18th PCS/E Conference
Innsbruck/Austria
2-5th October 2002

Generating an OLAP Cube for Hospital Data Analysis

Greinöcker A., Pfeiffer K.P.

Institut für Biostatistik und Dokumentation

Innsbruck, Austria

Keywords: data mining, Austrian-DRG, Minimum Basic Data Set, spatial data analysis

Abstract

In this paper the possibilities of data analysis of the LKF data using data mining procedures are presented. The focus will be set on preparing the data as Online Analytical Processing (OLAP) cubes. Based on this, the special possibilities of spatial data analysis, applied on hospital data, will be shown. As data basis, a representative random sample of Austrian LKF data of the year 2001 was used.

Objectives

The goal of the model is the preparation of LKF data for both microeconomic analysis on a hospital- or hospital department level and nationwide analysis on a macroeconomic level (e.g. financing of the hospital system).

On hospital level, this model should be applicable to different purposes: On the one hand, enabling the (explorative) processing of scientific questions of medical doctors (overview of diagnosis and procedures respectively), on the other hand business management decisions of the hospital management (resource planning and controlling respectively) should be supported. The model should give a detailed view on the data pool of the separate clinical departments. Also the possibility of an aggregated visualization should be possible.

Within the scope of nationwide analysis, differences of diagnosis and medical services concerning demographic and healtheconomic groups of variables should be recognized. In this context, spatial data visualization can offer an overview of the spatial distribution of different variables, like origin of the patients, occurrences of diagnosis and medical procedures.

Methods

Data were prepared in the context of OLAP for multidimensional data analysis. The data can be viewed from different aspects, and these views can easily be adapted to different questions. On the one hand, this enables a global view over the data, on the other hand, when navigating through the different dimensions, detailed views are also possible. The software, which was used for generating of this multidimensional model, was Cognos© Transformer©, and the frontend, which does the visualization and enables the navigation through the cube, is Cognos© Powerplay©.

For the analysis, the Minimum Basic Data Set (MBDS), which is included in the LKF data (= Austrian DRG system) and stored per patient, was used. The MBDS consists of inpatient data (hospital code, type of admission, date of admission, department code), patient specific data (date of birth, gender, main residence), medical data (main- and additional diagnosis, medical procedures (MEL) and results of the scoring..

After an error correction of the data, these were used as a starting point for the OLAP data model. As model, a Single View was used, which keeps all data, including all their hierarchical dimensions, in one single data table. Also the dataset had to be transformed from the relational structure to this flat table.

As dimensions, the residence of the patients (divided by province, district and postal code) departments and sub departments of the hospitals, demographic data of the patients, like age and gender, the overall length of stay, the main- and additional diagnosis (split up following the chapters and sub- chapters of the ICD-10 (International Classification of Diseases, Revision 10)) and also the MEL (split up following the Austrian catalogue of procedures), was used.

As demographic data pool, the data of the Austrian national census of 2001 was used. To enable comparability between the individual districts, data has to be standardized by age and gender.

As so called measures (these are the aggregates in the cell entries of the OLAP cube), the number of patients, age, the overall length of stay (in each case sum and average), was used. In a next step further data e.g. regarding the use of intensive care units or also data regarding socioeconomic factors will be included.

The datasets, which have been aggregated corresponding their geographical dimensions, have been integrated, with the assistance of the geographical information system ArcView, into the spatial data of the provinces and districts.

Results

Additionally, the following datasets for labelling and categorization were necessary:

- An ICD-10 table, where an assignment of the diagnosis codes to their parental group (chapter, sub- chapter, code without 4th character) is specified
- A procedure (MEL)-catalogue, where the MEL codes can be assigned to their parental hierarchy, as well as labels for all categories and codes
- An assignment of the postal codes to their corresponding districts and provinces. Because of the fact, that the assignment of the postal codes is not deterministic, the village with the most inhabitants has been used in case of equal postal codes.
- Names of districts, provinces and villages for the purpose of labelling
- Labels for the type of admission and discharge of the patients
- Labels for the codes of the main clinical departments
- Names of the health insurance institutions
- Assignment of the hospitals to their hospital types (e.g. special hospitals, extended standard service function, …)

In contrast with the analysis described by (Hristovski, 2000), the model was extended with specific data for Austria like codes of the hospital departments and the MEL data.

