Statistical Analysis and Visualisation: Application of Different Software Packages

Magdalena Les, Zbigniew Les
The University of Melbourne
Australian Housing and Urban Research Institute, Melbourne
m.les@ahuri.edu.au

Abstract

A system of integrated packages, as a tool for problem solving at university level, is proposed. In this paper, Excel™, SPSS™, MapInfo™ and Mathematica™ were used as an example of this system. Cluster analysis as an example of statistical analysis is used to the analysis of the migration problem. Visualisation as a main component of the presentation of the results and its role in education is described.

Keywords

visualisation, cluster analysis, a system of integrated packages, problem solving.

1. Introduction

Whenever new technology emerges, its proper incorporation into educational process, depends on understanding the new technology, its power and how it can be applied at different levels of education. In such a highly technologically developed world, one of the major goals in education becomes the development of problem-solving skills of the students. Visualisation has been always recognised as an intrinsic component of teaching-learning from primary to university levels of education (Zimmermann, 1991). As such it plays very important role in process of understanding of the researched phenomena and can also be seen as a means of the development of problem-solving skills.

The aim of the present paper is to show how compatible software packages can be used to solve a particular problem. As an example, a problem from demography is presented. Visualisation is used in the final stage of the analysis to present the findings in order to allow better understanding and interpretation of the results. Application of four software packages Excel, SPSS, MapInfo and Mathematica, as tools for problem solving, is described. The process of problem solving presented in this paper consists of three stages: collection and storing data in the form of transition matrices in Excel, statistical analysis of this data in SPSS and finally, visualisation of the results with MapInfo and animation using Mathematica. All software packages described in this paper work in the Windows environment. For the purpose of this paper**,** all software packages used will be called *'a system of integrated packages'*.

2. Problem solving and visualisation

In the highly technologically developed and growing in complexity world, one of the major goals in education becomes the development of problem-solving skills (Tuma, 1980; Funkhouser and Dennis, 1992). Domain specific knowledge and domain general strategies are very important components in the development of problem solving competence. Especially, the importance of general strategies, which could be applied across a range of tasks and domains, is need to be developed (English, 1992). Solving most of the problems is concerned with collection of data and their explanation. In this paper, in problem solving we are concerned with collection of data, performing an analysis on this data and presentation of the results.

In problem solving it is not only the results which are important but also the clear presentation and communication of these results, which can be seen as a part of explanatory process. Visualisation, as the way of presentation of the results, allows better understanding and give a better explanation of the researched phenomena. The term visualisation, for the purpose of this paper, will be understood as a means toward a better understanding of the concept or the problem. The aim of perception of visual information is to "form a mental image" which allows the better mental representation of knowledge. The three visualisation levels (the presentation, peer and personal levels) was presented in McCormick (1987). Visual images and animation shown in this paper can be regarded at the presentation or personal level depending on the knowledge which students possess. For the students who have only basic statistical and demographic knowledge, visualisation can be seen at the presentation level (in this case only partial interpretation of the results will be possible). For the students who possess deep statistical and demographic knowledge, it can explain demographic concepts such as intraurban migration (see next chapter) in the context of statistical analysis.

3. System of integrated packages

Many software applications have been developed to enhance teaching and learning in particular subject domains (Barta, Eccleston and Hambush, 1993). To achieve an intended educational aim specialist educational packages are being developed (see e.g. Craske, 1991) based on concepts taken from different models, e.g. Intelligent Tutoring Systems. These packages are based on utilisation of the student model, application of the teaching strategies and knowledge representation (Craske, 1991). Also to foster higher-order thinking skills, "an instructional shell for thinking skills that can be easily customised to multimedia content across a range of disciplines, from science and mathematics to the social science and humanities" can be built (Dede, 1992). This educational software is often focused on the learning-teaching of a given segment of knowledge rather than the overall problem solving process. To develop general strategies of the problem solving of students, which can be applicable across a range of tasks and domains, another approach is possible. One of the solutions, proposed in this paper, is to use *'a system of integrated packages'* which allow the complex analysis to be performed.

Packages such as Excel, SPSS or MapInfo are generally used separately, the end result being obtained within the particular software / package. However the possibility exists for integrated use of software packages as tools to solve a given problem**.** In this approach the results of analysis obtained in one package (for example, Excel) can be used as an input in another package (for example, SPSS). The problem to be solved is divided into a set of subproblems, which are solved using knowledge from appropriate domains and computer packages as problem solving tools.

As an example of utilising *'a system of integrated packages'* as a tool for problem solving, the analysis of migration problem is presented. In demography, migration can be studied at three different levels: interstate migration, intrastate and intraurban migration. The results allows the description of the percentage change of population in the capital cities which is directly connected with housing needs. To study migration, different mathematical models, methods of statistical analysis and visual representation of the results can be applied.

The problem of migration, the aim of which was identifying the clusters of Statistical Local Areas (SLAs) characterised by the same 'level' of migration, was solved by applying *'a system of integrated packages'*. In analysis of this problem, data were exported from one application package to another after performing different kind of calculations and analysis, and finally the results were visualised. There are many of the software packages which can be used at each stage of this analysis. An application of four selected software packages Excel (storage and preliminary calculations), SPSS (statistical analysis), MapInfo and Mathematica (visualisation), as tools for problem solving in demography is presented in the following chapters.

