Sociedad de Estadística e Investigación Operativa
Top (2005) Vol. 13, No. 1, pp. 1–74

Towards a Unified Territorial Design Approach – Applications, Algorithms and GIS Integration

Jörg Kalcsics
Universität des Saarlandes, Germany
E-mail: j.kalcsics@orl.uni-saarland.de
Stefan Nickel
Universität des Saarlandes and Fraunhofer Institut für Techno- und Wirtschaftsmathematik, Germany
E-mail: s.nickel@orl.uni-saarland.de
Michael Schröder
Fraunhofer Institut für Techno- und Wirtschaftsmathematik, Germany
E-mail: schroeder@itwm.fhg.de

Abstract

Territory design may be viewed as the problem of grouping small geographic areas into larger geographic clusters called territories in such a way that the latter are acceptable according to relevant planning criteria. In this paper we review the existing literature for applications of territory design problems and solution approaches for solving these types of problems. After identifying features common to all applications we introduce a basic territory design model and present in detail two approaches for solving this model: a classical location–allocation approach combined with optimal split resolution techniques and a newly developed computational geometry based method. We present computational results indicating the efficiency and suitability of the latter method for solving large–scale practical problems in an interactive environment. Furthermore, we discuss extensions to the basic model and its integration into Geographic Information Systems.

Key Words: Territory design, political districting, sales territory alignment, optimization algorithms, Geographical Information Systems, computational geometry.

AMS subject classification: 90C59, 90B80, 68U05, 90-02.

1 Introduction

Territory design may be viewed as the problem of grouping small geographic areas called *basic areas* (e.g. counties, zip code or company trading areas) into larger geographic clusters called *territories* in such a way that the latter are acceptable according to relevant planning criteria. Depending on the context, these criteria can either be economically motivated (e.g. average sales potentials, workload or number of customers) or have a

demographic background (e.g. number of inhabitants, voting population). Moreover spatial restrictions (e.g. contiguity, compactness) are often demanded. We note, that in the literature often the term alignment instead of design is used. As both expressions are interchangeable we will use the latter one throughout this paper.

Territory design problems are motivated by quite different applications ranging from political districting over the design of territories for schools, social facilities, waste collection or emergency services to sales and service territory design. However, the two main applications are political districting and sales and service territory design. In the former application, a governmental area, such as a city or state, has to be partitioned into a given number of territories. As each territory elects a single member to a parliamentary assembly, the main planning criteria is to have approximately the same number of voters in each territory, i.e. territories of similar size, in order to respect the principle of "one man–one vote". The task of designing sales territories is common to all companies which operate a sales force and need to subdivide the market area into regions of responsibility. Closely related to this is the problem of designing service territories for attending to customers or technical facilities. Typical planning requirements are to design territories which are similar in size, e.g. in terms of sales potentials or workload, or which reduce travel times within the territories needed to attend to customers or service incidents.

Due to legal regulations, shifting markets or the introduction of new products, territory design decisions have to be frequently re–evaluated. Especially for a large number of basic areas and territories this is a lengthy task and therefore an algorithmic optimization approach for expediting the process is often desired. For sales territories, well–planned decisions enable an efficient market penetration and lead to decreased costs and improved customer service while in terms of political districting, an algorithmic approach protects against politically motivated manipulations during the territory design process.

When reviewing the literature, one can observe that only few papers consider territory design problems independently from a concrete practical background. Hence the tendency in operations research to separate the model from the application and establish the model itself as a self–contained topic of research can not be observed. However, when taking a closer look at the proposed models for different applications, a lot of similarities can be

observed. Indeed, the developed models can often be, more or less directly, carried over to other applications.

Therefore, we will introduce a basic model for territory design and present in detail two approaches for solving the problem: a classical location–allocation approach combined with optimal split resolution techniques and a newly developed computational geometry based method. In the former one, which was already introduced in the mid sixties and has been extensively used since then, the territory design problem is modeled as a discrete capacitated facility location problem and solved by applying a location–allocation method. As territories of similar size are usually obtained in this approach by first solving a continuous, capacitated transportation problem and then rounding the fractional assignments in some non–optimal way in order to obtain non–overlapping territories, we present optimal techniques for resolving the non–integral assignments. Optimal in the sense that the maximal or average deviation of the territory size from the mean is minimal.

Since this method is not suitable for the use in an interactive environment, as our computational tests showed, we developed a new method, which is based on computational geometry and utilizes the underlying geographical information of the problem. The idea of this method is to recursively partition the region under consideration geometrically into smaller and smaller subproblems taking the planning criteria into account until an elemental level has been reached where the territory design problem can efficiently be solved for each of the elemental subproblems. Although the idea for this method was already briefly sketched in the literature, no details were given. We will present computational results indicating the efficiency and suitability of this method for solving large–scale practical problems in an interactive environment.

The rest of the paper is organized as follows. In the next section, we will present various applications for territory design problems found in the literature and identify basic features, common to all applications. Based on these observations we introduce in Section 3 a basic model, which covers most of the previously identified features. In the following section we review the existing literature on models and solution techniques for solving territory design problems and in Sections 5 and 6 present in detail two methods for solving such problems: a classical location–allocation approach combined with optimal split resolution techniques and a newly developed computational geometry based method. In the next section, computational

results for the two approaches are presented and in Section 8 we discuss extensions to the basic model and the integration of the presented methods into a Geographic Information System. The paper concludes with a summary and an outlook to future research.

2 Applications of territory design problems

In what follows we will present several applications which all have in common the task of subdividing the region under inspection into a number of territories, subject to some side constraints.

2.1 Political districting

The problem of determining political territories can be viewed as one of dividing a governmental area, such as a city or a state, into subareas from which political candidates are elected. This problem, usually referred to as *political districting*, is particularly important in democracies where each territory elects a single member to a parliamentary assembly. This is for example the case in Canada, New Zealand, most states in the U.S. and Germany. We note that in this context, territories are usually called *districts*, *census tracks* or *constituencies*.

In general, the process of redistricting has to be periodically undertaken in order to account for population shifts. The length of these periods varies from country to country, e.g. in New Zealand every 5 years, in Canada and the U.S. every decade. To aid this process, operations researchers have developed since the early sixties many different automatic and neutral districting procedures. For recent books on political districting, the reader is referred to Grilli di Cortona et al. (1999).

In the past, political districting has often been flawed by manipulation aiming to favor some particular party or to discriminate against social or ethnic minorities. Since the responsibility for approving state and local districting plans usually falls to elected representatives, plans are likely to be shaped implicitly, if not overly, by political considerations, e.g. to keep them in power. A famous case arose in Massachusetts in the early nineteenth century when the state legislature proposed a salamander–shaped electoral district in order to gain electoral advantage. The governor of the

state at that time was Elbridge Gerry and this practice became known as gerrymandering. See Lewyn (1993) for an interesting description of gerrymandering cases.

To prevent political interference in the districting process, many states have set up a neutral commission, whose functions include the drawing up of political boundaries satisfying a number of legislative and common sense criteria. Depending on the country or jurisdiction involved, these criteria may be enforced by legislative directive, judicial mandate or historical precedent. However, in the scientific literature related to political science, law or geography there is no consensus on which criteria are legitimate for the districting process, i.e. satisfy the neutrality condition. Some are predetermined by constitutional laws, while others are based on common sense and can be disputed. Moreover, it is unclear how they should be measured (Williams (1995)). In the following, we will present some of the most commonly used, see e.g. Williams (1995), George et al. (1997) and Bozkaya et al. (2003). We also refer to these references for additional criteria in districting which are not mentioned here. They can be placed under the three general headings: demographic, geographic and political.

Demographic criteria

Population and voter equality When designing electoral districts the main criteria is equity in order to respect the principle of "one man–one vote", i.e. every vote has the same power, and achieve an equitable presence of elected officials. However, depending on the country, deviations from the equal population target are permitted in order to take other criteria into account. Allowed relative deviations from the average range from 5% in New Zealand to 25% in Germany up to, in exceptional cases, 50% in Canada. In the U.S. however, population equality has been deemed by the courts to be very important and as a result the actual deviation now–a–days is in most cases less than 1%.

Minority representation The intention of this criterion is to ensure that minority voters have the same opportunity as other members of the electorate to participate in the political process and to elect representatives of their choice (Parker (1990)). Especially in the U.S., this criteria has become an important consideration during the last 30

years. Note, that this criteria may also be seen as political one.

Geographic criteria

Compactness A district is said to be geographically compact if it is somewhat round–shaped and undistorted. Although being a very intuitive concept, a rigorous definition of compactness does not exist. Niemi et al. (1990) and Horn et al. (1993) propose several measures to asses the compactness of a district, none of which is comprehensive.

Compact districts are desired since this reduces the possibility of gerrymandering. In fact, in the U.S., compactness has been defined as simply the absence of gerrymandering. Some authors however argue that the importance of compactness is less than that of other criteria, since gerrymandering is usually not a problem when an algorithm does not use political data (Garfinkel and Nemhauser (1970)).

Contiguity In general, territories have to be geographically connected. First, to protect once more against gerrymandering and second simply because of administrative reasons.

Boundaries and community integrity In many cases, districts should be designed such that they adhere to the boundaries of other political constituencies, like cities or counties, or match as closely as possible the boundaries of the previously existing electoral districts. Moreover topological obstacles like mountain–ranges or large bodies of water should be taken into account (George et al. (1997)).

Political criteria

Political data Although much disputed some authors decided to consider political data for the redistricting process. For example Bozkaya et al. (2003) employ a criteria to achieve socio–economic homogeneity across the districts to ensure a better representation of residents who share common concerns or views. Like Garfinkel and Nemhauser (1970), they argue that through the use of computer based methods, the possible subjective influence is minimal.

Table 1 provides a selection of articles for political districting problems solved with methods from operations research. It is indicated which of

the above mentioned criteria are considered and in addition the country to which the article refers to. Since all authors, without exception, take population equity into account, this criterion is omitted.

Reference	Country	Contiguity	Boundaries	Compact-ness	Political data
Hess et al. (1965)	USA	+	-	+	-
Garfinkel & Nem. (1970)	USA	+	-	+	-
Helbig et al. (1972)	USA	+	+	+	-
Bodin (1973)	USA	+	-	-	-
Bourjolly et al. (1981)	Canada	+	-	+	+
Nygreen (1988)	Wales	+	+	+	-
Hojati (1988)	Canada	-	-	+	-
George et al. (1997)	New Zealand	+	+	+	-
Ricca & Simeone (1997)	Italy	+	+	+	-
Mehrotra et al. (1998)	USA	+	+	+	-
Cirincione et al. (2000)	USA	+	+	+	+
Bozkaya et al. (2003)	Canada	+	+	+	+
Forman and Yue (2003)	USA	+	+	+	-

Table 1: *Selected operations research studies for political districting.*

2.2 Sales and service territory design

The important but expensive task of designing sales territories is common to all companies which operate a sales force and need to subdivide the market area into regions of responsibility. Closely related is the problem of designing service territories for attending to customers or technical facilities. Here, often quite similar criteria are employed for the design of territories for service staff.

Fleischmann and Paraschis (1988) report on a German manufacturer of consumer goods who delivers products to several thousand wholesalers. Sales promotions and advertising amongst the retailers is very important in the considered business and is carried out by sales agents, where each agent is in charge of a certain territory. The study was motivated by the impression of the company, that the 8 year old territories seemed to be inappropriate for today's business, mainly because of the uneven distribu-

tion of workload. (Hereby, the workload of a customer was expressed as an internal score taking into account sales value and frequency of visits.)

Blais et al. (2003) report on a home–care districting problem in the province of Quebec, Canada. In this case, local community health clinics are responsible for the logistics of home–care visits by health–care personnel, like nurses or physiotherapists, in a given area. If the area is too large, it has to be partitioned into territories, each looked after by a different multi–disciplinary team. Due to a change in the Quebec health care policies, the overall workload increased and became uneven between different territories. To alleviate this problem, the home–care services managers decided to increase their number and re–align them.

In general, there are several motivations for aligning existing or designing new territories. First an increase or decrease in the number of sales– or service–men obviously requires some adjustment of the territories. Other reasons are to achieve better coverage with the existing personnel or to evenly balance workload among them. Moreover customer shifts or the introduction of new products make it necessary to align territories.

In the following we present several commonly used criteria for sales territory design problems, see e.g. Zoltners and Sinha (1983).

Organizational criteria

Number of territories Often, the number of districts to be designed is predetermined by the sales force size designated by the company or planner, see e.g. Fleischmann and Paraschis (1988). In case the size is not self–evident, several methods are proposed to compute suitable numbers. For an overview see Howick and Pidd (1990).

In the past, several authors pointed out, that there exists an interdependency between the sales force size and the territory design and proposed models which assume a variable number of territories, see e.g. Drexl and Haase (1999).

