

Article

Analytical Estimation of Map Readability

Lars Harrie ^{1,†,*}, Hanna Stigmar ^{1,†} and Milan Djordjevic ²

¹ Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, SE-223 62 Lund, Sweden; E-Mail: hanna.stigmar@gis.lu.se

² Department of Geography, Faculty of Science and Mathematics, University of Nis, Visegradska 33, 18000 Nis, Serbia; E-Mail: milan.djordjevic@pmf.ni.ac.rs

† These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: lars.harrie@nateko.lu.se; Tel.: +46-46-2220155.

Academic Editor: Wolfgang Kainz

Received: 6 November 2014 / Accepted: 5 March 2015 / Published: 27 March 2015

Abstract: Readability is a major issue with all maps. In this study, we evaluated whether we can predict map readability using analytical measures, both single measures and composites of measures. A user test was conducted regarding the perceived readability of a number of test map samples. Evaluations were then performed to determine how well single measures and composites of measures could describe the map readability. The evaluation of single measures showed that the *amount of information* was most important, followed by the *spatial distribution* of information. The measures of *object complexity* and *graphical resolution* were not useful for explaining the map readability of our test data. The evaluations of composites of measures included three methods: threshold evaluation, multiple linear regression and support vector machine. We found that the use of composites of measures was better for describing map readability than single measures, but we could not identify any major differences in the results of the three composite methods. The results of this study can be used to recommend readability measures for triggering and controlling the map generalization process of online maps.

Keywords: cartography; map readability; usability; user test; supervised learning

1. Introduction

The readability of maps is important. Traditionally, cartographers were responsible for ensuring the readability of maps. However, because of the cartographic digital revolution, particularly the use of the Internet, many maps are not controlled by cartographers. Therefore, the possibility of measuring the readability of a map analytically is important: is it possible to define a measure or a composite of measures that describe map readability?

Measures of map readability have two main uses. The first usage concerns dataset specifications, in which producers often set thresholds of measures, such as the minimum size of an object and the minimum line width. The second usage is to trigger, control and evaluate the automatic generalization process. Although readability measures are frequent applied, relatively few user studies have advanced the understanding of their application. In this study, we calculate an extensive number of readability measures and evaluate the applicability of these measures in a user study. The aim of the evaluation is twofold. First, the evaluation aims to determine which measures are useful for explaining map readability. Because a map is a complex entity, a single measure is unlikely to explain its full readability; we must combine the measures as weighted readability formulas. Therefore, our second aim is to compare three composite methods of measures for describing map readability. Specifically, the aim of this study is to evaluate the use of measures, rather than develop new measures. The paper is organized as follows. First, the background of previous research in readability measures is provided. In Section 3, we describe the methodology, including details of the map samples, the user-test procedure, the participants in the user test, and the readability measures and composites. Section 4 presents the results of the user test and the evaluation of single and composite measures. The paper concludes with a discussion and conclusions.

2. Related Studies

2.1. Background

Map readability is a broad term. In this study, we conducted a user test in which we used the following definition: Map readability focuses on the possibility of discerning map symbols (separating individual symbols and separating symbols from the background) and on the ease of reading, interpreting, and comprehending a map.

The development of analytical measures of map readability is a relatively new field of research compared with the analysis of written text, for example. In the 19th century, Sherman [1] proposed that readability is affected by the length and organization of sentences, as well as by the choice of words. Readability formulas were introduced in the 1920s. These formulas were used to predict the difficulty of a text based on its contents. Gray and Leary [2,3], for example, investigated more than 200 elements of content, style, format and features of organization. By 1981, over 200 readability formulas had been published. These formulas have been validated by user tests, which have focused, for example, on readership, reading persistence and reading efficiency [3]. The obvious question is whether single measures or composites of measures, similar to those used to determine text readability, are useful for predicting map readability.

2.2. Readability Measures

In cartography and vision science, visual complexity has been identified and reduced. Visual complexity can be defined as a state in which excess items, or their representation or organization, lead to degradation of user performance [4]. This definition implies that factors such as the *amount of information/object complexity* (“excess items”), the *graphical resolution* (“representation”), and the *spatial distribution* (“organization”) of map objects determine the degree of readability of a map. In the cartographic literature, there have been a substantial number of measures proposed:

- *Amount of information*: the number of objects (e.g., [5–7]); the number of objects of a particular type [8,9]; the number of vertices [10–12]; the number of nodes, links and areas [11,12]; total length of links [12]; and occupied space [7,13].
- *Spatial distribution*: the distribution of objects [11]; object symmetry and organization [7]; entropy measures for objects and points [14,15]; homogeneity and number of neighbors [15]; density of objects [16]; and congestion measures [17].
- *Object complexity*: sinuosity [18,19]; total angularity [7]; and line connectivity [12,20].
- *Graphical resolution*: minimum size of points (on paper and on screens); minimum width of lines; and minimum separation of objects (see e.g., [21]). Additional measures include aspects of colors (e.g., contrast) of the visualized objects [7,22].

The measures above are mainly used for vector maps. For raster maps, other types of complexity measures are used. For example, Fairbairn [12] demonstrated that image compression is a valid measure for the structural complexity of raster maps. This result is extended by Jégou and Deblonde [23] who, among others, used a quad-tree representation of the raster map. The complexity of the image is then estimated by using the structure of the tree and the color value differences of adjacent pixels.

There have been some studies on composites of measures for map complexity. Fairbairn [12] argued that it would be advantageous to use composite measures to describe the complexity of vector maps. Rosenholtz *et al.* [4] and Rosenholtz *et al.* [24] presented three measures for describing properties of visualization: the first describes the visual complexity based on color, contrast and orientation; the second calculates a weighted sum of entropies; and the third indicates the density of edge pixels.

The readability of maps is related to visual distractors of images in general. In vision science, several studies have been performed regarding factors of efficient image searches. Researchers (using field-specific terminology) and others have investigated which distractors (unwanted spatial objects) affect the search of a specific target. He *et al.* [25] concluded that without distractors, the perception of spatial objects is limited by the visual resolution. However, when several objects are presented, the perception depends on the ability of attentional processes to isolate the objects. To guide a person’s attention, the target must be different from the distractors. Examples of such differences could be color, orientation and size [26]. In the scope of this paper, we could state that visual science has found that identifying and searching for map objects depend on the surroundings of the map objects. If there are many similar objects nearby, this identification/searching process is degraded, which will affect the readability of the map (*cf.* [5]).

2.3. Usability Tests of Readability Measures

Moacdieh and Sarter [27] provided an extensive review of methodologies for measuring readability (which they denote *level of clutter*) in graphics and images. In their conclusions, the authors state that measuring readability requires both a characterization of the display and/or subjective evaluation and an assessment of the performance of readability using performance outcome measures. Phillips and Noyes [5] tested the complexity of the topographic basis of geological maps. The aim of the study was to recommend methods of improving 1:50,000 geological maps. In this test, five versions of the same map with differing topographic bases were compared in terms of map readability. The authors found that the amount of information on a map reflects the map readability. The results supported the idea that numerous close-proximity objects of the same symbol style or color tend to create clutter. Rosenholtz *et al.* [24] tested three raster-map readability measures, *i.e.*, feature congestion, sub-band entropy and edge density, in map search tasks. The search tasks were performed on digital maps. A significant correlation existed between the mean log (reaction time) and readability measures. In the second experiment, four users were asked to identify a one-second displayed target (on a map) and indicate its orientation. Contrast thresholds were studied in relation to readability measures, and a significant correlation was noted. In a third experiment, color variability was studied. Eighteen maps of varying colors were created. Four users were asked to find a specific target on the maps as quickly as possible. It was noted that the reaction times were longer for maps with larger color ranges. Lohrenz *et al.* [28] also used raster-map readability measures, which were based on saliency and color. The authors found that low color density plus high saliency results in clutter. Stigmar and Harrie [29] evaluated 17 measures of the amount of information, spatial distribution and object complexity. Twelve test participants were interviewed regarding the readability of a number of test maps and were asked to rank maps according to the perceived readability. The results showed that some measures of the amount of information and spatial distribution corresponded well to the participants' opinions. The measures of object complexity did not show the same correspondence.

2.4. Readability Measures in Dataset Specifications and Cartographic Generalization

Map readability measures in dataset specifications are used for graphical resolutions (e.g., [30,31]), but the other measurement categories have not yet led to similar recommendations for map production. There are, for example, few rules regarding the total angularity of a line or the maximum number of vertices in a map region. Furthermore, to the authors' knowledge, no map specification includes composites of measures.

Stoter *et al.* [32] comprehensively studied map specifications for automated map generalization (*i.e.*, selection and simplified representation of details appropriate to the scale and/or purpose of a map [17]). In their results, the authors listed graphical measures for both single objects and groups of objects. However, measures that target the readability of a map were missing in the specifications. Many other studies in the field of generalization have both developed and used map readability measures, particularly in the context of evaluation. Early work was performed by McMaster and Shea [19]. These authors suggested the use of cartometric evaluation, e.g., measures of density, distribution and shape, for triggering generalization. More comprehensive studies have been

performed, e.g., the AGENT project [17]. In this project, several readability measures of both individual and groups of object (proximity, parallelism, congestion, *etc.*) were developed (see e.g., [33,34] for summaries and [17] for a detailed description). Recent overviews of the evaluation of generalization, in which readability measures are a component, are given in [35,36].

