Abstracting and Formalizing Model Generalization

Matthias Bobzien and Dieter Morgenstern
Institute of Cartography and Geoinformation
University of Bonn
Meckenheimer Allee 172, D-53113 Bonn, Germany
Phone: +49-228-73-3529
Fax: +49-228-73-7756
E-Mail: {bobzien,mostern}@ikg.uni-bonn.de

Keywords: model generalization, generalization function, abstracted generalization, formalized generalization, integrity constraints, morphism.

Abstract

In this paper the generalization process is interpreted as a function from a (ungeneralized) source set to a (generalized) target set, both sets being sets of geo-spatial features. The source set holds certain properties such as geometrical, topological and non-spatial properties as well as their combinations. These properties must be preserved by the generalization function so that the target set obtains the same properties as the source set. Holding these properties invariant, the generalization function forms a morphism between the two sets of geo-spatial features. The invariant properties are formalized and as a result integrity constraints are found. These allow a consistency check of a set of generalized data as well as support for the construction of the generalization function.

1 Introduction

Model generalization is the process of deriving a digital landscape model (DLM) of lesser resolution from a DLM of higher resolution [Mül91, MS99]. In the process of producing a map, model generalization precedes cartographic generalization [Sch99]. While model generalization mainly consists of statistical and filtering processes [BW88] it has shown that cartographic processes, such as amalgamation [Bob01], geometry-type change [Bob02, Sch02] and line simplification [MS99] are needed, too.

Within the last years the generalization research group at the Institute for Cartography and Geoinformation at the University of Bonn successfully worked

on automated model generalization. At the first stage of the project, topographic data based on ATKIS[1] Basis-DLM[2] has been generalized to DLM250[3] data [MS99]. A working prototype was developed. At the second stage of the project (the present stage) the acquired data of the first stage (the DLM250 data) will be generalized to DLM1000[4] data.

For obvious reasons the findings and results of the first stage will be used for the second stage of the project. This comprises data structures, generalization rules and generalization algorithms. (Still there are plenty of aspects that cannot be transferred one-to-one from the first stage to the second one.) To find the common aspects of both generalization stages we had to abstract the problem of model generalization. The result is the abstracted model generalization which is independent of a specific change of level of detail. The abstraction gets hold of the commonalities of both steps of model generalization.

Reaching the abstraction level, the idea emerged to go a step further and to formalize the abstraction. This would allow others who work on generalization applications to benefit from results which were initially conceived for model generalization only. Conceivable applications comprise generalization of geological maps, sea maps, road maps and possibly topographic maps.

The aim of this working paper is to formalize those aspects of the problem of model generalization that can be of use to other generalization applications. As a result integrity constraints emerge that can be verified easily and automatically. Figure 1 (a) depicts a usual work flow for problem solving: starting from a given problem (for example model generalization) one develops a conception to solve the problem and then implements the conception into a program. The conception includes data structures and algorithms.

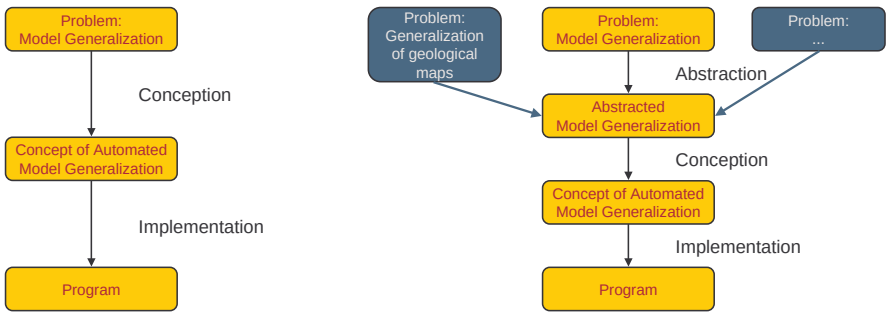

Figure 1: *Work flow for solving the generalization problem.*

To make the concept usable for other applications one has to insert an intermediate step between the problem and the concept. This is the abstraction of