Spatial data analysis can be enabled via integrating the data with ArcView, to get a spatial representation of the results – aggregated by their geographical groups (region, district, province). The spatial visualization and the possibilities, which spatial data analysis offers respectively, provides important decision support for health system planning (planning of requirements for hospitals and hospital departments).

Another possibility, which results from the combination of the MBDS data with the spatial data, is the visualization of the distribution of certain diseases. Weichbold (2000) performed such an analysis for hepatitis B and C. By this type of analysis, epidemiological problems can be considered on a spatial level.

An example of spatial visualization is shown in figure 1, where the average of the overall score of the LKF data of the year 2001, separated by the districts of Austria, is displayed.

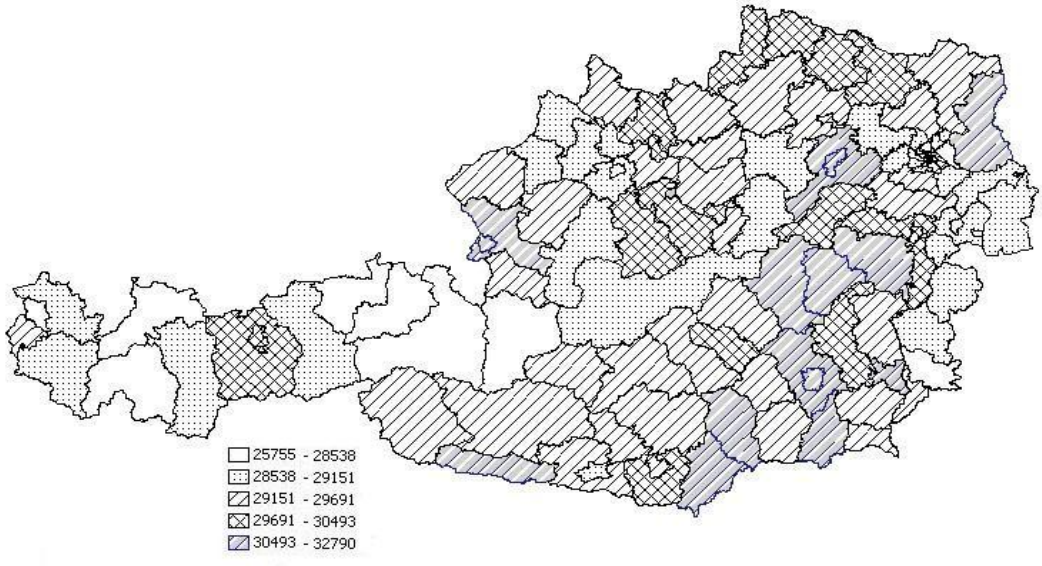


Figure 1: Average of the overall score separated by districts

A drill up (a jump to the parent hierarchy) from the districts to the provinces would result in the following figure:

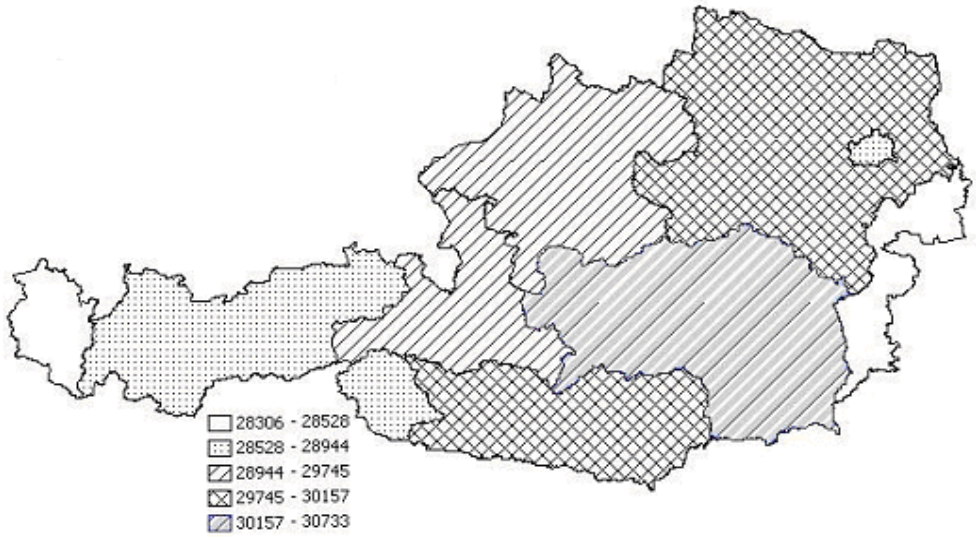


Figure 2: Average of the overall score separated by provinces

Discussion

The drawback, which results from the organization of the dataset as a Single View, is compensated for the profit of speed when accessing, compared to the star-scheme. Also, this structure sets no limits to how the dimensions should be organized. This opens up more flexibility to the user regarding the possibilities of analysis. This approach is only suitable for relatively small Data Warehouses and Data Marts respectively. Because of the very high redundancy, the size of the dataset increases inproportional. In this case, the model is easily transformed into a star- or snowflake-scheme by dividing the single view into pieces.

During the design of the model, the possibilities of easy adding of additional datasets should always be kept in mind. The model should offer the possibility to integrate on the one hand non-spatial datasets like laboratory data taken from the hospital information system, additional information about the patients and on the other hand spatial data, which means additional information about positions in space. An example for spatial data, which can be joined to the LKF- data, is the spatial distribution of ecological damage or the air quality or socioeconomic data. An example for the combination with non-spatial data can be found in Syrjälä (2001) : Here blood transfusion data have been combined with DRG -data. This enables statistical analysis as well as benchmarking and cost comparisons. From the DRG, information like the portion of transfused patients, the number of transfused units and the costs in the different DRG-groups can be extracted.

Of course if it would be desirable, if all functionalities, which have been introduced here, would be available combined in one single application. To safe the effort of implementing these, it would make sense to uxe the functionalities of other applications. For example, the OLAP- functionality of Cognos© Powerplay© could be combined with the possibility of spatial visualization and query of data from ArcView from ESRI. Also statistical and visualization functions from STATISTICA can be used. Technically, this can be realized with component oriented integration of GIS and statistical packages. Most of the modern windows applications offer their functionalities for a use outside of the application via components. The combination and integration respectively can be implemented in Visual Basic, C++ (ATL- library) or C#.

Unfortunately there is no unique primary key per patient available in the Austrian DRG data set because of data security reasons. That is why every person gets, when admitting, a new id assigned, which has the drawback, that no connection between the information after the readmission and the information from the previous stay can be found. Information about the whole medical history of a patient would enable still more possibilities of data analysis.

Another problem, which was already mentioned by Neubauer (2001), is that the residence of a patient is coded with the postal code, but the census data in Austria have geopolitical units, which makes an unique assignment impossible. As already mentioned, conditional remedial action takes the assignment of the region with the highest population to the postal code.

Conclusion

The quality of the MBDS data as well as their usability for scientific analysis is often called into question. Weichbold (2000), however, refers to an evaluation concerning the spatial distribution of this data, where they have been compared with data of a multicenter study. A remarkable agreement in the origin of the registered patients was observed. The experience with the prototype which was introduced here showed, that the MBDS dataset can be analyzed very well with the aid of this OLAP- model. In conjunction with geographical information systems, spatial views on the data could do good work in decision support on different levels of the health service. The introduction of additional variables and dimensions would be sensible (services per clinical department, unique patient id, …).

Prospect

In the future, the whole LKF data should be stored annually in a Data Warehouse with a relational structure. For this, it is necessary, that another primary key, the year, is added. Based on this data pool, the OLAP cubes should be automatically generated corresponding to the needs of the users. The scripts, which do the cube generation, should be implemented in a way, that an adaptation on modifications on the base model (the way data is stored in the Data Warehouse) can be performed easily.

Based on the OLAP cube, Data Mining procedures should be employed to find patterns, which could not be revealed yet. For example, statistically significant differences could be detected between hospitals concerning the length of stay in the same diagnosis group (Lavrac, 1999).

Interesting in this context would be the possibility of making the spatial navigation interactive. For example, a tool could offer the user the map as a GUI for spatial drill up or down functionalities and displaying the aggregates graphically (e.g. as pie charts on the map).

This analysis shows, that with an OLAP- solution, the Austrian-DRG-Data could be used more efficient for scientific, epidemiological and economical analysis.