Most measures of map readability have been geometrically oriented, e.g., based on the amount and distribution of information. This trend is consistent with most research on automated generalization that has focused on geometries. Brewer and co-authors [37,38] importantly suggested that we must not neglect the importance of the symbol style. Apart from improving the cartographic quality, the inclusion of symbol style changes could also reduce the work of maintaining multiple scale databases. A practical example of decreasing the need for geometric generalization by symbol style changes is provided by the National Land Survey of Sweden. The survey has used a partly transparent outer part of road symbols in small-scale maps to reduce the need for movement, e.g., moving building objects away from road objects. Additionally, Roth *et al.* [39] proposed that the traditional generalization operators that are geometry focused (see e.g., [40]) should be extended with new operators that focus on the changes in symbol styles. The authors introduce a new operator called *symbol* that includes, e.g., adjustment of color, adjustment of iconicity and adjustment of pattern. To trigger and control the new generalization operators, we need readability measures that focus on symbol styles. Defining readability measures of symbol styles is difficult because of the formalization. For example, color measures should include semantic rules (if themes are related, then they should have similar colors), contrast rules (support figure/ground in the map) and conventional rules (e.g., water is blue) [41].

2.5. Semantic Aspects of Readability Measures

It should be noted that most of the readability measures in the literature (as well as in this study) concern the syntactic component, rather than the semantic component, of map readability. Semantics are related to the perceived meaning of map symbols and are therefore difficult to measure. It has been argued (e.g., [42]) that it is not possible to measure the readability of a map because it is not possible to completely measure the semantic aspects. Some portions of information are not actually presented in the map but are derived from the reader's previous knowledge and intelligence. This statement is also supported by studies in visual science. Neider and Zelinsky [43], for example, performed a user test of search times as a function of clutter in images (the number of buildings in the scene). The authors found that low-level descriptions of the scene (similar to some of the measures described in 3.5 below) could not fully explain the search time; therefore, they conclude that conceptual aspects may also be important in determining the effects of clutter on searches. We are also convinced that the semantic level will affect readability, but we have a pragmatic view. If we can show that syntactic measures are useful for improving map readability (e.g., by controlling the map generalization process), then these measures should be used, even though they do not present the entire truth.

As described above, much research has focused on defining map-readability measures. However, comprehensive user surveys that target the applicability of these measures are lacking. Such studies are useful for determining which readability measures or composite measures should be included in the map specifications and used for triggering and controlling the automated map generalization process. We argue that such measures are particularly important for automated generalization in map services

based on user-generated data, such as OpenStreetMap [44]. A user of these services cannot expect the same level of positional accuracy and completeness as when using NMA services (*cf.* [45]), but they will expect high map readability.

3. Materials and Methods

3.1. Method Overview

This study compares the automated computation of map readability and perceived readability. The methodology comprises six main steps.

1. A number of map samples were created.
2. A user test was performed using these map samples.
3. Analytical readability measures were chosen.
4. Composites of the readability measures were selected.
5. An evaluation of how well the single readability measures could describe map readability was performed.
6. An evaluation of how well the composites of readability measures could describe map readability was performed.

In the remainder of this section, we describe the first four steps. The last two steps are given in the results section.

3.2. Materials—Creation of Map Samples

We decided that all map samples should be derived from topographic maps for two reasons. First, topographic maps are the most common maps and are also the base of other maps (such as thematic maps). Second, this study aims to evaluate the applicability of syntactic readability measures. Therefore, we preferred to use map samples containing feature types that are well-known to most of the participants. If there are unknown feature types in the map, there is a risk that these feature types would affect the participant's perception of the map readability. In other words, we want to avoid a situation in which the participant's shortcomings in semantic understanding would affect their judgment of the syntactic content.

The ideal map sample size is controlled by two opposing aspects. The first aspect is that the map samples should be as large as possible. Using small map samples would provide an unnatural setting for the participants. However, the map samples must be homogenous to produce reliable results. If the map samples are not homogenous, then it is difficult to evaluate what circumstances affect a user's perceived readability of a map. Additionally, because the readability measures are based on the entire map sample, it is problematic to evaluate their applicability if only parts of the map sample are unreadable.

Based on these considerations, we derived map samples from a topographical map database covering the vicinity of Helsingborg, Sweden. The map database consisted of layers in the scale range of 1:10,000–1:50,000 (from the National Land Survey of Sweden and the municipality of Helsingborg). First, 60 map regions were selected: 30 regions at a scale of 1:10,000 and the other

regions at a scale of 1:50,000. The map regions were chosen such that they presented relatively homogeneous areas, which were attained by a relatively small map size (3 × 2 cm). The map regions represented the most typical types of areas in the map (e.g., urban areas, recreational areas, industrial areas, and rural areas) and represented variable map readability. We classified the map samples into the following categories: *dense* (if the map sample has a generally dense information impression); *sparse*; *many object types*; *few object types*; *dense lines* (sub-regions contain dense line objects); *dense point objects* (sub-regions contain dense point objects); and *dense buildings*.

The maps for all 60 regions were compiled using three levels of detail (LOD 1–3, see Table 1 for details on the map information in each LOD) by selecting data layers with different resolutions. To conduct evaluations of graphical resolution measures, the maps were presented with two different symbol styles. The first symbol style, TS, (see Figure 1a–d) is a “traditional” Swedish style, often used for paper maps and for traditional-looking digital maps. The second symbol style, NS, (see Figure 1e–h) is a pale style developed by the Swedish National Land Survey for backdrop mapping.

Table 1. The map information included at the three levels of detail (LOD) of the map samples. The LODs are not standard products but are collections of data layers selected to fit the purpose of this study. Point data are in **bold**, line data are in *italics* and polygon data are in normal font.

Scale	LOD 1	LOD 2	LOD 3
1:50,000	<ul style="list-style-type: none"> • Railroad point symbols • Ancient remains and building point symbols • <i>Roads</i> • <i>Power lines</i> • <i>Hydrography</i> • <i>Contour lines</i> • <i>Protected areas</i> • Establishments (e.g., power plants) • Land cover 	<ul style="list-style-type: none"> • Railroad point symbols • <i>Roads</i> • <i>Power lines</i> • <i>Hydrography</i> • <i>Contour lines</i> • Protected areas • Establishments (e.g., power plants) • Land cover 	<ul style="list-style-type: none"> • Railroad point symbols • Ancient remains and building point symbols • <i>Roads (low resolution)</i> • <i>Power lines (low resolution)</i> • <i>Hydrography (low resolution)</i> • <i>Contour lines</i> • Protected areas • Establishments (e.g., power plants) • Land cover (low resolution)
1:10,000	<ul style="list-style-type: none"> • Road and railroad point symbols • Ancient remains and building point symbols • Buildings • <i>Boundaries (e.g., real-estate boundaries)</i> • <i>Roads</i> • <i>Power lines</i> • <i>Hydrography</i> • <i>Contour lines</i> • Establishments (e.g., power plants) • Land cover 	<ul style="list-style-type: none"> • Road and railroad point symbols • Ancient remains and building point symbols • Buildings • <i>Roads</i> • <i>Power lines</i> • <i>Hydrography</i> • <i>Contour lines</i> • Establishments (e.g., power plants) • Land cover (low resolution) 	<ul style="list-style-type: none"> • Road and railroad point symbols • Ancient remains and building point symbols • <i>Boundaries (e.g., real-estate boundaries)</i> • <i>Roads</i> • <i>Power lines</i> • <i>Hydrography</i> • <i>Contour lines</i> • Establishments (e.g., power plants) • Land cover (low resolution)

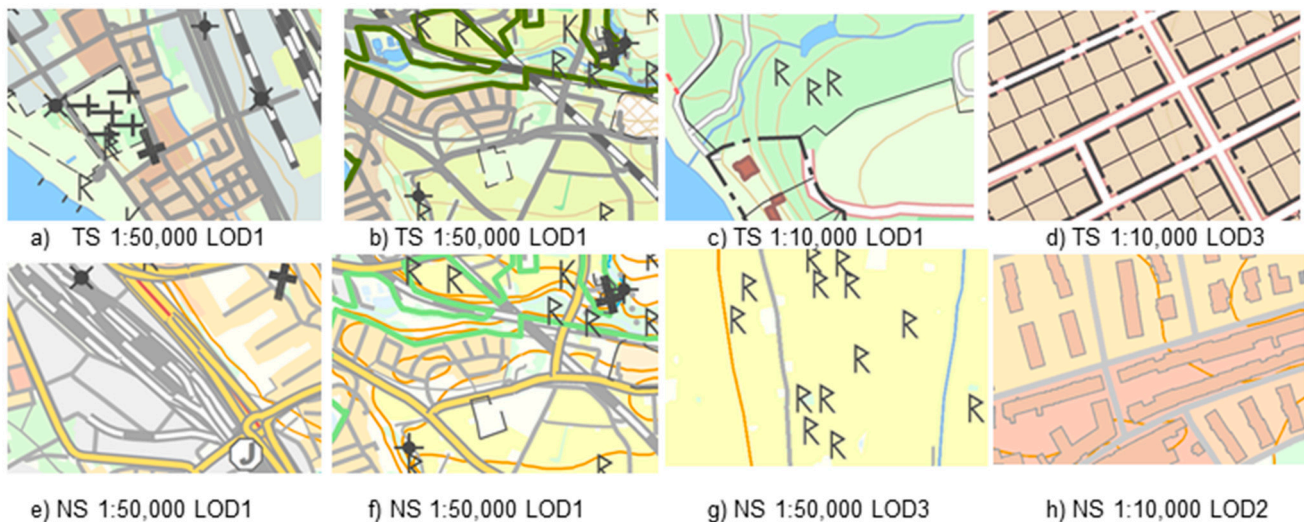


Figure 1. Eight map samples (a–h) used in the study. The maps are presented with two symbol styles (NS and TS) at two scales (1:50,000 and 1:10,000) and at three levels of detail ((LOD) 1–3). Maps b and f only differ in symbol styles.