Basic areas Sales territories are in most cases not designed based on single customers. In fact, customers are usually first aggregated into small areas which in turn then serve as a basis for the territory design process. Typical examples for basic areas are counties, zip code areas, predefined prospect clusters or company trading areas. As a result,

depending on the level of detail or aggregation, the complexity of the problem reduces considerably and in addition relevant planning data, like sales potentials or distances, is generally much easier to obtain or estimate. Especially the last characteristic has a strong impact on the effort of the planning process.

Exclusive assignment of basic areas In most applications basic areas have to be exclusively assigned to a territory. This requirement is motivated by several factors. Most notably, unique allocations result in transparent responsibilities for the sales representatives avoiding contentions among them and allowing for the establishment of long–term customer relations. The latter aspect often goes along with the desire to minimize arbitrary changes in territory boundaries. Moreover, staff–dependent performance reviews are easier to compile.

Locations of sales representatives As sales persons have to visit their territories regularly, their location, e.g. office or residence, is an important factor to be considered in the territory design process. Here, one has to decide wether locations of representatives are predetermined and should be kept or are subject to the planning process.
With respect to the latter case, Zoltners and Sinha (1983) remark, that center–seeking territory design approaches have the practical shortcoming, that most sales persons have strong preferences for home–base cities. Fleischmann and Paraschis (1988) however report in their case study, that management did not want sales persons residences to influence the definition of territories heavily, because addresses can frequently change.

Geographical criteria

These criteria are mainly motivated by the fact that sales representatives have to travel within their territories.

Contiguity Territories should be geographically connected.

Accessability Often a good accessibility of territories, e.g. to highways, or within territories, for example by means of public transportation, is required. Moreover, sometimes non–traversable obstacles like rivers or mountain ranges have to be accounted for.

Compactness In most applications, compact territories are an important design criterion. As Hess and Samuels (1971) point out, one way to improve a salesman's efficiency is to reduce his unproductive travel time. Compact territories usually have geographically concentrated sales (service) activity, therefore less travel, more selling (service) time and hopefully higher sales (better service levels). In other words, the term compactness expresses the desire for territories with minimal total travel times.

Most models try to achieve compactness by minimizing a weighted distance between basic areas and district centers. Whereas for political districting problems, usually squared Euclidean distances are employed to achieve compactness, Cloonan (1972) and Marlin (1981) point out that travel costs in a territory are more proportional to straight line distances. For even more accurate computations, Segal and Weinberger (1977) use network distances. Although minimizing the weighted distances does not necessarily guarantee (visually) compact territories, it is easily tractable and the results provided are usually satisfactory. Moreover, as the underlying motivation for compact sales and service territories is to minimize travel times, this approach is plausible. For an overview of more elaborate compactness measures see Niemi et al. (1990) and Horn et al. (1993).

In several applications, travel times within territories are in addition considered as part of an activity–related design criterion, e.g. workload. Ronen (1983) describes the case of a sales territory design problem for sparse accounts of a distributor in the Midwestern United States, where travel time is even the major design criterion, as the market area of the distributor covers almost five states.

Activity–related criteria

Geographic requirements on sales territories mainly focus on the travel aspect. However traveling is only a means to an end for the actual work to do, namely selling products or providing service.

Balance For sales and service territory alignment problems, often districts which are balanced relative to one or more attributes (called activity measures) are sought for. This criterion expresses a relation of territories among each other and is motivated by the desire of an even

treatment of all sales persons. For example in order to evenly distribute workload or travel times among the sales persons or service staff or for reasons of fairness in terms of potential prospects or profit.

Although several different sales territory related attributes have been discussed in the literature (see e.g. Hess and Samuels (1971), Zoltners and Sinha (1983)), one can observe, that only few authors consider more than one criterion simultaneously for designing balanced territories (Deckro (1977), Zoltners (1979), Zoltners and Sinha (1983)).

Apart from the desire for balanced territories, sometimes strict upper or lower bounds for the size of districts are given. For example on maximal travel times or minimal number of customers within the district.

Maximizing profit Especially for sales companies, profit is a major aspect in the planning process. Generally a limited resource of call time or effort is available and has to be allocated in a profit–maximizing way amongst a number of sales entities such as customers or prospects. See Howick and Pidd(1990) for a good overview of commonly used time–effort allocation methods.

To this end, several authors (Lodish (1975), Shanker et al. (1975), Glaze and Weinberg (1979), Zoltners and Sinha (1983), Skiera and Albers (1994) and Drexl and Haase (1999)) propose an integration of time–effort allocation and territory design methods in order to produce more profit and sales than can be obtained by merely using sales potential.

However, Ronen (1983) claims, that changing the solution of the strategic territory design problem is much more complicated and expensive than that of the operational time–effort allocation problem and therefore addresses them separately.

Although several methods for integrating time–effort allocation and territory design to maximize profit have been proposed, most procedures still consider the balancing requirement when designing districts. For example to evenly share workload, potential prospects or profit among their sales force. Skiera and Albers (1994) and Drexl and Haase (1999), among others, object that the balancing aspect, i.e. fairness or equity, is not the primary criterion for most companies. The main aim should be to maximize profits, regardless of

any balancing aspect. Skiera (1997) reports of (randomly generated) experiments, where the sales obtained by a pure profit–maximizing approach compared with one taking balance into account, were 5% – 14% higher.

In Table 2 a selection of, in our opinion, important articles for sales and service territory design problems solved with methods from operations research is provided. An extensive overview of models before 1990 can be found in Howick and Pidd (1990).

For each reference we mark wether a selected criterion is considered in the proposed model ("+") or not ("-"). For the organizational criteria "number of territories" and "locations of sales representatives", a "f" or "v" indicates if the number or locations are fixed or variable. "Mult" stands for models taking more than one activity measure for balance into account. Since the organizational criterion "exclusive allocation" and the geographic "compactness" requirement were considered by all authors without exception, these two are omitted in the table.

2.3 Other applications

Besides the most common problems of sales territory design and political districting, several authors report on various other closely related applications.

2.3.1 Territories for facilities providing service at a fixed location

In many cases, customers have to visit a (public) facility in order to obtain service, e.g. schools or hospitals.

School districts

Palermo et al. (1977) and Ferland and Guénette (1990) deal with the problem of assigning residential areas to schools. As an outcome of the

Reference	Applic.	Organizational		Geographic		Activity–related	
		Nb.	Loc.s	Contig.	Access.	Bal.	Profit
Hess & Samuels(1971)	Sales/serv.	f	v	-	-	+	-
Easingwood (1973)	Service	f	f	+	-	+	-
Shanker et al. (1975)	Sales	v	f	+	+	+	+
Segal & Weinberger (1977)	Service	f	f	+	-	+	-
Glaze & Weinberg (1979)	Sales	f	v	-	+	-	+
Zoltners(1979)	Sales	f	v	+	+	+/mult	+
Marlin (1981)	Service	f	f	-	-	+	-
Ronen (1983)	Service	f	f	-	-	+	-
Zoltners & Sinha (1983)	Sales	f	f	+	+	mult	-
Fleischmann & Paraschis (1983)	Sales	f	v	-	v	+	-
Skiera & Albers (1994)	Sales	f	f	+	+	-	+
Drexl & Haase (1999)	Sales	v	v	+	+	-	+
Blais et al. (2003)	Service	f	-	+	+	-	+

Table 2: *Selected operations research studies for sales and service territory design problems.*

planning process, all residential areas in the region under consideration are partitioned into a number territories, one for each school. Criteria generally taken into account are capacity limitations on and equal utilization of the schools, maximal or average travel distances for students, good accessability and racial balance (the latter especially in the U.S.).

Territories for social facilities

When planning territories for social facilities, like hospitals or public utilities, administrative units have to be aggregated into territories. As a result, it is determined for every inhabitant to which facility he should go in order to obtain service. Typically the number of inhabitants of each territory has to be within predetermined bounds in order to account for a good utilization and a limited capacity of the social facility. Moreover territories should be contiguous and the facilities should be easily accessible for all inhabitants of the respective territory, for example by public transportation. See e.g. Andria et al. (1979) and Minciardi et al. (1981).

2.3.2 On–site service territories

Several (public) institutions provide their service not at a fixed location but distributed over a geographic region or on–site where the service incident occurs.

Winter services and solid waste collection

Muyldermans et al. (2002) deal with the planning of winter gritting and salt spreading services. On a superior planning level, the region under consideration has to be partitioned into territories, where each territory contains at least one vehicle depot. Afterwards, vehicle routes for providing service are planned for each territory separately. The main design criteria for the territories are balance, in terms of travel distance, compactness and contiguity. Moreover, territories should allow the planning of "good" routes.

Closely related is the problem of solid waste disposal. In a first step, so–called sectors are determined, where each sectors consists of a set of streets or street segments in which waste has to be collected on a certain day. Afterwards, routes for the garbage trucks within the sectors are computed. According to Hanafi et al. (1999), the overall time for collecting garbage should be minimized (compactness), the time for collecting garbage should be approximately the same for all sectors (balance) and the sectors should be contiguous.

Whereas in the case study of Muyldermans et al. (2002), territories are required to be non–overlapping, Hanafi et al. (1999) reports that, depending on how often per week waste has to be collected, certain streets can belong to more than one sector, i.e. basic areas are not mutually exclusively assigned to sectors.

Emergency service territories

D'Amico et al. (2002) report on a case study for police district design, where police departments have to partition their jurisdiction into so–called command districts. After the districts have been fixed, an optimal number of patrol cars that should be on duty is assigned to each command district and the "goodness" of the districts in terms of several different

performance measures is assessed. A closely related problem is described by Baker et al. (1989), which face the task of designing so–called primary response areas for county ambulances.

As reported, the main design criteria for the territories are workload balance, geographical compactness and contiguity. However, what distinguishes these problems from the previously mentioned is an additional consideration of response times to calls for service, which should be minimized and/or have to be below certain threshold values. These considerations require the incorporation of queueing models and measures: officer or ambulance workloads constitute utilization of servers and response times constitute customer waiting times.

2.3.3 Electrical power districting

According to Bergey et al. (2003), the World Bank regularly faces the challenge of helping developing countries to move from state owned, monopolistic electric utilities to a more competitive environment with multiple electricity service providers. At that, they face the task of partitioning the physical power grid into economically viable territories (distribution companies). The main aim is to determine territories with approximately equal earning potential in order to provide an environment that will foster competition, and that are compact over a geographic region and therefore easier to manage and more economical to maintain. Moreover, the territories should be non–overlapping and contiguous.

2.3.4 Clustering and aggregation of spatial data

Clustering aims at the aggregation of data into classes. On the one hand, each class should comprise data with characteristics as similar as possible and on the other hand, the dissimilarities of data between different classes should be as large as possible. Although the basic task of aggregating smaller units into larger sets is the same for clustering and territory design, the motivations are quite different. The former strives for inner homogeneity of data while the latter aims at outer similarity. Therefore models for solving both problems are in general not compatible.

However for some aggregation problems, territory design models are

applicable. For example Simchi–Levi et al. (2003) formulate the following guidelines, among others, when aggregating demand points for location problems with the aim of reducing the complexity of the problem: aggregate demand points for 150 to 200 zones; make sure each zone has approximately an equal amount of total demand; place aggregated points at the center of the zone. These guidelines read as a classical center–seeking territory design problem.

3 Basic modeling of the territory design problem

Since the early sixties, many authors have investigated territory design problems and provided models for various applications. In the following we will focus on aspects that are shared by most of these models. They cover the essential aspects of territory design problems and can be applied to most of the applications that have been described in the previous section.

Focusing on basic modeling aspects might be considered as a disadvantage, since a user may find that some of his requirements are not reflected in such a model. However, there exist several reasons why general purpose models for territory design are worth studying:

1. Often such a model provides a sufficient approximation for the practical application. For example George et al. (1997) and Fleischmann and Paraschis (1988) report that the solutions obtained by their models were implemented in practice. Both models are rather similar and address the design of electoral districts and sales territories, respectively.

2. The models provide "good" solutions, which can in turn serve as a starting point for manual improvements or local search heuristics, which are able to take more complex criteria into account.

3. There exists a broad range of practical problems to which the models can be applied.

4. General purpose models can serve as a starting point for more complex models that take additional planning criteria into account, depending on the real–world situation.

Our objective is to provide algorithms that run in a general purpose geographical information system. Therefore we do not know the exact problem that a potential user has. Modeling only the most common and basic aspects of the territory design problem allows a wide applicability of the provided algorithms.

In the following we present 'building blocks' for basic models in territory design. Also some notation is introduced that is summarized at the end of this section.

