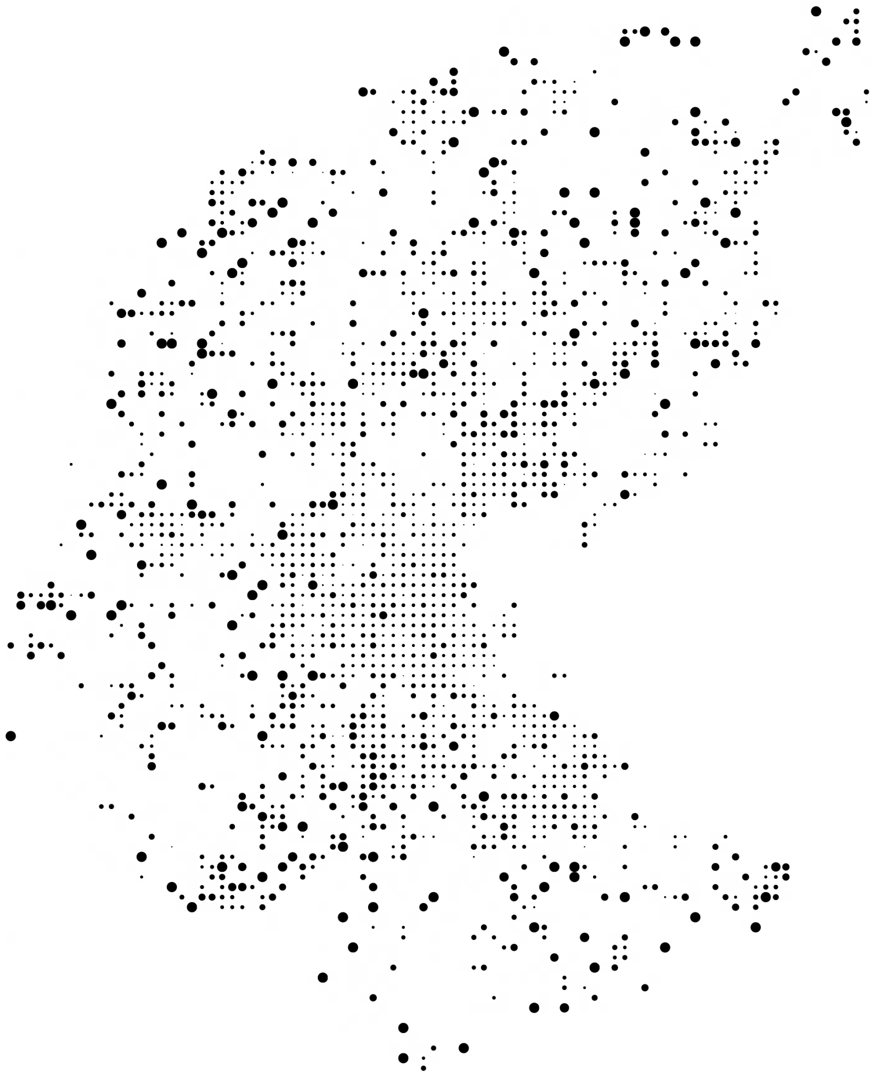


Fig. 96 Circle's area corresponds to function of predominated error.

5.2.4 Density vs Predominant Error

So, we added a representation of POIs' densities to the visualization. This was done by representing a density by color value of a corresponding cell, Fig. 97. This visualization compared density of POIs and predominant error per cell. So, we could say that saturated red cells with small circles meant that there were a lot of good classified POIs, since the predominant classification error is small even with that density amount. While, the cells with big circles and almost white color meant that there was big errors in classification, however there were a small number of classified POIs. The worst case would be in cells with big circles and red saturated colors. That would mean that there were many bad classified POIs.