3.3. Participants in the User Study

We sought experienced map users or those involved in GIS as test participants. The target user categories were thus GIS and geography professionals, GIS and geography students, and professionals who use maps in their work (e.g., surveyors). Of the participants, 47% were female and 53% male. The average age was approximately 30 years old (although the participants' ages ranged from younger than 20 to older than 70). Half (51%) of the participants were students, 23% were teachers or researchers, and 26% were “others”; 47% of the participants were Serbian students. The last group was the only one to complete the survey as part of a laboratory (*i.e.*, they did not volunteer to participate). All other participants answered on a voluntary basis; they were invited using e-mail lists for professionals in the targeted user categories. Overall, 37% of the participants were from Sweden, 56% were from Serbia and 7% were from other countries. Just over a quarter of the participants (27%) used maps every day, 59% used maps every week, 13% used maps every month, and 1% claimed that they never used maps.

3.4. Procedure of the User Study

The user test was designed as a web-distributed questionnaire. To test all 350 map samples without exhausting the participants, we developed seven tests with 50 map samples each. Most of the map samples were used in more than one part of the test (see below). The map samples were displayed randomly in the tests. Apart from the different maps, the seven tests were identical. The tests were given in order as the participants opened the test web page (*i.e.*, participant 1 was given test 1, participant 2 was given test 2, ..., participant 8 was given test 1, *etc.*). The test language was English, as we expected participants of different nationalities.

The first page provided an introduction, in which the participants were informed of the aim of the test, as well as the test process. We also provided a definition of map readability (see Section 2.1

above). This was followed by some personal profile questions and by the main test itself. The main test consisted of five parts:

1. readability evaluation of 17 map samples (see Appendix A),
2. readability ranking of 10×4 map samples,
3. readability evaluation of 17 map samples,
4. map task (for 10 map samples), and
5. readability evaluation of 16 map samples.

The test ended with a comment box in which the participants were asked to write their comments on the map samples or the test, if they had any. The entire test took approximately 20 min to complete.

The readability evaluations of test parts 1 and 3 consisted of 17 pages each, and test part 5 was 16 pages, with one map sample on each page. The definition of “readability” was given on the first page (*cf.* Section 2.1), and the participant was asked to assess the “readability” of the map sample as “very difficult to read”, “difficult to read”, “easy to read”, or “very easy to read”. The readability ranking (test part 2) consisted of 10 pages with four map samples on each page. The map samples used in this part of the test were a selection of the 50 maps used in the readability evaluations (parts 1, 3 and 5). The participants were asked to compare the map samples and order them according to their readability. We used the sequences from this part to compare the order of the same maps in the readability evaluations. The map task (test part 4) consisted of 10 pages, with one map sample on each page. The participants were asked to count the number of buildings or ancient remains on the map and to note the number. The map samples in this part of the test were taken from a 28-map set that was used in the map tasks in all seven versions of the tests. Both ranking and evaluation of the same map samples were included to determine whether the participants were consistent in judging the map readability. The purpose of the map task was to compare the performances of the participants to their answers in the readability evaluations.

After removing some outliers (e.g., when the participant assessed all map samples as “very easy to read”), 214 participants were included in the test. Some participants (18) did not complete the entire test, but we have used the results of the questions they did complete. We argue that these participants may not have had time to complete the entire test or may have been interrupted. Since their answers seem to be serious, we chose to include their answered questions; there is a minor tendency that these persons reveal the map samples as more difficult to read than the rest of the participants.

Because of the form of the test—a web-distributed questionnaire—the sizes of the map samples may have appeared different to participants based on the size and resolution of their computer screens. The map samples were designed for a 19-inch, 1280×1024 pixel screen, where the map samples were displayed at their original scales (1:50,000 and 1:10,000) (3×2 cm). To compute the sizes of the map samples evaluated by the participants, they were asked to provide their screen size and resolution. However, when evaluating the results by comparing the *perceived readability values* (see next section), we found no differences in the readability perceived by the participants using different screen sizes or resolutions. All the data were therefore evaluated together.

3.5. Readability Measures

In this study, we used analytical readability measures defined for different *types of information* in the map. Based on their geometrical properties, we defined the following four types of information (cf. [46]):

- *minor objects* consisting of small, stand-alone point, line and area objects;
- *line networks* consisting of line objects forming networks (such as roads, rivers and boundaries);
- *tessellation objects* consisting of area objects forming tessellations (partitions), such as land use; and
- *field-based data* consisting of contour lines.

The measures belong to the following *categories*: amount of information, spatial distribution, object complexity and graphical resolution (cf. Section 2). The measures are listed in Table 2 and described in the forthcoming section. All measures were computed for each map sample. The selection was based on a literature search and previous experience [29,47]. In the test, we restricted ourselves to vector-based measures, *i.e.*, we did not use raster-based measures of complexity (as in e.g., [12,23]).

Table 2. The measures and their applications to the types of information (rows) and measure types (columns).

Type of Information	Amount of Information	Spatial Distribution	Object Complexity	Graphical Resolution
Minor objects	<ul style="list-style-type: none"> • Number of objects • Number of vertices • Object line length 	<ul style="list-style-type: none"> • Spatial distribution of objects • Spatial distribution of vertices 	<ul style="list-style-type: none"> • Object size • Line segment length 	<ul style="list-style-type: none"> • Brightness difference • Hue difference
Line networks	<ul style="list-style-type: none"> • Number of objects • Number of vertices • Object line length 		<ul style="list-style-type: none"> • Line segment length 	<ul style="list-style-type: none"> • Brightness difference • Hue difference
Tessellation objects	<ul style="list-style-type: none"> • Number of objects • Number of vertices • Object line length 		<ul style="list-style-type: none"> • Object size • Line segment length 	<ul style="list-style-type: none"> • Brightness difference • Hue difference
Field-based data	<ul style="list-style-type: none"> • Number of objects • Number of vertices • Object line length 		<ul style="list-style-type: none"> • Line segment length 	
All objects	<ul style="list-style-type: none"> • Number of object types • Number of objects • Number of vertices • Object line length 	<ul style="list-style-type: none"> • Proximity indicator • Proximity value 		

The readability measures used should ideally be defined based on the visual presentation of the data. Therefore, insignificant points in the objects (from a visual perspective) were removed with Douglas and Peucker’s algorithm [48] using a threshold of 1.0 meter.

3.5.1. Measures of the Amount of Information

The measures of the amount of information are defined in a map region; in the present study, this region is the entire map sample (3×2 cm). All the measures, apart from the number of object types, should be normalized to the size of the region in map space. Thus, all the values of these measures are divided by 6 cm^2 .

* *Number of objects* is the total number of objects in the region. Compound objects consisting of multiple objects were not allowed; thus, each geometric object was counted as one object. Furthermore, each link in a network was defined as one object.

* *Number of vertices* is the total number of object break-points for all objects in the region.

* *Object line length* is the total line length of all objects in the region. The total length of the boundary is used for area objects, while the boundary of the minimum-bounding rectangle of the point symbol is used for point objects.

* *Number of object types* is the number of all object types in the region. An object type is defined as a group of objects that have the same symbol style.

3.5.2. Measures of Spatial Distribution

* *Spatial distribution of objects* (HI_{SD_obj}) is a normalized version of Li and Huang's [15] geometric measure (see earlier work by Sukhov [49,50]). The measure is based on the Voronoi cells of the objects and is defined as the following entropy (cf. [51]):

$$HI_{SD_obj} = \frac{\sum_{i=1}^n p_i \log p_i}{\sum_{i=1}^n \frac{1}{n} \log \frac{1}{n}} = \frac{\sum_{i=1}^n p_i \log p_i}{\log \frac{1}{n}} \quad (1)$$

where p_i is the ratio between the area of Voronoi cell i and the area of the map, and n is the number of objects.

