The Excel Data Mining Add-in. Applications in Audit and Financial Reports

Daniel HOMOCIANU,
“Alexandru ioan Cuza” University of Iași,
E-mail: daniel.homocianu@feaa.uaic.ro

Dinu AIRINEI,
“Alexandru ioan Cuza” University of Iași,
E-mail: adinu@uaic.ro

Abstract
Performance reasons in decision making based on business data usually require a good management of multiple data formats and also processing speed, flexibility, portability, automation, power of suggestion and ease of use. The paper comes with theoretical ideas and practical examples in favor of using the Excel Data Mining Add-in's for the aforementioned reasons. Most of the examples include figures linked to video scenarios constructed by the authors and part of an interactive online list with eighteen pieces. Together they contribute to understanding most of the requirements to fulfill in order to have valid examples and useful results.

Keywords: business and financial data, spreadsheets, Data Mining (DM), examples

JEL Classification: C61, D81, D83, M42

To cite this article:
Homocianu, D. and Airinei, D. (2017). The Excel Data Mining Add-in. Applications in audit and financial reports, Audit Financiar, vol. XV, no. 3(147)/2017, pp. 451-468, DOI: 10.20869/AUDITF/2017/147/451

To link to this article:
http://dx.doi.org/10.20869/AUDITF/2017/147/451
Received: 17.03.2017
Revised: 13.04.2017
Accepted: 20.04.2017
The Excel Data Mining Add-in. Applications in Audit and Financial Reports

Ph D researcher Daniel HOMOCIANU
Ph D professor Dinu AIRINEI

The Research Department, Faculty of Economics and Business Administration (FEAA),
“Alexandru Ioan Cuza” University of Iaşi (UAIC),
daniel.homocianu@feaa.uaic.ro
The Department of Accounting, Business Information Systems and Statistics, FEAA, UAIC,
adinu@uaic.ro

Abstract
Performance reasons in decision making based on business data usually requires a good management of multiple data formats and also processing speed, flexibility, portability, automation, power of suggestion and ease of use. The paper comes with theoretical ideas and practical examples in favor of using the Excel Data Mining Add-in’s for the aforementioned reasons. Most of the examples include figures linked to video scenarios constructed by the authors and part of an interactive on-line list with eighteen pieces. Together they contribute to understanding most of the requirements to fulfill in order to have valid examples and useful results.

Key words: business and financial data, spreadsheets, Data Mining (DM), examples

JEL Classification: C61, D81, D83, M42

Introduction
This paper starts from some techniques used by most Data Mining tools when dealing with large data from databases and presents the advantages of using spreadsheets as client applications (msdn.microsoft.com/.../dn282385.aspx). The last ones are so familiar to the end users and have an interface that integrates programming or scripting languages for office applications such as VBA meaning Visual Basic for Applications for Microsoft Excel (msdn.microsoft.com/.../ee814737.aspx), and Google Apps Script for Google Sheets (developers.google.com/.../sheets), many functions and advanced facilities for processing, analysis, representation and simulation all based on the interactivity and dynamics principles with great impact on users’ ability to perceive, interpret, understand and manage complex information in different cases.

The concept of Data Mining essentially means the supervised identification of undiscovered patterns and hidden relationships in huge data sets (searchsqlserver.techtarget.com). Inmon which is a well-known guru in data warehousing (computerweekly.com) gave one of the most concise definitions of a Data Mining (Inmon and Linstedt, 2014) namely analysis of large quantities of data to find patterns such as groups of records, unusual records and dependencies. The Data Mining initiatives usually come from marketing and retail sales departments and are suited for organizations having very large databases (Airinei, 2002).

This concept is closely related to that of data oriented Decision Support Systems (DSS) and Business Intelligence (BI) – especially the one for strategic purposes (dssresources.com/...id=174) that requires huge amounts of data (bi-insider.com). Although the BI term is known as a set of concepts and methods for improving decision-making emerging in the 90’s (Howard Dresner from the Gartner Group - dssresources.com/.../dsshistory.html), the evidences from the specialty literature indicate approaches
from 15 years earlier (Cleland and King, 1975; Pearce, 1976) containing clear references to BI, business planners and managers and decision making.

As concluded by Dan Power (dssresources.com/...id=199), Data Mining tools include: case-based reasoning, data visualization (mostly graphs, trees, and clusters), fuzzy queries and analyzes, genetic algorithms, and neural networks.

Starting a few years ago we are witnessing implementations of this concept and related models not only in applications dedicated to database and data warehouse management systems, but in modules of spreadsheet applications that are working with these above as suggested even from this paper’s title. This seems obvious when thinking that such dedicated products allowed the construction of DM structures and models starting simply from one table (usually as aggregation of many others from a database).

The applicability of the theoretical and practical elements of this article in auditing, especially the one of performance (Fraser, 1998) and financial reports is justified starting from a specific need to valorize the existing data structures (often data records in tables and tables in databases) and get rapidly and at minimum cost reports able to present clear information on causality related to effectiveness (actual / estimated results compared to those proposed) and efficiency (consumed resources compared to achieved / estimated results).

The concrete examples in this article support certain conclusions drawn from the literature review, namely: the utility of approaching the audit engagements by using data mining techniques (Vintilescu Belciug et.al., 2010) as a complement to traditional methods of risk analysis and intervention on site, the consecration of existence of possible areas of integration between data mining and audit processes grouped by stages (Sirikulvadhana, 2002) such as: planning, execution, documentation and completion or by specific examples (Wang si Yang, 2009) as neural networks for: risk assessment, finding errors and fraud, determining the going concern of a company, evaluating financial distress, and making bankruptcy predictions and decision trees for: analysis of bankruptcy, bank failure, and credit risk), the advance of the latest software tools that implement data mining algorithms and the fact that many users considered them until recently being not very friendly (Chersan et.al., 2013) and requiring technical skills advanced enough.

The paper also aims to eliminate some confusions on using time stamp values (e.g. calendar dates, parts of it or replacement values) when operating on time series containing business data (e.g. sales amount recognized as a factor of direct influence for the level of certain financial indicators such as the operating income).

The research methodology

The source data for the examples presented in this paper come from two Microsoft samples databases. The first set of examples was created starting from an Access database file called foodmart (sites.google.com/.../supp4excel2datamining) originally available on the installation CD of a previous version of Microsoft (MS) SQL Server. The second one is from a MS SQL Server sample database called „AdventureWorksDW2012_Data.mdf” already installed and prepared for use inside a Windows 8.1 32 bits virtual machine (y2u.be/Xs2SWtBqdzI) that we have used for this article. This machine benefited from the Microsoft Imagine / formerly Dream Spark educational software license for all applications installed inside and it was optimized for Oracle Virtual Box. In fact SQL Server 2012 (or 2008) is a prerequisite for the installation of the Excel Data Mining add-in which is detailed in the second video tutorial (playlist mentioned below). Although they serve for building the examples and related video support materials (tutorials – the playlist created by the authors and available at goo.gl/JDDtFp), such data only have a guide purpose in this research with high applicative nature, the similarities to reality being merely coincidental.

From intuitive patterns to deep analysis starting from simple sources of data as tables

The first example we have chosen was meant to classify by generating a decision tree where the estimated variable was a categorical one with two possible values (house_owner: Yes or No – Y or N)
depending on some other fields (see figure 1) containing information about customers (an export to Excel from the customer table in the foodmart database).

