FURTHER CONSIDERATIONS ON SPREADSHEET-BASED AUTOMATIC TREND LINES

DANIEL HOMOCIANU
PH. D RESEARCHER, ALEXANDRU IOAN CUZA UNIVERSITY OF IASI, FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION, DEPARTMENT OF RESEARCH
e-mail: dan.homocianu@gmail.com; daniel.homocianu@feaa.uaic.ro

Abstract

Most of the nowadays business applications working with data sets allow exports to the spreadsheet format. This fact is related to the experience of common business users with such products and to the possibility to couple what they have with something containing many models, functions and possibilities to process and represent data, by that getting something in dynamics and much more than a simple static less useful report.

The purpose of Business Intelligence is to identify clusters, profiles, association rules, decision trees and many other patterns or even behaviours, but also to generate alerts for exceptions, determine trends and make predictions about the future based on historical data. In this context, the paper shows some practical results obtained after testing both the automatic creation of scatter charts and trend lines corresponding to the user's preferences and the automatic suggesting of the most appropriate trend for the tested data mostly based on the statistical measure of how close they are to the regression function.

Key words: *Spreadsheets, Business Intelligence (BI), Microsoft (MS) Office Excel trends, Google Sheets Trends*

JEL Clasiffication: *C130, D810, D830*

1. Introduction

The spreadsheet type applications (e.g. MS Office Excel, Polaris Office Sheet, Google Sheets) are already known for their power to optimize decisional problems (e.g. custom defined decision matrices, trend lines [1], simple and multiple [2] regression), for solvers, for interaction with web/data mining tools [3] and for many other facilities offered to common business users.

As set of concepts and methods to improve decision-making [4] mentioned since the 90's, Business Intelligence and its insight making features [5] usually rely on data warehouse (DW), data mining, analytic and query tools. Starting a few years ago, to use many of these tool features in MS Excel and exploit the spreadsheet well-connected format [6] with its newest ability to make data geo-referencing [7] is something possible via Power Pivot, Power Query, Power View & Power Map. Moreover by its later integration with SharePoint [8], Excel follows the idea that the content must be organized by content management systems and applications in order to ensure its effective management and easy retrieval and delivery in different formats [9].

2. Data to test by getting different trend functions with different levels of fit

Because the preparation of most business decision making scenarios involving historical data (e.g. fig.1, where *OrdYM* [10] means order's year and month) is usually time consuming in terms of finding out the optimum choice for certain limits of parameters, we believe a minimum support is needed, at least under the form of automatic suggestions.

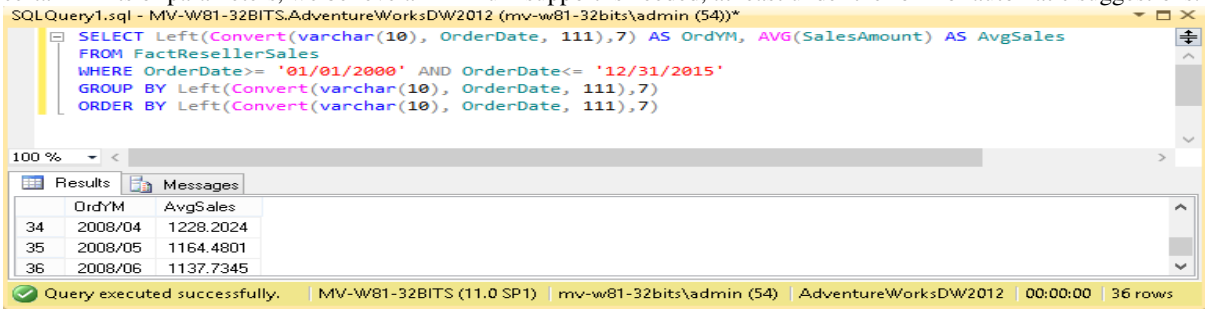


Figure1. Query based intermediary source data - 36 resulting rows with average values for prediction using time limits

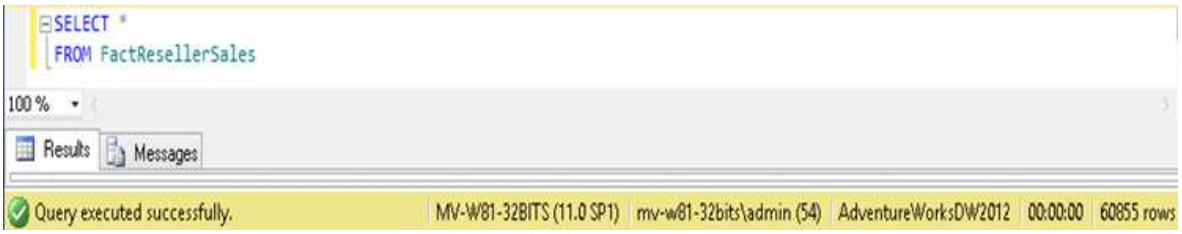


Figure 2. The primary data source (a 60855 rows table in a MS SQL Server database) of the previous GROUP BY query (fig.1) to be adapted for dynamic behavior in a Visual Basic prototype [11]

Instead of letting the user test those 5 types of equations available in Excel trend line's options (except Moving Average which does not allow forward prediction) and those 5 options for the polynomial degree (values from 2 to 6), we can automatically get R-squared (R-sq) by using MS Visual Basic (VB) for each of those 9 cases and suggest the one with the maximum value [12]. The examples used in the paper refer to sales data from a MS sample database named AdventureWorksDW2012 (see figs.1 and 2), but the automatic calculations can be extended to any two columns (X and Y values) user defined datasets.

```
Imports System.Data.SqlClient
Imports Microsoft.Office.Interop
Public Class Form1
    Dim conn_str = "server=MV-W81-32BITS;database=AdventureWorksDW2012; integrated security=sspi"
    Dim q = "SELECT Left(Convert(varchar(10), OrderDate, 111),7) AS OrdYM, " _
            & "AVG(SalesAmount) AS AvgSales " _
            & "FROM FactResellerSales "
    Dim lc = "WHERE OrderDate>="
    Dim rc = " AND OrderDate<="
    Dim goby = " GROUP BY Left(Convert(varchar(10), OrderDate, 111),7) " _
             & "ORDER BY Left(Convert(varchar(10), OrderDate, 111),7)"
    Dim tbl = "FactResellerSales"
```