* *Spatial distribution of vertices* (HI_{SD_ver}) is similar to the spatial distribution of objects; it is based on a Voronoi diagram of the vertices:

$$HI_{SD_ver} = \frac{\sum_{i=1}^k p_i \log p_i}{\log \frac{1}{k}} \quad (2)$$

where p_i is the relative size of Voronoi cell i , and k is the number of vertices.

* *Proximity value* (PV) determines whether disjoint objects are too close to each other:

$$PV = \frac{\sum_{i=1}^n \sum_{j=i+1}^n \text{Area of intersection between buffers around object } i \text{ and object } j}{\text{Area of region}} \quad (3)$$

where n is the number of disjoint objects. The buffer size is based on the symbol size and a required minimum separation of 0.3 mm between the symbols of the objects. Note that this measure only addresses disjoint objects; objects that are connected should not be added to the proximity value.

* The proximity indicator is defined as the number of object pairs for which the shortest distance between the objects is less than a set threshold value (0.2 mm). The objects in the pair must be disjoint. This measure is computed for distances between different minor objects, between minor objects and lines, and between different field-based objects.

3.5.3. Measures of Object Complexity

* *Object size* reflects the object size distribution. In many cases, the interest is not the size of the smallest object but whether there are many small objects. Therefore, we list all the objects according to size and use the size of the 30% percentile object (*i.e.*, 30% of the objects are smaller than the size of the measure).

* *Line segment length* concerns all the line segments in the line and area objects, and it reflects the line segment length distribution of all of these segments. In this study, we use the value of the 10% percentile as a measure of the line segment length.

3.5.4. Measures of Graphical Resolution

In the user studies, we used a topographical map database to generate the map samples. Therefore, the database was compiled by adhering to the standard limits of the minimum size of objects, *etc.* Therefore, we restricted the measures of graphical resolution to colors, particularly the measures of *brightness difference* and *hue difference*. Both refer to only one color per object; if an object is multi-colored, then the most dominant color is used. The measures are based on colors expressed in the RGB system, where each component (*red*, *green* and *blue*) is defined by a value between 0 and 255.

* *Brightness difference* (Δbr) is defined as the absolute difference in brightness (br_1 , br_2) for two neighboring objects (*cf.* [52]):

$$\Delta br = |br_1 - br_2|$$

where

$$br_1 = \frac{(red_1 \cdot 299) + (green_1 \cdot 587) + (blue_1 \cdot 114)}{1000} \quad (4)$$

red_1 *etc.* are the colour components for the first object, and br_2 is defined analogously.

* *Hue difference* (Δh) is obtained by (*cf.* [52]):

$$\Delta h = |red_1 - red_2| + |green_1 - green_2| + |blue_1 - blue_2| \quad (5)$$

The formulas for brightness and hue difference are defined for single neighborhood relationships. In this study, we used the mean values for all neighborhood relationships to describe the differences in brightness and hue in the map. We applied measures for different information types as follows:

- *minor objects*—each possible pair of non-disjoint minor objects and tessellation objects constitutes a neighborhood relationship;
- *line networks*—each possible pair of non-disjoint line network objects and tessellation objects constitutes a neighborhood relationship; and
- *area tessellations*—neighbors are simply neighboring polygonal objects.

3.5.5. Implementation of the Analytical Measures

The measures were implemented in a Java program based on the open-source packages JTS Topology Suite (JTS; [53]) and the OpenJUMP platform [54]. To create Voronoi regions (for evaluating the spatial distribution of points and objects), we used the c-program Triangle [55,56].

3.6. Composite Methods

The second aim of the evaluation is to compare composite methods to find the most appropriate option. The first task was to select the composite methods to be compared.

A common approach for compositing is to set up a number of criteria that must all be satisfied. In this study, the criterion for map readability is that the map's readability measures are less than a particular threshold. To test this approach, we included *threshold evaluation* as one of the composite methods. Another common composite approach we would like to evaluate is a linear combination of measures. Therefore, we included the composite method *multiple linear regression*.

Multiple linear regression requires that the map samples used for creating the regression relationship have a numerical value that describes the map readability (as our perceived readability values). One could anticipate a situation in which the map samples are classified as either readable or non-readable. In such a case, traditional multiple linear regression is not feasible. One composite method that could use such a training set is the *support vector machine* (SVM; [57]). In the study, we would like to determine whether we lose information by using a training set only based on a classification (rather than our perceived readability values); therefore, we included SVM in our study.

Support Vector Machine

Support vector machines (SVMs) are a supervised learning technique. SVM was first developed in the 1970s [57] but was not given much attention until the 1990s. SVMs have been used for classification problems in pattern recognition and object tracing. Originally, SVMs were designed to be used for only two classes. Now, however, several approaches have been proposed for multiclass classifications.

SVMs construct a dividing hyperplane based on the properties of training samples. The distance from the hyperplane to the nearest training data points (of each class) should be maximal. To find this hyperplane, so-called support vectors are used (see Figure 2). Not all training samples need to contribute to the hyperplane. For many data samples, however, it is not possible to completely separate the data by linear boundaries. In these cases, cost measures are introduced to “penalize” some data points that do not fit the linear border (*cf.* [58]). When a linear approach is not suitable, SVMs can use a nonlinear classification, which maps the feature vectors into a higher dimensional space where they may be more easily separated [59].

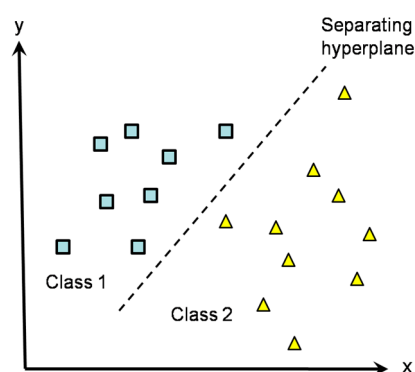


Figure 2. A hyperplane separates two classes in a two-dimensional space (x and y).

4. Result

4.1. Perceived Map Readability

The test answers were collected and grouped according to the test task. The answers to the readability evaluation (test parts 1, 3 and 5) were converted into numerical values of 1 to 4, corresponding to the readability assessments made by the participants (1 for “very difficult to read”, 2 for “difficult to read”, etc.). For each map sample, the mean of the numerical values of the readability was computed; this value is denoted as the *perceived readability value*. The perceived readability values were used in the evaluation of single measures.

For the evaluation of the measures, we also categorized the map samples as either *readable* or *non-readable* (Table 3). In this classification, we used a mean perceived readability value of 2.5 as the threshold value. This is a reasonable threshold because the participants graded these map samples as “very difficult to read” or “difficult to read” more frequently than “easy to read” or “very easy to read”.

The answers between the different groups of participants were quite consistent. If we compare the answers from only the Serbian students with those of the entire group, there was a 6% mismatch in the classification in readable/non-readable map samples.

Table 3. The number of readable and non-readable map samples.

Map Symbol Type	Non-Readable	Readable
NS	61	114
TS	49	126
All	110	240

Examples of perceived readability values, standard deviation of perceived readability values and perceived readability classes for the map samples shown in Figure 1a–h are listed in Table 4.

Table 4. Examples of perceived readability values (PRV), standard deviation of perceived readability values (Std) and readability classification for the eight map samples shown in Figure 1. It should be noted that maps b and f only differ in symbol styles. Thus, the map with the traditional style (TS) is regarded as more readable than the map with the pale style (NS).

Map Sample (from Figure 1)	a	b	c	d	e	f	g	h
Scale	1:50,000	1:50,000	1:10,000	1:10,000	1:50,000	1:50,000	1:50,000	1:10,000
Symbol style	TS	TS	TS	TS	NS	NS	NS	NS
PRV	2.48	1.83	3.21	3.03	2.30	1.64	3.40	2.96
Std	0.53	0.54	0.69	0.80	0.82	0.73	0.50	0.64
Readability classification	Non-readable	Non-readable	Readable	Readable	Non-readable	Non-readable	Readable	Readable

The results of the readability ranking (test part 2) were arranged as an ordered sequence of maps from the most readable to the least readable. The mean ranking by all the participants was computed for each sequence. This sequence was compared with the order of the same maps given by the perceived

readability values. The sequences only differed in a few instances. Hence, the assessments of readability were very similar when the participants evaluated the readability of one map at a time and when they evaluated a number of maps together. This result confirms the reliability of the perceived readability values.

The results of the map task (4), counting of the number of buildings or ancient remains on the maps, were transformed into the number of correct and incorrect answers for each map. The results showed a rather low proportion of correctly counted map objects. However, most of the time the error was only one or two objects. It could also be noted that most of the map samples used in this part of the test were perceived as non-readable (“very difficult to read” and “difficult to read”) in the readability assessment.

4.2. Correlation of Single Measures of Map Readability

The aim of the evaluation of single measures is to identify the measures that showed a good correlation with the perceived readability of the user test. This was performed by comparing the computed values of each measure with the *perceived readability value*. All of the map samples were used in the evaluation of the graphical resolution measures. For the other types of measures, we used only the TS map samples, as this symbol style was found to be more readable (discussed later in this section).

