Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 438–447

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Data mining and visualization for decision support and modeling of public health-care resources

Nada Lavrač [a,b,*], Marko Bohanec [a], Aleksander Pur [c], Bojan Cestnik [a,d], Marko Debeljak [a], Andrej Kobler [e]

[a] *Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia*
[b] *University of Nova Gorica, Nova Gorica, Slovenia*
[c] *Ministry of Interior Affairs, Štefanova 2, Ljubljana, Slovenia*
[d] *Temida, d.o.o. Ljubljana, Slovenia*
[e] *Slovenian Forestry Institute, Ljubljana, Slovenia*

Received 1 July 2006
Available online 10 November 2006

Abstract

This paper proposes an innovative use of data mining and visualization techniques for decision support in planning and regional-level management of Slovenian public health-care. Data mining and statistical techniques were used to analyze databases collected by a regional Public Heath Institute. We also studied organizational aspects of public health resources in the selected Celje region with the objective to identify the areas that are atypical in terms of availability and accessibility of public health services for the population. The most important step was the detection of outliers and the analysis of availability and accessibility deviations. The results are applicable to health-care planning and support in decision making by local and regional health-care authorities. In addition to the practical results, which are directly useful for decision making in planning of the regional health-care system, the main methodological contribution of the paper are the developed visualization methods that can be used to facilitate knowledge management and decision making processes.

Keywords: Data mining; Decision support; Knowledge discovery; Knowledge management; Visualization; Applications to health-care

1. Introduction

Effective medical prevention and good access to health-care resources are important factors that affect citizens' welfare and quality of life. As such, these are important factors in strategic planning at the national level, as well as in planning at the regional and local community levels. Large quantities of data collected by medical institutions and governmental public health institutions can serve as a valuable source of evidence that needs to be taken into account when making decisions about priorities to be included in regional strategic health-care plans.

Slovenian regional public health institutes (PHIs), coordinated by the national Institute of Public Health (IPH), are an important part of the system of public health in Slovenia. Their functions are public health monitoring, organizing public health-related activities and proposing and implementing actions for maintaining and improving public health. PHIs themselves coordinate a regional network of hospitals, clinics, individual health professionals and other health-care resources involved in particular health-care activities. Data at all levels are collected, and a national-level data warehouse is maintained at the national IPH.

[*] Corresponding author. Fax: +386 1 477 3315.
E-mail addresses: nada.lavrac@ijs.si (N. Lavrač), marko.bohanec@ijs.si (M. Bohanec).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.10.003

This paper describes an application of data mining and decision support in public health-care, carried out in Slovenia within a project called MediMap. The goal of MediMap was to improve health-care knowledge management through data mining and decision support integration [3]. *Data mining* [4] is concerned with finding interesting patterns in data. Data mining includes predictive data mining algorithms, which result in models that can be used for prediction and classification, and descriptive data mining algorithms for finding interesting patterns in the data, like associations, clusters and subgroups. Data mining is typically applied to knowledge discovery in large and complex databases and has been extensively used in knowledge management [1] and industrial and business problem solving [2]. On the other hand, decision support [5,6] is concerned with helping decision makers solve problems and make decisions. As indicated by the results of recent research [3], data mining and decision support integration can lead to improved solutions in practical applications.

Health-care is a knowledge-intensive domain, in which neither data gathering nor data analysis can be successful without using knowledge about both the problem domain and the data analysis process. This indicates the usefulness of integrating data mining with decision support techniques [3,5] to promote the construction of effective decision criteria and decision models supporting decision making and planning in public health-care. The integration of the data mining and decision support approaches, as well as the novel visualization techniques developed for the purpose of this health-care application, have facilitated knowledge management and improved decision support.

In MediMap, we mainly used descriptive data mining methods and combined them with visualization and multi-criteria decision support techniques to improve the management of data and knowledge at the Public Health Institute of the Celje region. The main objective of MediMap was to set up appropriate models and tools to support decisions concerning regional health-care, aimed to serve also as a reference model for other regional PHIs. We approached this goal in two phases: first, we analyzed the available data with data mining techniques, and second, we used the results of data mining for a more elaborate study using decision support techniques. In the first phase we focused on the problem of directing the patients from primary health-care centers to specialists. In the second phase we studied organizational aspects of public health resources in the Celje region with the goal to identify the areas that are atypical in terms of availability and accessibility of public health services.

The paper is organized as follows. Section 2 presents data mining and decision support used as the main technologies used for knowledge management in this application. Section 3 presents the data that was used for the analysis of the Celje health-care resources. The results of applying data mining and decision support techniques, and the visualization of the results, are presented in Section 4. Section 5 concludes by summarizing the main results and by presenting plans for further work.

2. Data mining and decision support for knowledge management

Data mining [3,4] is concerned with finding models and patterns from the available data. Data mining includes predictive data mining algorithms, which result in models that can be used for prediction and classification, and descriptive data mining algorithms for finding interesting patterns in the data, like associations, clusters and subgroups.

Decision support [3,5] is concerned with helping decision makers solve problems and make decisions. Decision support provides a variety of data analysis, preference modeling, simulation, visualization and interactive techniques, and tools such as decision support systems, multiple-criteria modeling, group decision support and mediation systems, expert systems, databases and data warehouses. Decision support systems incorporate both data and models.