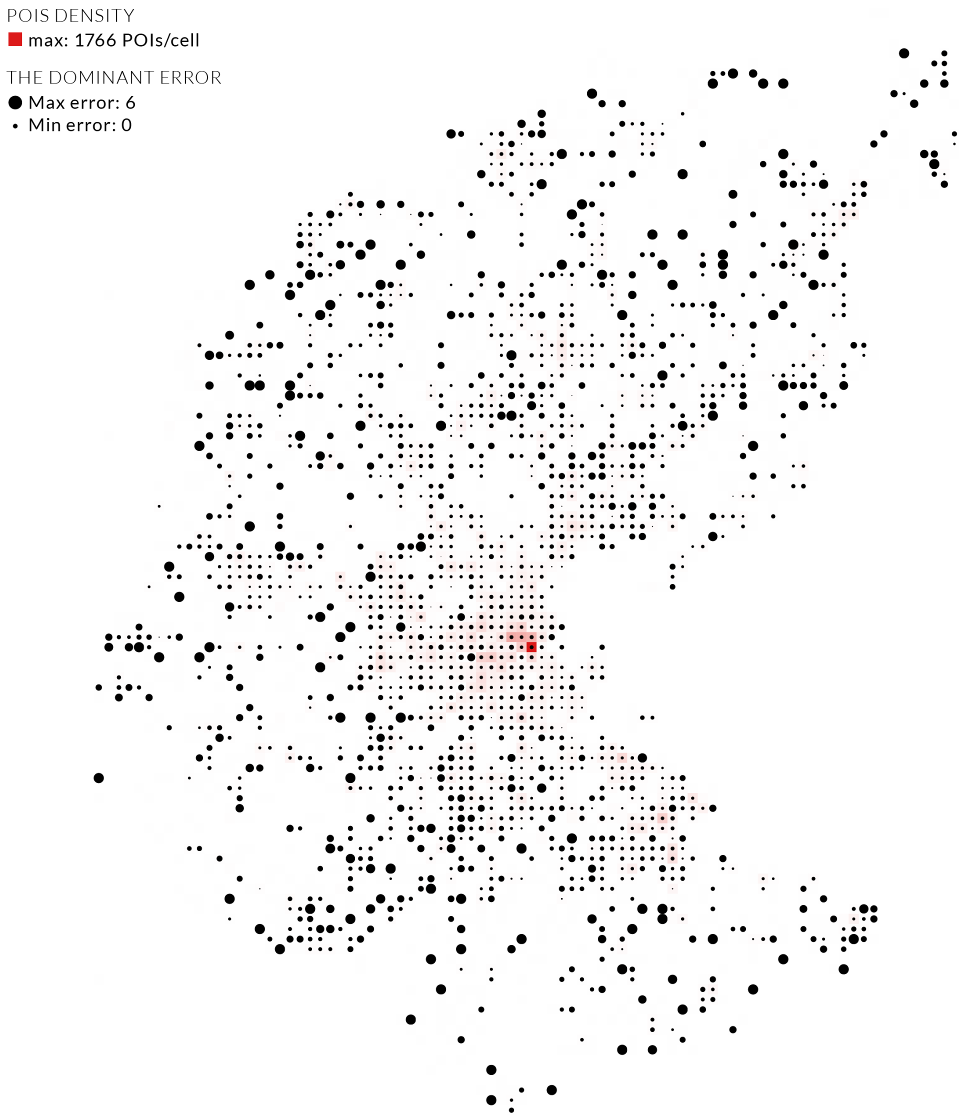


Fig. 97 Visualization of classification error and comparison with density of classified POIs. Value of red color varies in function of density and circle area in function of predominant error.

Extension 5.2.5

The visualization methods described above can easily be extended to include the performance of practical application of the POI data — in this case, the employment size estimation model from. Figure x shows the resulting visualization for the study of the area of Cambridge. As we can see from this figure, there is some relation between POI classification error and estimated employment size error. Basically, here we plotted the visualization described above over a map, which administrative areas were painted in blue color with varying value—the high dense areas were represented with saturated blue and the less dense areas were painted with light blue.