The evaluation is performed by setting up a regression formula of the form:

$$y = \alpha + \beta \cdot x \quad (6)$$

where x is the perceived readability value, y is the value of the measure, and α and β are the regression parameters. The number of perceived readability values that can be explained by the measure is computed by R :

$$R = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (7)$$

where

y_i is one perceived readability value;

\hat{y}_i is an estimate of the readability value based on the value of the measure using the regression (Equation (6));

\bar{y} is the mean value of all perceived readability values.

We also tested whether the perceived readability values were independent of the single measures. First, we introduced the null hypothesis that the readability values were independent of the single measure and performed a two-sided statistical test by computing the *p-value*. The *p-value* is the probability of obtaining the same values in the null hypothesis and in the data material. In practice, whether the regression parameter β is equal to zero is tested (*i.e.*, the perceived readability values are independent of the values of the measures). If the *p-value* is less than 0.01, then there is a strong argument against the null hypotheses.

4.2.1. Result

The evaluation result of the single measures is provided in Table 5. Generally, the measures of the amount of information had the highest R -values and best explain the readability of the maps. The

second-best category was the spatial distribution measure. Measures of object complexity and graphical resolution could not explain the map readability.

Table 5. Evaluation of single measures. Note that in some cases, the measure is equal to zero for a map sample (e.g., the measure *Spatial distribution of objects* is zero when no minor objects are present). These samples are then removed for evaluating this particular measure.

Readability Category	Readability Measure	Types of Information	R-value	p-value
Amount of information	Number of object types	All objects	0.52	$1.0 \cdot 10^{-13}$
Amount of information	Number of objects	All objects	0.47	$8.3 \cdot 10^{-11}$
Amount of information	Number of vertices	All objects	0.48	$1.7 \cdot 10^{-11}$
Amount of information	Object line length	All objects	0.62	$2.8 \cdot 10^{-20}$
Spatial distribution	Proximity indicator	All objects	0.40	$4.2 \cdot 10^{-8}$
Spatial distribution	Proximity value	All objects	0.26	$5.5 \cdot 10^{-4}$
Spatial distribution	Spatial distribution of objects	Minor objects	0.12	0.16
Spatial distribution	Spatial distribution of vertices	Minor objects	0.04	0.67
Object complexity	Object size	Minor objects	0.06	0.59
Object complexity	Line segment length	Line networks	0.11	0.21
Graphical resolution	Brightness difference	Tessellation objects	0.10	0.19
Graphical resolution	Hue difference	Tessellation objects	0.10	0.19

4.2.2. Discussion

Based on the *p*-values, we can reject the null hypothesis that the perceived readability values are independent of the values of the amount of information. This confirms previous findings by Phillips and Noyes [5], Rosenholtz *et al.* [24], and Stigmar and Harrie [29]. However, even if the perceived readability values depend on the measured values, the degree of explanation is low (the *R*-values are comparatively small); it is not possible to explain readability only by measuring the amount of information. This statement can be illustrated by the best overall measure—object line length. For this measure, the following regression relationship was obtained:

$$\text{perceived_readability_value} = 3.24 - 0.035 \cdot \text{object_line_length} \quad (8)$$

From Figure 3, we can see that map samples with long object line lengths generally have low perceived readability. However, the opposite case is not that clear. A map with short object line lengths could have low perceived readability. The difficulty in reading these maps may stem from other map properties (e.g., cluttering by point objects). Specifically, it is not possible to identify a relationship between map readability and a single measure; thus, defining the readability based on composites of measures is needed.

The object complexity and graphical resolution may poorly explain the perceived readability because the cartographic data were the appropriate resolution and symbol style for the scale used. If we had used much more detailed data in maps of the same scale, then this type of measure may have been more important.

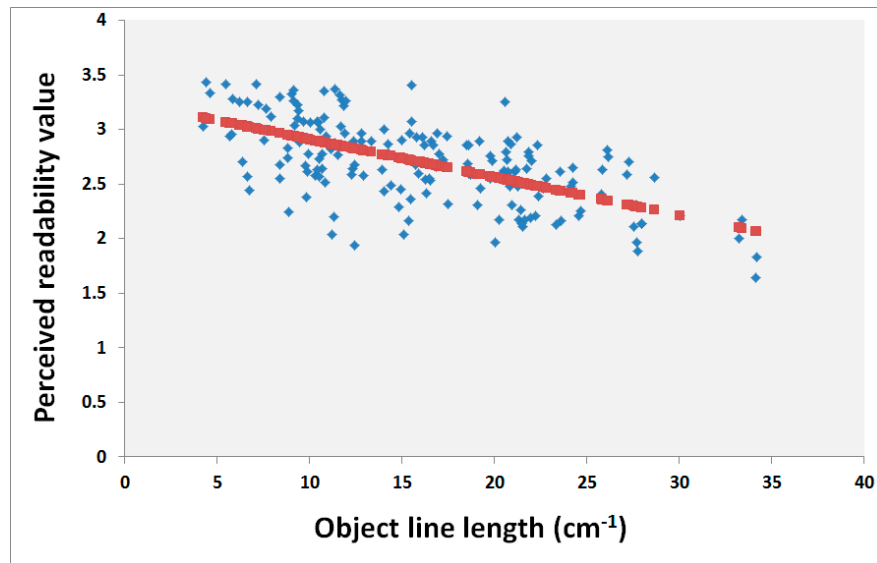


Figure 3. Relationship between the perceived readability value and the object line length.

4.2.3. Evaluating Brightness and Hue Measures

The color measures in this study were used to compare the pale symbol style (NS) with the traditional style (TS). The evaluation showed that the NS map samples were more difficult to read than the TS samples (Table 3). The color measures for the tessellation objects in the NS maps were $\overline{\Delta br} = 8$ (mean difference in brightness for all tessellation objects and map samples) and $\overline{\Delta h} = 35$ (mean difference in hue for all tessellation objects and map samples); the corresponding values for the TS maps were $\overline{\Delta br} = 15$ and $\overline{\Delta h} = 52$.

The difference in the perceived readability values (for maps of the same region but with different symbol styles) is particularly large for maps that only contained land use information. The maps shown in Figure 4 had perceived readability values of 1.9 (NS) and 3.0 (TS). For these maps, the color measures for the tessellation objects were $\overline{\Delta br} = 7$ and $\overline{\Delta h} = 28$ for the NS map and $\overline{\Delta br} = 25$ and $\overline{\Delta h} = 84$ for the TS map.

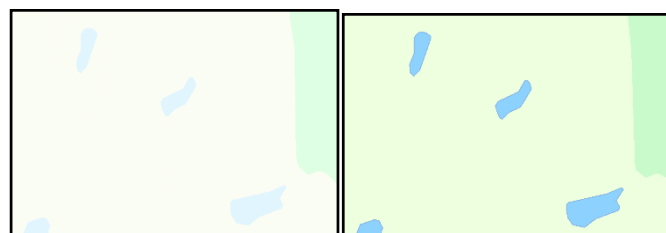


Figure 4. The same map area with different symbol styles. The TS map sample (**right**) was found to be more readable than the NS map sample (**left**).

4.3. Evaluation of Composites of Measures

4.3.1. Selection of Measures in the Composites

The choice of measures for the composites was based on the following criteria:

- (1) measures from as many categories (*i.e.*, amount of information, spatial distribution, object complexity and color) as possible should be included;
- (2) only measures for which we can reject the null hypotheses that the perceived readability value is independent of the measure are included (*i.e.*, the *p*-value must be less than 0.01 in Table 5);
- (3) the correlation between the measures should not be too high.

Because the measures of the amount of information best explain the perceived readability (*cf.* Table 5), we decided to select two measures from this category. The three best candidates were *object line length*, *number of vertices* and *number of object types*. The second-best category was spatial distribution. The best candidate is the *proximity indicator*. All of these measures are defined for all objects (*cf.* Table 5). There are no measures in the categories of object complexity and graphical resolution that meet the second criteria.

We investigated the correlation between the four candidate measures using the following correlation formula:

$$\text{correlation}(m_1, m_2) = \frac{\sum_{j=1}^{n_m} (m_{1j} - \bar{m}_1)(m_{2j} - \bar{m}_2)}{\sqrt{\sum_{j=1}^{n_m} (m_{1j} - \bar{m}_1)^2 \sum_{j=1}^{n_m} (m_{2j} - \bar{m}_2)^2}} \quad (9)$$

where m_1 and m_2 are two measures, n_m is the number of map samples, m_{1j} is the value of measure 1 for map sample j , and \bar{m}_1 is the mean value of measure 1 for all map samples. This coefficient is equal to 1 for a perfect correlation and 0 for no correlation.

Table 6 provides the correlation coefficient values for the four candidate measures. High correlations were found between the measures describing the amount of information. For example, the correlation between *number of vertices* and *object line length* was 0.92. Based on these correlation values, we decided not to use the measure *number of vertices*. The final list of measures was:

- m_1 = Object line length (all objects)
- m_2 = Number of object types (all objects)
- m_3 = Proximity (all objects).

Table 6. Correlations (Equation (10)) between measures. The computations are based on the 175 TS map samples.