Figure 3. Two imports & six global variables including some parts of the previous query tested in SQL Server and used in MS VB together with custom time limits both to suggest the best function and predict a future value by using it [13]

3. How to automatically generate a trend line by using VB and Excel?

As can be seen in figs.3 and 4, in order to automatically generate a scatter chart and a trend line first we will use a sequence of code lines (fig.4 - first 21 lines inside the rectangle) also needed when suggesting the most appropriate equation for our data (in front of the VB code lines presented in fig.5).

This VB code sequence is responsible for: (1.) creating a dynamic query (*lq*) made by linking a first fixed part (*q*) to the fixed part of its criteria (*lc* & *rc*), the variable one (values of the date time pickers *DTP1* & *DTP2*) and its Group By and Order By sections (*goby*); (2.) defining and initializing a SQL connection (*conn*), a connection string (*conn_str*), a data set (*ds*) and a data adapter (*da*) needed for automatic connection and retrieval of our data set after querying a table within a SQL Server database; (3.) defining and creating a new Excel.application (*oXL*) and workbook object (*oBook*); (4.) getting the number (*n*) of rows (minus 1) of our data set and (5.) calculating each item of a 2 * n corresponding matrix (*matr*).

Next 8 lines (fig.4) are meant to add a new sheet, the cells with chart's source data, a scatter line chart (size, type, source data and name of series as legend: Average Sales) and they also get the function type (the value of the *cbo* combo box into cboSel) and the number of periods for forward prediction (the value of the *tbo* text box into *fwd*).

Last 20 lines (fig.4) are trying to separately treat the polynomial case (extract the value of the *cbo_ordpoly* combo box into *op* if *cboSel=3*, meaning getting its degree in case of polynomial), create a trend line, get the property containing the R sq's value, fill the Excel's column C with consecutive values via the "For" loop, effectively extract the R-sq's value and the equation and grant the current user full control of the newly created workbook.

```
Private Sub Button3_Click(sender As Object, e As EventArgs) Handles Button3.Click
    Dim lq = q _
            & lc & "'" & DTP1.Value.ToShortDateString.ToString() & "'" _
            & rc & "'" & DTP2.Value.ToShortDateString.ToString() & "'" _
            & goby
    Dim conn = New SqlConnection(conn_str)
    conn.Open()
    Dim da = New SqlDataAdapter(lq, conn)
    Dim ds = New DataSet
    da.Fill(ds, tbl)
    Dim oXL As Object        ' Excel application
    Dim oBook As Object      ' Excel workbook
    Dim oSheet As Object     ' Excel Worksheet
    Dim oCh As Object        ' Excel Chart
    oXL = CreateObject("Excel.application")
    oBook = oXL.Workbooks.Add
    Dim n = ds.Tables(tbl).Rows.Count - 1
    Dim matr(n, 1)
    For i = 0 To n
        matr(i, 0) = ds.Tables(tbl).Rows(i).Item("OrdYM")
        matr(i, 1) = ds.Tables(tbl).Rows(i).Item("AvgSales")
    Next
    oSheet = oBook.Worksheets.Item(1)
    oSheet.Range("A1").Resize(n + 1, 2).Value = matr
    oCh = oSheet.ChartObjects.Add(150, 40, 800, 400).Chart
    oCh.ChartType = Excel.XlChartType.xlXYScatterLines
    oCh.SetSourceData(oSheet.Range("A1:B" & (n + 1).ToString))
    oCh.SeriesCollection(1).name = "Average Sales"
    Dim cboSel = Me.cbo.SelectedItem.ToString
    Dim fwd = Me.tbo.Text
    If cboSel = "3" Then
        Dim op As Integer
        op = Int32.Parse(Me.cbo_ordpoly.SelectedItem)
        oCh.SeriesCollection(1).Trendlines.Add(Type:=cboSel, order:=op, Forward:=fwd, Backward:=0, DisplayEquation:=False, DisplayRSquared:=True).select()
    Else
        oCh.SeriesCollection(1).Trendlines.Add(Type:=cboSel, Forward:=fwd, Backward:=0, DisplayEquation:=False, DisplayRSquared:=True).select()
    End If
    Dim R_sq = oCh.SeriesCollection(1).Trendlines(1).DataLabel.text
    For i = 0 To n
        oSheet.Range("C" & (i + 1)) = (i + 1).ToString
    Next
    oSheet.Range("D1") = R_sq
    oSheet.Range("E1") = Mid(R_sq, 6, Len(R_sq) - 5)
    oCh.SeriesCollection(1).Trendlines(1).DisplayRSquared = False
    oCh.SeriesCollection(1).Trendlines(1).DisplayEquation = True
    Dim T_eq = oCh.SeriesCollection(1).Trendlines(1).DataLabel.text
    oSheet.Range("F1") = T_eq
    oXL.Visible = True
    oXL.UserControl = True
End Sub
```

same sequence
behind
the **Suggest.. button**

Figure 4. The source code (behind the Excel prediction button) responsible for generating a simple regression based prediction (chosen type of equation and polynomial degree) starting from a custom time interval [14]

4. How to automatically suggest via hidden nonpersistent trend lines generated in Excel?

We are using the same initial code sequence explained above (first 21 lines in fig.4). Next 13 lines (fig.5) are used to define and set the values of a 2 columns on 9 rows matrix (*trend*) which contains the equation types, their