[1] ATKIS: Authoritative Topographic-Cartographic Information System (of Germany) [ATK89].
[2] Basis-DLM: base digital landscape model; level of detail about 1:25,000.
[3] DLM250: derived digital landscape model; level of detail about 1:250,000.
[4] DLM1000: derived digital landscape model; level of detail about 1:1,000,000.

the problem (see figure 1 (b)). The abstracted model generalization allows other generalization applications to adapt to the conception of automated model generalization. There are two possible ways for these applications to get adapted: either they have to be abstracted as well and then this abstraction has to be adjusted to the abstracted model generalization, or the abstracted model generalization has to be mapped onto the other problem. These sketched ways are beyond the scope of this paper.

However, the abstracted model generalization will be represented without using expert knowledge of model generalization itself. This is achieved by using mathematical terms and conceptions. Still examples from model generalization are used to explain the developed concepts.

A related approach that formalizes the process of generalization was presented by Ai and van Oosterom [AvO01]. We compared this approach to ours and found the following differences:

1. Ai and van Oosterom provide a mapping of relationships. We consider this as problematic in respect to the mathematical formalization. In our approach the relationships are kept for different data sets.

2. Ai and van Oosterom differentiate between the cardinalities of the mapping (1:1, n:1, n:m). Again we see here a problem for formalization. We propose a substitution of n:m mappings by clustering combined with a 1:1 mapping. This leads to a true mathematical function.

3. Ai and van Oosterom provide an analysis of topological, distance and orientation relationships, while we intend to produce integrity constraints comprehending spatial and non-spatial relationships.

2 The Generalization Function

On our way to the abstraction of model generalization we state the following: For the purpose of automated generalization the ungeneralized source data and the generalized target data are linked together. This results from the observation that every feature of the target is derived from one or more features of the source set. The linking has to be done in such a way that the relationships of objects of both data sets can be identified. On the database level this leads to multiple representation databases [Kil00].

Figure 2 depicts a simple example of such a representation. On the left the source data is shown and on the right the generalized target data is indicated. The relationship between both data sets is depicted by double-headed arrows (not every relationship is shown in this example). Observe that this is not a 1:1 relationship. In an adequate representation one is able to relate the generalized feature to a given source feature and the other way round. Such a representation is very useful

a) for analysing topographic situations during the generalization process itself (this might be necessary while constructing the target set) and

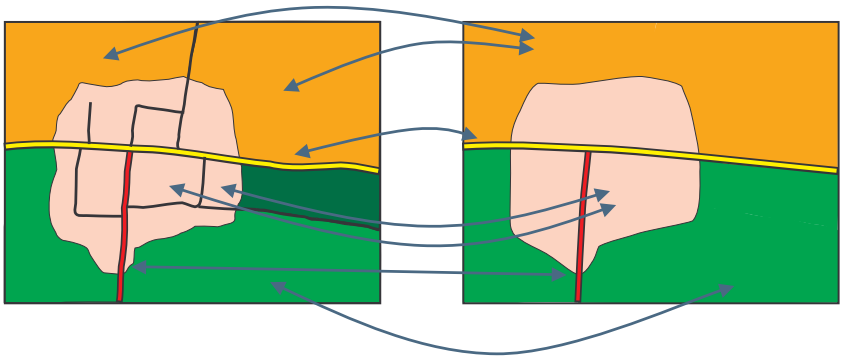

Figure 2: *A multi-scale representation of ungeneralized and generalized data, linked together.*

b) for automated updating of the generalized data. When features of the source set are changed, the necessary changes in the generalized data set can be derived more easily.

The illustration of the relationships by double-headed arrows leads canonically to a *relation* which is a mathematical representation of the linkage between the two data sets: Let the two sets of features be interpreted as mathematical sets. Then the generalization forms a relation between these two sets. In the following we denote the source set of ungeneralized features as S and the target set of generalized features as T. Then the generalization forms a relation $R \subset S \times T$ which means that the generalization is a subset of all possible paired combinations of elements of S and T. Elements of the relation are ordered pairs of elements of each set.