Data mining and decision support can be integrated to better solve data analysis and decision support problems. In *knowledge management* [1], such integration is interesting for several reasons. For example, in data mining it is often unclear which algorithm is best suited for the problem. Here, we require some decision support for data mining. Another example is when there is a lack of data for the analysis. To ensure that appropriate data is recorded when the collection process begins it is useful to first build a decision model and use it as a basis for defining the attributes that will describe the data. These two examples show that data mining and decision support can complement each other, to achieve better results. Different aspects of data mining and decision support integration have been investigated in [3].

3. Public health data

To model the Celje regional health-care system, we first wanted to better understand the health-care resources and their connections in the Celje region. The location of this region on the map of Slovenia is shown in Fig. 1. The Celje region is composed of 11 communities, further divided into 34 local communities.

For the purpose of MediMap, data mining techniques were applied to the data of 11 community health centers (CHCs) of the Celje region. The dataset consisted of three databases:

- The health-care providers database,
- The out-patient health-care statistics database (patients' visits to general practitioners and specialists, diseases, human resources and availability), and
- The medical status database.

To model the processes of a particular CHC (the patient flow), we used additional data describing the directing of patients to other CHCs or specialists.

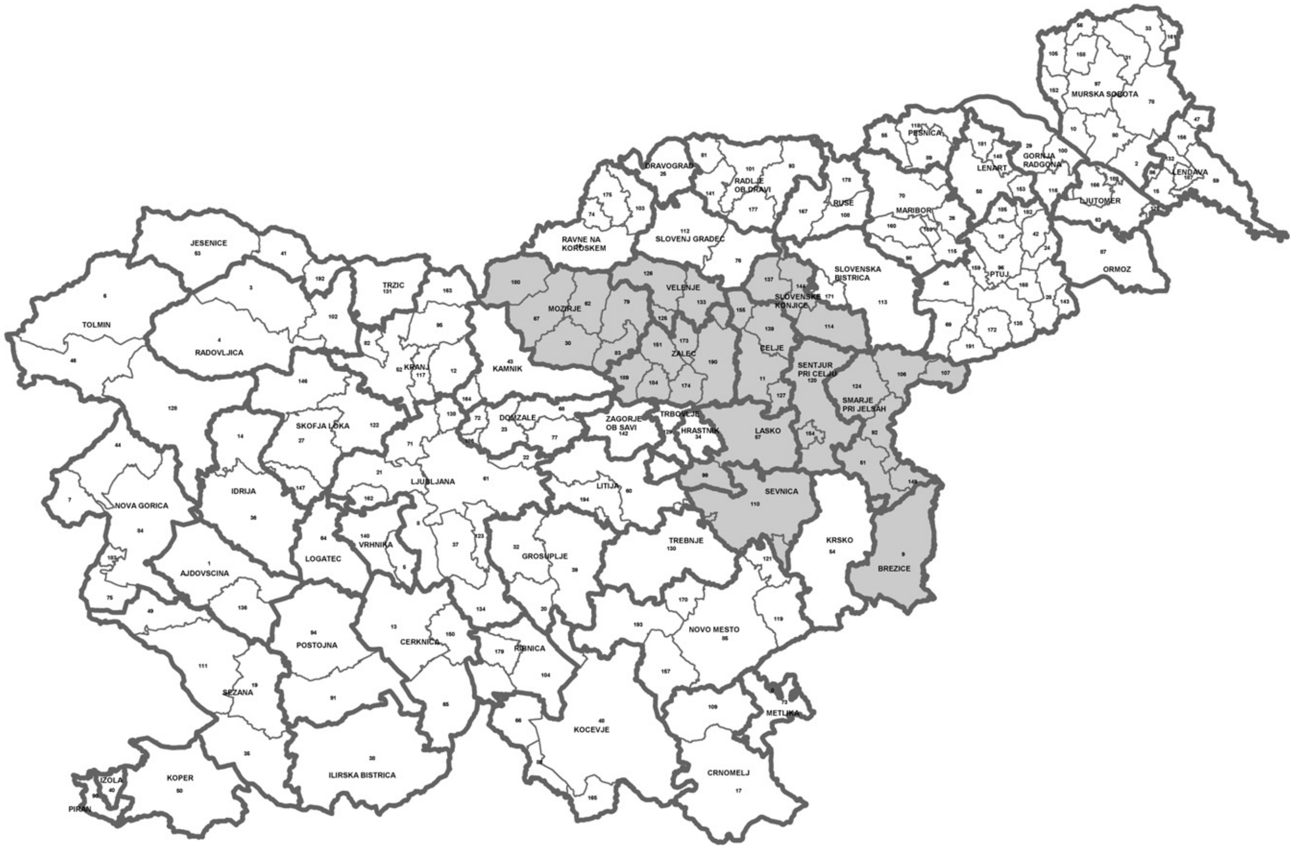


Fig. 1. The Celje region, located on the map of Slovenia.

4. Results of analyses

This section presents the detected similarities of community health centers of the Celje region, the analysis of the availability and accessibility of various public health-care resources, as well as the achieved results of decision support and visualization allowing for more advanced planning of health-care resources.