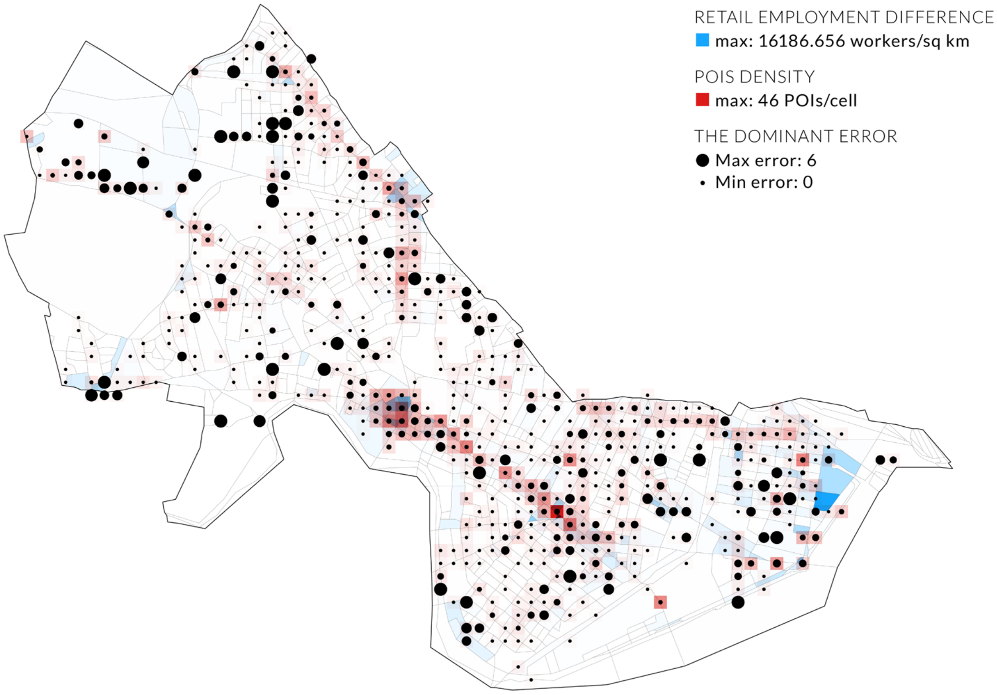


Fig. 98 Visualization of relation between POI density, red color, estimated disaggregated employment size error (blue), and POI classification error (black circles) for the area of Cambridge, MA.

5.3 Discussion

In this chapter we presented diverse visual techniques to represent semantic information, such as visualization of names of POIs, and the methods to improve performance and efficiency. We demonstrated different approaches to find a hull of a set of points. Finding the organic shape of a cluster is important to reveal its uniqueness. For that purpose we presented our method that was based on simulation of rubber band. Then we exemplified the usefulness of the algorithm by applying it on clusters of POIs. Additional semantic information was then added, by placing names of POIs on a path defined by a polygon of a shape. Furthermore, all the characters were placed on a guiding path, so that the resulted words had an organic form that accompanies the path. Finally, we discussed a visual variable that we used to encode the relevance of a category—typographic weight. Yet, a visual improvement of the model was presented.

Further work will focus on implementation of web application in JavaScript programming language. This will allow the user to explore the data via web browser. Furthermore, the visualization will be enhanced with supplementary graphics that will present more statistical information, such as number of POIs per cluster, related clusters to the given one, etc. Finally, we will explore other approaches to present names of POIs, based on the maps of Paula Scher. So, we will experiment filling our found shapes of clusters with the typography by placing the words in organic compositions.

The second part consisted in visualizing the accuracy of automatic classification. With the understanding of spatial distribution and density of POIs we proceeded to improve visualization, in order to make it more rigorous. This led us to the use of small multiples, where we could compare the shifts in relationship between various levels of classification error. Later on we realized that developed visual methods only shows a spatial distribution of POIs classification error. With the observation that the error distribution is more or less homogeneous we proceeded to remove the density dimension. The resulted visualization showed us the accuracy of classifier of aggregated POIs by grid cells. Further comparison with the density of POIs shows if classification error was critical or not. Yet, we visualized estimated employment dense at Block level. Then we cross-visualized the first representation with this one, in order to understand if one observation influences the other. The visualizations developed in this part of project were submitted and accepted for the International Journal on Advances in Intelligent Systems, as a visual proof of concepts developed by researchers [45].

Chapter 6
Conclusions

In the dissertation we presented different aspects of information visualization regarding to the representation of semantic and raw data. In order to develop efficient visualizations we investigated the state of the art, which covers theory and existing visualization techniques.

According to the Bertin's theory in order to efficiently represent information the designer should consider a content, the information to be transmitted, and a container, properties of graphic system. So, the information comprises the invariant and variable part. Variables, or components, of information consist in one or more elements, or categories. The relationship between components or elements defines four levels of organization: associative, selective, ordered or quantitative. The associative and selective information belong to the qualitative or nominal organization level.