References

Hristovski, D., Rogac, M., Markota, M.: Using Data Warehousing and OLAP in Public Health Care, Proceedings/ AMIA Annual Symposium. AMIA Symposium, 2000, Pages 369 – 373

Lavrac, N.: Selected techniques for data mining in medicine; Artificial Intelligence in Medicine 16 (1999) 3 – 23

Neubauer, G.: Hospital discharge data, disease frequency and statistical modelling; ROeS-Seminar Mayrhofen, 24. – 26. September 2001

Syrjälä, M. T.: Transfusion Practice in Helsinki University Central Hospital: an analysis of diagnosis- related groups (DRG); Transfusion Medicine, 2001, 11, 423 – 431

Weichbold, V, Bertel, A., Pelzer, A., Pfeiffer, K.P., Ferenci, P.: Hepatitis B und C: Inzidenz und regionale Verteilung der Hospitalisationen in Österreich; Wiener klinische Wochenschrift (2000) 112/23; 995 – 1001

Präsentation

18th International Case Mix Conference
PCS/E 2002, Innsbruck, Austria
2.-5. October 2002

Presentation hold on Friday, Session 2.1/06

PCS/E 2002

Generating an OLAP Cube for Healthcare Data

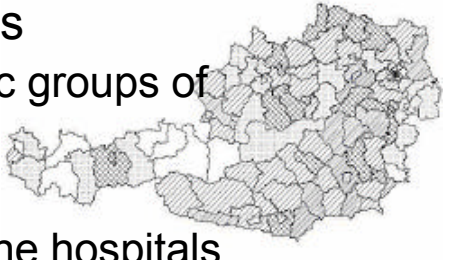
Albert Greinöcker

Institute of Biostatistics and Documentation

University of Innsbruck

Motivation

- Austrian DRG- Data usable for data analysis using OLAP
- Possibilities
  - Hospital- or hospital department level
    - Processing of different questions of medical doctors
      - Overview of diagnosis and procedures
    - Business management decisions of the hospital management
      - Resource planning and controlling
    - Detailed **and** global views possible
  - Regional or nationwide analysis
    - Differences of diagnosis and medical services
      - Concerning demographic and socio-economic groups of variables
    - Spatial data visualization
      - E.g. Visualization of the catchment areas of the hospitals



Methods

- OLAP
  - How can healthcare data be transformed into an OLAP model ?
  - Which possibilities do arise ?
- Spatial data analysis
    - How can explorative spatial data analysis be applied on MBDS-data ?
- Data Basis
  - Sample of Minimum Basic Data Set (MBDS) 2001
    - Administrative data
    - Medical data
    - Scoring data
  - Austrian national census data of 2001
    - To standardize by age and gender
- Software used
  - Cognos Transformer
  - Cognos Powerplay
  - ESRI ArcView

What is OLAP ? (1)

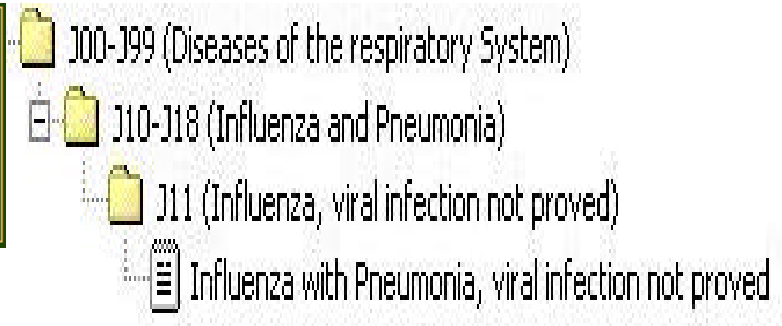
- "**O**n**l**ine **A**nalytical **P**rocessing"
- OLAP enables an user to extract and view data easily and selectively from different points-of-view
  - Fast, because aggregations are pre-calculated
  - Interactive
  - Easy to use
    - No SQL-statements necessary
  - Integrated in a shared environment
    - Multiple users can access the same cube at the same time
    - Analysis on different aspects with the same cube possible
- Generate a detailed **and** global view out of one cube

Sample dimensions (1)

- Age
  - Age groups with steps of five years
    - 1 year-steps
      - Month
        - » days
- Residence (geographical dimension)
  - Province
    - District
      - Postal Code or Community
  - Problem here:
    - Assignment from postal code to the political districts is ambigious

Sample dimensions (2)