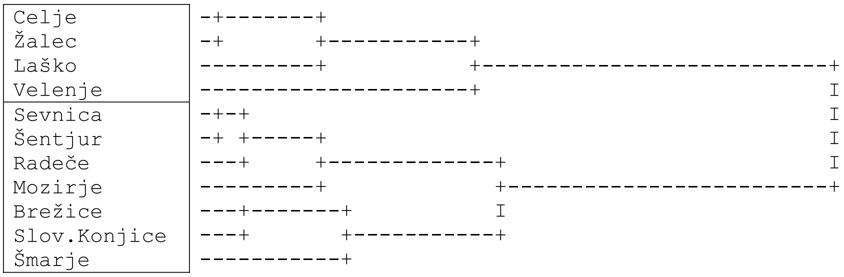
4.1. Detecting similarities of community health centers with data mining

The goals of this analysis were to detect the similarities between CHCs, and to detect the atypical CHCs. Similarities between CHCs were analyzed according to four different categories: (a) patients' age categories, (b) patients' social categories, (c) organization of the CHC and (d) employment structure of the CHC (Table 1). The categories (a)–(c) are described by five attributes and (d) is described by four attributes. The attributes of categories (a) and (b) are numeric and represent relative frequencies (e.g., value x of the attribute pre-school means that in a given CHC x% of patients are pre-school children).

For each category, similarity groups were constructed using four different clustering methods: agglomerative classification [7], principal component analysis [7], the Kolmogorov–Smirnov test [8], as well as the quantile range test and polar ordination [9]. An illustration of clusters, generated by Ward's agglomerative hierarchical clustering

Table 1
Description of categories and attributes used in analyzing the similarities between CHCs

	Categories	Attributes				
a	Patients' age	0–6 (pre-school)	7–19 (school)	20–49	50–64	⩾65
b	Patients' social status	Blue-collar workers	Farmers	Pensioners	Unclassified	Other
c	Organization of a CHC	Years of operation	Contacts per hour	Contacts per practice	Number of surgeries	Contacts per employee
d	Employment structure of a CHC	Education level	Time since professional exam	Time since first employment	Average age of employee	



```
Celje         -+-------+
Žalec         -+       +-----------+
Laško         ---------+           +--------------------------+
Velenje       ---------------------+                          I
Sevnica       -+-+                                            I
Šentjur       -+ +-----+                                      I
Radeče        ---+     +-------------+                        I
Mozirje       ---------+             +------------------------+
Brežice       ---+-------+           I
Slov.Konjice  ---+       +-----------+
Šmarje        -----------+
```

Fig. 2. Results of hierarchical clustering of CHCs.

using the Euclidean distance measure, is given in Fig. 2. According to the maximal inter-cluster dissimilarity, the methods splits the CHCs into two top-level clusters, Cluster 1 formed of upper four CHCs and Cluster 2 of the bottom seven CHCs.

The similarities of community health centers were presented and evaluated by PHI Celje domain experts. In several cases the results confirmed already known similarities, while the experts could not find obvious explanations of the results of clustering. To explain the main differences between clusters, we have transformed the result of clustering into a classification task, and used a decision tree learning algorithm (J48 WEKA implementation of the well-known C4.5 learner [10]) to get a decision tree distinguishing the two classes (the two top-level clusters). To illustrate the approach, take the two top-level clusters of Fig. 2, considered as two disjoint classes. Fig. 3 shows a decision tree in which only the most informative attribute, distinguishing between the two groups of health centers, is an attribute of category (a): the age of patients. Community health centers in which pre-school children (PreSc) constitute more than 1.41% of all visits to the center form Cluster 1 (consisting of seven health centers). The experts' explanation is that these centers lack specialized pediatrician services, hence pre-school children are frequently treated by general practitioners. This is undesirable and indicates the need for corrective health-care management decisions. Despite the simplicity of the presented result achieved, and the approach taken, the combination of clustering and decision tree learning turned out to be useful for achieving a better explanation of the results achieved, which were satisfactory to the health-care experts.

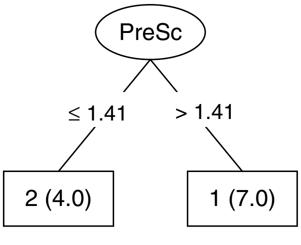


Fig. 3. A decision tree representation of the two clusters from Fig. 2, offering an explanation for the grouping into two classes (class 1 consisting of seven CHCs and class 2 consisting of four CHCs).

Averages over four clustering methods per category were used to further try to detect the similarities between the CHCs of the Celje region (Fig. 4). The results of these experiments confirmed some similarities between community health centers, but the similarity matrix did not provide novel explanations to the experts.

To further analyze the differences between the health centers, we have developed a different visualization method, enabling the analysis of the typicality of CHCs based on the comparison of the estimated number of patients that can be handled by a CHC (its capacity estimated by the number of employed staff) and the actual number of patients handled by the CHC. The outcome, shown in Fig. 5, was very much appreciated by the experts. The figure presents some atypical CHCs, deviating from the diagonal line, such as CHC Brežice and Žalec, which have insufficient staff compared to the number of actual patients requiring health-care services.

In summary, the results of these experiments confirmed some similarities between community health centers and pointed out atypical community health centers together with their properties that required corrective management activities. This part of the analysis, enabling decision support through the visualization of deviating/atypical CHCs, turned out to be most appreciated by the collaborating experts.

4.2. *Availability and accessibility of public health-care resources*

The goal of this analysis was to detect the local communities that are underserved concerning general practice health services—this means that the population in these areas has less than a generally accepted level of services available for the population. We evaluated 34 local communities in the Celje region. The evaluation was based on the ratio of the capacity of health-care services available to patients from the community and the demand for these services by the population of the same area.