On the other hand the designer has at his disposal the graphic system with visual units, aka marks. They can vary in position on paper or screen and in size, texture, value, color, orientation and shape. So, there are eight visual variables, which can be assigned to the point, line or area implantation. The point, while associated to a mark, can vary in all visual variables. Marks that indicate line can vary in size, value, texture, color, orientation of its components and shape of detail. The area can vary in value, texture and color. The way how the designer utilizes the two planar dimensions defines diagrams, networks, maps or symbols. Visual variables as information variables have their levels of organization and there are visual variables that are best suited for each level. In other words, there are variables that are most efficient to represent each class of information. So, associative variables are shape, orientation, color and texture, while value and size are dissociative. Selective information is best represented with all variables, except a shape. Ordered values are badly represented with shapes, colors and orientation. Quantitative values are efficiently represented just with size.

Cleveland's theory of graphical perception distinguishes the information as quantitative and categorical, which can be described as scale and physical information. His study defines the taxonomy of graphical perception. So, the user can read a graph by table look-up, while decoding scale information, and by pattern perception, while decoding physical information. The table look-up is done by scanning, interpolating and matching operations, while the pattern perception is done by detection, assembly and estimation. The last is divided in three progressive levels: discrimination, ranking and rationing.

The methodology of a design process was based on the one proposed by Ben Fry, which consists of seven stages: acquire, parse, filter, mine, represent, refine and interact. The process, of course, is not linear. On representation step the designer may return to filter or acquire stage, and on interaction step the data can be mined or the representation can be refined again. Later, we proposed our adaptation of this model. So, from our perspective, as designers, the process should start after the mining stage and progress as follows: analyze the data; encode/map information to the visual variables representing it with basic visualization techniques; represent given information in more suited way; provide methods of interacting; verify the usefulness of the representation.

The second part of the state of the art chapter focused on visualization techniques, such as diagrams, maps and interaction. In diagrams we covered a simple line chart and its variation such as sparklines. Then we introduced time-series with examples of multiple line charts, such as planetary movement graph, Playfair's time-series graph that visualizes the trade between England and Denmark and Norway, with further Cleveland criticism of inaccuracy of multiple lines graphs. Then we introduced a stacked area technique, which was exemplified with a graph designed by Steve Chappel and Reebe Garofalo, and one modern application of this technique produced by Moritz Stefaner. Finally, we covered small multiples. This high density graphic is very efficient in interpretation, which shows shifts in data.

Next we covered a graphical construction called maps. First of all we introduce basic theory of cartography. So, maps provide the ability to visually display physical locations. This information is called geographic coordinates, and is described by latitude and longitude, an absolute precise location on earth's surface. The transformation of this data from a spherical surface to a flat plane is called map projection. The process can be visualized as a projection of each point on a glob to a cylinder, cone or plane. We focused on normal cylindrical projection, more precisely on variation of a Mercator projection popularized by major map web services such as Google Maps and OpenStreetMaps, referenced as EPSG:3857.

Then we introduced a concept of subjective maps, which was exemplified by the antique map of Catal Hoyuk settle. This map was drawn based on the subjective knowledge of habitants about their village. The following maps were anamorphic maps, which display statistical information by deforming land areas or distances. The anamorphic map produced by Émile Cheysson represents travel times from Paris to various places in France by deforming its spatial size. Yet, we introduced a novel variation of anamorphic maps technique, which was applied in the UrbanCyclr visualization.

We also presented another type of subjective maps: schematic maps. These maps represent road or subway network systems using abstract graphic elements. As in the example of tube map produced by Henry Beck, he used only vertical, horizontal or on 45 degree diagonal lines. The omission of geographic detail makes tube map to be extremely readable, even in great distances. We also presented semi-schematic maps, such as Peuntinger Map. This map displays locations, orientations and distances of geographic features as accurate as possible, however the map has no similarities with reality. The usefulness of subjective maps was justified by comparing with Dymaxion Map produced by Buckminster Fuller. This map does not represent any obvious visual distortion of relative shapes and size, however from political, stereotypical, cultural and other points of view this map is useless.

Finally we discussed about typographic maps, such as maps of Paula Scher and Axis Maps. They use words to represent the information about countries, cities, streets, rivers, etc. This type of maps uses subjective information that is true for one group of people and totally unclear to others.

In the following small subsection we briefly presented topographic and contour maps. This was not particularly interesting to us, since this technique is useful to represent elevations of earth's surface. However, contour maps sometimes are used to represent statistical information. For example the map produced by Louis-Léger Vauthier displays population densities by use of contour lines.