Figure 1. Example of export followed by the use of classification option of the Excel Data Mining add-in and the configuration of the input fields

The results (fig.2) of this classification above by using default settings (Microsoft Decision Trees algorithm and 30 percent of data for testing) consist in: (1) a decision tree and (2) a dependency network that indicate the most important variables that influence the houseowner value, namely marital_status (married or single – M or S) and yearly_income (eight thresholds: '$10K - $30K', '$30K - $50K', '$50K - $70K', '$70K - $90K', '$90K - $110K', '$110K - $130K', '$130K - $150K', '$150K +'), in this order of importance.
Figure 2. Example of result of classification starting from data in a table with customers and made by using the Microsoft Decision Trees algorithm

![Diagram of decision tree with variables like marital_status and yearly_income and outcomes like buyer or not buyer.]

Source: The video tutorial created by the authors: y2u.be/Nx9xqCX1DjY

As seen above (left side of fig.2) the branches that indicate a higher probability for Yes (Y – houseowner) are darker, the rest of them being colored with a lighter shade. We can also observe that the houseowner as a variable depends essentially on the marital_status (right side of fig.2 – the slide bar on Strongest Links) and then on the yearly_income (the slide bar on All Links). And that can also be deduced directly from the decision tree in which the node closest to the root expresses a test (inf.ucv.ro) corresponding to the marital_status attribute. When clicking on marital_status=M (terminal node) we have got a probability more than 74% in all ten tests we have done in the same configuration (input columns, column to analyze, algorithm, percentage of data for testing).

Figure 3. Examples of discriminative analysis after applying the logistic regression (profs.info.uaic.ro) for the same conditions above and specifying those two already identified major impact input variables and some of their values

![Logistic regression results with variables like state_province and city, and outcomes like yes or no for the presence of a variable.]

Source: The video tutorial created by the authors: y2u.be/-6jzQuyTilo
In the previous images (fig.3) we tried to show how we have predicted probabilities that the customers fall into those two categories of the binary response (onlinecourses.science.psu.edu): house owner or not, depending on some explanatory variables and their values. We have done the discriminative analysis partially captured above (fig.3) starting from another algorithm, namely Logistic Regression implemented by Microsoft using a variation of the Neural Network algorithm (msdn.microsoft.com/.../ms174828.aspx) which is easier to train.

Figure 4. Example of results of the “Copy to Excel” functionality

Source: The video tutorials created by the authors: y2u.be/Nx9xqCX1DjY and y2u.be/-6jzQuyTjlo

The “Copy to Excel” functionality helped us to send the results back to Excel as new sheets containing screen shots (left side of fig.4 for decision trees) or most importantly data sets with visual effects usually involving conditional formatting done automatically (e.g. discriminative analysis based on logistic regression - right side of fig.4).

Based on the results described above (figs.1-4) on can develop similar examples to address also the problem of customer classification (corresponding to the acceptance / maintaining phase of the audit approach) into one of two categories: acceptable / unacceptable, starting from a validated log of such decisions, in tabular format and including many other descriptive attributes (geography, industry, average number of employees, turnover, evolutions of certain indicators, amount of fee, etc.).
Cumulating historical data and using descriptive fields from many tables of a database

The dynamic and interactive reports responding to many information needs that we are so familiar with as well as the older static ones as snapshots of information at precise moments and generating more questions than answers (Rasmussen et.al., 2002) may use both current and historical data. The 1st category is represented by data from Transaction Processing Systems (TPS) and commonly referring to the current year while the second essentially means data involving a larger period as time reference. The proportion of using those two categories essentially depends on decisional needs (at operational, tactical or strategic level). For minimizing the redundancy and dependency of data or because of storage space and write speed needs (deshpande.mit.edu) the schema of a traditional relational data source is usually thought as many tables obtained by applying the principles of normalization (w3schools.in). Moreover, because of further performance reasons (read, respectively write speed needs) historical data must be separated from current data. Both categories essentially include records from transaction tables (e.g. expenses, sales, exams, etc.) the difference being made by the value of the time stamp. That explains why those tables loaded only with historical data are being renamed with a time indication, archived and separated from the rest of the transactional system in order to improve its operational (current) performance. When needing large amounts of historical data for analyzes based on ad-hoc queries the systems must do vice versa by aggregating into a single table (source for a fact table in a data warehouse) all the records from the historical archives of the transaction tables (of the same type as the resulting one). In most cases, that generates the advantage of an increased potential to identify patterns. But it also comes with difficulties related to putting data together in a common and consistent format especially when the applications and the structure of the data source have also changed in time.

Figure 5. Gathering both transactional and descriptive data by using two MS Access SQL queries in cascade

Source: The video tutorial created by the authors: y2u.be/kTuYLuv3Eo
The figure 5 is presenting an example of inventory data gathering (applications including freight audit) in two major steps corresponding to two SQL queries in Microsoft Access: 1st - based on cumulating (UNION clause) the records from two transaction tables of the same type and corresponding to just two years (1997 and 1998) and adding an necessary id column (inventory_id with values generated automatically – AutoNumber type) in the resulting persistent table (INTO clause); 2nd - based on temporarily retrieving values of descriptive fields from all the tables related or suitable for a relation (fig.5 – INNER JOIN clause) with the one resulting from the 1st query above, namely inventory_fact. In this case the resulting tabular data consisting in the second set of just 11352 records won’t get into a persistent table of the database (a kind of de-normalization - searchoracle.techtarget.com) otherwise needed to save time at the expense of storage space and it will serve for external export (Excel) just after executing / running the query itself.

Figure 6. Results of consecutively using 2 Data Mining models - derived target field with only 2 values

Source: The video tutorials created by the authors: y2u.be/4nOMMRoC2BU and y2u.be/wce_aoTTsbw
Moreover, for speed of design reasons we have chosen all source fields without selecting them explicitly but indicating that by using the most flexible wildcard character, asterisk / “*”, after the table name (fig.5), both SQL and design mode (techrepublic.com). For the same reasons above the new derived column needed for analysis (fig.6 - output attribute for both models: classification-top and logistic regression-bottom) was then defined directly in Excel by using the IF function (top of fig.6, ord<>shp as 1 or 0 meaning that units_ordered and units_shipped are different compared to each other or equal).

**Association rules for identifying behavioral patterns**

In the theory and practice of data warehouses and multidimensional modeling the examples below reminds of the “snowflake” schema meaning that the source for a dimension (perspective of analysis based on descriptive columns organized in hierarchies) is not represented by just a single table but many related ones (in one-to-many relations: e.g. product category, product subcategory, and product – fig.7) able to support the analysis with more than just one descriptive field per dimension. In order to be able to apply the association rules algorithm in this case below we have also needed repetitive values for the SalesOrderNumber field to be associated to different product categories / subcategories / names.

Figure 7. Gathering both types of data: transactional about sales and descriptive about products by using a single MS SQL Server query

```
SELECT [SalesOrderNumber], [EnglishProductCategoryName],
[EnglishProductSubcategoryName], [EnglishProductName], [UnitPrice]
FROM [dbo].[FactInternetSales], [dbo].[DimProductCategory], [dbo].
[DimProductSubcategory], [dbo].[DimProduct]
WHERE [dbo].[FactInternetSales].[ProductKey]=[dbo].[DimProduct].[ProductKey] AND
[dbo].[DimProduct].[ProductSubcategoryKey]=[dbo].[DimProductSubcategory].
[ProductSubcategoryKey] AND [dbo].[DimProductSubcategory].[ProductCategoryKey]=
[dbo].[DimProductCategory].[ProductCategoryKey]
```
The main reason for gathering those descriptive data residing in multiple tables from the database in the example above (fig.7) is to determine association rules type “If I buy the product X, I will buy the product Y too.” in the purchasing behavior (FactInternetSales source table) and the most important dependencies (fig.8).