For this analysis, the following novel measures and criteria were proposed. In our case, the *capacity* of health-care services is defined as available time of health-care services for patients in the given community, and *demand* means the number of accesses to health-care services from patients

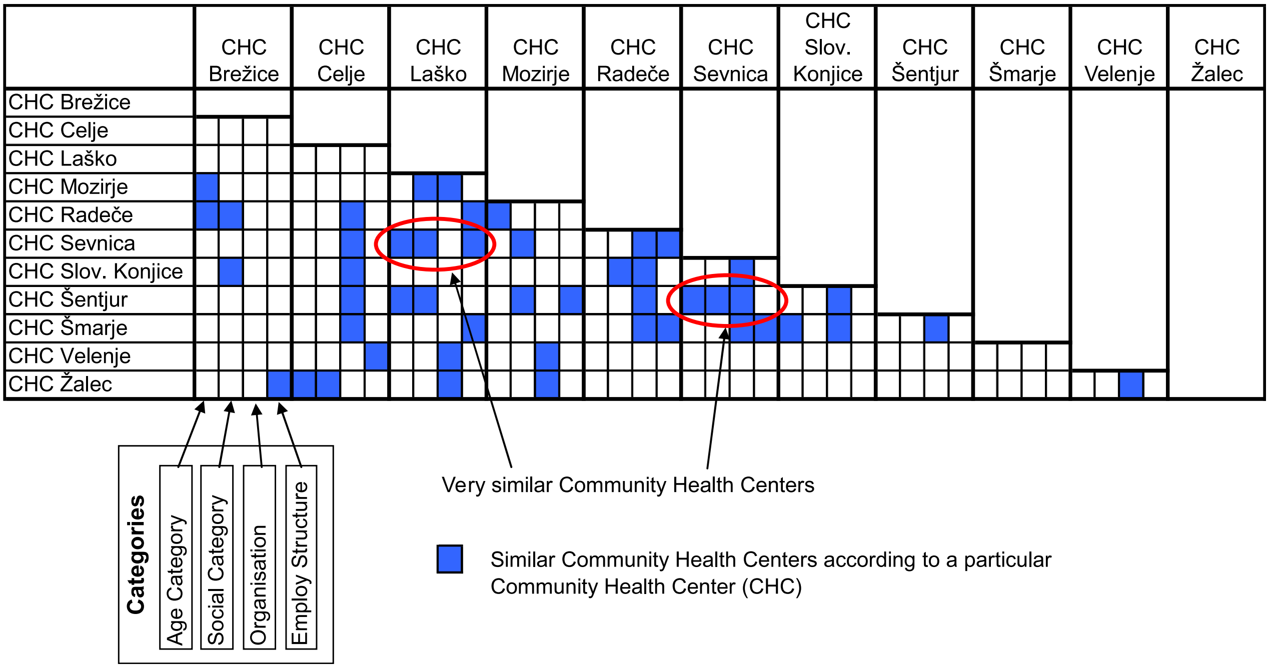


Fig. 4. The similarity matrix of community health centres in the Celje region.

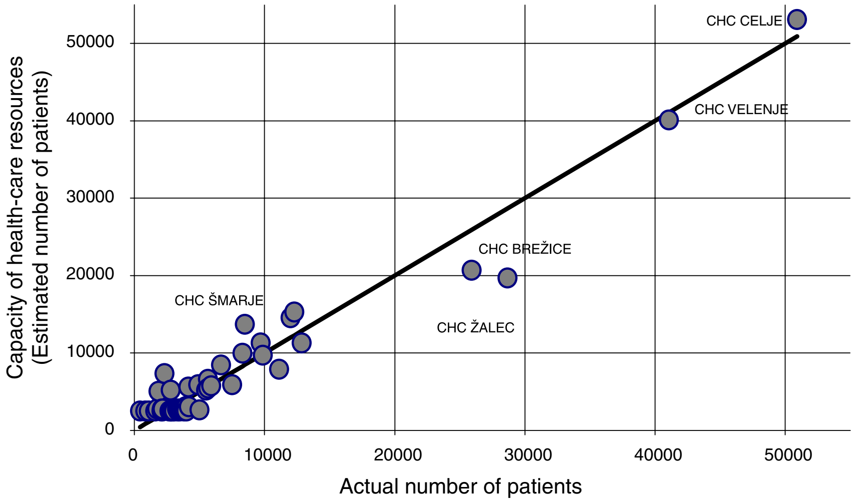


Fig. 5. Detecting atypical Celje region health-care resources (deviating from the diagonal line).

from the community. Therefore, our main criterion for the evaluation of the health-care system for patients in a community is actually the demand/capacity ratio, computed by the average time of available health services per access of a patient from the given community.

In the definition of this measure, called AHSP (availability of health services for patients)

$$\mathrm{AHSP} = \frac{\sum t_i}{p_c} \tag{1}$$

variable t_i denotes the total working time of health-care service i in community c, and p_c the number of accesses to health-care services of patients from community c. By setting the appropriate expert-defined threshold, this measure can turn into a criterion that can be used for decision support.