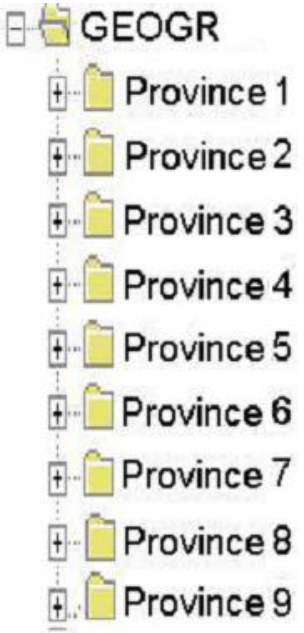
- ICD-10
  - Main Chapters
    - Subchapters
      - Subsubchapters
        - » 4th digit



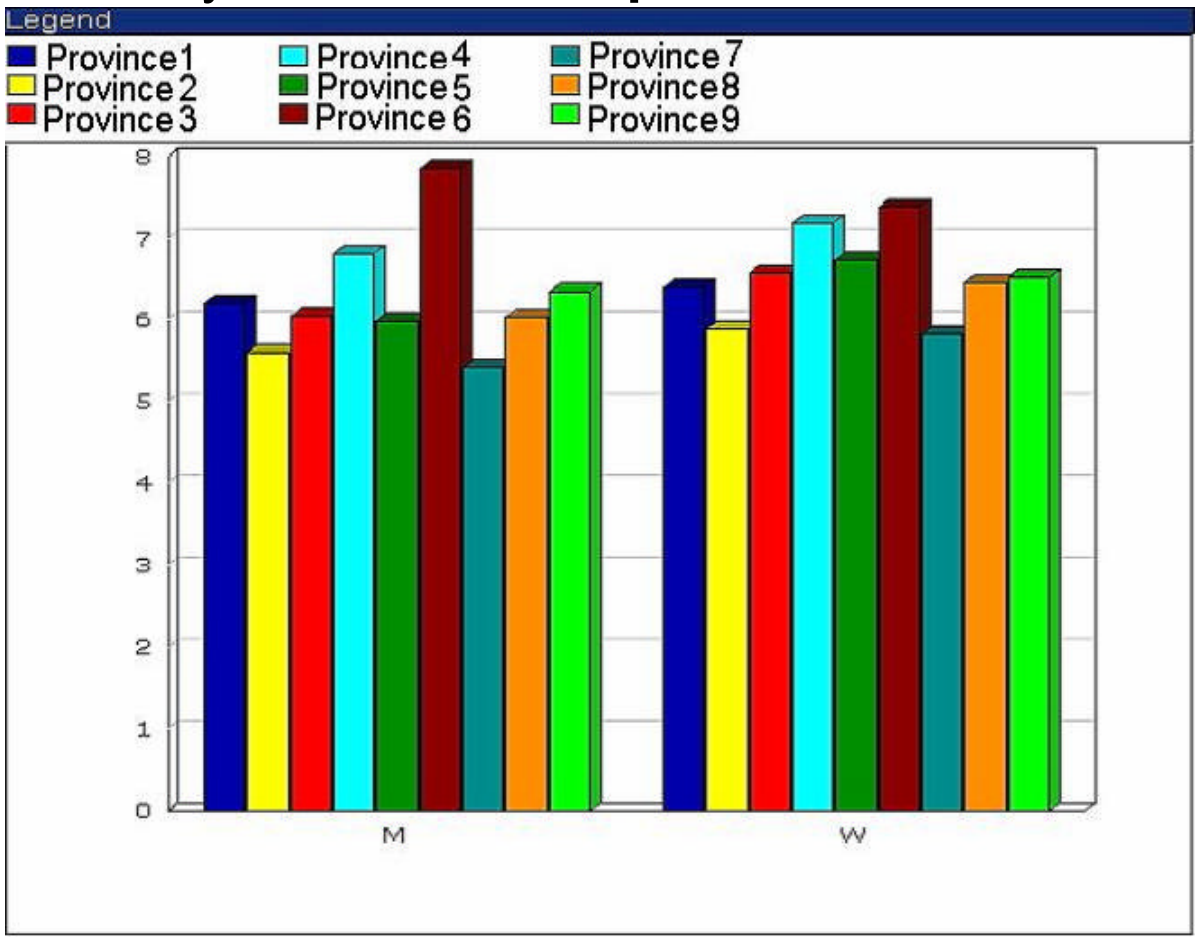
Measures, aggregates & needed datasets

- Facts (measures)
  - LOS (Length Of Stay)
  - Overall Score
  - Age
  - Gender
- Aggregate functions
  - Mean
  - Min
  - Max
  - Count
- Additional necessary datasets:
  - An ICD-10 table
  - A procedures (MEL)-catalogue
  - An assignment of the postal codes to their corresponding districts and provinces
  - Additional datasets for the purpose of labelling

