OLAP FOR HEALTH STATISTICS: HOW TO TURN A SIMPLE SPREADSHEET INTO A POWERFUL ANALYTIC TOOL

Barbara Artnik [(1)], Gaj Vidmar [(2)], Jana Javornik [(3)]

(1) University of Ljubljana, Faculty of Medicine, Institute of Social Medicine
(2) University of Ljubljana, Faculty of Medicine, Institute of Biomedical Informatics
(3) Government of the Republic of Slovenia, Institute of Macroeconomic Analysis and Development

ABSTRACT

Over the last ten years, Online Analytical Processing (OLAP) has become a very popular tool for interactive analysis of multidimensional information. Providing online operation and flexible summarising, tabulating and charting options, it has become an essential part of the decision support process in corporate setting. Our aim is to demonstrate how easily applicable and useful OLAP can be in the public sector. To achieve that, we used data compiled from different sources for the purpose of exploring the relations between causes of death (according to ICD-10) and socio-economic characteristics (educational level, marital status, profession, etc.) for selected years 1992, 1995 and 1998 in Slovenia. Using a standard personal computer and the Windows® platform, the application was implemented in Microsoft® Excel 2000, without any programming. After data cleansing (elimination of incorrect entries, duplicates and inconsistencies based on exploratory statistical methods), the case-based spreadsheet data was instantly converted into an OLAP application with the user-friendly pivot table technology. A bonus of this approach is that the results can be made directly accessible over the WWW by publishing the workbook to a web server. Provided that the user has Microsoft® Internet Explorer and Microsoft® Office 2000 installed, all the drill-in, drill-out and dimension-swapping capabilities are accessible within the browser, while the data source remains fully protected. Privacy constraints are respected since all the information is only provided at the aggregate level. Even though our dataset provides exhaustive coverage of mortality at the national level, storage and processing capabilities did not prove to be an issue. Hence, we argue that OLAP methodology should find a place in health statistics. With proper data collection and/or transformations, informative comparisons within the study population and with international databases become readily accessible. The key advantage of OLAP over relational database management systems and ordinary tables is interactive browsing of multidimensional and hierarchical data, while OLAP can also aid data integrity checking and reporting.

OLAP V ZDRAVSTVENI STATISTIKI: KAKO SPREMENITI PREGLEDNICO V ZMOGLJIVO ANALITIČNO ORODJE

V zadnjih desetih letih je OLAP (Online Analytical Processing) postal priljubljeno orodje za interaktivno analizo večrazsežnostnih podatkov. Zaradi sprotnega odziva in raznolikih možnosti povzemanja in grafičnega prikazovanja podatkov je v gospodarstvu postal nepogrešljiv del sistemov za podporo vodstvenemu odločanju. S prispevkom skušamo pokazati, da je OLAP uporaben in koristen tudi v javnem sektorju. Uporabili smo podatke, zbrane iz različnih virov za potrebe preučevanja povezanosti vzroka smrti (po MKB-10) s socio-ekonomskimi značilnostmi (stopnjo izobrazbe, zakonskim stanom, poklicem, idr.) za izbrana leta 1992, 1995 in 1998 v Sloveniji. Na običajnem osebnem računalniku z operacijskim sistemom Windows® smo celotno aplikacijo izdelali v orodju Microsoft® Excel 2000 brez kakršnegakoli programiranja. Po prečiščenju podatkov (odstranitvi napačnih vnosov, podvajanj in neskladij v podatkih na podlagi eksploratornih statističnih analiz) smo preglednico s podatki za vsako posamezno osebo zlahka pretvorili v aplikacijo OLAP z uporabo vrtilnih tabel. Dodatna prednost tega pristopa je, da je moč rezultate takoj objaviti v svetovnem spletu – delovni zvezek zgolj prenesemo na spletni strežnik. Če ima uporabnik nameščeno potrebno programje (Microsoft® Internet Explorer in Microsoft® Office 2000), so vse zmožnosti vrtanja navzdol in navzgor ter zamenjevanja razsežnosti dostopne znotraj spletnega brskalnika, pri čemer izvorni podatki ostanejo zaščiteni. Tudi zaupnost osebnih podatkov je zavarovana, saj se vse analize izvajajo na nivoju združenih podatkov. Čeprav podatki obsegajo vse registrirane smrtne primere v državi, zmožnosti strojne in programske opreme niso predstavljale omejitve. Zato menimo, da bi se metodologija OLAP lahko hitro uveljavila v statistiki tudi na področju zdravstva. Če podatke ustrezno zberemo in predelamo, postanejo analize znotraj preučevane populacije in primerjave z mednarodnimi zbirkami podatkov dostopne dobesedno v trenutku. Bistvena prednost OLAP-a pred sistemi za upravljanje z relacijskimi podatkovnimi zbirkami in običajnimi tabelami je interaktivnost pregledovanja večrazsežnostnih in hierarhično organiziranih podatkov, OLAP pa lahko pomaga tudi pri zagotavljanju kakovosti podatkov in izdelavi rednih poročil.