From the results in figs. 7 and 8 we can understand why the applications of the algorithms able to identify association rules can contribute to audit and fraud detection and prevention. As example, if the set of inputs would have attributes such as: Claim id, Insurance type, Name of the insurance product, Name of the insured person, Insurer, Name of the examiner agent and Solution (total or partial loss and reject) and the algorithm would identify “IF Casco insurance, Insured person X and Examiner agent Y THEN total loss” as association with high probability and importance, it would not necessarily mean a fraud alarm but it would worth at least the effort to investigate further.

Figure 8. Results of applying the Microsoft Association Rules algorithms (the associate option of the add-in)

Source: The video tutorial created by the authors: y2u.be/3_8E01hnSD0
Forecasting starting from aggregated historical data

For more historical data than in the previous example (fig.5) we have considered to create a special forecasting scenario closer to reality. We have started from scratch with a new example involving data on 36 months in four calendar years, this time by using a simple SQL query on a single table but with ORDER BY and GROUP BY clauses for sorted results and aggregations meaning computing aggregated values as: sums, averages, total counts, counts for a specified condition and so on. In our case those were averages on every month of an year combined into a single numerical field derived / composed by passing from left to right in the specific order: years to months corresponding to larger to smaller units (fig.9 - just like in Microsoft’s data sample which is provided when installing the Data Mining add-in).

Figure 9. Historical data aggregation (time stamp style from the Microsoft’s sample) using a SQL Server query (GROUP BY clause) followed by simply copying results to use them in forecasting (Excel’s Data Mining add-in)

Source: The video tutorials created by the authors: y2u.be/RjTwGROD0TI and y2u.be/qHJ3Zm3JBT4
After the steps described above (figure 9) and several other processing operations (fig.10) we will get to a set of data suitable for forecasting implemented by using the Microsoft Time Series algorithms as a combination of ARIMA (Auto-Regressive Integrated Moving Average - optimized for improving accuracy in long-term predictions) and ARTXP (Auto-Regression Trees with Cross-Prediction - optimized for predicting the next likely value in a time series - msdn.microsoft.com/.../bb677216.aspx) algorithms.

Figure 10. Deriving and explaining the correct time stamps as full dates internally stored (Excel) as numbers in right format data sources as support for undistorted forecasting

Source: The video tutorial created by the authors: y2u.be/e0SkDwG9mNY

We have also thought at automatically deriving the correct (the MM/DD/YYYY format translated into an integer number - fig.10) time stamp labels and we presented more details about their comparative behavior when getting trend line functions and forecasting results with this Excel Data Mining add-in (last three tutorials in the aforementioned playlist).

Support for querying persistent Data Mining models

First of all, persistent in this context refers to a model defined the way it will be deployed and stored on the server (SQL Server Analysis Services – a module other than the Database Engine) and available for querying (fig.11).

Figure 11. Data Mining eXtensions (DMX) sales prediction query examples (SQL Server Analysis Services) based on a DM time series model from a wrong format data source (text time stamp: 200815 /15th month in 2008)

Source: The video tutorial created by the authors: http://y2u.be/qHJ3Zm3JBT4
The **Data Mining** add-in available in Excel offers many advantages over the direct use of **SQL Server Analysis Service**. Among others, on can mention here: speed of use of Excel’s tabular environment and formula language, possibility of many exports / imports as / from spreadsheets starting from different database formats and to indirectly involve multiple source tables by using the Structured Query Language (SQL), the possibilities of exploiting the resulting structures and models directly (the “copy to Excel” option), by using queries (SQL DMX extension - figs.12 and 13) or programmatically (fig.12). Last two are conditioned by activating persistency when defining models (*use temporary model* option unchecked - fig.13 vs. figs.1 and 2).

Figure 12. Rough example of how to programmatically query a well-defined Data Mining model by using a DMX query in Visual Basic (VB).NET preceded by testing most of it on SQL Server Analysis Services

Source: The authors’ projection resulting after development attempts with VB and SQL Server
This last advantage reminds us that the programmatic generation (Airinei and Homocianu, 2009) of Excel dashboards and scorecards by using suggestive representations, warning indicators and dynamic formatting with support for BI has been simplified a lot since the 2007 version of the Microsoft Office suite. Combining that with the ability to programatically determine behavioral patterns and generate predicted values starting from high performance and easy to use tools such as this Data Mining add-in available for Office 2010, 2013 and 2016 promises much in terms of productivity. All these advances were defined after many years of using dedicated and now well-known technologies (e.g. SQL Server tested by authors since early 2000).

When it comes to spreadsheet products (dssresources.com//.../sshistory.html) such as: VisiCalc, Lotus 1-2-3, Microsoft Excel, Microsoft Works Spreadsheet, Sun Open Office Spreadsheets, Polaris Office Sheet, and Google Sheets the average experience of final users is up to decades. Furthermore the easiness of using these applications even just as interface instruments to connect to data from databases and data warehouses and display it was an objective reason to continue with testing the Data Mining component that led to making this article.

By using a way of reporting which identifies itself with a sequence of steps which borrow their names from those eighteen support tutorials and also some techniques previously defined namely: E2P4CAFR (Homocianu, 2015), ACCORD / CADRE (Homocianu and Airinei, September 2014) and S-DOT (Homocianu and Airinei, August 2014) on can reach in stages, but with a minimal number of steps to follow some representations that are dynamic, interactive, suggestive, based on causality and rooted in the current reality and in the history defined by data stored in the organization’s data sources.
**Conclusions**

We can conclude that the possibilities of the Excel Data Mining add-in component are above the expectations of a business analyst, offering the advantage of integrating identified classification patterns, association rules and predictions with the support for connectivity to various data formats, data validations, advanced graphical representations, geographical referencing, automatic conditional formatting and key performance indicators (KPI), pivot and power pivot tables and charts, automatic solving of optimization problems (solver) and the DAX (Data Analysis eXpressions) language together with the traditional formula language thereby increasing the chances of defining dashboards based on simulations, analyzes and Data Mining models truly useful for audit staff interested in performance monitoring.

We hope we have identified many real motivations to choose this Microsoft add-in for the Office suite as a near real time Data Mining tool, beyond many other recommendations available in the specialized literature and practice.

Beyond effective examples of working with well-known software applications available for a considerable range of users and providing advanced methods for analysis, query and representation of current and historical data particular to support tools for Data Mining and Business Intelligence, the paper also provides a brief theoretical description necessary in order to understand a rapid way of generating complex dynamic reports as dashboards based on analyzes and Data Mining models starting especially from sales and financial data.

The video tutorials developed by the authors, integrated in a playlist, and successively referenced in the paper prove the attempts to enrich the aforementioned way of reporting and to ensure the minimization of the number of steps required when trying to implement similar examples.

Overall, the article is trying and we hope that it succeeds to convey by clear examples some desirable traits as speed, simplicity, capacity of synthesis, transparency, flexibility and availability in reporting with support in Data Mining, as key elements of performance in preparing financial statements and supporting audit activities.