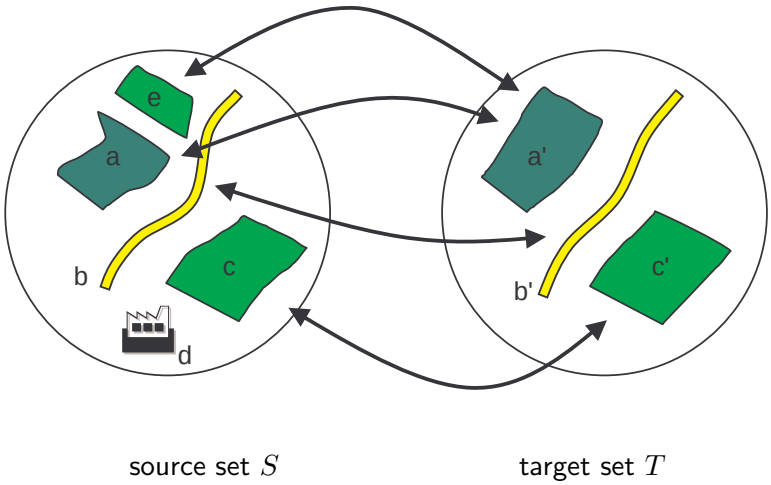

Figure 3: *Example of a generalization relation R between S and T: $R \subset S \times T$.*

Figure 3 depicts an example: The sets S and T contain five and three features,

respectively. Road b and grassland c of set S are generalized to road b' and grassland c' of set T. Forest a and grassland e are amalgamated to forest a'. Factory d is omitted in the target set. So in the example the relation R consists of four pairs: $R = \{(a, a'), (b, b'), (c, c'), (e, a')\}$.

We now analyse the cardinality of this relation. Usually an element of the source set will be generalized to one or zero elements of the target set. In the latter case the object will be omitted. In rare cases an object might be splitted by the generalization process, or the assignment might be not unique, and thus the feature will be related to two or more elements of the target set. Such a situation is referred to as a m:n relationship [AvO01]. In model generalization such cases should not exist. If such a case appears nevertheless, there will be possibilities to transform the ambiguous relation into an unambiguous one. Clustering of features for example may lead back to a 1:1 relationship.

So we assume that an element of S is generalized to not more than one element of T. Hence the generalization relation will not be a one-to-many relation, and thus the relation forms a *function* (or mapping) out of S into T. The example of figure 3 is transformed into the new context (of a function) in figure 4. The *generalization function* $f : S \to T$ maps elements of the source set into the target set. In this example we have $f(a) = a'$, $f(b) = b'$, $f(c) = c'$ and $f(e) = a'$. The mapping of the factory $f(d)$ is not defined.

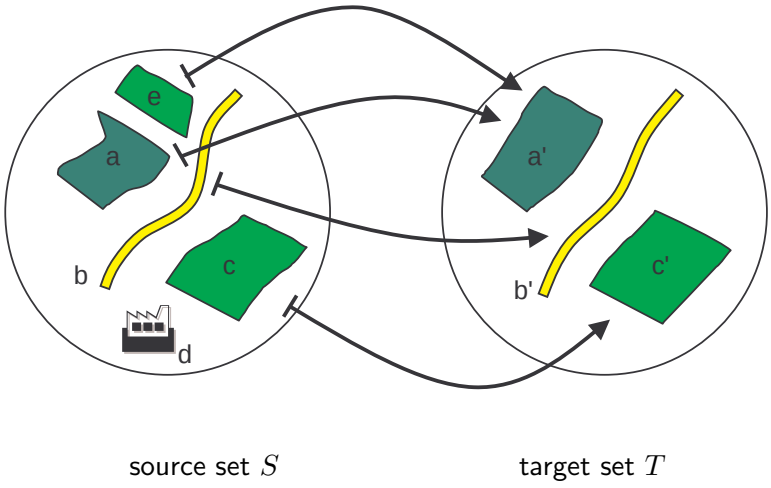

Figure 4: *Example of a generalization function f out of S into T: $f : S \to T$.*

Observe that

a) the function is *not injective*; this means, several elements of S may be mapped onto one and the same element of T (in the example forest a and grassland e are amalgamated into forest a'),

b) the function is *surjective*; that is: each element of T has at least one originating object in S. This represents that the generalized data set cannot contain additional information to that in the source set.

c) the function is *partial*, i.e. *not total*; elements of the source set may exist that are not mapped onto any object of the target set. In the example this holds for factory d which is omitted in the generalized data set.