Basic areas A territory design problem encompasses a set V of *basic areas*, sometimes also called sales coverage units. These *BA*s are geographical objects in the plane: points (e.g. geo–coded addresses), lines (e.g. street–sections) or geographical areas (e.g. zip–code areas, counties, predefined company trading areas). In the latter case the geographical areas are generally given as polygons. See Figure 1 for an example of basic areas defined by zip–code regions. In case of non-point objects, a basic area $v \in V$ is represented by its center with coordinates (x_v, y_v).

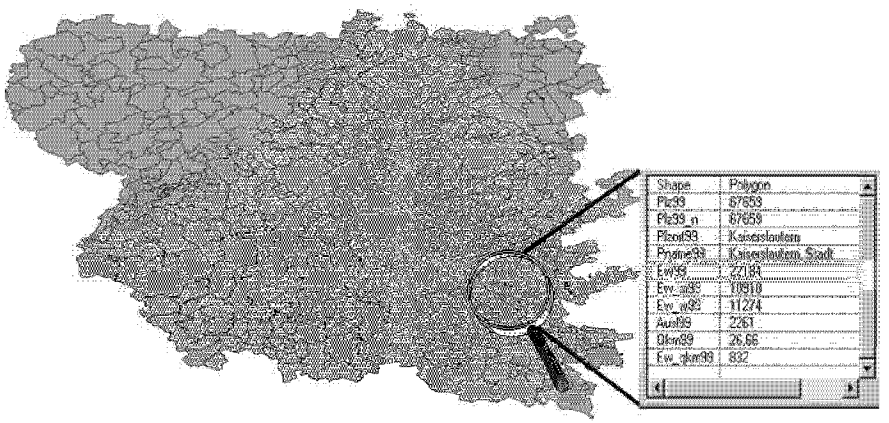

Figure 1: *German zip–code areas and associated demographical data. 'EW' abbreviates 'Einwohner' (inhabitants).*

For territory design problems usually one or more quantifiable attributes, called activity measures, are associated with each of the basic areas. Typical examples are workload for servicing or visiting

the customers within the area, estimated sales potential or number of inhabitants. See Figure 1 for an example.

We will assume here, that for each basic area $v \in V$ just a single activity measure $w_v \in \mathbb{R}_+$ is given. This may also be an aggregation of different values.

Territory centers In general, a center is associated with each territory. This may be some specific site, e.g. a salesman residence or office, or simply the geographical center of the territory. In general, the center is identical with the center of one of the basic areas comprising the territory. Therefore, we denote by $V_c \subset V$ the set of territory centers. In our model these centers can either be predetermined and fixed or subject to planning.

Number of territories For the remainder of the paper we assume that the number of territories is given in advance and is denoted by p. This is not a severe restriction since the algorithms presented below can be adapted to handle the number of territories as a planning parameter.

Unique assignment of basic areas We require every basic area to be contained in exactly one territory. Hence, the territories define a partition of the set V of basic areas. Let $B_i \subseteq V$ denote the i-th territory, then

$$B_1 \cup \cdots \cup B_p = V \text{ and } B_i \cap B_j = \emptyset, i \neq j.$$

Balance All territories should be balanced with respect to the activity measure. Hereby the activity measure (or size) of a territory is the total activity measure of the contained basic areas. Formally, $w(B_i) = \sum_{v \in B_i} w_v$ is the size of B_i.

Due to the discrete structure of the problem and the unique assignment assumption, perfectly balanced territories can generally not be accomplished. Therefore a common way to measure balance is to compute the relative percentage deviation of the district sizes from their average size μ. The larger this deviation is, the worse is the balance of the territory.

Contiguity In order to obtain contiguous districts, explicit neighborhood information for the basic areas is required. Although there exist several models which are based on a neighborhood graph for the basic areas, we will not incorporate this graph into our considerations.

However, the described solution procedures can easily be extended to take a neighborhood graph of the basic areas into account.

Compactness We model compactness, depending on the solution method, in two different ways. The first is to minimize the total weighted distance

$$\sum_{i=1}^{p} \sum_{v \in B_i} w_v d_{iv}$$

(Euclidean, squared Euclidean or network–based) from district centers to basic areas. For the geometric approach we will derive a compactness measure based on convex hulls to achieve compact territories (see Section 6 for details).

Objective The objective can be informally described as follows: partition the set V of basic areas into a number p of territories which satisfy the specified planning criteria like balance, compactness and contiguity.

To end this section we summarize the notation introduced above.

V	set of basic areas
(x_v, y_v)	coordinates of $v \in V$
$w_v \in I\!R_+$	activity measure of $v \in V$
V_c	set of territory centers
p	number of territories
$B_i \subset V$	i-th territory
$w(B_i)$	size of B_i
$\mu = w(V)/p$	average size of territories
d_{iv}	distance between v and the center of B_i

4 Overview of solution techniques

Many territory design approaches have appeared in the literature ranging from location–allocation and set–partitioning methods over divisional algorithms to local search methods and meta heuristics. For extensive reviews, see Howick and Pidd (1990) and Ricca and Simeone (1997).

4.1 Location–allocation methods

The first mathematical programming approach was proposed by Hess et al. (1965) for a center–seeking political districting problem. They modeled the problem as a capacitated $p-$median facility location problem. In this problem, we are given a number of already existing facilities (customers) and a number of candidate locations for new facilities. Furthermore, with every customer a demand for a specific product or service is associated which has to be satisfied. Moreover, the new facilities have a limited capacity for satisfying the customer demands. The task is now to locate a certain number of new facilities (e.g. plants, warehouses) among the candidate locations and allocate the already existing facilities to them, taking the capacity of the new facilities into account, such that the demands of the customers are satisfied "efficiently" from or at the new facilities.

Applied to the territory alignment problem, basic areas correspond to customers and their demand to the activity measure attributed to the basic area. Candidate locations for the new facilities are all basic areas. The new facilities to be located are also called *territory centers*. When solving the model, simultaneously new facilities are located among the candidate locations, i.e. the basic areas, and basic areas are allocated to the new facilities, i.e. territory centers. Here, for each territory a center will be located. All the basic areas allocated to the same new facility constitute a territory with the new facility as its center. (Note that this center is not necessarily the geographical center of the territory.) The capacity of the new facilities is chosen in such a way that the districts obtained by solving the problem are well balanced.

Unfortunately, due to its combinatorial complexity, the practical use of this model is fairly limited. To this end, Hess et al. (1965) (and subsequently Hess and Samuels (1971) in their GEOLINE model) used a *location–allocation* heuristic to solve the problem. In this heuristic, the simultaneous location and allocation decisions of the underlying facility location problem are decomposed into two independent phases, a location and an allocation phase, which are iteratively performed until a satisfactory result is obtained. In the location phase the centers of the territories are chosen while in the allocation phase the basic areas are assigned to these centers.

Location phase

There exist several approaches for determining a new configuration of territory centers. A fairly simple and commonly used method is to solve in each territory resulting from the last allocation phase a 1−median problem. See e.g. Fleischmann and Paraschis (1988), George et al. (1997). Alternatively, one can take the territory centers of the previous iteration and perturb them utilizing some local search technique to obtain a new configuration of centers, see e.g. Kalcsics et al. (2002). Hojati (1996) proposes to determine new centers based on the solution of a Lagrangean subproblem. It shows that the choice of territory centers has a considerable impact on the resulting territories in that a "bad" selection of centers will seldom yield acceptable territories.

Allocation phase

In most cases, the problem of allocating basic areas to territory centers is formulated as a capacitated assignment problem, see e.g. Hess et al. (1965) and also Section 5. While the balancing requirement is generally included as a side constraint, compact and contiguous territories are tried to be obtained by minimizing the sum of weighted distances between basic areas and territory centers. For political districting problems, authors tend to use squared Euclidean distances (e.g. Hess et al. (1965), Hojati (1996), whereas for sales territory design problems, largely straight line (Cloonan (1972), Marlin (1981)) or network distances (Segal and Weinberger (1977), Zoltners and Sinha (1983)) are employed. Marlin (1981) observes for his problem, that using squared Euclidean instead of straight line distances produces compact but disconnected territories. He concludes that the success of squared Euclidean distances depends on the ability to redefine territory centers and is not appropriate for the case of fixed centers. A similar phenomenon was observed by Hojati (1996). Although the model can easily be extended, e.g. to balance more than one activity measure, only those criteria can be incorporated which can be formulated in linear terms. This excludes for example more complex measures of compactness.

The assignment problem now is usually tackled by relaxing the integrality constraints on the assignment variables and solving the resulting capacitated transportation problem using specialized algorithms, like network flow methods, which are suitable for solving large scale problems.

Using this approach, George et al. (1997) solved a problem with up to 25000 basic areas. However, solving the relaxed problem yields optimal solutions which satisfy the balancing constraints but usually assign portions of basic areas to more than one territory center. To this end, Hess and Samuels (1971) proposed a simple tie breaking rule, named *AssignMAX*, which exclusively assigns the so–called *split areas* to the territory (center) which "owns" the largest share of the split area. In their applications they found, that a rate, i.e. mean number of areas per territory, of $n/m \geq 20$ was more than adequate to provide territories whose size was within $\pm 10\%$ of the average. Fleischmann and Paraschis (1988), however, report that for their application this simple heuristic gave very poor results. For about 50% of the resulting territories the restriction on the size of the territories was violated, in many cases heavily. (The mean number of areas per territory was approximately 8.) To this end, they presented a more sophisticated split resolution technique which tries to maximize the number of split areas that can be resolved without violating the size restriction on the territories. However, in this way, not all splits could be resolved automatically and some manual postprocessing was required. A quite similar idea to resolve split areas was proposed by Hojati (1996). Optimal split resolution techniques minimizing the maximal, total or standard deviation from the average are proposed by Schröder (2001) and will be discussed in more detail in Section 5.

In order to avoid split areas at all, Zoltners and Sinha (1983) propose a slightly different approach. They model the allocation problem as an integer program utilizing so–called SCU–adjacency trees and solve it using Lagrangian relaxation and subgradient optimization.

A completely different allocation approach is to sequentially assign basic areas to territory centers based on distance, i.e. a basic area will be allocated to closest territory center. This minimal distance allocation yields disjoint, compact and often connected, however, usually not well balanced territories as the balance criterion is completely neglected when deciding about the allocation. The attractiveness of this method, denoted as *AllocMinDist*, primarily lies in its simplicity and computational speed. See Kalcsics et al. (2002).

4.2 Divisional methods

Among these methods, the *successive dichotomies* strategy of Forrest (1964) and the *wedge–cutting* method of Chance (1965) can be mentioned. In the latter case, every district has the shape of a slice of cake and thus touches both the center and the boundary of the region under consideration. However, this approach does not pay much attention to compactness.

Forrest solves the problem by using the principle of diminishing halves. The idea of these types of methods is to iteratively partition the region under consideration into smaller and smaller subproblems, where a subproblem is defined by a set of basic areas and the number of territories, this set has to be partitioned into. The iteration stops if a level has been reached where the territory design problem for each of the subproblems can be solved easily; usually, if the subproblems have to be partitioned just into one territory and therefore already constitute a territory. Hence, given a subproblem, the basic operation is to divide the set of basic areas of the subproblem in a suitable way into two "halves". Unfortunately, Forrest did not provide any details on how he performed this division.

A simple, yet efficient way is to place a straight line in the plane through the set of basic areas of the subproblem, separating it into a right and a left half. The line should placed in such a way, that the two resulting subproblems are likely to yield contiguous, compact and well balanced territories upon further partitioning. See Section 6 for more details.

4.3 Other approaches

Several other methods have been proposed over the past decades, see also Howick and Pidd (1990) and Ricca and Simeone (1997).

Set–partitioning models Garfinkel and Nemhauser (1970) proposed a set partitioning based approach to tackle the problem. In a first step, candidate territories are generated which are contiguous, compact and have a total electorate within the tolerance and, in a second step, territories are selected from the set of candidates to optimize the overall balance of the district plan. See also Garfinkel (1968).

Mehrotra et al. (1998) picked up this model, merely exchanging the objective function by one which minimizes the overall compactness of

the territories. They developed a column generation algorithm, which is capable to consider many more potential districts than the initial approach of Garfinkel and Nemhauser and applied it to a districting problem with up to 50 basic areas. Similar approaches are taken by Shanker et al. (1975) and Nygreen (1988).

A major advantage compared to location–allocation methods is, that almost any criterion can be applied on the generation of candidate districts. However, due to the combinatorial complexity, set–partitioning models have not been used with more than 100 basic areas.

Criteria methods These are manual approaches, which provide sales management with data and an objective for sales territory alignment, but do not provide a methodology for actually designing the territories. See e.g. Easingwood (1973), Lodish (1975).

Eat–up In this approach, one territory after the other is extended at its boundary through successively adding yet unassigned, adjacent basic areas to the district, until it is sufficiently large. See e.g. Mehrotra et al. (1998).