**References**

1. Airinei, D., *Depozite de date (Data Warehouses)*, Polirom Publishing, Iași, 2002, p.202
2. Airinei, D., Homocianu, D., *The Geographical Dimension of DSS Applications*, Scientific Annals of “Alexandru Ioan Cuza” University of Iași, Tome LVI, 2009, p.637-642
3. Chersan, I., Carp, M., Mironiuc, M., *Data mining – o provocare pentru auditorii financiari (Data Mining – A Challenge for Financial Auditors)*, Financial Audit Journal, No.10, 2013, pp.57-64
4. Cleland, D., King, W., R., *Competitive Business Intelligence Systems*, 6/1975, Business Horizons Journal, pp.19-28
5. Fraser, L., E., *Public Sector Audit - Business Integration and Causal Analysis*, Quality Audit Conference, February 26-27, 1998, Louisville, KY, vol.7
6. Homocianu, D., Airinei, D., *Business Intelligence facilities with applications in audit and financial reporting*, Financial Audit Journal, No. 9/2014, pp.17-29
7. Homocianu, D., Airinei, D., *Consolidating source data in audit reports*, Financial Audit Journal, No.8/2014, pp.10-19
8. Homocianu, D., *Excel Power Pivot's Applications in Audit and Financial Reports*, Financial Audit Journal, No.11/ 2015, pp.127-138
9. [http://bi-insider.com/business-intelligence/operational-bi-vs-strategic-bi/](http://bi-insider.com/business-intelligence/operational-bi-vs-strategic-bi/)
10. [http://dssresources.com/faq/index.php?action=artikel&id=174](http://dssresources.com/faq/index.php?action=artikel&id=174)
11. [http://dssresources.com/faq/index.php?action=artikel&id=199](http://dssresources.com/faq/index.php?action=artikel&id=199)
12. [http://dssresources.com/history/dsshistory.html](http://dssresources.com/history/dsshistory.html)
13. [http://dssresources.com/history/sshistory.html](http://dssresources.com/history/sshistory.html)
14. [http://inf.ucv.ro/documents/rstoean/5.%20Arbori%20de%20decizie.pdf](http://inf.ucv.ro/documents/rstoean/5.%20Arbori%20de%20decizie.pdf)
15. [http://goo.gl/JDDtFp](http://goo.gl/JDDtFp)
16. [http://profs.info.uaic.ro/~val/statistica/StatWork_10.pdf](http://profs.info.uaic.ro/~val/statistica/StatWork_10.pdf)
17. http://searchoracle.techtarget.com/tip/Optimizing-database-performance-part-2-Denormalization-and-clustering
18. http://searchsqlserver.techtarget.com/definition/data-mining
19. http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse
20. http://www.techrepublic.com/article/10-tips-for-using-wildcard-characters-in-microsoft-access-criteria-expressions/
21. http://www.w3schools.in/dbms/database-normalization/
22. http://y2u.be/Xs2SWtBqdzI
23. https://deshpande.mit.edu/portfolio/project/hybrid-dbms-optimized-read-intensive-applications
24. https://developers.google.com/apps-script/guides/sheets
25. https://msdn.microsoft.com/en-us/library/bb677216.aspx
26. https://msdn.microsoft.com/en-us/library/dn282385.aspx
27. https://msdn.microsoft.com/en-us/library/ms174828.aspx
28. https://msdn.microsoft.com/en-us/library/office/ee814737.aspx
29. https://onlinecourses.science.psu.edu/stat504/node/149
30. https://sites.google.com/site/supp4excel2datamining2017af/d
31. Inmon, W., H., Linstedt, D., Data architecture: A primer for the data scientist. Big Data, Data Warehouse and Data Vault, 2014, Morgan Kaufmann, MA, pp.336
32. Pearce, F., T., Business Intelligence Systems - need, development, and integration, vol.5 (2-3)/1976, Industrial Marketing Management Journal, pp.115-138
33. Rasmussen, N., H., Goldy, P., S., Solli, P., O., Financial Business Intelligence - Trends, Technology, Software Selection and Implementation, John Wiley and Sons, Inc., New York, 2002, pp.98-99
34. Sirikulvadhana, S., Data Mining As A Financial Auditing Tool (Thesis), Swedish School of Economics and Business Administration, 2002, pp.49-57, https://pdfs.semanticscholar.org/2612/f764664796f911e9ff9a79b7bb9de84bf16c.pdf
35. Vintilescu Belciug, A., Crețu, D., Gegea, C. (2010), Utilizarea tehniciilor de data mining ca metodă complementară în audit (Using data mining techniques as a complementary method in audit), Financial Audit Journal, No. 7, 2010, pp.30-35
36. Wang, J., Yang, J., G., S., Data Mining Techniques for Auditing Attest Function and Fraud Detection, Journal of Forensic & Investigative Accounting Vol. 1, Issue 1, January, 2009
Componenta Excel Data Mining. Aplicații in audit și raportări financiare

cercetător dr. Daniel HOMOCIANU
profesor universitar dr. Dinu AIRINEI

Departamentul de Cercetare, Facultatea de Economie și Administrarea Afacerilor (FEAA), Universitatea “Alexandru Ioan Cuza” din Iași (UAIC), daniel.homocianu@feaa.uaic.ro
Departamentul de Contabilitate, Informatică Economică și Statistică, FEAA, UAIC, adinu@uaic.ro

Rezumat
Argumentele de performanță în luarea deciziilor bazate pe date economice solicită uzuul un management bun al multiplelor formate de date și, de asemenea, viteză de procesare, flexibilitate, portabilitate, automatizare, putere de sugestie și ușurință de utilizare. Lucrarea vine cu idei teoretice și exemple practice în favoarea utilizării componentei Excel Data Mining pentru motivele menționate anterior. Cele mai multe exemple includ figuri legate la scenarii video construite de autori și parte a unei liste on-line interactive cu optsprezepe piese. Împreună ele contribuie la înțelegerea celor mai multe cerințe care trebuie îndeplinite pentru a avea exemple valide și rezultate utile.

Cuvinte cheie: date economice și financiare, foi de calcul, Data Mining (DM), example

Clasificare JEL: C61, D81, D83, M42

Introducere
Lucrarea începe de la o serie de tehnici utilizate de cele mai multe instrumente Data Mining pentru volume mari de date din baze de date și prezintă avantajele utilizării foiilor de calcul ca aplicații client (msdn.microsoft.com/.../dn282385.aspx). Acestea din urma sunt foarte familiare utilizatorilor finali și au o interferă care integrează limbaje de programare sau de scripting pentru aplicații de birou precum VBA însemnând Visual Basic pentru Aplicații în cazul Microsoft Excel (msdn.microsoft.com/.../ee814737.aspx) și Google Apps Script pentru foi de calcul Google Sheets (developers.google.com/.../sheets), multe funcții și facilități avansate de procesare, analiză, reprezentare și simulare toate bazate pe principii de interactivitate și dinamică cu impact considerabil asupra capacității utilizatorilor de a percepe, interpreta, înțelege și gestiona informații complexe în diferite situații. Conceptul de Data Mining înseamnă esențialmente identificarea supervizată a tipare nedescoperite și relații ascunse în seturi foarte mari de date (searchsqlserver.techtarget.com). Inmon care este un bine cunoscut guru în depozite de date (computerweekly.com) a dat una dintre cele mai concise definiții ale Data Mining (Inmon și Linstedt, 2014) și anume analiza unor cantități mari de date pentru a găsi tipare precum grupuri de înregistrări sau înregistrări și dependențe neobișnuite. Inițiativele Data Mining vin de obicei de la departamentele de marketing și vânzări cu amănuntul și sunt potrivite pentru organizații care au baze de date de dimensiuni foarte mari (Airinei, 2002). Acest concept este strâns legat de acela de sistem de asistare a deciziilor (Decision Support Systems - DSS) orientate spre date și de Business Intelligence (BI) – în special acela pentru obiective strategice (dssresources.com/...id=174) care solicită volume uriașe de date (bi-insider.com). Deși termenul de BI term este cunoscut ca și set de concepțe și metode de îmbunătățire a luării deciziilor care a apărut în anii ’90 (Howard Dresner din grupul Gartner - dssresources.com/.../dsshistory.html), probele din literatura de specialitate indică abordări cu 15 ani mai devere (Cleland și King, 1975; Pearce, 1976), care conțin referințe clare la BI, planificatorii / administratorii afacerii și manageri respectiv la luarea deciziilor.
După cum concluziile lui Dan Power (dssresources.com/...id=199), instrumentele Data Mining includ: raționamentul bazat pe caz, vizualizarea datelor (în special grafice, arbori și grupuri), interogări și analize vagi (fuzzy), algoritmi genetici și rețele neuronale.