Notice that the AHSP measure does not take into account that many patients access health services in neighbouring or even more distant communities. Moreover, some of communities do not have their own health-care services at all. Since the migrations of patients into other communities are an important factor, we proposed a novel measure, $\mathrm{AHSP_m}$, which takes migrations into the account. $\mathrm{AHSP_m}$ is an average of available time per access of a patient in all of the health services, depending on the amount of patients from community c that each service received. First, we defined two variables: X_c—the available time per access of a patient from community c and Y—the health service (which can have values 1, 2, 3,...). We got the desired average with the help of the law of total expectation:

$$E(X_c) = \sum_i E(X_c|Y=i)P(Y=i) \tag{2}$$

The term $E(X_c|Y=i)$, which we denote by a_i, is the available time per access of a patient from community c at health-care service i. It can be calculated as the ratio of the total working time of health-care service and the total number of visits. The probability $P(Y=i)$ that the patient visited health-care service i can be stated as the ratio of the number of accesses of patients from community c to health service i (denoted p_{ci}), and the total number of accesses of patients from community c (already defined as p_c). Consequently, we can write the new criterion as

$$\mathrm{AHSP_m} = \sum_i a_i \frac{p_{ci}}{p_c} = \frac{1}{p_c} \sum_i a_i p_{ci} \tag{3}$$

The evaluation of communities in the Celje region using the AHSP and $\mathrm{AHSP_m}$ measures is shown in Figs. 6 and 7, respectively. The color intensity represents the availability of health services for patients: the darker the color, the higher the health-care availability in the community (measured in hours per visit).

Notice the advantage of the modified measure proposed in Eq. (3): the main difference between the evaluations employing the two measures is namely noticeable in communities which do not have own health-care services, like Braslovče, Tabor, Dobje and Solčava. If the migrations of patients to neighbouring communities are not considered, then it looks as if the inhabitants of these communities were without access to health-care resources (Fig. 6). Therefore, $\mathrm{AHSP_m}$ (Fig. 7) is a more realistic evaluation measure and by appropriately setting the threshold values, can be turned into an appropriate criterion to be used by health-care decision makers. Having determined different threshold values, a geographical representation of the results was made possible. This result visualization was very well-accepted by the health-care experts.

To further refine the analysis concerning the availability of health-care services and for the purpose of its visualization (Fig. 8), we introduced two additional measures. The AHS (availability of health services) measure was defined, aimed at measuring the availability of health-care services for the population from a community. More precisely, AHS is defined as the available time of health-care services per population g_c from community c, considering the migrations:

$$\mathrm{AHS} = \frac{1}{g_c} \sum_i a_i p_{ci} \tag{4}$$

The next measure RAHS (rate of accesses to health services), defines the rate of accesses to health-care services for population g_c from community c:

$$\mathrm{RAHS} = \frac{p_c}{g_c} \tag{5}$$

In this case, $\mathrm{AHSP_m}$ is defined as the ratio between the availability of health services for population from the community and the rate of visiting the health services:

$$\mathrm{AHSP_m} = \frac{\mathrm{AHS}}{\mathrm{RAHS}} \tag{6}$$

All these measures, and the derived criteria based on threshold values, give us some very interesting indicators of health conditions and the availability of health-care services in different communities. Using a novel visualization method developed for this purpose, they can be

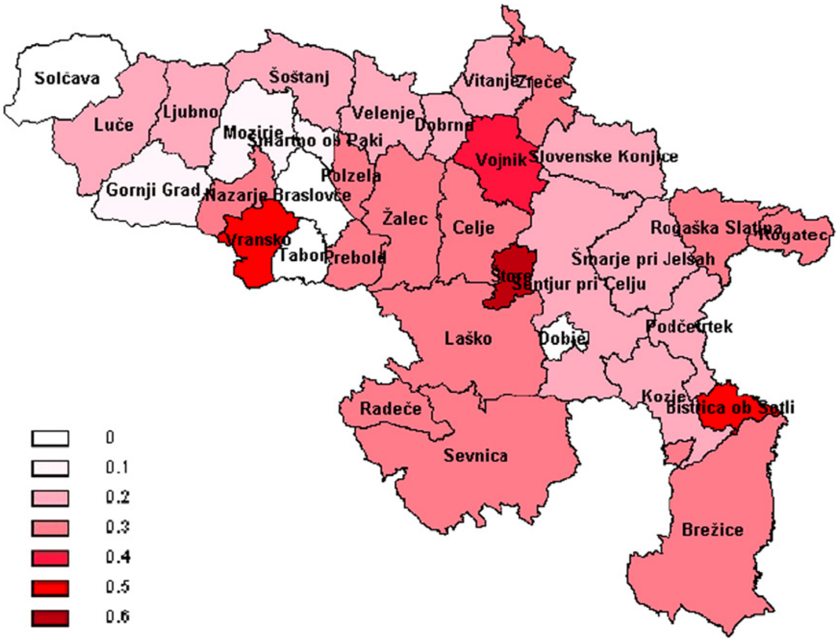


Fig. 6. Availability of health services (AHSP), measured in hours per visit, in the Celje region in 2003.

