University of West Hungary

Simonyi Karoly Faculty of Engineering, Wood Sciences and Applied Arts

Institute of Informatics and Economics

Decision support and its relationship with the random correlation phenomenon

Ph.D. Dissertation
of

Gergely Bencsik

Supervisor:

László Bacsárdi, PhD.

2016

DECISION SUPPORT AND ITS RELATIONSHIP WITH THE RANDOM CORRELATION PHENOMENON

Értekezés doktori (PhD) fokozat elnyerése érdekében
a Nyugat-Magyarországi Egyetem Cziráki József Faanyagtudomány és Technológiák
Doktori Iskolája

Írta:
Bencsik Gergely

Készült a Nyugat-Magyarországi Egyetem Cziráki József Doktori Iskola

Informatika a faiparban

programja keretében.

Témavezető: Dr. Bacsárdi László
Elfogadásra javaslom (igen / nem)

(aláírás)

A jelölt a doktori szigorlaton .... % -ot ért el,

Sopron, ...

....
a Szigorlati Bizottság elnöke

Az értekezést bírálóként elfogadásra javaslom (igen /nem)

Első bíráló (Dr. ....) igen /nem

(aláírás)

Második bíráló (Dr. ....) igen /nem

(aláírás)

A jelölt az értekezés nyilvános vitáján............% - ot ért el

Sopron, 2016

....
a Bírálóbizottság elnöke

A doktori (PhD) oklevél minősítése...................................

....
Az EDHT elnöke

STATEMENT

I, the undersigned Gergely Bencsik hereby declare that this Ph.D. dissertation was made by myself, and I only used the sources given at the end. Every part that was quoted word-for-word, or was taken over with the same content, I noted explicitly by giving the reference of the source.

Alulírott Bencsik Gergely kijelentem, hogy ezt a doktori értekezést magam készítettem, és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Sopron, 2016

…………………………………………….

Gergely Bencsik

Abstract

Mankind has always pursued knowledge. Over the philosophy questions by Paul Gauguin—"Where do we come from, what are we, where are we going?"—the science may answer these questions. In every scientific field, empirical and theoretical researchers are working to describe natural processes to better understand the universe. Gauguin's questions were modified by scientists: Are these two variables correlated? Do several independent datasets show some connection to each other? Does a selected parameter have effect on the second one? Which prediction can we state from independent variables for the dependent variable? But the seeking of knowledge is the same.

The constantly increasing data volume can help to execute different analyses using different analyzing methods. Data itself, structures of data and integrity of data can be different, which can cause a big problem when data are uploaded into a unified database or Data Warehouse. Extracting and analyzing data is a complex process with several steps and each step is performed in different environments in most cases. The various filtering and transformation possibilities can make the process heavier and more complex. However, there is a trivial demand for comparison of the data coming from different scientific fields. Complex researches are the focus of the current scientific life and interdisciplinary connections are used to better understand our universe.

In the first part of our research, a self-developed Universal Decision Support System (UDSS) concept was created to solve the problem. I developed a universal database structure, which can integrate and concatenate heterogenic data sources. Data must be queried from the database before the analyzing process. Each algorithm has its own input structure and result of the query must be fitted to the input structure. Having studied the evolution line of databases, Data Warehouses and Decision Support Systems, we defined the next stage of this evolution. The Universal Decision Support System framework extends the classic Data Warehouse operations. The extended operations are: (1) create new data row [dynamically at the data storage level], (2) concatenate data, (3) concatenate different data rows based on semantic orders. Reaching universality is difficult in the logic and presentation layer, therefore we used an "add-on" technique to solve this problem. The set of transformation and analyzing methods can be extended easily. The system capabilities are used in three different scientific fields' decision support processes.

The second part of our research is related to analyzing experiences and data characteristics performed in the Universal Decision Support System. Nowadays, there are several methods of analysis to describe different scientific data with classical and novel models. During the whole analysis, finding the models and relationships mean results yet then comes the prediction for the future. However, the different analyzing methods have no capability to interpret the results, we just calculate the results with the proper equations. The methods itself does not judge: the statements, whether the correlation is accepted or not, are made by experts. Our research focuses on how it is possible to get different inconsistent results for a given question. The results are proved by mathematical methods and accepted by the experts, but the decisions are not valid since the correlations originated from a random nature of the measured data. This random characteristics—called Random Correlation—could unbeknown to the experts as well. But this phenomenon needs to be handled to make correct decisions. In this thesis, different methods are introduced with which Random Correlation can be analyzed and different environments are discussed where Random Correlation can occur.

Kivonat

Az emberiség mindig is kereste a választ a honnan jövünk, mik vagyunk, hová megyünk kérdésekre. Paul Gauguin kérdéseire talán a tudomány fogja megadni a választ. Minden tudományterületen folynak elméleti és tapasztalati kutatások, hogy leírják a természetben zajló folyamatokat. A cél minden esetben, hogy jobban megismerjük a minket körbevevő univerzumot. A kutatók azonban átfogalmazzák Gauguin kérdéseit, úgy, mint például hogy összefügg-e két változó, két független adathalmaz korrelál-e egymással, egyik paraméternek van-e valamilyen hatása a másik paraméterre, független változók alapján milyen jóslást mutatnak az eredmények a függő változóra vonatkozóan. Ugyanakkor minden esetben a keresett tudás visszavezethető Gauguin kérdéseire.

A kísérletek során mért adatok kritikus szerepet töltenek be a tényeken alapuló döntések meghozatalában. Ezen adatsorok egységes tárolása nem minden esetben triviális, különösen, ha más céllal, ebből fakadóan más környezetben történt adatszerzésről van szó. Ebben az esetben az adatok struktúrájukban, integritási szintjükben mások lehetnek, amelyek nehezítik az egységes adatbázisba vagy adattárházba történő beillesztésüket. Az adatok kinyerése és értelmezése többlépcsős folyamat. A nyers adatokon értelmezett szűkítések és transzformációk ezután tipikusan már egy másik rendszerben kerülnek megvalósításra, végül a már transzformált adatokon értelmezzük a vizsgálati módszert, amely egyrészt adott elemzési területhez (matematika, statisztika, adatbányászat) kapcsolódik, másrészt általában ugyancsak különböző rendszerben implementálnak. Másik oldalról egyértelmű igény van több tudományterületen mért adatok összevetésére, interdiszciplináris kapcsolatok kutatására.

Kutatásom első részében a fentiekből indultam ki és adtam egy általam kifejlesztett egységes megoldást. Létrehoztam egy univerzális adatbázis struktúrát, amely a különböző forrásokból érkező heterogén adatokat fogadni és összefűzni is képes. Az univerzális lekérdező felület segítségével az egyes módszerek különféle bemeneti struktúráját állíthatjuk össze. Kutatásom során tanulmányoztam az adatbázisok és adattárházak evolúciós vonalát, amelynek eredményeképpen egy új állomást definiáltam. Az univerzális döntéstámogató keretrendszer funkciói kibővítik a hagyományos adatbázis és adattárház műveleteket, amelyek: (1) új adatsor létrahozása [dinamikusan az adatbázis struktúrában], (2) adott adatsor összefűzése, (3) különböző adatsorok összefűzése adott szemantikai összerendelést követve. A logikai és megjelenítési szintű univerzalitás elérése nehézkes, ezért a szakirodalomban követett „add-on” technikákat alkalmaztam, ugyanakkor maximálisan törekedtem a könnyű bővíthetőségre. A rendszer képességeit három különböző tudományterületen végzett elemzéssel mutatom be.

Disszertációm második része az univerzális elemző keretrendszerben történő elemzések tapasztalatai alapján a módszerek és adatok karakterisztikájára vonatkozik. Az adatelemzési folyamat során a modell és a kapcsolatok megtalálása már önmagában eredményt jelent, ezután következik a – lehetőleg minél pontosabb – jóslás a jövőre nézve. A különböző adatelemzési módszerek önmagukban nem képesek értelmezni az eredményeket, vagyis az adott matematikai képlettel csak kiszámoljuk az eredményeket. A módszerek önmagukban nem ítélkeznek, a megállapításokat, miszerint elfogadjuk vagy sem az adott összefüggést, mindig egy elemző személy teszi meg. Kutatásomban azt vizsgáltam, hogy mi történik akkor, ha az elemző által keresett összefüggés matematikailag bizonyítható, de az adott elemzési döntés mégsem helytálló, mivel a matematikai összefüggés a mért adatok olyan véletlenszerűségéből adódik, amelyet az elemző személy sem ismerhet. Az olyan összefüggéseket, amelyek a véletlenszerűség következményeként létrejönnek, véletlenszerű kapcsolatoknak neveztem el.

Acknowledgement

I am heartily thankful to my mother *Klára Jordanits* for the support and encouragement provided so many years. Without her continuous support, this work would not have been born.

I am thankful to my supervisor, *Dr. László Bacsárdi* (associate professor at the Institute of Informatics and Economics, University of West Hungary) his support and guidance that helped to complete my thesis.

As a young researcher, I was able to visit K.U. Leuven twice a few years ago. I am really grateful for their pleasant environment which inspired me in my research related to the Universal Decision Support System.

I would like to thank *Prof. Jos Van Orshoven* (professor at the University K.U. Leuven) to review my dissertation. I also thank *Dr. Ákos Dudás* (senior lecturer at the Budapest University of Technology and Economics) to review my dissertation. They advices help me a lot to make my dissertation better.

I want to thank my colleague, *Attila Gludovátz* who helped me a lot in my research.

Contents

List of Figures

List of Tables

Key Terms

Analyzing Session. A process in the Universal Decision Support System, which starts from the query phase and ends with the presentation of the results.

ANOVA. It is a statistical test to determine whether the data groups' means different are or not.

Big data. A large dataset, which satisfies Volume, Velocity and Variety (3V) properties.

Business Intelligence. Variety of models, methods and software solutions used to organize and analyze raw data.

Database. An organized collection of data. Databases are managed with Database Management Systems.

Data Warehouse. Repository of all kinds of Enterprise data.

Decision. The act or process of deciding. Choosing from several decision alternatives.

Decision Support System. According to Sprague's definition, DSSs are *"an interactive computer based system that help decision-makers use data and models to solve ill-structured, unstructured or semi-structured problems".*

DSS-generator. According to the definition of Sprague and Watson, DSS generator is a *"computer software package that provides tools and capabilities that help a developer build a specific DSS"*

ETL. Extract, Transform, Load. This component is responsible to extract data from the original data source, transform the structure from the original to the data warehouse pre-defined structure and upload data with new structure into the data warehouse at the end.

Ionogram. An ionogram describes the current state of the Ionosphere (a layer of the Earth's atmosphere).

Knowledge Discovery. According to the definition of Fayyad, Knowledge Discovery is a *"non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".*

Multi-Criteria Decision Analysis. A tool that performs complex analyses, i.e., support ill- or non-structured decisions

NoSQL. Next Generation Databases, mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable.

Random Correlation. A process to determine the random impact level of a given analysis (*Session*).

Session. A process in Universal Decision Support System. The steps of the process are the following: (1) integrating data, (2) uploading data into database (3) query sub-process, (4) analysis [data manipulation phase) and (5) presentation of the results.

Universal Decision Support System (UDSS). The concept provides data- and model-based decision alternatives. It supports to solve all structured and semi-structured problem, i.e., the nature of the data and the methods as well as the goal of the decision are general.

Common symbols

Common symbols used in Chapter 5 (Random Correlations).

μ: expected value.

σ: deviation.

α: significance level.

F-value: the result value of ANOVA.

k: number of columns.

n: number of data items of each column.

r: range.

t: number of performed methods.

Ω-model: Random Correlation method, all possible input candidates are generated.

C-model: Random Correlation method, it shows how much data are needed to find a correlation with high possibility.

1. Introduction

Data has an important role nowadays. Data about people, environments and everything are measured. Mobil phones, smart technologies and all kinds of sensors transmit data to databases. Data are analyzed by analysts and are used to create personal offers, marketing plans and perform other activities to better understand customer behaviors. Data play an important role in the field of industry as well. Since industry is a very large area, data collectors, sensors, data management and analyzing processes are becoming a more and more dominant area among scientific fields. This tendency is supported by the American Internet Industry and European Industry 4.0 approaches as well. According to the predictions, there will be 50-100 billion sensors and then measured data will be available on the Internet. The data and the related countless analysis possibilities mean several challenges not just in computer science, but in all kinds of scientific fields. However, more and more complex data and analyzing capabilities also mean challenges. But can the old models be applied in the new environment? Can all methods be used in the so-called big data environment? Are the results satisfying and precise? Should we apply modifications to get better decision alternatives? It is possible that this complexity requires new models and methods.

1.1. Problem specification

There are a lot of data and data rows is easy to collect nowadays. Related to that, the standard research methodology is defined by many state-of-art publications [1, 2]. Specialized research methodologies also appear corresponding to the given research fields [3, 4]. In general, an analyzing session starts with the data preparations (collect, clean and/or transformation), continues with choosing analyzing method and finally, the result is presented and interpreted. If we have a lot of data item, we talk about big data, which can provide more analyzing possibilities and more precise results, as we would expect. But a lot of contradictory results were born in different scientific fields and the literature contains many inconsistent statements.

In biology, squids size analyzes generated opposite results. It was reported by Jackson and Moltaschaniwskyj that squids got bigger [5] than before. The research target areas were the northeast seaboard of North America, the Pacific coast of South America, West Africa, the European Atlantic and Mediterranean oceans. But another research proved that the squids' size is getting smaller [6]. It is true, that in [6], the squids' size is presented based on the difference between tropical (small body size) and sub-tropical (large body size) zones, however, there is an overlap between the previously areas in [5]. There is a common author in both papers, but the results can regard as contradictory. Zavaleta et al. stated that grassland soil has more moisture [7]. According to Liu et al., grassland soil must face against less moisture [8]. Church and White showed out a significant acceleration of sea level [9]. However, comparing the results with [9], Houston and Dean results show us see-level deceleration [10]. According to one research group, the Indian rice yields to increase [11], while another reports decrease [12]. One research team stated that coral inland atolls are sinking [13], while another reported rising [14]. According to McCaffery and Maxell, Columbia spotted frogs population is growing [15], while McMenamin et al. reported population decline [16]. One research group result was that warm boosts Chinese locust outbreaks [17], while another team stated the same with cold [18]. According to Drinkwater research, the cod population is thriving [19] but another group stated cod population decreasing [20].

In medicine, the salt consumption is always generating opposite publications. There are papers supporting it and do not disclose any connection between consumption and high blood pressure [21]. Another research group states that the high salt consumption causes not only high blood pressure but kidney failure as well [22]. An Eastern African country, Burundi is heavily hit by malaria disease. Two contradictory results were published about the number of the malaria patients. According to [23] report, the malaria is increased, while Nkurunziza and Pilz showed contradictory results [24]. Further researches are performed in malaria at global level. Martens et al. estimate 160 million more patient in 2080 [25] while others report global malaria recession [26].

In forestry, Fowler and Ekström stated that UK has more rain in the recent years [27] than before. According to Burke et al., UK has not just simple droughts, but further droughts is predicted [28]. Held et al. stated that Sahel, a transition zone between Sahara and savanna in the north part of Africa, has less rain [29]. However, another research group suggested more rain for Sahel [30]. In Sahel local point of view, Giannini's result was that it may get more or less rain [31]. Crimmins et al. stated that plants move downhill [32], while Grace et al. suggested opposite result: plants move uphill [33]. Dueck et al. dealt with plant methane emission. They result was that this emission is insignificant [34]. Keppler et al. stated that this emission is significant and they identify plants as the important part of the global methane budget [35]. Contradictory results are in leaf index research as well. Siliang et al reported leaf area index increase [36], while other research mentioned leaf area index decrease [37]. According to Jaramillo et al. Latin American forests have thrived with more carbon dioxide [38] but Salazar et al.'s projection is that Latin American forest decline [39]. One research group presented more rain in Africa [40], while another reported less rain [41]. According to Flannigen et al., Boreal forest fires may continue decrease [42] but Kasischke et al.'s projection was increasing of fires [43]. Three different results can be found about bird migration. According to one, bird migration is shorter [44]. The second presents long migration time [45]. The third reported that bird migration is out of fashion [46]. Two publications with contradictory title were published related to Amazon rainforest green-up [47, 48].

In sociology, there are arguments related to data-based analysis and because these results do not produce the real predictions, a new methodology was proposed [49]. Another example is based on questionnaires and scores. Seeking the answer for internet addiction, Lawrence et al. showed that the increasing Internet using among young people increases the chance of depression [50]. But Cai-Xia Shen et al. showed that Internet is critical for daily satisfaction of the children [51]. Relating to Internet, Massively Multiplayer Online Role-Playing Games (MMORPG) always generate contradictory results. Quoting Brian and Wiemer-Hastings, *"Research on Internet addiction has shown that users can become addicted to it"* [52]. However, Yee result states that *"Oftentimes, both the media and researchers into media effects collapse all video gamers into a simplistic archetype. While this facilitates making sweeping generalizations of potentially deviant behaviors or consequences (i.e., addiction and aggression), this strategy inevitably ignores the important fact that different people choose to play games for very different reasons, and thus, the same video game may have very different meanings or consequences for different players"* [53]. According to Doughty et al., Stone Age hunters may have triggered past warming [54]. Smith et al. stated the same, but past cooling is included in their statement [55].

In Earth science, Schindell et al. stated that winters could getting warmer in the northern hemisphere [56]. According to other opinion, winters are maybe going to colder there [57]. Knippertz et al. deal with wind speeds and they concluded that wind speed become faster [58]. Another resource group stated that wind speed is declined by 10-15% [59]. According to the third opinion, the wind speed speeds up, then slows down [60]. Many research was performed about the debris flows in Swiss Alps. One research group states that debris flows may increase [61] but another group's results were that it may decrease [62]. Another research group published that it may decrease, then increase [63]. In Charland et al. research, we can read that San Francisco is getting foggier [64]. However, according to another opinion, Pacific coast of California has less fog [65]. Miller and Vernal stated that Northern Hemisphere ice-sheets grow [66]. In the Intergovernmental Panel on Climate Change (IPCC) fourth assessment report contained the opposite result: ice sheets of northern hemisphere are declining [23]. In research of North Atlantic Ocean, Boyer et al. stated that it became saltier [67] from 1955 to 2006. According to another result, it became less salty in recent decades [68]. Knutson et al. suggested that North Atlantic cyclone frequency is decreasing [69]. The counter-result was formulated by the authors of the Global Climate Projection report [70]. In the same report, we can read that Indian monsoons are getting wetter. Chung and Ramanathan presented that Indian monsoons are getting drier [71]. About the Gulf Stream speed, one group stated that it slows [72], another group reported small amount of increasing speed [73]. According to Burnett et al. Great Lakes have more snow [74]. Mortsch and Quinn reported less snow [75]. One research group result presents slowdown of Earth rotation [76], while another group stated that Earth rotates quicker [77]. According to Martin et al., the avalanches hazard is decreasing in mountains [78]. However, more avalanches are expected by another research team [79].

Nosek et al. repeat *98* + *2* psychology researches (two were repeated by two individual group) [80]. Only *39%* of the publications showed the same significant results as before. In another cases, contradictory results between the repeated research and the original research, came out. The authors of the original publications were part of the repeated research as well to secure the same research methodology performed before in the original case. The 270 authors' paper main conclusions were:

- *The most noted scientific journals review processes are not so solid*. They would not to decide that the results are good or bad, they do not want to confute the results. This approach is the same as our opinion: as we mention before, we do not deny real correlations.
- *Discover more cheat-suspicious result*. Nosek's work is part of a multi-level research project. During the other phase of the main project, cheat-suspicious results were found.
- *Another scientific area has the same reproduction problem, not just psychology*. We summarized a lot of contradictory results in this section, but also in Nosek's paper, there are references about non-solid results.
- *They urge cooperation between scientists*. Nosek et al. encourage researchers to build public scientific databases, where data, which the scientific results and conclusion based on, are available. Since UDSS main goal was to support any kind of scientific researches, the concept is suitable to be a scientific warehouse.

The above mentioned researches focus on the same topics but they have different, sometimes even contradictory results. This shows us how difficult the decision making could be. Our research focuses on how the inconsistent results could be originated. This does not mean that one given problem cannot be approached with different viewpoints. We state that there are circumstances, when the results could born due to simple random facts. In other words, based on parameters related to data items (e.g., measured items range, mean and deviation) and analyzing method (e.g., number of methods, outlier analysis) can create such environment, where the possible judgment is highly determined (e.g., data rows are correlated or non-correlated, pendent or independent). Our goal was the examination of these situations and analyze where and how the contradictory results can be born. Based on our results, a new phenomenon named *Random Correlation* (RC) is introduced.

RC can appear each scientific fields. To analyze the RC behavior on various data sets, a Decision Support System is needed to be implemented. In general, the trivial DSS implementation approach is the following: (1) problem definition, (2) design of data collecting methods which has effect on database structure (3) design of DSS functions (4) implementation and (5) test and validation. To ease creating DSSs, new DSS solutions are implemented and the technology is continuously evolving. The DSS design phases can be shorter than earlier and the component approach can be applied as generic implementation, however, not all kinds of modifications can be managed easily. Diversity in data nature and decision goals can eventuate problems as well as environment heterogeneity or performances. If a Data Warehouse is used in a research, the structure of the Data Warehouse must be modified if a new data row shows up and all kind of modifications eventuate a new project. In a company, if a new production machine is used, then new processes must be implemented. Due to this, new data will be measured which leads to the partial or total redesign of the old DSS system. Handling data originating from different fields could be a difficult task. Each scientific field has its own characteristics of data and methods of analyses. They differ in data storage, data queries, data transformation rules, in one word in whole analysis process. However, to answer RC questions, we need to handle differences uniformly. Therefore, we need to build a system with universal purposes.

The Universal Decision Support System (UDSS) concept and Random Correlation (RC) are the two main parts of this interdisciplinary dissertation.

1.2. Outline

The dissertation is organized as follows.

Chapter 2: This chapter deals with the overview of the related literature. In Section 2.1, the applied methods are reviewed like the normality statistics tests, the test of Bartlett, ANOVA and regression techniques. In Section 2.2, our focus is on the literature of the Decision Support Systems. In Section 2.2.1, the DSS history is presented. The current DSS solutions are presented in details. The DSS definition approaches are summarized in Section 2.2.2., and the various DSS classification and components are discussed in details in Section 2.2.3. DSS common models and methods, DSS-generator, Data Warehouse and ETL, Multi-criteria approach, NoSQL and Business intelligence solutions are discussed in each further subsection of Section 2.2 respectively.

Chapter 3: Specific research objectives are defined in Section 3.1. In Section 3.2, we overview the generalization process, which lead us to Universal Decision Support concept and Random Correlations. Related methods and models are summarized in Section 3.3

Chapter 4: In this chapter, the Universal Decision Support System concept is discussed in details. In Section 4.1, we present the Universal Decision Support System architecture. In Section 4.2, the steps of an analyzing session are introduced. From Section 4.3 to 4.6, each UDSS elements and their implementation are detailed. Results related the UDSS are presented in Section 4.7. Three main analyzing processes are performed with UDSS: (1) decisions processes related to forestry is introduced in Section 4.7.1, (2) ionogram automatic and semi-automatic processing is presented in Section 4.7.2, while (3) vendor selection decision support processes is discussed in Section 3.3.

Chapter 5: Random Correlation theory is discussed in this chapter. We created the Random Correlation framework with the following parts (1) definition, (2) parameters, (3) models and methods, (4) classes and (5) standard RC analyzing process. The framework is introduced in Section 5.1. In Section 5.2, Analysis of Variance is analyzed in the view of Random Correlation. During our research, increasing total possibility space problem arises. We proposed space reducing techniques to solve this problem in Section 5.3. The result of ANOVA is presented in Section 5.4, while results related to regression techniques are introduced in Section 5.5.

Chapter 6: The main results of this dissertation is concluded in this chapter.

This dissertation follows the spelling of the US English.

2. Overview of related literature

The literature overview has two main parts. First, we summarize models and methods which we use during RC analysis. Second, we summarize Decision Support System (DSS) literature. We review the solutions available on the market as well as the possibilities which can be used to build a universal DSS.

2.1. Applied methods

In this section, the mathematical backgrounds of the used algorithms are presented. Since RC is a new theory, we start with the basic (classic) analyzing methods. It would seem that the applied methods overview is basic, however, the precise equations and the theoretical background must be introduced to understand the RC calculation processes and to reproduce the research easily. Summarized methods are behind of our self-developed the Space Reducing Techniques. We highlight those equations and steps, which are used during the implementation of RC analyzing session.

Two main analyzing methods, Analysis of Variance and regression techniques are analyzed in the view of RC. Their assumptions, which are statistical tests mainly, are also discussed. All statistical tests' calculation process has the following steps:

(0) **Check the assumptions (if any)**. It is possible that the method has conditions of adaptation. If these assumption is not passed, the given analyzing method cannot be used.

(1) **Define H_0**. It is called null hypothesis. Every test has its own H_0. For example, the data items follow the normal distribution.

(2) **Define H_1**. It is called alternative hypothesis. Every test has its own H_1. In the end of test, either H_0 or H_1 is accepted, but we cannot accept both at a time. For example, the data items do not follow the normal distribution.

(3) **Define significance level (α)**. This is the level of the probability of rejecting the null hypothesis, however, it is true. [Type I error.]

(4) **Identify critical value**. Based on α and the given statistical distribution, the *critical value* can be determined.

(5) **Calculation process ending with a value**. Each test has its own calculation process ending with the value of *test statistic*.

(6) **Comparison**. We compare the *test statistic* and the *critical value*. According to the comparison result, accept or reject H_0 and related to that, reject or accept H_1.

(7) **Conclusion**. We make the conclusion, for example, if H_0 is accepted, then the data items follow the normal distribution at significance level α.

2.1.1. Normality

The normal distribution is the basic statistical distribution. Many test condition includes that the data must follow the normal distribution. In the case of normality check, the main question is, whether the measured values follow the normal distribution or not.

2.1.1.1. Classic Chi-Square (χ^2) Test for a Normal Distribution

The Classic Chi-Square Goodness of Fit Test for a Normal Distribution is one of the oldest test. Based on [81], this test is allowed if the following three basic conditions are met for the experiment (*Step 0*):

- Data sample are chosen randomly;
- The studied variable is categorical;
- The number of data items in each category is at least *5*.

The H_0 states that the data items follow the normal distribution, H_1 states they do not follow the normal distribution (*Step 1* and *2*). We define the significance level α (*Step 3*). The *critical value* can be determined in the χ^2 table with the degrees of freedom (*Step 4*). In the case of this test, the degrees of freedom (*df*) is calculated as $\chi^2_{1-\alpha} = v$, where $v = l - b - 1$. The b is the number of the given distribution parameters, which we estimate from the sample. Now, $b = 0$, therefore $v = l - 1$.

According to *Step 5*, we need to calculate the two parameter of normal distribution $N(\mu, \sigma)$ in the case of normality,: the expected value (μ) and the deviation (σ). In the case of Chi-Squere test, we have observed frequencies (O_i), and expected frequencies (E_i). The O_i is determined, i.e., these values are measured, however, the E_i must be calculated. The main though behind the test is that calculation of E_i is originated to standard normal distribution. Therefore, we create discrete sets and standardize the sets limits with Eq. 1.

$$Z = \frac{x_i - \mu}{\sigma}, \qquad \text{Eq. (1)}$$

where x_i is the limit of the given set. Based on these results, we seek the proper values from the *Z*-table. Since the table values is cumulated, we can subtract the neighboring values, so we can calculate that possibilities in each set, which are the expected possibilities (E_i) in the given set. Multiplying these possibilities with the number of all data (*N*), we get that number, which the expected number of the data in each set, when the given data row would follow the normal distribution. The last step of the calculation phase is to calculate the test statistic with the Eq 2.

$$\sum_{i=1}^{l} \frac{(O_i - E_i)^2}{E_i}, \qquad \text{Eq. (2)}$$

where *l* is the number of classes. If the test statistic is smaller than the critical value (*Step 6*), then we accept H_0 and can make the conclusion that the data items follow the normal distribution at significance level α. Contrarily, we accept H_1 and then, data do not follow the normal distribution statistically (*Step 7*).

2.1.1.2. D'Agostino-Pearson test

The other method used for normality check is D'Agostino-Pearson omnibus test [82]. The H_0 states that we do not have reason to reject that the data follows the normal distribution, since H_1 states that they do not follow [*Step 1* and *2*]. We remark that the H_0 do not state the data follow the normal distribution obviously. However, if H_1 is accepted, then we can state that data do not follow the normal distribution unequivocally. The α has the same meaning as before (*Step 3*). The χ^2 table contains the *critical value*, but in this case, the degrees of freedom is always *2* regardless the sample size (*Step 4*). D'Agostino did not proof that explicitly, most of the experts says this is an empirical approach rather. The calculation process is based on the sample distribution skewness and kurtosis (*Step 5*). The moment coefficient of skewness is calculated with Eq. 3.

$$skewness: g_1 = \frac{m_3}{m_2^{3/2}}, where$$
$$m_3 = \frac{\sum_i^n (x_i - \bar{x})^3}{n} n \; and \; m_2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n} n, \qquad \text{Eq. (3)}$$

and $\bar{x}$ is the mean and *n* the sample size, m^2 is the variance and m^3 is called the third momentum of the data set. The Eq. 4 is the measure of skewness, if we have data of the entire population. But we have only sample from the population in most cases. The sample skewness is defined with the following equation:

$$G_1 = \frac{\sqrt{n * (n-1)}}{n-2} * g_1. \qquad \text{Eq. (4)}$$

However, sample skewness G_1 just approximates the population skewness. In other words, there is an error between the population skewness and the sample skewness. This error must be noted during the calculation. G_1 must be divided with standard error of skewness (*SES*):

$$test\ statistics\ (skewness): Z_{g_1} = \frac{G_1}{SES}, where\ SES = \sqrt{\frac{6 * n * (n-1)}{(n-2) * (n+1) * (n+3)}}. \qquad \text{Eq. (5)}$$

We can also say that the G_1 measures the amount of sample skewness. The bigger the number, the bigger the skew of the sample. At the same time, Z_{g_1} is a probability. This probability means whether the population (based on sample skewness G_1) is skewed or not. The bigger the number of Z_{g_1}, the bigger the possibility that population is skewed.

The moment coefficient of kurtosis calculation process is similar to skewness:

$$g_2 = a_4 - 3, where$$
$$a_4 = \frac{m_4}{m_2^2}\ and \qquad \text{Eq. (6)}$$
$$m_4 = \sum_i^n (x_i - \bar{x})^4 / n\ and\ m_2 = \sum_i^n (x_i - \bar{x})^2 / n.$$

Again, $\bar{x}$ is the mean and *n* is the sample size. The m_4 is called the fourth momentum. Based on Eq. 7, the sample excess kurtosis is calculated as follows:

$$G_2 = \frac{n-1}{(n-2) * (n-3)} * [(n+1) * g_2 + 6]. \qquad \text{Eq. (7)}$$

We can divide G_2 with standard error of kurtosis (SEK) to get the test statistic for kurtosis:

$$test\ statistic\ (kurtosis): Z_{g_2} = \frac{G_2}{SEK}, where\ SEK = 2 * SES = \sqrt{\frac{n^2 - 1}{(n-3) * (n+5)}}. \qquad \text{Eq. (8)}$$

Now, we calculated the D'Agostino-Pearson omnibus test statistic finally:

$$K^2 = {Z_{g_1}}^2 + {Z_{g_2}}^2. \qquad \text{Eq. (9)}$$

This K^2 follows χ² distribution with *df* = 2. Based on the K^2 test statistic value, we need to calculate $\chi^2_{(df=2)} > K^2$ from the χ^2 with p values (S*tep 6*). The lower this *p-value*, the higher the chance to reject H_0. The *p-value* is always between 0 and 1, it can be interpreted based on the following rule of thumb:

- **Small p-value (*p* < 0.05)**. This case means there is strong evidence against H_0. Reject H_0 is advisable.
- **Large p-value (*p* > 0.05)**. This case indicates weak evidence against H_0, rejecting H_0 is not advice.
- **If p value is close to 0.05**. Making the decision can be hard in this case. Research methodologies recommend to publish exact *p-value*.

In *Step 7*, if we have small *p-value*, we can conclude that the data set is not a normal distribution, otherwise (large *p-value*) it follows the normal distribution.

2.1.2. Bartlett test

Bartlett test is a homogeneity test for variances [83]. The assumption is that the data set must follow the normal distribution (*Step 0*). H_0 states that the variances are equal, H_1 indicates that at least one variance differs (*Step 1* and *2*). The α means the same as before (*Step 3*). As we will see in *Step 5*, the *b* test statistic follows the χ^2 distribution with *df* = *k* – 1, where *k* is the number of sample (*Step 4*).

The *b* test statistic is calculated as follows (Step 5):

$$b = \frac{(N-k) * \ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1) * \ln(S_i^2)}{1 + \frac{1}{3*(k-1)} * \left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)}, \qquad \text{Eq. (10)}$$

where $N = \sum_{i=1}^{k} n_i$, S_i^2 is the sample variances and $S_p^2 = \frac{1}{N-k} * \sum_{i=1}^{k}(n_i - 1) * S_i^2$ is the pooled estimate for the variance. If the *b* test statistic is smaller than the χ^2_{k-1} *critical value* (*Step 6*), then we can conclude that the variances are homogenous, otherwise, they are not (*Step 7*).