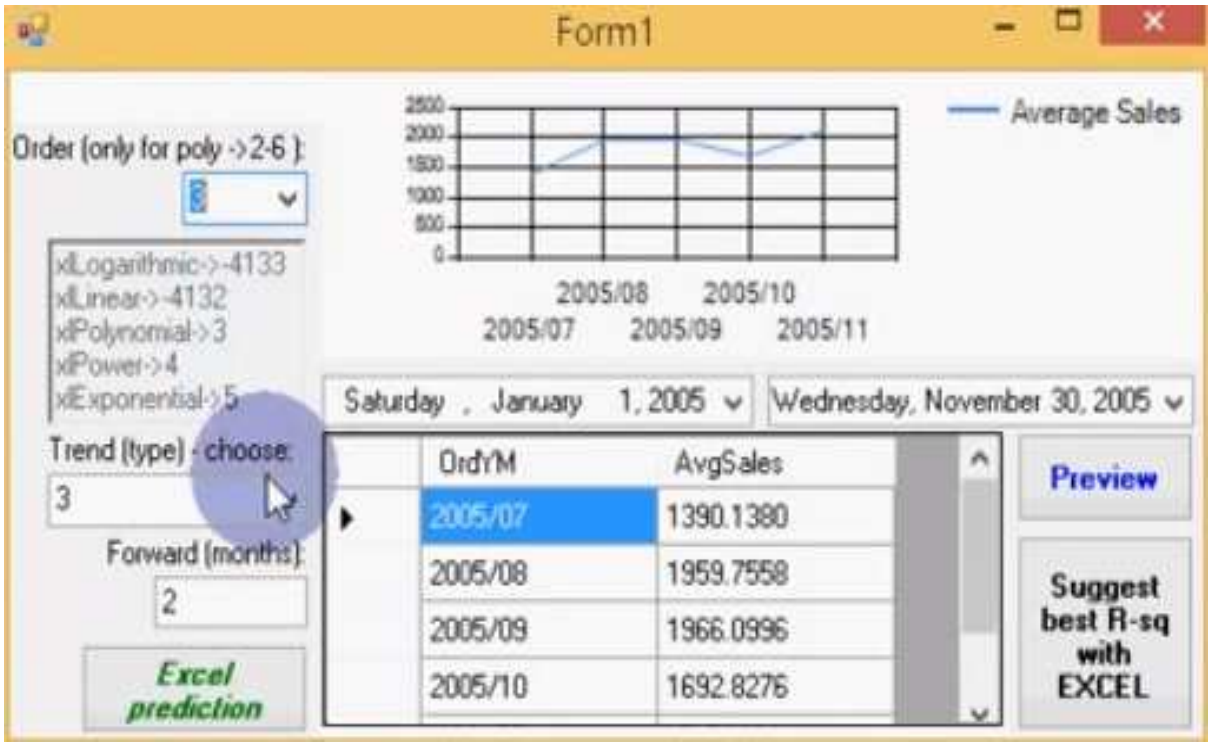
corresponding VB codes (-4133, -4132, 3, 4 or 5) & the order parameter (or the polynomial degree: 2, 3, 4, 5 or 6).

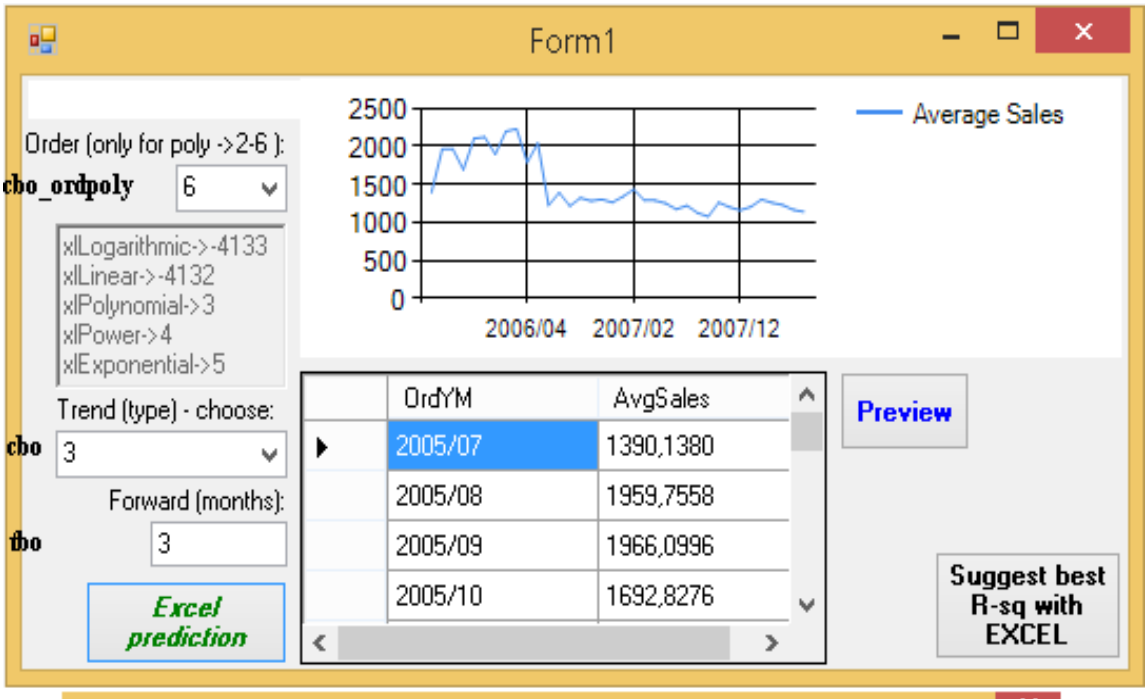
```
Dim trend(8, 1) As String
trend(0, 0) = "-4133"
trend(0, 1) = "Logarithmic"
trend(1, 0) = "-4132"
trend(1, 1) = "Linear"
For k = 2 To 6
 trend(k, 0) = "3"
 trend(k, 1) = "Polynomial" & k.ToString
Next
trend(7, 0) = "4"
trend(7, 1) = "Power"
trend(8, 0) = "5"
trend(8, 1) = "Exponential"
Dim o As Integer
Dim max_R_sq = ""
Dim mem_type As Integer
For k = 0 To 8
 oSheet = oBook.Worksheets.Add
 oSheet.Range("A1").Resize(n + 1, 2).Value = matr
 oCh = oSheet.ChartObjects.Add(150, 40, 800, 400).Chart
 oCh.ChartType = Excel.XlChartType.xlXYScatterLines
 oCh.SetSourceData(oSheet.Range("A1:B" & (n + 1).ToString))
 oCh.SeriesCollection(1).name = "Average Sales"
 If k >= 2 And k <= 6 Then
  o = Int32.Parse(Mid(trend(k, 1), 11, 1))
  oCh.SeriesCollection(1).Trendlines.Add(Type:=trend(k, 0), order:=o, Forward:=1, Backward:=0, DisplayEquation:=False, DisplayRSquared:=True).select()
 Else
  oCh.SeriesCollection(1).Trendlines.Add(Type:=trend(k, 0), Forward:=1, Backward:=0, DisplayEquation:=False, DisplayRSquared:=True).select()
 End If
 Dim R_sq = oCh.SeriesCollection(1).Trendlines(1).DataLabel.text
 R_sq = Mid(R_sq, 6, Len(R_sq) - 5)
 If R_sq > max_R_sq Then
  max_R_sq = R_sq
  mem_type = k
 End If
Next
MsgBox("Your Data indicate the " & trend(mem_type, 1).ToString & " trend from those 5 on the upper-left, with a max trust of: " & max_R_sq)
oBook.Close(savechanges:=False)
oXL.Quit()
End Sub
```