4. Data storage and preliminary calculations

The collected data need to be stored and further calculations and analysis need to be done. There are many powerful tools for storing, displaying, presenting and managing data, such as Paradox, dBase or Excel. Excel is a powerful calculation spreadsheet with many build-in options / functions which allow very fast and easy calculations to be done. Data stored and preliminarily processed in Excel can be easily exported into other systems such as statistical packages SPSS or SAS for further analysis.

Excel is especially useful for storing small data matrices on which further calculations need to be performed. The basic demographic data file presented in this paper was stored in Excel. Data matrices represented intraurban migration on the level of Statistical Local Areas (SLAs) in the borders of Melbourne statistical division. 'Flows' from the origin SLAs to destination SLAs (matrices 'origin by destination' for the 5-year periods 1976, 1981, 1986 and 1991) were expressed in absolute numbers. To perform further analysis, the transition probabilities depicting 'flows' of people 'out' of the particular SLAs or 'in' particular SLAs were calculated (in Excel). These transition probabilities were used as an input data file in SPSS for further statistical analysis.

5. Statistical analysis of data

To solve a problem of identifying Melbourne's SLAs (Statistical Local Areas) which represent the similar level of intraurban migration, selection of the tool and proper method of analysis is essential. In this case as a tool the SPSS package was used. SPSS for Windows, release 6, is a comprehensive and flexible statistical analysis and data management system. Convenient options in SPSS for converting data files from dBase (.db) or Excel (.xls) formats allows for immediate performing statistical analysis on data files created in other software packages (Norusis, 1993).

The Excel data file was converted into SPSS where the aim of statistical analysis was to establish relatively homogeneous groups of objects (SLAs) in order to find patterns of 'flows' within Melbourne statistical division area. This was done using cluster analysis. There are many methods of cluster analysis with the main two approaches being agglomerative and divisive hierarchical clustering (Anderberg, 1973). In the case of the

demographic data used, the most appropriate approach was an agglomerative, hierarchical cluster analysis. The Ward's method was selected for cluster formation. Ward's method calculates the means for each variable under study. Then, for each case, the squared Euclidean distance to the cluster means is calculated according to the formula:

distance(SLA_i,SLA_j)= $\sum_i (x_i - y_i)^2$, x_i - coordinates of SLA_i, y_j - coordinates of SLA_j;

where i=1,2,...,n and j=1,2,...,n (n=58 is the number of Melbourne's SLAs). As a result, the SLAs which are characterised by the similar 'flows' of the number of people "in" or "out" of the particular SLAs were identified. The results of cluster analysis were represented in a visual form called a dendrogram. However, the raw dendrogram (see Figure 1) is not easy to interpret in terms of identifying the patterns of intraurban migration. In order to represent the results in a visual form the MapInfo package was used to create shaded maps of Melbourne's SLAs (see Figure 2).

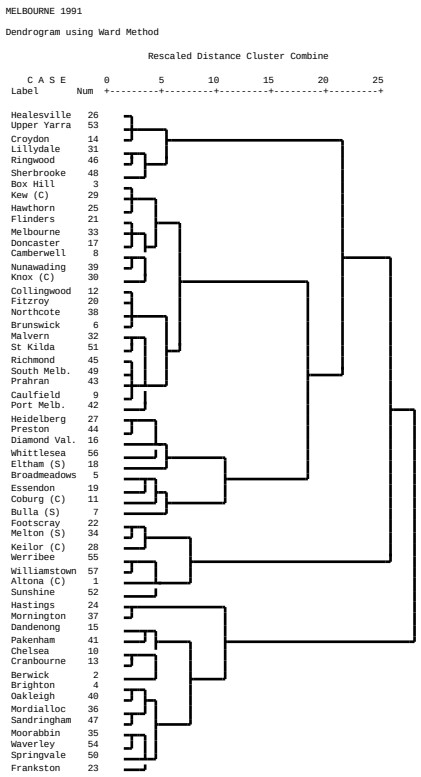

Figure 1. Dendrogram (tree) representation of the clusters of SLA's which shows the similar level of intraurban mobility

To visualise the data (in MapInfo) which represented the dendrogram in SPSS, an additional variable indicating the cluster membership of the particular SLA in Melbourne was created and saved in working data file in SPSS. The number of clusters, which best depict differences in segmented sample, was selected on the basis of analysis of the dendrogram (based on the theory of cluster analysis). The variable, which represented the best results of clustering, in this case, when the original set of SLAs is divided into five clusters, was selected (see Figure 1). This newly created

variable, with the names of SLAs assigned to it, was used as a key variable to map clusters in MapInfo.

6. Visualisation of the results

6.1 Application of MapInfo

To represent, in a visual form, data which has been created as a result of cluster analysis MapInfo package was used. MapInfo for Windows is a powerful desktop mapping system which allows the user to generate colour maps by using data supplied with MapInfo or imported to MapInfo from another application. Data imported into MapInfo from another software are not assigned to geographical objects such as maps or areas incorporated in MapInfo. To be able to represent graphically this data, they must first be geocoded in the MapInfo package. Geocoding is a process of assigning point objects to records in a table.