Începând de acum câțiva ani suntem martori implementării acestui concept și modelelor asociate nu doar în aplicațiile dedicate sistemelor de gestiune a bazelor și depozitelor de date, dar și în module ale aplicațiilor tip foi de calcul care le folosesc așa cum se sugerează chiar din titlul lucrării. Pare evident când ne gândim că asemenea produse dedicate au permis construirea de structuri și modele DM chiar plecând de la un singur tabel (uzual ca agregare a mai multor tabele dintr-o bază de date).

Aplicabilitatea elementelor teoretice și practice din acest articol în audit, în special cel al performanței (Fraser, 1998) și raportări financiare este justificat plecând de la o nevoie specifică de a valoriza structurile existente de date (adesea înregistrări în tabele și tabele în baze de date) și de a obține rapid și cu cost minim rapoarte care pot prezenta informații clare despre cauzalitate în relație cu eficacitatea rezultate efective / estimate comparate cu cele propuse) și eficiența (resurse consumate comparate cu rezultate obținute / estimate).

Exemplele concrete din acest articol susțin anumite concluzii deșprinse din analiza literaturii de specialitate și anume: utilizarea abordării misiunilor de audit folosind tehnici data mining (Vintilescu Belciug et.al., 2010) complementar metodelor clasice de analiză a riscurilor și intervenției la fața locului, consacramentul unor zone posibile de integrare a data mining cu procesele de audit grupate pe faze (Sirikulvadhana, 2002) precum: planificarea, execuția, documentarea și finalizarea sau pe exemple specifice (Wang si Yang, 2009) ca: rețele neuronale pentru: evaluarea riscului, găsirea erorilor și a fraudelor, determinarea preocupărilor pentru continuitatea activității unei firme, evaluarea dificultăților financiare, precum și realizarea de predicții de faliment și arborii de decizie pentru: analiza de faliment, de faliment bancar și a riscului de credit), progresul celor mai noi instrumente software care implementează algoritmi data mining și faptul că mulți utilizatori le considerau până nu demult nu foarte prietenoase (Chersan et.al., 2013) și necesitând abilități tehnice destul de avansate.

Lucrarea își propune și eliminarea unor confuzii privind utilizarea valorilor tip ștampilă temporală (datatea / estimate) în audit, precum și realizarea de predicții de faliment a bază de date a mai multor tabele în baze de date, dar și în module ale aplicațiilor tip foi de calcul care le folosesc așa cum se sugerează chiar din titlul lucrării. Pare evident când ne gândim că asemenea produse dedicate au permis construirea de structuri și modele DM chiar plecând de la un singur tabel (uzual ca agregare a mai multor tabele dintr-o bază de date).

Aplicabilitatea elementelor teoretice și practice din acest articol în audit, în special cel al performanței (Fraser, 1998) și raportări financiare este justificat plecând de la o nevoie specifică de a valoriza structurile existente de date (adesea înregistrări în tabele și tabele în baze de date) și de a obține rapid și cu cost minim rapoarte care pot prezenta informații clare despre cauzalitate în relație cu eficacitatea rezultate efective / estimate comparate cu cele propuse) și eficiența (resurse consumate comparate cu rezultate obținute / estimate).

Exemplele concrete din acest articol susțin anumite concluzii deșprinse din analiza literaturii de specialitate și anume: utilizarea abordării misiunilor de audit folosind tehnici data mining (Vintilescu Belciug et.al., 2010) complementar metodelor clasice de analiză a riscurilor și intervenției la fața locului, consacramentul unor zone posibile de integrare a data mining cu procesele de audit grupate pe faze (Sirikulvadhana, 2002) precum: planificarea, execuția, documentarea și finalizarea sau pe exemple specifice (Wang si Yang, 2009) ca: rețele neuronale pentru: evaluarea riscului, găsirea erorilor și a fraudelor, determinarea preocupărilor pentru continuitatea activității unei firme, evaluarea dificultăților financiare, precum și realizarea de predicții de faliment și arborii de decizie pentru: analiza de faliment, de faliment bancar și a riscului de credit), progresul celor mai noi instrumente software care implementează algoritmi data mining și faptul că mulți utilizatori le considerau până nu demult nu foarte prietenoase (Chersan et.al., 2013) și necesitând abilități tehnice destul de avansate.

Lucrarea își propune și eliminarea unor confuzii privind utilizarea valorilor tip ștampilă temporală (date calendaristice, părți din ea sau valori înlocuitoare) la exploatarea serilor de timp cu date economice (de exemplu, volumul vânzărilor recunoscut ca factor de influență directă a nivelului anumitor indicatori financiari precum rezultatul din exploatare).

**Metodologia de cercetare**

Datele sursă pentru exemplele prezentate în această lucrare provin din două modere de baze de date Microsoft. Al doilea set de exemple a fost creat plecând de la un fișier bază de date Access denumit foodmart (sites.google.com/.../supp4excel2datamining) inițial disponibil pe CD-ul de instalare a unei versiuni anterioare de Microsoft (MS) SQL Server. Al doilea provine dintr-o bază de date mostră de tip MS SQL Server denumită „AdventureWorksDW2012_Data.md” deja instalată și pregătită pentru utilizare într-o mașină virtuală (y2u.be/Xs2SWtBqdzl) Windows 8.1 pe 32 biți pe care am folosit-o pentru acest articol. Aceasta mașină a beneficiat de licența Microsoft Imagine / fostă Dream Spark de software educațional pentru toate aplicațiile instalate în interiorul ei și a fost optimizată pentru Oracle Virtual Box. De fapt SQL Server 2012 (sau 2008) este o componentă necesară pentru instalarea Excel Data Mining care este detaliată în al doilea tutorial video (lista menționată mai jos). Deși servesc construirii de exemplu și materiale video suport corespunzătoare (tutoriale – lista creată de autori și disponibilă la adresa goo.gl/JDDtFp), asemenea date au doar caracter instructiv în această cercetare de natură pronunțat aplicativă, asemănările cu realitatea fiind doar o coincidență.

**De la tipare intuitive la o analiză profundă plecând de la surse simple de date precum tabelele**

Primul exemplu pe care l-am ales a fost menit să clasifice prin generarea unui arbore de decizie în care variabila estimată a fost una de tip categoric având două valori posibile (proprietar de casă / houseowner: Da/Yes sau Nu/No – Y sau N) în funcție de alte câmpuri (vezi figura 1) conținând informații despre clienți (un export în Excel din tabelul cu clienți customer din baza de date foodmart).
Rezultatele (fig.2) acestei clasificări de mai sus cu reglajele implicite (algoritmul Arbori Decizionali Microsoft și 30 procente din date pentru test) sunt: (1) un arbore decizional și (2) o rețea de dependențe ce indică cele mai importante variabile ce influențează valoarea atributului houseowner și anume: starea civilă / marital_status (căsătorit/married sau nesimțit – M sau S) și venitul anual / yearly_income (opt praguri in mii/K de dolari$/: 'S10K - $30K', '$30K - $50K', '$50K - $70K', '$70K - $90K', '$90K - $110K', '$110K - $130K', '$130K - $150K', '$150K +'), în această ordine a importanței.
Așa cum se observă în stânga fig.2 ramurile ce indică o probabilitate mai mare pentru Da (Y – homeowner) sunt mai închise, restul fiind colorate cu o nuanță mai deschisă. Putem de asemenea observa că variabila homeowner depinde esențial de marital_status (dreapta fig.2 – bara de derulare pe lagăturile cele mai puternice/Strongest Links) și apoi de yearly_income (bara pe toate lagăturile/All Links). Aceasta se poate deduce și direct din arbrele decizional în care un nod mai aproape de rădăcină exprimă un test (inf.ucv.ro) aferent atributului marital_status. La click pe marital_status='M' (nod terminal) am obținut o probabilitate peste 74% în toate cele zece teste făcute în aceeași configurație (coloane de intrare respectiv de analizat, algoritm, procentaj dat de test).