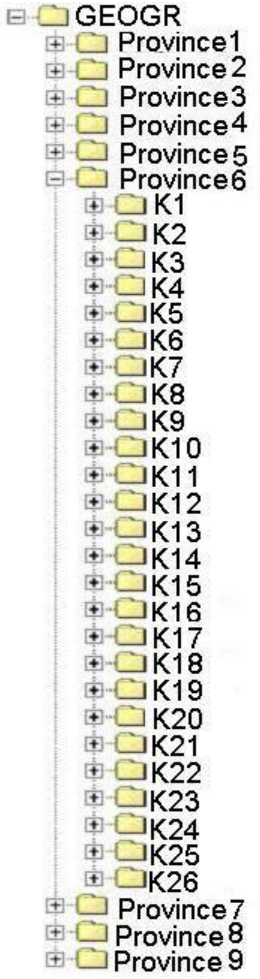
Drill down example (1)



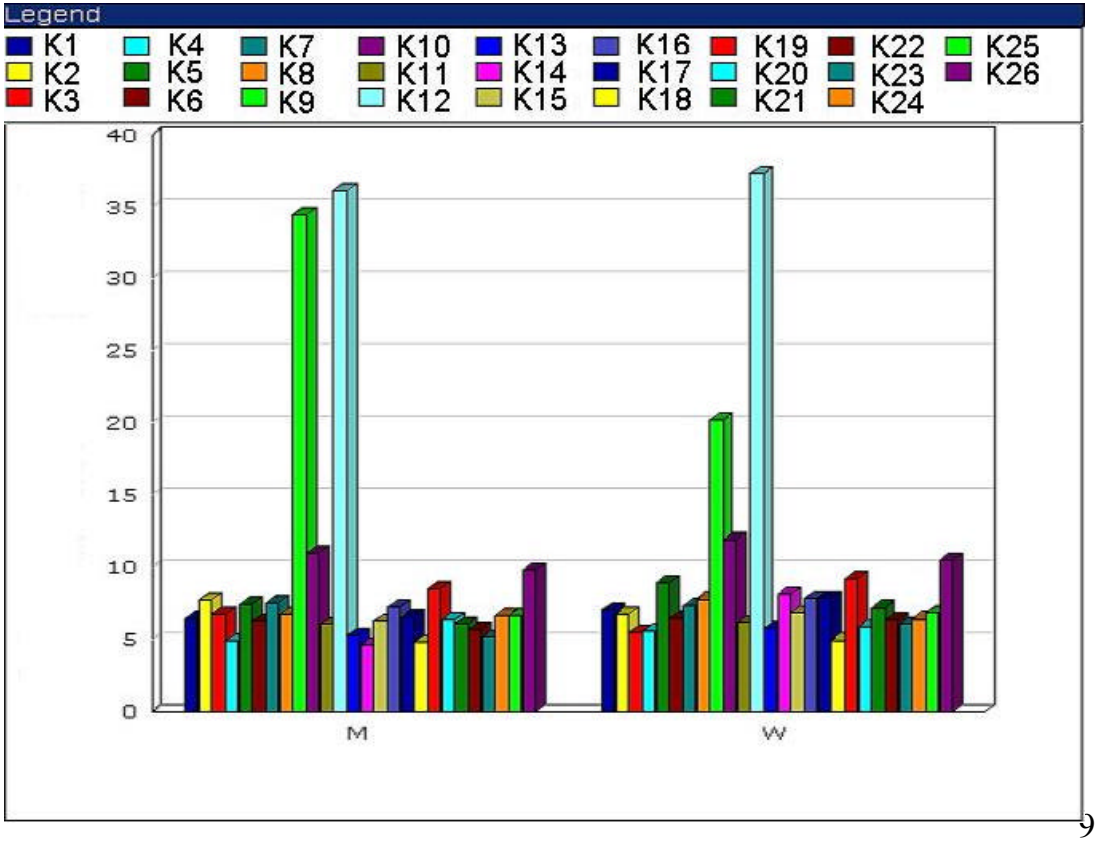
- Compare gender concerning average length of stay in the different **provinces of Austria**