Figure 5. The source code (behind the Suggest.. button) returning the most appropriate type of equation for our data based on R sq's values automatically given by Excel [15]

The second "For" loop in fig.5 (starting from the 17th line) adds a new sheet for each test of those 9 meant to verify which function from all those 5 is the most appropriate for our time filtered data set and which polynomial degree from another 5 options (only for polynomials). Instead of letting the user to fully control the spreadsheet (fig.4) after automatically creating scatter charts and corresponding trend lines, this loop (fig.5) verifies the R-sq (*R_sq* >

max_R_sq) in all 9 cases retaining the maximum value and the corresponding type and order (*trend* [*mem_type*, 1] and *o*, if polynomial), then gives a feed-back message box and closes Excel not saving anything (*savechanges:*=False).





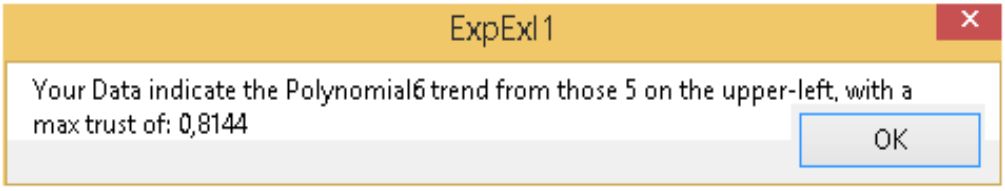
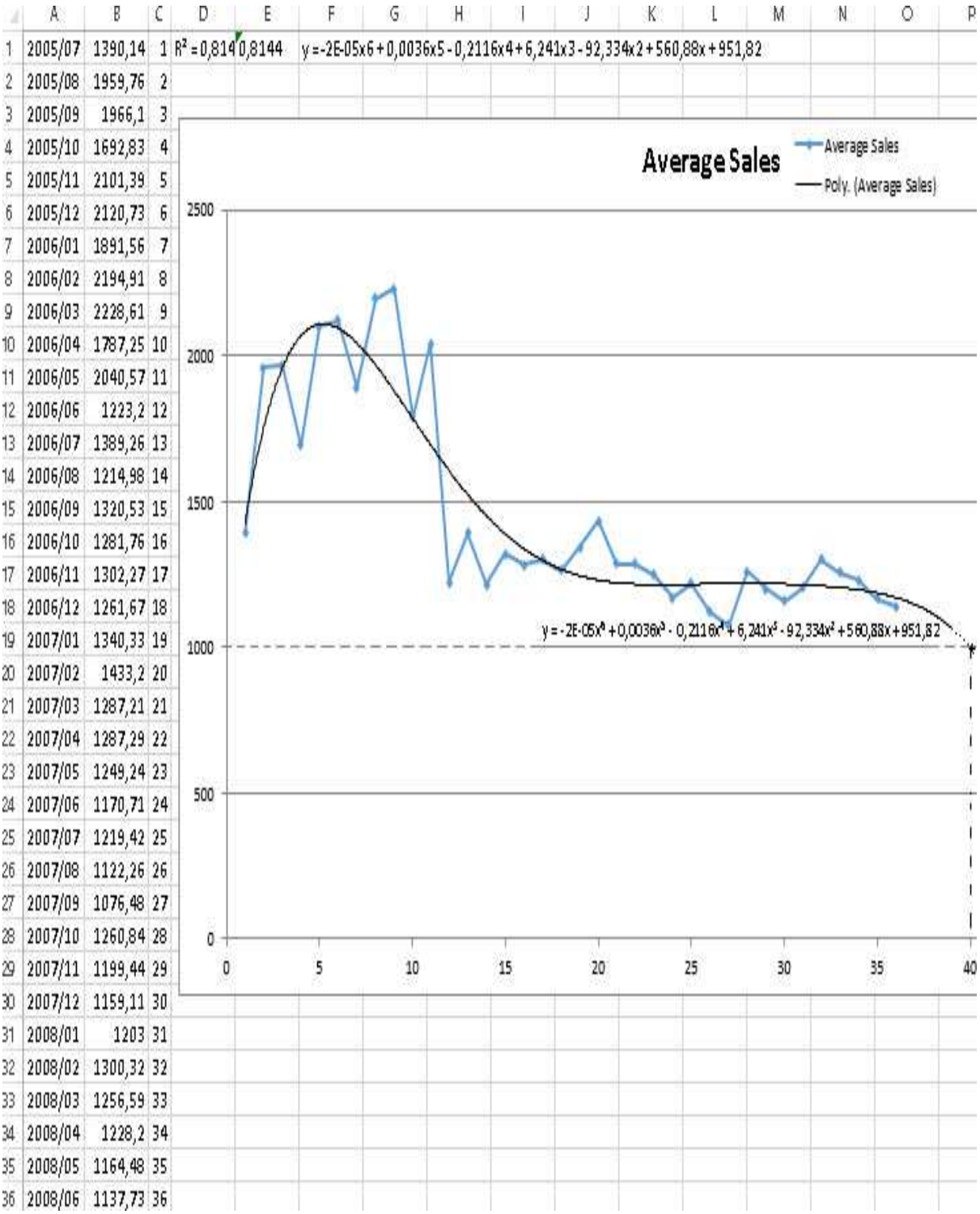