Statistical maps are meant to represent quantitative data. One of the first uses of statistical maps was a chart produced by Edmond Halley, displaying a trade of winds. This map represents the direction and force of winds by line strokes. One of the worthy uses of statistical maps was a dot map produced by Dr. John Snow. In this visualization the deaths from cholera were represented by dots and water pumps by crosses. This map led Snow to discover sources of cholera. With this example we demonstrated the usefulness of visualization.

Another useful kind of maps is flow maps. These maps say little about the path, but include the information about what is flowing, in what direction and how much is being transferred. We exemplified this technique in low data dense map, produced by Charles Minard, of exports of French wine, and a high dense map of transportation, produced by Henry Harness. The flow maps demonstrate the concept of time-spatial representation. The classical Minard's flow map of Napoleon's march on Moscow exemplifies the efficient representation of data related to time and space. In modern visualizations the flow can be represented by animation. In the example of Lisbon's traffic, Pedro Cruz represents the flow of taxis by animating line traces.

Finally we presented a technique used in statistical maps known as heatmaps. On one example we saw how this technique can be used to create an image that reveals hotspots of searches using Mircrosoft's Live Search Map. On the other example this technique is used to visualize the travel times from a selected point on a city.

The following sub-section covered zooming and panning interaction techniques. We introduced a simple and local zoom technique. With the simple zoom and panning the user can bring closer and move the data space. With a variation of local zoom technique, aka fish-eye, the user can bring closer a part of an image retaining the whole image untouched. This was implemented in data lenses visualization by Pedro Cruz.

So, we started an urban mobility project with the data set that contained locations of bus stops and metro stations, and counts of passenger, as well as means and standard deviations aggregated by similar weekdays. We defined that the visualization must use simple and clear visual language, must be interactive and run in real time, which allows the user to explore the relationship between the crowd and urban space through the visualization of a public transport use. We also introduced the tools we used, such as Processing and its libraries, QGIS and OpenStreetMaps shape files.

Our first attempt to represent data was based on simple representation using a bubble graph, which represented the count of passengers. Later we demonstrated that the use of area lead to more accurate conclusions. Then we experimented with simple graphical elements. Based on the dot map of disease we proceeded to represent the data in similar way, but using lines. The idea and visual results appeared interesting to us, though the high amount of visual clutter made the visualization inefficient, yet, it did not work for metro station, since the metro stations belong to various streets. So, we went back to the mining stage and calculated new values that represented the anomalies in public transport use.

In the following experiment we encoded our data with irregular triangles. Due to their pointing nature we encoded the positive values with triangles pointing up and negative values with triangles pointing down. The size and opacity of these represented the corresponding absolute value. This quick experiment led us to the conclusion that our data should be represented using more sophisticated mechanisms. Further experimentation led us to the conclusion that the zoom technique can be used to transform the representation from one state to another or to morph from one visualization model to another. We discussed in detail a transformation of the same model, which in each state efficiently represent the data.

So, we introduced a metaball based visualization model. The developed representation presents strong pattern perception characteristics at general views and accurate estimation of information at zoomed views. Due to the nature of the metaball equation extra pixels are added to the visualization, which creates a blobby aspect of representation and add a subjective aspect to the visualization. Since the calculation was computationally expensive, we relayed on hardware acceleration by using GPU and shaders to produce visual artifacts.

Finally, we developed an application that uses the developed visualization model along with a supplementary graphic. This chart concerns the temporal aspect of the data. Using a bar graph we visualized summed positive and negative deviations aggregated by time intervals.

The second project focused on visualization techniques that help understand land use. So, we presented diverse visual techniques to represent semantic information such as visualization of names of POIs and the methods to improve performance and efficiency. We demonstrated different approaches to find a hull of a set of points. Finding an organic shape of a cluster was important to reveal its uniqueness. For that reason we presented our method, which was based on simulation of rubber band.

Then we exemplified the usefulness of the algorithm by applying it on clusters of POIs. Additional semantic information was then added, by placing names of POIs on a path defined by a polygon of a shape. Finally, we discussed a visual variable that we used to encode relevance of a category—typographic weight. Finally, a visual improvement of the model was presented.

The second part consisted in visualization of the accuracy of automatic classification. With the understanding of spatial distribution and density of POIs we proceeded to improve visualization, in order to make it more rigorous. This led us to the use of small multiples, where we could compare the shifts in relationship between various levels of classification error.