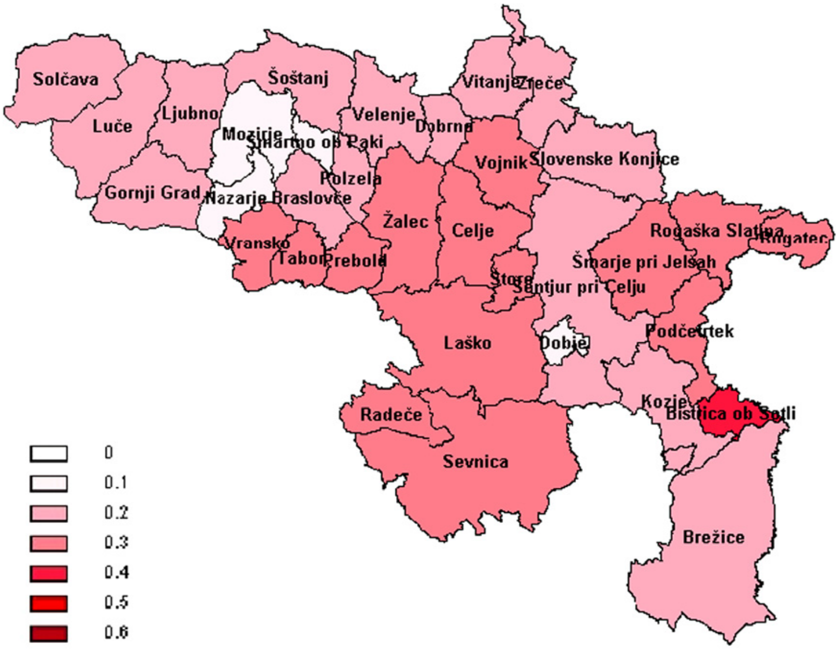


Fig. 7. Availability of health services in Celje in 2003, measured in hours per visit, considering the migrations of patients to neighboring communities ($AHSP_m$).

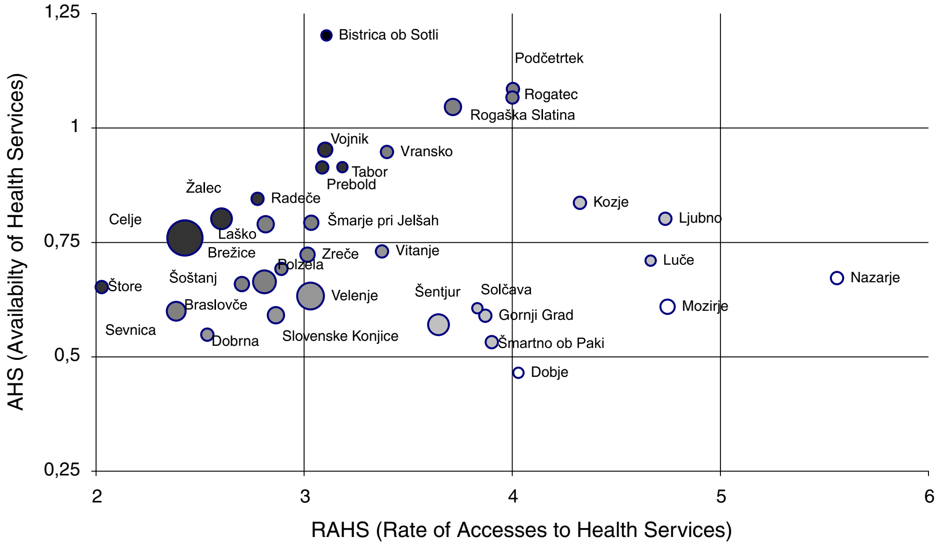


Fig. 8. Available time of health-care services per population by community.

conveniently presented as shown in Fig. 8. Four measurements are actually shown in the chart: RAHS along the horizontal axis, AHS along the vertical axis, $AHSP_m$ as dot color intensity, and population size g_c as dot diameter. Communities with average values of RAHS and AHS appear in the middle of the chart. The outliers represent unusual communities regarding health-care. Communities at the left side of the chart have lower rate of accesses to health services and the ones at the right side have higher access rates. Communities with lower values of AHS are located at the bottom, and those with higher values at the top. The dark-colored communities have higher values of $AHSP_m$ than the light-colored ones.

Consequently, by proposing a novel visualization of this multi-criteria problem, Fig. 8 enables the discovery of implicit and interesting knowledge about health-care services in different communities. For example, the reason for a high value of $AHSP_m$ in communities at the left side of the chart (e.g., Štore) could be the low rate of accesses to the nearest health services, caused by inappropriate medical procedures in these services. A possible reason for the low value of $AHSP_m$ in communities at the right side (Naz-

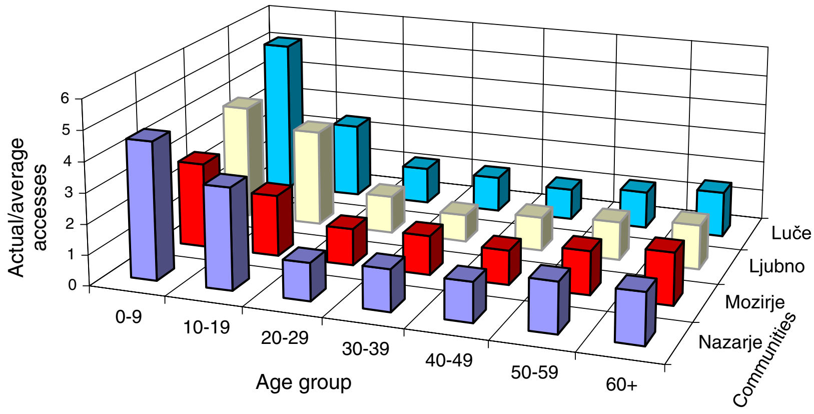


Fig. 9. The ratio between the actual and the average accesses to health services (in year 2003).

arje, Mozirje, Luče in Ljubno) might be high rates of accesses to health services. Further expert analysis was motivated based on this multi-criteria result visualization.