Figura 3. Exemple de analiză discriminatorie după aplicarea regresiei logistice (prof.s.info.uaic.ro) pentru aceleași condiții de mai sus și specificând cele două variabile de intrare cu impact major deja identificate și o serie de valori pentru ele

Sursa: Tutorialul video creat de autori: y2u.be/Nx9xqCX1DjY

Sursa: Tutorialul video creat de autori: y2u.be/-6jzQuyTjlo
In imaginile anterioare (fig.3) am încercat să punem în evidență cum anume am preconizat probabilitățile ca clienții să facă parte din cele 2 categorii de răspuns binar (onlinecourses.science.psu.edu): proprietar de casă sau nu, funcție de anumite variabile explicative și de valorile lor. Am realizat analiza discriminatorie partial redată mai sus (fig.3) plecând de la un alt algoritm și anume cel de regresie logistică, implementat de Microsoft folosind o variație a algoritmului de tip rețele neuronale (msdn.microsoft.com/.../ms174828.aspx) care este mai ușor de instruit.

Figura 4. Exemplu de rezultate ale funcționalității “Copiați în Excel” (“Copy to Excel”)
Cumularea datelor istorice și folosirea câmpurilor descriptive din tabelele bazei de date

Rapoartele dinamice și interactive care răspund la multe nevoi informaționale și cu care suntem atât de familiazat dar și cele statice mai vechi ca și fotografii ale informațiilor la momente precise și generând mai multe întrebări decât răspunsuri (Rasmussen et.al., 2002) pot utiliza atât date curente cât și istorice. Prima categorie este reprezentată de date din sistemele de procesare a tranzacțiilor (TPS) și care se referă în mod obișnuit la anul curent în timp ce a doua presupune esențial date implicând o perioadă mai mare de timp. Propoziția utilizării celor două categorii depinde esențial de nevoile decizionale (la nivel operațional, tactic sau strategic). Pentru minimizarea redundanței și a dependenței datelor sau din rațiuni de spațiu de stocare și de nevoie de viteză de scriere (deshpande.mit.edu) schema unei surse de date tradiționale de tip relațional este gândită uzual ca mai multe tabele obținute prin aplicarea principiilor normalizării (w3schools.in). Mai mult, din cauza unor motive suplimentare de performanță (nevoi de viteză de citire respectiv de scriere) datele istorice trebuie separate de cele curente. Ambele categorii includ esențial înregistrări din tabele cu tranzacții (de exemplu cheltuieli, vânzări, examene, etc.) diferența fiind dată de valoarea știmpelei temporale. Aceasta explică de ce aceste tabele încărcate doar cu date istorice sunt redenumite cu un indicativ de timp, arhiveate și separate de restul sistemului tranzacțional pentru a-i îmbunătăți performanța operațională (curentă). Când este nevoie de volume mari de date istorice pentru analize bazate pe interogări ad-hoc, sistemele trebuie să procedeze invers prin agregarea într-un singur tabel (ursă pentru un tabel de fapte într-un depozit de date) a tuturor înregistrărilor din arhivele istorice ale tabelelor cu tranzacții (de același tip ca cel rezumat). În cele mai multe cazuri, aceasta generează avantajul unui potențial crescut de identificare de tipare. Dar vine și cu dificultăți legate de punerea laolaltă a datelor într-un format comun și consistent în special atunci când aplicațiile și structura sursei de date s-au modificat și ele în timp.

Figura 5. Acumularea de date tranzactionale și descriptive folosind două interogări SQL MS Access în cascadă

Sursa: Tutorialul video creat de autori: y2u.be/kTuYLuv3Eo
Figura 5 prezintă un exemplu de acumulare de date de inventar (aplicații inclusiv în auditul transportului de marfă) în doi pași majori corespunzând celor două interogări SQL în Microsoft Access: prima – bazându-se pe cumularea (clauza UNION) înregistrărilor din două tabele cu tranzactii de același tip și corespunzând doar celor doi ani (1997 și 1998) respectiv pe adăugarea unei coloane id necesare (inventory_id cu valori generate automat – de tip AutoNumber) în tabelul persistent rezultat (clauza INTO); a doua – bazându-se pe extragerea temporară a valorilor câmpurilor descriptive din toate tabelele aflate în legătură sau potrivite pentru o legătură (fig.5 – clauza INNER JOIN) cu cel care rezultă din prima interogare de mai sus, și anume inventory_fact. În acest caz datele tabelare rezultate constând în al doilea set de doar 11352 înregistrări nu vor intra într-un tabel persistent al bazei de date (un fel de denormalizare - searchoracle.techtarget.com) în alte situații necesar persistent al bazei de date în detrimentul spațiului de stocare și va servi pentru export extern (Excel) imediat după executarea / rularea interogării în sine.

Figura 6. Rezultatele folosirii consecutive a 2 modele Data Mining - câmp țintă derivat cu 2 valori posibile

Sursa: Tutorialele video create de autori: y2u.be/4nOMMRoC2BU și y2u.be/wce_a0TTsbw
Mai mult, din motive legate de viteză de proiectare am ales toate câmpurile sursă fără selectarea lor explicită dar indicându-le folosind cel mai flexibil caracter de căutare și anume asteriscul / “*”, după denumirea tabelului (fig.5), atât în modul SQL cât și în cel de proiectare asistată / design (techrepublic.com). Din aceleași motive de mai sus, noua coloană derivată necesară pentru analize (fig.6 - atribut de ieșire pentru ambele modele: clasificare-sus și regresie logistică-jos) a fost ulterior definită direct în Excel folosind funcţia IF (partea superioară a fig.6, ord<>shp ca 1 sau 0 adică unitățile comandate / units_ordered și unitățile expediate / units_shipped sunt diferite respectiv egale).

Reguli de asociere pentru identificarea de modele comportamentale

În teoria și practica depozitelor de date și a modelării multidimensionale exemplele de mai jos amintesc de schema “fulg de nea” însemnând că sursa unei dimensiuni (perspectivă de analiză bazată pe coloane descriptive organizate în ierarhii) nu este reprezentată doar de un singur tabel ci de mai multe aflate în legătură (relații „unu la multe”: de exemplu categoria de produs / product category, subcategoria / product subcategory și produsul / product – fig.7) și capabile să asigure suportul unei analize pentru mai mult decât un singur câmp descriptiv pe dimensiune. Pentru a putea aplica algoritmul regulilor de asociere în acest caz de mai jos, am avut nevoie și de valori repetitive pentru câmpul Numărul comenzii de vânzare / SalesOrderNumber de asociat la diferite categorii, subcategorii sau nume de produse.