Later on, we realized that developed visual methods only show a spatial distribution of POIs classification error. We observed that the error distribution is more or less homogeneous, so we removed a density dimension. The resulting visualization showed us the accuracy of classifier of aggregated POIs by grid cells. Further comparison with the density of POIs showed if classification error was critical or not. Additionally, we visualized estimated employment dense at Block level. Then we cross-visualized the first representation with this one, in order to understand if one observation influenced the other.

The dissemination of the results of this dissertation resulted in a publication on the SIGGRAPH 2013 conference [46] and on the incorporation of some of the developed visualizations on a paper of the International Journal on Advances in Intelligent Systems [45]. Further dissemination activities will involve the submission of a paper to one of the main international conferences on Visualization and on the submission of a paper to a top ranked international journal.

Bibliography

[1] J. Bertin, Semiology of graphics : diagrams, networks, maps, 1st ed. Redlands, Calif.: ESRI Press : Distributed by Ingram Publisher Services, 2010.

[2] E. Lupton and J. C. Phillips, Graphic Design. Princeton Architectural Pr, 2008.

[3] W. S. Cleveland, The Elements of Graphing Data, Revised. New Jersey: Hobart Press, 1994.

[4] B. Fry, Visualizing data, 1st ed. Beijing; Cambridge: O'Reilly Media, Inc., 2008.

[5] A. Marcus, "Improving the User Interface," 28-Apr-1999. [Online]. Available: http://webword.com/interviews/marcus.html. [Accessed: 31-Aug-2013].

[6] E. R. Tufte, Beautiful evidence, 1st ed. Cheshire, Connecticut: Graphics Press, 2006.

[7] E. R. Tufte, The visual display of quantitative information, 2nd ed. Cheshire, Connecticut: Graphics Press, 2001.

[8] W. S. Cleveland, "Graphical Perception," in The Elements of Graphing Data, New Jersey: Hobart Press, 1994.

[9] I. Spence and H. Wainer, "William Playfair," in Encyclopedia of Social Measurement, vol. 3, San Diego, CA: Academic Press, 2005, pp. 71–79.

[10] E. R. Tufte, Visual Explanations: Images and Quantities, Evidence and Narrative. Connecticut: Graphics Press, 1997.

[11] M. Stefaner, "Visual Tools For the Socio-Semantic Web," University of Applied Sciences Potsdam, Potsdam, 2007.

[12] E. R. Tufte, Envisioning Information. Cheshire Connecticut: Graphics Press, 1990.

[13] M. M. Yavuz, Ed., Sense of Patterns, 2011. [Online]. Available: http://casualdata.com/senseofpatterns/. [Accessed: 2013].

[14] E. A. Murphy, "One Cause? Many Causes? The Argument from the Bimodial Distribution," Journal of Chronic Diseases, vol. 17, p. 309, 1964.

[15] P. Bil'ak, “Family planning, or how type families work,” www.typotheque.com, 2008. [Online]. Available: https://www.typotheque.com/articles/type_families. [Accessed: 14-Aug-2013].

[16] M.-J. Kraak and F. Ormeling, Cartography, 3rd ed. Guilford Press, 2010.

[17] X. Chen, “Seeing differently: cartography for subjective maps based on dynamic urban data,” Massachusetts Institute of Technology, 2011.

[18] M. Friendly and D. J. Denis, “Milestones in the history of thematic cartography, statistical graphics, and data visualization,” Retrieved, Aug. 2009.

[19] B. Attila, Ed., SubMap x UrbanCyclr (2.1), 2011. [Online]. Available: http://submap.kibu.hu/. [Accessed: 16-Aug-2013].

[20] B. R. Fuller, “Life Presents R. Buckminster Fuller's Dymaxion World,” Life, pp. 41–55, Mar. 1943.

[21] R. Bringhurst, The Elements of Typographic Style. Hartley & Marks, 2013.

[22] E. Lupton, Thinking with Type, 2nd revised and expanded edition. Princeton Architectural Press, 2010.

[23] P. Scher, Ed., Typographic Maps. [Online]. Available: http://www.paulaschermaps.com/. [Accessed: 23-Aug-2013].

[24] Axis Maps. [Online]. Available: http://store.axismaps.co.uk/. [Accessed: 23-Aug-2013].

[25] C. Ratti, A. Biderman, and L. Liu, “Urban Mobility Landscape: Real Time Monitoring of Urban Mobility Patterns.”