Clustering Deckro (1977) proposed an approach, where each basic area is initially treated as a single district. Then iteratively pairs of districts are merged together forming new and bigger territories until the prescribed number of districts is reached.

Multi–kernel growth This method starts by selecting a certain number of basic areas as "seeds"(centers) for the districts. The algorithm then successively adds to each center neighboring basic areas, in order of decreasing distance, until the desired territory size is reached. See e.g. Bodin (1973).

Local search The well known local search techniques are heuristic methods, which try to improve an existing territory plan by successively shifting basic areas between neighboring territories with the aim of minimizing a weighted additive function of different planning criteria.

A simple approach is employed by Bourjolly et al. (1981). Algorithms based on simulated annealing are proposed by Browdy (1990), Macmillan and Pierce (1992) and D'Amico et al. (2002). Ricca (1996) develops descent, simulated annealing and tabu search algorithms.

The latter technique has been successfully applied in the recent papers of Bozkaya et al. (2003) and Blais et al. (2003). See also the upcoming book of Bozkaya et al. (2005).

Genetic algorithms Genetic algorithms for solving territory design problems have been introduced recently by Forman and Yue (2003) and Bergey et al. (2003). The former authors utilize a technique based on an encoding and on genetic operators used to solve Traveling Salesman Problems. The encoding chosen is a path representation and a single chromosome travels through each basic area, and as the areas are traversed, territories are formed by the sequence of basic areas. Bergey et al. (2003) use their own representation and, moreover, incorporate a simulated annealing method to improve results.

5 A location–allocation method with optimal split resolution

Hess et al. (1965) were the first to model the problem of designing political districts as a mixed integer linear program. Essentially the model is discrete capacitated facility location problem.

Trying to solve this $\mathcal{NP}$-hard MIP with a commercial solver was not possible at the time of the paper of Hess et al. and is today still not suitable for a decision support system in an interactive environment. The reason is that the running time of the solver to find an optimal solution and to prove its optimality depends in a non-predictable way on the problem data. The user in the interactive environment on the other hand will not accept such a behavior of his software tool.

Therefore to solve their MIP in a heuristic fashion, Hess et al. use a location-allocation procedure. In the location phase the centers of the districts are chosen while in the allocation phase the basic areas are assigned to these centers. The location part is simple, in each territory resulting from the last allocation phase a 1-median problem is solved.

Here we are concerned mainly with the allocation phase. In the following we will discuss how Hess et al. solve it and present a technique for optimal split resolution. The material in this section is taken from Schröder (2001).

5.1 The allocation problem

Let V be the set of basic areas and $V_c \subset V$ the set of p territory centers. For $v \in V$ and $i \in V_c$ let d_{iv} be the distance from i to v. The average territory size is $\mu = w(V)/p$. Let $\tau > 0$ be a tolerance value for the deviation of the actual sizes of the territories from μ.

With the assignment variables

$$x_{iv} = \begin{cases} 1 & \text{if basic area } v \text{ is assigned to center } i \\ 0 & \text{otherwise} \end{cases}$$

Hess et al. (1965) proposed the following allocation model:

$$\min \quad \sum_{v \in V} \sum_{i \in V_c} w_v \, d_{iv} \, x_{iv} \qquad \text{(ALLOC)}$$

$$\text{s.t.} \quad \sum_{i \in V_c} x_{iv} = 1 \qquad \forall \, v \in V \qquad (5.1)$$

$$(1-\tau)\mu \leq \sum_{v \in V} w_v \, x_{iv} \leq (1+\tau)\mu \qquad \forall \, i \in V_c \qquad (5.2)$$

$$x_{iv} \in \{0,1\} \qquad \forall \, v \in V, i \in V_c \qquad (5.3)$$

The objective function minimizes the overall distance, weighted with the activity measure, from the basic areas to the respective territory centers. The model tends to produce compact and also geographically connected territories. Constraints (5.1) ensure that each basic area is allocated to exactly one territory center (disjointness criterion). By (5.2) we enforce that the size of each of the territories is within the predefined tolerance (balance criterion). In total we have a quadratic number of decision variables and a linear number of constraints.

The outcome of the model depends a great deal on the parameter τ. The smaller the tolerance τ is the better the balance of the obtained territories. Unfortunately, if τ is too small, i.e. the upper and lower bounds on the size of the districts in constraints (5.2) are very tight, then territories tend to be no longer compact and connected; the problem might even become infeasible. In addition the complexity of the problem and therefore the time for solving it generally increases the smaller τ is. On the other hand the larger the tolerance is the worse the balance of the territories will be.

The way of Hess et al. to overcome this problem is as follows. They set the tolerance to $\tau = 0$ and relax the integrality constraint on the assignment variables $x_{iv} : x_{iv} \in [0, 1]$. Then x_{iv} is the fraction of basic area v allocated to territory center i.

This relaxed problem is a linear program which is basically a classical transportation problem.

$$\min \quad \sum_{v \in V} \sum_{i \in V_c} w_v \, d_{iv} \, x_{iv} \qquad \text{(TRANSP)}$$

$$\text{s.t.} \quad \sum_{i \in V_c} x_{iv} = 1 \qquad \forall \, v \in V \qquad (5.4)$$

$$\sum_{v \in V} w_v \, x_{iv} = \mu \qquad \forall \, i \in V_c \qquad (5.5)$$

$$x_{iv} \geq 0 \qquad \forall \, v \in V, i \in V_c \qquad (5.6)$$

Problem (TRANSP) can efficiently be solved using specialized network algorithms. In the optimal solution of (TRANSP) the territories are perfectly balanced. On the other hand the criterion of disjointness is generally not satisfied, due to the continuous variables in (TRANSP). A basic area v for which more than one variable x_{iv}, $i \in V_c$ has positive values is called *split area* or just *split*. For a basic (optimal) solution of (TRANSP) it is easy to prove that there are at most $p - 1$ splits.

To establish the criterion of disjointness after the solution of (TRANSP) it is necessary to round for every split its fractional variables to one (one variable) or zero (the other variables). This yields disjoint territories but on the other hand destroys their perfect balance. Since there are many possibilities for the rounding, we want to find one that results in as much as possible balanced territories. We call this the *split resolution problem.*

5.2 The split resolution problem

Let $V^S \subset V$ be the set of splits in the optimal solution of (TRANSP). We want to resolve the splits while keeping the territories as balanced as possible. To quantify this objective we have different possibilities:

$$\text{Minimize } \max\{|W_i - \mu| : i \in V_c\} \text{(maximum deviation)} \qquad (\text{SPRES}_\infty)$$

$$\text{Minimize } \sum(|W_i - \mu| : i \in V_c) \text{(total deviation)} \qquad (\text{SPRES}_1)$$

$$\text{Minimize } \sum((W_i - \mu)^2 : i \in V_c)(\text{equiv. to standard deviation}) \qquad (\text{SPRES}_2)$$

where $W_i = w(B_i)$ is the size of the territory with center i.

To further examine the problem of split resolution we need a definition, given first by Fleischmann and Paraschis (1988): Let (x_{iv}) be an optimal basic solution of (TRANSP). The graph $T^S = (U^S, E^S)$ with $U^S = V_c \cup V^S$ and $E^S = \{(i, v) : 0 < x_{iv} < 1\}$ is called *split adjacency*. The edges of the split adjacency correspond to the fractional variables in the solution (x_{iv}) of (TRANSP). Clearly, T^S is cycle-free, i.e. a forest. Figure 2 visualizes the definition of the split adjacency.

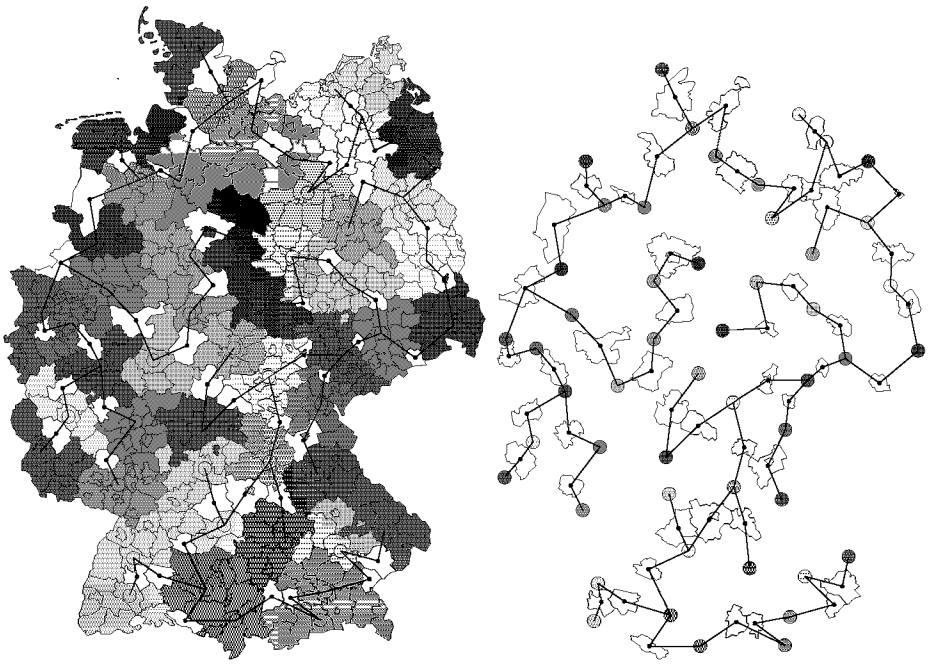

Figure 2: *Left: structure of the optimal solution of (TRANSP). The edges shown correspond to fractional variables. Right: the corresponding split adjacency.*

In Hess and Samuels (1971) and also in George et al. (1997) a simple rule is proposed for split resolution: They assign every split v fully to the center i for which x_{iv} is maximum. Schröder (2001) calls this heuristic for resolving the splits *AssignMAX*.

In Schröder (2001), page 147, it is proven that AssignMAX has no finite guarantee for minimizing the maximum deviation (i.e. problem (SPRES_∞)). An example is given in which the maximum deviation resulting from AssignMAX grows as $d^2 - d$, while it is possible to resolve the splits in a way that yields a maximum deviation of $2d$. Here $d \geq 4$ is the maximum node degree in the split adjacency used in the example.

This result is more of theoretical interest since in split resolution problems coming from real world data we can assume that the degrees of the nodes in the split adjacency are not very large. The following theorem gives an upper bound on the maximum deviation of the territory sizes from μ in terms of the degrees in the split adjacency.

Theorem 5.1. *Let* $\bar{\delta}_V = \max\{deg(v) : v \in V^S\}$, $\bar{\delta}_I = \max\{deg(i) : i \in V_c\}$ *(degrees in* T^S*) and* $\bar{w} = \max\{w_v : v \in V^S\}$. *Then*

$$\max\{|W_i - \mu| : i \in V_c\} \leq (1 - 1/\bar{\delta}_V)\bar{\delta}_I\bar{w},$$

where W_i *is the size of the territory with center* i *after the application of AssignMAX.*

Proof. For a given $v \in V^S$ let $i \in V_c$ be the center to which v is assigned by AssignMAX. By the definition of this heuristic

$$x_{iv} = \max\{x_{i'v} : i' \in V_c, x_{i'v} > 0\}.$$

The number of positive $x_{i'v}$, $i' \in V_c$ is equal to $\deg(v)$ in T^S. Therefore $x_{iv} \geq 1/\deg(v)$ and by assigning v to i the size of the territory with center i increases by at most $w_v(1 - 1/\deg(v))$. It follows that the maximum size of a territory is at most $(1 - 1/\bar{\delta}_V)\bar{\delta}_I\bar{w}$ larger than μ.

If on the other hand a split v is not assigned to center i then $x_{iv} \leq 0.5$. Consequently the minimum size of a territory is at most $(1/2)\bar{\delta}_I\bar{w}$ smaller than μ. Since $\bar{\delta}_V \geq 2$ this is bounded by $(1 - 1/\bar{\delta}_V)\bar{\delta}_I\bar{w}$. □

5.3 Optimal split resolution

Solving the problem of split resolution in an optimal fashion means that we solve one of the problems (SPRES_1), (SPRES_2) and (SPRES_∞) to optimality. The following result, proved in Schröder (2001), shows the computational complexity of these problems.

Theorem 5.2. *(SPRES$_1$) and (SPRES$_2$) are $\mathcal{NP}$-hard, while (SPRES$_\infty$) can be solved polynomially.*

This result assumes that the degrees in the split adjacency can become arbitrarily large. If we assume that in practical problems the maximum degree in T^S is bounded, the algorithm we present next for the optimal solution of the split resolution problem runs in linear time. This algorithm is generic and solves (SPRES$_*$), where $*$ is one of $1, 2, \infty$.