3 Invariant Properties

Apart from the specified characteristics of the generalization function, both sets of geo-data can be characterized, too. The source set holds certain properties such as

a) non-spatial properties (referring to attributive and class information, also called *semantics* [ATK89])

b) topological properties,

c) geometrical properties,

d) combined properties (combinations of the preceding properties).

The geometrical properties together with the topological constraints form the spatial properties. All four categories of properties must be preserved by the generalization function (and thus during the generalization process) so that the target set obtains the same properties as the source set. With these properties held invariant, the generalization function forms a *morphism* between the two sets of geo-spatial features. A morphism is an important mathematical concept that describes a structure-commuting function. Figure 5 depicts the principle of a morphism in our context: Both the source set and the target set hold the same properties, thus the generalization function is invariant in respect of these properties. In the following several invariant properties will be elaborated. They will be formalized and illustrated by examples of simplified data.

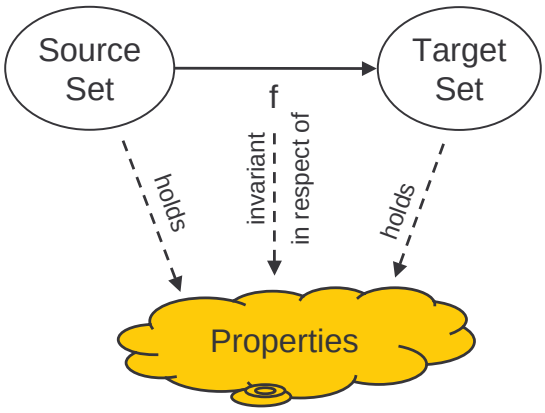

Figure 5: *Principle of a morphism: The generalization function is invariant in respect of certain properties.*

3.1 Invariant Non-Spatial Properties

Just ask yourself the following question which is visualized in figure 6: Why will a road be generalized to a road and not to a railway track? The answer is given straight forward: It's because a road is not a railway track. Both sets of geodata, the source set and the target set, are composed of classes [Nye91]. Roads and railway tracks are examples for two of these classes. The classes are disjoint subsets of the sets S and T, so $S = C_1 \cup C_2 \cup \ldots \cup C_n$ with $C_1 \cap C_2 \cap \ldots \cap C_n = \emptyset$ and $T = C'_1 \cup C'_2 \cup \ldots \cup C'_m$ with $C'_1 \cap C'_2 \cap \ldots \cap C'_n = \emptyset$. Each element of S and T belongs to exactly one of these classes. These are the invariant properties of S and T.

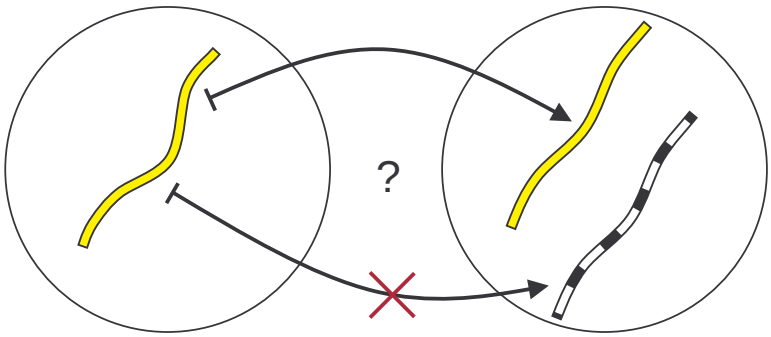

Figure 6: *Invariant non-spatial properties: Why will a road be generalized to a road and not to a railway track?*