Figure 6. The most appropriate type of equation (Polynomial 6th degree, message box on bottom) for our data indicated by the *Suggest*.. button – different user interfaces with (newer - up) & without time filters (older - bottom) [16]



	A	B	C
1	2005/07	1390,14	1
2	2005/08	1959,76	2
3	2005/09	1966,1	3
4	2005/10	1692,83	4
5	2005/11	2101,39	5
6	2005/12	2120,73	6
7	2006/01	1891,56	7
8	2006/02	2194,91	8
9	2006/03	2228,61	9
10	2006/04	1787,25	10
11	2006/05	2040,57	11
12	2006/06	1223,2	12
13	2006/07	1389,26	13
14	2006/08	1214,98	14
15	2006/09	1320,53	15
16	2006/10	1281,76	16
17	2006/11	1302,27	17
18	2006/12	1261,67	18
19	2007/01	1340,33	19
20	2007/02	1433,2	20
21	2007/03	1287,21	21
22	2007/04	1287,29	22
23	2007/05	1249,24	23
24	2007/06	1170,71	24
25	2007/07	1219,42	25
26	2007/08	1122,26	26
27	2007/09	1076,48	27
28	2007/10	1260,84	28
29	2007/11	1199,44	29
30	2007/12	1159,11	30
31	2008/01	1203	31
32	2008/02	1300,32	32
33	2008/03	1256,59	33
34	2008/04	1228,2	34
35	2008/05	1164,48	35
36	2008/06	1137,73	36

Figure 7. The simple regression based prediction chart & trend line (chosen equation type: see cbo - fig.6) automatically created in Excel for 3 periods (forward: see tbo) [17]

By doing all these the user can get the false idea that this application prototype does the required tests behind all by itself when in fact the Excel application is the one responsible for computing the R-squared values when put to

work from code lines. This is a reason why we must choose honest and complete names for our user interface controls.

5. Higher degree for polynomial trend lines by using Google technology

For this paper's query based full sample data set (all 36 resulting lines - fig.1) the coefficient of determination for any tested equation (Excel trend lines) hardly gets near 0.8 which is usually not so bad. But we can test other facilities, functions (e.g. LINEST, LOGEST, FORCAST, TREND) [18], equation types or even a higher (more than 6) polynomial degree that Excel does not natively allow for trend lines in its common configuration (except add-ins).