Our algorithm for optimal split resolution is a dynamic programming procedure and is based on the absence of cycles in the split adjacency. W.l.o.g we can assume that the split adjacency T^S is a tree (otherwise we insert some arbitrarily chosen edges into the forest until it is connected). Further we select some node $v_0 \in V_c$ and consider T^S as a rooted tree with root v_0.

The procedure works in T^S from the leaves to the root and computes the optimal value of the selected objective function. Thereby it makes use of the fact that there are two different types of nodes in T^S corresponding to splits and territory centers. Every edge in T^S joins two nodes of different type.

The process of split resolution is equivalent to the solution of a tree partitioning problem for the split adjacency: partition T^S into subtrees that contain exactly one node from V_c. More precisely the node set U^S is partitioned into subsets that induce a subtree of T^S and have a one-element intersection with V_c.

To proceed we need some notation. Let $T' = (U', E')$ be a subtree of T^S. T' is feasible if $|U' \cap V_c| = 1$. We associate a cost with T', depending on the selected objective function:

$$c(T') := c(U') := \begin{cases} |w(U') - \mu| & \text{for (SPRES}_1\text{) and (SPRES}_\infty\text{)} \\ |w(U') - \mu|^2 & \text{for (SPRES}_2\text{)} \end{cases}$$

The cost measures the deviation of the weight of the subtree from the average weight. Here we define the weight of a node $i \in V_c$ corresponding to a center as the total weight of this node and all uniquely assigned nodes v in V, i.e. with $x_{iv} = 1$.

Since we are searching for the optimal partition of T^S into feasible subtrees, we define the cost of a partition $\pi = \{T'_i : i \in V_c\}$ as a generalized

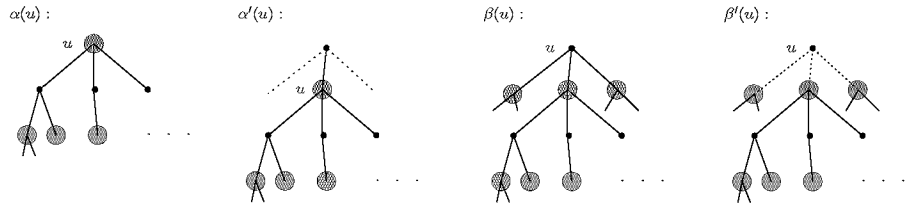

Figure 3: *Partial optima in the dynamic programming procedure for the partitioning of T^S. Circles depict center nodes, dots are split nodes.*

sum of the cost of the subtrees

$$c(\pi) = \bigoplus_{i \in V_c} c(T_i')$$

where

$$\oplus = \begin{cases} + & \text{for (SPRES}_1\text{) and (SPRES}_2\text{)} \\ \max & \text{for (SPRES}_\infty\text{)} \end{cases}$$

A partition π of minimal cost corresponds to a split resolution with most balanced territories, where the deviation is measured by the selected objective function (SPRES$_*$).

We introduce some further notation related to the rooted tree T^S. For $u \in U^S$ let T_u denote the subtree rooted in u, i.e. the subtree containing u and all its descendants. pre(u) is the father of u and S_u is the set of sons of u.

In the recursion of the dynamic programming procedure we denote the optimal value of the objective function for different partial problems in the following way (see Figure 3).

- For $u \in V_c$, $\alpha(u)$ is the cost of an optimal partition of T_u.
- For $u \in V_c$, $\alpha'(u)$ is the cost of an optimal partition of $T_u \cup \{\text{pre}(u)\}$.
- For $u \in V^S$, $\beta(u)$ is the cost of an optimal partition of T_u.
- For $u \in V^S$, $\beta'(u)$ is the cost of an optimal partition of the forest $T_u - \{u\}$.

To work form the leaves to the root in T^S we use the following recursions. If u is a leaf then $u \in V_c$ (since all nodes corresponding to splits have degree ≥ 2 in T^S) and

$$\alpha(u) = c(\{u\}) \qquad \alpha'(u) = c(\{u, \text{pre}(u)\}).$$

If $u \in V^c$ is not a leaf

$$\alpha(u) = \min_{S' \subseteq S_u} \left\{ c(\{u\} \cup S') \oplus \bigoplus_{u' \in S'} \beta'(u') \oplus \bigoplus_{u' \in S_u - S'} \beta(u') \right\} \tag{5.7}$$

$$\alpha'(u) = \min_{S' \subseteq S_u} \left\{ c(\{u, \text{pre}(u)\} \cup S') \oplus \bigoplus_{u' \in S'} \beta'(u') \oplus \bigoplus_{u' \in S_u - S'} \beta(u') \right\} \tag{5.8}$$

Here S' denotes the set of sons that are included in the subtree for the center node u.

If $u \in V^S$

$$\beta(u) = \min_{u' \in S_u} \left\{ \alpha'(u') \oplus \bigoplus_{u'' \in S_u - \{u'\}} \alpha(u'') \right\} \tag{5.9}$$

$$\beta'(u) = \bigoplus_{u' \in S_u} \alpha(u') \tag{5.10}$$

The verification of (5.7) to (5.10) is straightforward. Since nodes of center type and nodes of split type alternate in T^S equations (5.7) to (5.10) can be applied recursively until finally $\alpha(v_0)$ is found. This is the cost of an optimal partition. This partition (and hence the optimal split resolution) can then be determined by backwards calculation.

The calculation of $\alpha(u)$ and $\alpha'(u)$ according to (5.7) and (5.8) generally requires exponential time, whereas the calculation of $\beta(u)$ and $\beta'(u)$ (equations (5.9) and (5.10)) can be done in linear time. In summary we find:

Theorem 5.3. *An optimal split resolution can be computed in* $O(\bar{\delta}_I 2^{\bar{\delta}_I} |V_c| + \bar{\delta}_V |V^S|)$ *steps.*

For bounded maximum degree in the split adjacency this is linear. Note that also AssignMAX requires linear time. For problem (SPRES_∞) the optimization problems (5.7) and (5.8) can be solved in polynomial time, regardless of the degree of nodes in T^S. More precisely the following theorem is proved in Schröder (2001).

Theorem 5.4. *($SPRES_\infty$) can be solved in $O(\bar{\delta}_I^3 |V_c|^2 \log \bar{w} + \bar{\delta}_V |V^S|)$ steps.*

Equations (5.7) to (5.10) yield also a better worst case estimate for (SPRES_∞) than AssignMAX (cf. Theorem 5.1).

Theorem 5.5. *Let $\bar{w} = \max\{w_v : v \in V^S\}$. For the optimal solution of ($SPRES_\infty$) we have*

$$\max\{|W_i - \mu| : i \in V_c\} \leq \bar{w},$$

where W_i is the size of the territory with center i.

The proof is not difficult and works by showing that for all $u \in U^S$ $\alpha(u), \alpha'(u) \leq \bar{w}$ and $\beta(u), \beta'(u) \leq \bar{w}$, respectively. This can be done by induction using equations (5.7)–(5.10) written down for the case of (SPRES_∞). See Schröder (2001), page 161, for details.

5.4 The location phase

The allocation method proposed in the preceding section can be combined with any method to determine good territory centers. Approaches for this task have been shortly described in Section 4.1. We refer to the literature cited there for more information on the location phase.

5.5 Comment on the location-allocation method

The location-allocation method has the advantage that the solution of the MIP (ALLOC) is not required. Instead in each iteration the linear program (TRANSP) has to be solved. This can be done rather efficiently. However we found that the running time is still too high for the solution of large-scale problems with many thousands of basic areas in an interactive environment.

We can make the allocation step much faster by assigning every basic area to the nearest center. This means that we drop constraints (5.2). This

AllocMinDist heuristic has a greatly reduced running time. However, as one would expect and as the computational results in Section 7 show, the balance of the territories obtained is not satisfactory.

Therefore in the next section we present a new heuristic based on geometric ideas. It has the desirable property of being very fast (comparable to location-allocation with *AllocMinDist*) and producing territories that are balanced comparable to location-allocation with (TRANSP) and split resolution with AssignMAX.

6 A computational geometry based heuristic

Although the idea of using methods of computational geometry has already been mentioned in the literature, Forrest (1989), no details were given.

The idea of the method presented here is to recursively partition the complete problem geometrically using lines into smaller subproblems until an elemental level is reached where we can efficiently solve the territory design problem for each of the elemental subproblems. The solutions to these problems then directly yield a solution for the original problem.

A territory is given by a subset B of V. The heuristic strives to align territories that are balanced with respect to the activity measure. Ideally we would have $w(B) = w(V)/p$ for every district, but in general this is not possible due to the discrete nature of the problem. Therefore we assume that a lower bound L and an upper bound U for the activity measure of a territory are given. For example, L and U can be calculated from a maximally allowed deviation $\tau > 0$ from average size by

$$L = (1 - \tau)w(V)/p \quad \text{and} \quad U = (1 + \tau)w(V)/p. \tag{6.1}$$

A territory B is called *feasible* if $L \leq w(B) \leq U$.

In the following we will often identify (sets of) basic areas with (sets of) points in the plane.

6.1 Main ideas of the heuristic

The basic operation of this heuristic is to divide a subset $V' \subseteq V$ of the basic areas, i.e. points, into two "halves" V'_l and V'_r by placing a line in the plane

within this set of points. V_l' (V_r') are then defined as the set of points, i.e. basic areas, located left (right) of the line. By this we partition the territory design problem for V' into two disjoint subproblems, one for V_l' and one for V_r'. These subproblems are then solved independently from one another again by dividing each of them along a line. This iterative partitioning into subproblems gives the heuristic its name: successive dichotomies (termed in Ricca and Simeone (1997)).

Since a problem that is not trivial generates two subproblems, the problems our heuristic examines are related according to a binary tree. The root of the tree is the problem we start with and the leaves correspond to territories.

Figure 4 shows an example of partitioning the original problem with basic areas V into two disjoint subproblems with basic areas V_l and V_r, respectively. Figure 5 illustrates the two subproblems generated by the solution of a problem.

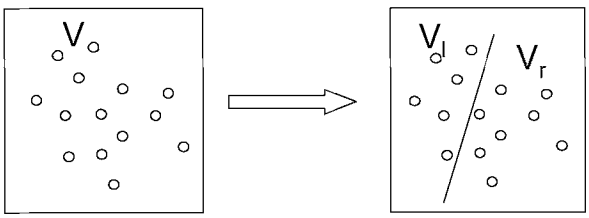

Figure 4: *A partitioning of problem V into two disjoint subproblems with basic areas V_l and V_r, respectively.*

Our heuristic explores the binary tree with nodes corresponding to problems and terminates when all leaves are generated. Two questions need to be answered:

- How do we perform the partitioning of a problem into subproblems? (Section 6.2)
- How do we explore the tree? (Section 6.3)

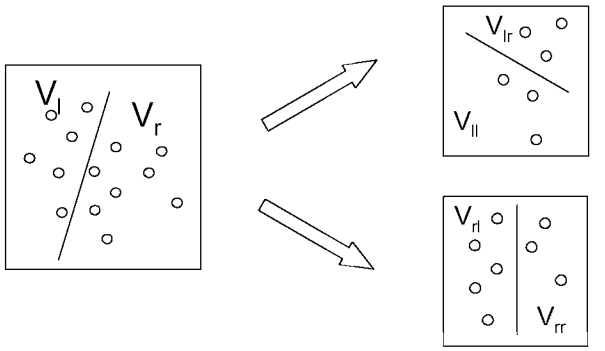

Figure 5: *Every problem generates two subproblems.*

6.2 Solving the basic problem

The initial problem is to partition V into p territories. An instance of a (sub-)problem (called *basic problem*) is defined by $V' \subseteq V$ and a positive number $p' \leq p$. A basic problem with $p' = 1$ is trivial, since then V' defines a territory. If $p' > 1$ we solve the problem by choosing a line and numbers $p'_l, p'_r \geq 1$ with $p'_l + p'_r = p'$. This yields two new subproblems (V'_l, p'_l) and (V'_r, p'_r) that replace problem (V', p').

The method to solve the basic problem is designed to take the criteria of balance and compactness into account. Further it is important that it runs fast, since many basic problems need to be solved in the search tree.

In an instance of the basic problem we are given a subset V' of basic areas as points in the plane and a number $p' \geq 1$. As already mentioned above we have to make two decisions:

1. Select $p'_l, p'_r \geq 1$ with $p'_l + p'_r = p'$.
2. Select a line to partition V' into two subsets V'_l and V'_r of points left of and right of the line, respectively.