Drill down example (2)

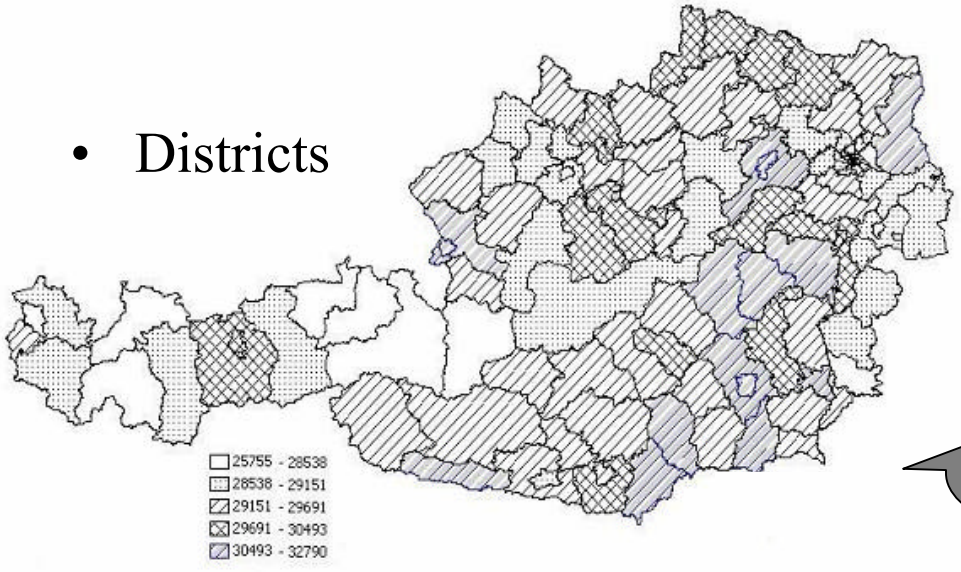


- Compare gender concerning average length of stay in the different **hospitals of one province of Austria**