The generalization function has to ensure that an element of a class of the source set is reasonably mapped into an appropriate class of the target set. This constraint must be preserved by the generalization function. For this purpose the classes of both sets have to be related, too. An example of a relation CR between classes is shown in table 1. Written mathematically, the relation CR consists of: $CR = \{$(state road, road), (county road, road), (railway track, railway track), (woodland, woodland), $\}$. Observe that, as the example of roads shows, both a state road and a county road might be generalized to a road without further distinction. Thus the relation between the classes need not be a one-to-one relation.

source set class	**target set class**
state road	road
county road	road
railway track	railway track
woodland	woodland

Table 1: *Relation CR between classes of the source set S and the target set T.*

We now are able to formalize the above given constraint as follows: For each element $x \in C_i \subset S$ and $x' \in C'_j \subset T$ the generalization function f must hold:

$$f(x) = x' \qquad \Longrightarrow \qquad (C_i, C'_j) \in R.$$

If this constraint is violated, the generalization function will not be correct. Further constraints can be formulated for the attributive information of the classes.

3.2 Invariant Topological Properties

There are several topological properties that have to be held invariant by the generalization function. In a topological data model [HG94, PG97], as we propose to use for model generalization [Bob00], several topological errors such as overshoot, undershoot or sliver polygons will be solved on the database level. Therefore we will concentrate on topological relationships here.

Topological relationships can be defined either by the means of graph theory [Eve79, Har72] or by the 4- or 9-intersection model by Egenhofer et al. [EF91, EH92]. The latter has the advantage of a finer distinction between topological relationships while the former simplifies some analysis of topological relationships—if the topology is already modelled on the database level.

The first example of a topological property is the *neighborhood relationship*. For our purpose we define neighborhood by adjacency and/or incidence of topological elements (nodes, edges, faces). We subsume the individual relationships under the common neighborhood relation N that contains all pairs of neighboring elements within one geo-data set. N is *symmetric*, so if (x, y) is element of N, then (y, x) is an element of N, too. We define N as *reflexive*, which means that each geo-object is adjacent to itself, so for each x, (x, x) is in N. The reason for this predefinition is given further in this example.

The generalization function has to ensure that the mapping of two neighboring elements of S are neighbors in T, too. Formalized, this constraint reads as follows: For each pair of elements $x, y \in S$ with $f(x) = x' \in T$ and $f(y) = y' \in T$ must hold:

$$(x, y) \in N \qquad \Longrightarrow \qquad (x', y') \in N.$$

Figure 7 shows an example of this constraint. All neighborhoods of set S are preserved by the generalization function. Observe that for the pair of features $(a, d) \in N$ the mappings are identical (feature $a' \in T$). This is an example for the need of N being reflexive, as $(f(a), f(e)) = (a', a')$ has to be element of N.

A second example of a topological relationship is the *inclusion relation* I. A pair of objects x, y are topologically related in respect of I when x lies inside of y. In contrast to the neighborhood relation, I is asymmetric (if x is inside of y, then y is not inside of x). Nevertheless the formalized constraint is similar: For each pair of elements $x, y \in S$ with $f(x) = x' \in T$ and $f(y) = y' \in T$ must hold:

$$(x, y) \in I \qquad \Longrightarrow \qquad (x', y') \in I.$$

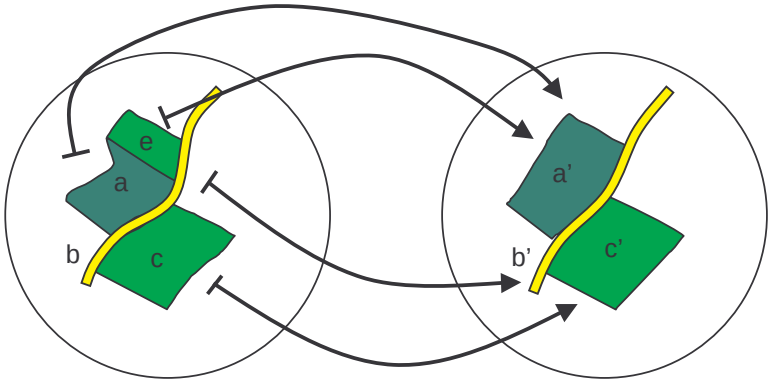

Figure 7: *Invariant topological properties: Neighborhood relationship*

3.3 Invariant Geometrical Properties

The generalization of geometry seems to be the largest part of generalization research. This can be seen by the numerous approaches that deal with this subject. However, the geometrical generalization plays an inferior role in model generalization, since graphical presentation of geo-data is not the major purpose of digital landscape models [BW88, Sch99]. Hard problems of map generalization, such as displacement, are therefore not an issue in model generalization.