2.1.3. Analysis of Variances (ANOVA)

The ANOVA is used to determine whether the groups' averages are different or not and it is applied widely in different scientific fields. There are three assumption of ANOVA adaptation (*Step 0*):

1. Sampling must be done randomly;
2. Each group must follow the normal distribution (normality check);
3. Variances must be statistically equal (homogeneity check).

The null hypothesis H_0 states that the averages are equal statistically and the alternative hypothesis H_1 declines the equality statistically (*Step 2* and *3*). The significance level has the same meaning (*Step 4*). Since we have *k* groups and in each group there are *n* values, therefore we have df_1 and df_2 (*Step 5*). The first regards the number of groups, therefore df_1 = k – 1. The second is related to individual group values, in each group, the *df* is *n* – 1, we have *k* groups, so $df_2 = k * (n-1)$. As we can see in Table 5, the *F* test statistic follow the Fisher distribution with *df₁* and *df₂*. The given *critical value* can be sought in Fisher table with $F_{(df_1,df_2)}$.

The calculation process of ANOVA is summarized in Table 1.

Table 1: ANOVA test statistic calculation process [84]

Difference	Total differences	Degrees of freedom	Average differences	F value
Inner in group	*SSB*	*k – 1*	*MSB*	$\frac{MSB}{MSW} = F$
Outer in group	*SSW*	*k * (n – 1)*	*MSW*	
Total	*SST=SSB+SSW*			

The following expressions were used: $SSB = \sum_{j=1}^{k} n_j * \ (\bar{x}_j - \bar{\bar{x}})^2$ and $SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{i,j} - \bar{x}_j)^2$, *k* stands for the number of columns, *n* is the number of rows, *F* is the test statistic.

If the *F* test statistic is smaller than the $F_{(df_1,df_2)}$ critical value, then we can conclude that the sample means are equal at the significance level α (*Step 6* and *7*). If *F* is bigger, then the means are not equal.

2.1.4. Regression techniques

In our research, regression is analyzed in the view of RC. The main goal is to find the entity which best fits for the given data points P$_i$(*x*, *y*). The kind of entity is dependent on what kind of regression we use.

Although regression is not a statistical test, it has assumptions as well [85]:

1. The values of *Y* are independent, in other words, the observations are independent.
2. For each $x_i \epsilon X$ values, the $Y_i|_{x_i}$ values distribution is normal.
3. Each $Y_i|_{x_i}$ variances are equal, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_n^2$.
4. The error is a random variable with normal distribution with expected value (μ) 0, and given variance (σ^2), $\varepsilon \sim N(0, \sigma^2)$.

In this assumptions, *X* and *Y* are populations (not samples), and one x_i has one Y_i population with *l* elements, $i = 1 \dots n$, and *n* is the number of points.

In our research, we used the regression techniques summarized in Table 2.

Table 2: Used regression techniques [86]

Type	Sough entity	Solving equations
Linear	$y = a * x + c$	$a\sum x_i^2 + b\sum x_i = \sum x_i * y_i,$ $a\sum x_i + b * n = \sum y_i,$
Quadratic	$y = a * x^2 + b * x + c$	$a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2 = \sum x_i^2 * y_i,$ $a\sum x_i^3 + b\sum x_i^2 + c\sum x_i = \sum x_i * y_i,$ $a\sum x_i^2 + b\sum x_i + c * n = \sum y_i,$

Exponential	$y = a * b^x$	$\log y = \log a + x * \log b\,, originate\ to\ linear$
Logarithmic	$y = a + b * \ln(x)$	$a = \frac{\sum y_i - b \sum \ln x_i}{n}$ $b = \frac{\sum y_i * \ln x_i - \sum y_i * \sum \ln x_i}{n \sum (\ln x_i)^2 - (\sum \ln x_i)^2}$

Since we can always perform these calculations, i.e., we always find the best fitting entity, the quality of this entity is a question. In other words, we need a measurement value, which show us how good the fitting quality is. This quality is measured by coefficient of determination (*r^2*). The following equation is used for every type of used regression techniques to get *r^2*:

$$r^2 = 1 - \frac{SSE}{SST}, \qquad \text{Eq. (11)}$$

where $SSE = \sum_i^n (y - \hat{y})^2$,

$\hat{y}$ is the predicted value based on the best fitting entity,

n is the number of points,

$$SST = \sum_i^n (y - \bar{y})^2, \qquad \text{Eq. (12)}$$

where $\bar{y}$ is the mean of the measured y_i.

For the r^2 value, there are several rules of thumb to decide how strong or week the correlation is. One rule is to divide (0-1) interval into four sections:

$0 < r^2 < 0.25$, there is no connection;

$0.25 < r^2 < 0.5$, the connection is weak;

$0.5 < r^2 < 0.75$, the connection is satisfying;

$0.75 < r^2 < 1$, the connection is strong.

According to other rule, $r^2 < 0.5$, the connection is weak or there is no connection and $r^2 > 0.8$, the connection is strong.

2.2. Literature overview of Decision Support Systems

2.2.1. History of DSSs

DSSs' origin goes back in the mid of 20th century when military goals were dominant. One of the oldest DSS was SAGE (Semi-Automatic Ground Environment). This system was designed to unify different images about a wide area; it was used during cold war. With computers in a network; SAGE was the largest computer ever built. According to another approach, the first DSS can be originated in LEO I (Lyons Electronic Office) in 1951. The task of this DSS was to handle daily orders, calculating production requirements and it had some reporting function. In the public sector scientists began to study methods and computerized quantitative models, which can support decision support [87, 88, 89, 90]. However, the military using was still in focus, SAGE was used until ‘80s. One of the first dissertation related to DSS was introduced by Morton [91]. It involves DSS building and implementation with computer, including demo decision making for management. In the industry, more DSS development was started but their operation how started in later decades. In parallel, several scientists started their research which led to different theoretical DSS frameworks.

In ‘70s, DSS were evolving together with technological conditions and used to support different business processes DSS was launched in portfolio management [92]. Brandaid DSS was implemented for decision support of marketing [93]. Keen and Scott Morton’s book titled “Decision support systems: an organizational perspective” was one of the first architectural, models and methods summary on the field of DSS [94]. In this decade, the first definitions of DSS appeared as well. To the end of decade, the Executive Information System (EIS) and Executive Support System (ESS) definitions were also born [95]. By the end of the decade, researchers used not just the term Decision Support System but wrote about its “evolution” [96].

In the ’80s, DSS implementation and design pattern frameworks were introduced. The framework of Bonczek et al. contains four main parts: (1) a language system module, on which the DSS can be programmed, (2) a presentation layer, which is responsible for the visualization of results, (3) a knowledge system, which is a container to store the knowledge related to problem solving, and (4) business logic, which solves the specific problem for a decision session [97]. Another important milestone of this decade was created by Sprague [98]. Based on his paper, further explanation of building effective DSS was discussed in [99]. It is an overview of all entities and methodologies, which can be the parts of a DSS. For example, how the data can be handled, which design patterns exist, which analysis are available, which structural and architectural possibilities can be implemented. This research by Sprague and Watson determined the next decades and led to the phenomena of DSS-generator. Based on frameworks, several DSSs were built. IFPS (Interactive Financial Planning System) was used widely until the mid of 90s and this DSS is used in the education as well [100]. Different class of DSS is also created. Group Decision Support System (GDSS) term was born by DeSanctis and Gallupe who developed a GDSS called SAMM [101]. The Spatial Decision Support Systems (SDSS) can also be originated from the end of ‘80s [102]. The technology point of view supported the DSS evolution as well. IBM launched DB2 on its MVS mainframe in 1983. At the end of the decade, more handbooks and publication related to DSS and more future directions were publicated [103, 104]. In 1989, Gartner group proposed the term Business Intelligence as an umbrella, which provides “concepts and methods to improve business decision making by using fact-based support systems”. IBM Data Warehouse architectural and theoretical background was created [105].

From 1990 to 2000, the terms “Spatial Decision Support System”, “Business Intelligence” and “Data Warehouse” were spread all of the world [106, 107, 108]. Related to these terms, Codd et al. defined OLAP [109]. By 1997, the world’s biggest production data warehouse, called Teradata, was built by the firm Walmart [110]. The system is still in operation and the firm is one of the biggest decision support system vendor. Teradata was the first implementation, which could handle very large data volume, i.e., the first system with big data solutions. The GDSS was also evolved in the ‘90s. Shakun proposed an evolutionary system design for GDSS [111]. The requirements of DSS architecture and DSS roles during the whole decision process were discussed and defined [112, 113]. Another big impact was the appearance of the World Wide Web in the mid ‘90s, when every bigger DSS vendor started to develop web-based DSS solutions related to their “old” systems.

After 2000, decision support system arrived to various fields of science. Considering the new available technology tools, Car et al. discuss about the new generation of DSS [114]. They took Web 2.0 and modular developing into their focus and proposed a new classification of DSS. Besides the evolving technology, new conceptual approaches arrived: a new Hypothesis Management Framework was proposed by Gosliga and de Voorde [115]. In 2007, Power and Sharda stated that model-driven DSS are mainly built based on quantitative techniques [116]. Multi-criteria decision analysis come to the front. Fuzzy, multi-criteria approach, analysis concepts and other earlier technologies can be combined with web and other new technologies [117, 118, 119].

DSSs are spreading out to more and more scientific field e.g., medicine, biology, economics, earth science and forestry.

In medicine, Bates et al. summarized ten commandments to build evidence-based medicine DSSs [120]. These commandments sum up the specific properties of medicine DSSs:

1. *Speed*. Mainly in medicine, there are a lot of situations when time is critical.
2. *Anticipate needs and deliver in real time*. This means that sharing information electronically is not enough, the information must be anticipated by the system.
3. *Fit into the user's workflow*. Based on the authors' experiences, guidelines and alerts were rarely used by the users. Therefore, understanding clinical workflow by the user is important. In other words, DSSs are needed to focus on patient as well.
4. *Little things can make a big differences*. In the general viewpoint of DSS, developers intend to suggest that a given decision is the right one based on the given characteristics. But from the viewpoint of human factors, little parameter value differences can make big differences in the patient's body.
5. *Recognize that physicians will strongly resist stopping*. This means that if clinicians must make a decision their act is based on their belief mainly. For example, if they have to make a decision between two treatments *a* and *b*, they choose *a* for example based on their belief even if there is no evidence that *a* is better than *b* in the given situation.
6. *Changing direction is easier than stopping*. DSS can propose possibilities to change the treatment direction, if doctors are not sure about the diagnoses or treatments.
7. *Simple intervention work best*. Simple and easy interpretation of information can be the best as described in Point 3, but some modification possibilities must be implemented. According to the authors' example, "use aspirin in patients' status-post myocardial infarction unless otherwise contraindicated" message from DSS is not usable without some modification.
8. *Ask for additional information only when you really need it*. Cases when clinicians need particular information must be differentiated from cases when the providers do not give a piece of information about the patient, i.e., more irrelevant information can hide the matters of diagnosis or treatment.
9. *Monitor impact, get feedback, and respond*. Authors proposed this commandment for *action-based* design of this kind of DSS. Actions must be performed at every end of a process related to the patient, e.g., a treatment ended.
10. *Manage and maintain your knowledge-based systems*. It is critical to check the decision accuracy and give feedback to the system and simultaneously improve decision support processes based on clinicians' experiences.

Another medicine DSS methodology was proposed by Sim at el [121]. Based on methodologies, specific DSSs were built. The developing of DXplain was started in 1987, but further developments were added by Edward et al. in 2005, including transition to the web, expansion of database and newer feature such as focus and disease comparison [122]. Graber and Mathew proposed Isabel [123] and their results were the following "The clinical decision support system suggested the correct diagnosis in 48 of 50 cases (96%) with key findings entry, and in 37 of the 50 cases (74%) if the entire case history was pasted in. Pasting took seconds, manual entry less than a minute, and results were provided within 2–3 seconds with either approach."

Beside the classification precision, speed is also an important factor which is in accordance with the first commandments of Bates et al. This property is a medicine DSS specialty. Another class of medicine DSS is based on Fuzzy logic and neural networks, because there are more less-structured problems in this scientific field. Doctors' knowledge and experiences can be determinative parameters to support the decision, i.e., make the diagnoses. Saleh et al. proposed a DSS with Fuzzy logic to detect breast cancer [124]. Gago et al. developed a knowledge-based system called INTCare [125].

The origins of DSS can be related to economy (e.g., LEO I) and the economy is determinating it in our century as well. In 2002, Power summarized basic concepts about economy DSSs for manager [126]. He introduced an extended DSS framework, decision-making processes, design and development process as well as architectures. Goodwin et al. proposed a DSS for Quote generation [127], which architecture is summarized in Fig. 1.

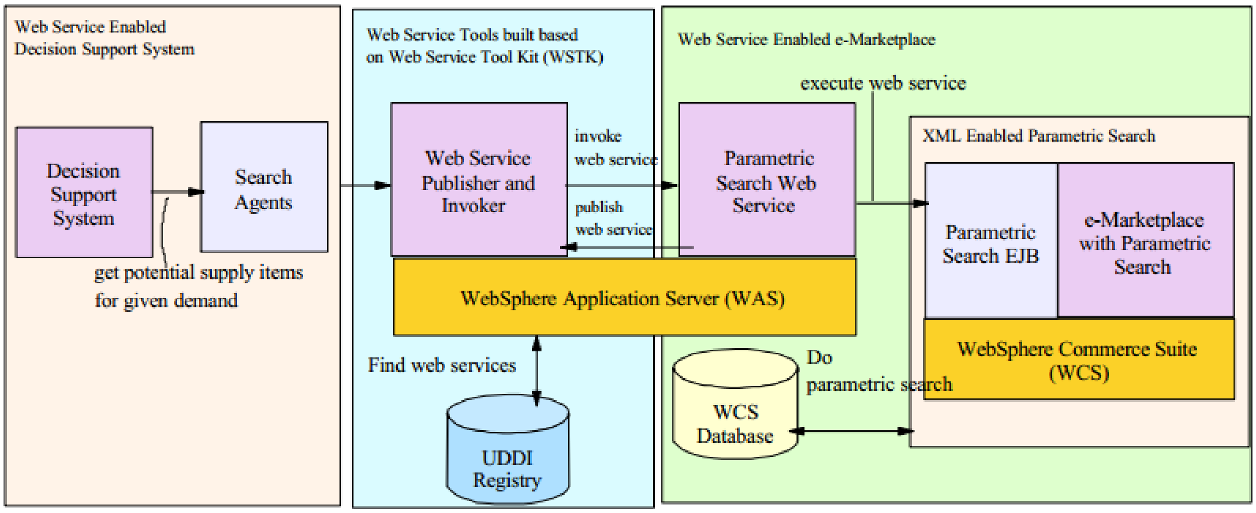

Web Service Enabled
Decision Support System
Decision Support System
Search Agents
get potential supply items
for given demand
Web Service Tools built based
on Web Service Tool Kit (WSTK)
Web Service Publisher and Invoker
invoke web service
publish web service
WebSphere Application Server (WAS)
Find web services
UDDI Registry
Web Service Enabled e-Marketplace
execute web service
Parametric Search Web Service
XML Enabled Parametric Search
Parametric Search EJB
e-Marketplace with Parametric Search
WebSphere Commerce Suite (WCS)
Do parametric search
WCS Database

Figure 1: Architecture of Quote generation DSS implemented by Goodwin et al. [127]

There are other systems to support decision for stock management. Luo et al. created Multi-Agent System called MASST for Stock trading early 2000 [128]. Kulak proposed FUMAHES system to support decision for material handling equipment [129]. This system contains five modules:

1. *Database for material handling*. It contains move and storage equipment types, which are trucks, vehicles or storage systems for example
2. *Database for manufacturing system requirements*. This module stored "classic" data about material equipment handling.

3. *Knowledge base*. FUMAHES authors studied literature and communicate with experts related to material equipment handling to determine rules. This rules are stored in a knowledge base.
4. *Inference engine*. This entity makes a connection the modules before and search solution candidates.
5. *Multi-attribute decision making module*. Final decision based on candidates is made with this module.

Wen et al. built an automatic stock decision support system [130]. They used box theory and support vector machine techniques to analyze buy-and-sell operation related to Microsoft and IBM. The used process can be seen in Fig. 2.

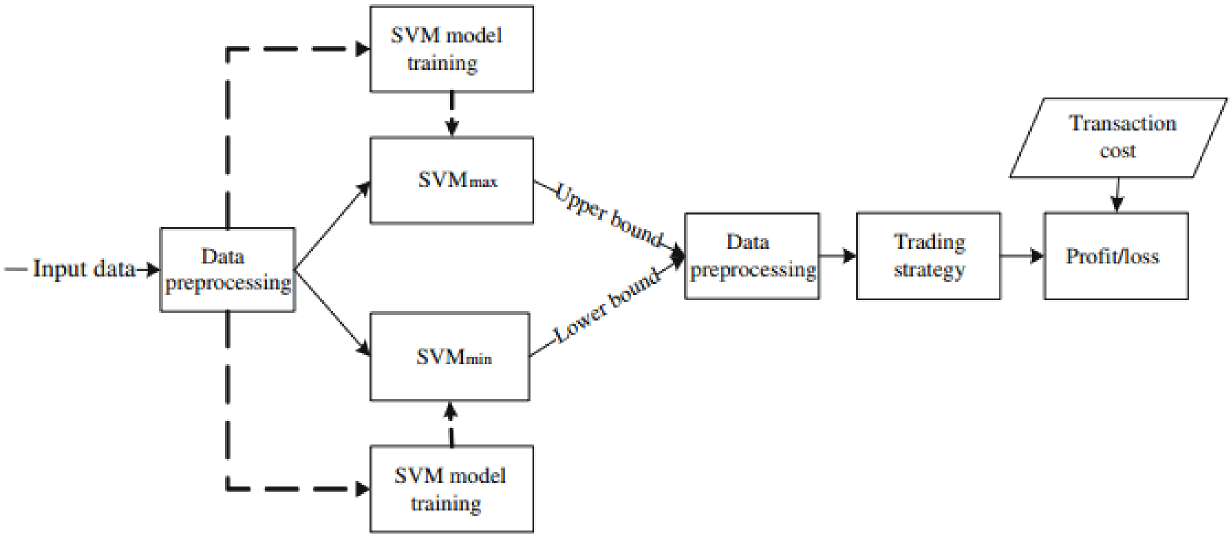


Figure 2: Automatic stock DSS implemented by Wen et al. [130]

Vincent proposed a DSS called Multi-level and Interactive Stock Market In-vestment System (MISMIS), which can perform forecasting based on time series [131]. Istudor and Dutá proposed a web-based group DSS [132]. Based on initialize parameters, core equity, loan capital, turnover, operating income, operating expenses, interest expenses and profit tax, with nine self-developed equations, the system is capable to determine leverage effect of indebtedness and some details related to that. In portfolio selection, Ghasemzadeh and Archer created the PASS system [133].

There were more systems related to the basic enterprise resource planning processes, but DSS with ERP II approach and supply chain management appeared as well. ERP II supports cooperation between companies, i.e., a company allows another company to access some data. For example, a vendor is authorized to see the customer company purchase orders. The customer company does not need to deal with purchase process and not only the items will be in the right place in time (Just in Time), but the vendor can propose its inner processes based on data related to customer’s needs. Achabal et al. dealt with this example and created a system to support vendor management inventory processes [134]. Their research is related to another phenomenon named Supply Chain Management (SCM). Inside a company, the main processes are connected to each other, and they are imaged mainly horizontally. However, the SCM shows the connection between the various types of companies and represents it vertically instead. Based on SCM, Wang et al. used analytic hierarchy process and multi criteria technique to support product-driven supply chain selection [135]. Biswas and Narahari developed decision support for supply chains through object modelling (DESSCOM) system [136]. This DSS contains two main parts: (1) DESSCOM-model, which can create the SCM at the desired abstraction level and (2) DESSCOM-Workbench, which solves the problem with various decision tools. Ezzeddine et al. proposed an agent-based framework called i-SEEC to support co-operation between companies in supply chain [137]. This solution is handled by ontologies, which is a good way to universality.

In forestry, Battaglia et al. investigate carbon, water and nitrogen model for forest growth and developed CABALA DSS [138]. The main goal of this system is analyzing and forecasting biomass allocation, i.e., simulating the growth of young trees. It is based on experiments (data), DSS (i.e., CABALA) and knowledge capture triple, and experts are in the middle of this triple. Orshoven et al. proposed a DSS based on similar attributes named as Afforest, which can answer the 'Where', 'How', 'How long', and 'What if' questions related to environment performance (EP) of afforestation such as carbon sequestration in soil, nitrate leaching and groundwater recharge [139]. It was further developed by a Belgian research group who introduced ForAndesT which has relational database [140]. The Belgian team continued developing the system to OSMOSE, which is a framework to generate decision possibilities for land use planning [141]. The database structure and the analyzed questions are similar as in ForAndesT system. The next Belgian DSS was Sim4Tree, which has functionality such as forest development, ecosystem services, economic performance, user-defined rules as well as climate scenarios [142]. The system architecture follows the three-layer design concepts as it is illustrated in Fig. 3.

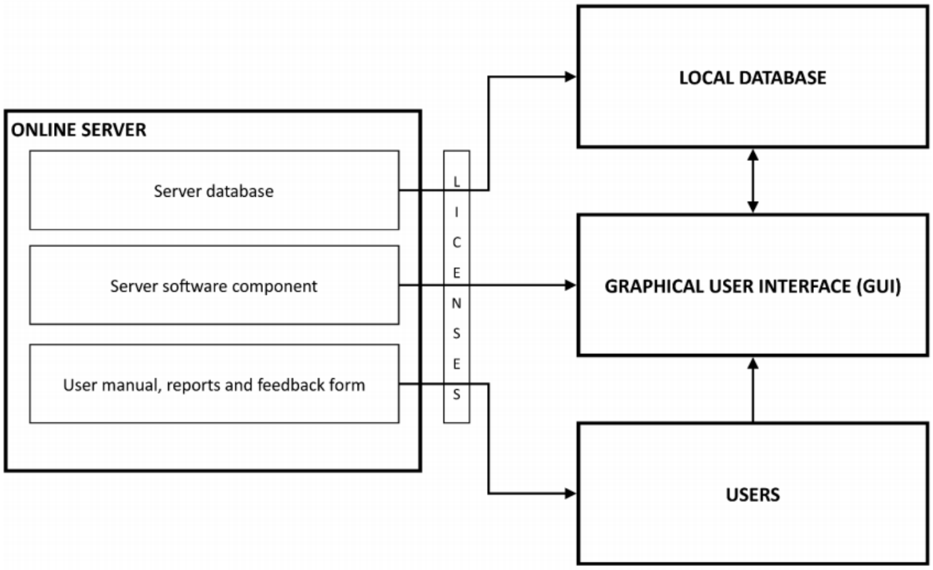

ONLINE SERVER
Server database
Server software component
User manual, reports and feedback form
LICENSES
LOCAL DATABASE
GRAPHICAL USER INTERFACE (GUI)
USERS

Figure 3: Sim4Tree architecture [142]

In Finland, where forestry is very important, MELA was developed by Finnish Forest Research Institute to support decision for production and management scenarios related to the overall defined goals [143]. The system has two main parts; the MELA Core and Extensions. The core can connect to the extensions. An extension can be defined by the users. This architectural design can be powerful, because scientists using MELA can implement their extension only. Twery et al. extended their NED project with new capabilities related to integrated forest ecosystem management. NED-2 focuses on goal-driven decision process, which means that land parameters are needed to be known and the rules defined by the users has the highest priority around this parameters [144]. Their system architecture is summarized in Fig. 4.

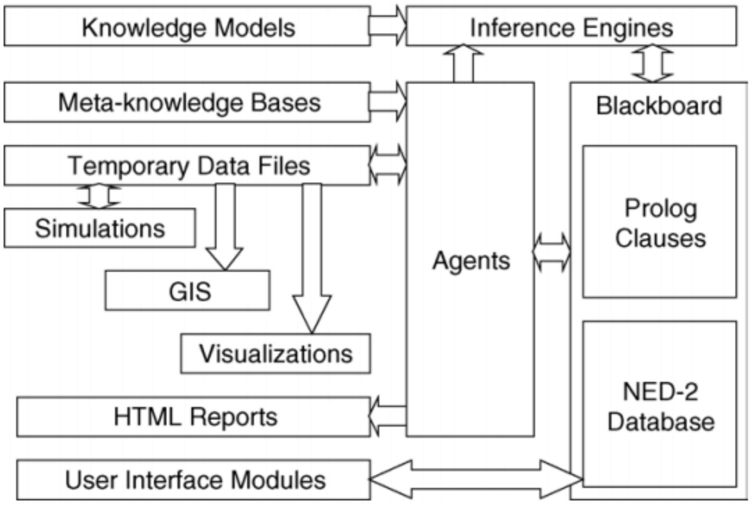


Figure 4: NED-2 architecture [144]

DSS were developed on another fields of forestry as well. Bonazountas et al. introduced their system in order to support decision for forest fires [145]. Their DSS contains eight modules:

- *The Data Acquisition (DA) module*. It handles data sources and transform them into a common format.
- *Satellites on area*. These are not "inner" modules of the system. However, forest fire decision support process must be started from these satellites, which provide data for DA. Meteorology and other data source can also be part of this data collection phase.
- *The Fuel Mapping (FM) module*. It supports to identify areas where fuel for fire could be.
- *The Scenarios Generation (SG) module*. It performs scenarios based on available data and different rules defined by the user.
- *The Socio-Economic Risk characterization module (SRM)*. It determines socio-economic risks in the area.
- *The Probabilistic Planning (PP) module*. A set of methods of analysis, scenarios are created by *PP* module.
- *The Valuation (VAL) module*. Based on economic, environmental and social scores, it computes a ranking.
- *The User Interface (UI) module*. It provides data management interface and make the connection between the user and the system.

This DSS was introduced and validated in the island of Evoia, Greece, which area is very vulnerable to forest fire. Kok et al. created Elbe-DSS to support decision for integrated river-based management and the system was tested in the German section of the river Elbe [146]. Using the system, four main problems were analyzed such as water quality, flood risk, river navigation and riverscape ecological value problem.

2.2.2. DSS definitions

DSS has many definitions. One of the first definition is created by Little [147]. He stated that DSS is a "model-based set of procedures for processing data and judgments to assist a manager in his decision making". It can be seen that model-driven DSSs were the first kind of DSSs. In [148], DSSs are "interactive computer-based systems that help decision makers utilize data and models to solve unstructured problems". Keen and Scott Morton defined DSS as a "couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semi-structured problems" [94]. This definition indicates a manager's point of view and therefore it is related to business decisions rather. Mann and Watson defined DSS as "an interactive system that provides the user with easy access to decision models and data in order to support semi-structured and unstructured decision-making tasks [149]. According to Bidgoli, DSS is "a computer-based information system consisting of hardware/software and human element designed to assist any decision-maker at any level". However, the emphasis is on semi-structured and unstructured tasks [150]. Finley interprets DSS phenomena in a wider sense: it is "a computer-based system that aids the process of decision making" [151]. Another definition is more specific, Turban stated that DSS is "an interactive, flexible, and adaptable computer-based information system, especially developed for supporting the solution of a non-structured management problems for improved decision making. It utilizes data, provides an easy-to-use interface, and allows for the decision maker's own insights." [152]. In 2005, Turban developed his definition: "[DSS] is a computer-based system that combines models and data in an attempt to solve semi-structured and some unstructured problems with extensive user involvement" [153]. According to Sprague and Watson, DSS is a computer-based system, which help solving ill-structured problem based on data and analysis models based on that data [154]. Sauter stated that data with different sources are uploaded in the DSS database layer and analyses can be performed with specific models and methods [155]. This definition is related to Data Warehouse and ETL (Extract, Transform, Load) [see later]. However, there are less trivial definitions. Keen stated in 1980 that "there can be no definition of Decision Support System, only of Decision Support" [156]. Later in 1998, Schroff emphasized this nature of DSS and he stated that there is no way to give a trivial DSS definition [157]. There is no commonly accepted definition of DSS nowadays.

2.2.3. DSS classification and components

Over time, DSS types have been classified by many researchers with different point of view. In late '70s, Donovan and Madnick proposed two classes of DSS: (1) ad hoc DSS and (2) institutional DSS [158]. Alter identified seven individual classes based on his research [159]. Hackathorn and Keen created three distinct classes for DSS [160]:

1. *Personal DSS*. It is used for personal problem solution.
2. *Group DSS*. Group of researchers can make a decision together with it.
3. *Organizational DSS*. It refers not just to one group but the entire organization.

Holsapple and Whinston defined five classes: (1) text-oriented DSS, (2) database-oriented DSS, (3) spreadsheet-oriented DSS, (4) solver-oriented DSS and (5) rule-oriented DSS [161]. Hättenschwiller differentiates three DSS classes [162]:

- *Passive DSS*. These systems cannot produce explicit decision at the end of the analyzing session.
- *Active DSS*. Implicit solutions are provided by these systems.
- *Cooperative DSS*. Users can modify and re-think the results generated by the DSS. This result can be sent back to the DSS for validation. This cycle can be repeated until the desired result is found. This kind of DSS concept will be used by our research as well.

Power proposed enterprise-wide DSS, which support managers to make decision based on enterprise data contained in data warehouse, and desktop DSS, which serves personal decision making [163]. In [164], Power developed his DSS classification, which is the most current DSS classification nowadays. According to it, there are five classes based on history evolution:

- *Model-driven DSS*. Early DSS were model-driven DSS. Financial and simulation models based on mathematics are in the focus.
- *Data-driven DSS*. Different experimental data are uploaded to the database (or data warehouse), these data can be queried. Then analyzing method can be applied.
- *Communication-driven DSS*. Similar to GDSS, decision making is supported by communication between clients. Network and communication technologies are the main components of these systems.
- *Document-driven DSS*. These type of DSS focus on documents such as scanned documents, images or voice and video materials. Document search and document handling are the first priority in the system.
- *Knowledge-driven DSS*. The system can suggest solution for the given problem. Technical and professional forums and blogs can be part of these DSSs.

According to [165], a new classification is needed based on the new technologies appeared after 2000. The authors differentiate five plus one classes of DSSs based on their network paradigm:

- *None*. Simply standalone application.
- *Single link*. A data collecting application, for example a machine or sensor.
- *LAN*. Information can be got on LAN only (local network data, machines, sensors).
- *Enterprise network*. DSS are used on enterprise-wide data sets or resources.
- *Internet*. Information are found on the Internet.

The author of [165] proposed a sixth class, which is the *Interoperably connected* class. This class supports new multiple interfaces, user-defined data sources and web service resources implemented with Web 2.0 and semantic web technologies.

Complete DSS framework was also proposed. Tripathi names four subsystems, which can be used to build a DSS [166]:

1. *Data Management subsystem*. Data are stored in this layer and it can be related to data warehouse.
2. *Model Management subsystem*. Set of models and methods, which can be performed on.
3. *Knowledge-Based Management subsystem*. This module can support user and/or organization knowledge to make better decision.
4. *User Interface subsystem*. This component is responsible for the communication between the decision makers and the system itself.

The interaction between these components is summarized in Fig. 5.

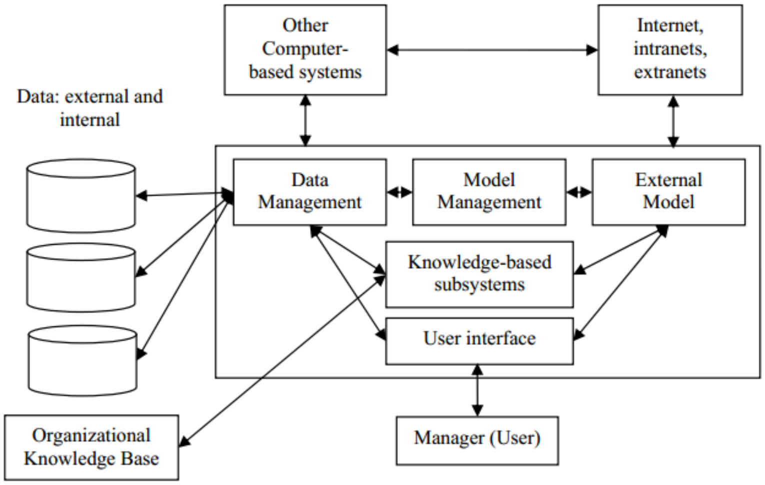

Other Computer-based systems
Internet, intranets, extranets
Data: external and internal
Data Management
Model Management
External Model
Knowledge-based subsystems
User interface
Organizational Knowledge Base
Manager (User)

Figure 5: Connection between components according to Tripathi [166]

According to Raheja and Mahajan, there are hardware and software components in DSS [167]. The software components can be divided as Data Base Management System and Model Base Management System. The former supports similar purposes as Tripathi's: store data, perform queries, updating data and security hierarchy. The Model Base Management System contains two sub-components:

1. *Pre-written computer program*. This group is a set of pre-defined models and methods, e.g., regression analysis. However, it can also be a self-programmed algorithm for special manager purposes as well.
2. *Model building blocks*. Some entities of pre-written computer program can be used to build an ad-hoc application.

The previous sections gave an overview about DSS solutions, definitions and components. Some entities were used in our new Universal Decision Support System concept. All of the applied models are summarized in the next section.

2.2.4. DSS-generator

According to Sprague and Watson, DSS generator is a "*computer software package that provides tools and capabilities that help a developer build a specific DSS*" [13]. The architectural overview is illustrated in Fig. 6. There are building blocks, i.e., tools, which can be combined with each other to build a specific DSS. Specific DSS is built for a specific problem, but it is less useful to support another kind of decision except what it was built for. If we would like to solve another set of problems, then we need to build another specific DSS with building blocks and DSS generators. This means that DSS generator is a framework and using and combining framework elements all kind of specific DSS, i.e., any kind of problem can be solved and any kind of decision can be supported.