In this case, data imported from SPSS were geocoded on the basis of the names of SLAs. After successful geocoding, the map of Melbourne with incorporated SLA's in the borders of Melbourne statistical division was 'shaded by value' showing very clearly patterns created by SLAs grouped after cluster analysis in SPSS (see Figure 2). The old maxim that "a picture is worth one thousand words", is the best exemplification of the whole process of visualisation of the results of statistical cluster analysis.

Exemplification of dendrogram visualised in the form of a shaded map

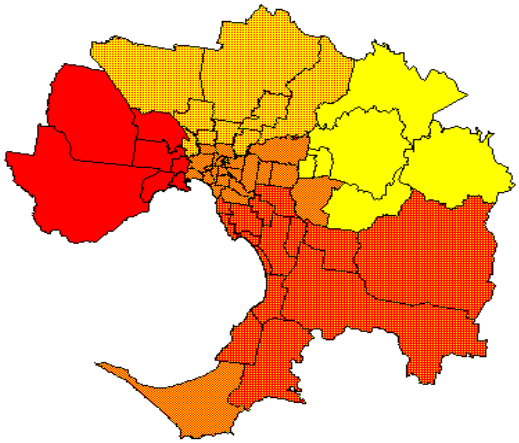

Figure 2. Shaded SLA's of Melbourne showing 'intraurban mobility' between years 1986-1991.

6.2 Application of Mathematica

To obtain a visualisation of the results in the form of animated pictures, which show changes of patterns of intraurban migration through the years 1976-1991, a Mathematica as a presentation tool was used. Mathematica is a computer package which has very broad application, including numerical or symbolic calculations, modelling and analysis of data and as a programming language (Wolfram, 1993).

In this paper, Mathematica is used only as a visual system to animate data, which is not typical uses of this kind of software. The overall assumption was that a moving picture conveys more information that a static one so a good animation contributes to better understanding of the results. The basic idea is to generate a sequence of "frames" which can be next displayed in rapid succession. Usually a standard Mathematica graphics functions can be used to produce frames; the mechanism for displaying the frames as a movie depends on the used Mathematica interface. The sequence of frames for a movie can be obtained by using the MapInfo package. After transforming into a bitmap format, the pictures (frames) are put in a sequence of cells and then animated using the command from the menu. This is a simplest and fastest way to get animated graphics which show, in this particular case, the intraurban movement in Melbourne's SLAs over time.

7. Conclusions

Use of a system of integrated software packages as a tool for problem solving and explanation of the results of analysis was presented. This system can be used as a means to teach a problem solving in a *given domain*. The main advantages of such an approach is that it is possible to incorporate knowledge from many domains as well as a range of existing computer software packages as tools for problem solving. A disadvantage of this approach is the longer time needed to become familiar with a set of software packages. Students have to require the skills of using many specialised packages, which contain different chunks of information, effectively. Knowledge not only of the material but also familiarity with the computer packages is needed. An application of *'a system of integrated packages'* as a tool for problem solving seems to be a promising proposal for the development of general strategies of the problem solving of the students. It can be controversial, but the knowledge developed and experience gained using 'a system of integrated packages' can deepen the understanding of the studied topic and give better insight into solving complex problems.

8. References

Anderberg, M. (1973). *Cluster analysis for applications*. Academic Press, New York.

Barta, B., Eccleston, J. and Hambush, R. (Eds.) (1993). *Computer mediated education of information technology professionals and advanced end-users*. North-Holland, Amsterdam.

Craske, N. G. (1991). Knowledge representation and architecture in intelligent tutoring systems. Unpublished Phd Thesis, La Trobe University, Melbourne.

Dede, Ch. J. (1992). The future of multimedia: Bridging to virtual worlds, *Educational Technology*, May, pp. Vol. XXXII, No. 5, pp. 54-60.

English, L. (1992). Children's use of domain-specific knowledge and domain-general strategies in novel problem solving, *British Journal of Educational Psychology*, Vol. 62, pp. 203-216.

Funkhouser, Ch. and Dennis, R. (1992). The effect of problem-solving software on problem solving ability, *Journal of Research on Computing in Education*, Vol. 24, No. 3, pp. 38-348.

McCormick, B. H., DeFanti T. and Brown, M. (Eds.) (1987) Visualisation in scientific computing, *Computer Graphics*, 21, Vol. 26, No. 6

Norusis, M.J. (1993). *SPSS for Windows, Professional Statistics, Release 6.0.* (software) SPSS Inc., Chicago

Tuma, K (1980). *Problem solving and education: Issues in teaching and research,* Lawrence Erlbaum, Englewood Cliffs, NJ

Wolfram, S. (1993). *Methematica. A system for doing mathematics by computer*. Addison-Wesley Publishing Company, New York

Zimmermann, W. and Cunningham, S. (Eds.) (1991) *Visualization in Teaching and Learning Mathematics.*