Readability Measure	a	b	c	d
(a) Object line length	1.0			
(b) Number of vertices	0.92	1.0		
(c) Number of object types	0.72	0.71	1.0	
(d) Proximity indicator	0.37	0.20	0.11	1.0

4.3.2. Composite I: Threshold Evaluation

To compute the threshold value for each measure, we utilized the regression relationships computed in Section 4.2. The computations of the threshold value were based on where the regression line intersected the threshold for a readable map (*i.e.*, when $y = 2.5$ in Equation (6); *cf.* categorization of readable maps in Section 4.1). By using this approach, we would obtain a threshold value of 21.7 cm^{-1} for the object line length. According to Figure 3, these values classify too many map samples as non-readable (which is also the case for the other two measures). We made some tests of different values and found that an

increase of 10% of the threshold values where appropriate (*i.e.*, gave many correct classified map samples); hence, the threshold values (T_i) is defined as follows:

$$T_i = \frac{2.5 - \alpha_i}{\beta_i} \cdot 1.10 \quad (10)$$

where α_j and β_j are the regression parameters estimated for measure j (*cf.* Equation (6)).

Table 7 contains the estimated threshold values. This table also contains information on the number of maps that were classified as non-readable according to each measure. The total number of map samples that did not meet at least one of the threshold values was 39, which should be compared with the total number of map samples that were perceived as non-readable in the user test (49).

Table 7. Threshold values for classifying a map sample as non-readable.

Measure	Threshold	Number of Map Samples Classified as Non-Readable
Object line length	>23.9 cm ⁻¹	23
Number of object types	>17.4	19
Proximity indicator	>71.2	11

4.3.3. Composite II: Multiple Linear Regression

The composite method of multiple linear regression was included to investigate whether a linear combination of several measures could describe the readability of the map samples. The following regression formula was used:

$$perceived_readability_value = \alpha + \beta_1 \cdot m_1 + \beta_2 \cdot m_2 + \dots + \beta_n \cdot m_n \quad (11)$$

where

α , β_i are regression parameters, m_i is the value for measure i , and n is the number of measures used in the regression. The regression parameters are determined by a least-squares' fit of the data sample, in this case, by the perceived readability values and the measures in Section 4.3.1.

After the parameters were determined, the regression parameters were used to classify the map samples. This classification was made by applying Equation (11) to all map samples. If the estimated value was less than 2.5, then the map sample was classified as non-readable. The computations for multiple linear regression was performed by a Matlab script.

4.3.4. Composite III: Support Vector Machine

In this study, we used the linear approach of SVMs, in which two classes (non-readable and readable) were available, *i.e.*, we used settings similar to Figure 2. We used all the map samples as a training set. For the computations, we used the SVM tool in the Bioinformatics Toolbox in Matlab [60], along with personal Matlab scripts.

4.3.5. Results of the Measure Composites

Table 8 presents the result of the measure composites. Note that we used the same map samples for training and testing the composite methods (*e.g.*, the same map samples to determine the regression parameters that are later used for the evaluation). By doing this, we overestimate the amount of

correctly classified map samples. For the composite methods, multiple linear regression, and support vector machine, we performed several tests in which we divided the map samples into two parts, for example, 145 map samples for training and 30 for evaluation. Typically, the number of correctly classified map samples is approximately 2%–5% lower than that shown in Table 8. However, the random selection of map samples in each category substantially affects the results (a single evaluation can vary between 65% and 95% correctly classified map samples). Therefore, it is not straightforward to compare the composite methods, which was the main aim of this study; hence, we included all the map samples in both the training and evaluation datasets.

Table 8. Percentage of correctly classified map samples for each composite method using the measures m_1 = object line length, m_2 = number of object types and m_3 = proximity indicator.

Measures Used in the Composites	Threshold Evaluation	Multiple Linear Regression	Support Vector Machine
m_1	76	75	73
m_1, m_2	78	78	74
m_1, m_2, m_3	78	83	79

In the supplementary material, the results of the three composite methods are listed for each map sample. From this list, we can conclude that there are minor differences between the methods classifying readable and non-readable samples.

5. Discussion

5.1. User Tests

The data acquisition process, *i.e.*, the user test, is important. In this study, we used a web-distributed test, which provided us with a large number of participants. This extensive material has been valuable when performing the described evaluations. However, every method has disadvantages. One disadvantage of our user test is that we were not able to observe or talk to the participants during the test. At the end of the test, we provided a comment box for which the participants were able to comment on the test or maps. However, only a few participants took advantage of this feature. Therefore, we do not know the attitudes of most of the participants towards the maps, which might have provided valuable qualitative data. Another disadvantage of user tests based on judgment is that they might answer differently than they would in real life [61]. In future studies, it is therefore important to include other user test methods to reflect the different aspects of participants' performances.

In the user studies we divided the participants into seven groups where the groups studied different maps. If there are biases between the groups this will potentially affect the classification of the map samples, especially since we used a single threshold value to distinguish between readable and non-readable maps. These circumstances are likely the reason that seemingly similar map samples are classified differently (see *e.g.*, the map samples Trad10_GL1_04, Trad10_GL2_04 and Trad10_GL3_04 in the supplementary material).

In the user test, we used a scale of four possible answers, where we later classified two as readable and the other two as non-readable maps. In this way, we forced the user to decide whether the map was

readable or not. Forcing the user to answer without a neutral choice is debatable, and we are aware that our choice of excluding a neutral choice might bias our results. Furthermore, we used the mean value of all answers to decide whether a map was readable or non-readable. We also tested with using the median values. The difference between these measures was fairly small. Five readable map samples (using mean values) were classified as non-readable using medians, and the same amount was misclassified in the other direction. All of the map samples that were classified differently using mean and medians had a perceived readability value between 2.41 and 2.58. In our study we preferred to use the mean value since it provided us with the possibility to use standard multiple linear regression with the (mean) perceived readability values as dependent variable (*cf.* Equation (11)).

5.2. Composites of Measures

There are no major differences between the results of the three composite methods (Table 8). The percentage of correctly classified maps is mainly dependent on the ability or inability of the measures to explain readability. However, there are a few things that we should note. The threshold evaluation is appealing because it is conceptually easy and logical. If all the threshold constraints are met, then the map is simply classified as readable. A challenge of this method is setting the threshold. In this study, we set the threshold values according to a common formula for all measures (Equation (10)). We tested the threshold values through manual modifications to obtain a somewhat better result, but we preferred to continue with Equation (10) in the evaluation. In principle, it would also be possible to write an optimization routine to define the optimal threshold values (according to the map samples).

The results of *multiple linear regression* (MLR) and *support vector machine* (SVM) are similar (Table 8). There are well-known methods for determining the regression parameters in MLR and the hyperplanes in SVM. One advantage of SVM is its ability to handle the situation in which training datasets only have information on whether the training map samples are readable or non-readable (a standard MLR requires numerical readability values). Additionally, the MLR method is likely more sensitive for outliers in the test data, which could be map samples with uncommon measurement values.

We also performed experiments with the artificial neural network Biased ARTMAP [62,63]. Biased ARTMAP is an unsupervised learning classification (clustering of map samples in measurement space) followed by supervised classification (determining whether each map sample cluster contains readable or non-readable map samples). However, the use of Biased ARTMAP produced significantly worse results than the other composite methods, possibly because Biased ARTMAP, and similar methods, relies on clusters in the input data. However, the readable/non-readable map samples do not form clusters in the measurement space (*cf.* Figure 2) but rather determine the perceived readability by threshold values.

5.3. Evaluation of the Study

In our composite study (Table 8), approximately 80% of the map samples were classified correctly based on the three best available measures. The accuracy was possibly limited by the following factors:

- (1) The readability measures are inadequate.
- (2) The symbol design was not good.
- (3) The best composites methods were not used.

- (4) We must consider the semantic aspects of map reading.
- (5) We should have used fuzzy classification rather than crisp classification.

A short discussion of each of these factors is stated below.

One could argue that we included all relevant measures in this study. However, raster-based measures (e.g., [12,23]) and measures of complexity and graphical resolution (developed in e.g., [17]) are missing. Of course, we may still need to develop new measures that are missing in the literature. If we study map samples that were perceived as non-readable (by the participants) but classified as readable (by the measures), then we find samples with regions of dense lines (foremost railway lines) and point objects (*cf.* Figure 5 and supplementary material). Here, we would need better measures to include these properties. One might also argue that the problem of the readability of these maps is not related to geometry but to the symbol style; hence, readability measures that better capture the symbol style are needed. It might also be so that the maps in Figure 5 are perceived as non-readable because of that there are several similar objects in a neighborhood, which complicate the search for a spatial object (see e.g., [5,26]).

The design of the symbols is surely an important aspect of map readability. In our result we revealed that map samples including cadastre boundaries often were classified as non-readable. This is especially the case for small real-estates such as the right map in Figure 5. In this case the chosen map symbol for the cadastre boundary is not appropriate for the size of the cadastre units.



Figure 5. Maps that were often misclassified as readable: **(left)** maps with dense lines, **(middle)** maps with dense/overlapping symbols and **(right)** map with bad symbol types (of the cadastre boundaries).