[26] P. M. Cruz, “Boundaries in information visualization - towards information aesthetics,” University of Coimbra, Coimbra, 2010.

[27] D. Fisher, “Hotmap: looking at geographic attention.,” IEEE Trans Vis Comput Graph, vol. 13, no. 6, pp. 1184–1191, 2007.

[28] C. Lightfoot and T. Steinberg, Eds., Travel-time Maps. [Online]. Available: http://www.mysociety.org/2006/travel-time-maps/. [Accessed: 22-Aug-2013].

[29] P. M. Cruz, “Data Lenses,” pmcruz.com, 22-Apr-2012. [Online]. Available: http://pmcruz.com/information-visualization/data-lenses. [Accessed: 16-Aug-2013].

[30] F. Calabrese and C. Ratti, "Real Time Rome," Networks and Communication Studies, vol. 20, 2006.

[31] A. Vaccari, M. Martino, F. Rojas, and C. Ratti, "Pulse of the city: Visualizing Urban Dynamics of Special Events," presented at the Proc. GraphiCon, 2010, pp. 64–71.

[32] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, "The Livehoods Project: Utilizing Social MEdia to Understand the Bynamics of a City," 2012.

[33] Livehoods, 2012. [Online]. Available: http://livehoods.org/. [Accessed: 17-Aug-2013].

[34] B. Fry and C. Reas, Eds., Processing, 2001. [Online]. Available: http://processing.org/. [Accessed: 24-Aug-2013].

[35] Java Map Projection Library. [Online]. Available: http://www.jhlabs.com/java/maps/proj/. [Accessed: 24-Aug-2013].

[36] OpenStreetMaps. [Online]. Available: http://www.openstreetmap.org/. [Accessed: 24-Aug-2013].

[37] Quantum GIS. [Online]. Available: http://www.qgis.org/. [Accessed: 24-Aug-2013].

[38] J. F. Blinn, "A generalization of algebraic surface drawing," ACM Transactions on Graphics (TOG), vol. 1, no. 3, pp. 235–256, 1982.

[39] A. C. D. C. O. Alves, "Semantic enrichment of places - Understanding the meaning of public places from natural language texts," 2012.

[40] A. O. Alves, F. Rodrigues, and F. C. Pereira, "Tagging space from information extraction and popularity of points of interest," in Ambient Intelligence, Springer, 2011, pp. 115–125.

[41] J. Oliveirinha, F. Pereira, and A. Alves, "Acquiring semantic context for events from online resources," in Proceedings of the 3rd International Workshop on Location and the Web, 2010.

[42] L. Byron, Ed., Mesh. [Online]. Available: http://leebyron.com/else/mesh/. [Accessed: 31-Aug-2013].

[43] JTS Topology Suite, 2006. [Online]. Available: http://www.vividsolutions.com/jts/JTSHome.htm. [Accessed: 31-Aug-2013].

[44] Blob Detection. [Online]. Available: http://www.v3ga.net/processing/BlobDetection/index-page-home.html. [Accessed: 31-Aug-2013].

[45] F. Rodrigues, A. Alves, E. Polisciuc, and S. Jiang, “Estimating Disaggregated Employment Size from Points-of-Interest and Census Data: From Mining the Web to Model Implementation and Visualization,” Journal On Advances In Intellegent Systems, vol. 6, pp. 41–52, 2013.

[46] E. Polisciuc, A. Alves and P. Machado, Visualizing Urban Mobility, Poster Session at SIGGRAPH conference, Anaheim, 2013.

Título:

Visual Tools For Urban Mobility Study

Design strategies for visualization and analysis of raw and semantic data to understand urban mobility

Dissertação de mestrado em Design e Multimédia
Faculdade de Ciência e Tecnologia
Universidade de Coimbra

Autor:

Evgheni Polisciuc (*evgheni@student.dei.uc.pt*)

Orientadores:

Ana Alves (*ana@dei.uc.pt)*

Artur Rebelo (*arturr@dei.uc.pt)*

Penousal Machado (*machado@dei.uc.pt)*

Juri:

Pedro Cruz (*pmcruz@dei.uc.pt*)

Nuno Coelho (*ncoelho@dei.uc.pt*)

Impressão

case de impressão Dual Print

quantidade de 9 cópias

formato A4

Coimbra, 4 de Setembro de 2013

• U C •