What matters in digital landscape models is the position accuracy of geo-objects. The desired accuracy is given by the resolution level of the digital landscape model. Simplified geometry of generalized data has to be within a certain *tolerance corridor*. Such a tolerance corridor is used in several line simplification algorithms such as the algorithms developed by Douglas/Peucker or Lang [MS92]. These algorithms guarantee that the position of a generalized line does not deviate from the original line by more than a certain amount. Of course the mentioned algorithms fulfil more than this. Among other things they try to preserve the characteristics of the line. Topological errors may occur as a result of line simplification. See the next section for further details to this issue.

In the context of abstracting the model generalization, here the task is to use already developed algorithms such as the above mentioned to formulate the geometrical constraints that have to hold during generalization. The tolerance corridor will be one of the main aspects of the abstraction. This is subject of ongoing research and therefore will be deferred to a later publication.

3.4 Invariant Combined Properties

Combined properties comprise properties that cannot be classified into one of the preceding properties alone. We present three examples of combined properties of which two are combinations of topological and non-spatial properties and one is a combination of topological and geometrical properties. We just

formulate the constraints but do not formalize them completely here.

The first example is the property of *area coverage*. In a digital landscape model it is desired that the whole surface within the acquired area is covered by area features. Thus, given a topological tessellation of the surface by a set of faces, the property reads: for each face of the set of faces there must exist at least one area feature that is connected to this face. The generalization function must not destroy the area coverage, thus the above property must hold in the generalized target set, too.

The second example is the property of *readability* or *net connectivity*. Roads for example form a network, and each road of this network is topologically connected to every other road of that network by a series of roads. Other networks are formed by railway tracks, rivers or power lines. The readability can be expressed by a relation E that contains all pairs of mutually reachable features of one network within one data set. This relation is transitive which means that if $(x, y) \in E$ and $(y, z) \in E$, then $(x, z) \in E$, too. The connectivity must not be destroyed by the generalization function. So the mappings of two mutually reachable features must be reachable in the generalized data set, too.

The last example is a combination of topological and geometrical properties. As mentioned in the last section, the line simplification may lead to topological inconsistencies such as self intersection, intersection with other lines or sidedness alteration [Saa99]. Several approaches exist that deal with the preservation of topological consistency during line simplification [Saa99, dBvKS98]. In our context the task is to formulate constraints that guarantee the topological consistency *and* the fulfillment of the requirements of line simplification. An informal formulation might run like the following: for each line of the target set it is necessary that it is within the tolerance corridor of the related line of the source set. It also has to have no self intersection and no intersection with other lines of the target set. Furthermore each mapped point feature that lies in the incident faces of the source line has to lie in the mapping of this face (i.e. it must not change its sidedness).

4 Conclusion

Motivated by the wish to use an already evolved conception for model generalization in other generalization applications, model generalization was abstracted in this paper. The generalization process was interpreted as a function out of an ungeneralized source set into a generalized target set. This generalization function has to preserve certain properties of the source set while constructing the target set. These properties are then called invariant. The function thus forms a morphism between the two sets of geo-data with respect to the invariant properties.

The invariant properties were denominated and formalized. This resulted in the definitions of integrity constraints that allow a consistency check of a set of generalized data as well as support for the construction of the generalization function or for the implementation of a generalization concept. The integrity constraints were formulated as logical expressions that can be verified easily

and automatically.

Further work has to be done on those aspects that have not been formalized yet. This includes the set of geometrical properties as well as the combined properties. The next step after that will be to find a generalization application that can benefit from the developed abstraction. Possible candidates are the generalization of geological maps and the generalization of road maps.