Figura 7. Acumularea ambelor tipuri de date: tranzaționale despre vânzări și descriptive despre produse folosind o singură interogare MS SQL Server

```
SELECT [SalesOrderNumber], [EnglishProductCategoryName],
[EnglishProductSubcategoryName], [EnglishProductName], [UnitPrice]
FROM [dbo].[FactInternetSales], [dbo].[DimProductCategory], [dbo].
[DimProductSubcategory], [dbo].[DimProduct]
WHERE [dbo].[FactInternetSales].[ProductKey]=[dbo].[DimProduct].[ProductKey] AND
[dbo].[DimProduct].[ProductSubcategoryKey]=[dbo].[DimProductSubcategory].[ProductSubcategoryKey] AND
[dbo].[DimProductSubcategory].[ProductCategoryKey]=[dbo].[DimProductCategory].[ProductCategoryKey]
```

Sursa: Tutorialul video creat de autori: y2u.be/2rW2wK77HD8
Motivul principal pentru acumularea acestor date descriptive din multiplele tabele ale bazei de date din exemplul de mai sus (fig.7) este de a determina reguli de asociere în comportamentul de cumpărare de tipul “Dacă cumpăr produsul X, voi cumpăra și produsul Y.” (tabelul sursă FactInternetSales) precum și cele mai importante dependențe (fig.8).

Rezultatele din fig.7 și 8 ne pot face să înțelegem de ce aplicațiile algoritmilor de identificare de reguli de asociere pot contribui în audit și detectarea și prevenirea fraudelor. Cu titlu de exemplu, dacă setul de date de intrare ar avea atribute precum: Identificator daună, Tip produs de asigurare, Nume asigurat, Asigurator, Nume agent constatator și Soluționare (daună totală sau parțială și respingere) iar algoritmul ar identifica “DACĂ asigurare Casco, Asigurat X și Agent constatator Y ATUNCI daună totală” ca asociere cu probabilitate și importanță mari, aceasta nu ar însemna neapărat o alarmă de fraudă dar ar merita măcar efortul de a face investigații suplimentare.

Figura 8. Rezultatele aplicării algoritmilor cu reguli de asociere Microsoft (opțiunea asociază / associate)
Prevederi plecând de la date istorice agregate

Pentru mai multe date istorice decât cele din exemplul precedent (fig.5) am luat în considerare crearea unui scenariu special de prevederi mai aproape de realitate. Am pornit de la zero cu un nou exemplu care solicită date pe 36 de luni din patru ani calendaristici, de data aceastea folosind o interogare simplă SQL pentru un singur tabel dar cu clauze ORDER BY și GROUP BY pentru rezultate sortate și aggregă însemnând calcularea de valori agregate ca: sume, valori medii, număr total de apariții, număr de apariții pentru o condiție specificată, etc. În cazul nostru acestea au fost medii lunare pe ani combinate într-un singur câmp numeric derivat / compus prin trecerea de la stânga la dreapta în ordinea specifică: de la ani la luni corespunzând cele de la unități mai mari la unități mai mici (fig.9 – exact ca în mostra de date a Microsoft care este furnizată la instalarea componentei Data Mining).

Figura 9. Agregare de date istorice (stilul ștampilei temporale din mostra Microsoft) cu o interogare SQL Server (clauza GROUP BY) și o simplă copiere a rezultatelor de folosit în previziune (componenta Excel Data Mining)

Sursa: Tutorialele video create de autori: y2u.be/RjTwGROD0TI și y2u.be/qHJ3Zm3JBT4
După pașii descriși mai sus (figura 9) și alte câteva operații de prelucrare (fig.10) vom ajunge la un set de date potrivit pentru previziuni implementate folosind algoritmul de serii de timp Microsoft ca o combinație a algoritmilor ARIMA (medie mobilă auto-regresivă și integrată - optimizată pentru creșterea acurateții în predicțiile pe termen lung) și ARTXP (arbore de auto-regresie cu predicție încrușită optimizată pentru estimarea următoarei valori probabile într-o serie de timp - msdn.microsoft.com/.../bb677216.aspx).

Figura 10. Derivarea și explicarea ștampilelor temporare corecte ca date complete stocate intern (Excel) ca și numere în surse de date în format potrivit ca suport pentru previziuni nedistorsionate

Sursa: Tutorialul video creat de autori: y2u.be/eO5KdwG9mNY

Ne-am gândit de asemenea la derivarea automată a etichetelor ștampilelor temporale corecte (formatul LL/ZZ/AAAA tradus într-un număr întreg - fig.10) și am prezentat mai multe detalii despre comportamentul lor comparativ la obținerea de funcții de trend și previziunea rezultatelor cu această componentă Excel Data Mining (ultimele trei tutoriale din lista menționată anterior).

**Suport pentru interogarea de modele Data Mining persistente**

În primul rând, persistent în acest context se referă la un model definit astfel încât să fie procesat și stocat pe server (SQL Server Analysis Services / servicii de analiză – modul diferit de motorul de baze de date / Database Engine) și disponibil pentru interogare (fig.11).

Figura 11. Exemple de interogări Data Mining eXtensions (DMX) de previziune a vânzărilor (SQL Server Analysis Services) pe baza unui model DM de serii de timp plecând de la o sursă de date în format greșit (ștampilă temporală text: 200815 / a 15-a lună în 2008)

Sursa: Tutorialul video creat de autori: http://y2u.be/qHJ3Zm3JBT4
Componenta *Data Mining* din Excel oferă multe avantaje față de utilizarea directă a *SQL Server Analysis Service*. Printre altele, putem menționa aici: viteza de utilizare a mediului tabelar și a limbajului de formule din Excel, posibilitatea multor exporturi / importuri în / din noi de calcul tabelar plecând de la diferite formate de baze de date și de a implica indirect multiple tabele sursă folosind limbajul structurat de interogare (SQL), posibilitățile de exploatare a structurilor și modelelor obținute: direct (opțiunea *copy to Excel*), cu interogări (extensia DMX a SQL - fig.12 și 13) sau programatic (fig.12). Ultimele două sunt condiționate de activarea persistentei la definirea modelurilor (opțiunea de folosire a unui model temporar / use temporary model nebijată - fig.13 comparativ cu fig.1 și 2).

Figura 12. Exemplu brut de interogare programatică a unui model *Data Mining* bine definit folosind o interogare DMX în Visual Basic (VB).NET precedată de testarea a mare parte din ea în *SQL Server Analysis Services*

*Sursa: Proiecția autorilor obținută după încercări de dezvoltare cu VB și SQL Server*
Figura 13. Exemplu de interogare de predicție DMX (proprietar casă / houseowner) pe baza unui model de clasificare persistent cu arbori de decizie (visualizarea generică a conținutului în fundal)