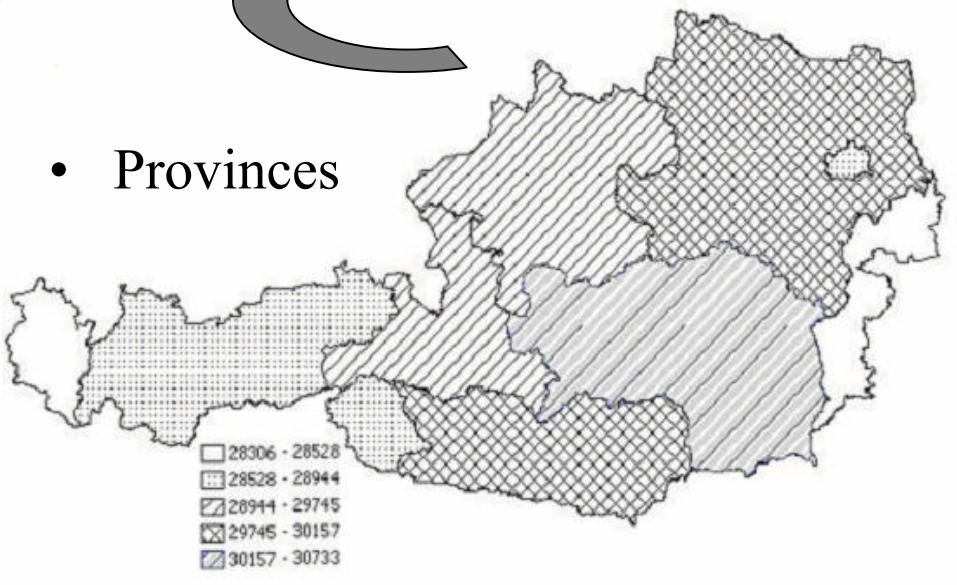


Spatial data analysis

- Districts



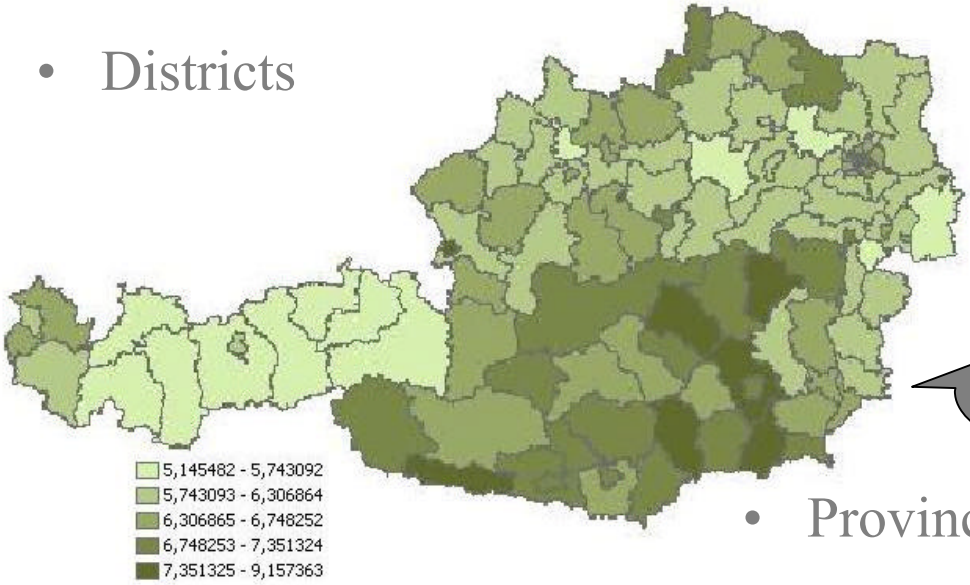
- Provinces



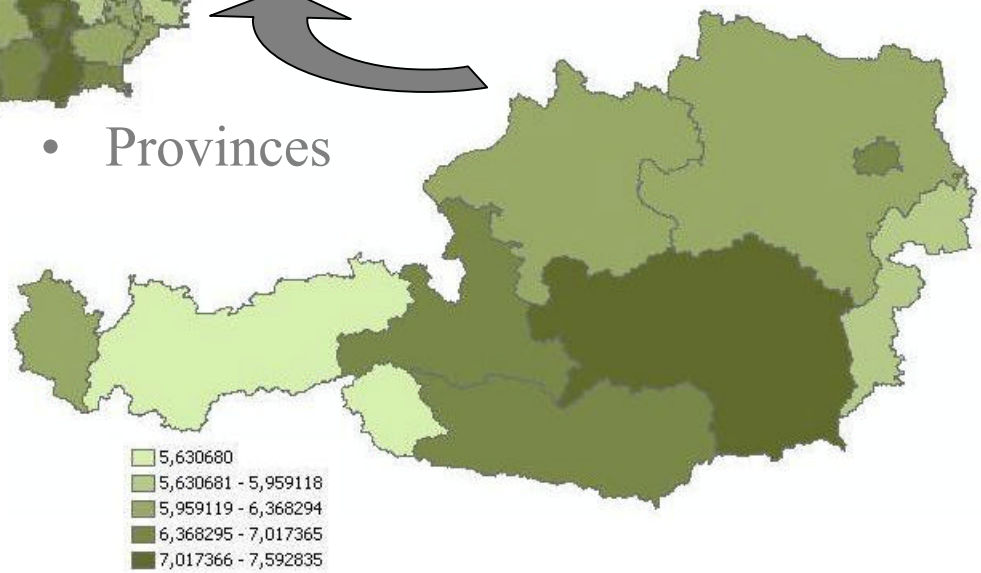
- Average of the overall score of LKF data of the year 2001

Spatial data analysis

- Districts



- Provinces



- Average length of stay of the inpatients of the year 2001

Conclusion

- Problem found
  - No unique personal identifier
    - 1 person ⇔ many admissions per year possible
    - After re-admisssion no follow up
- Assessment
  - MBDS data offers good usability for analysis using OLAP and explorative spatial data analysis
  - Usable for health politics decision support
  - Support for hospital management

Outlook

- The introduction of additional variables and dimensions would be sensible
  - Combine with other spat. distributions
- Integration of all data over the years in one data warehouse
  - Time series analysis
- Data mining procedures should be employed
- Tests on statistical significance should be integrated
- Making spatial navigation interactive
  - the map as a graphical user interface
    - for spatial drill up or down functionalities