In [99] authors stated that there are "two basic objectives of the DSS Generator: (1) to permit quick and easy development of a wide variety of specific DSS; and (2) the Generator must be flexible and adaptive enough to facilitate the iterative design process by which specific DSS can respond quickly to changes". The UDSS concept not only extends this phenomenon but specifies "building blocks" more precisely.

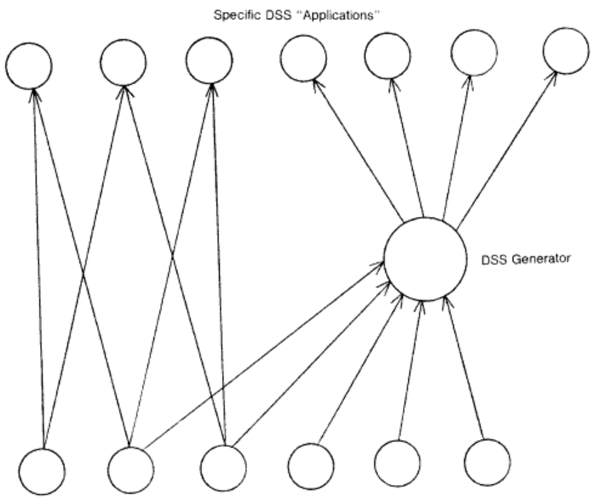


Figure 6: DSS-generator concept [12]

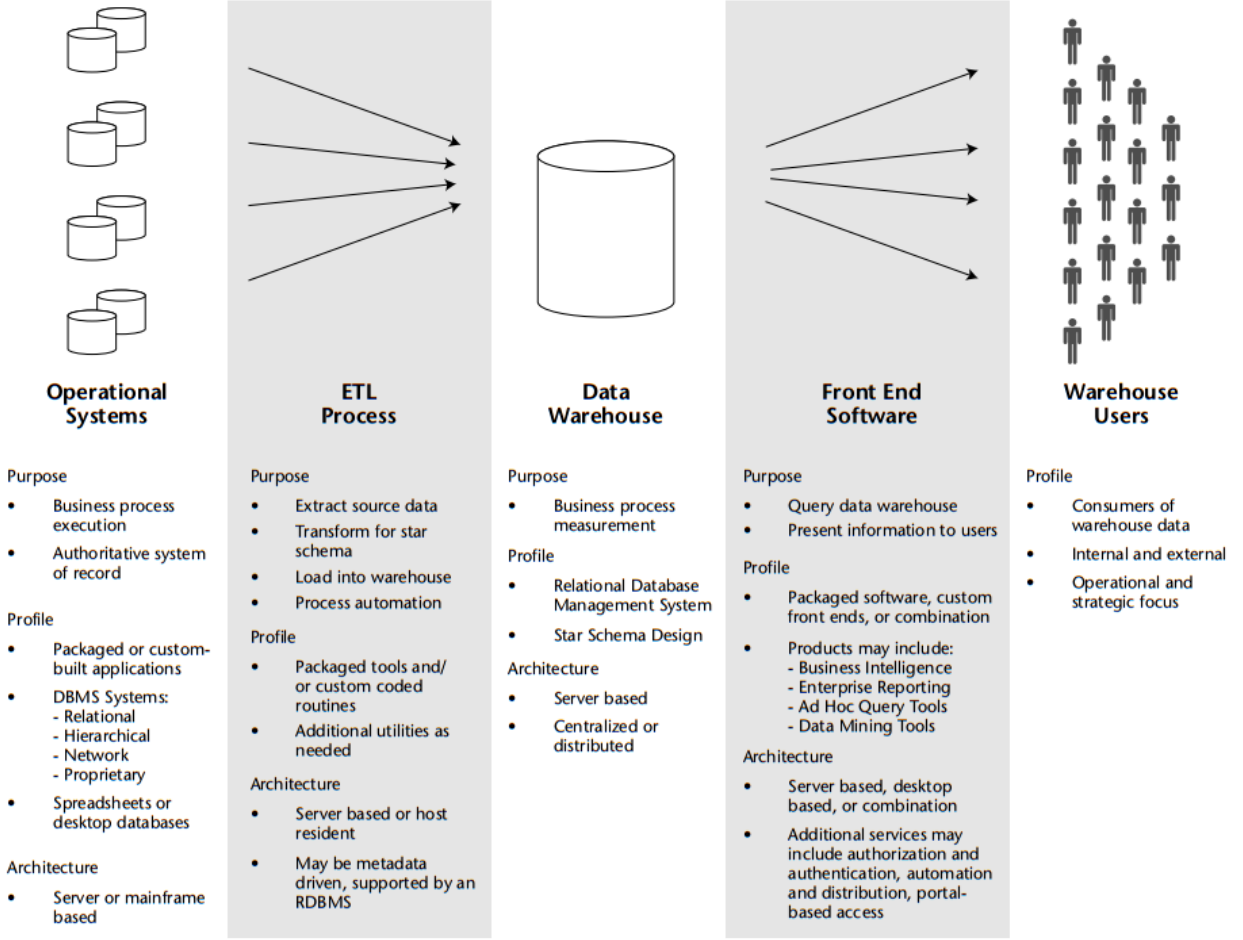
For example, Microsoft Office Excel is a spreadsheet-based application with data manipulation functions and an end-user productivity tool. Excel connectivity capabilities with other software and analysis possibilities are evolved a lot, moreover, own problem solving processes can be implemented with its VBA programming language. Since Excel can generate a specific DSS, therefore Excel can be viewed as a DSS-generator. However, if we would like to analyze specific data from measurement whose outputs are not Excel files, ETLs (Extract, Transform, Load) must be written for each data row. It is possible to store data in different Excel files, however, if a new data row is appeared, a new ETL must be written and this modification can effect on some other parts of the system as well. Another way that Excel is connected to other software, e.g., a Database Management System, however, in this case, the whole analyzing process is not in one system. This *Best-of-Breed* design concept means that every individual system is connected to each other with interfaces. However, each system has its own structure and process, one system cannot see into another solution, therefore the whole decision support processes is not integrated. From this point of view, this Best-of-Breed design pattern is less effective. In the view of UDSS, we would like to extend DSS generator, and create a concept, which can handle well-structured as well as ill-structured data sources and problems in one robust system.

Since DSS generator was introduced, many publications were written about this topic. Yeo and Nah proposed a DSS-generator and built a specific DSS for management game with their generator [168]. Dong introduced a self-developed web-based framework to integrate different components, such as data, models and visualization techniques [169]. Keenan proposed Geographic Information System (GIS) as a DSS-generator [170]. Savic et al. developed GANETXL, a general purpose DSS generator [171].

DSS-generator can also be related to Random Correlation (RC). Random Correlation means a factor of the proven result's endurance. If specific DSS is built with DSS-generator, then the decision support process can be affected by random correlation. Combining DSS generator tools, the random correlation chance can be increased. The level of randomness must be determined.

2.2.5. Data warehouse & ETL

Data Warehouse is a powerful tool to analyze data. There are several entities of the data warehouse concept, but in the point of UDSS view, architecture and data modelling are the most important. These two characteristics are well summarized by Adamson [172]. A typical data warehouse architecture is summarized in Fig. 7.



Operational Systems
Purpose
• Business process execution
• Authoritative system of record
Profile
• Packaged or custom-built applications
• DBMS Systems:
- Relational
- Hierarchical
- Network
- Proprietary
• Spreadsheets or desktop databases
Architecture
• Server or mainframe based
ETL Process
Purpose
• Extract source data
• Transform for star schema
• Load into warehouse
• Process automation
Profile
• Packaged tools and/ or custom coded routines
• Additional utilities as needed
Architecture
• Server based or host resident
• May be metadata driven, supported by an RDBMS
Data Warehouse
Purpose
• Business process measurement
Profile
• Relational Database Management System
• Star Schema Design
Architecture
• Server based
• Centralized or distributed
Front End Software
Purpose
• Query data warehouse
• Present information to users
Profile
• Packaged software, custom front ends, or combination
• Products may include:
- Business Intelligence
- Enterprise Reporting
- Ad Hoc Query Tools
- Data Mining Tools
Architecture
• Server based, desktop based, or combination
• Additional services may include authorization and authentication, automation and distribution, portal-based access
Warehouse Users
Profile
• Consumers of warehouse data
• Internal and external
• Operational and strategic focus

Figure 7: A typical Data Warehouse architecture [172]

In the center of the data warehouse concept stands data modelling. The most common model is the star scheme with two data tables types:

1. *Dimension tables*. They are organized around the companies' processes. In these tables, embedded structure like space and time is typical. These are dimensions in the traditional way as well, that is why it is named as 'dimension table'. In a company environment, space and time were the more frequent dimensions. There can be other dimensions, such as financial. In enterprise resource planning systems, we can define and handle tailor-made dimensions. The hierarchical implementation is very frequent in this kind of tables. For example, in the case of time, there are columns of time, hour, day, week, month, quarter, and year.
2. *Fact tables*. They contain data related to the transaction and foreign keys to the dimension tables. At data level, the term transaction means a unit of work, represent any changes in the database.

We use the wider meaning of transaction. It means all kinds of information processing, which can be divided into individual operations.

A typical star scheme can be seen in Fig. 8.

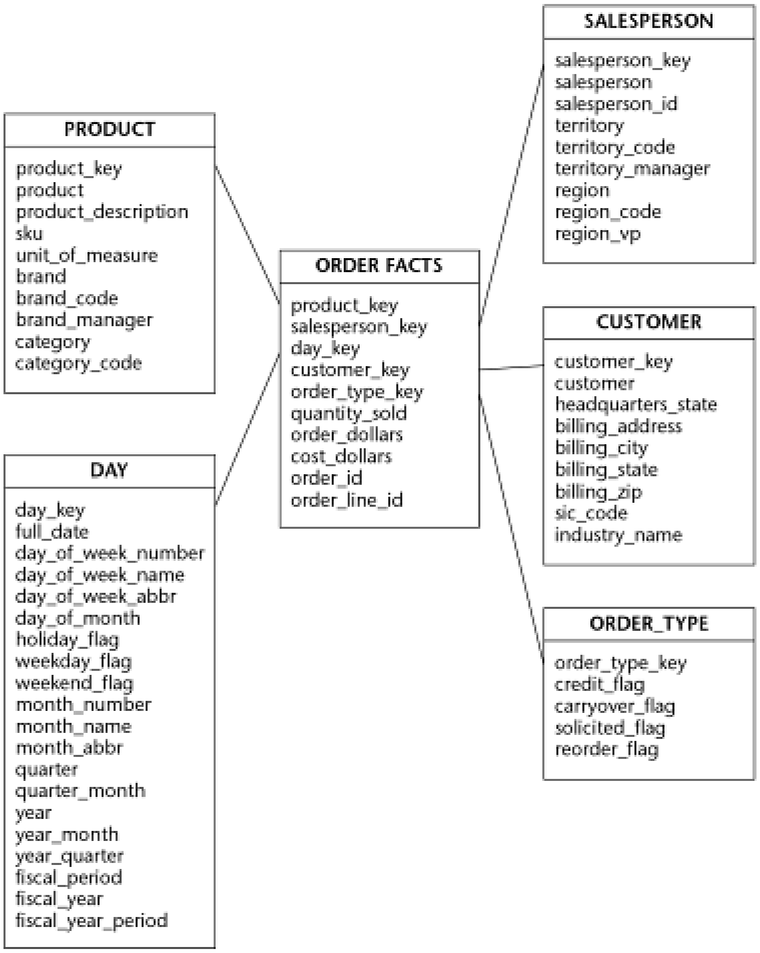


Figure 8: Star schema [172]

There is one fact table in the middle with columns to store transaction data and the foreign keys. Through these keys the fact table is connected to the dimension tables. In time, optimized developments of star scheme were created. One of them is the snowflake scheme where some dimension tables are normalized and new tables are created. The next step was the galaxy scheme where fact tables share dimension tables.

The data warehouse concept is used widely, but it was developed for business to handle large scale transaction data and to perform analysis on data. Companies developed their own data warehouse and used it for decision support. But if a star structure is defined for a given data warehouse, then the modification i.e., creating new dimension table or connections inside the data layer is hard and a re-defined data warehouse is needed. This can take a long time and re-design requests a new project generally. The traditional '*delete*' function is also missing, performing data version operations is more typical. The need of adding an extra data row is important in many scientific field. For example, we measured *n* data row to time *t*, but we would like to measure a new *n+1* data row from *t*. In this case, data warehouse structure re-design is needed. Data are bounded to each other with dimension (e.g., time, place) and characteristics. Insert another data row with other characteristics into the pre-defined data warehouse structure may cause a lot of changes in the structure, e.g., creating new tables, making connection between old and new tables etc. The situation is similar when a given data row's attributes are changed. The data warehouse concept is a methodology but for the given conditions (data format, measured data structure etc.) at time *t*. There is no methodology for building a system for universal purposes.

In Fig. 8, an ETL (Extract, Transform, Load) component can be seen, which is responsible to extract data from the original data source (binary files generally, outputs of the testers and/or sensors), transform the structure from the original to the data warehouse pre-defined structure and upload data with new structure into the data warehouse at the end. The problem is similar as earlier. A new data row does not mean only the redefinition of the data warehouse structure but a new ETL must be written as well. Except some cases in which the original data format is similar, but this is not the typical case. Based on our experiences with data of small and midsize companies, data of Geodetic and Geophysical Institute and data of Forest Research Institute, written proper ETLs were difficult. We do not state that a total universal ETL can be written, but ETL must be implemented in one integrated system. Our UDSS concept gives the possibility to create ETLs. Scientific fields have their own data (and used analyzing methods and research methodologies), but if we would like to analyze data rows coming from different scientific fields, we need to write new ETLs and re-design the given data warehouse.

Since it is possible to have large number of data, data warehouse is a good start to handle big data. We used ETL, OLAP entities and data warehouse building methodology. However, UDSS concept solves the previous mentioned problems with a new data structure and universal ETL concept.

Big data is a relatively new scientific area, experts state that 50-150 billion sensors will be available on the Internet in 2020. In this new field, the Industrial Internet and the Industry 4.0 was born in the USA and in Europe respectively. But decision making based on big data is not trivial. The more data we have the larger the possibility that random correlation occurs. According to some predictions, any kind of result can be created with large amount of data, which does not reflect the reality and can be led to false decision. More data does not secure precise result and it can cause the unwanted effect of random correlation.

2.2.6. Multi-criteria

The Multi-Criteria Decision Analysis (MCDA) is a tool that performs complex analyses, i.e., support ill- or non-structured decisions. A standard MCDA process is summarized by Dodgson [173]. The MCDA steps are as follows:

1. Define the environment of the MCDA (e.g., goals, key decision makers).
2. Define the possibilities, which enables to reach the goals
3. Define the objectives and criterion related to each option.
4. Rank the options with their expected assessments.
5. Define the weights between criteria and sign the importance level of each criteria.
6. Create decision alternatives (calculate alternatives, this operation is performed typically in a DSS).
7. Evaluate the MCDS results.
8. If necessary, perform analysis with changing scores or weights, or make the final decision.

During the MCDA process, different weights can be defined between parameters, and different scenarios can be created based on these various parameter values. After creating decision scenarios, an evaluation process is the next phase. It is performed based on the pre-defined evaluation model. At the end, scenarios are ranked and the final decision can be made.

These general steps are used by Tzeng et al. who analyzed alternative energies used in public transport such as busses [174]. They defined 11 parameters, such as energy supply, energy efficiency, air pollution, noise pollution, industrial relationship, costs of implementation, cost of maintenance, vehicle capability, road facility, speed of traffic flow and sense of comfort. Weights were connected to each parameter, and the result was that the best choice is the electric bus with exchangeable battery. A lot of MCDA algorithms exist, such as PROMETHEE, ELECTRE and Analytical Hierarchy Process (AHP). These MCDA algorithms seek and give a rank of scenarios starting with the best scenarios. Pohekar and Ramachandran made an overview about the common used MCDA methods [175]. Their result was that AHP method is the most used technique followed by PROMETHEE and ELECTRE algorithms. Ho et al. dealt with supplier selection [176] and stated that the most used technique is the individual approaches but integrated approach like AHP is also very popular. Baizyldayeva et al. dealt with MCDA at technical level and compare implementation of MCDA algorithms [177].

In the view of UDSS, MCDA can be performed at two levels. MCDA parameters can be viewed as data rows, and can be chosen freely. Moreover, the number of parameter can be changed easily, just new data rows must be queried during the analyzing process. At the second level, in the Logic, the proper techniques, e.g., PROMETHEE or AHP can be used. In a next iteration phase, a ranking algorithm can be performed. It is possible to implement new MCDA algorithms and/or ranking method as well. From the Multi Criteria technique, the iterative data manipulation characteristics and new techniques implementation possibilities are used in our UDSS concept.

MCDA can be affected by Random Correlation. Increasing the number of parameters and defining various weights between parameters can increase the chance of Random Correlation. The different evaluation models can lead to contradictory result, which can be originated to randomness. Optimal and sub-optimal results could be valid for the given situation, however, Random Correlation possibility level must be determined.

2.2.7. NoSQL

According to the definition of the community of NoSQL, “Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable” [178]. It means that NoSQL (sometimes it is translated with not only SQL) do not work with classic relational data models. Several characteristics apply such as schema-free and easy replication support. NoSQL is based on BASE concept, which contains three properties:

- *Basically available*. This property means that the new value of the data is not always available. It is possible that the data is in inconsistent or changing state.
- *Soft state*. The system state can be changed without any input.
- *Eventually consistent*. The system guarantees that all reads reach the latest write’s value, if there are no further updates.

In the standard “classic” database view, instead of BASE, ACID concept is applied:

- *Atomicity*. All tasks of a transaction are performed or none of them are.
- *Consistency*. All protocols of the system must be observed. Unlike BASE concept, the database must always be in the consistent state.
- *Isolation*. None of the transactions can access to the transactions that are in the unfinished state.
- *Durability*. If a transaction is complete, then it is marked as complete and it will survive any system failures.

There are several types of NoSQL, such as:

- *Key-value stores*. Simplest type, there is a simple programming interface with *get*(*key*), *put*(*key*, *value*) and *delete*(*key*) methods.
- *Document stores*. It stores semi-structured data. There are no pre-defined structures and join operation. It uses JSON (JavaScript Object Notation) data interchange format. It can be used for logging and as a data layer of Content Management Systems (CMS).
- *Wide column stores*. Similar to SQL database management systems, data are stored in columns. However, data are stored in key and value pair in each column. Sometimes, a timestamp is stored near key-value pairs. Because the columns, it is a pre-defined but because of the key-values pairs it is flexible schema.
- *Graph databases*. There is a $G = (V, E)$ graph, where V contains the set of vertices and E contains the set of edges. In this type of NoSQL, data are stored in the vertices and data relationships are stored in the set of E.

All of the biggest IT vendors have NoSQL solutions. Microsoft offers DocumentDB from the cloud Azure. IBM has Informix, Cloudant, Lotus/Domino software. Oracle NoSQL Database is based on the *key-values* pairs. But there are other NoSQL solutions. MongoDB, as well as Couchbase are document stores NoSQL implementations. Cassandra is used by Facebook and it is a wide column store. Graph NoSQL databases are HBase, OrientDB and Titan. Solution Hadoop is not just a database, but an environment, a software ecosystem, which can cooperate with NoSQL databases such as HBase.

All kinds of NoSQL are prepared for big data and they support fast data processing even in the case of big data. With the help of NoSQL, several data types can be stored. Therefore, NoSQL can be the data layer of UDSS. However, NoSQL has some disadvantage. The most NoSQL implementations do not support ACID behavior.

The use of NoSQL BASE concept for UDSS purpose is limited. Performing analyses after each other, updating the database with new results and semi results, updating data items can be complicated because of the *basically available* property. If only one wrong (non-updated) value are in the calculation process, the result of the analysis can be no fair. Other inconvenient characteristics also exist. For example, MongoDB has only restricted support for UTF-8, therefore handling string search in different languages can be a problem. In most cases, only the 64-bit operating system is suggested.

Having summarized, NoSQL is a great new direction and further UDSS data layer research can start. But the characteristics of NoSQL is not suitable for the current UDSS concept, so we choose the standard, long-standing relational database.

2.2.8. Business Intelligence

Business Intelligence (BI) is an extensive term. It means variety of models, methods and software solutions used to organize and analyze raw data. BI as a discipline includes online analytical processing, data mining, analyses' techniques and reporting. The main goal is to turn raw data into information, and therefore better decisions can be made. Implementing BI is complex and it can be very expensive. Therefore, small and midsize companies do not allow themselves to buy it. However, mainly every companies use BI at certain level.

A lot of vendors are on the market. Microsoft, Cognos, Business Objects, Oracle and SAS have their own BI solution to support all kinds of operations. For example, Microsoft has MSSQL server for building data warehouse including report services, SharePoint for supporting teamwork, Dynamics NAV as an ERP solution, Dynamics CRM for customer relationship management and Office package including Excel. Not all kinds of solutions must be purchased to get BI functions, it is possible to do analyses with Excel. The BI "parts" can be integrated with each other, for example, Dynamics CRM data can be integrated with Excel, customer behavior analyses can be done in Excel, the results can be uploaded into MSSQL. In general, BI solution can be combined with even such parts which are another vendor's implementation. This is an advantage in the view of UDSS, because BI parts as building blocks can be integrated and different analysis processes can be executed from data level to the reporting (presentation) level. Even the closed source BI solution can be extended with specific developments, for example, with unique analysis processes. The other BI services remains untouched and they can cooperate with the new unique analysis techniques.

There are some disadvantages of BI. The price of such a system can be really expensive. If we buy more and more BI services for specific decision support, then the price is increasing. The unique developments are also at high cost. The system installation and configuration need strong informatics background. A possible solution for that problem can be the cloud technology. In this case, a pre-configured BI environment can be used from the cloud. To access to cloud environment, we need only pay monthly fees. However, this fee is not so high than the installation, the configuration and the upkeep costs.

Another disadvantage is that researchers cannot see the whole operation in most of the cases. Researchers just input data items and then get the result, the operation of the method of analysis is hidden. We can assume that BI implementations are appropriate, but mistakes can be done with the best intension as well. The fixes come as updates. The implementation code cannot re-use or modify. This means that if a researcher would like to extend the given algorithm with one plus step, for example, divide *n – 2* instead of *n*, then it cannot be done. Including with this step, the whole algorithm must be re-implemented. The biggest problem is that we cannot check the methods' assumptions. For example, in Excel, a proper ANOVA is implemented. However, if we execute ANOVA in Excel, the process does not check ANOVA's assumptions (normality and equality of variances). There are Excel add-ons to perform these assumptions, however, in the most cases, we must pay for them. Otherwise, we must implement the assumptions in Excel. If we have several data rows, we would like to perform analyses on them, the whole analysis process is going more and more complex if we would like to check assumptions. Therefore, validation phase is difficult. Back to the "ANOVA in Excel" example, we must ask whether the assumptions have been checked or not. If not, the can be questionable. One possible solution is keep the source open. However, open BI solutions' support ends in most cases, because it is difficult to obtain money for open source BI.

BI solutions can be used for universal purposes, but they support business analyses in most cases. They are less for scientific purposes, since the deep validation needed for researchers has only a little support.

3. Specific research objectives and methodology

3.1. Specific research goals

The data sets have an important role in the view of Random Correlations. Having studied the used datasets, we can state that the number of data items can be seen as a large sample statistically, however, it is not enormously big. There are several definition of Big Data. The so called 3Vs definition is the most accepted. The *3Vs* term is an acronym, which means Volume, Velocity and Variety.

Volume. Data has exponentially growth. There are not just text files nowadays, but video, audio, large images and social channels data are also available. Large data eventuate the term Big Data.

Velocity. It implies data update property. Nowadays, we would like to see data items at once. For example, users pay attention recent social media data updates. However, they do not care data, which are older than one week for example.

Variety. Data sources can be stored in variate formats including Excel files, simple text files or CSV file format and any kind of binary file.

Our UDSS concept is mainly related to Variety property of the *3Vs* definition. UDSS concept is used in three use cases, which will be presented precisely later. Velocity can be also important from our UDSS point of view, but during the analyses of the given use cases, velocity property is less determining. Because of the amount of use cases' data, our environment can be considered as "big data inspired" instead of real big data. The "big data inspired" environment means that data volume is a statistical population, however, it is not so large that it eventuates performance problems at technical level. Therefore, standard database management systems can handle this number of data. There is less need to deal with performance, therefore the standard SQL-based relational database structure is chosen. After implementation of our decision support system, analyses of RC behavior can be started. Therefore, we have to solve two specific problems.

Problem 1. *Universal Decision Support System concept and architecture*. Taking universality in the focus, new design patterns must be applied. The problem is that many DSS are problem-specific and cannot be generalized without major changes at conceptual, logical and physical level. Database structure must be universal. It means that all kind of data must be stored in one database with one structure. Since data have different structure, the transformation between the original state and the new universal database must be ensured. Data are analyzed not only with one method but with many different ones. The problem is that yet undiscovered analyzing models cannot be implemented at present time. The set of analyzing method must be extendable and in the meantime, the other components of UDSS (for example, database and querying processes) must be untouched. Since analyzing can be very complex (e.g., applying data transformations, performing analyzing method and then performing another analyzing technique), the data manipulation methods can be performed after each other many times. Presentation of the results must be also part of the system. However, the different methods of analysis have different output structures. Each possible structure must be presented in the proper form to the analysts.

Problem 2. *Random Correlations*. Performing analysis processes in UDSS, we experienced that contradictory results can be produced. Based on the same data sets, the data can be manipulated in such a way that we get a given result. But using another analyzing process, another result can be produced. After both results had been proven, we followed the proper research methodologies. However, the different results led to different, sometimes contradictory decisions. We analyzed the circumstances, in which such kinds of result-pairs can occur. Since many algorithms can be executed after each other and some kinds of algorithms can be parameterized, the number of possible analysis is countless. The main question is that these countless analyzing possibilities including “big data inspired” environment can have an effect on the endurance of the results. Due to the continuously increasing data volume analysis with different methods, it is possible that the result can occur randomly. In this case, the UDSS approach (i.e., extending the DSS capabilities to perform much more analysis), the “big data inspired” environment, the complex research methodologies and Random Correlation face each other.

3.2. Generalization

In our Universal Decision Support system (UDSS) concept, we merged the previously discussed DSS approaches. The different DSS definitions have some common characteristics:

- Definitions are related to computers.
- DSSs help to solve structured, semi-structured or unstructured problems.

As we have seen before, there are many DSS classifications. According to Donovan and Madnick classification, there is no difference between ad-hoc and institutional DSSs in the UDSS point of view. UDSS is created to unify ad-hoc DSSs. Each ad-hoc system operation can be mapped to UDSS. The institutional class is a wider class. Analysis can be done with different point of view. The algorithms are behind the different scientific viewpoints. Therefore, Group, Institutional and Communication-driven DSS classes can be mapped to algorithmic level.

Holsapple and Whinston’s first three classes and Power document-driven class is based on documents, which can be stored in the data layer. These classes’ main operations are search and view operations, since the needed information are in one of the documents. To get the information, these documents need to be read. If documents are transformed into pure data (using character recognition for example), then we return to the “classic” level: pure data can be stored in data tables according to their proper types. Therefore, handling documents can be done by simply storage of documents or extract data from documents. This can be a part of a whole analysis starting from data extraction ending with pure data uploading.

Other classifications are based on models and methods. This approach is the nearest class to our concept. As the model-driven DSS, the UDSS focuses on data, mathematical proofed results, and the most important phase, the validation. There are some common aspects in the case of classification:

- DSSs are classified by the specific aspects of entities. These are mainly the systems which names end as “-driven”, e.g., river-driven, forestry-driven, disease-driven, data-driven DSSs.
- DSSs are classified based on decision types. These systems not always produce exact decision.
- Based on the previous two characteristics, there are mixed models.

The UDSS concept is based on the view of knowledge discovery process. Fayyad et. Al. define Knowledge Discovery in Databases as “the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [179]. The basic process is illustrated in Fig. 9.

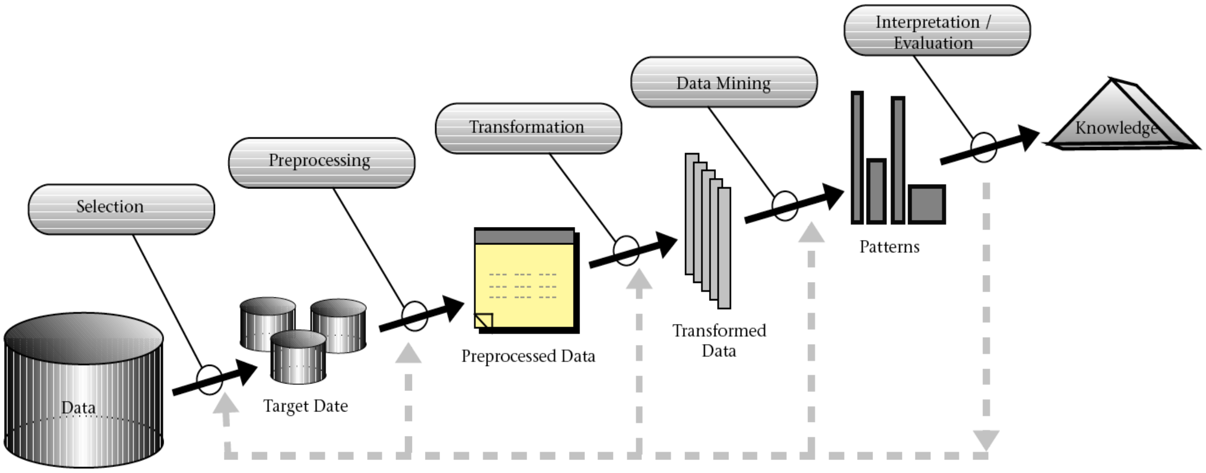


Figure 9: Knowledge Discovery process [179]

In the concept of UDSS, The Knowledge Discovery process is extended with two properties:

1) The processes in Fig. 15 (preprocessing, transformation and data mining) can be seen as models and methods in the UDSS point of view. The transformation phase has two approaches. (1) Data item values itself are not changed, e.g., filtering, data structure changes. (2) Data values can be changed after transformations, e.g., a data row is replaced with its average, or a lognormal transformation is performed on every data item. This characteristic is detailed in Section 2.4.
2) The *Data mining* phase is related mainly the analysis and data rows can be analyzed in various way, e.g., with mathematical methods (regression techniques), statistical methods (statistical tests, Analysis of Variance), number theory and graph algorithm (sorting, searching, graph traversal). Practical introduction will be presented in Section 2.4.

Studying DSSs literature, the definition and classification common grounds lead us to the self-designed UDSS concept. In the UDSS concept, different DSS definitions and classifications are defined at a higher abstraction level.

Definition. Universal Decision Support System (UDSS) concept provides data- and model-based decision alternatives. It supports to solve all structured and semi-structured problem, i.e., the nature of the data and the methods as well as the goal of the decision are general.

Universal Decision Support System includes:

- *Data management*. It takes care all of data including collecting data, loading into the common data layer (ETL functions), handling data-level permissions and it supports data queries.
- *Models and methods*. It supports data manipulation operations. However, the set of operations is not finite. If a method is not implemented in the system, then the system must be able to call that method from other system or ensure about the implementation of the method inside of the system.

UDSS does not make always a decision. Sometimes it just supports the decision making process and the users of the system make the decision itself. Therefore, communication (interfaces) between the user and the system is important. We defined users as scientists and experts.

With re-designing Knowledge Discovery process (see Fig. 15), the five steps of our general research process is defined as follows:

1) Identify, compose the problem;
2) Design the research;
3) Collect data;
4) Analyze these data with a method (or methods);
5) Present and interpret the results.

The first two steps include the definition of the problem, which we would like to solve. Then scientists can design the given research, define measurements, data and metadata. These data determine the database structure later. We assume that researchers know about their problems, they can describe them and they know which data should be measured and collected. In other words, they can carry out their whole experiments design in this first phases.

The collect data phase means getting data to data layer. It can be done by different ways, i.e., data can go through on more stations until they reach the data layer. This is similar to ETL, however, ETLs are responsible for uploading from outer source into Data Warehouse. In the UDSS, the ETL concept is extended and ETLs imply not just the last step of data uploading, but the whole data path from creating the data to the final loading into the data layer. The easiest solution is the direct upload into the data layer from the devices (sensors) but it cannot be performed in all cases. Therefore, it is possible that data go through several structures until become loaded into data layer.

The fourth contains a set of algorithms. In the view of UDSS concept, there is no difference between a preprocessing method, e.g., cleaning data, a transformation rule, e.g., lognormal conversion, and data mining techniques, e.g., clustering. Moreover, not just data mining techniques but other algorithms can be used in *Step 4*.

Results are presented in the presentation layer (*Step 5*). The main goal of UDSS is to get better decision based on facts (data). In general, new results are born if we have new data row or current data are analyzed with new methods. In this view, presentation layer has less importance than *Steps 1-4*, since the decision is supported by these steps. Presentation layer is responsible for the interaction between the user and the system. It is used to define analyzing process rules. The other functionality of this layer is result visualization.

If the results are not satisfying, the steps can be repeated. In this data-driven point of view, there are three possibilities:

1. We have “direct” data, related to the main research topic (main data row).
2. Not all data is strictly related to research topic. Data rows contain non-directly related information, e.g., time, place, or another dimension-like entities.
3. There is no data related to the research topic (theoretical research).