Sursa: Proiecția autorilor obținută după încercări de dezvoltare cu SQL Server

Acest ultim avantaj ne amintește că generarea programatică (Airinei și Homocianu, 2009) de tablouri de bord și tabele de scoruri Excel folosind reprezentări sugestive, indicatori de alertă și formatele dinamice cu suport pentru BI-s-a simplificat mult începând cu versiunea 2007 a pachetului Microsoft Office. Combinarea acesteia cu abilitatea de a determina programat modele comportamentale și de a genera valori previzionate plecând de la instrumente de înaltă performanță și uşor de folosit precum această componentă Data Mining disponibilă pentru Office 2010, 2013 și 2016 promite mult în termeni de productivitate. Toate aceste progrese au fost definite după mai mulți ani de utilizare de tehnologii dedicate și acum bine-cunoscute (de exemplu SQL Server testat de autori la începutul anilor 2000). Pentru produsele de tip foioșă de calcul tabelar (dssresources.com/.../sshistory.html) precum: VisiCalc, Lotus 1-2-3, Microsoft Excel, Microsoft Works Spreadsheet, Sun Open Office Spreadsheets, Polaris Office Sheet și Google Sheets experiența medie a utilizatorilor finali este de până la zeci de ani. Mai mult, ușurința de utilizare a acestor aplicații chiar și doar ca instrumente de interfață pentru conectarea la datele din baze de date și depozite de date și afișarea lor a fost un motiv obiectiv pentru a continua cu testarea componentei Data Mining care a condus la conceperea acestui articol. Folosind o modalitate de raportare care se identifică cu o sevență de pași care își împrumută numele de la cele optisprezece tutoriale suport și, de asemenea, unele tehnici anterior definite și anume: E2P4CAFR (Homocianu, 2015), ACCORD / CADRE (Homocianu și Airinei, septembrie 2014) și S-DOT (Homocianu și Airinei, august 2014) se poate ajunge în etape, dar cu un număr minim de pași de urmat la anumite reprezentări care sunt dinamice, interactive, sugestive, bazate pe cauzalitate și înradăcinate în realitatea curentă și în istoria definită de datele stocate în sursele de date ale organizației.
Concluzii

Putem concluziona că posibilitățile componentei Excel Data Mining sunt peste așteptările unui analist de afaceri, oferind avantajul integrării tiparelor de clasificare identificate, a regulilor de asociere și predicțiilor cu suportul pentru conectivitate la formate variate de date, validările de data, reprezentările grafice avansate, referențieri geografice, formatările condiționate automate și indicatorii cheie de performanță (KPI), tabelele și graficile de tip pivot și power pivot, rezolvarea automată de probleme de optimizare (solver) și limbașul DAX (expresii de analiză a datelor / Data Analysis eXpressions) împreună cu limbașul tradițional de formule, sporiț astfel sănsele definirii de tablouri de bord fundamentate pe simulări, analize și modele Data Mining cu adevărat utile pentru personalul de audit interesat de monitorizarea performanței.

Sperăm că am identificat multe motivații reale pentru alegerea acestei componente Microsoft pentru pachetul Office ca un instrument Data Mining aproape în timp real, dincolo de multe alte recomandări disponibile în literatură și practica de specialitate.

Dincolo de exemple substanțiale de lucru cu aplicații software bine cunoscute, disponibile pentru o gamă largă de utilizatori și care asigură metode avansate de analiză, interogare și reprezentare a datelor curente și istorice specifice instrumentelor suport pentru Data Mining și Business Intelligence, lucrarea furnizează și o scurtă descriere teoretică necesară înțelegerii unei modalități rapide de generare de rapoarte complexe și dinamice precum tablourile de bord fundamentate pe analize și modele Data Mining plecând în special de la date despre vânzări și financiare.

Tutorialele video dezvoltate de autori, integrate într-o listă și referite succesiv în această lucrare demonstrază încercările de îmbogățire a modului de raportare anterior menționat și de a asigura minimizarea numărului de pași necesari atunci când se încearcă implementarea de exemple similare. În ansamblu, articolul încercă și sperăm că reușește să transmită prin exemple clare anumite caracteristici dorite cum ar fi: viteză, simplitate, capacitate de sinteză, transparentă, flexibilitate și disponibilitate în raportarea fundamentată pe Data Mining, ca și elemente cheie de performanță în pregătirea situațiilor financiare și sprijinirea activităților de audit.

Referințe

1. Airinei, D., Depozite de date, Editura Polirom, Iași, 2002, p.202
2. Airinei, D., Homocianu, D., The Geographical Dimension of DSS Applications, Analele Științifice ale Universității “Alexandru Ioan Cuza” din Iași, Tome LVI, 2009, p.637-642
3. Chersan, I., Carp, M., Mironiuc, M., Data mining – o provocare pentru auditorii financiari, Revista „Audit Financiar” nr. 10, 2013, p.57-64
4. Cleland, D., I., King, W., R., Competitive Business Intelligence Systems, 6/1975, Business Horizons Journal, pp.19-28
5. Fraser, L., E., Public Sector Audit - Business Integration and Causal Analysis, Quality Audit Conference, February 26-27, 1998, Louisville, KY, vol.7
6. Homocianu, D., Airinei, D., Business Intelligence facilities with applications in audit and financial reporting, Revista „Audit Financiar”, nr. 9/2014, p.17-29
7. Homocianu, D., Airinei, D., Consolidating source data in audit reports, Revista „Audit Financiar”, Nr.8/2014, p.10-19
8. Homocianu, D., Excel Power Pivot’s Applications in Audit and Financial Reports, Revista „Audit Financiar”, Nr.11/ 2015, p.127-138
9. http://bi-insider.com/business-intelligence/operational-bi-vs-strategic-bi/
10. http://dssresources.com/faq/index.php?action=artikel&id=174
11. http://dssresources.com/faq/index.php?action=artikel&id=199
12. http://dssresources.com/history/dsshistory.html
13. http://dssresources.com/history/sshistory.html
14. http://inf.ucv.ro/documents/rstoean/5.%20Arbori%20de%20decizie.pdf
15. http://goo.gl/JfD4fF
16. http://profs.info.uaic.ro/~val/statistica/StatWork_10.pdf
17. http://searchoracle.techtarget.com/tip/Optimizing-database-performance-part-2-Denormalization-and-clustering
18. http://searchsqlserver.techtarget.com/definition/data-mining
19. http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse
20. http://www.techrepublic.com/article/10-tips-for-using-wildcard-characters-in-microsoft-access-criteria-expressions/
21. http://www.w3schools.in/dbms/database-normalization/
22. http://y2u.be/Xs2SWtBqdZI
23. https://deshpande.mit.edu/portfolio/project/hybrid-dbms-optimized-read-intensive-applications
24. https://developers.google.com/apps-script/guides/sheets
25. https://msdn.microsoft.com/en-us/library/bb677216.aspx
26. https://msdn.microsoft.com/en-us/library/dn282385.aspx
27. https://msdn.microsoft.com/en-us/library/ms174828.aspx
28. https://msdn.microsoft.com/en-us/library/office/ee814737.aspx
29. https://onlinecourses.science.psu.edu/stat504/node/149
30. https://sites.google.com/site/supp4excel2datamining2017af/d
31. Inmon, W., H., Linstedt, D., Data architecture: A primer for the data scientist. Big Data, Data Warehouse and Data Vault, 2014, Morgan Kaufmann, MA, p.336
32. Pearce, F., T., Business Intelligence Systems - need, development, and integration, vol.5 (2-3)/1976, Industrial Marketing Management Journal, p.115-138
33. Rasmussen, N., H., Goldy, P., S., Solli, P., O., Financial Business Intelligence - Trends, Technology, Software Selection and Implementation, John Wiley and Sons, Inc., New York, 2002, p.98-99
34. Sirikulvadhana, S., Data Mining As A Financial Auditing Tool (Thesis), Swedish School of Economics and Business Administration, 2002, pp.49-57, https://pdfs.semanticscholar.org/2612/f764664796f911e9ff9a79b7bb9de84bf16c.pdf
35. Vintilescu Belciug, A., Crețu, D., Gegea, C. (2010), Utilizarea tehnicii de data mining ca metodă complementară în audit, Revista „Audit Financiar” nr. 7, 2010, p.30-35
36. Wang, J., Yang, J., G., S., Data Mining Techniques for Auditing Attest Function and Fraud Detection, Journal of Forensic & Investigative Accounting Vol. 1, Issue 1, January, 2009