4.3. Decision support for planning health-care resources

Additional analysis of these rates can be provided by a chart shown in Fig. 9. The chart shows the ratio of actual rates of accesses of health services and expected rates for age groups of the population in the communities. This ratio is used in order to simplify the detection of unusual rates of accesses to health services. The expected rate of accesses to health services is the average rate of population in an age group. For example, the access to health services of the population aged between 0 and 9 years is almost five times more frequent than of the population aged between 20 and 29 years. The age group of population from communities is measured along the horizontal axis. Thus, the chart shows that in these communities the rate of accesses to health services of the population under 20 is unusually high. This finding motivated further analysis, which showed that the main reason for the high value of $\mathrm{AHSP_m}$ in these communities is the absence of paediatric services, which was later confirmed by the health-care experts.

A further view on the disparity of health-care in the communities is provided in Fig. 10. There, the evaluation of health services is based on the ratio between the health-care capacity and demand. In our case the demand means the number of accesses to health services, and is measured along the horizontal axis. The capacity is proportional to the working time of health services, and is measured along the vertical axis. Some of the health services are denoted by an identification number and the community name. The regression line represents the expected working times of health services, with respect to the number of accesses. The working times of the health services under the regression line, like Nazarje and Mozirje, are too short, and of those above the regression line are too long. Consequently, this chart can serve for supporting decisions in planning the capacity and working times of health services. Methodologically, the aim of this chart is to highlight the CHCs that lie far away from the regression line rather than to accurately construct the regression line. In our case, we constructed the regression line using all the CHCs, without

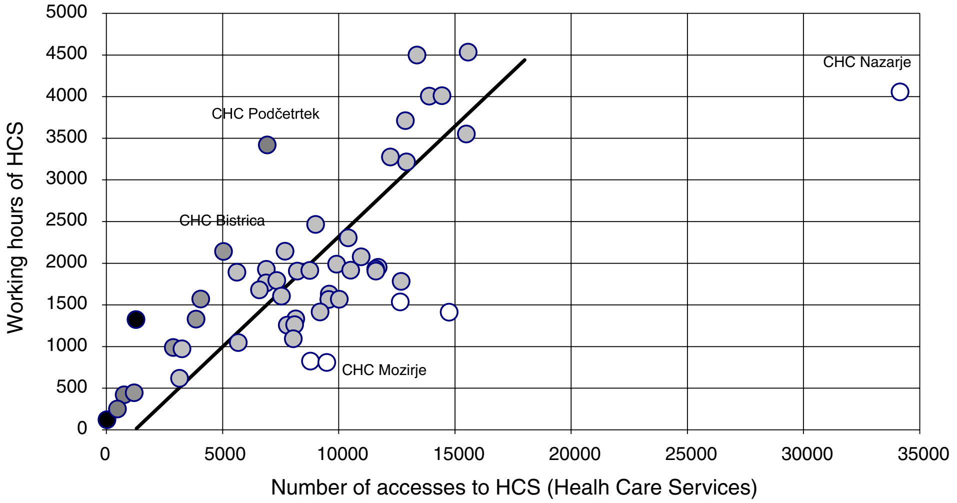


Fig. 10. Evaluation of health services: the ratio between the health-care capacity and the demand.