An interesting research direction is to identify a pattern between misclassified maps and the properties of the map samples (*cf.* supplementary material). We studied the relationships between the classification between the map samples and the outcomes of the study. A χ^2 test (with 5% significance) indicates that map samples that are dense and have dense lines is more likely to be misclassified (than map samples in general), but it is hard to make any clear conclusion. If we would have used a finer categorization of the properties of the map samples, we could perhaps have been able to identify the map samples in Figure 5; however, our categorization (which occurred before the analysis) was too broad.

We can conclude that the three composite methods used provided similar results, despite the differing foundations of the methods. We can also conclude that composite methods based on clustering in measurement space, such as Biased ARTMAP [62,63], are not appropriate. Whether there are other composite methods that would provide substantially better results are, to the authors' knowledge, not very likely.

The readability measures in this study are at the syntactic level (*cf.* Section 2.5), *i.e.*, they measure graphical complexity. By studying the examples of misclassified map samples, we can observe the

following. Map samples that are perceived as readable but classified as non-readable often cover a common geographic pattern (Figure 6 and supplementary material). For map samples that cover more unusual geographic regions, the situation is often the opposite: the map samples are perceived as non-readable but classified as readable (Figure 6 and supplementary material). Hence, it seems as that the map reader's ability to interpret the map in a geographical context are important, *i.e.*, we cannot neglect the semantic aspects of the map samples when we study map reading. This result is indicated even though we deliberately chose map regions that did not contain strange geographic features (*cf.* Section 3.2). One can argue that this result is an obvious result, e.g., it is well-established that readability of images in general is affected by the understanding of the context [43].

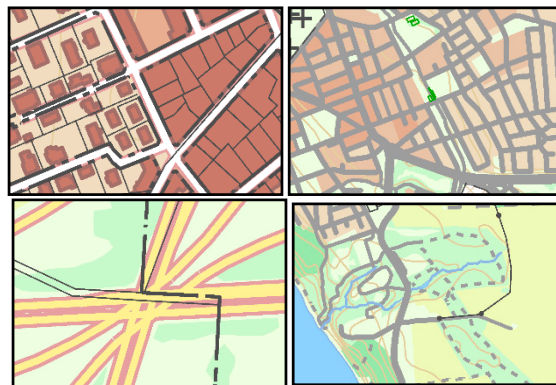


Figure 6. The two maps on the top are perceived as readable but are classified as non-readable; they both include a common geographic pattern. The two maps on the bottom are perceived as non-readable but are classified as readable; they both cover a more unusual geographic pattern.

Finally, in our study, we classified all the map samples as either readable or non-readable. By studying the perceived readability values for misclassified map samples (*cf.* supplementary material), we can conclude that many misclassified map samples have a perceived readability value close to the threshold (as described in Section 4.1, we used a perceived readability value of 2.5 as the threshold). One could experiment to see whether the use of a fuzzy classification scheme would improve the results, but this objective is outside the scope of our paper.

6. Conclusions

In this study, we evaluated *single measures* of map readability and methods for describing readability by the *composites of measures*. In the evaluation of the single measures, we found that measures of the amount of information were correlated with perceived map readability, which confirms the results of Phillips and Noyes [5], Rosenholtz *et al.* [24] and Stigmar and Harrie [29]. The best correlation was given by *object line length*, *number of object types* and *number of vertices*. For the measures of spatial distribution, *proximity indicator* and *proximity value* showed the best results. We could not reject the hypothesis that the perceived readability value is independent of the measures of object complexity or graphical resolution. However, the map samples used in the tests all obeyed the basic rules of object complexity (proper level of generalization) and graphical resolution (suitable

symbol style). Finally, it seems as our study lacked appropriate measures for identify non-readable maps due to dense lines and dense/overlapping symbols.

The result shows that the use of measure composites is better for describing map readability than single measures. By using the best measure *object line length*, we could correctly classify approximately 75% of the map samples as readable/non-readable. When we added two other measures, *number of object types* and *proximity indicator*, the amount of correctly classified samples increased to approximately 80%. We could not identify any major differences in the three composite methods we evaluated. However, we can conclude that the threshold evaluation method would require more work on optimizing the threshold values. We could also conclude that the support vector machine is a suitable method because it only requires a test dataset in which each map is classified as either readable or non-readable (while the multiple linear regression method requires numerical readability values).

By studying the map samples that were not correctly classified by the composite methods, we can conclude the following. It seems as the map reader's ability to understand the geographic context (represented in the map samples) affects his/her ability to read the map; that is, we cannot fully explain the map readability by the graphical complexity of the map.

A practical recommendation, based on this study, is that map producers should not solely use graphical resolution readability measures in their map specifications. Producers should complement these measures with measures of amount of information, such as *object line length*. Furthermore, the map generalization process should be triggered and controlled by readability measures for the amount of information and possibly the spatial distribution (which is sometimes already considered).

Our comparison of the two symbol styles (denoted TS and NS, *cf.* Section 3.2 and Figure 4) indicates that the pale style is less readable than the traditional style. The color differences in these two styles are also nicely captured by the hue and brightness measures. This result is interesting in the context of backdrop mapping on the web. Commercial services (Google Maps, Bing Maps, *etc.*) have used pale colors for a long time because they allow users to add their own thematic information on top of the maps. In recent years, many of the national mapping agencies have also provided services with pale colors to support the addition of thematic information. This development is positive overall, but one must be careful to not lose the readability of maps. According to our findings, readability loss is a risk.

Acknowledgments


Financial support from the Vinnova project Planeringsportalen, Lund University, Lantmäteriet and from the Erasmus Mundus External Cooperation Window—Basileus Project is gratefully acknowledged. Thanks to Florian Sallaba for support with the computations of SVM, Joakim Eriksson for guidance on user tests, Markus Nyström for creating the web interface for the user studies, Anders Ek for support with the symbol styles, and of course thanks to all the participants in the user tests. Lantmäteriet and the city of Helsingborg are acknowledged for providing geographic data. The authors also thank the anonymous reviewers for good comments that improved the paper.

Author Contributions

Lars Harrie and Hanna Stigmar designed the study, did the literature review and wrote the paper together. Lars Harrie computed the measures and performed the evaluation. Hanna Stigmar designed and conducted the user survey. Milan Djordjevic worked with the evaluation.

Appendix

Please study the map for a short while. Then estimate to what degree you find it readable (based on the objects, their properties, organization and symbolization).



How readable is the map?

a. Very easy to read

b. Easy to read

c. Difficult to read

d. Very difficult to read

Appendix A. Example of page in readability evaluation in the web questionnaire.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Sherman, A.L. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*; Ginn & Co.: Boston, UK, 1893.
2. Gray, W.S.; Leary, B. *What Makes a Book Readable*; Chicago University Press: Chicago, IL, USA, 1935.
3. DuBay, W.H. *The Principles of Readability*, 2004. Available online: <http://www.impact-information.com/impactinfo/readability02.pdf> (accessed on 19 October 2014).

4. Rosenholtz, R.; Li, Y.; Mansfield, J.; Jin, Z. Feature congestion: A measure of display clutter. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR, USA, 2–7 April 2005; pp. 761–770.
5. Phillips, R.J.; Noyes, L. An investigation of visual clutter in the topographic base of a geological map. *Cartogr. J.* **1982**, *19*, 122–132.
6. Wolfe, J.M. Guided search 2.0: A revised model of visual search. *Psychon. Bull. Rev.* **1994**, *1*, 202–238.
7. Oliva, A.; Mack, M.L.; Shrestha, M.; Peeper, A. Identifying the perceptual dimensions of visual complexity of scenes. In Proceedings of the 26th Annual Meeting of the Cognitive Science Society, Chicago, IL, USA, 5–7 August 2004.
8. Töpfer, F.; Pillewizer, W. The principles of selection. *Cartogr. J.* **1966**, *3*, 10–16.
9. Schnur, S.; Bektaş, K.; Salahi, M.; Çöltekin, A. A comparison of measured and perceived visual complexity for dynamic web maps. In Proceedings of the GIScience, Zurich, Switzerland, 14 September 2010–17 September 2010.
10. Woodruff, A.; Landay, J.; Stonebraker, M. Constant information density in zoomable interfaces. In Proceedings of the Advanced Visual Interfaces '98, L'Aquila, Italy, 24–27 May 1998; pp. 57–65.
11. MacEachren, A.M. Map complexity: Comparison and measurement. *Am. Cartogr.* **1982**, *9*, 31–46.
12. Fairbairn, D. Measuring map complexity. *Cartogr. J.* **2006**, *43*, 224–238.
13. Frank, A.U.; Timpf, S. Multiple representations for cartographic objects in a multi-scale tree—An intelligent graphical zoom. *Comput. Graph.* **1994**, *18*, 823–829.
14. Bjørke, J.T. Framework for entropy-based map evaluation. *Cartogr. Geogr. Inf. Syst.* **1996**, *23*, 78–95.
15. Li, Z.; Huang, P. Quantitative measures for spatial information of maps. *Int. J. Geogr. Inf. Sci.* **2002**, *16*, 699–709.
16. Hangouet, J.F. *Voronoi Diagrams on Segments—Properties and Tractability for Generalization Purposes*; Technical Report for AGENT; Cogit, IGN: Saint-Mandé, France, 1998.
17. AGENT. Project Homepage. Available online: <http://agent.ign.fr> (accessed on 22 January 2015).
18. João, E.M. *Causes and Consequences of Map Generalisation*; Taylor & Francis: London, UK, 1998.
19. McMaster, R.B.; Shea, K.S. *Generalization in Digital Cartography. Resource Publications in Geography*; Association of American Geographers: Washington, DC, USA, 1992.
20. Mackaness, W.A.; Mackechnie, G.A. Automating the detection and simplification of junctions in road networks. *GeoInformatica* **1999**, *3*, 185–200.
21. Spiess, E. The need for generalization in a GIS environment. In *GIS and Generalization, Gisdata 1*; Müller, J.-C., Lagrange, J.-P., Weibel, R., Eds.; Taylor & Francis: London, UK, 1995; pp. 31–46.
22. Eley, M.G. Color-layering and the performance of the topographic map user. *Ergonomics* **1987**, *30*, 655–663.
23. Jégou, L.; Deblonde, J.-P. Vers une visualisation de la complexité de l'image cartographique. *Cybergeo: Eur. J. Geogr.* **2012**, *600*, doi:10.4000/cybergeo.25271.
24. Rosenholtz, R.; Li, Y.; Nakano, L. Measuring visual clutter. *J. Vis.* **2007**, *7*, 1–22.
25. He, S.; Cavanagh, P.; Intriligator, J. Attentional resolution and the locus of visual awareness. *Nature* **1996**, *383*, 334–337.
26. Wolfe, J.M.; Horowitz, T.S. What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* **2004**, *5*, 1–7.