3.3. Used models and architecture patterns

In general, there are two main approach of DSS implementation: (1) create a specific DSS and (2) create a DSS entity with a DSS-generator. The latter will be introduced in DSS-generator section. The specific DSS has two common characteristics:

1. **Different characteristics**. Every scientific field has its own properties. For example, in medicine, time can be very important. Therefore, it is always a critical factor of medicine DSS implementation design. In forestry, time can also be important, but in a different point of view. DSSs' performance is not so important, because a relatively long decision processes has less or no effect on the real results. Because of the nature of this field, decision making at given point of time is made for the next decades or even century. A typical example is the growth of the trees: they grow relative slow so the results of the decision can be seen only in 10-50 years later.
2. **Technical implementation**. Different elements were used by the DSS implementations. Early implementations were mainframe-like. From the '90s, relational Database Management Systems were used. Another technical element is the World Wide Web, however standalone applications programmed in high-level programming language are also important. The Data Warehouse, ETLs, Multi Criteria, big data inspired environment and data mining techniques implementation are also related to the applied design patterns.

The purpose of the UDSS's concept is to unify the different characteristics, i.e., all kinds of data originating from different scientific field must be handled, data can be manipulated and analyzed to create one or more decision scenarios.

The results are presented according to the output structure of the analyzing phase. Two more elements have effect on the UDSS. One was the ForAndest database, which structure is summarized in Fig. 10.

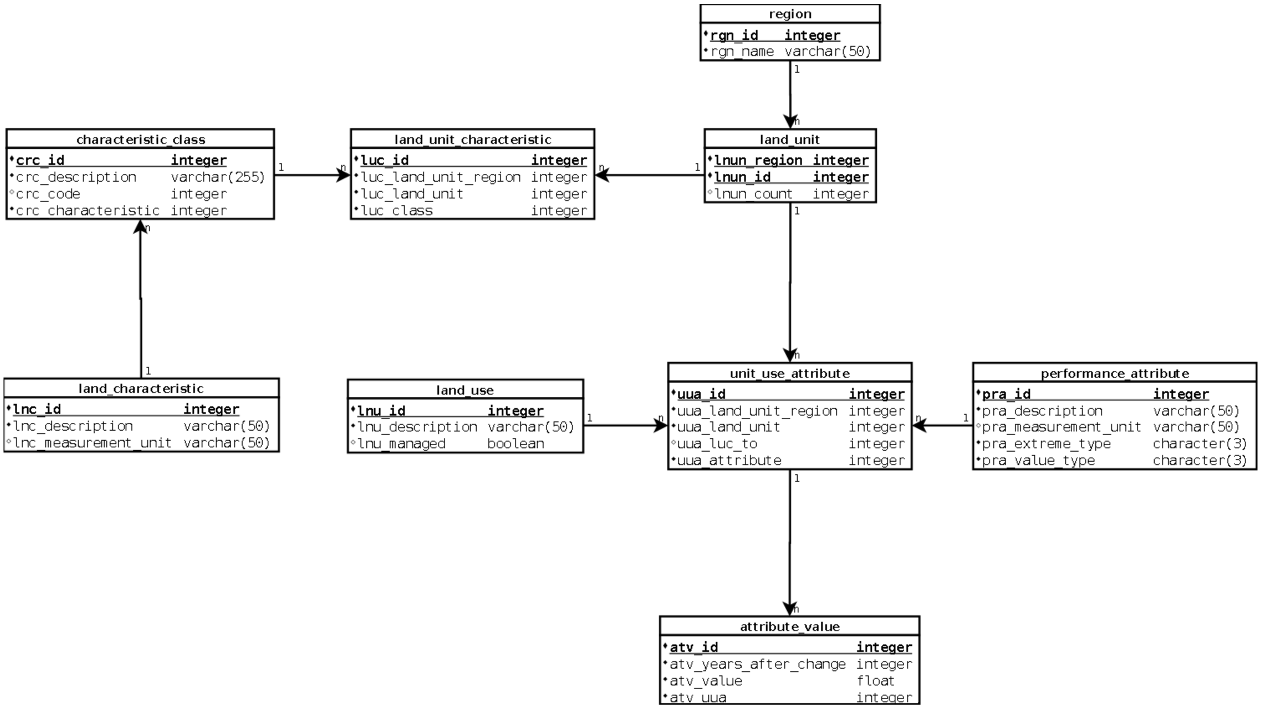


Figure 10: ForAndesT database structure

There are two table types: (1) entity tables and (2) connection tables. Each DSS parameter, such as *land characteristics*, *region*, *performance attribute*, *land use* and *attribute value* are stored in a table individually. The *characteristic class* and *land unit* tables are aggregate tables, e.g., every *land characteristic* is related to one *characteristic class*.

According to connection tables, such as *land unit characteristics* and *unit use attribute*, the proper DSS parameter values can be queried after the join operations. This is similar to our concept, however, our concept is extended to a higher abstraction level: the universal decision support system (UDSS) concept supports solving problems not only in forestry, but in other scientific fields as well.

We are also used extension design, which is used in MELA for example. There is a general core, and the different extensions are on this core in the architectural view, if scientists would like to use their own methods, they can implement it as an extension. In a higher abstraction level, it is possible to define entire decision support processes starting from data connection, through analyzing with a self-developed method and ending with presentation of the results. It is also possible to modify some parts of the system, while other parts of the system stay untouched. These extended viewpoints are also used in the case of UDSS. Extension-like design is very similar to Raheja and Mahajan's Model Base Management System component: we can use pre-defined methods, but we can implement new techniques. Moreover, some blocks from the core can be used to implement these new algorithms.

The architectural propositions can be summarized as:

- *Data Management layer*. Measured data are stored in a database, moreover, all kind of data management, queries and data security solutions are part of this layer.
- *Models and methods*. Data cleaning methods, aggregation techniques, analyzing algorithms, in one word, data manipulation possibilities are needed to create decision alternatives.
- *User Interface*. This module supports interaction between the system and the decision makers, e.g., input the algorithm parameters as well as present the results.

This summary is similar to the tree-layer architecture. However, more specific components are needed to provide the universality.

4. Universal Decision Support System

4.1. Architecture

The overview of our UDSS is based on the three-layer architecture as it is summarized in Fig. 11.

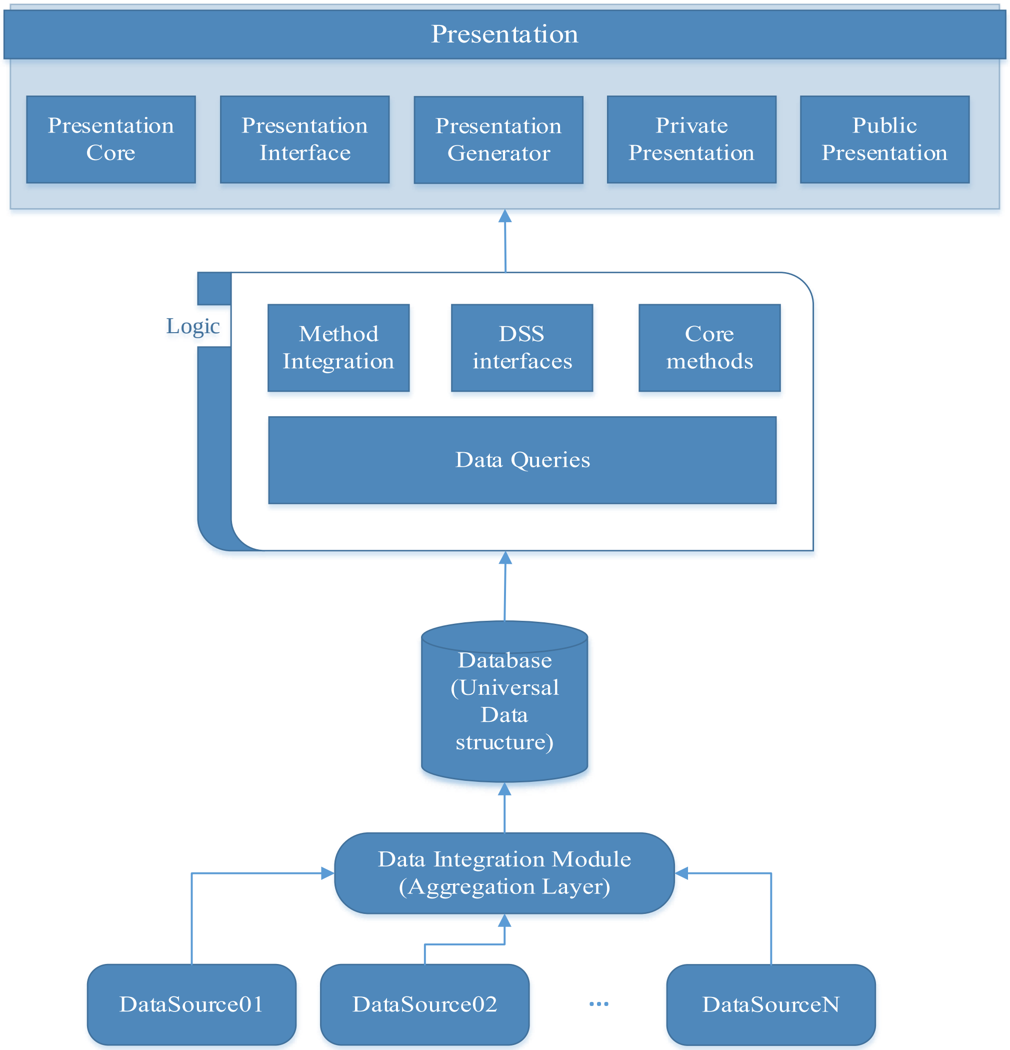


Figure 11: Universal Decision Support System concept architecture

Data storage management should be universal. The given data structure is changed to the structure of the database in a transformation phase. The database structure must be able to contain all of kinds of data items. In other words, transformation rule must exist between any unique data row and the universal database structures.

Definition. Universal Database structure (UDB) is a structure, which can receive and store any kinds of data, or at least there are trivial rules to transform original data formats to the universal database structure.

Data Integration Module supports researchers to create their own transformation rules between their new data source and the database structure.

Definition. Data Integration Module (DIM) is a rule-based interface to define transformation rules to change original data structure into the UDB structure.

If the database structure is fixed, the query module can be simpler, since we know all information about the UDB. In this case, a universal database structure could be defined as it is introduced in Section 4.4.

Definition. Data Queries (DQ) is a module executes queries to get the desired raw data from database.

Since many transformation and/or analyzing process can be performed on raw data, DQ is responsive for "raw" data queries primarily. But in the technical point of view, filtering, and a few function, such as average, summary and other operation on raw data can also be the task of DQ.

Methods and algorithms are used to manipulate data. The processing phase supports successive execution of methods.

Definition. Data Manipulation Module (DMM) is a set of algorithms, which can be performed to get scientific results, i.e., to support decision.

The Logic layer has three components.

Definition. Core Methods (CM) contains algorithms, which are already in the system.

Definition. Method Integration (MI) supports the method integration process.

Core Methods and Method Integration are enough to implement an UDSS. If an already implemented method exists in in another ad-hoc DSS, then it is possible to import that method into UDSS. Therefore, the Decision Support System Interfaces should be a part of an UDSS. If we have an old DSS, and we would like to build a new DSS following our UDSS concept, the old system's methods can be utilized partly or completely. Open source software implementation can also be used with standardized interfaces.

Definition. Decision Support System Interfaces give support to invoke algorithm implemented in another system.

The presentation core (PC) is a set of views related to the given algorithm, i.e., each data manipulation method can have a default view.

Definition. Presentation Core (PC) is a set of views.

Each algorithm has its own output entities, e.g., a linear regression result can be presented in a coordinate system. But if we would like to get the strength of correlation, then a view with r^2 can be also defined. Presentation generator supports scientists to define how they would like to see the results.

Definition. Presentation generator (PG) supports the user to define presentation form based on entities of the performed algorithm.

If researchers would like to use another presentation subsystem, there must be a way to define interfaces.

Definition. Presentation interfaces support the presentation subsystem integration process.

User Interface Module (UIM) is needed for the communication between the user and the system. MI can be programmed in Algorithm Definition Language or in a standard programming language, i.e., Java, C#, etc. If a scientist knows (or learned) this language, then he or she is able to define all of the methods, and a source generator is able to generate an executable file based these defined rules. This is true for PG as well.

Definition. User Interface Module (UIM) helps the communication between the user and the system.

4.2. Analyzing session

We define users as experts and scientists. We assume that the given problem definition and the given experiment are known by scientists, therefore they start to use the system with *Step 3*. In the classical view, metadata are data about data but we use it in an extended interpretation in UDSS. If we would like to perform an experiment and analyze our data, there is at least one data row. However, there are other data which values describe the environment of the experiment. With this approach, all analyzing structure can be defined, which enables us to perform an analyze process related to:

- The main data itself;
- Its metadata itself;
- The main data row and its metadata;
- Metadata rows only;
- Two or more different main data rows;
- Metadata, each is related to different main data.

Unlike to ad hoc DSSs, we can compare not only more main data rows but one main data row's metadata with another main data row's metadata as well. This increases the number of the possible comparisons and can lead to complex analyses. For example, we would like to make statements about the wood color, then the measured wood color value is the main data row and the other data rows are metadata, i.e., "data about the main data row". If we would like to analyze ionosphere with ionograms, time can be important, i.e., daily ionograms have be differentiated from ionograms at night. In this case, time is a metadata, which can affect the analyses of ionograms. The previous examples will be further discussed in Section 2.5. The classical metadata has less influences on the results, but metadata can be very important in the interdisciplinary property of UDSS. In UDSS concept, there are pre-defined processes to develop scientific results. The steps are summarized in Fig. 12.

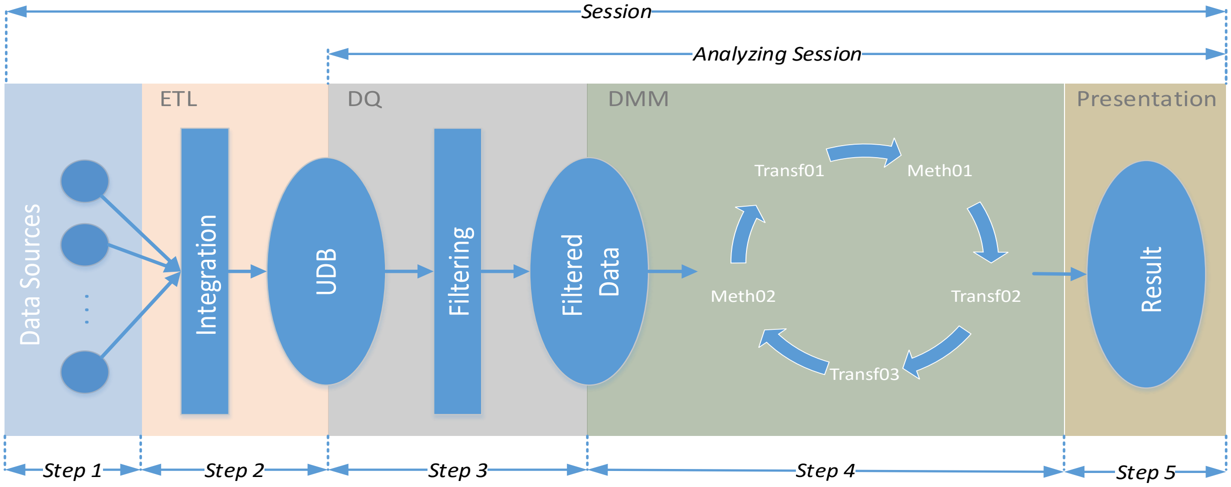


Figure 12: General Session process

Measured data and metadata can be stored in different formats. These data items can be integrated and uploaded into the system's database (*Step 1*). If parts of one data row are scattered, then users can define the concatenate rule, i.e., ETL functions (*Step 2*). This step must be done only once. Several solutions give data integration tool. If any dataset changes or another datasets arrives, many data connection parameters have to be changed, e.g., if a new data row appears in the original source, this modification must be signed in the connection. This process can be difficult having a few data source already and it is hard if there is large number of sources. It is better to store datasets inside the system in one universal data structure. Another important characteristic is that new data rows (main or metadata) can store in UDSS database without system modification and the same is true when attaching a new metadata to an existing main data row. In other DSS, more entities such as database structure, data queries, decision processes must be changed to achieve this.

The pre-defined database structure is the base of our system. The data layer operations are:

- **Create**. User creating a new main data row.
- **Attach**. User attaching a new metadata row to the existing main data row.

The *Attach* operation also means that we would like to add a new data row to the main data row.

Filtering methods must be applied in the query process (*Step 3*) to get specific data. Since we have one big robust database with pre-defined structure, the data queries are simpler.

In the concept of UDSS, transformation methods and analyses have the same meaning (*Step 4*). If the system core includes the transformation or methods of analysis, then users can use them after each other until the desired result is appeared. If the users would like to use algorithms, which is not part of the system core, then the new method must be imported into system. This can be done in several ways.

- **Using core methods**. Analysts would like to use methods which are in the core.
- **Import new method**. The system provides user interface to write a new method in a common language. The new data manipulation techniques must be written in a common language, which can be an Algorithm Description Language (ADL) or a programming language.
- **Import new method in file format**. The new method is imported in file format. In the file, algorithms are written in a common language.
- **Invoke new method**. The new method is invoked from another system through interfaces. Interfaces must also be defined in a common language.

In the cases of the import processes and the "invoke new method" operation, the new method is imported into the system core. With this, its next use will be simpler.

The result is presented to the user (*Step 5*). Presentation operations are:

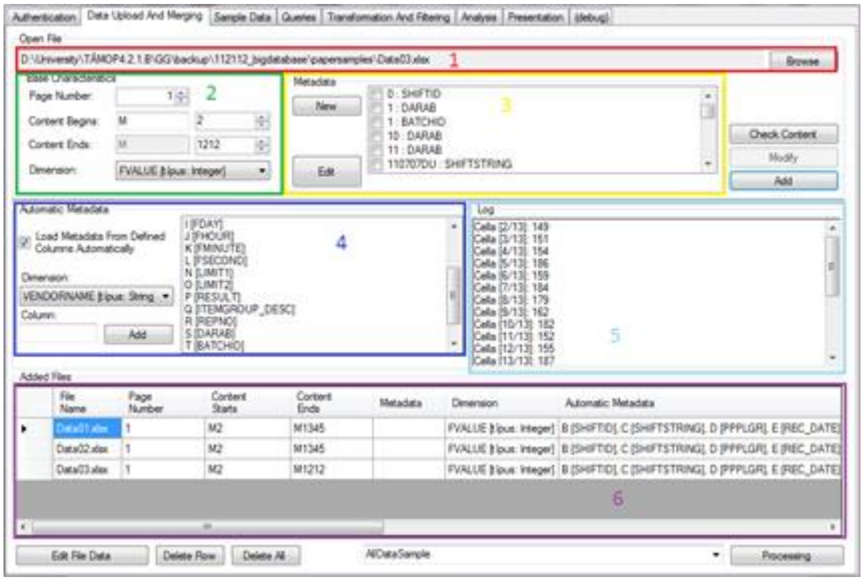
- **Using default view**. Researcher would like to use the default view of the last performed algorithm. This view is in the Presentation Core.
- **Import new view**. Researcher would like to use his or her own view written in a common language.
- **Import new view in file format**. Import the new desired view in a file format. New views are also written in a common language.
- **Send result data to another system**. The result data are sent to another system through UDSS interfaces.

Each method has its own basic view. It is also possible to define a new view for the given algorithm. In the case of *send result data to another system* operation, presentation data are imported into another system and the result visualization is executed in this another system.

Not only one good solution exist since different scenarios can be created based on different decision parameters and s. This is named as multi criteria. Therefore, the process can be repeated from *Step 3*, since the different approaches must be based on the same data and the difference is mainly in the applied methods, which are performed after each other. Since data uploading must be performed once, the other steps are more frequent. Therefore, the process from *Step 3* to *Step 5* is called an Analyzing *Session* (AS). The process from *Step 1* to *Step 5* is called a *Session*.

4.3. Data Integration Module

Data items are stored in their original state and either documentation is available or data owners know every information of their data structure. There must be a clear transformation between the original data structure and our database structure. Defining these rules is the functionality of the Data Integration Module (DIM).



Open File: D:\University\TÁMOP4.2.1.B\GG\backup\112112_bigdatabase\papersamples\Data03.xlsx

Page Number: 1

Content Begins: M 2

Content Ends: M 1212

Dimension: FVALUE [tipus: Integer]

- 0 : SHIFTID
- 1 : DARAB
- 1 : BATCHID
- 10 : DARAB
- 11 : DARAB
- 110707DU : SHIFTSTRING

☑ Load Metadata From Defined Columns Automatically

Dimension: VENDORNAME [tipus: String]

Column:

I [FDAY]
J [FHOUR]
K [FMINUTE]
L [FSECOND]
N [LIMIT1]
O [LIMIT2]
P [RESULT]
Q [ITEMGROUP_DESC]
R [REPNO]
S [DARAB]
T [BATCHID]

```
Cella [2/13]: 149
Cella [3/13]: 151
Cella [4/13]: 154
Cella [5/13]: 186
Cella [6/13]: 159
Cella [7/13]: 184
Cella [8/13]: 179
Cella [9/13]: 162
Cella [10/13]: 182
Cella [11/13]: 152
Cella [12/13]: 155
Cella [13/13]: 187
```

	File Name	Page Number	Content Starts	Content Ends	Metadata	Dimension	Automatic Metadata
▶	Data01.xlsx	1	M2	M1345		FVALUE [tipus: Integer]	B [SHIFTID], C [SHIFTSTRING], D [PPPLGR], E [REC_DATE]
	Data02.xlsx	1	M2	M1345		FVALUE [tipus: Integer]	B [SHIFTID], C [SHIFTSTRING], D [PPPLGR], E [REC_DATE]
	Data03.xlsx	1	M2	M1212		FVALUE [tipus: Integer]	B [SHIFTID], C [SHIFTSTRING], D [PPPLGR], E [REC_DATE]

AllDataSample

Figure 13: Data integration and Uploading Module

The following example is used to illustrate this functionality. (Here we give you just a brief overview about the example, it will be detailed in Section 4.7.3.) A wood production company would like to analyze their vendors' performance. The data about the vendors' performance are stored in scattered files, and we would like to upload them into our UDB with our universal DIM. In Fig. 13, the uploading of vendor data and definition of uploading rules is illustrated.

We marked six sections in Fig. 13, Users can browse different files, e.g., Excel files (*Section 1.*). In this example, there are three files with scattered data rows (*Section 6*). Experts can define the main data row items and data and metadata dimensions (*Section 2)*. Dimensions are very important, because data are for interpretation of main data row and metadata items. In other words, if we have a data item value *5*, we need to recover the information, what does this *5* mean. If that item dimension is centigrade then *5* means temperature was *5*, but if we would like to analyze earthquake magnitudes, then the dimension related to *5* must be defined as magnitude scale. Both data and metadata values have dimensions.

Metadata can be attached to each record individually (*Section 3*), but if each main data has the same type of Metadata (not the same value, but the same type of dimension) it also can be assigned automatically (*Section 4*). Main data row items from the original source can be checked (*Section 5)*. The *fact* has similar meaning as in the context of data warehouse. During the uploading process, a *fact name* can be given which identifies this dataset in the UDB.

To analyze this example deeper, the following are illustrated in Fig. 13 following rules:

- We opened three Excel files (data was scattered in three files), the last opened file name can be seen in *Section 1*.
- We defined the main data row in column *M* from *x* to *y*, where *x* is the start of the main data and *y* is the end of the main data in each file. The *x* and *y* can be different in the various files such as column identifier as well. For example, the *x* was *2* and the y was *1212* in the third file.
- We determined the dimensions both main data and metadata items. Main data row dimension was "*fvalue*".
- We chose the automatic Metadata uploading and we connected each kind of metadata to the proper dimension. For example, there are values in the column *E* about "*rec_date*" dimension and in the column *Q* "*itemgroup_desc*" similarly. Each main data value has its own "*rec_date*" and "*itemgroup_desc*" value.
- We set the name of the whole dataset "*AllDataSample*".

We assumed that experts have knowledge about their data structure. Users know what "*fvalue*", "*rec_date*" and "*itemgroup_desc*" and other dimensions' mean. But the type of the dimensions should be defined from a programmer's point of view as well (*Section 3*, *edit* button). In our example "*fvalue*" is integer, "*rec_date*" is date and "*itemgroup_desc*" is string.

At the end, DIM module unifies the scattered data rows (both main and metadata rows) based on rules defined by the user and it shapes the data structure, which is ready to be uploaded in UDB. If the research is in the planning phase, when user can define the way of data collecting, then it is better to upload measured data into the UDB directly. However, most of the cases, data collecting phase is defined already and data must be uploaded into the UDB with DIM ETL. In our example, we used Excel, but other DIM function can also be written in UDSS Common Language.

4.4. Database structure

The foundation of the UDSS concept is the database structure, which is summarized in Fig 14. The tables named *User*, *UserConnectRole*, *Role*, and *RoleConnectFact* are important in user authority. Our UDSS solution is role-based, each user has at least one role. Since all data are stored in the UDSS database, the authority system provides that incompetent user cannot access data related to another user. In Fig. 14, roles are connected to facts, which represent both main data and metadata.

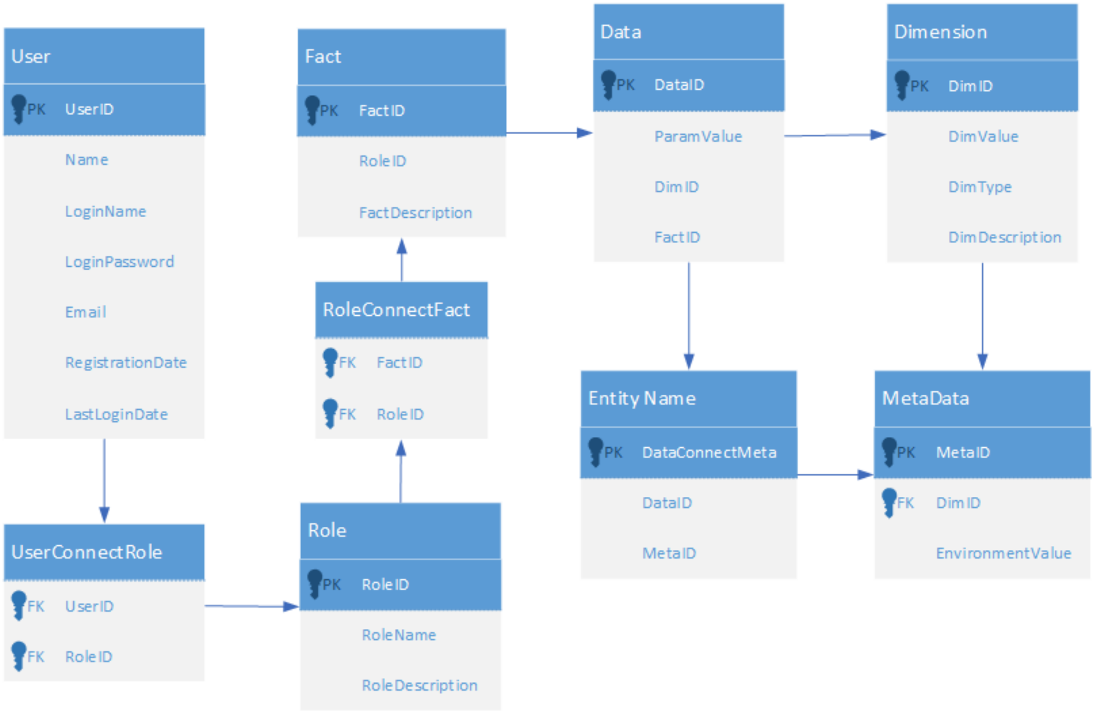


Figure 14: Structure of our UDB

Facts are the summary name of the data set. The main data row is stored in the *Data* table, metadata in *Metadata*, and dimensions is *Dimension* table respectively. Since each main data value is related to one fact, therefore *FactID* column in the *Data* table creates the connection between facts and main data items as foreign key. The *DataConnectMeta* provides the connection between data and its metadata items. Each individual main data values can have more metadata values. Primary keys identify one row in the tables, therefore we can define their own different metadata values for each main data items individually. Most database management systems can handle only one data type for one column, such as integer, double, string, date, etc. It is possible to store every item in string format, and the data type can be stored in *Dimension* table, but every value must be converted to the proper data type at the end of query process, which can take very long time if the number of desired data is large. Therefore, the concept must be implemented according to the given system types. Database concept extension is illustrated in Fig. 15. The main differentiates are that *Data* and *MetaData* tables are implemented according to types. Four main types are used in our system, *integer*, *double*, *string* and *date*. If more types are needed, it is possible to define as many tables as the types are during the implementation phase. Data values are stored according to their types; therefore, it is no need to convert them during or after the query process. The database concept has two characteristics: (1) compression property and (2) the changeability of data rows.

The following example illustrates the first characteristics. If we have one fact which has *z* data items and each data items have its own metadata values n_1, n_2… n_z, then the size of *Fact* table is 1 and the size of *Data* is *z*. However, the size of *Metadata* table is not necessary $n_1 + n_2 + … + n_z$, because compression of the table is possible. If two main data item have the same metadata value and dimension, there is no need to create one row in the metadata table, only the connection has to be indicated in the *DataConnectMeta* table. By compressing the *Metadata* table, the size of connection tables can grow significantly. Because of the connection table's size, the query process can take large amount of time. In our big data inspired environment, we did not experience large query times. The data management was satisfying with a standard computer setup and a database management system.

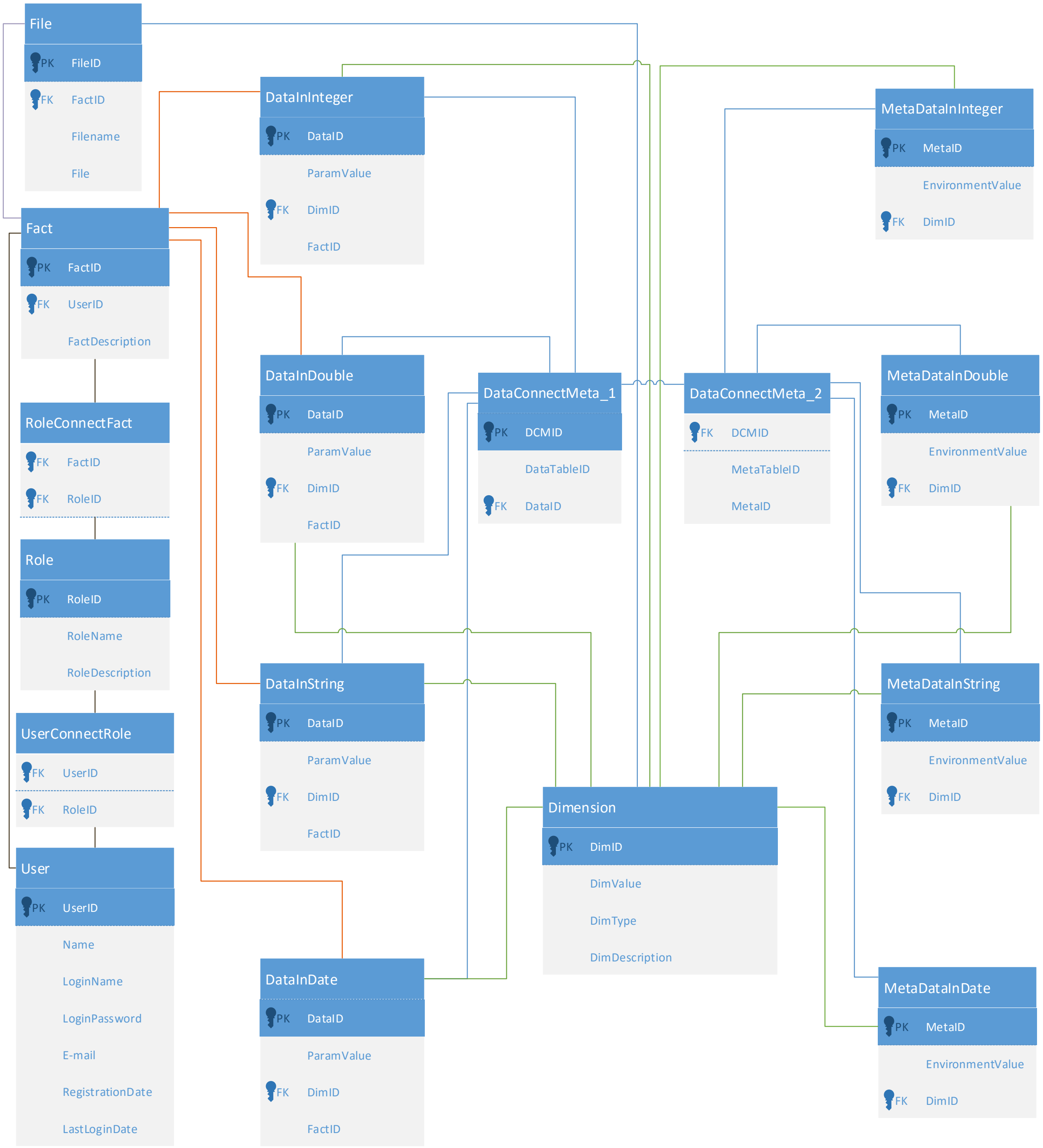


Figure 15: Final structure of our UDB entity

The second characteristic is the changeability of data row. It is not always clear which row is the main data row. But as we will see, this is not important in our concept. The main data row items' IDs provide the relationship between all data items though connection table and therefore our database concept is also universal. We can simply query all data items, but more important, the original relationships between data are preserved with this implementation. Therefore, every data row can be the main data row. Each data row can be main data row and even the term “main data row” will be less important in the future since this main data row will appear as a black box to the users in the query module. It is not a critical question which is the main data row and which are the metadata. But if experts would like to create two or more main data rows of the same measurement, they can do it as well. The results of queries will be the same when there would be just one. Experts can describe their problems in different ways in this system and therefore this system works as a framework.