References

[ATK89] AdV-Arbeitsgruppe ATKIS. *ATKIS-Gesamtdokumentation.* Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), Hannover, 1989.

[AvO01] Tinghua Ai and Peter van Oosterom. A map generalization model based on algebra mapping transformation. In Walid G. Aref, editor, *Proceedings of the Ninth ACM International Symposium on Advances in Geographic Information Systems*, pages 21–27, November 9–10 2001.

[Bob00] Matthias Bobzien. Implementationsaspekte der Modellgeneralisierung. In *Mitteilungen des Bundesamtes für Kartographie und Geodäsie*, number 17, pages 15–25, Frankfurt am Main, 2000.

[Bob01] Matthias Bobzien. Flächenzusammenfassung in der Modellgeneralisierung. In *Mitteilungen des Bundesamtes für Kartographie und Geodäsie*, number 20, pages 19–29, Frankfurt am Main, 2001.

[Bob02] Matthias Bobzien. Geometry-type change in model generalization – a geometrical or a topological problem? Unpublished working paper for the Joint ISPRS/ICA Workshop on Multi-Scale Representations of Spatial Data, Ottawa, July 7–8 2002. Available via Internet: `http://www.ikg.uni-hannover.de/isprs/workshop/abstract-bobzien.pdf`, 2002.

[BW88] Kurt E. Brassel and Robert Weibel. A review and conceptual framework of automated map generalization. *International Journal of Geographical Information Systems*, 2(3):229–244, 1988.

[dBvKS98] Mark de Berg, Marc van Krefeld, and Stefan Schirra. Topologically correct subdivision simplification using the bandwidth criterion. *Cartography and Geographic Information Science*, 25(4):243–257, 1998.

[EF91] Max J. Egenhofer and Robert D. Franzosa. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991.

[EH92] Max J. Egenhofer and John R. Herring. Categorizing binary topological relationships between regions, lines, and points in geographic databases. Technical report, Department of Surveying Engineering, University of Maine, Orono, 1992.

[Eve79] Shimon Even. *Graph Algorithms*. Pitman, London, 1979.

[Har72] Frank Harary. *Graph Theory*. Addison-Wesley, Reading, Massachussetts, 1972.

[HG94] Günter Hake and Dietmar Grünreich. *Kartographie*. De Gruyter, Berlin, New York, 1994.

[Kil00] Tiina Kilpeläinen. Maintenance of multiple representation databases for topographic data. *The Cartographic Journal*, 37(2):101–107, December 2000.

[MS92] Robert B. McMaster and K. Stuart Shea. *Generalization in Digital Cartography*. Association of American Geographers, Washington D.C., 1992.

[MS99] Dieter Morgenstern and Dietrich Schürer. A concept for model generalization of digital landscape models from finer to coarser resolution. In *ICA-Proceedings 1999*, Ottawa, 1999.

[Mül91] J.-C. Müller. Generalization of spatial databases. In David J. Maguire, Michael F. Goodchild, and David W. Rhind, editors, *Geographical Information Systems*, pages 457–475. Longman Scientific & Technical, 1991.

[Nye91] Timothy L. Nyerges. Representing geographical meaning. In Barbara P. Buttenfield and Robert B. McMaster, editors, *Map Generalization*, pages 59–85. Longman Scientific & Technical, 1991.

[PG97] Lutz Plümer and Gerhard Gröger. Nested maps — a data model for spatial aggregates. *GeoInformatica*, I(3), 1997.

[Saa99] Alan Saalfeld. Topologically consistent line simplification with the Douglas-Peucker algorithm. *Cartography and Geographic Information Science*, 26(1):7–18, 1999.

[Sch99] Nadine Schuurman. Critical GIS: Theorizing an emerging science. *Cartographica*, 36(4), 1999. Chapter 4.

[Sch02] Dietrich Schürer. *Ableitung von digitalen Landschaftsmodellen mit geringerem Strukturierungsgrad durch Modellgeneralisierung*. PhD thesis, Institut für Kartographie und Geoinformation, Universität Bonn, 2002.