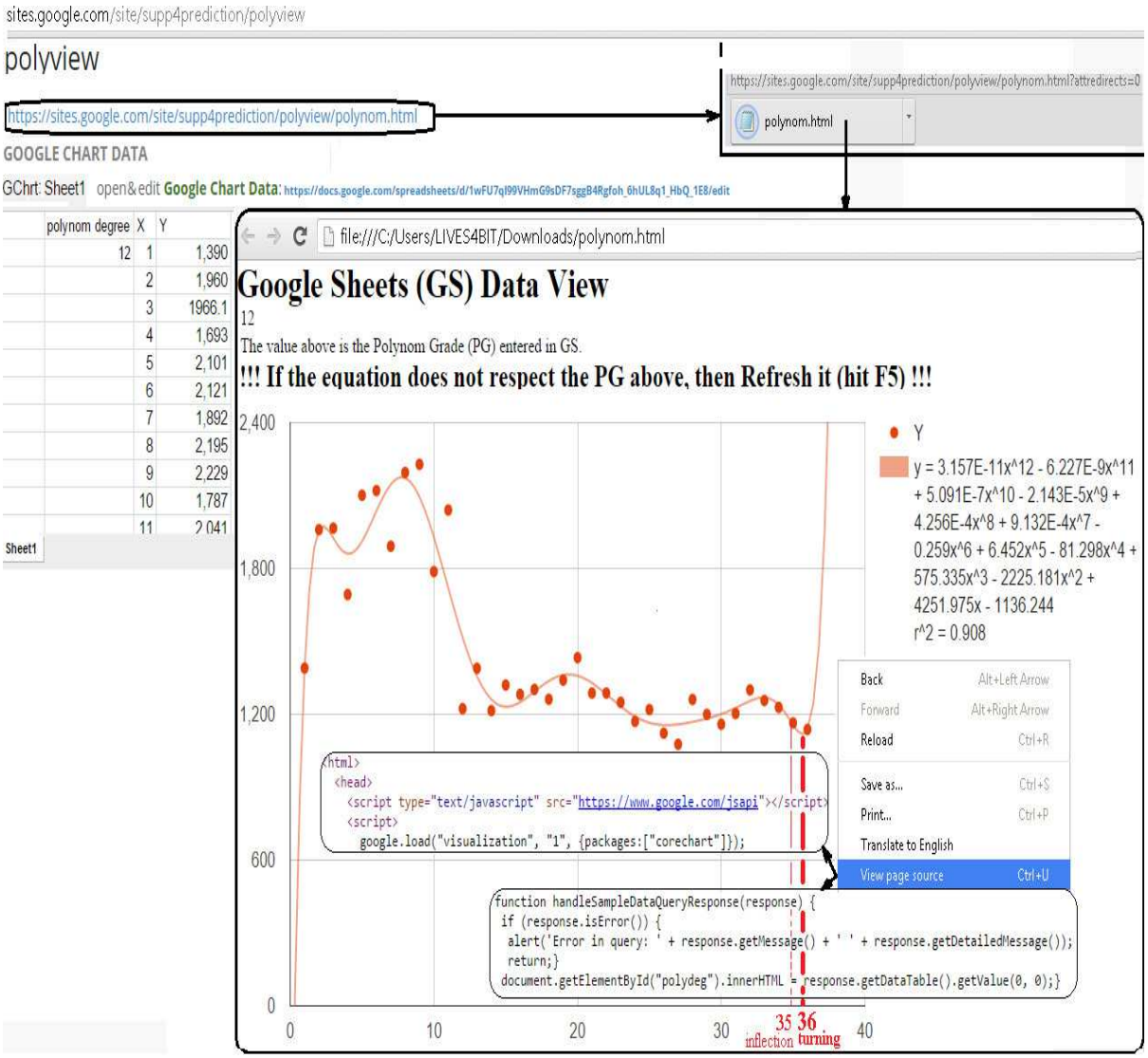


Figure 8. Simple regression (more than a 6 degree polynomial) generated in Java Script via Google Sheets [19]

Beeing curious of what it is happening with R-sq when establishing a higher polynomial degree, we have also tested a Java Script sequence able to query a Google sheet (fig.8 - fixed lenght data set with the same values of the 2nd column as the ones given by the 2nd column of the query in fig.1) and allowing us to programmatically generate dynamic polynomial trend lines with a degree grater than 6 and their corresponding R-sq, all in one .html file. We have also tested a higher polynomial degree inside a Google Sheets add-on programmatically generated by the author (fig.9 - variable lenght data set) via the Google Visualization API [20] and the Html Service [21] (mostly because of deprecated although still working Google UiApp [22]).

A result which we can get from this new model resumed in the last two examples (figs.8 and 9) showing an increased R-squared value (0.908) for the 12th degree polynomial, is the predicted Y value for an X near 40. In fact Y suddenly (last inflection point somewhere near the X value of 35 and last turning [23] point around 36 - fig.8) and

apparently tends to increase to infinite. And because of that it dramatically contradicts the previous expectations (figs.6 and 7 - 6th degree polynomial based model with R-quared of 0.814) both in terms of predicted value (Y~1000 for an X near 40 - fig.7) and last moment changes (slight decrease over 36 - fig.7).