As it was illustrated in our example, scattered vendor performance data are uploaded into UDB. The main data row was *fvalue* as integer. Therefore, these data are stored in *DataInteger* table. The *rec_date* was date type, while *itemgroup_desc* was string type. In this case, data are stored in *MetaDataDate* and *MetaDataString* tables accordingly. The relation of the main data row and these metadata rows are created in *DataConnectMeta1* and *DataConnecMeta2* connection tables. The first is responsive for the main data types, the second is for metadata types. This means that the connection is created in the case of all individual main data value and related metadata value according to types. The compression of Dimension is possible in this example.

4.5. Query process

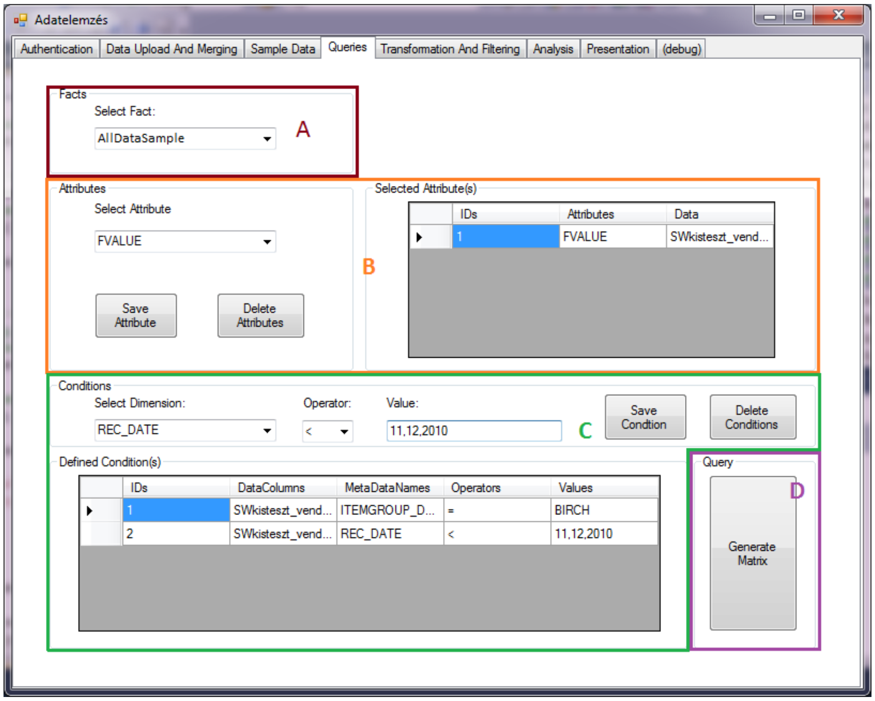
Because of the fix database structure, the query module can be simpler than the data integration process and database structure. The only critical parameter is that the query system must support different outputs, because different methods of analysis have different input structure. The query structure and the first method’s input structure must be matched. If we would like to perform several methods after each other, then the output-input structures must be matched as well. In the case of structure mismatch, shaping of the proper structure must be supported through the user interface. After the query process, data are organized in matrix structure named as analysis matrix. In the case of structure mismatch, the structure of the matrix is manipulated by the user until the proper input structure of the next method is created. SQL aggregation functions are in this part of the system.

To query a data row, we need to define the following parameters:

- **Fact**. User defines which dataset will be queried. Dataset is identified by its name [*Section A in Fig. 16*].
- **Data row**. User defines the desired data row(s) of the selected fact. There is no difference between main data row and metadata. Data rows appear filtered according to previous selected Fact [*Section B in Fig. 16*].
- **Filters**. Filtering condition(s) for data items can be defined during query process [*Section C in Fig. 16*].

If an analyst would like to have more data row, then the previous process starts again. Each defined data row is represented as a column in the result matrix. Therefore, there is possible to analyze main data row-metadata, metadata-metadata and even a dataset data (i.e., fact) row with another data set data row.

In the background, the system generates SQL queries [*Section D in Fig. 16*]. The query engine translates the user’s query definition to SQL language. The conditions are concatenate after SQL *where* term.



Facts

Select Fact: AllDataSample

Attributes

Select Attribute: FVALUE

Selected Attribute(s)

	IDs	Attributes	Data
▸	1	FVALUE	SWkisteszt_vend...

Conditions

Select Dimension: REC_DATE

Operator: <

Value: 11,12,2010

Defined Condition(s)

	IDs	DataColumns	MetaDataNames	Operators	Values
▸	1	SWkisteszt_vend...	ITEMGROUP_D...	=	BIRCH
	2	SWkisteszt_vend...	REC_DATE	<	11,12,2010

Query

Figure 16: Self-developed UDSS DQ example

In the view of DQ, the types of query are:

- **Main data row itself**. All data items are in the analysis matrix according to fact main data.
- **Main data row with filter defined for its own**. Same as the previous case, but filters are defined for the main data row.
- **Main data row with filter defined for metadata**. The filters are defined for metadata values, but we still would like to get main data items which satisfy metadata conditions.
- **Metadata itself**. A metadata row of a given data set (fact) is in the analysis matrix. All values related to that metadata will be queried.
- **Metadata row with filter defined for its own**. Similar to the previous one, but some conditions are defined for the given metadata values.
- **Metadata row with filter defined for main data row**. Filters are defined for the given metadata.
- **Metadata row with filter defined for another metadata**. Condition is defined for metadata *A*, but the proper metadata *B* values will be queried.

The query process can be performed with two phases: (1) query preparation and (2) creation of the matrix. In the first phase, *DataIDs* in *Data* tables are evaluated. If filtering is applied, then only the *DataIDs* which fulfill the condition are queried. Since there can be linear combination of the options above, the DQ performs the condition queries row by row. At the end of each given query, the section set operator is applied for that *DataIDs*, which satisfy the conditions. For example, conditions are defined for two individual metadata rows and we would like to query main data values. But further conditions are also defined for the main data row directly. In such a case, all queries are tracked back to *DataIDs*, and in the end of the first phase we have that *DataIDs*, which satisfies all conditions. Each query works with *DataIDs*, therefore *DataID* is considered as *superkeys*.

In the second phase, the analysis matrix is created by the DQ, i.e., the proper data or metadata values defined by the user are queried according to the first phase *DataIDs* set. The user cannot see this query engine (it is a black box) and main data row and metadata term will not be relevant. This matrix will be manipulated by models and methods performed after each other.

4.6. Logic and Presentation Layer

Preserving the universality is a challenging task in the Logic Layer. The system must be prepared to use not just the pre-defined algorithms, but any kind of transformation methods, mathematical models, analyzing processes and data mining techniques. There are three requirements, which must be fulfilled in a system working with the UDSS concept.

First, the UDSS must be suitable to support such algorithms which are not developed yet but could be used inside the system in the future. In second and more frequent case, the scientist would like to change some steps during the standard analyzing process or modify the given algorithm according to scientist desire. The third challenge is to apply models and methods after each other. Each method has its own input and output structure, and the output structure can be the input of the next applied technique. From the UDSS point of view, combinations of analysis can be executed to assuring the universality and leads to more complex analyses and to more precise decisions. A standard process in Logic Layer is summarized in Fig. 17.

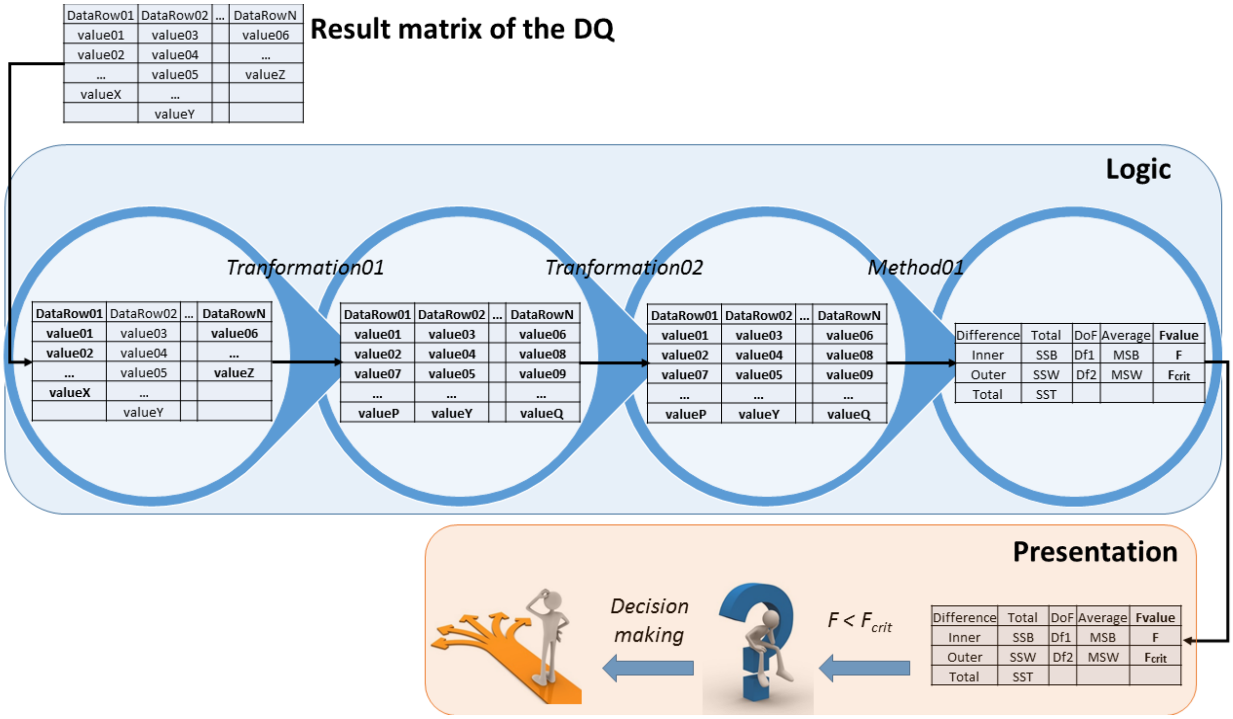

Result matrix of the DQ
DataRow01 | DataRow02 | ... | DataRowN
value01 | value03 | | value06
value02 | value04 | | ...
... | value05 | | valueZ
valueX | ... | |
| valueY | |
Logic
Tranformation01
Tranformation02
Method01
DataRow01 | DataRow02 | ... | DataRowN
value01 | value03 | | value06
value02 | value04 | | ...
... | value05 | | valueZ
valueX | ... | |
| valueY | |
DataRow01 | DataRow02 | ... | DataRowN
value01 | value03 | | value06
value02 | value04 | | value08
value07 | value05 | | value09
... | ... | | ...
valueP | valueY | | valueQ
DataRow01 | DataRow02 | ... | DataRowN
value01 | value03 | | value06
value02 | value04 | | value08
value07 | value05 | | value09
... | ... | | ...
valueP | valueY | | valueQ
Difference | Total | DoF | Average | Fvalue
Inner | SSB | Df1 | MSB | F
Outer | SSW | Df2 | MSW | Fcrit
Total | SST | | |
Presentation
Decision making
F < F_crit
Difference | Total | DoF | Average | Fvalue
Inner | SSB | Df1 | MSB | F
Outer | SSW | Df2 | MSW | Fcrit
Total | SST | | |

Figure 17: Standard process of Data Manipulation Module and Presentation Layer

In the DMM phase, there is a “starter” matrix as the result of the DQ process. Each row represents a data row, which comes from different data sets, i.e., from different facts. DMM methods can be applied on this starter matrix. The analysts can manipulate each row individually. Therefore, the matrix can be shaped to the input structure of the next method.

In Fig. 17, an example performing a standard DMM process is illustrated. We would like to perform Analysis of Variance (ANOVA) for example. After data rows are queried, it is possible that data rows' length is not matched, e.g., measurement was made in different granularity of time or place to analyze data, the data rows must be transform into a common dimension. In Fig. 17, all data rows are transformed into the same length. To achieve this, we applied some kind of aggregation or drill down-like transformation on *DataRow01* and *DataRowN*. In ANOVA, there are assumptions which must be fulfilled before we apply it. One of these conditions is normality assumption, therefore *Transformation02* is performed because of that. If we assume that all data rows passed this test, then ANOVA can be the next method. Based on the result of ANOVA, the final decision can be made.

It is possible to invoke methods from other system through UDSS interfaces by the following way. The DQ generates the analysis matrix. Methods existing in other system could be used to manipulate this matrix values. In our UDSS, we implemented interfaces with CL for *WEKA* and the *R* software. The methods of analysis are performed with *WEKA* or *R*, but the other analysis steps are done in UDSS. Both systems must be installed on the same computer for the analyses.

The presentation layer is summarized in the *Presentation* section of Fig. 17. Let us further study our previous ANOVA example. As a result, we got an *F value.* If this value is greater than the *F critical* value, then we can make the decision since at least one of the means differs from the other means statistically.

The implementation of Presentation Generator is similar to the Logic Method Integration. Different views can be defined in CL and they can be assigned to a method. More than one view can be assigned to a method, however, each method must have a default view. These views are stored in Presentation Core.

The presentation phase can be performed in other system. In this case, the result matrix is sent to the presentation software. Simple example is the copy-paste operation. For example, if the result of DMM is a table, then this table can be copied and pasted into Excel. In Excel, various presentation tools are available for the user.

4.7. Validations and results

4.7.1. Use Case I: UDSS operation with current implementation

Because ForAndesT had influence on our UDB concept, therefore we examine how it is possible to implement decision support processes of the ForAndesT in our UDSS concept.

Afforest data sources were Excel files, while ForAndest has a relation database as well as Osmose does. There are some differences between the systems but their functioning principle is very similar. Therefore, all three systems' decision processes can be implemented in our UDSS system.

In forestry, there are couple of questions, which we would like to answer with a given DSS. These questions can be classified into the following types:

- *"What" question*. Under the current circumstances (current land use type), what the land units' performance will be. Land use type means tree species.
- *"What if" question*. What the performance of a land unit would be, if initial land use type is converted into a new one, e.g., afforestation technique is changed replacing one tree species to another one.

- *"Where" question*. Which land units are the best option under the user constrains.
- *"How" question*. Which silvicultural technique is the best for the given land unit under the user constraints.
- *"How long" question*. This question is related to the time interval. In ForAndesT, it shows which rotation length is the best.
- *"When" question.* When should the initial land use type replaced with a new one to get the best performance. It is similar to "what if" type of question, but the best performance is sought.

Not all systems can answer all type of questions. The more the system evolved the more questions can be answered. For example, ForAndest can answer "*Where*", "*How*" and "*How long*" questions, but OSMOSE can answer several types of question.

The process of analysis can be summarized as follows:

- **Define the question**.
  The systems can answer several types of questions in forestry.
- **Define environment performances (EPs)**.
  These are the input characteristics of the given land. ForAndest has five Eps, such as run-off production, sediment production, carbon sequestration in soil, carbon sequestration in biomass and income.
- **Define performance attributes.**
  These are the outputs. Based on the inputs of land use types, we would like to get result for these attributes. In ForAndest, these target attributes are carbon sequestration in soil, nitrate leaching and groundwater recharge.

We followed the steps of General Session process discussed earlier in Section 4.2 and illustrated in Fig. 12. First, we dealt with data management. ForAndesT has its own pre-defined database structure, therefore transformation rules need to be defined between ForAndesT and our UDB entity. As we mentioned earlier, any data row can be defined as main data row. We chose *region* as a main data row and the rest of the rows are metadata. But other definitions are also possible, e.g., other data row can be chosen as main data row.

The rules of data structure transformation are summarized in Table 3.

Table 3: Correspondences between ForAndesT and UDB entity

Tables	**Type in ForAndest**	**Type in UDB entity**
land characteristics	Data	Metadata
region	Data	Main data
performance attribute	Data	Metadata
land use	Data	Metadata
attribute value	Data	Metadata
characteristic class	Aggregate	Metadata
land unit	Aggregate	Metadata
land unit characteristics	Connection	None
unit use attribute	Connection	None

The *land unit characteristics and unit use attribute* tables are connection tables, and there is no need to handle them in UDB. The join operation is performed differently in UDB than in ForAndesT, and key values are generated by the UDSS automatically.

The next step is the DMM phase. The goal of DSSs in forestry is to answer a given type of question. Each type of question is an algorithm. In this point of view, we need to write these algorithms in CL and upload into the core. Belgian DSS were written in C#, in the same language as our UDSS entity. Therefore, it is not necessary to write the related code again.

We show an example for the "Where" question. The method behind answering this question is the Iterative Ideal Point Threshold (IIPT) developed by Annelies et al [141]. First, we need to choose performance attributes for which different weights can be assigned. Next, we have to define which attributes will be minimalized or maximized. Since we have more attributes with weights and optimization questions, therefore IIPT and these DSS systems have *multi-criteria* concept. Having performed IIPT, we get a sub-optimal answer. This answer shows which land units are suitable for the user defined performance attributes. It is rare that a land unit satisfying the preferences is found at the first run. As the name of IIPT algorithm indicates, iterative search is performed based on Eq. 13.

$$goal\_value = optimal\_value \pm iteration\_nr * \left(\frac{\max\_weight}{weight\_ES}\right) * \left(\frac{range}{\#iteration}\right), \qquad \text{Eq. (13)}$$

where

optimal_value is minimum or maximum performance value according to the user definition;

iteration_nr is the number of iteration;

max_weight is the maximum weight among all selected attributes' weights;

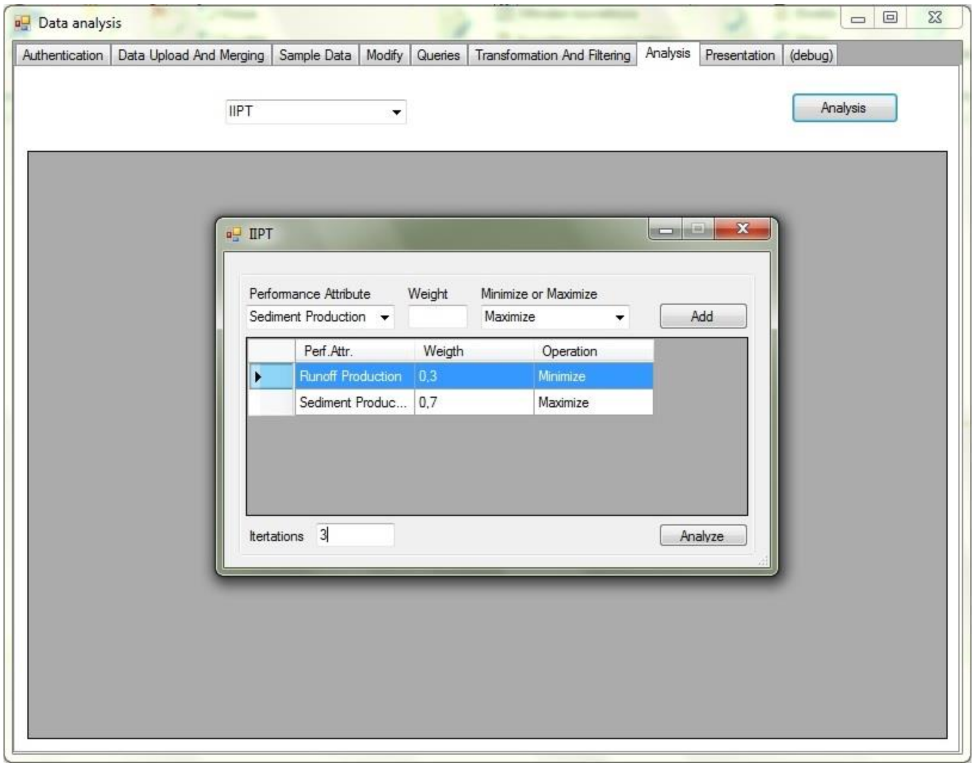
weight_ES is the weight of the given attribute;

range is the difference between the maximum and minimum value of the given attribute numeric co-domain;

#iteration is the actual iteration number.

The number of iterations is also defined by the user and this influences the number of sub-optimal solutions.

To perform an analyzing session, proper data must be queried. We need (1) land units, because we seek those land units, which satisfy the conditions, and (2) performance attributes. Because we seek best land units for the conditions, there is no need for any kind of transformation or for other pre-DMM techniques. We chose two performance attributes, *Runoff* and *Sediment*. The weights are set to *0.3* and *0.7*, and we minimize *Runoff* and maximize *Sediment*. The iteration number is set to *3*. The IIPT parameterization overview can be seen in Fig. 18.



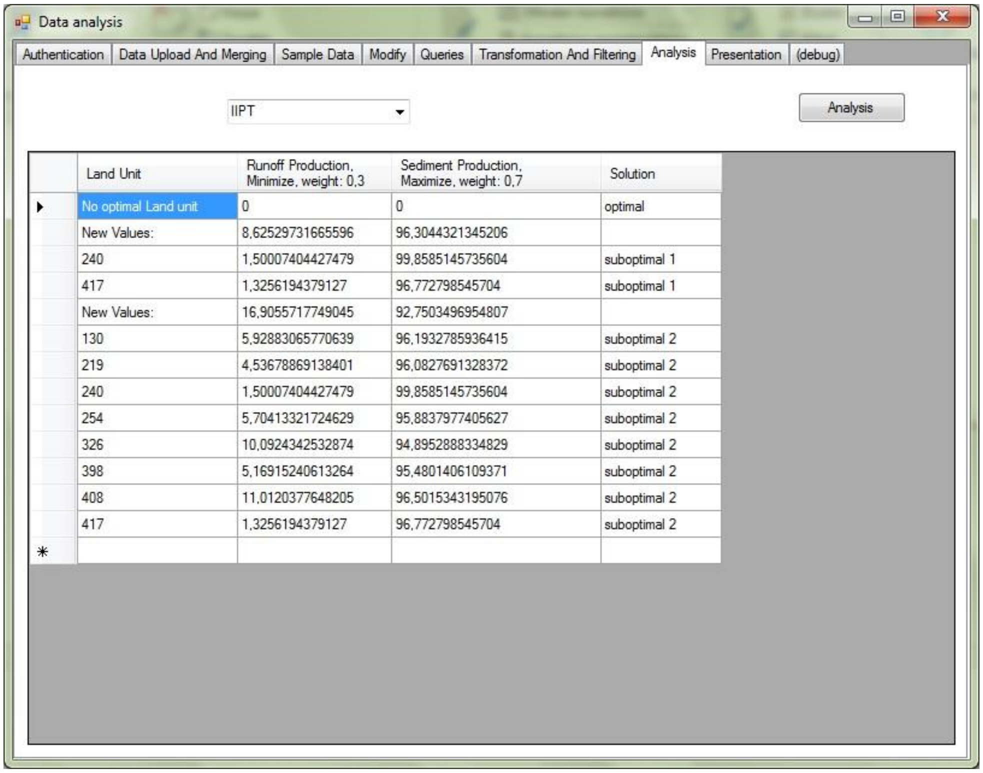
Performance Attribute: Sediment Production
Weight:
Minimize or Maximize: Maximize

Perf.Attr.	Weigth	Operation
Runoff Production	0,3	Minimize
Sediment Produc...	0,7	Maximize

Iterations 3

Figure 18: IIPT parameterization in our UDSS entity

After performing the IIPT, the result is illustrated in Fig. 19.



Land Unit	Runoff Production, Minimize, weight: 0,3	Sediment Production, Maximize, weight: 0,7	Solution
No optimal Land unit	0	0	optimal
New Values:	8,62529731665596	96,3044321345206	
240	1,50007404427479	99,8585145735604	suboptimal 1
417	1,3256194379127	96,772798545704	suboptimal 1
New Values:	16,9055717749045	92,7503496954807	
130	5,92883065770639	96,1932785936415	suboptimal 2
219	4,53678869138401	96,0827691328372	suboptimal 2
240	1,50007404427479	99,8585145735604	suboptimal 2
254	5,70413321724629	95,8837977405627	suboptimal 2
326	10,0924342532874	94,8952888334829	suboptimal 2
398	5,16915240613264	95,4801406109371	suboptimal 2
408	11,0120377648205	96,5015343195076	suboptimal 2
417	1,3256194379127	96,772798545704	suboptimal 2

Figure 19: IIPT result

There is no optimal solution in the first iteration, i.e., no land unit satisfies the conditions. The new *goal_values* are calculated. In the second iteration, we have two hits, i.e., two land units with *ID 240* and *417*. In the third iteration, the *goal_values* are more permissive (one is increasing and the other is decreasing), therefore more land units fit.

Although the Afforest, ForAndest and Osmose DSSs are similar, there are several differences between these systems, such as question types, regions, EPs or Performance Attributes. New system developments were carried out based on the given experiences with the older systems and new types of question were developed. For example, data of other regions is collected and Osmose can answer two more types of question than ForAndest (“*How long*” and “*When*”).

The UDSS concept solves this problem since there is no need to implement a new system. Since UDSS concept offers the following solutions:

- All data with proper uploading rules can be uploaded into the UDB.
- If new data are measured, then it can be attached to the given main data row.
- If we would like to use a new method, i.e., answer a new type of question, just the method itself must be created or called.
- Data and methods can be combined easily.
- Multi-criteria analysis can be performed on forestry data.

4.7.2. Use Case II: Ionogram processing

The UDSS helps not only existing forest decision support systems, but can be used in other fields like earth science, vendor selection, and production optimalization. In this section, UDSS was used to determine relevant area of the ionograms.

Ionosphere’ layers are created by ionized gas with sunbeams. During the process, neutrons turn into positive or negative atom depending on they lose or receive an extra electron. This process is not stable, ionized atoms react with each other losing and getting electrons. Based on the ion density, the ionosphere can be divided into more layers as it is illustrated in Fig. 20.

The lowest layer is *D-layer*, and going higher there are *E-*, *F1-* and *F2* layers. There is difference in the appearance of the layer in daytime and at night, because the sunbeams and radiations related it like UV have strong effect on the ionization of atoms. This is the reason, why there is no *D-layer* at night. In the view of UDSS, *F1* and *F2* layers are important, because these two layers are present in day and also at night. Therefore, the analysis of ionograms focus on these two layers mainly.

The ionosphere measurement is made with the ionosonde. The data of ionosonde can be visualized as an ionogram, i.e., the ionosonde’s output is the ionogram. The ionogram is a binary picture, which contains a lot of noise and the relevant areas. The relevant areas have two main parts: (1) the *ordinary* component and (2) the *extraordinary* component. An example of an ionogram is illustrated in Fig. 21, green is the *ordinary* and red is the *extraordinary* component.

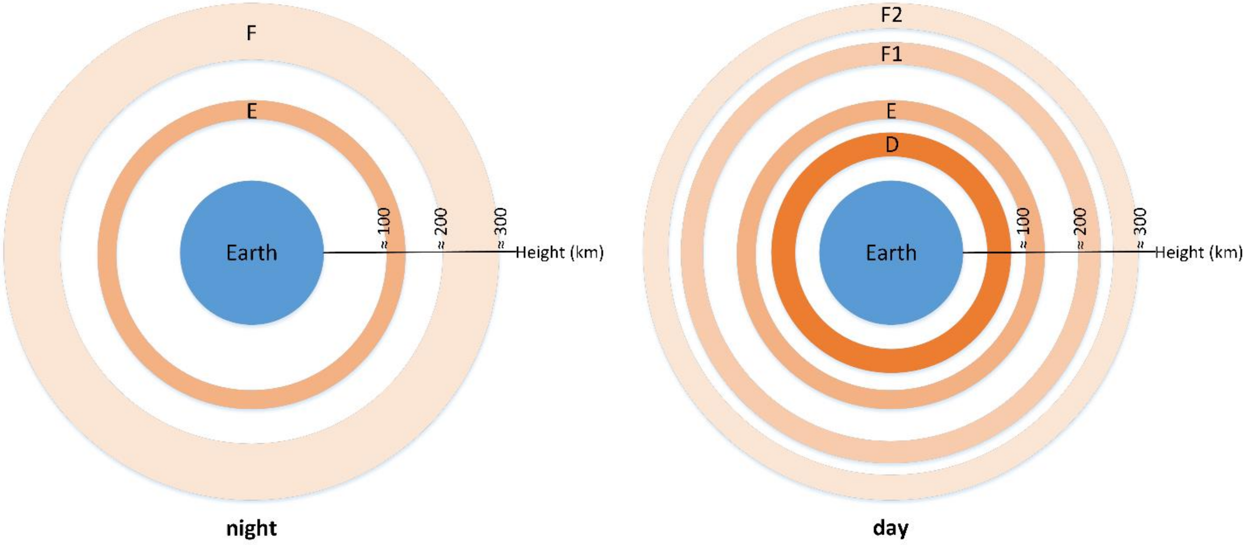


Figure 20: Ionosphere layers day and night time

One of the challenges is to separate these two components from noises. The analysis of an ionogram has two main phases: (1) cleaning the data (filtering the noise from the picture) and (2) performing the desired analyzing techniques for the two components. Another obstacle is the diversity of ionograms: general models cannot be defined. This is why the automatic processing of ionograms is not trivial. There are different partial solutions but none of them works in every ionogram.

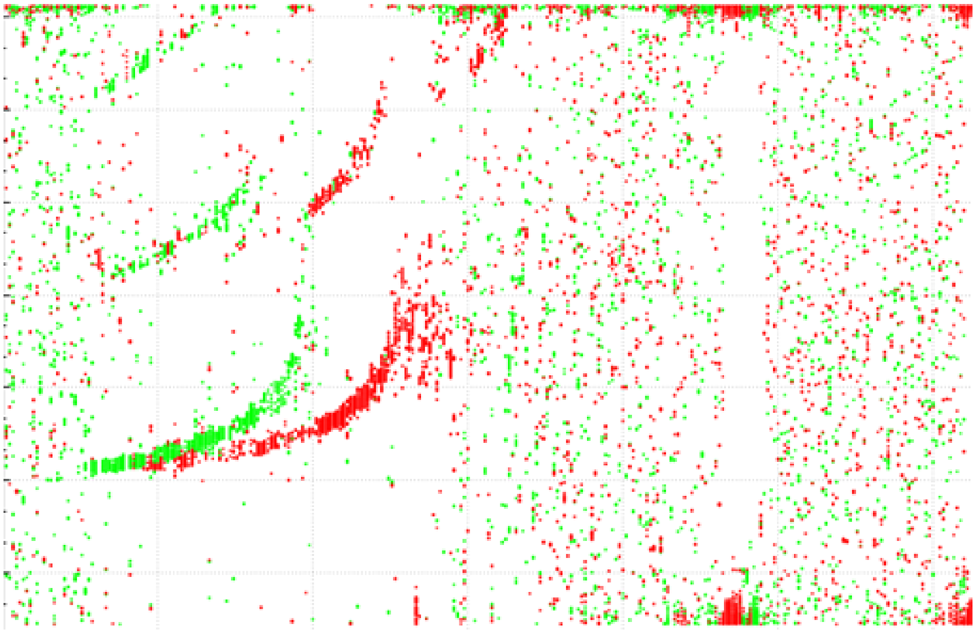
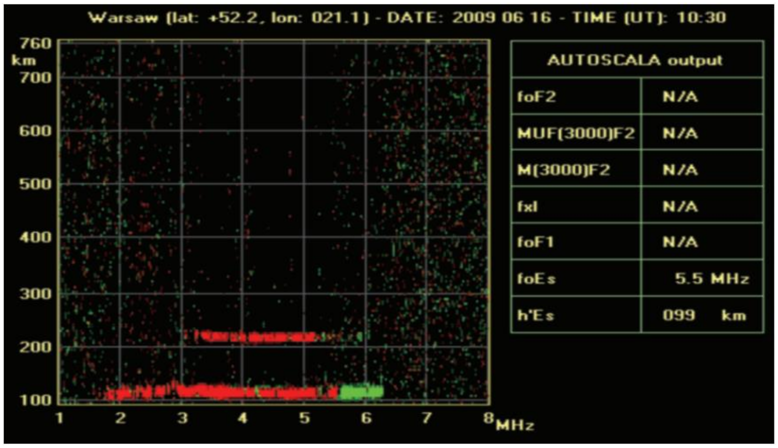


Figure 21: An ionogram example

One of the most popular ionogram analyzing software package is AutoScala developing continuously by an Italian research group. The software was introduced by Pazzopane et al in [180]. The functionality of AutoScala is illustrated in Fig. 22. Using AutoScala, maybe the most ionogram researches were done by this team [181, 182, 183, 184].