1. INTRODUCTION

In the early eighties, new methodologies for searching information in existing databases were developed, including knowledge discovery in databases (KDD) and *online analytical processing* (OLAP). The term OLAP was introduced almost a decade ago in a report commissioned by a software vendor [1], but a less controversial contemporary definition is *fast analysis of shared multidimensional information* (FASMI), endorsed by Nigel Pendse [2]. In addition to multidimensionality of the data, key features of OLAP are online operation, built-in and programmable analytical capabilities, and different presentational and reporting options. Common characteristics of KDD and OLAP algorithms are their operation on large datasets and results, which mainly consists of aggregated information from existing databases, not known in advance. The key advantage of OLAP over *relational database management systems* (RDBMS) and ordinary tables is interactive browsing of multidimensional and hierarchical data, while OLAP can also facilityte data integrity checking and improve reporting.

First OLAP implementations were limited to large enterprise and scientific datasets handled by proprietary systems. Today, numerous commercial systems are available and almost all RDBMS and statistical packages include support for OLAP. This powerful technology is hence available to anyone dealing with large datasets, but at quite high price – not only in terms of software, but also or even more so in terms of implementation/consultancy costs.

2. EXCEL'S KEY FEATURES FOR DATABASES AND OLAP

Since its introduction in 1987, Microsoft® Excel has developed into the most popular and versatile spreadsheet application on the software market. Regarding database capabilities, the major developments were the introduction of multi-sheet workbooks and pivot tables in 1993 (version 5), new VBA interface and data validation in 1995 (version 8/97) and pivot charts in 1999 (version 9/2000).

2.1 Excel and databases

The starting point for database functionality in Excel is that any worksheet or part of a worksheet can serve as a database once a header row of field names is followed by the data rows below. In real-life applications, it is strongly recommended to use entire worksheets as data tables, especially because that enables the simple yet powerful use of the AutoFilter feature, while separate sheets should be used for reports, pivot tables and charts. With Advanced Filters, of course, selection and searching is extended and refined, while simple Conditional Formatting instantaneously adds visual analytical functionality to any table, regardless of the data aggregation level.

Being oriented towards Microsoft® Office products in this article, we cannot avoid the issue of Excel vs. Access. In spite of obvious limitations (quantity of data, one user at the time), Excel has major advantages over Access as regards reporting. As soon as detailed formatting and/or complex

® *Microsoft and Windows are registered trademarks of Microsoft Corporation. Oracle is a registered trademark of Oracle Corporation. SAP is a registered trademark of SAP AG.*

analyses are required in a report, or the report data should be used to generate other reports, Excel is virtually the only option. Generally speaking, the analogy of Excel vs. Access being like a car vs. a truck is an informative summary of the comparison of the two programmes – in terms of developers ("drivers"), speed, capacity, costs, adaptability and required organisational support. In this context, one can think of Oracle®, Microsoft® SQL Server or SAP® as trains, boats, airplanes.

Since this article aims at stressing what users can do with Excel without any programming, using only built-in capabilities and bundled add-ins, we should point to three worksheet functions, accessible via simple formulas, that provide a true break-through in terms of database functionality: SUMPRODUCT, INDEX and MATCH. They are primarily designed for reporting, but they can also be essential in the pre-processing phase of an OLAP application, or can be used to provide the data to be subsequently processed by a pivot table.

Another database-related tool, which is a part of the standard Microsoft® Office package, is Microsoft® Query. As it is not a separate application in the latest versions, it is even more hidden from the typical user. However, like the other components of the package, it has become a truly professional and powerful tool over the years. We have not enough space to go into any detail here, but it should be stressed that its relatively simple operation provides Excel workbooks with full functionality of a relational database.

2.2 Pivot Tables and Pivot Charts

Pivot Tables are the key feature for doing OLAP in Excel without VBA programming. In a stepwise wizard-guided process, the user defines the data source and the fields that define pages, columns, rows and data (i.e., measures) of the pivot table. Various table options can be set along the way or after the table had been created. Instead of giving any further description, we refer the reader to Excel's extensive help and the web resources listed below.