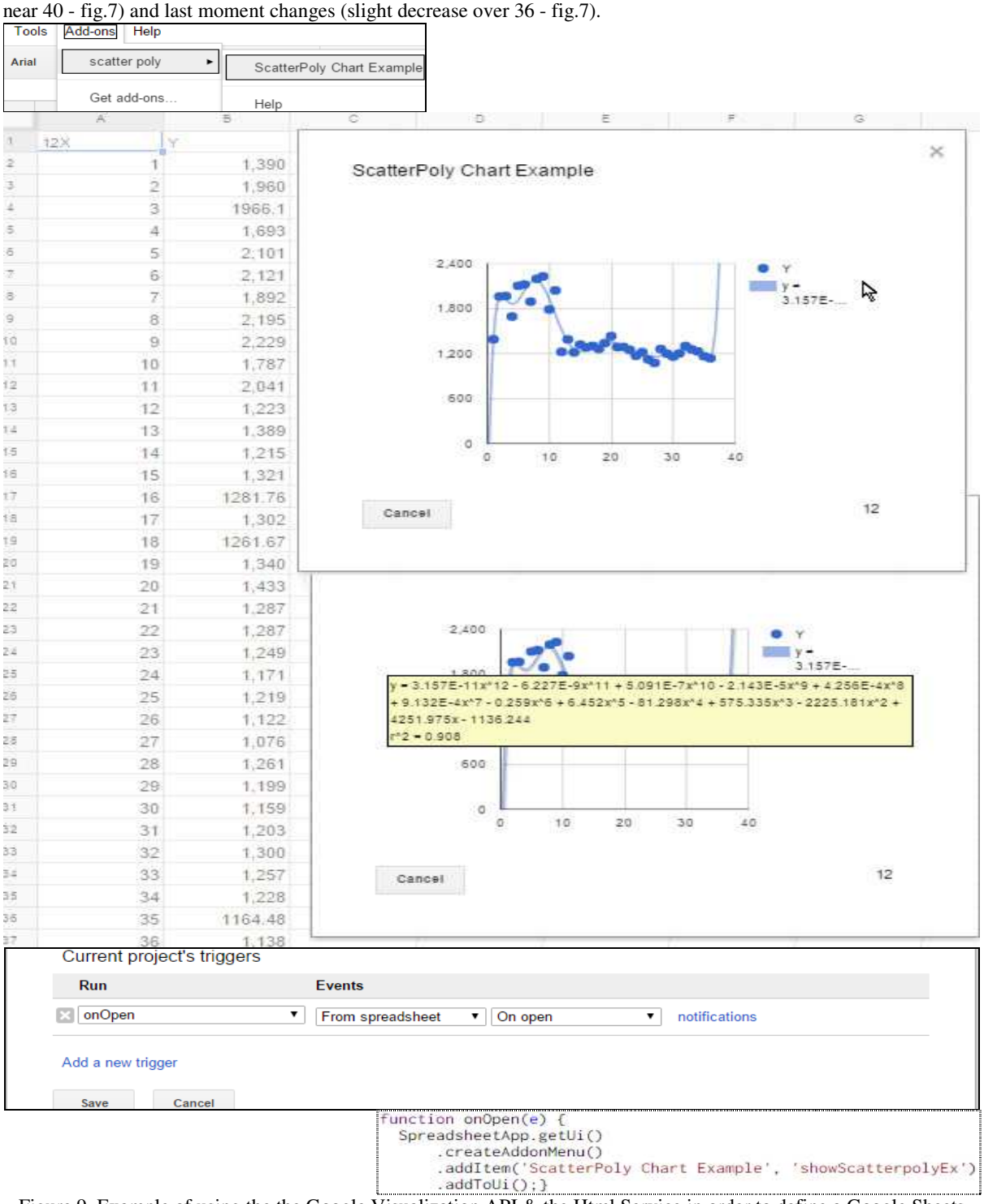


Figure 9. Example of using the the Google Visualization API & the Html Service in order to define a Google Sheets chart add-on (polynomial trends even over 6th degree) [24]

6. Limitations of this approach

First of all, the prediction when considering only the R-sq values of the Excel trend lines can lead to different

trends than intuitively expected. We have proved this for something that apparently seemed to be a descending line (fig.10). However the linear regression indicates a high value for R-sq~0.98. But we have also got a 3rd degree polynomial function with max R-sq (1 or 100% - see fig. 10). Can we jump to conclusions? First of all, we may need to have more data in order to make a decent prediction.

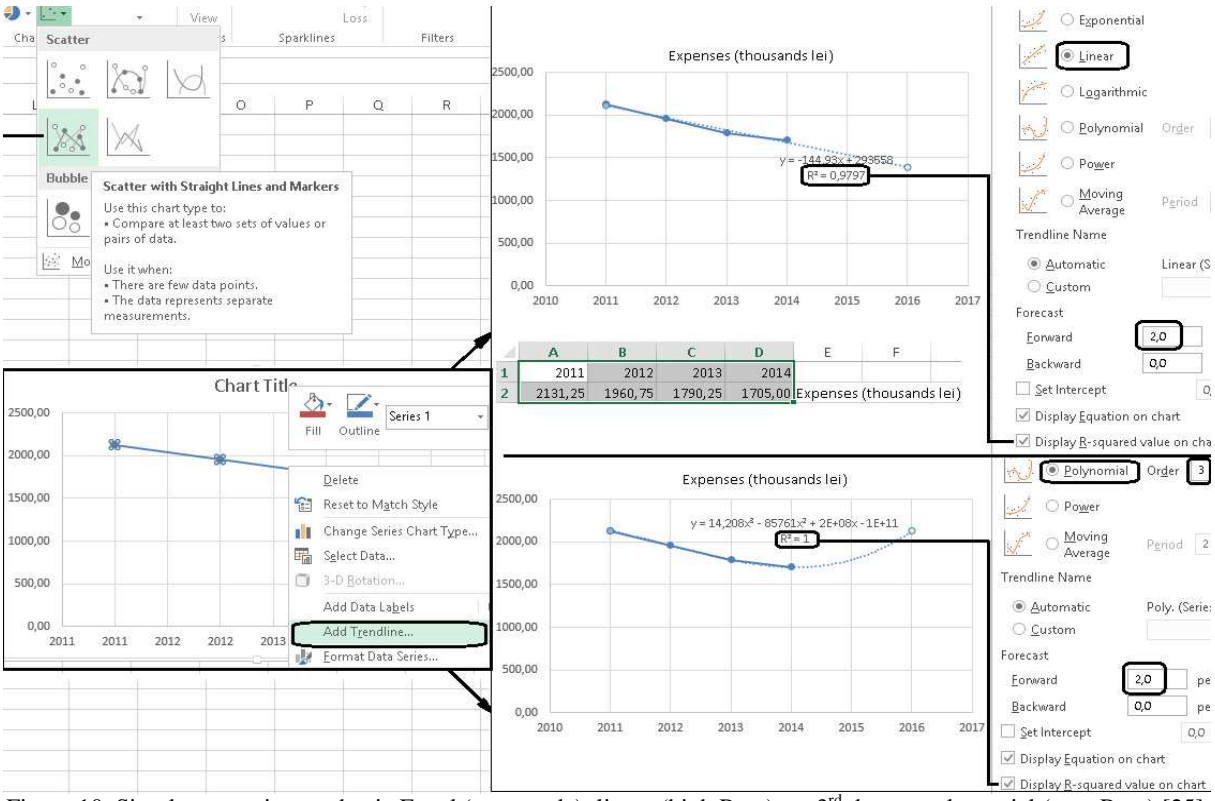


Figure 10. Simple regression made via Excel (user mode): linear (high R-sq) vs. 3rd degree polynomial (max R-sq) [25]

In fact, the first major type of limitations is strictly related to the interpretation of the values of the coefficient of determination (R-squared). In other words, low R-squared values are not always bad and high R-squared values are not always good. In most cases we additionally need to know two things: the result of the F-test of overall significance able to determine if the relationship is statistically significant [26] and the measure of randomness for residuals (Y) vs. fitted plots (X) [27]. Some [28] even consider that for certain cases we can have the worst model with the higher R-squared and the best one with an R-squared of 0. The mathematics professor Joseph Nystrom [29] from Worcester State University, Massachusetts, provides an inspired explanation of the R-squared concluding that its value represents the percentage of the differences in Y (variability in dependent Y) that can be explained by differences in X (variability of independent X). In other words, by the model itself. He also say that the residuals as difference between 100% and the value of R-squared represent that part remaining unexplained by the model. In order to better predict, Stephanie Glen [30] from the University of Nottingham and Keith Bower [31] from the University of Washington, advise us to use adjusted R-sq - meaning the R-sq adjusted for a number of variables and predicted R-sq - showing the measure of how the model would be at predictiong future values.