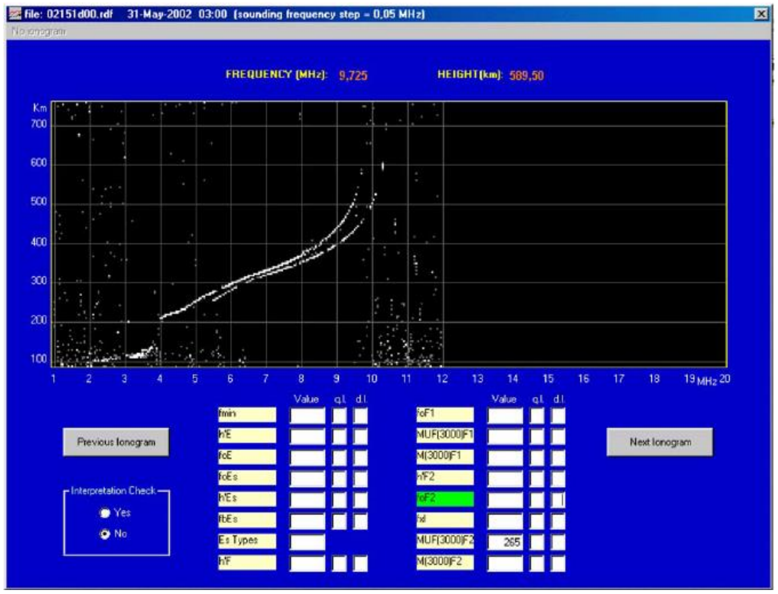


Warsaw (lat: +52.2, lon: 021.1) - DATE: 2009 06 16 - TIME (UT): 10:30

AUTOSCALA output	
foF2	N/A
MUF(3000)F2	N/A
M(3000)F2	N/A
fxI	N/A
foF1	N/A
foEs	5.5 MHz
h'Es	099 km

Figure 22: Ionogram processing in AutoScala, source: [180]

Another application is Interpre, which contains semi-automatic evaluating function as well [185]. The automatic model can be incorrect, therefore manual re-calibration can be necessary. In Interpre, it is possible to define a starting point, and based on this point, another ionogram information (see later, in Table 2) are calculated. But Interpre can only process so-called INGV ionosonde measurements. The user interface of Interpre is illustrated in Fig. 23.



file: 02151d00.rdf 31-May-2002 03:00 (sounding frequency step = 0,05 MHz)

No ionogram

FREQUENCY (MHz): 9,725 HEIGHT(km): 589,50

MUF(3000)F2 265

Interpretation Check
○ Yes
● No

Figure 23: Interpre ionogram processing

The currently used precise evaluation process of ionograms was defined long time ago [100]. There are different methods to create the ionograms but all of them are images from technical point of view and the evaluation process is the same for all ionograms. However, an ionogram is described by several parameters which values needs to be determined. This process is not trivial because of the diversity of ionogram, i.e., diversity of the F1 and F2 curves and the appearing noises. Based on [186], the used parameters are summarized in Table 4.

Table 4: Used parameters when analyzing ionogram data [186]

Parameter	Description
fmin	The lowest frequency at which an ordinary echo is observed on the ionogram
foE	ordinary critical frequency of *E* layer
foEs	*Es* layer highest ordinary frequency, a mainly continuous *Es* trace is observed
fbEs	The blanketing frequency of layer used to derive *foEs*
fminF	Minimum frequency of *E* trace Equals *fbEs* when E presents
foF1	*F1* layer's ordinary critical frequency
foF2	*F2* layer's ordinary critical frequency
fxI	The highest frequency of spread *F* traces (both ordinary and extraordinary).
h'E	*E* layer ordinary minimum virtual height
h'Es	The minimum virtual height of the layer used to derive *foEs*
tpEs (type of Es)	A characterization of the shape of *Es* trace
h'F	*F1* layer's ordinary minimum virtual height
h'F2	*F2* layer's ordinary minimum virtual height

For our case study, the data of ionograms were produced by Geodetic and Geophysical Institute of the Research Centre for Astronomy and Earth Sciences in Hungary. The ionograms are measured in every half-hour. The measured data are in string format. Each row of the file starts with the frequency following the measured value. These values are integers with the following co-domain:

- ***0***, if there is no data.
- ***1***, which means Ordinary component data.
- ***2***, which means Extraordinary component data.
- ***3***, if there is some kind of noise.

Using the concept of UDSS, a session can be defined to analyze ionograms. First, the data stored in files must be uploaded into the UDB entity. This process is performed with DIM. We define transformation rules as follows: the frequency is the main data row and the other values are metadata. We named the new data set as "Ionograms", the main data row is frequency, i.e., 1000, 1025, 1050 etc., and the metadata are 0, 1, 2 and 3. The file name is needed as metadata to identify which ionogram values are related to the given file. To help the work of the researchers, the original files can be stored in the *File* table of UDB structure.

In the query phase, we need all data items. But if we would like to analyze just an *ordinary* component, we need to query only that data. In DMM phase, two-level analysis must be performed, i.e., two phase are needed. In the first phase, the Connected-Component Labeling algorithm was used to determine the relevant areas. This method can classify the data of ionogram and provides related components into one class. The classes containing large number of data must be kept since these represent the relevant areas. Other classes with only a few data can be dropped, and as a result, only "larger" classes remain. (The value of "large" and "few" is defined by the user.) In the second phase, various analyzing methods can be performed according to the desired research.

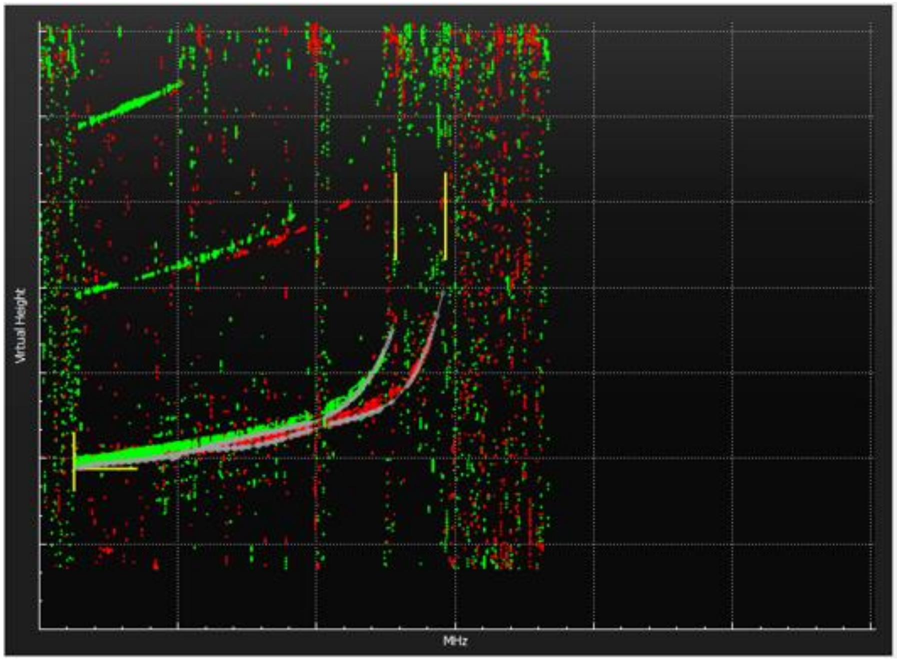


Figure 24: Ionogram best fitting

In ionogram research, the curve of best fit is sought. The white curves are the best approximations of *F1* and *F2* in the example presented in Fig 24. With the least squares approximation technique, the ionosphere consistency can be determined. This is the decision support part of the process. The approximations support to determine the right state of ionosphere at a given time. Ionogram components shape can be various, therefore it is possible that the curve of best fit is not found at the first time. For example, it is possible that not the fourth-degree equation makes the best result, but a fifth- or higher-degree fitting.

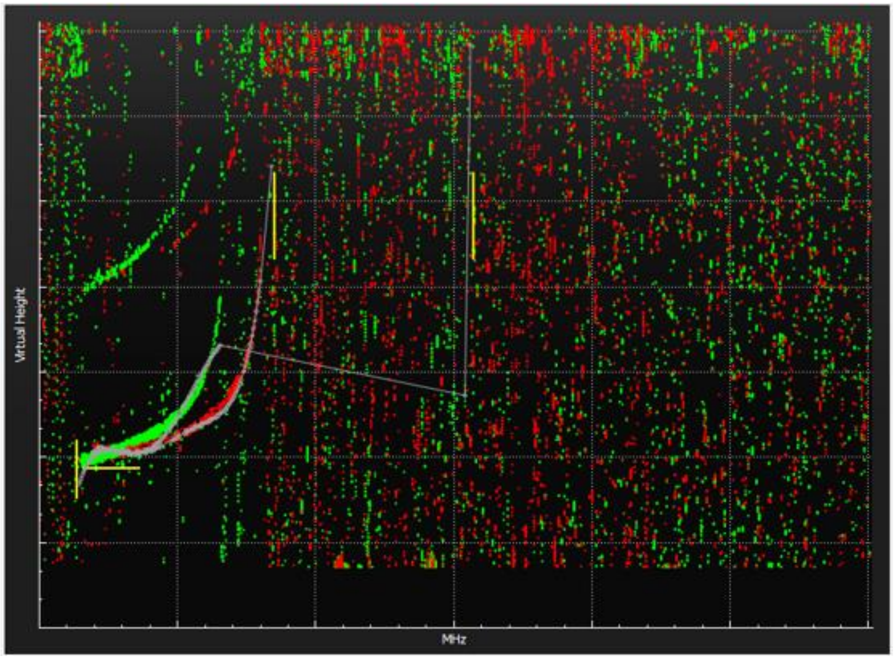
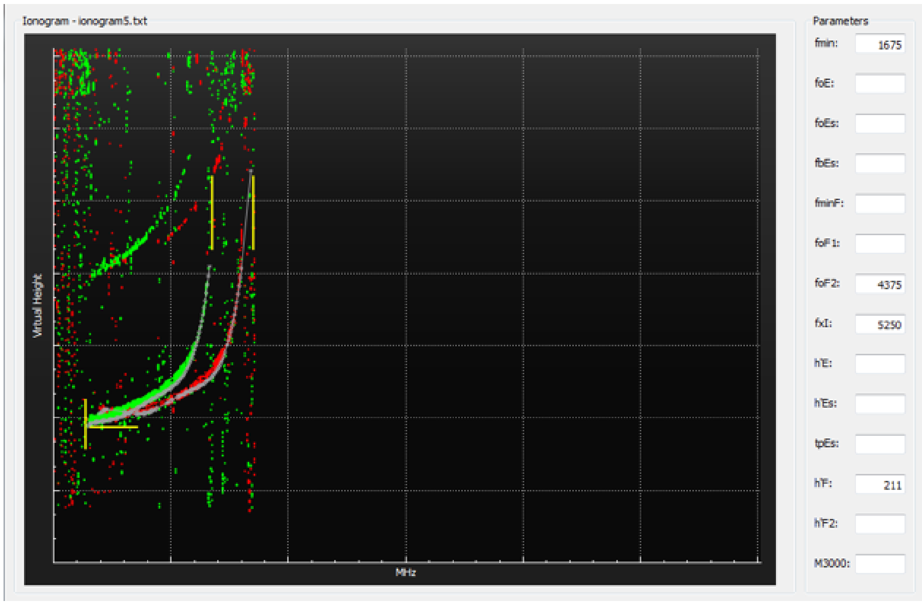


Figure 25: Worst case Ionogram evaluation

Sometimes the fitting algorithm does not work well on the specific ionogram due to its shape and further iterations are needed to get the results. Such a case is illustrated in Fig. 25 when the used techniques cannot find the fitting.

If we use a filtering as a new DMM method in a new iteration, the result will be satisfying as illustrated in Fig. 26.



Parameters

fmin:	1675
foE:	
foEs:	
fbEs:	
fminF:	
foF1:	
foF2:	4375
fxI:	5250
h'E:	
h'Es:	
tpEs:	
h'F:	211
h'F2:	
M3000:	

Figure 26: Good Ionogram evaluation with filtering

At the end of our analyzing session, the parameters of the ionogram are determined as illustrated in the right side of Fig. 31. Further tuning can be performed through Presentation Interface, i.e., values at the right side can be defined by analysts.

This example shows us how the UDSS can solve semi-structured problem. Using UDSS, the following advantage are:

- Universal Database can also store ionogram data.
- Original pictures are preserved, modification (analyzing) can be restored or saved.
- All kinds of ionogram can be analyzed.
- Several algorithms can be executed in each phase. (We used Connected-Component Labeling algorithm).
- Both automatic and manual evaluation can be performed (semi-structured problem solving).

4.7.3. Use Case III: supplier performance analysis

In the third case study, the optimization capability of UDSS is discussed. There is a wood production company where the color of the wooden material is the most important question, because the next production steps depend on it. Vendors ship boards to the company from the forest. Before the shipment, there are some preprocessing of the wood which are very crucial since some factor like drying time can has effect on the color of the wood. Therefore, the boards' color is measured in the first step of the production line at this company.

At a local company, a camera records pictures. Pictures are processed with computer and at the end, the color of wood is determined. This value is named ***fvalue***. For each *fvalue*, the following information are available:

- ***date*** **and** ***time***. The date and time of the measurement performed by the camera.
- ***limit1*** **and** ***limit2***. The low (high) level whereby the result is *dark* (*light*).
- ***result***. This value can be *dark*, *ok*, and *light* based on measured value (*fvalue*), *limit1* and *limit2*.
- ***pieces***. How much same *fvalue* were measured after each other.
- ***itemgroup_desc***: Wood species of the boards.
- ***vendorname***: The name of the vendor.

The *result* column is generated based on *fvalue*, *limit1* and *limit2*. If the measured value is below the *limit1* then the board is *dark*, between the limits is *ok* and if above the *limit2* then *light*. If in one set of the boards have the same measured values, then the *pieces* counter increase by 1. If a value of *pieces* is *3* for example, then it means that the same record was measured with *3* times. Not all *3* records are stored in database, just the value of *pieces*, therefore database is smaller with *pieces* than without *pieces*. The database contains about one million records. The most important question of the company was the rank of vendors.

First step in our session is data uploading with ETL. We had one big Excel file with all data. With our ETL entity, transformation rules can be defined: (1) main data row is *fvalue* and (2) other information are metadata. There is no need to query all data. There are several wood species and we analyze only one at a time, therefore filtering operations can be applied for tree species. In this example our focus was on *birch*. According to the experts of the company, the measured value is not a real value if it is lower than *134* or higher than *199*. We query *fvalue* and *pieces* with the following three conditions:

(1) *Itemgroup_desc = birch*;
(2) *fvalue < 200*;
(3) *fvalue > 133*.

After querying, data manipulation starts. First, the "real" data must be recovered from our universal database. For this, each measured value (row) will be multiplied with the number in *pieces* column and this produces all of the measured data. The method of analysis is ANOVA. This method compares groups' means. Vendors are the groups now, and the measured values (main data row) are in the analysis matrix. The precise operation of ANOVA is discussed in Section 3.3. If ANOVA shows that the means are statistically equal, then we cannot rank them. However, if it shows that means are not equivalent, then we can select vendors, whose means are best with *Duncan* test for example. Duncan test ranks the groups' means starting with the smallest. Calculating critical values, we can filter all set of groups, which means can statistically be seen as equal. Identical groups (vendors) can be determined and if we had such a list, then the company can order shipments from the best vendors. The nearest the mean is to the center of the "ok" range, the better a vendor is. This range starts from *163* and end with *173*.

This test has two conditions: (1) data items must follow the normal distribution and (2) variances must be homogeneous. Before applying ANOVA on the data, its conditions must be checked.

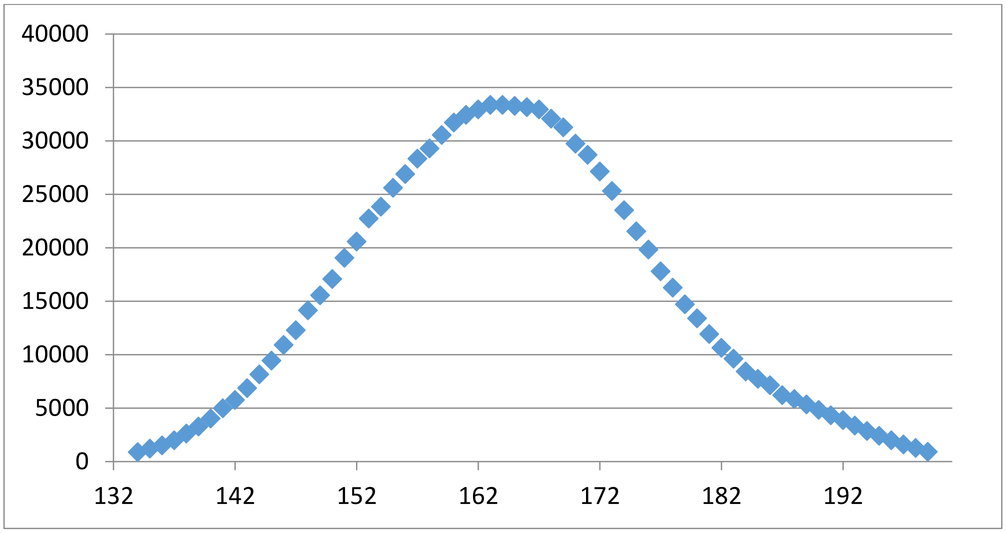


Figure 27: Characteristics of all data items

Fig. 27 is generated by the Presentation Layer. The curve seems to be similar to the normal bell shape. Therefore, we assume that this condition is satisfied. The homogeneity of the variances can be checked with Bartlett-test. The means and variances of the vendors are detailed in Table 5.

Table 5: means and deviations of the vendors

Vendors	Means	Variance	Counts
A	162.77	8.93	30593
B	166.53	8.96	56731
C	164.11	10.97	11776
D	157.82	12.18	11418
E	174.52	10.55	35758
F	162.9	11.38	7484
G	168.51	12.5	194004
H	160.2	11.54	60779
I	164.83	10.85	77569
J	162.95	11.74	427304
M	162	11.53	15870
N	166.99	12.83	41754
O	161.39	10.13	33223
P	165.41	11.08	9454

Based on Table 5, we can assume that variances are equal. But the properties of normality and homogeneity of variances must be proved. The classic *normality chi* test failed and we got the result value *2694.289* with the *chi critical value 55.76* at level α = *0.05*. The value is much bigger as the critical value, but we would expect to be lower based on the values of Fig. 27. We also checked normality with D'Agostino-

Pearson Omnibus test, which analyzes the skewness and kurtosis and comes up with a single *p-value*. Unfortunately, this test is failed as well. We used *Bartlett-test* to check homogeneity of variances. The result *15773.06* was much bigger than the critical value *59.334* at the same significance level.

There are related works where instead of these two prerequisites only the following condition is used: ANOVA must be robust. Accepting this permissive condition, ANOVA can be performed and the result is summarized in Table 6.

Table 6: ANOVA results

Parameters	***SS***	***df***	***MS***	***F***	***p value***	***$F_{critical}$***
Between groups	10389968.79	13	799228.4	5998.189	0	1.720166
Inside groups	135071078.7	1013705	133.245			
Total	145461047.5	1013718				

This *F* value is much bigger than the critical value (α = *0.05*). However, we can assume that the means are not regarded as equal (the smallest value is *157.82* and the biggest is *174.52* in Table 5), and ANOVA supports this theorem, but we can also see that *F value* is unduly big. To further investigate this interesting situation, *Duncan* test was applied. This test selects vendors, whose means are considered to be equal statistically. In other words, inside each groups generated by *Duncan* test, the means are regarded as equivalent statistically, i.e., if we execute ANOVA on each group selected by *Duncan* test, then the calculated *F* value will be lower than *F critical* value. However, no group was found, which means there are no means which are equal statistically according the Duncan test (α = *0.05*). However, vendor F, J and M have nearly the same numerical values nearly in Table 5.

We come to the conclusion that the large number of the data somehow affects on tests' output. In many publications, very large data sets are not checked for normality because of the central limit theorem. The central limit theorem does not state that the larger the sample size, the closer it approximates a normal distribution. In the *R* software, *normality* test can only be done with the limit of 5000.

Since the "*ok*" range is between *163* and *173*, therefore *G* is the best vendor and second best one is *B*. Some vendors' performances are the same (*C* and *I*, or *F*, *J*, *M*). ANOVA also suggests the same results but the result of ANOVA cannot be accepted always because the *F value* is very huge. These contradictory results led us to our new theory named *random correlation*.

In this case study, the UDSS advantages were:

- Supplier selection data can also be stored in UDSS.
- Several algorithms can be performed.
- Our data and analyzing process cannot be performed in ad-hoc DSS, however, vendor selection problem defined in another ad-hoc DSS can be solved using UDSS.
- UDSS can solve optimalization problem as well.
- UDSS can be used in company environment. (This case study is related to production department).

5. Random Correlation

Data are collected to analyze them and based on analysis results, decision alternatives are created. After making the decision, we act according to the selected decision, e.g., we define a correlation between two parameters based on the coefficient of determination or initiate production with the current settings of the new production machine. All decision processes have a validation phase, however, validation can generally only be done after the decision has been made. If we have made a false decision, then the consequences lead to a false correlation or to refuse materials. But how is it possible to make a wrong decision based on data and with precise mathematically proven analysis methods? Our answer is a new theorem, which starts from the point that correlations between parameters and decision alternatives with following proper decision methodologies can be born randomly and this randomness is hidden also from analysts. Studying the state-of-the-art literature about statistical analyzes, we haven't found any literature dealing with this phenomenon. So we started our work on the theory of Random Correlation.

Having studied the literature, we found a lot of contradictory results. The related literature is presented in Section 1.1. According to our state-of-the-art reviews, correlations have not been analyzed from the viewpoint of RC.

5.1. Random Correlation Framework

In this section, random correlation framework and its components are presented. The RC framework has three elements. First, parameters, which is used to describe data structure, will be introduced. Second, calculation methods and models will be presented, with which random correlation can be measured. Random correlation can be triggered by several cause. The third part is the main classes of RC, where each cause is represented by a class.

5.1.1. Definition

The main idea behind the random correlation is that data rows as variables present the revealed, methodologically correct results, however these variables are not truly connected, and this property is hidden from researchers as well. In other words, the random correlation theory states that there can be connection between data rows randomly which could be misidentified as a real connection with data analyses techniques.

There are lot of techniques to measure result's endurance, such as r^2, statistical critical values etc. We do not intent to replace these measurements with RC. The main difference between "endurance measurement" values and RC is the approach of the false result. If we have a good endurance of the result, we strongly assume that the result is fair or the sought correlation exists. RC means that under the given circumstances (see *Parameters* Section), we can get results with good endurance. We can calculate r^2 and critical values, we can make the decision based on them, but the result still can be affected by RC. If we have the set of the possible inputs, the question is how the result can be calculated at all.

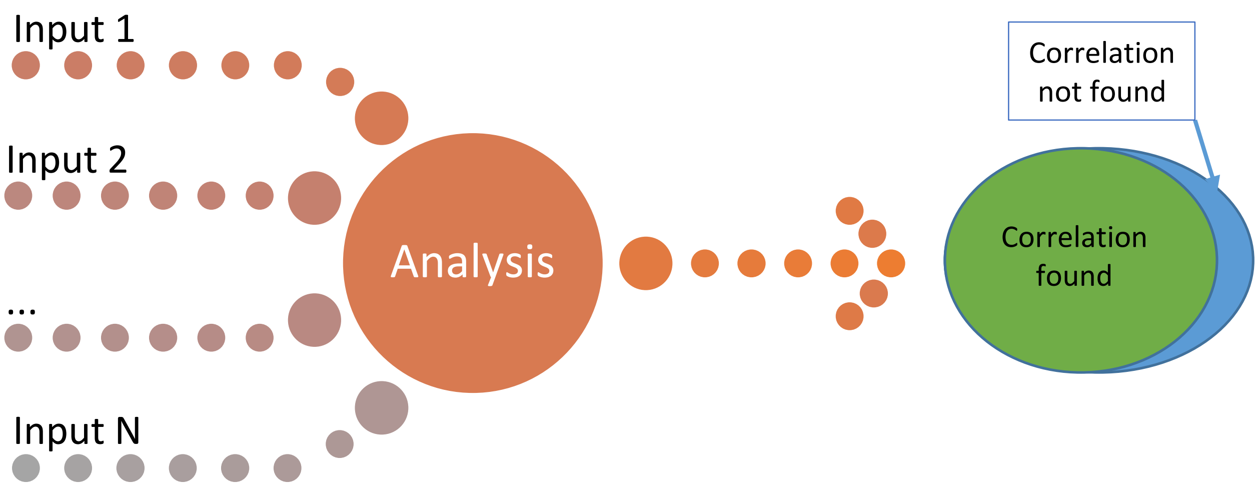


Figure 28: Sematic figure of Random Correlation

As we can see in Fig. 28, RC means that inputs as measured data, and the given method or combinations of methods determine forth the rate of the “correlated” and “non-correlated”. In Fig. 33, we can see that in the set of results, the “correlation found” is highly possible, independent of which “endurance measurement” method is performed. The question is, how such result sets can yell. Which circumstances can cause “pre-defined” results? In our research, we define a Random Correlation Framework to analyze such situations.

5.1.2. Parameters

Every measured data has its own structure. Data items with various, but pre-defined form are inputs for the given analysis. We need to handle all kind of data inputs on the one hand, and to describe all analyzing influencing environment entities on the second hand. For example, if we would like to analyze a data set defined in UDSS with regression techniques, then we need the number of points, their *x* and *y* coordinates, the number of performed regressions (linear, quadratic, exponential) etc. Having summarized, the random correlation framework parameters are:

- ***k***, which is the number of data columns;
- ***n***, which is the number of data rows;
- ***r***, which is the range of the possible numeric values;
- ***t***, which is the number of methods.

To describe all structure, matrix form is chosen. Therefore, parameter *k*, which is the number of data rows [also the columns of the matrix], and *n*, which is the number of data items contained in the given data row [also the rows of the matrix], are the first two random correlation parameters.

The third parameter, range *r* means the possible values, which the measured items can take. To store these possibilities, the lower (*a*) and upper (*b*) bounds must only be stored. For example, *r*(1,5) means the lower limit is 1, the upper limit is 5 and the possible values are 1, 2, 3, 4 and 5. Range *r* is not a very strict condition because the measured values intervals can be defined many times, these values are often between these lower and upper bounds. A trivial way to find these limits, when *a* is the lowest measured value and *b* is the highest one. They can be sought non-directly as well. These bounds are determined by an expert in this case. E.g., a tree grows every year, but it is impossible to grow 100 meters from practical point of view. The longitude line is infinity, but it is possible to define *a* and *b*. Although in our work integers are used, it is possible to extend this notation for real numbers since the possible continuous nature of the measured data. The continuous form can be approximated with discrete values. In this case, the desired precision related to *r* can be reached with the defined number of decimals. The sign *r*(1,5,¨) means the borders are the same as before, but this range contains all possible values between 1 and 5 up two decimals.

Parameter *t* is the number of methods. We assume that if we execute more and more methods, the random correlation possibility increases. For example, if *t* = 3, that means 3 different methods are performed after each other to find a correlation. This *t = 3* could be interpreted several ways related to the specific random analyzing process. We have the following four interpretations for *t*:

1) Interpretation 1: The number of methods;
2) Interpretation 2: The input parameter's range of the given method;
3) Interpretation 3: The decision making according to divisional entity level (output parameter);
4) Interpretation 4: The outlier analysis.

Interpretation 1 is trivial since it represents the number of methods. *Interpretation 2* means the following: an input AS parameter can influence the results. In this case, the more input parameters' values are used during the analysis, the higher the possibility that true results born. For example, in statistics, the significance level (α) can be chosen by the user. However, different levels of α have a precise statistic background, it is possible to increase this level by the scientists, which cause H_0 true sooner or later according to Bonferroni. In other words, Bonferroni claimed that if we have more and more data rows, we have connections between them at higher probability. But the type I error is in the background, which increases in the case of more data rows. Type I error means that we reject H_0, however it is true [188]. Extending his theory, we state that if we use the correction of Bonferroni, then we still can have random correlations.

Interpretation 3 is similar to the second one but this regards to output parameters. It is such an entity, which value can influence the decision. While *Interpretation 2* refers to a calculated number, *Interpretation 3* means that entity, which will be compared with the calculated value. For example, in the case of regressions, choosing r^2 level can be different. There are rules to define which result is "correlated" or "non-correlated". However, these rules are not common and there can be agreed that results with *0.8* or *0.9* are "correlated", but *0.5* or *0.6* are not so trivial. This kind of output divisional entities values have strongly effect on decisions.

Interpretation 4 is necessary most of the cases, however, by performing more and more outlier analysis, the *t* can increase heavily. It is trivial, that the junk data must be filtered. For example, as we mentioned in Section 3.5.3, trivial valuable data are between *134* and *199* in the example of vendor selection. However, not all cases is so simple. In case of ionogram in Section 3.2.2., semi-automated analyzing is better, because of the various data shapes, i.e., ionogram curves, it is hard to make a decision about which points should be kept and which not. However, in the view of RC, the main problem is still that if we perform more and more techniques then we will get some kind of mathematically proved result. Moreover, by combining (1)-(4), we get a result anyway, decreasing the chance of "non-correlated". For example, combining regression techniques with outlier analysis, the "no good" points filter possibilities can raise and the result gets seemingly better. However, the result has no choice but becoming "correlated", which can be led us to RC.

Since we do not know every parameters and every variable range, two classes were defined:

1. **Closed system**. All analysis parameters are known, the correlation is sought between these attributes.
2. **Opened system**. We do not know all possible parameters related to the research. The number of possible variables is infinite.

Random correlations can be interpreted in the viewpoint of both systems. Although the opened system's RC factor can be determined uneasily, sometimes it is possible. The closed system's RC factor can be determined with RC models and methods.

5.1.3. Models and methods

In this section, that methods are introduced, with which the RC occurrence possibility can be calculated. In the context of random correlation, there are two main models:

(1) We calculate the total possibility space [Ω];
(2) We determine the chance of getting a collision e.g., find a correlation.

In the case of (1), all possible measurable combinations are produced. In other words, all possible *n*-tuples related to *r(a,b)* are calculated. Because of parameter *r*, we have a finite part of the number line, therefore this calculation can be performed. That is why *r* is necessary in our framework. All possible combinations must be produced, which the researchers can measure during the data collection. After producing all tuples, the method of analysis is performed for each tuple. If "correlated" judgment occurs for the given setup, then we increase the count of this "correlated" set S_1 by 1. After performing all possible iterations, the rate *R* can be calculated by dividing S_1 with $|\Omega|$. *R* can be considered as a measurement of the "random occurring" possibility related to RC parameters. In other words, if *R* is high, then the possibility of finding a correlation is high with the given method and with related *k*, *n*, *r* and *t*. For example, if *R* is *0.99*, "non-correlated" judgment can be observed only *1%* of the possible combinations. Therefore, finding a correlation has a very high possibility.

Contrarily, if *R* is low, e.g., *0.1*, then the possibility of finding a connection between variables is low. We accept the rule of thumb, that correlation possibility should be lower than the "non-correlated" case. However, there is a third option as well. If the correlated and non-correlated judgments can be meaningful in the view of the final result. For example, in the case of Analysis of variance (ANOVA) both H_0 and H_1 can be meaningful, therefore *R* should be around of 0.5. Related to the whole possibility space, this RC model is named *Ω-model*. The steps of *Ω-model* is summarized in Fig 29.

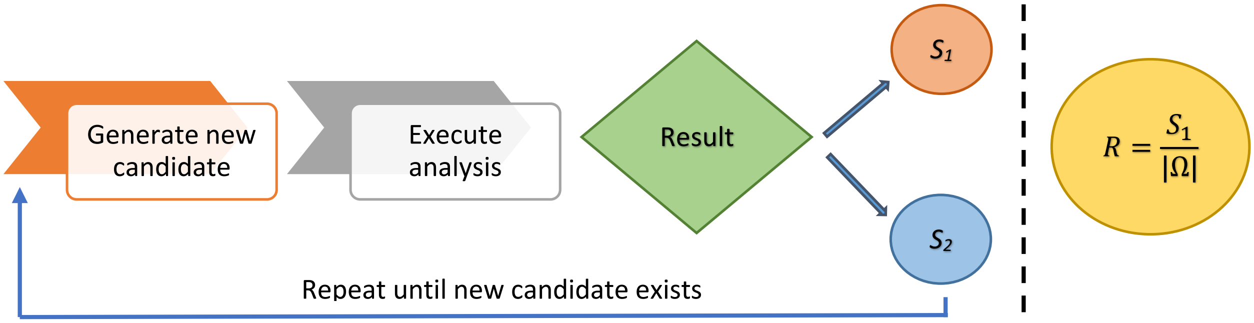


Figure 29: *R* calculation process

In Fig. 29, S_1 represents that correlation was found, while S_2 represents that correlation has not been found. The algorithms' pseudo code is the following:

```
REPEAT
    Get(newCandidate);
    Execute(method);
    IF findCorrelation(true)
        $S_1 = S_1 + 1$;
    ELSE
        $S_2 = S_2 + 1$;
UNTIL Exists(NewCandidate);
$R = S_1 / |\Omega|$;
```

The calculation of S_2 is not necessary, because during the final calculation only S_1 is used. It does not matter that S_1 or S_2 is the numerator. If we choose S_2 as numerator, the result *R* shows the rate of the "non correlated". In the case of S_1, the *R* means the rate of "correlated".

In the case of (2), rate *C* is calculated. This shows how much data are needed to find a correlation with high possibility. Researchers usually have a hypothesis and then they are trying to proof their theory based on data. If one hypothesis is rejected, scientists try another one. In practice, we have a data row *A* and if this data row does not correlate with another, then more data rows are used to get some kind of connection related to *A*. The question is how many data rows are needed to find a certain correlation. We seek that number of data rows, after which correlation will be found with high possibility. This method is named *Θ-model*. There is a rule of thumb stating that from 2 in 10 variables (as data rows) correlate at high level of possibility, but we cannot find any proof, it rather is a statement based on experiences. The calculation process can be different depending on the given analyzing method and RC parameters.