How p'_l and p'_r are selected depends on whether p' is even or odd. If p' is even we simply set $p'_l = p'_r = p'/2$. If p' is odd we consider two cases,

$$p'_l = \left\lceil \frac{p'}{2} \right\rceil, p'_r = \left\lfloor \frac{p'}{2} \right\rfloor \quad \text{and} \quad p'_l = \left\lfloor \frac{p'}{2} \right\rfloor, p'_r = \left\lceil \frac{p'}{2} \right\rceil.$$

For the second decision one could imagine to consider more general methods to define a partition of V' than separating the points along a line. But it has two advantages to do it in this way, stated in the following propositions.

Proposition 6.1. *If the basic problem is solved by partitioning V' into two parts by a separating line, we yield territories which have (when considered as subsets of V) pairwise disjoint convex hulls.*

Proof. Obviously V'_l and V'_r have disjoint convex hulls. Apply this argument recursively. □

In Proposition 6.1 we assume that none of the points in V' is located on the separating line. However the statement stays true if we assume that points located on the line are always included into V'_l.

Proposition 6.2. *If no three points in V' lie on a common line, the number of partitions of V' along a line is bounded by $|V'|^2$.*

Proof. Each such partition is induced by an ordered pair (v, u) of (not necessarily distinct) points in V' in the following way: If $v \neq u$ they define a unique ordered line, we define V'_l to contain all points left of the line together with v and u. If $v = u$ and v can be separated from the other points in V' by a line we set $V_l = \{v\}$. □

Generating partitions

Proposition 6.2 limits the number of partitions that are interesting to be considered by a quadratic term. Unfortunately this is too much for large scale problems. Therefore we decided to examine not all of these partitions. Instead only those partitions of V' are considered that are generated by lines with some fixed directions. Only a small number of directions is used. This seems to be rather restrictive but we found that it produces very good results. The major advantage however is that all partitions generated by lines of a fixed direction can be examined very fast.

To explain why, let the direction be given by $\alpha \in [0, \pi)$, the angle of the line with the positive x-axis. Let us first assume that $\alpha = \pi/2$, i.e. we consider separating lines parallel to the y-axis. Before examining

the partitions generated by such lines, we sort the points in V' by non–decreasing x-coordinate x_v. It is clear that every possible partition along a line parallel to the y-axis divides this sorted sequence into a left and a right part. To examine all partitions we only have to examine all subdivisions of the sequence into a left and a right part. Thus there are as many (nontrivial) partitions as points in the sequence minus one, which is linear in $|V'|$.

See Figure 6 for an example. Every vertical line, i.e. a line parallel to the y-axis, through each of the points generates a partition of the original problem into two subproblems.

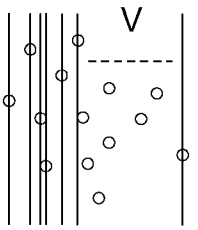

Figure 6: *Possible partitions of the problem using vertical lines.*

If α is different from $\pi/2$ the same idea applies after rotating the coordinate system so that the line through the origin with angle α becomes the y-axis. Then again we sort the points by non–decreasing x-coordinate. The new x-coordinates of the points after rotating the coordinate system are given as

$$x_v \sin\alpha + y_v \cos\alpha.$$

Before we explain how we examine a partition of V' under the criteria of balance and compactness, we summarize how the partitions that we consider are generated.

Step 1. Select a number $N \geq 2$ of line directions to consider.

Step 2. For $i = 0, 1, \ldots, N-1$ let $\alpha_i = \pi/i$.

Step 3. Consider direction α_i: sort the points in V' by non–decreasing value of $x_v \sin\alpha_i + y_v \cos\alpha_i$. Let this sequence be denoted $v_1, v_2, \ldots, v_s$.

Step 4. For $k = 1, 2, \ldots, s-1$ "examine" the partition given by

$$V'_l = \{v_1, \ldots, v_k\} \text{ and } V'_r = \{v_{k+1}, \ldots, v_s\}.$$

We repeat steps 3 and 4 for all N directions. We found that $N = 8$ or 16 provide good results.

Examining a partition

The quality of a partition V'_l, V'_r of V' and $p'_l + p'_r = p'$ depends mainly on two factors.

- What are the average activity measures $w(V'_l)/p'_l$ and $w(V'_r)/p'_r$? (Balance)
- How 'compact' are (the convex hulls of) V'_l and V'_r? (Compactness)

Balance

First we discuss balance. Ideally we would have $w(V'_l)/p'_l = w(V'_r)/p'_r = w(V')/p'$, since then we would finally get territories with exactly the same size. Due to the discrete nature of the problem this is generally not possible.

Therefore we try to come as close as possible to perfect balance. In the sequence $v_1, \ldots, v_s$ of Step 3 above we determine an index k such that

$$w(\{v_1, \ldots, v_{k-1}\}) < (p'_l/p')\, w(V') \text{ and } w(\{v_1, \ldots, v_k\}) \geq (p'_l/p')\, w(V') \tag{6.2}$$

(If all w_v are positive k is unique.) Only the partition generated for this value k is considered further. All other partitions generated in step 4 above are not balanced enough. But also the partition for k will be discarded if it is *infeasible*. This is the case when $k < p'_l$ or $s - k < p'_r$ or when $w(V'_l)/p'_l$ or $w(V'_r)/p'_r$ is not in the interval $[L, U]$.

Consequently for each direction α_i we consider between zero and two feasible partitions, depending also on whether p' is even or odd. All these partitions are then ranked according to the compactness criterion presented next.

Compactness

Every partition we consider is generated by a line L which is defined by the direction α_i and the location of basic area v_k according to (6.2). To find a

compact territory design we use a measure that is based on the following reasoning. The segment of L that lies "within" V' will contribute to the total length of territory borders in the final territory layout. If we try to make this segment short we can hope to end up with a small total border length, and therefore with a compact layout.

For example in Figure 7 two possible choices of lines for partitioning a given set of basic areas into two smaller subproblems are illustrated. Intuitively the line on the right–hand side will yield more compact territories than the one on the left–hand side.

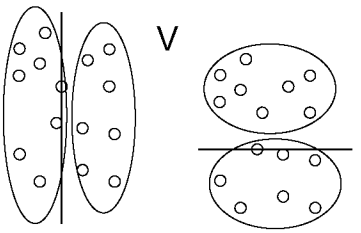

Figure 7: *Two possible partitions with different compactness.*

Since V' is a discrete set of points, it is not clear however what the length of the intersection of line L with V' is. We use two methods to define this length.

The first one is to use the length of the intersection of L with the convex hull C of V'. By convexity if v_k is inside of C we see that L intersects C in two points. The Euclidian distance of these two points defines the length of the segment (Figure 8, left picture). If v_k is a vertex of C the length can be zero. See Figure 8 left–hand side picture.

Using the convex hull works well if the points in V' are uniformly distributed within C. Typically this is the case when $|V'|$ is not very large. But consider the example of Figure 8, right–hand side. Here the convex hull does not describe the distribution of the points in V' well, and the length of the segment of L within the hull does not give the right information on how long the part L within V' is.

Therefore if $|V'|$ is large we use a different method to measure compactness. We consider all points of V' that are close to L, i.e. lie in a stripe whose width is a fixed percentage of the width of the hull C. (Width

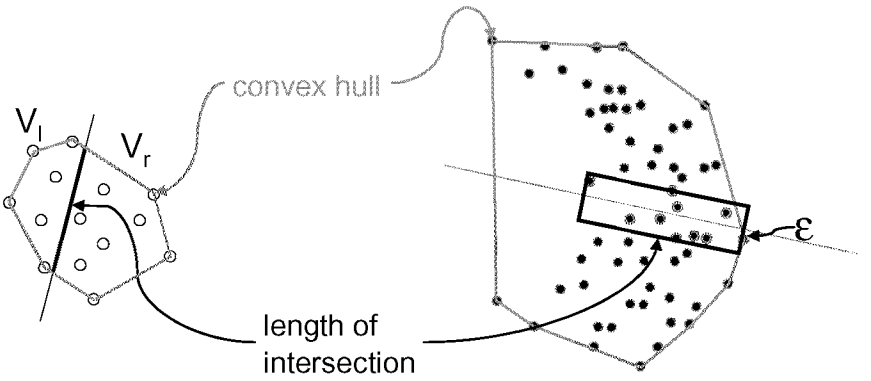

Figure 8: *Two options to determine the length of the intersecting line: intersection with the convex hull or using a stripe.*

measured orthogonally to L.) We project these points onto L and define the largest distance between any two of the projections as length of the segment of L that lies within V'.

6.3 Search tree to find a good territory design

In the last section we explained how a basic problem is "solved" by partitioning it into two subproblems. Since several line directions are considered we have different partitions into subproblems. In order to choose an appropriate partition for the subdivision all of them are ranked by heuristic measures for balance and compactness.

The straightforward "greedy" approach to realize just the best partition according to this ranking is however not sufficient. Even though only the best balanced partition among those possible for a certain line direction is considered there is no guarantee that one does not encounter an infeasible subproblem later, i.e. one with $w(V')/p'$ outside the interval $[L, U]$.

Therefore we incorporated a backtracking mechanism into our heuristic. It allows to revisit a basic problem at a higher level to revise the division made there. We then perform the next partition according to our ranking and continue the search. In the following we explain the search mechanism more formally.

The search tree

A node $\varphi = (V_\varphi, p_\varphi)$ of the search tree corresponds to a basic problem: subdivide V_φ into p_φ territories. The status of a node can be *active, inactive* or *isLeaf*. The latter is the case if p_φ equals one. The root node is $\varphi_0 = (V, p)$. It is active at the beginning of the search.

In each iteration of the search, an active node $\varphi = (V_\varphi, p_\varphi)$ is selected. If $p_\varphi = 1$ the status of the node is set to *isLeaf*. Otherwise it is set to *inactive*, and by using the highest ranked unused feasible partition for the basic problem of this node, two new active nodes $\varphi_l = (V_{\varphi_l}, p_{\varphi_l})$ and $\varphi_r = (V_{\varphi_r}, p_{\varphi_r})$ are generated. Here V_{φ_l} (V_{φ_r}) are the basic areas left (right) of the dividing line. Now all feasible partitions for the two generated subproblems φ_l and φ_r are computed and ranked due to the balance and compactness criterion.

The search terminates when no active nodes are left. The set of leaf nodes then corresponds to a territory plan. If for an active node φ and its corresponding basic problem no feasible partitions are left, a backtrack operation is performed as follows. The father node of φ, which is inactive, is made active again and all its descendant nodes are deleted. Afterwards the search continues with this node.

If the situation occurs that we have to backtrack from the root node it is proved that the problem is infeasible. This occurs if L and U are defined too constraining (under the selected number of line directions).

Limiting the size of the search tree

The search encounters at least $2p - 1$ active nodes until it terminates, this is the minimum number of nodes of a binary tree with p leaves. Due to backtracking operations however the number of nodes examined can be much larger. Especially proving infeasibility of the problem requires to examine all feasible partitions for all basic problems, in general this number is exponential.

Therefore it is necessary to limit the search. One possibility is to stop after a maximum number of nodes has been explored and to output that no feasible (with respect to L and U) territory design has been found.

We chose another way since it seemed better to us to output some result,

even an infeasible one, instead of no result. Therefore, after a certain node limit is reached, we decrease L and increase U by some amount and thus enlarge the number of feasible partitions. We do not restart the search so the relaxed bounds apply only to newly generated nodes of the search tree. The change in L and U is made in such a way that the tolerance τ (see (6.1)) is doubled. This relaxation mechanism is repeated a few times if necessary. If then the search still does not terminate, we finally set L to zero and U to infinity. Afterwards the search performs no more backtracking and terminates quickly.

6.4 Outline of the successive dichotomies heuristic

Now a rough outline of the successive dichotomies heuristic will be given:

Input Set of basic areas V with corresponding activity measures w_v, $v \in V$, and number of territories p. Parameter τ.

Step 1 Initialization
Compute the values $L = (1 - \tau)\, w(V)/p$ and $U = (1 + \tau)\, w(V)/p$.
Set the status of the root node $\varphi_0 = (V, p)$ to *active* and compute and rank all feasible partitions.

Step 2 **While** there are *active* nodes left

Let $\varphi = (V_\varphi, p_\varphi)$ be an *active* node.

/ Leaf node */*
If $p_\varphi = 1$ **then** set the status of φ to *isLeaf*. **Continue** with *Step 2*.

/ Backtrack */*
If there are no more feasible partitions for φ left **then**

If $\varphi = \varphi_0$ is the root node
then relax L and U and compute and rank again all feasible partitions of φ_0. **Continue** with *Step 2*.
else set the father node φ_f of φ to *active* and delete all descendant nodes of φ_f. **Continue** with *Step 2*.

/ Partition */*
Implement the highest ranked partition creating two new new *active*

nodes $\varphi_l = (V_{\varphi_l}, p_{\varphi_l})$ and $\varphi_r = (V_{\varphi_r}, p_{\varphi_r})$. Delete this partition from the list of partitions of node φ. Compute and rank all feasible partitions of φ_l and φ_r.
Set φ to *inactive*.