Another class of limitations is related to the fact that many business variables are interrelated. Thereby it might happen to go wrong in estimating evolution of such variables by considering only one way relation of determination between them.

7. Conclusions

The article brings further considerations starting from previously published [32] ideas and examples. It also shows how to: integrate a specific decision support application prototype with spreadsheets, automatically export certain data based on dynamic queries, create charts and get the best fitted functions for our scenario.

One big reason for the whole approach was to show that sometimes it is not absolutely necessary to "reinvent the wheel". That was applicable also for estimating the evolution of various dependent variables and the main limitations of the approach were considered and shortly presented in a special section of the paper.

In order to ensure the full comprehension of all examples mentioned within the article, we decided to add some references to eight (initially three - first three) step-by-step video tutorials (those contaning the word "index" in the URL) grouped in a short playlist and following the LIVES4IT (Homocianu, 2015) approach [33].

Overall, the main purpose of this paper was to enrich the original ideas and examples and because of that one of practical nature without claiming completeness.

8. Acknowledgments

This paper was funded by "Alexandru Ioan Cuza" University of Iaşi (UAIC), Faculty of Economics and Business Administration (FEAA), The Research Department.

9. Bibliographical references

[1] ncsu.edu/labwrite/res/gt/gt-reg-home.html
[2] excel-easy.com/examples/regression.html
[3] **Greavu-Serban V.**, *Analysis method of research papers published for audit domain, based on titles and keywords*, Annals of the „Constantin Brâncuşi" University of Târgu Jiu, Economy Series, Issue 4/2015, pp.54-55, www.utgjiu.ro/revista/ec/pdf/2015-04/08_Greavu.pdf
[4] dssresources.com/history/dsshistory.html
[5] **Airinei D., Homocianu D.**, *Globalization & Business Intelligence-Reloaded*, Scientific Annals of the Alexandru Ioan Cuza University of Iasi, 2010, Economic Sciences Series, pp.373, ssrn.com/abstract=2381813,
[6] **Homocianu D.**, *Sistemele de asistare a deciziilor în contextul societăţii cunoaşterii*, Editura Universităţii "Alexandru Ioan Cuza", Iaşi, 2009, pp.121, ssrn.com/abstract=2384380
[7] **Homocianu D., Airinei D.**, *Business Intelligence facilities with applications in audit and financial reporting*, Audit Financiar, pp. 21, Issue 9 (117)/2014, ssrn.com/abstract=2502552
[8] **Pucilowski A.**, *SharePoint as a Web Content Management System*, 2011, cognifide.com/blogs/sharepoint/sharepoint-as-a-wcms
[9] **Sireteanu A., Airinei D., Andone I., Homocianu D.**, *Implementing a Web Content Management System for an Educational Institution*, Economy Informatics Journal, 1-4/2008, pp.95, economyinformatics.ase.ro/content/en8/sireteanu%20a%20sa.pdf
[10] w3schools.com/sql/func_convert.asp
[11] stackoverflow.com/questions/20845465/ctrln-doesnt-open-new-query-window-in-management-studio-ms-sql-2008
[12] support.office.com/en-us/article/Choosing-the-best-trendline-for-your-data-1bb3c9e7-0280-45b5-9ab0-d0c93161daa8
[13] youtube.com/watch?v=aD1mDysZPUY&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ&index=4
[14] youtube.com/watch?v=EeZ--EEoP8U&index=6&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ
[15] sites.google.com/site/supp4ecotrend2015/download/fig5.png?attredirects=0&d=1
[16] youtube.com/watch?v=I_yEs6s0SFg&index=1&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ
[17] youtube.com/watch?v=YU0j4lTvX6s&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ&index=3
[18] k2e.com/tech-update/tips/160-predicting-the-future-with-trendlines-in-excel
[19] youtube.com/watch?v=wvYkQQAYTUY&index=7&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ
[20] computerworld.com/article/2593623/app-development/application-programming-interface.html
[21] mogsdad.wordpress.com/2015/07/19/converting-from-uiapp-chart-service-to-html-service-google-visualization-api/
[22] **Homocianu D., Airinei D.**, *On-Line Dynamic Dashboards in Audit Activities*, Audit Financiar, pp. 91-100, Issue 5 (125)/2015, ssrn.com/abstract=2602661
[23] teacherschoice.com.au/Maths_Library/Calculus/stationary_points.htm
[24] youtube.com/watch?v=9h6ERGZdV0o&index=8&list=PLkA3hbHQGQUC_O9KXr7crGMkxOY-z6pkZ
[25] sites.google.com/site/supp4ecotrend2015/download/other_tests.png
[26] blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
[27] carbon.ucdenver.edu/~mas/coursemtls/resids.pdf
[28] people.duke.edu/~rnau/rsquared.htm
[29] youtube.com/watch?v=IMjrEeeDB-Y
[30] youtube.com/watch?v=KjRrdb2x6dA
[31] keithbower.com/Miscellaneous/Some%20Misconceptions%20about%20R-Sq.htm
[32] **Homocianu D.**, *Spreadsheets as Decision Support Tools – Case Study on Automatic Trend Lines*, in Proceedings of The 12th International Conference ECOTREND, Targu-Jiu, Romania, 2015, ssrn.com/abstract=2698102

[33] **Homocianu D.**, *The LIVES4IT Approach on Access to Documentation Resources, Education, Training and Research*, in Proceedings of The 6th International Conference on Information Science and Information Literacy, Sibiu, Romania, 2015, bcu.ulbsibiu.ro/conference/proceed.htm, ssrn.com/abstract=2602711