During the *C*-model calculation process, we generate all possible candidate (*Ω*) based on RC parameters first. We create individual subsets. It is true for each subset that every candidate in the given subset is correlated with each other. We generate candidates after each other and during in one iteration we compare the current generated candidates with all subsets' all candidates. If we find a correlation between the current candidate and either candidates, then the current candidate goes to the proper subset which the "correlated" candidate belongs to. Otherwise, a new subset is created with one element, i.e., with the current candidate. *C* is the number of subsets.

C show us that how many datasets must be measured during the research to get a correlation with at least two datasets for sure. The pseudo code of the *C*-model is the following:

```
counter = 1;
flag = TRUE;
Create($H_i$);
Put($H_i$, firstCandidate);
REPEAT
        newCandidate = Generate();
        FOR($i$ = 1; $i$ < $H_{counter}$; $i$++)
                FOR($j$ = 1; $j$ < $|H_i|$; $j$++)
                        currendCandidate = $H_{i,j}$;
                        ExecuteMethod(newCandidate, currentCandidate);
                        IF ExecuteMethod == TRUE THEN
                                $H_i$.ADD(currentCandidate);
                                flag = false;
        IF flag = false THEN
                counter = counter + 1;
                Create($H_{counter}$);
                $H_{counter}$.ADD(newCandidate);
                flag = TRUE;
UNTIL Exists(NewCandidate);
C = counter;
```

Based on value *C*, we have three possible judgements:

- *C* is high. Based on the given RC parameters, it must be lots of dataset to get a correlation with high probability. This is the best result, because the chance of RC is low.
- *C* is fair. The RC impact factor is medium.
- *C* is low. The worst case. Relatively few datasets can produce good correlation.

5.1.4. Classes

RC can be occurred because of different causes. Data, the research environment, the methods of analysis can be different, therefore, classes are added to the framework. Each class is represented by a cause, which along the RC as phenomena can be appeared.

Class 1. Different methods can be applied for the given problem. If we cannot find good results with one method, then we choose another one. The number of analysis can be multiplied not just with the number of chosen methods but with methods' different input parameters range and seeking and removing outliers. There could be some kind of error rate related to method's results. It is not defined when the data are not related to each other. When we use more and more methods with different circumstances, e.g., different parameters and error rates, then we cannot be sure whether a true correlation was found or just a random one. If we increment the number of the methods, then there will be such a case when we surly find a correlation.

Class 2. Two (or more) methods produce opposite results. But this is not detected, since we stop at the first method with satisfying result. It is rather typical finding more precise parameters based on the "correlation found" method. When methods are checked for this class, there can be two possibilities: (1) two or more methods give the same "correlated" result and (2) one or more methods do not present the same results. In option (1), we can assume that the data items are correlated truly with each other. In option (2), we cannot make a decision. It is possible, that the given methods present the inconsistent results occasionally or they always produce the conflict near the given parameters and/or data characteristics. There is a specific case in this group, when one method can be inconsistent with itself. It produces different type of results near given circumstances, e.g., sample size.

Class 3. The classic approach is that the more data we have, the more precise results we get. But it is a problem if a part of the data rows produces different results than the larger amount of the same data rows. For example, a data row is measured from start time t_0 to time t, another is measured from start time t_0 to time $t + k$, and the two data sets make inconsistent result. This is critical since we do not know which time interval we are in data collection. For this problem, the cross validation can be a solution. If all subsets of the data row for the time period t are not perform the same result, we only find a random model at high probability level. If they fulfill the "same result" condition, we find a true model likely.

This third group has another concept, which is slightly similar to the first one. We have huge amount of data sets in general, and we would like to get correlation between them. In other words, we define some parameters, which was or will be measured, and we analyze these data rows and create a model. We measure these data items further [time $t + k$]. If the new result is not the same as the previous one, we found a random model. The reason could be that a hidden parameter was missed from the given parameters list at the first step. It is possible, that the value of these hidden parameters change without our notice, and the model collapse. This kind of random correlation is hard to predict.

5.1.5. Analyzing process

RC tells that it is feasible to get the mathematically proven results so, that another results could not come out at higher possibility, just the given one. If we would like to perform a research starting from data management ending with results and publication, then the given process can (should) be analyzed in the view of random correlation as well. In the RC framework, to perform an RC analyzing session related to the given analyzing procedure, six steps were defined.

1. Introducing the analyzed method's basic mathematical background;
2. Introducing which random correlation class contains the given case;
3. Define what is exactly understood under the method's random correlation;
4. Define and choose random correlation's parameters;
5. Calculations and proving;
6. Validation with simulations and interpretation.

Step 1 is an optional one, but some basic overview of the given method could be necessary for the further analyses. *Step 6* refers for a computer program in general, but it always includes making decision about the given RC analyzing.

Now, we have a standard RC analyzing session and the suitable RC entities (parameters, calculation methods and classes) must be chosen in each step. From now, we focus on practice and analyze methods from the point of RC view.

5.1.6. Simple example: χ2 test

In almost all statistical test, normality is the first assumption. The normality test is used to check whether the given measured data follow the normal distribution or not. Several test is known, χ^2 is maybe the oldest.

The χ^2 test mathematical background is summarized in Section 2.1.1.1. [*Step 1*]. The χ^2 test belongs to *Class 3*: increasing the number of data supports contradictory results [*Step 2*]. In the view of RC, we seek that circumstances, where χ^2 test accept the normality (H_0), and where does not (H_1) [*Step 3*]. We use *k* and *n* RC attributes, where *k* is the number of classes, and *n* is the number of data [*Step 4*]. Our calculation process is based on Eq. 3 [*Step 5*]. Because of the square function, the numerator is positive either way. The denominator (E_i) is also positive, therefore the result of division is positive. Each class (*k*) have this error, and at the end, we sum up the positive errors. Adding large number of *k* errors, the result will also be large at high possibility level. The χ^2 distribution curve is different according to degrees of freedom and therefore the critical value is also growing. It is illustrated in Fig. 30.

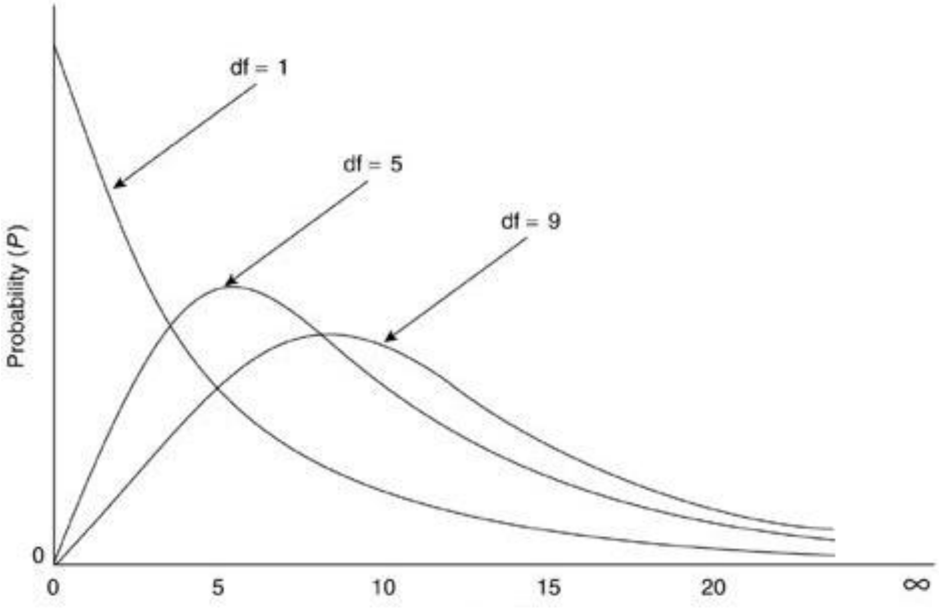


Figure 30: χ^2 distribution with different degrees of freedom

According to curves and degrees of freedom, the critical value, which is the area under the curve (integral), does not grow as fast as it would according to *k*. Several integral values can be seen in Table 7.

Table 7: Several critical value of χ^2 with degrees of freedom

df	1	4	9	19	39	59	119	999
integral	**3.841**	**9.488**	**16.919**	**30.144**	**54.572**	**77.931**	**145.461**	**1073.643**

The bigger *k* and *n* are*,* the bigger the chance that the χ^2 test refuses normality is. If we have *k* = 5 groups, and in each group have the error of 1.5, the sum of error is 7.5, which is lower than the χ critical value 9.488 with $df = 5 - 1 = 4$ at α = 0.05 and the test accepts H_0. If we have 40 classes, as we had in Fig. 27, then the error is $40 * 1.5 = 60$. The χ critical value is 54.572, which is lower than 60, therefore H_1 is accepted. Further increasing *k*, the difference between the critical value and the calculated value (sum of errors) will be larger. Increasing *n*, the margin 1.5 is very good, because if an O_i value is 160.2 (vendor H in Table 5), the margin 1.5 from E_i is very good. In the case of small *n*, 1.5 can be seen as large margin objectively. If the O_i would be 3, then E_i is 4.5, which is half as big again as 3. Increasing both *k* and *n*, the errors can be larger which induces that χ^2 test punishes larger amount of data [*Step 6*]. It can lead to that analyzing with small *k* and *n* can show normality property in the most cases, while it declines normality in the case of big data inspired environment. According to our view, χ^2 normality test cannot be used in the case of big data inspired environment, and it is affected by RC. Our opinion is supported by *R* statistical software as well, because χ^2 test cannot be performed more than 5000 data. However, 5000 can also be large. This RC property of χ^2 test is hidden also from the analyzers. We replaced χ^2 test with D'Agostion Pearson normality test to check normality, if normality is an assumption. D'Agostion Pearson test is based on the skewness, which tells us the amount of distortion from the horizontal symmetry and the kurtosis, which is the measurement of how tall and sharp the central shape is. Since D'Agostion Pearson test has another mathematical background, therefore the problem of the χ^2 is not appeared.

5.2. RC analysis: ANOVA with *Ω*-model

According to RC analysis' steps, the first step is to give a short overview of the analyzed method. In the case of ANOVA, the summary was introduced before [*Step 1*]. As for the random correlations, ANOVA belongs into *Class 2* and it has a specific place: both H_0 and H_1 can be meaningful *[Step 2]*. The "non-correlated" can be defined as the means are statistically similar [H_0], therefore the influencing variable has no effect on the subject ["non-correlated"]. The H_1 means that variable has influence ["correlated"]. In this case, the random correlation means that the H_0 or the H_1 can take priority over against the other according to parameters defined later in Step 4. The seeking rate should be around 0.5 [*Step 3*]. Small deviation is allowed, but huge distortion is dangerous.

The following entities are used in this case *[Step 4]*: (1) Number of measurement (*k*); (2) Number of data item related to one measurement; (3) Range (*r*) where *r* contains the lower bound *a* and upper bound *b*. Since ANOVA is in *Class 2*, parameter *t* is not used. To calculate all combination and perform ANOVA in each case *[Step 5]*, it is practical to develop a computer program. But investigating other RC environments, UDSS concept can solve this task better with the Data Manipulation Module.

ANOVA assumptions must be handled in the implementation. ANOVA first assumption is that sampling must be done randomly. In the view of RC, we can assume it is passed. Another basic assumption is normality. Therefore, all produced candidates must be checked in the view of normality.

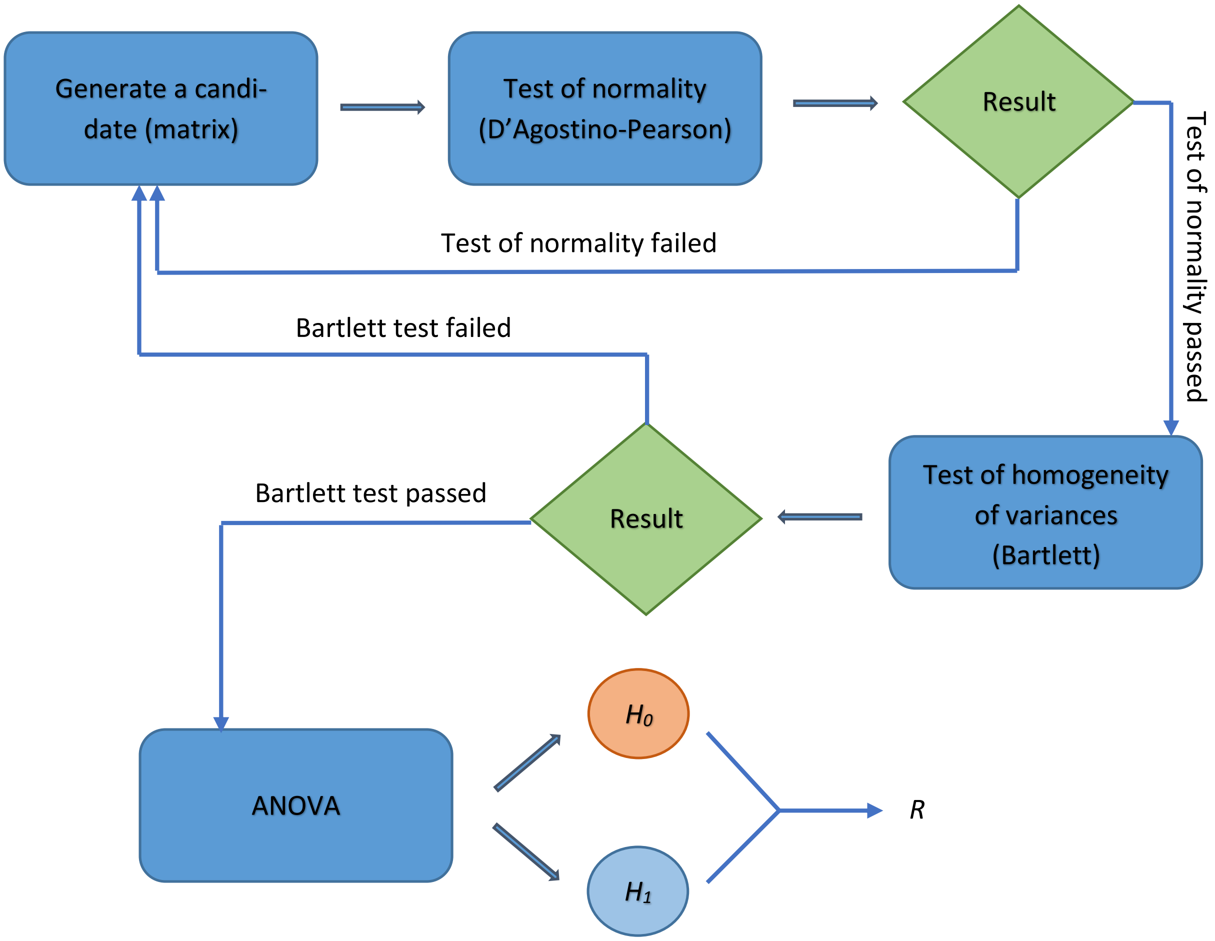


Figure 31: ANOVA RC process overview

The combinations, which do not follow the normal distribution, must be deleted in the set of candidates. We used the D'Agostino-Pearson test to check normality. Another assumption can be that the variances must be statistically equal. This was checked with Bartlett test. Finally, we calculate the rate *R [Step 6]*. This *R* can also be calculated by the same program.

In Fig. 31, one candidate is a matrix, because in the case of ANOVA, there are several measurements. Parameters *k*, *n*, and *r* are used and the calculation method is the calculation of the total possibility space. The *R* means the balancing indicator between H_0 and H_1. If *R* is not balanced, for example, H_0 chance is 0.95 according the whole possibility space, that means this kind of possible variations are more frequent then the others, e.g., H_1's. This means that creating the "non-correlated" result is much easier.

Unfortunately, the problem of RC problem cannot be solved with sampling. To illustrate this, let us study the following example. We know that ANOVA's input is always a matrix. If we have *k = 3*, *n = 4* and *r*(*1*,*3*), the first six candidates are the followings:

$$\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}(I)\ \begin{bmatrix}2&1&1\\1&1&1\\1&1&1\end{bmatrix}(II)\begin{bmatrix}3&1&1\\1&1&1\\1&1&1\end{bmatrix}(III)\begin{bmatrix}1&2&1\\1&1&1\\1&1&1\end{bmatrix}(IV)\begin{bmatrix}1&3&1\\1&1&1\\1&1&1\end{bmatrix}(V)\begin{bmatrix}1&1&2\\1&1&1\\1&1&1\end{bmatrix}(VI)$$

Producing candidates continues until the last candidate is produced, i.e., all items have the value *3*. We perform ANOVA for every candidates that will eventuate in one result for each candidate, i.e., H0 or H_1 hypothesis will be accepted in each case. What does sampling mean? Sampling is a subset of the total possibility space (*Ω*) and a given sample is one of the candidates. From the viewpoint of the analysts, sampling is one of the applied methods. But the total possibility space is not always known for the analysts, therefore RC can be hidden easily.

5.3. Total possibility space problem and candidates generates

5.3.1. Overview

Before we present the numerical results of RC analysis, the total possibility space will be examined. Related to *Ω*, increasing value of RC parameters cause exponentially growing space, e.g., value of parameter *k*, *n* and *r*. If we analyze *k = 4*, *n = 8*, *a = 1* and *b = 5*, then the whole space to be calculated is $5^{32} = 2.328 * 10^{22}$. These huge numbers cannot be evaluated in real time even with a computer program. Producing all combinations in greedy way is not an option. Some Space Reducing Techniques (SRT) must be performed. SRT depends heavily on the given method. Therefore, in each case of method, the own space reducing algorithm (SRA) must be developed. The SRT is a set, and an SRA is an item in this SRT set. A self-developed SRA named Finding Unique Sequences algorithm (FUS) is applied for ANOVA.

5.3.2. Finding Unique Sequences algorithm

Based on Table 5, we get *F* value on division *MSB* and *MSW* and each of these be calculated by division with degrees of freedoms [$k - 1$ and $k * (n - 1)$]. The degrees of freedoms can be seen as constant in the view of each case. Therefore, we need to calculate only *SSB* and *SSW* values, then the further divisions are determined and produced the given *F*. That can be seen in the next matrixes.

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 3 \\ 3 & 2 & 1 \end{bmatrix} (a) \quad \begin{bmatrix} 3 & 3 & 3 \\ 2 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix} (b)$$

Since the (*a*) and (*b*) matrixes have same *SSB*, *SSW* and degrees of freedoms, the *MSB* and *MSW* are also the same including *F* value.

ANOVA compares the inner and the outer variances. The outer one is the variance of group's means, but one group's mean is determined by data values, which are related to the inner variance. Therefore, we continue with the inner variance examination. This means, that we must calculate only one column total possibility space and this calculation is repeated *k* times. We remark that is not enough to store the *SSW* only, because one *SSW* can belong to one *m* mean, but one mean can belong to several *SSW*. If we have $(i)[1, 2, 2]$ and $(ii)[1, 1, 3]$ the $SSW_{(i)} = 0.6666$, $SSW_{(ii)} = 2.6666$, while the means are the same *1.6666*. This leads to different *F* values. Therefore, we must store *mean – SSW* tuples. This is the first level of decreasing of total space calculation.

The second level means that there is no need to produce one column's all possibilities either. We need to produce those combinations which have different *SSW*. This can be performed with repeated combination technique. The total possibility space must be equal with these two level decreasing therefore we need to store the frequency for each *mean – SSW* tuple as well. The frequency of one tuple can be calculated with the repeated permutation:

$$\frac{n!}{s_1! * \; s_2! \ldots s_i!}, \qquad \text{Eq (14)}$$

where *n* is the number of elements, s_i is the number of repetitions. This number is not large generally. We produce one group's all repeated combinations, then we calculate each combination's *mean*, *SSW* and *frequency*. We can calculate in order *SSB*, *MSW*, *MSW*, *F* based on these triples. The frequency of *F* can be calculated the follow:

$$F_{i,k,n,a,b} = \left(\prod_{i}^{k} C(SSW_i)\right) \; * C(m_i), \qquad \text{Eq (15)}$$

where *C(SSW_i)* is the count of the SSW frequency and *C(m_j)* is the count of the given means combination.

A given *F* is compared with the $F_{critical}$, and since the frequency of that comparison is known, we can define how many times this judgment occurs. If this judgment is zero, then the number of zeroes is increased. If we have one, then the number of ones is increased. A good validation process is comparing the all possibility space number with the sum of zeroes and ones, which can be equaled. The procedure of this algorithm is summarized in Fig. 32.

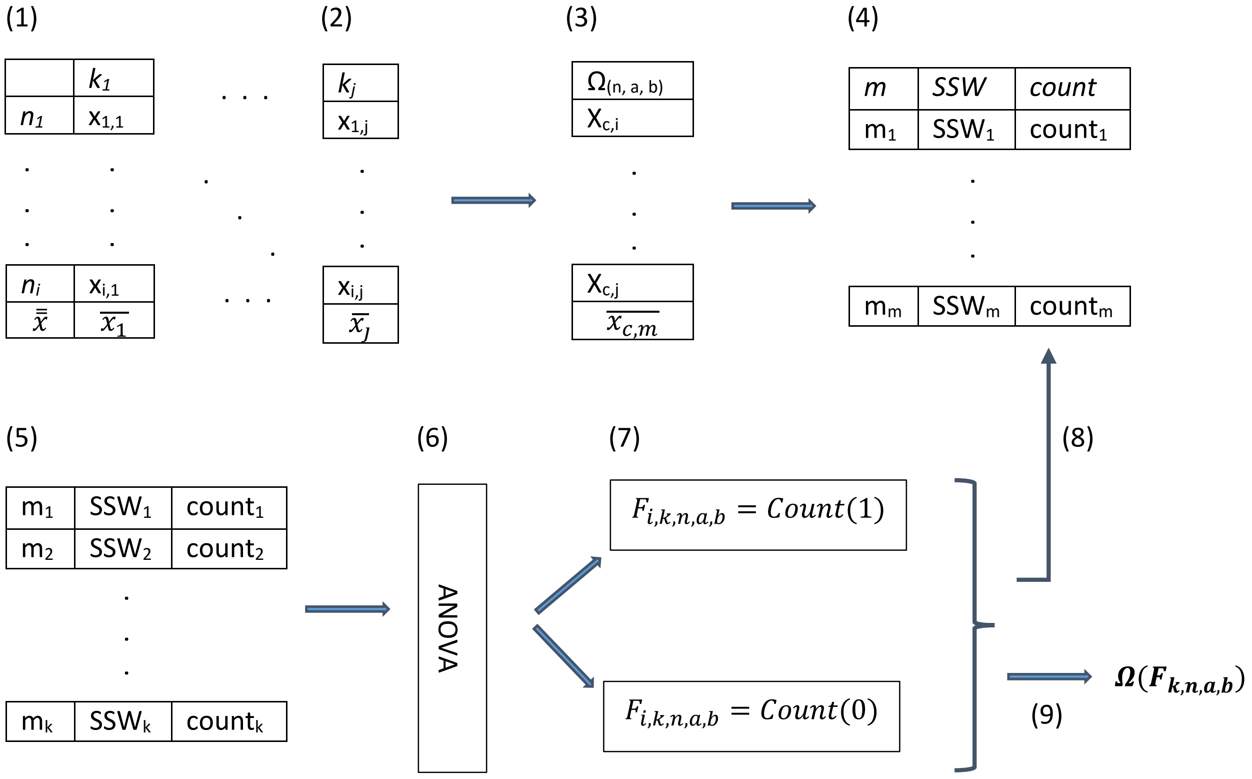


Figure 32: χ^2 distribution with different degrees of freedom

After producing the triples, all possible subsets are calculated which have *k* elements. The number of ones or zeroes are registered according to the ANOVA judgment based on the given subset [Step 7]. This operation is iterated [Step 8] until all repeated combination are created, the rate *R* can be determined [Step 9].

The whole process can be calculated at two levels based on the following equation:

$$\frac{(w+v-1)!}{v! * (w-1)!}, \qquad \text{Eq (16)}$$

where *v* is *n* and *w* is $r = b - a + 1$, in the first calculation phase [triples count], and *v* is the calculated number of the first case and *w* is *k* in the second phase [all possibilities]. The number of triples can be produced quickly in general, but the second step's calculation takes a lot of time in the case of large value of parameters. It follows that increasing *n* raises the calculation time indirectly, while increasing *k* raises it directly. Increasing range has affect the first step because we can make more combination related to *n*. In our calculation example with *k = 4*, *n = 8*, *a = 1*, *b = 5*, the total possibilities was $2.328 * 10^{22}$, if we use our method then we need to use eq. (3) twice. First we need to calculate the number of triples, which is 126 according to the example. The *v = 792* and *w = k = 4* in the second calculation, which is $1.651 * 10^{10}$. If we increase *k* only by 1 then this number will raise two magnitudes [10^{12}]. That means despite the decreasing property of the reviewed method the whole calculation can take a lot of time unfortunately. Because of that we have limitation seeking rate *R*. However, this SRT allows to examine ANOVA with bigger RC parameter values then in the case of greedy way.

5.3.3. Candidates producing and simulation level

It is possible that SRTs cannot grant enough space reduction in the case of huge RC parameter values. Therefore, simulation techniques must be applied to approximate the seeking of *R*.

Level 1. The trivial way to generate data rows randomly according to given *k*, *n* and *r*. We perform the analysis and notify the number of “correlated” and “non-correlated” cases. Based on these numbers, an *R’* can be calculated. Based on the definition of possibility, *R* is approximated by *R’*. This is the fastest way to get an estimated *R*, however, calculating *R’* would be precise only if the iteration *i* is large enough. After a certain level, performing *i* iterations cannot be possible in real time.

Level 2. The first phase of the SRT can be used to get more precise estimation of *R*. Because of the square function, the SRT first phase can be done quickly as we mention before. The problem is related to *k*, that is all *k* subsets must be produced from the result of the SRT first phase. However, if we produce all first phase candidates, i.e., use repeated permutation, and next, use simulation technique, i.e., randomly chosen k subsets in *i* iteration, then the second phase has an input, which contain only the accepted normal candidates. Therefore, more precise *R’* can be determined.

Level 3. The first phase candidates preparation and the related frequency *F* can be combined. At *Level 2*, *k* data rows are chosen, but its weight is 1, i.e., one judgment is calculated. If a data row was chosen in an iteration, and since its *F* is known, we can define a weight based on this *F*. For example, in *k = 3* case, at *Level 2*, just one judgment is, however, at *Level 3*, $F_1 * F_2 * F_3$ judgments are produced. In other words, when three given data rows are selected, then as a matter of fact, all their permutations are chosen as well, because in the first phase, one row represents a combination with its own all permutations, i.e., frequency *F*. We know that all F_k have the same result as in the case of *Level 2*. Therefore, we get more information in one iteration. In the next iteration, these 3 data rows [or neither their permutations of course] cannot be selected. This level produce more precise approximation of *R* with *i* iteration because of the known frequencies. The rate of *Level 3* is signed with *R**. *R** approximates *R* better than *R'*. This level can be only use when the first phase can be calculated in real time.

5.4. ANOVA results related to Ω (*R*)

In this section, we calculate the whole possibility space and determine rate *R* for ANOVA. First, SRT is used, the *Level 2* simulation is applied.

Table 8. ANOVA results with *r*(1, 3)

r(1,3)	**k = 3**	**k = 4**	**k = 5**	**k = 6**	**k = 7**	**k = 8**
n = 4	*1*	*0.9991*	*0.9997*	*0.9999*	*0.9999*	*0.9999*
n = 5	*0.9877*	*0.9923*	*0.9946*	*0.9957*	*0.9968*	*0.9976*
n = 6	*0.9744*	*0.9788*	*0.9818*	*0.9847*	*0.9864*	*0.9876*
n = 7	*0.9699*	*0.9741*	*0.9768*	*0.9797*	*0.9817*	*0.9831*
n = 8	*0.9630*	*0.9661*	*0.9685*	*0.9707*	*0.9721*	*0.9736*
n = 9	*0.9563*	*0.9594*	*0.9611*	*0.9621*	*0.9635*	*0.9645*
n = 10	*0.9656*	*0.9695*	*0.9722*	*0.9741*	*0.9756*	*0.9768*

In the case of small *n*, *k*, and *r*, the rates show high possibility in the favor of H_0. If the RC parameters values are small, then whatever researchers measure, the result will be H_0 at very high possibility, the compared data are statistically equal on average. The specific case is *n = 4* and *k = 3*, because the chance of H_0 is 1, which means there is no such input matrix eventuating H_1. However, according to big data inspired environment, larger *n*, *k* and *r* should be used. Therefore, the parameter values are changed in the further calculations. The results with different *r*, *k*, and *n* are shown in Table 9.

Table 9: ANOVA R values with different *r*, *k*, and *n*

r(a, b)	**k**	**n**	**R**
(1, 5)	3	4	0.9758
(1, 5)	3	5	0.9584
(1, 5)	3	10	0.9774
(1, 5)	4	5	0.9580
(1, 5)	4	9	0.9598
(1, 5)	5	6	0.9594
(1, 10)	3	4	0.9547
(1, 10)	3	5	0.9577
(1, 10)	4	4	0.9550

We can conclude that ANOVA favors H_0 with small *k* and *n*. Range *r* has no effect on rate.

FUS is performed on all combinations according to RC parameters. This provides true *R*s. However, the R^* is also noted since the average error rate of the approximation can be determined by comparing *R* and R^*. The significance level $\alpha=0.05$.

Table 10. ANOVA results using FUS and simulation

r(a, b)	*k*	*n*	*R*	*R**	*r(a, b)*	*k*	*n*	*R**
(1, 3)	3	30	**0.9523**	**0.9344**	(1, 3)	4	100	**0.9151**
(1, 3)	3	50	**0.9544**	**0.9737**	(1, 3)	7	100	**1.09E-9**
(1, 3)	5	10	**0.9722**	**0.9629**	(1, 3)	10	100	**0**
(1, 3)	5	15	**0.9604**	**0.9899**	(1, 3)	10	500	**0**
(1, 5)	3	10	**0.9774**	**0.9241**	(1, 5)	4	100	**0.5889**
(1, 5)	4	5	**0.9580**	**0.9782**	(1, 5)	5	100	**0.0040**
(1, 5)	4	9	**0.9598**	**0.9537**	(1, 5)	7	100	**7.19E-19**
(1, 10)	3	5	**0.9577**	**0.9437**	(1, 10)	4	10	**0.9717**
(1, 10)	4	5	**0.9565**	**0.9671**	(1, 10)	4	19	**0.9601**

Table 10 has two main parts. In the first part, *R* and R^* can be compared. The results show that approximation of R^* is appropriate. To calculate R^*, 1000 iterations were performed. In the second part, such cases are shown in which *R* cannot be calculated with FUS either. In these cases, only R^* can be calculated in real time.

First, the rates are high in the favor of H_0. But in the case of large enough *k* and *n*, the rates are heavily turn to H_1. If the same experiment is performed with relatively small RC values, getting the result H_0 and the "non-correlated" judgment is very high. Contrarily, the chance of H_1 is increased with large enough values and "correlated" decision will be accepted at high possibility. However, this is a paradox in the view of big data inspired environment. In general, if we have a conclusion with smaller number of data items, then more data should enhance the conclusion. However, our results show that we can get contradictory results comparing cases with few data and big data inspired environment. By increasing *k*, the chance to find statistically equal data rows after each other in *k*-times can be "difficult". In other words, increasing *k*, the chance to choose one data row (the *k*th) is high, which is not equal statistically with the other already chosen data rows (*1*, …, *k-1*). The answer can be proven by *Θ-method*. However, this answer does not affect the conclusion about the contradictory property of ANOVA.

If wide range is chosen, e.g., *r(1, 10)*, and *n* high enough (more than *30*), the candidates cannot be stored in memory because of their count. However, the *r(1, 10)* results which are illustrated in Table 10 suggest that the distortion is independent from parameter *r*.

5.5. Regression results related to Ω (*R*)

In this section, the regression techniques are described in the viewpoint of RC. Our standard analyzing steps are followed and noted in brackets. We summarized the basic mathematical equations in Section 2.1.4. [*Step 1*]. Regressions are in the first class of random correlation [*Step 2*]. If we have a set of data items and more and more regression techniques are used, then the chance of finding a correlation will be increased [*Step 3*]. Therefore, *t* is a critical parameter in this case [*Step 4*]. The *k* can be skipped, since we always have two columns, i.e., *x* and *y* coordinates, and therefore *k* = 2 is constant. We have two parameter *r*: r_1 determines the range of *x* values [$r_1(a_1, b_1)$]), while r_2 stands for range of the *y* values [$r_2(a_2, b_2)$]). The count of the candidates is $r_1 * r_2$ in the first phase. The calculation of Ω is based on *k*, *n* and *r* [*Step 5*]. The earlier described FUS (Section 3.5.2) can be used only partly: the first level of reduction cannot be applied since the order of the coordinates is important. For example, the *x'* = {2, 1, 2} and *y'* = {1, 3, 1} do not give the same r^2 as x = {1, 2, 2} and y = {1, 1, 3}. Therefore, all possibilities need to be directly produced in the first phase. However, the second level of reduction can be used without any modification. We perform regression techniques and seek the best fitting line or curve. If we find high r^2 with either of them, then we increase the count of "correlated" class, i.e. the number of 1's. Oppositely, 0's is increased by one. The acceptance level can be changed, we defined as $r^2 > 0.7$. Also in regression case, the simulation deal with applying conditions. Assumptions are detailed in Section 4.2.4. We can assume that independent property is satisfied. The normality is checked with D'Agostion-Pearson test. The homogeneity of variances is checked with Bartlett-test. Since we create all possible candidates, then X and Y can be seen as populations. The simulation can be run and R can be determined [*Step 6*].