/ Limiting search tree */*
If some (node) limit is exceeded **then** relax L and U.

Output A territory design made up of all nodes with status *isLeaf*.

In Figure 9 an example of two sales territory alignments of German zip–code areas (indicated as points) into 70 territories created by applying the above heuristic is presented. Two different sets of line directions were used: one with 2 (left–hand side image) and the other with 16 directions (right–hand side picture).

Figure 9: 70 *territories based on German zip–code areas.*

6.5 Combination of successive dichotomies with optimal split resolution

Here we want to sketch how the methods in Sections 5 and 6 can be combined to yield a modified successive dichotomies heuristic with a priori worst case bound on the deviation of the territory sizes from the average size.

The idea how to combine the methods is fairly simple, and relates to the way in which the basic problem is solved in 6.2. There we partition the subset V' into two subsets V'_l and V'_r of approximately equal size. Now, if in this process it is allowed to 'split' the elements in V', it is clearly possible to achieve $w(V'_l) = w(V'_r)$ by splitting at most one $v \in V'$.

Solving the basic problem with splitting allowed in this way during the search, we end up with territories of exactly the same size. Further if we consider these territories and the splitted basic areas as nodes of a graph, having edges between $v \in V$ and B if a part of v belongs to B, we yield a split adjacency similar to the one in section 5.2.

The important observation here is that this split adjacency is again cycle-free. The reason is the recursive partitioning of V in the search. Therefore all the results of section 5.3 on optimal split resolution apply here also. Particularly the worst case bound $\max\{w_v : v \in V^S\}$ for the maximum deviation of district sizes from average in terms of the maximum size of the splits holds.

Note that this modification of the successive dichotomies heuristic prevents also the necessity of backtracking. By combining successive dichotomies with optimal split resolution we yield a procedure for solving the territory design problem that is very fast and for which the balance of the resulting territories can be estimated in advance.

7 Computational results

In the following, we will present results indicating the computational efficiency of the successive dichotomies heuristic. They give a good idea of the performance of the method, both in running times and solution quality, and indicate its suitability for the use in an interactive environment. Actually, the results presented here are just an extract of our tests. They

are representative for the typical behavior of the algorithms and the conclusions that we draw are in fact based on many computational tests of the algorithms.

In the following, we will compare the successive dichotomies heuristic, named *Dicho*, with two location–allocation based methods. The first one, called *Inter*, employs the AllocMinDist method in the allocation phase and a local search technique based on Teitz and Bart's interchange method (Teitz and Bart (1968)) in the location part. The second heuristic, called *Split*, uses (TRANSP) in the allocation phase and resolves split areas using the AssignMAX method. The location phase utilizes a Lagrangean relaxation method.

The three heuristics were tested on problems of different sizes in terms of the numbers of basic areas, starting with 100 up to 1000. For each number of basic areas several problem instances were generated. Each instance was solved with varying numbers of territories and different activity measures. The instances were created using real–world data obtained from the GIS *ArcView*. Basic areas correspond to German zip–code areas and the activity measures are different demographic figures. Every test problem was solved with each of the three heuristics *Inter*, *Split* and *Dicho* and the running time of the respective method and the quality of the resulting territories in terms of maximal relative percentage deviation of a territory from the average were obtained. For each heuristic these two values were then averaged over all problem instances with the same number of basic areas. The results are depicted in Figures 10 and 11.

Comparing the two location–allocation methods one can easily see that the better solution quality in terms of maximal deviation of the *Split* heuristic is traded off against a considerably larger solution time compared to the *Inter* heuristic. Moreover the *Dicho* algorithm outperforms in average the other two heuristics with respect to running time and solution quality for almost all problem sizes. This underlines the quality and speed of the successive dichotomies heuristic and stresses its suitability for the use in an interactive environment.

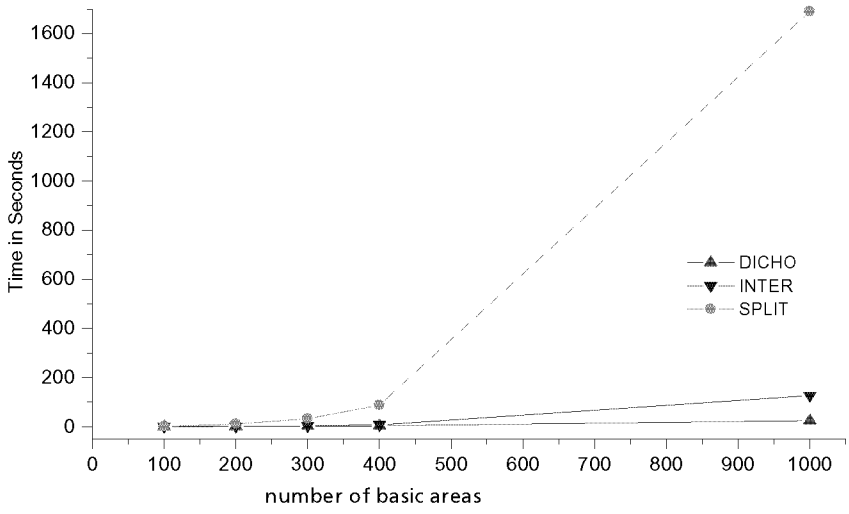

Figure 10: *Computational results for a varying number of basic areas in terms of running times*

8 Extensions and GIS–Integration

The heuristics presented in Sections 5 and 6 can be extended to take additional planning criteria or problem characteristics into account. For example

Several activity measures The heuristics can be extended to handle several activity measures. In this case while examining a partition (see section 6.2) one has to take all activity measures into account when determining the best balanced partition(s).

Prescribed and forbidden territory centers Another extension is the consideration of prescribed and forbidden territory centers. That means we are already given some fixed territory centers at the beginning which have to be taken into account or, the other way around, some basic areas are not allowed to be selected as district centers.

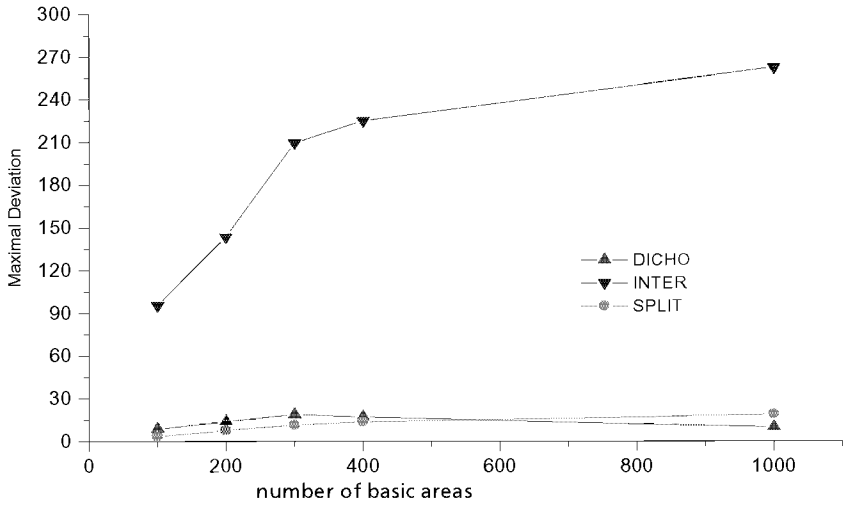

Figure 11: *Computational results for a varying number of basic areas in terms of solution quality.*

Prescribed territories In case some districts are already given at the beginning of the planning process the methods can be adapted to take the already existing territories into account and possibly add additional *BA*s to them.

Connectedness Finally, if neighborhood information about the basic areas is given, i.e. we know if two basic areas are neighboring or not, then the heuristics can be extended in such a way that they always yield connected territories (if possible).

Although the above mentioned extensions can be applied to location–allocation heuristics as well as to the successive dichotomies method one has to note that the latter heuristic is far more flexible in terms of incorporating extensions which do not rely on linear relationships, for example complex measures for compactness.

Moreover the heuristics can be incorporated into a larger framework in order to apply them to different practical planning problems, as outlined

in the introduction. For example scenarios where not all basic areas have to be (or can be) partitioned into territories due to budget constraints. Or applications where a limit on the maximal allowed geographic extend of the territories has to be taken into account.

In addition, the number p of territories need not be fixed in advance. Instead, the algorithm will choose the appropriate number of districts in such a way that the planning criteria are best fulfilled. For example, partition the basic areas in the region under consideration into as few as possible territories such that the size of all territories is below a certain maximal bound.

Integration into GIS

Enhanced with these extensions the successive dichotomies heuristic is the algorithmic base of a commercial software product for geo-marketing called *BusinessManager*. The BusinessManager is an extension of ESRI's ArcView GIS and has been developed by *geomer GmbH* together with *Fraunhofer ITWM*.

In the BusinessManager the user can solve general territory design problems. The interaction is integrated with the GIS so the user can access data from arbitrary shape files. The basic areas can be defined by points, lines, polygons etc, depending of the planning problem under consideration. Figure 12 shows a screenshot of the BusinessManager software.

The user benefits from this integration of optimization algorithms into a GIS in several ways. Firstly, GIS are a common tool in geo–marketing and the user has access to all GIS functionality to work on his planning problem. Secondly the seamless integration of territory planning heuristics allows the user to access these methods without being an expert in Operations Research. After the computations performed by the heuristics in the background are finished an immediate visualization of the results in the GIS allows the user to examine the proposed solution. Then he has the option to manually adjust the solution or to change the planning parameters and start a new run of the optimization engine. It is this interactive type of work with the heuristics that requires the fast generation of solutions, often mentioned in this paper.

The technical side of the integration of the heuristics into the GIS is

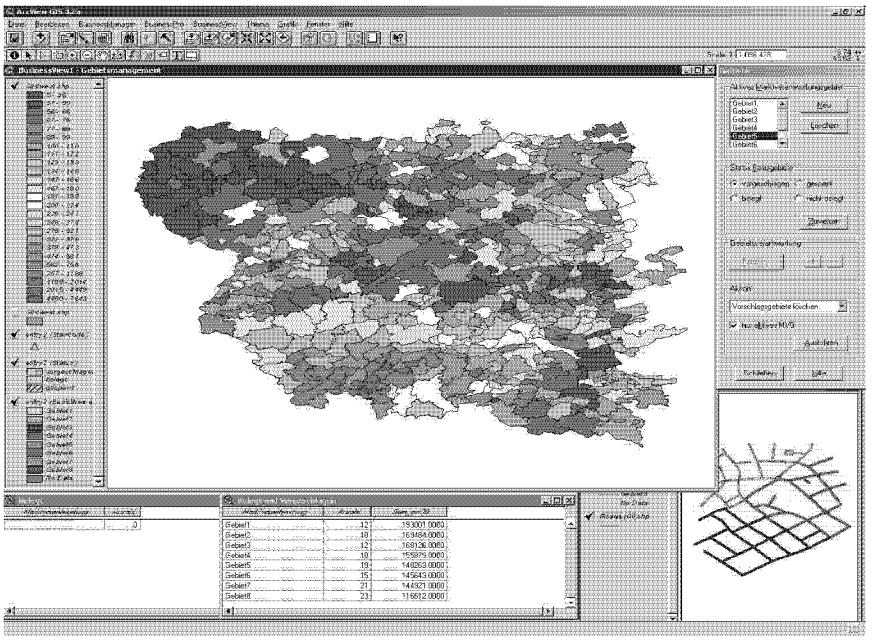

Figure 12: *Screenshot of the BusinessManager software.*

sketched in Figure 13. While the user interaction and data management is all done within the GIS there is the optimization engine as an underlying part. We found it useful to distinguish two layers in the optimization engine. The lower layer contains the implementation of the heuristics. Since generally the planning problem specified by the user in the GIS can not be mapped in a direct way to one of the heuristics (or there had to be a quite large number of them), an intermediate layer contains the so-called *scenario manager*. This layer selects and combines the algorithms in the heuristics layer that are suited to produce an answer to the user's planning problem. As an example the scenario layer calls the successive dichotomies heuristic repeatedly with varying p if the number of territories is asked as part of the planning result by the user.