27. Moacdieh, N.; Sarter, N. Display clutter a review of definitions and measurement techniques. *Hum. Factors* **2014**, *3*, doi:10.1177/0018720814541145.
28. Lohrenz, M.C.; Trafton, M.R.; Beck, M.R.; Gendron, M.L. A model of clutter for complex, multivariate geospatial displays. *Hum. Factors* **2009**, *51*, 90–101.
29. Stigmar, H.; Harrie, L. Evaluation of analytical measures of map readability. *Cartogr. J.* **2011**, *48*, 41–53.
30. Swiss Society of Cartography. *Topographic Maps—Map Graphics and Generalisation*; Cartographic Publication Series No. 17; Swiss Society of Cartography: Bern, Switzerland, 2005.
31. National Land Survey of Sweden. *HMK Handbok Geodesi, Kartografi, Lantmåteriverket, Sweden*; National Land Survey of Sweden: Gävle, Sweden, 1996. (In Swedish)
32. Stoter, J.; Burghardt, D.; Duchêne, C.; Baella, B.; Bakker, N.; Blok, C.; Pla, M.; Regnauld, N.; Touya, G.; Schmid, S. Methodology for evaluating automated map generalization in commercial software. *Comput. Environ. Urban Syst.* **2009**, *33*, 311–324.
33. Bard, S. Quality assessment of cartographic generalisation. *Trans. GIS* **2004**, *8*, 63–81.
34. Ruas, A.; Duchêne, C. A prototype generalisation system based on the multi-agent system paradigm. In *Generalisation of Geographic Information: Cartographic Modelling and Applications*; Mackaness, W.A., Ruas, A., Sarjakoski, L.T., Eds.; Series of International Cartographic Association, Elsevier Science: Amsterdam, The Netherlands, 2007; pp. 269–284.
35. Mackaness, W.A.; Ruas, A. Evaluation in the map generalisation process. In *Generalisation of Geographic Information: Cartographic Modelling and Applications*; Mackaness, W.A., Ruas, A., Sarjakoski, L.T. Eds.; Series of International Cartographic Association, Elsevier Science: Amsterdam, The Netherlands, 2007; pp. 89–111.
36. Stoter, J.; Zhang, X.; Stigmar, H.; Harrie, L. Evaluation and usability of map generalisation outputs. In *Abstracting Geographic Information in a Data Rich World, Lecture Notes in Geoinformation and Cartography*; Burghardt, D., Duchêne, C., Mackaness, W., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 259–297.
37. Brewer, C.A.; Buttenfield, B.P. Framing guidelines for multi-scale map design using databases at multiple resolutions. *Cartogr. Geogr. Inform. Sci.* **2007**, *34*, 3–15, doi:10.1559/152304007780279078.
38. Brewer, C.A.; Buttenfield, B.P. Mastering map scale: Workloads using display and geometry change in multi-scale mapping. *Geoinformatica* **2010**, *14*, 221–239, doi:10.1007/s10707-009-0083-6.
39. Roth, R.E.; Brewer, C.A.; Stryker, M.S. A Typology of Operators for Maintaining Readable Map Designs at Multiple Scales. *Cartogr. Perspect.* **2011**, *68*. Available online: <http://cartoperspectives.org/carto/index.php/journal/article/view/cp68-roth-et-al/18> (accessed on 12 January 2015).
40. Shea, K.; McMaster, R. Cartographic generalization in a digital environment: When and how to generalize. In Proceedings of the AutoCarto, Baltimore, MD, USA, 2–7 April 1989.
41. Christophe S. Creative colours specification based on knowledge (COLorLEGenD system). *Cartogr. J.* **2011**, *48*, 138–145.
42. MacEachren, A.M. *How Maps Work, Representation, Visualization, and Design*; The Guilford Press: New York, NY, USA, 1995.

43. Neider, M.B.; Zelinsky, G.J. Cutting through the clutter: Searching for targets in evolving complex scenes. *J. Vis.* **2011**, *14*, 1–16.
44. Sester, M.; Arsanjani, J.J.; Klammer, R.; Burghardt, D.; Haunert, J.H. Integrating and generalizing volunteered geographic information. In *Abstracting Geographic Information in a Data Rich World*; Lecture Notes in Geoinformation and Cartography; Burghardt, D., Duchene, C., Mackaness, W., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 119–155.
45. Al-Bakri, M.; Fairbairn, D. Using geometric properties to evaluate possible integration of authoritative and volunteered geographic information. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 349–370, doi:10.3390/ijgi2020349.
46. Van Smaalen, J.W.N. Automated Aggregation of Geographic Objects: A New Approach to the Conceptual Generalisation of Geographic Databases. Ph.D. Dissertation, Wageningen University and Research Centre, Wageningen, The Netherlands, 2003.
47. Harrie, L.; Stigmar, H. An evaluation of measures for quantifying map information. *ISPRS J. Photogramm. Remote Sens.* **2009**, *65*, 266–274.
48. Douglas, D.H.; Peucker, T.K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Can. Cartogr.* **1973**, *10*, 112–122.
49. Sukhov, V.I. Information capacity of a map entropy. *Geodesy Aerophotogr.* **1967**, *10*, 212–215.
50. Sukhov, V.I. Application of information theory in generalization of map contents. *Int. Yearb. Cartogr.* **1970**, *10*, 41–47.
51. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Champaign, IL, USA, 1964.
52. Techniques for Accessibility Evaluation and Repair Tools. W3C Working Draft. Available online: <http://www.w3.org/TR/AERT> (accessed on 19 October 2014).
53. Java Topology Suite. Available online: <http://www.vividsolutions.com/jts/jtshome.htm> (accessed on 19 October 2014).
54. OpenJUMP. Available online: <http://www.openjump.org/> (accessed on 19 October 2014).
55. Shewchuk, J.R. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In *Applied Computational Geometry: Towards Geometric Engineering, Lecture Notes in Computer Science*; Lin, M.C., Manocha, D., Eds.; Springer-Verlag: Berlin, Germany, 1996; Volume 1148, pp. 203–222.
56. Shewchuk, J.R. Delaunay refinement algorithms for triangular mesh generation. *Comput. Geom. Theor. Appl.* **2002**, *22*, 21–74.
57. Vapnik, V. *Estimation of Dependences Based on Empirical Data [in Russian]*; Nauka: Moscow, Russian, 1979.
58. Fradkin, D.; Muchnik, I. Support vector machines for classification. In *Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*; Abello, J., Carmode, G., Eds.; American Mathematical Society: Providence, RI, USA, 2006; Volume 70, pp. 13–20.
59. Tso, B.; Mather, P.M. *Classification Methods for Remotely Sensed Data*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2009.
60. Matlab, Bioinformatics Toolbox 3.6. Available online: <http://www.mathworks.com/products/bioinfo/> (accessed on 19 October 2014).

61. Cleveland, W.S.; McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc.* **1984**, *79*, 531–554.
62. Carpenter, G.A.; Grossberg, S.; Reynolds, J.H. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Netw.* **1991**, *3*, 565–588.
63. Carpenter, G.A.; Gaddam, C.S. Biased ART: A neural architecture that shifts attention toward previously disregarded features following an incorrect prediction. *Neural Netw.* **2010**, *23*, 435–451.

© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).