Pivot Charts are designed using the same wizard as the Pivot Tables. We do not present an example here, because no simple visualisation would be truly informative and more appropriate than a table for our example data. However, there are cases when Excel's pivot charts are a useful OLAP tool, if they are designed by a person with solid background in statistics and data presentation.

2.3 Web resources for developers

The use of Excel for data warehousing, reporting and OLAP can considerably reduce IT costs also because of the minimum training costs for the developers. A huge body of instructions and tips is publicly accessible thanks to a number of enthusiasts who maintain extremely valuable resource websites dedicated to Excel. Here we list a selection, ordered alphabetically by author surname:

- J. F. Lacher
  http://lacher.com/toc/tutpiv.htm
- P. Leclerc
  http://www.excel-vba.com/
- T. Mehta

http://www.tushar-mehta.com/

- C. Pearson
  http://www.cpearson.com/excel.htm
- J. Peltier
  http://www.geocities.com/jonpeltier/Excel/index.html
- D. Steppan
  http://geocities.com/dsteppan/ExcelTop.html
- J. Walkenbach
  http://www.j-walk.com/ss/excel/index.htm

3. EXAMPLE APPLICATION

The data for the example application were compiled from different sources for the purpose of exploring the relations between causes of death (according to ICD-10) and socio-economic characteristics (educational level, marital status, profession, etc.) for selected years 1992, 1995 and 1998 in Slovenia as part of the PhD project of the first author, which is related to her previous work [3].

The initial data-cleansing phase consisted of elimination of incorrect entries, duplicates and inconsistencies based on exploratory statistical methods.

We designed two separate pivot tables, each with a single measure, in order to avoid formatting problems Excel exhibits when the multiple measures feature is used. In Figure 1, the use of several row and column dimensions is demonstrated and the categorical outcome is reported as percentage using the standard work-around of counting the number of cells with non-empty ID field. The active window demonstrates how filtering works.

In Figure 2, the measure is the average of a numerical field. Instead of containing zero value, empty cells are clearly marked, which is a straightforward option. The selected cell is aimed at demonstrating how drill-in and drill-out is achieved by double-clicking. The usefulness of the "preserve formatting" option is evident from both figures.

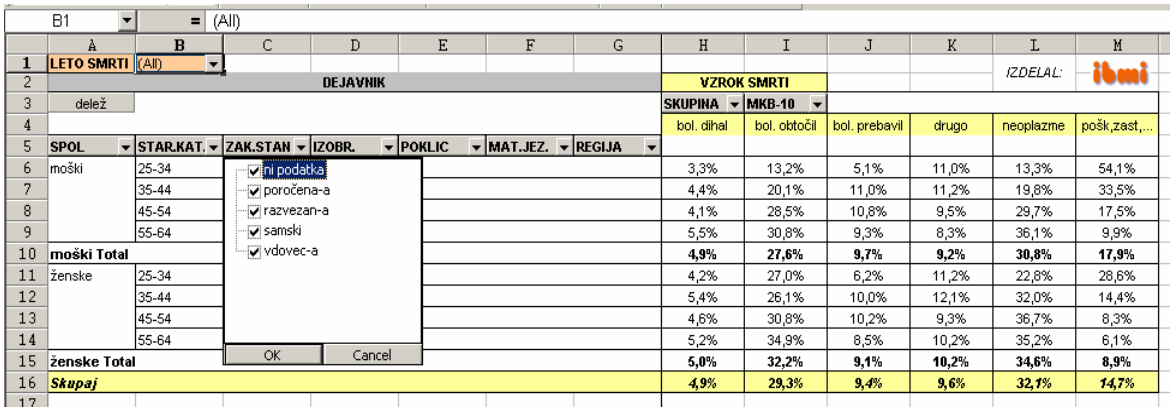

Figure 1. Sample pivot table screenshot – measure is percentage of cases.