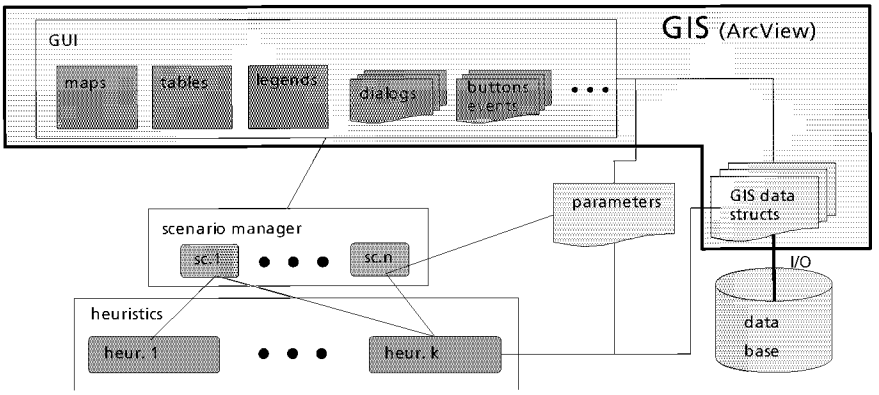

Figure 13: *Integration of the heuristics into ArcView GIS.*

9 Conclusions

Problems of territory alignment arise in different application areas. Our main purpose in this paper was to show how the planner can be adequately supported to solve such problems in practice and the role of operations research in this process.

From the many proposed algorithmic ideas how to solve territory design problems, two heuristic approaches are discussed more in-depth in this paper. The first one is based on the location–allocation principle. In the allocation phase we contribute a method to solve the split resolution problem in an optimal fashion. However, even if the instances of the transportation problem in the allocation phase can be solved efficiently, we found that on large-scale problems the running times were still to high.

Therefore we detail out a heuristic approach, so far only sketched in the literature, that solves the territory design problem geometrically. This successive dichotomies heuristic proves to be very fast and flexible. All the necessary requirements for the heuristic are basic procedures in computational geometry: polygon–line intersection, convex hull, coordinate transformation (rotation) etc. These are computationally fast and easy to implement. Especially in contrast to most of the classical methods there is no need for specialized solvers or methods for mathematical programming problems like for the transportation problem. Despite its simplicity the

successive dichotomies heuristic provides surprisingly good results.

We further described how the heuristics for territory design have been integrated into a commercial software for geo–marketing. This allows planners to use these heuristics for various planning problems in territory design without being operations research experts.

On the other hand, as indicated in Section 8, there is still a lot of work to do for operations researchers in the area of territory design, both on the theoretical and the practical side.

References

Andria F., Chiancone P. and Piraino S. (1979). Die Anwendung der Dynamischen Optimierung bei der Sozial–Sanitären Bezirkseinteilung. *Zeitschrift für Operations Research* B 23, 33–43.

Baker J.R., Clayton E.R. and Moore L.J. (1989). Redesign of Primary Response Areas for County Ambulance Services. *European Journal of Operational Research* 41, 23–32.

Bergey P.K., Ragsdale C.T. and Hoskote M. (2003). A Simulated Annealing Genetic Algortihm for the Electrical Power Districting Problem. *Annals of Operations Research* 121, 33–55.

Blais M., Lapierre S.D. and Laporte G. (2003). Solving a Home–Care Districting Problem in an Urban Setting. *Journal of the Operational Resesearch Society* 54, 1141–1147.

Bodin L.D. (1973). A Districting Experiment with a Clustering Algorithm. *Annals of the New York Academy of Sciences* 19, 209–214.

Bourjolly J.-M., Laporte G. and Rousseau J.-M. (1981). Découpage Électoral Automatisé: Application à l'Île de Montréal. *INFOR* 19, 113–124.

Bozkaya B., Erkut E. and Laporte G. (2003). A Tabu Search Heuristic and Adaptive Memory Procedure for Political Districting. *European Journal of Operational Research* 144, 12–26.

Bozkaya B., Erkut E., Laporte G. and Neuman S. (2005). *Political Districting: Solving a Multi-Objective Problem Using Tabu Search.* Kluwer.

Browdy M.H. (1990). Simulated Annealing: An Improved Computer Model for Political Redistricting. *Yale Law and Policy Review* 8, 163–179.

Cirincione C., Darling T.A. and O'Rourke T.G. (2000). Assessing South Carolina's 1990s Congressional Districting. *Political Geography* 19, 189–211.

Chance C.W. (1965). Political Studies: Number 2 – Represetation and Reappointment. Department of Political Science, Ohio State University, Columbus.

Cloonan J.B. (1975). A Note on the Compactness of Sales Territories. *Management Science* 19, 469.

Deckro R.F. (1977). Multiple Objective Districting: A General Heuristic Approach Using Multiple Criteria. *Operational Research Quarterly* 28, 953–961.

Drexl A. and Haase K. (1999). Fast Approximation Methods for Sales Force Deployment. *Management Science* 45:1307–1323, 1999.

D'Amico S.J., Wang S.-J., Batta R. and Rump C.M. (2002). A Simulated Annealing Approach to Police District Design. *Computers and Operations Research* 29, 667–684.

Easingwood C. (1973). A Heuristic Approach to Selecting Sales Regions and Territories. *Operational Research Quarterly* 24, 527–534.

Ferland J.A. and Guénette G. (1990). Decision Support System for a School Districting Problem. *Operations Research* 38, 15–21.

Fleischmann B. and Paraschis J.N. (1988). Solving a Large Scale Districting Problem: A Case Report. *Computers and Operations Research* 15, 521–533.

Forman S.L. and Yue Y. (2003). Congressional Districting Using a TSP–Based Genetic Algorithm. In: Cantu-Paz E., Foster J.A., Deb K., David L., Rajkumar R. (eds.), *Genetic and Evolutionary Computation–GECCO 2003. Proceedings*, Leture Notes in Computer Science 2723. Springer Verlag, 2072–2083.

Forrest E. (1964). Apportionment by Computer. *American Behavioral Scientist* 23, 23–35.

Garfinkel R.S. (1968). *Optimal Political Districting.* PhD thesis, The Johns Hopkins University, 1968. Also as working paper # 6812, College of Business Administration. Univversity of Rochester.

Garfinkel R.S. and Nemhauser G.L. (1970). Optimal Political Districting by Implicit Enumeration Techniques. *Management Science* 16, 495–508.

George J.A., Lamar B.W. and Wallace C.A. (1997). Political District Determination Using Large-Scale Network Optimization. *Socio-Economic Planning Sciences* 31, 11–28.

Glaze T.A. and Weinberg C.B. (1979). A Sales Territory Alignment Program and Account Planning System. In: Bagozzi R. (ed.), *Sales Management: New Developments from Behavioral and Decision Model Research.* Marketing Science Institute, Cambridge, 325–343.

Grilli di Cortona P., Manzi C., Pennisi A., Ricca F. and Simeone B. (1999). *Evaluation and Optimization of Electoral Systems.* SIAM Monographs on Discrete

Mathematics and Applications.

Hanafi S., Freville A. and Vaca P. (1999). Municipal Solid Waste Collection: An Effective Data Structure for Solving the Sectorization Problem with Local Search Methods. *INFOR* 37, 236–254.

Helbig R.E., Orr P.K. and Roediger R.R. (1972). Political Redistricting by Computer. *Communications of the ACM* 15, 735–741.

Hess S.W. and Samuels S.A. (1971). Experiences with a Sales Districting Model: Criteria and Implementation. *Management Science* 18, 41–54.

Hess S.W., Weaver J.B., Siegfeldt H.J., Whelan J.N. and Zitlau P.A. (1965). Nonpartisan Political Redistricting by Computer. *Operations Research* 13, 998–1008.

Hojati M. (1996). Optimal Political Districting. *Computers and Operations Research* 23, 1147–1161.

Horn D.L., Hampton C.R. and Vandenberg A.J. (1993). Practical Application of District Compactness. *Political Geography* 12, 103–120.

Howick R.S. and Pidd M. (1990). Sales Force Deployment Models. *European Journal of Operational Research* 48, 295–310.

Kalcsics J., Melo T., Nickel S. and Gündra H. (2001). Planning Sales Territories - a Facility Location Approach. Operations Research Proceedings 2001. Springer Verlag, 141–148.

Lewyn M.E. (1993). How to Limit Gerrymandering. *Florida Law Review* 45, 403–486.

Lodish L.M. (1975). Sales Territory Alignment to Maximize Profit. *Journal of Marketing Research* 12, 30–36.

Macmillan W. and Pierce T. (1992). Optimization Modelling in a GIS Framework: The Problem of Political Districting. Specialist meeting, April 16-18, 1992. National Center for Geographic Information and Analysis.

Marlin P.G. (1981). Application of the Transportation Model to a Large–Scale "Districting" Problem. *Computers and Operations Research* 8, 83–96.

Mehrotra A., Johnson E.L. and Nemhauser G.L. (1998). An Optimization Based Heuristic for Political Districting. *Management Science* 44, 1100–1114.

Minciardi R., Puliafito P.P. and Zoppoli R. (1981). A Districting Procedure for Social Organizations. *European Journal of Operational Research* 8, 47–57.

Muyldermans L., Cattrysse D., van Oudheusden D. and Lotan T. (2002). Districting for Salt Spreading Operations. *European Journal of Operational Research* 139, 521–532.

Niemi R.G., Grofman B., Carlucci C. and Hofeller T. (1990). Measuring Compactness and the Role of a Compactness Standard in a Test for Partisan and Racial Gerrymandering. *Journal of Politics* 52, 1155–1181.

Nygreen B. (1988). European Assembly Constituencies for Wales: Comparing of Methods for Solving a Political Districting Problem. *Mathematical Programing* 42, 159–169.

Parker F.R. (1990). *Black Votes Count.* Chapel Hill: The University of North Carolina Press.

Palermo P.C., De Giorgi C. and Tagliabue G. (1977). An Interactive Approach to the Facility Location Districting Problem. *Adv. Operations Research*, 341–346.

Ricca F. (1996). Algorthmi di Ricerca Locale per la Distrettizzazione Elettorale. *Atti Giorante AIRO* 634–637.

Ricca F. and Simeone B. (1997). Political Districting: Traps, Criteria, Algorithms and Tradeoffs. *Ricerca Operativa AIRO* 27, 81–119.

Ronan R.(1983). Sales Territory Alginment for Sparse Accounts. *OMEGA The International Journal of Management Science* 11, 501–505.

Schröder M. (2001). *Gebiete Optimal Aufteilen.* Ph.D. Thesis, University of Karlsruhe, 2001. `http://www.ubka.uni-karlsruhe.de/eva`.

Segal M. and Weinberger D.B. (1977). Turfing. *Operations Research* 25, 367–386.

Simchi-Levi D., Kaminsky P. and Simchi-Levi E. (2003). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies.* McGraw–Hill/Irwin.

Shanker R.J., Turner R.E. and Zoltners A.A. (1975). Sales Territory Design: An Integrated Approach. *Management Science* 22, 309–320.

Skiera B. (1997). Wieviel Deckungsbeitrag Verschenkt Man Durch Eine Gleichartige Einteilung der Verkaufsgebiete? *Zeitschrift für Betriebswirtschaftliche Forschung* 49, 723–746.

Skiera B. and Albers S. (1994). Costa: Ein Entscheidungs-Unterstützungs-System zur deckungsbeitragsmaximalen Einteilung von Verkaufsgebieten. *Zeitschrift für Betriebswirtschaft* 64, 1261–1283.

Teitz M.B. and Bart P. (1968). Heuristic Methods for Estimating Generalized Vertex Median of a Weighted Graph. *Operations Research* 16, 955–961.

Williams J.C. Jr. (1995). Political Redistricting: A Review. *Papers in Regional Science* 74, 13–40.

Zoltners A.A. (1979). A Unified Approach to Sales Territory Alignment. In: Bagozzi R. (ed.), *Sales Management: New Developments from Behavioral and Decision Model Research.* Marketing Science Institute, 360–376.

Zoltners A.A. and Sinha P. (1983). Sales Territory Alignment: A Review and Model. *Management Science* 29, 1237–1256.

DISCUSSION

B. Bozkaya
University of Alberta, Canada

This article deals with the general territory design (or districting) problem, and provides a "unified" (as the authors call it) modeling approach. Overall, I think the article is quite readable and fairly easy to follow. It has a review article flavor, but it also includes authors' contributions to the solution of the districting problem.

There are two main aspects of this article: the first one is the authors' efforts to review and bring together models from various districting literature. Examples include political districting, sales and service territory design, and school districting. The second aspect is the authors' contribution in terms of solution methodology. My review regarding these two aspects plus other issues is provided below.

Model

The first part of the article covers in detail the existing literature on territory design, including models and solution techniques. The authors' main motivation is to identify main elements of these models and turn them into a "unified" model that also performs well (computationally) in a GIS environment. While this is an effort well appreciated, there are two problems with this approach. For one thing (and based on my industry experience), often the applied models have many specific elements that solving a simpler model fast may not have too much of a practical value. If we consider the modeling elements included in the authors' model (balance, compactness, contiguity - though not explicitly enforced -, and non-overlapping districts), it is easy to see that the authors have only taken an "intersection" of the many models in the literature, and not quite have "unified" them. Plus