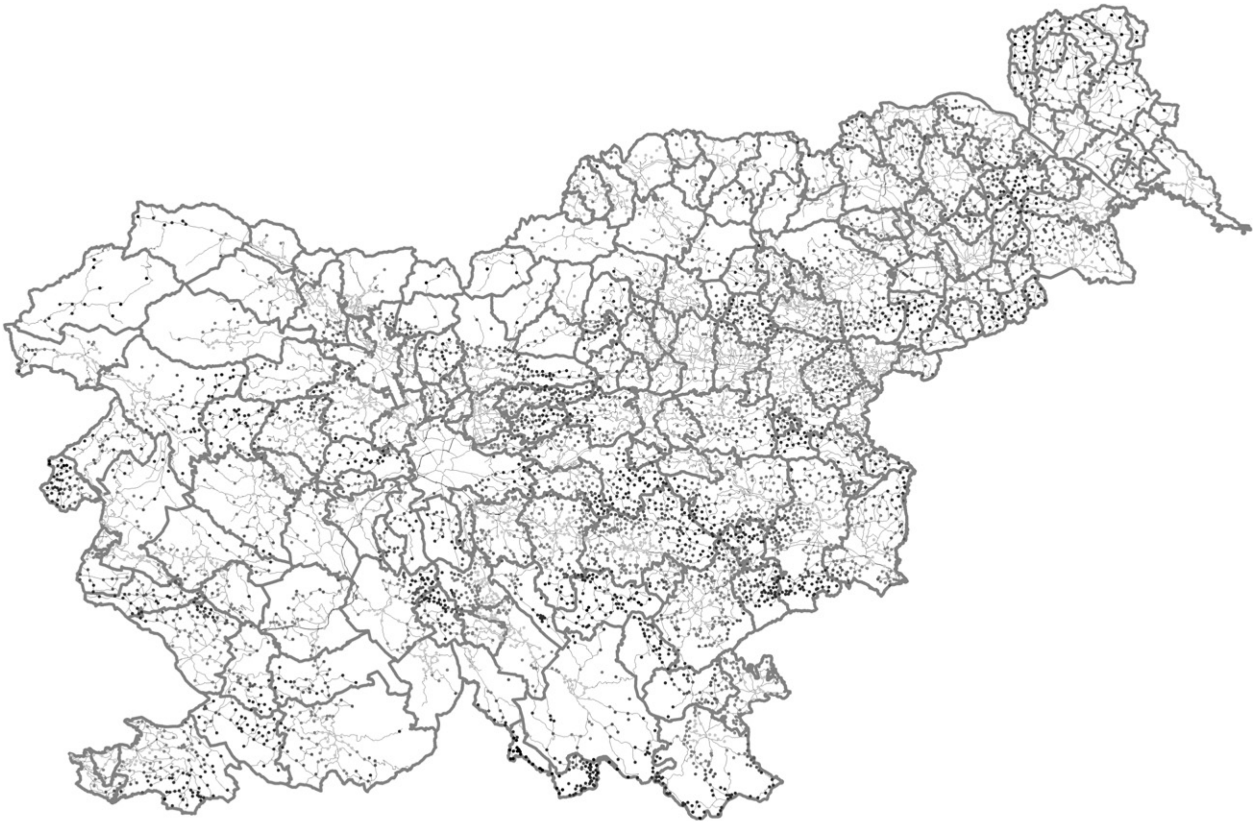


Fig. 11. CHC accessibility map for gynaecology for all the regions of Slovenia.

discarding any outliers as would be the case in the ordinary linear regression.

4.4. Decision support through GIS visualizations

GIS data can be used to visualize the national road network, detailed by road category, and the locations of community health centers of the map of Slovenia. Instead of the raw data visualization, presented by roads leading to the closest CHC for a citizen at a given location, we upgraded the road visualizations by computing the CHC accessibility through the so-called road "resistance" measure, which is anti-proportional to the average travel speed (which is—in turn—proportional to the road category). The following road categories, allowing for different access speeds, were taken into the account: highways (120 km/h), main roads (80 km/h), regional roads (60 km/h) and local roads (50 km/h). This lead to the development of the CHC "access" map, which enables the visualization of areas of Slovenia with low CHC access capacity. Such visualization enables the decision maker to see areas which have low accessibility to primary health services, possibly developing new health-care facilities in such regions. A sample access map for gynaecology for Slovenia is shown in Fig. 11. Each dot represents a settlement (town/village) and its intensity corresponds to the accessibility of the nearest gynaecological health service: the darker the dot, the lower the access capacity.

5. Conclusions

The use of data mining and decision support methods, including novel visualization methods, can lead to better performance in decision making, can improve the effectiveness of developed solutions and enables tackling of new types of problems that have not been addressed before. A real-life application of this approach in public health-care was shown in this paper, following some of the guidelines for public health management recommended in [11,12].

In the MediMap project we have developed methods and tools that can help regional public health institutes (PHIs) and the national Institute of Public Health (IPH) to perform their tasks more effectively. Tools and methods were developed for the reference case of PHI Celje and tested on selected problems related to health-care organization, accessibility of health-care services to the citizens and the health-care providers work. The main achievement was the creation of the model of the availability and accessibility of health services to the population of a given area. With the proposed model it was possible to identify the regions that differ from the average and to consequently explain the causes for such situations, providing many benefits for health-care planning and management processes.

In addition, the national IPH has used the results of this study to identify missing data that should be included in the improved protocol of public health data gathering at the national level, as the study indicated that additional—more detailed, but relatively easy to obtain—data from the community health centres was needed. This finding was valuable for the IPH, as this institution is in charge of defining the national data model and prescribing national data gathering rules and procedures.

In further work, we will extend this analysis to other regions of Slovenia. We will focus on the development of decision support tools for modeling of health-care providers using data mining. We wish to implement the developed methodology so that it can be regularly used for decision support in organizations responsible for the health-care network: the national Ministry of Health, the national IPH, and the regional PHIs.

Acknowledgments

We acknowledge the financial support of the Public Health Institute Celje and the Slovenian Ministry of Higher Education, Science and Technology. We also express our thanks to other members of the MediMap project team Tanja Urbančič, Mitja Jermol, Tadeja Kopač, Branko Kavšek and Primož Lukšič who also contributed to the results described in this paper.

References

[1] Smith RG, Farquhar A. The road ahead for knowledge management: an AI perspective. AI Magazine 2000;21(4):17–40.

[2] Biere M. Business intelligence for the enterprise. Engelwood Cliffs, NJ: Prentice Hall PTR; 2003.

[3] Mladenić D, Lavrač N, Bohanec M, Moyle S, editors. Data mining and decision support: integration and collaboration. Dordrecht: Kluwer; 2003.

[4] Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Los Altos, CA: Morgan Kaufman; 2006.

[5] Mallach EG. Decision support and data warehouse systems. New York: McGraw-Hill; 2000.

[6] Turban E, Aronson JE, Liang TP. Decision support systems and intelligent systems. 7th ed. Englewood Cliffs, NJ: Prentice Hall; 2004.

[7] Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier; 1998. p. 317–341.

[8] Zar JH. Biostatistical analysis. Englewood Cliffs, NJ: Prentice Hall; 1999. p. 478-481.

[9] Ludwig JA, Reynolds JF. Statistical ecology: a primer of methods and computing. New York: Wiley Press; 1988. p. 337.

[10] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. Los Altos, CA: Morgan Kaufmann; 2005.

[11] Niven PR. Balanced scorecard for government and nonprofit agencies. John Wiley and Sons Inc; 2003.

[12] The European Health Report, Health Systems Performance Assessment Methods, Annex 1, 2005. http://www.euro.who.int/document/e76907.pdf).