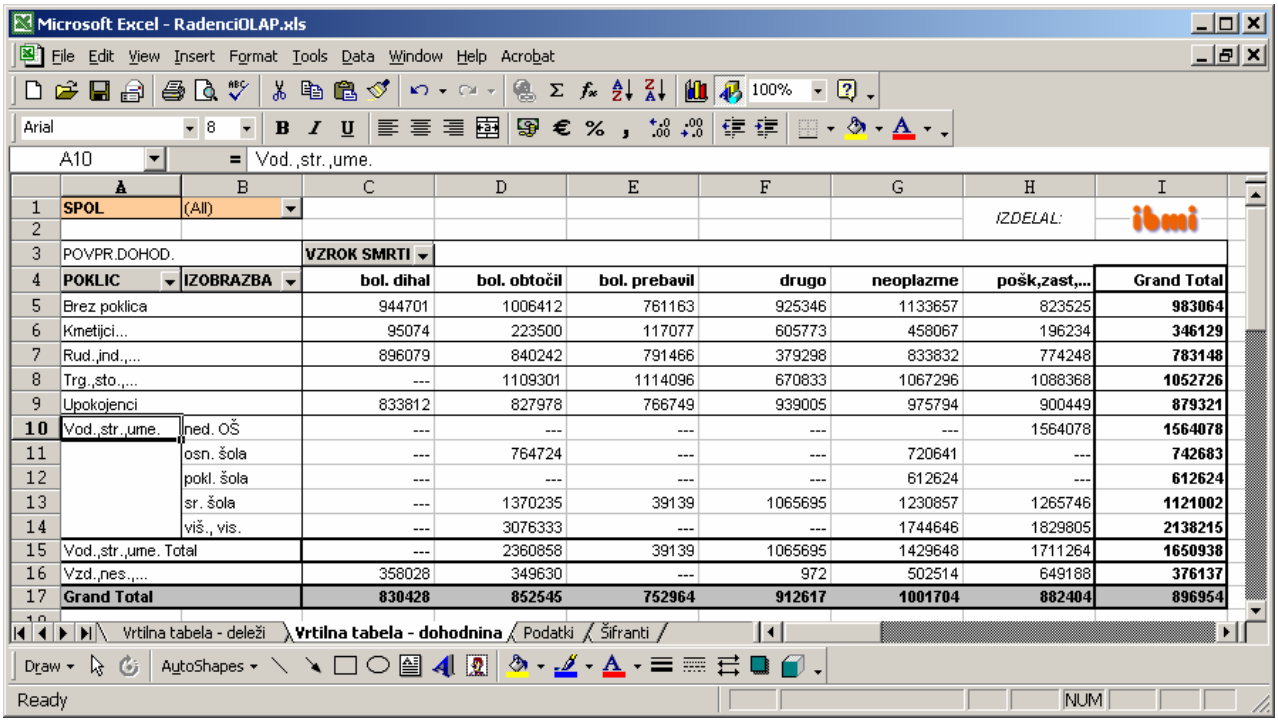

Figure 2: Sample pivot table screenshot – measure is average of a numeric field.

Even though Microsoft® Excel 2000 is included or at least mentioned in his review, Thomsen [4] argues that spreadsheets cannot provide adequate OLAP functionality. He claims that they completely fail to meet four key OLAP requirements: multiple dimensions, hierarchies, dimensional calculations and separation of structure and representation.

Rather then getting into a lengthy argumentation, we encourage anyone interested in challenging this opinion to try out Excel's capabilities and see for himself/herself that this is not really the case. In our opinion, the only serious problem are hierarchies. Hence, it is possible to conclude that simple OLAP applications with a limited number of dimensions can be adequately, quickly and easily implemented in Excel.

Another advantage of the described approach is that the results can be made directly accessible over the WWW by publishing the workbook to a web server. Provided that the user has Microsoft® Internet Explorer and Microsoft® Office 2000 installed, all the drill-in, drill-out and dimension-swapping capabilities are accessible within the browser, while the data source remains fully protected. If the application is properly designed (data worksheets hidden), privacy constraints are respected since all the information is only provided at the aggregate level. Contrary to widespread belief, storage and processing capabilities are not a serious issue with this approach with up to tens of thousands of records.

4. CONCLUSION

At virtually no cost, a functional OLAP application can be developed with Microsoft® Excel, based on the Pivot Table/Pivot Chart facility.

Database and OLAP functionality in Excel is a widely accessible technology and we believe that do-it-yourself decision-support systems based on it could become widely applied in the field of public health, as well as in official statistics, accounting, actuarial work, retail business and elsewhere.

5. REFERENCES

[1] Codd, E.F., Codd, S.B., Salley, C.T. (1993). Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. [Codd & Associates Technical Report, for Arbour Software, now Hyperion Solutions White Paper]

[2] Pendse, N. (2001). *What is Olap?* [URL http://www.olapreport.com/fasmi.htm, part of The OLAP Report website]

[3] Artnik, B., Premik, M. (2001). Health inequality in Slovenia. *Medical Acrhives*, *55* (1), 37–39.

[4] Thomsen, E. (2002). *OLAP Solutions: Building Multidimensional Information Systems* (Second Edition). New York: John Wiley.