We summarize the results in Table 11 when linear and exponential regressions were performed only, so $t = 2$.

Table 11. Results in the case of t = 2

t = 2	**r_1(1,5);r_2(1,3)**	**r_1(1,10);r_2(1,3)**	**r_1(1,3);r_2(1,5)**	**r_1(1,3);r_2(1,10)**
n = 7	*0.0527*	*0.0474*	*0.1453*	*0.2629*
n = 8	*0.0479*	*0.0375*	*0.1348*	*0.2597*
n = 9	*0.0462*	*0.0280*	*0.1334*	*0.2538*

We increased n from 5 to 10, changed r_1 and r_2, used t = 4. The results are summarized in Table 12.

Table 12. Rates of r^2 *with t = 4*

t = 4	**r_1(1,5);r_2(1,3)**	**r_1(1,10);r_2(1,3)**	**r_1(1,3);r_2(1,5)**	**r_1(1,3);r_2(1,10)**
n = 5	*0.2873*	*0.3071*	*0.3122*	*0.3288*
n = 6	*0.2092*	*0.2161*	*0.2530*	*0.3239*
n = 7	*0.1387*	*0.1379*	*0.2204*	*0.3102*
n = 8	*0.1142*	*0.1027*	*0.1947*	*0.3029*
n = 9	*0.1057*	*0.0796*	*0.1894*	*0.2927*

Further regression analyzing results are summarized in Table 13.

Table 13: Further regression results

$r_1(a_1, b_1)$; $r_2(a_2, b_2)$	**n**	**R**
(1, 3);(1, 3)	*5*	*0.3465*
(1, 3);(1, 3)	*10*	*0.1087*
(1, 5);(1, 5)	*5*	*0.5332*
(1, 5);(1, 5)	*8*	*0.2491*
(1, 5);(1, 5)	*9*	*0.2196*
(1, 4);(1, 6)	*5*	*0.4153*
(1, 4);(1, 6)	*8*	*0.2406*
(1, 4);(1, 6)	*9*	*0.2248*
(1, 6);(1, 4);	*5*	*0.5419*
(1, 6);(1, 4);	*8*	*0.2472*
(1, 6);(1, 4);	*9*	*0.2147*

By comparing Table 11 and 12, it can be concluded that the two extra methods increased the chance of $r^2 > 0.7$, sometimes even doubled it. This means that it is possible to improve the rates by increasing *t*. The case of $r_1(1,3);r_2(1,10)$ is very stable around *0.3*. It is an important result, since the "correlated" and "non-correlated" judgments have the same chance in each cases. On the other side, it is rightful assumption, that the theoretically rate cannot be around *0.5* in regression case, because *0.5* would mean that the "correlated" judgment is not more than a simple coin fifty-fifty rate. To say "correlated", the rate must be stricter. Therefore, the parameters related to rate *0.3* could also be suitable.

If *n* is increased, the *R* decreases. If we assume further decreasing and observe the *30* sample rule of thumb, then the chance to get $r^2 > 0.7$ is small. If we agree, that the connection between two variables must have smaller probability (not fifty-fifty), then regression techniques seems not be too sensitive to RC.

6. General discussion and conclusion

At the beginning of our research, we have realized there are many contradictory result in the fields of science. Our goal was to re-produce and analyze such environment, where contradictory results exist. Therefore, a decision support system was built with universal purposes.

We proposed a new approach to solve differences in (1) data structures, (2) data management, (3) methods and (4) presentation visualizations. We started UDSS from Sprague DSS-generator and we extended this theorem specifying the UDSS framework. We do not just used the DSS-generator term, but identified the tools and criteria, with which universality can be reached. We focused on scientific needs of DSS to solve ill-, semi- and fully-structured problem. UDSS is different from Data Warehouse and ad-hoc DSSs. Since Data Warehouse has its own data representation (star scheme) with operations (data cube), and ad-hoc DSSs focus on one or just few problem, then UDSS is proposed to extend the systems' capabilities. Extended operations, such as *create*, *attach* in data layer, *import new method* in DSS logic etc., can be performed without making any modification in UDSS.

Total universality cannot be reached easily but we believe that an analyst must always seek after universality, and DSSs must be as universal as possible. Based UDSS concept, we implement a DSS used in three very different kind of scientific field: (1) in forestry, (2) in earth science and (3) in economics. The main conclusions of the case studies are:

- **Unified management of different data**. The UDSS solution data layer supports:
    - *Data Extract-Transform-Load functions*. It is also possible to define ETL rules by the researchers.
    - *Universal Database operations*. UDB can store any kind of data and its structure preserves the connections between data and metadata. Data are stored in one robust system, which is an advantage in the view of analysis, because data sets are not scattered and can be handled in one place.
    - *Query operations*. Analysts can query either just a part of one data row or different data rows as well. Analyzing between different scientific field data, i.e. different data rows, can be performed, which lead to interdisciplinary researches.
- **Unified models and method management (Logic)**. Data can be analyzed with different models and methods, i.e., UDSS can solve different kind of problems related to different scientific fields, unlike ad-hoc DSS. The results can be presented according to analysts' needs.
- **Zero modification approach**. There is no need to modify the system if any kind of new data, method or other entities show up during the research processes. Researches can be extended, which can be easily handled in UDSS.

The main results of our research in the field of UDSS are the following.

- We proposed a new flexible universal database structure. Since the classic data warehouse structure does not support later modifications and its designation is mainly related to the business world, the suggested universal database can store any kind of data without modifications. The connections between data items, i.e., metadata and dimensions, are preserved.
- We designed the concept of the universal Data Integration Module which enables to gather information (data and metadata) from scattered data sources.

- We designed a generic Data Manipulation Module, which can store any kind of analyzing methods and a flexible Presentation Layer, which can be extended according to analysts' design.
- Using three scenarios, we showed that the UDSS concept extends ad-hoc DSSs capabilities with unified data management, with analyzing model (filtering, transformation, and analyzing methods) independency and with presentation independency. Demonstrating the capabilities of the UDSS, we showed that a specified forestry decision problem, an automatic and semi-automatic ionogram processing problem and a vendor ranking problem can be solved in the same universal environment.

Having performed analysis about the vendor performance, we also experienced such inconsistency. If the number of possible analysis is high, then results can be occurred just randomly. The level of randomness can be increased by large data volume, i.e., with Big Data. Random Correlation framework is created to analyze how big the random chance in the case of the given research is. The framework contains:

- **Definition**. The precise definition of Random Correlation.
- **Parameters**. Parameter *k* is number of columns, *n* is number of rows, *r* is the range and *t* is number of performed methods.
- **Models and Methods**. There are total possibility space (*R*) and Collision (*C*) methods to analyze research results in the view of RC.
- **Classes**. They are three classes (1) increasing performed methods problem (2) contradictory class and (3) big data inspired class.
- **RC analyzing session**. We defined six steps to analyze RC.

Each research can be analyzed with the created RC analyzing session. We analyzed ANOVA and regression techniques with our Finding Unique Sequences algorithm. ANOVA is sensitive to RC, because in the case of big data inspired environment, H_1 outcome chance increases, while with small data volume, H_0 chance is larger than *90%* on average. Having increased the number of performed regression techniques, we experienced small growth of number of accepted r^2 (> 0.7). However, regression techniques is not so sensitive to RC in the case of big data inspired environment. In the case of large number of data, the possibility of RC is low if we find a correlation between two data rows. However, we cannot calculate all possible candidates because of the large total possibility space.

Our results in the field of Random Correlation.

- We defined Random Correlation. There are methodologically correct results in which the variables are not truly connected. We defined four parameters and three classes, which can describe the environment to analyze random impact level.
- We defined two models: (1) the total possibility calculation and (2) the collision probability. The main though of the first one is to calculate all possible results and analyze the impact of randomness. The second one answers how many more data or analyzing methods must be used to get the desired "random" result.
- We used the RC framework to analyze Analysis of Variance (ANOVA) statistical test to determine how big the random impact can be. We showed that ANOVA is very sensitive for random correlation at high possibility.
- We used the RC framework to analyze regression techniques to determine how big the random impact can be. We showed that regression techniques have a less random impact level.

- We have given two integrated frameworks, i.e., UDSS and RC, to seek real results with the given random impact levels.

It is important to remark that we do not deny that real connections exist. We state that RC factor should be taken into account for research. Therefore, the standard research methodology, such as design research, collecting data, executing analysis, interpreting results, should be extended with the step of *RC analysis*. It means that scientists may calculate RC factor based on data and given method of analysis.

To distinguish real correlations from random correlations, we recommend for all scientists to always calculate how big the possibility of RC can be. Moreover, it is important to analyze whether the results space balanced or not. The RC factor could be attached to every scientific results.

Appendix

Acronyms

DSS *Decision Support System*

UDSS *Universal Decision Support System*

RC *Random Correlation*

MCDA *Multi-Criteria Decision Analysis*

BI *Business Intelligence*

KD *Knowledge Discovery process*

UDB *Universal Database structure*

DIM *Data Integration Module*

DQ *Data Queries*

DMM *Data Manipulation Module*

CM *Core Methods*

MI *Method Integration*

PC *Presentation Core*

PG *Presentation generator*

UIM *User Interface Module*

AS *Analyzing Session*

IIPT *Iterative Ideal Point Threshold*

List of publications

Publications related to this PhD Dissertation

Universal Decision Support System

[B1] Jos M.F. Van Orshoven, Vincent Kint, Anja Wijffels, René Estrella, **Gergely Bencsik,** Pablo Vanegas, Bart Muys, Dirk Cattrysse, Stefaan Dondeyne, "Upgrading Geographic Information Systems to Spatio-Temporal Decision Support Systems", *Mathematical and Computational Forestry & Natural-Resource Sciences*, vol. 3, pp36−41, 2011

[B2] **Gergely Bencsik**, Attila Gludovátz, László Jereb, "Integrált informatikai elemző keretrendszer alkalmazása a magyar felsőoktatásban", Proc. of Informatika a felsőoktatásban konferencia 2011, pp1040−1047, Debrecen, Hungary, 2011 *(In Hungarian)*

[B3] **Gergely Bencsik,** Attila Gludovátz, László Jereb, "Adaptation of analysis framework to industry related economic problems", Proc. of The Impact of Urbanization, Industrial and Agricultural Technologies on the Natural Environment: International Scientific Conference on Sustainable Development and Ecological Footprint, pp1−6, Sopron, Hungary, 2012

[B4] **Gergely Bencsik**, Attila Gludovátz, "Univerzális elemző keretrendszer gazdasági alkalmazásának lehetőségei különös tekintettel a termelővállalatokra", Proc. of International Symposium on Business Information Systems (OGIK 2012), pp36−37, Győr, Hungary, 2012 (In Hungarian)

[B5] **Gergely Bencsik**, Attila Gludovátz, László Jereb, "Decision support framework with wood industrial application", Proc. of 8th International PhD & DLA Symposium, pp154, Pécs, Hungary, 2012

[B6] **Gergely Bencsik,** Attila Gludovátz, "Adaptation of a universal decision support system in forestry", Proc. of Implementation of DSS tools into the forestry practice, *Technical University Zvolen*, pp37−49, 2013

[B7] **Gergely Bencsik,** László Bacsárdi, "Towards to decision support generalization: the Universal Decision Support System Concept", Proc. of IEEE 19th International Conference on Intelligent Engineering Systems (INES), pp277−282, Bratislava, Slovakia 2015

Random Correlation

[B8] **Gergely Bencsik**, László Bacsárdi, "Statisztikai adatok közötti véletlen összefüggések tanulmányozása", Proc. of Informatika a felsőoktatásban konferencia 2014, pp403−412, Debrecen, Hungary, 2014 *(In Hungarian)*

[B9] **Gergely Bencsik,** László Bacsárdi ,"Effects of Random Correlation on ANOVA and Regression", Proc. of the 9th International Conference on Knowledge, Information and Creativity Support Systems, pp396−401, Lemesos, Cyprus, 2014

[B10] **Gergely Bencsik,** László Bacsárdi, “New Methodology to Analyze the Random Impact Level of Mathematically Proved Results”, Proc. of 15th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), pp33−38, Budapest, Hungary, 2014

[B11] **Gergely Bencsik,** László Bacsárdi, “Novel methods for analyzing random effects on ANOVA and regression techniques”, Advances in Intelligent Systems and Computing, *Springer*, pp499-509, 2016, ISSN 2194-5357

Other publications

[B12] Attila Gludovátz, **Gergely Bencsik**, “Egy felsőoktatási képzés Balanced Scorecard alapú mintarendszer működésének demonstrálása”, Proc. of 5th International Symposium on Business Information Systems, pp88, Győr, Hungary, 2007 *(In Hungarian)*

[B13] **Gergely Bencsik**, Attila Gludovátz, László Bacsárdi, “Tudásmenedzsment módszerek faipari alkalmazása”, Proc. of. Inno Lignum konferencia, Sopron, Hungary, 2010 *(In Hungarian)*

[B14] **Gergely Bencsik**, “Analyzing the Adaptability Condition of Decision Support Systems and Data mining in Forestry”, Short Time Scientific Mission report, COST FP0804, Leuven, Belgium, 2010

[B15] Márton Edelényi, **Gergely Bencsik**, Attila Gludovátz**,** “Adaptation possibilities of knowledge management tools in higher education”, Proc. of. Szellemi tőke, mint versenyelőny avagy A tudásmenedzsment szerepe a versenyképességben, pp883−897, Komárno, Slovakia, 2010

[B16] **Gergely Bencsik**, Attila Gludovátz, “Experience with universal data analyses”, Proc. of Forest Management Decision Support Systems (FORSYS) Conference, COSTFP0804, pp53−54, Umea, Sweden, 2013

[B17] **Gergely Bencsik**, “Virtual Environment to simulate business processes related to ERP systems”, Presentation on Microsoft Dynamics Convergence 2014 conference, Barcelona, Spain, 2014

[B18] Gergely Pieler, **Gergely Bencsik**, “Relevant and related area extraction from binary images”, Proc. of 11th International Symposium on Business Information Systems, Budapest, Hungary, 2014

[B19] Péter Kiss, **Gergely Bencsik**, “Virtual Economic Environment”, Proc. of 11th International Symposium on Business Information Systems, Budapest, Hungary, 2014

[B20] Péter Kiss, **Gergely Bencsik**, László Bacsárdi, “From ERP trainings to business: a new approach of simulations in economics”, Proc. of 12th International Symposium on Business Information Systems, pp49−50, Veszprém, Hungary, 2015

[B21] Attila Gludovátz, **Gergely Bencsik**, László Bacsárdi, “IT challenges of a production system”, Proc. of 12th International Symposium on Business Information Systems, pp31, Veszprém, Hungary, 2015

References

[1] J. A. Khan, “Research methodology,” APH Publishing Corporation, New Delphi, pp334, 2008, ISBN: 8131301362

[2] J. Kuada, “Research Methodology: A Project Guide for University Students,” Samfundslitteratur, Frederiksberg, pp139, 2012, ISBN: 8759315547

[3] P. Lake, H. B. Benestad, B. R. Olsen, “Research Methodology in the Medical and Biological Sciences,” Academic Press, London, pp519, 2007, ISBN: 0123738741

[4] A. Mohapatra, P. Mohapatra, “Research methodology,” Partridge Publishing, India, pp124, 2014, ISBN: 148281790X

[5] G. D. Jackson, N.A. Moltschaniwskyj, "Spatial and temporal variation in growth rates and maturity in the Indo-Pacific squid Sepioteuthis lessoniana (Cephalopoda: Loliginidae)," *Marine Biology*, vol. 140, pp747−754, 2002

[6] G. T. Pecl, G. D. Jackson, “The potential impacts of climate change on inshore squid: biology, ecology and fisheries,” *Reviews in Fish Biology and Fisheries*, vol. 18, pp 373−385, 2008

[7] E. S. Zavaleta, B. D. Thomas, N. R. Chiariello, Gregory P. Asner, M. Rebecca Shaw, Christopher B. Field, "Plants reverse warming effect on ecosystem water balance,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp9892–9893, 2003

[8] W. Liu, Z. Zhang, S. Wan, “Predominant role of water in regulating soil and microbial respiration and their responses to climate change in a semiarid grassland,” *Global Change Biology*, vol. 15, pp184–195, 2009

[9] J. A. Church, N. J. White, “A 20th century acceleration in global sea-level rise, “ Geophysical Research Letters, vol. 33, pp1−4, 2006

[10] J.R. Houston, R.G. Dean, “Sea-Level Acceleration Based on U.S. Tide Gauges and Extensions of Previous Global-Gauge Analyses,” *Journal of Coastal Research*, vol. 27, 409−417, 2011

[11] P. K. Aggarwal, R. K. Mall, “Climate Change and Rice Yields in Diverse Agro Environment of India. II. Effect of Uncertainties in Scenarious and Crop Models on Impact assessment,” *Climatic Change*, vol. 52, pp331−343, 2002

[12] J. R. Welch, J. R. Vincent, M. Auffhammer, P. F. Moya, A. Dobermann, D. Dawe, “Rice yields in tropical/subtropical Asia exhibit large but opposing sensitivities to minimum and maximum temperatures,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp14562−14567, 2010

[13] J. A. Church, N. J. White, J. R. Hunter, “Sea-level rise at tropical Pacific and Indian Ocean islands,” *Global and Planetary Change*, vol. 53, pp 155−168, 2006

[14] A. P. Webb, P. S. Kench, “The dynamic response of reef islands to sea-level rise: Evidence from multi-decadal analysis of island change in the Central Pacific,” *Global and Planetary Change*, vol. 72, pp234−246, 2010

[15] R. M. McCaffery, B. A. Maxell, “Decreased winter severity increases viability of a montane frog population,” *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), vol. 107, pp 8644−8649, 2010

[16] S. K. McMenamina, E. A. Hadly, C. K. Wright, “Climatic change and wetland desiccation cause amphibian decline in Yellowstone National Park,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp16988−16993, 2008

[17] G. Yu, H. Shen, J. Liu, “Impacts of climate change on historical locust outbreaks in China,” *Journal of Geophysical Research*, vol. 114, pp1-11, 2009

[18] L. C. Stige, K.-S. Chan, Z. Zhang, D. Frank, N. C. Stenseth, “Thousand-year-long Chinese time series reveals climatic forcing of decadal locust dynamics,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp16188−16193, 2007

[19] K. F. Drinkwater, “The response of Atlantic cod (Gadus morhua) to future climate change,” *ICES Journal of Marine Science*, vol. 62, pp 1327–1337, 2005

[20] R. A. Clark, C. J. Fox, D. Viner, M. Livermore, “North Sea cod and climate change – modelling the effects of temperature on population dynamics,” *Global Change Biology*, vol. 9, pp1669−1680, 2003

[21] L. Hooper, C. Bartlett, G. D. Smith, S. Ebrahim, “Systematic review of long term effects of advice to reduce dietary salt in adults,” *British Medical Journal*, vol. 325, pp628–632, 2002

[22] S. Pljesa, “The impact of Hypertension in Progression of Chronic Renal Failure,” *Bantao Journal*, vol. 1, pp71-75, 2003

[23] Climate Change 2007, Impacts, Adaptation, vulnerability, report, 2007

[24] H. Nkurunziza, J. Pilz, “Impact of increased temperature on malaria transmission in Burundi,” *International Journal of Global Warming*, vol. 3, pp78−87, 2011

[25] P. Martens, R.S. Kovats, S. Nijhof, P. de Vries, M.T.J. Livermore, D.J. Bradley, J. Cox, A.J. McMichael, “Climate change and future populations at risk of malaria,” *Global Environmental Change*, vol. 9, pp89−107, 1999

[26] P. W. Gething, D. L. Smith, A. P. Patil, A. J. Tatem, R. W. Snow, S. I. Hay, “Climate change and the global malaria recession,” *Nature*, vol. 465, pp342−345, 2010

[27] H. J. Fowlera, M. Ekstrom, “Multi-model ensemble estimates of climate change impacts on UK seasonal precipitation extremes,” *International Journal of Climatology*, vol. 29, pp385−416, 2009

[28] E. J. Burke, R. H. J. Perry, S. J. Brown, “An extreme value analysis of UK drought and projections of change in the future,” Journal of Hydrology, vol. 388, pp131−143, 2010

[29] I. M. Held, T. L. Delworth, J. Lu, K. L. Findell, T. R. Knutson, “Simulation of Sahel drought in the 20th and 21st centuries,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp1152–1153, 2006

[30] R. J. Haarsma, F. M. Selten, S. L. Weber, M. Kliphuis, “Sahel rainfall variability and response to greenhouse warming,” *Geophysical Research Letters*, vol. 32, pp1−4, 2005

[31] A. Giannini, “Mechanisms of Climate Change in the Semiarid African Sahel: The Local View,” *Journal of Climate*, vol. 23, pp743−756, 2010

[32] S. M. Crimmins, S. Z. Dobrowski, J. A. Greenberg, J. T. Abatzoglou, A. R. Mynsberge, “Changes in Climatic Water Balance Drive Downhill Shifts in Plant Species’ Optimum Elevations,” *Science*, vol. 331, pp324-327, 2011

[33] J. Grace, F. Berninger, L. Nagy, “Impacts of Climate Change on the Tree Line,” *Annals of Botany*, vol. 90, pp537−544, 2002

[34] T. A. Dueck, R. de Visser, H. Poorter, S. Persijn, A. Gorissen, W. de Visser, A. Schapendonk, J. Verhagen, J. Snel, F. J. M. Harren, A. K. Y. Ngai, F. l. Verstappen, H. Bouwmeester, L. A. C. J. Voesenek, A. van der Werf, “No evidence for substantial aerobic methane emission by terrestrial plants: a ^{13}C-labelling approach,” *New Phytologist*, vol. 175, pp29−35, 2007

[35] F. Keppler, J. T. G. Hamilton, M. Braß, T. Röckmann, “Methane emissions from terrestrial plants under aerobic conditions,” *Nature*, vol. 439, pp187−191, 2006

[36] L. Siliang, L. Ronggao, L. Yang, “Spatial and temporal variation of global LAI during 1981–2006,” *Journal of Geographical Sciences*, vol. 20, pp323−332, 2010

[37] G. P. Asner, J. M. O. Scurlock, J. A. Hicke, “Global synthesis of leaf area index observations: implications for ecological and remote sensing studies,” *Global Ecology and Biogeography*, vol. 12, pp191−205, 2003

[38] C. Jaramillo, D. Ochoa, L. Contreras, M. Pagani, H. Carvajal-Ortiz, L. M. Pratt, S. Krishnan, A. Cardona, M. Romero, L. Quiroz, G. Rodriguez, M. J. Rueda, F. de la Parra, S. Morón, W. Green, G. Bayona, C. Montes, O. Quintero, R. Ramirez, G. Mora, S. Schouten, H. Bermudez, R. Navarrete, F.

Parra, M. Alvarán, J. Osorno, J. L. Crowley, V. Valencia, J. Vervoort, “Effects of Rapid Global Warming at the Paleocene-Eocene Boundary on Neotropical Vegetation,” Science, vol. 330, pp957−961, 2010

[39] L. F. Salazar, C. A. Nobre, M. D. Oyama, “Climate change consequences on the biome distribution in tropical South America,” *Geophysical Research Letters*, vol. 34, pp1−6, 2007

[40] M. Hulme, R. Doherty, T. Ngara, M. New, D. Lister, “African climate change: 1900–2100,” Climate Research, vol. 17, pp145−168, 2001

[41] A. P. Williams, C. Funk, “A westward extension of the warm pool leads to a westward extension of the Walker circulation, drying eastern Africa,” *Climate Dynamics*, vol. 37, pp2417−2435, 2011

[42] M.D. Flannigan, Y. Bergeron2, O. Engelmark, B.M. Wotton, “Future wildfire in circumboreal forests in relation to global warming,” *Journal of Vegetation Science*, vol. 9, pp469−476, 1998

[43] E. S. Kasischke, N. L. Christensen, B. J. Stocks, “Fire, Global Warming, and the Carbon Balance of Boreal Forests,” *Ecological Applications*, vol. 5, pp437−451, 1995

[44] M. E. Visser, A. C. Perdeck, J. H. Van Balen, C. Both, “Climate change leads to decreasing bird migration distances,” *Global Change Biology*, vol. 15, pp1859−1865, 2009

[45] N. Doswald, S. G. Willis, Y. C. Collingham, D. J. Pain, R. E. Green, B. Huntley, “Potential impacts of climatic change on the breeding and non-breeding ranges and migration distance of European Sylvia warblers,” *Journal of Biogeography*, vol. 36, pp1194−1208, 2009

[46] F. Pulido, P. Berthold, “Current selection for lower migratory activity will drive the evolution of residency in a migratory bird population,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp 7341−7346, 2010

[47] A. R. Huete, K. Didan, Y. E. Shimabukuro, P. Ratana, S. R. Saleska, L. R. Hutyra, W. Yang, R. R. Nemani, R. Myneni, “Amazon rainforests green-up with sunlight in dry season,” *Geophysical Research Letters*, vol. 33, pp33−39, 2006

[48] A. Samanta, S. Ganguly, H. Hashimoto, S. Devadiga, E. Vermote, Y. Knyazikhin, R. R. Nemani, R. B. Myneni, “Amazon forests did not green-up during the 2005 drought,” *Geophysical Research Letters*, vol. 37, pp1−5, 2010

[49] M. Savage, R. Burrows, “The Coming Crisis of Empirical Sociology,” *Sociology*, vol. 41, pp885−899, 2007

[50] L. T. Lam, Z.-W. Peng, “Effect of Pathological Use of the Internet on Adolescent Mental Health. Archives of Pediatrics and Adolescent,” *Medicine*, vol. 164, pp164−174, 2010

[51] C.-X. Shen, R.-D. Liu, D. Wang, “Why are children attracted to the Internet? The role of need satisfaction perceived online and perceived in daily real life,” *Computers in Human Behavior*, vol. 29, pp185−192, 2013

[52] B. D. NG, P. Wiemer-Hastings, “Addiction to the Internet and Online Gaming,” *Cyberpsychology & Behavior*, vol. 8, pp110−113, 2005

[53] N. Yee, “Motivations for Play in Online Games,” *Cyberpsychology & Behavior*, vol. 9, pp772−775, 2006

[54] C. E. Doughty, A. Wolf, C. B. Field, “Biophysical feedbacks between the Pleistocene megafauna extinction and climate: The first human-induced global warming?,” *Geophysical Research Letters*, vol. 37, L15703, 2010

[55] F. A. Smith, S. M. Elliott, S. Kathleen Lyons, “Methane emissions from extinct megafauna,” *Nature Geoscience*, vol. 3, pp374−375, 2010

[56] D. T. Shindell, R. L. Miller, G. A. Schmidt, L. Pandolfo, “Simulation of recent northern winter climate trends by greenhouse-gas forcing,” *Nature*, vol. 399, pp452−455, 1999

[57] V. Petoukhov, V. A. Semenov, “A link between reduced Barents-Kara sea ice and cold winter extremes over northern continents,” Journal of Geophysical Research, vol. 115, D21111, 2010

[58] P. Knippertz, U. Ulbrich, P. Speth, “Changing cyclones and surface wind speeds over the North Atlantic and Europe in a transient GHG experiment,” *Climate Research*, vol. 15, pp109−122, 2000

[59] R. Vautard, J. Cattiaux, P. Yiou, J.-N. Thépaut, P. Ciais, “Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness,” Nature Geoscience, vol. 3, pp756−761, 2010

[60] I. Bogardi, I. Matyasovzky, “Estimating daily wind speed under climate change,” Solar Energy, vol. 57, pp239−248, 1996

[61] M. Rebetez, R. Lugon, P.-A. Baeriswyl, “Climatic Change and Debris Flows in High Mountain Regions: The Case Study of the Ritigraben Torrent (Swiss Alps),” *Climatic Change*, vol. 36, pp371−389, 1997

[62] M. Stoffel, M. Beniston, “On the incidence of debris flows from the early Little Ice Age to a future greenhouse climate: A case study from the Swiss Alps,” *Geophysical Research Letters*, vol. 33, L16404, 2006

[63] M. Stoffel, M. Bollschweiler, M. Beniston, “Rainfall characteristics for periglacial debris flows in the Swiss Alps: past incidences–potential future evolutions,” *Climatic Change*, vol. 105, pp263−280, 2011

[64] A. Charland, B. Lebassi, J. Gonzalez, and R. Bornstein, “Cooling of maximum temperatures in coastal California air basins during 1969–2005: monthly and extreme value trends,” 20th Proceedings of Conference on Probability and Statistics in the Atmospheric Sciences, Atlanta, Georgia, 2010

[65] J. A. Johnstonea, T. E. Dawson, “Climatic context and ecological implications of summer fog decline in the coast redwood region,” *Proceedings of the National Academy of Sciences of the United States of America*,” vol. 107, pp4533−4538, 2010

[66] G. H. Miller, A. de Vernal, “Will greenhouse warming lead to Northern Hemisphere ice-sheet growth?,” *Nature*, vol. 355, pp244−246, 1992

[67] T. Boyer, S. Levitus, J. Antonov, R. Locarnini, A. Mishonov, H. Garcia, S. A. Josey, “Changes in freshwater content in the North Atlantic Ocean 1955–2006,” *Geophysical Research Letters*, vol. 34, L16603, 2007

[68] R. Curry, C. Mauritzen, “Dilution of the Northern North Atlantic Ocean in Recent Decades,” *Science*, vol. 308, pp1772−1774

[69] T. R. Knutson, J. J. Sirutis, S. T. Garner, G. A. Vecchi, I. M. Held, “Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions,” *Nature Geoscience*, vol. 1, pp359−364, 2008

[70] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, H. L. Miller (eds.), “IPCC, 2007: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,” Cambridge University Press, Cambridge, United Kingdom and New York, USA, pp996

[71] C. E. Chung, V. Ramanathan, “Weakening of North Indian SST Gradients and the Monsoon Rainfall in India and the Sahel, “ *Journal of Climate*, vol. 19, pp2036−2045, 2006

[72] H. L. Bryden, H. R. Longworth, S. A. Cunningham, “Slowing of the Atlantic meridional overturning circulation at 25°N,” *Nature*, vol. 438, pp655−657, 2005

[73] J. K. Willis, “Can in situ floats and satellite altimeters detect long-term changes in Atlantic Ocean overturning?,” *Geophysical Research Letters*, vol. 37, L06602, 2010

[74] A. W. Burnett, Matthew E. Kirby, Henry T. Mullins, William P. Patterson, “Increasing Great Lake–Effect Snowfall during the Twentieth Century: A Regional Response to Global Warming?,” *Journal of Climate*, vol. 16, pp3535−3542, 2003

[75] L. D. Mortsch and F. H. Quinn, “Climate Change Scenarios for Great Lakes Basin Ecosystem Studies,” *Limnology and Oceanography*, vol. 41, pp903−911, 1996

[76] O. de Viron, V. Dehant, “Effect of global warming on the length-of-day,” *Geophysical Research Letters*, vol. 29, pp1146−1149, 2002

[77] F. W. Landerer, J. H. Jungclaus, J. Marotzke, “Ocean bottom pressure changes lead to a decreasing length-of-day in a warming climate,” *Geophysical Research Letters*, vol. 34, L06307, 2007

[78] E. Martin, G. Giraud, Y. Lejeune, G. Boudart, “Impact of a climate change on avalanche hazard,” *Annals of Glaciology*, vol. 32, pp163−167, 2001

[79] M. Keiler, J. Knight, S. Harrison, “Climate change and geomorphological hazards in the eastern European Alps,” *Philosophical Transactions of The Royal Society*, vol. 368, pp2461−2479, 2010

[80] B. Nosek et al. (Open Science Collaboration),“Estimating the reproducibility of psychological science,” *Science*, vol. 28, aac4716, 2015

[81] G. W. Corder, D. I. Foreman, “Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach,” Wiley, pp294, 2009, ISBN: 047045461X

[82] R. B. D'Agostino , A. Belanger, R. B. D'Agostino, Jr, “A Suggestion for Using Powerful and Informative Tests of Normality,” The American Statistician, vol. 44, pp316−321, 1990

[83] N. J. Salkind, “Encyclopedia of Research Design,” SAGE Publications, pp1776, 2010, ISBN: 1412961270

[84] H. Sahai, M. I. Ageel, “Analysis of variance: fixed, random and mixed models,” Springer, pp742, 2000, ISBN: 0817640126

[85] J. P. Hoffmann, K. Shafer, “Linear Regression Analysis: Applications and Assumptions”, second edition, NASW Press, pp240, 2015, ISBN: 0871014572

[86] G. D. Garson, “Curve Fitting & Nonlinear Regression,” Statistical Associates Publishers, pp58, 2012, ASIN: B00942WWA6

[87] R.C. Raymond, “Use of the Time-sharing Computer in Business Planning and Budgeting,” *Management Science*, vol. 12, pp. 363–381, 1966

[88] E. Turban, “The Use of Mathematical Models in Plant Maintenance Decision Making,” *Management Science*, vol. 13, pp. 342–359, 1967

[89] G. L. Urban, “SPRINTER: A Tool for New Products Decision Makers,” *Industrial Management Review*, vol. 8, pp.43–54, 1967

[90] C. C. Holt, G. P. Huber, “A Computer Aided Approach to Employment Service Placement and Counselling,” *Management Science*, vol. 15, pp.573–595, 1969

[91] M. S. S. Morton, “Computer-Driven Visual Display Devices – Their Impact on the Management Decision-Making Process,” Doctoral Dissertation, Harvard Business School, 1967

[92] T. P. Gerrity Jr., “Design of Man-Machine Decision System: “An Application to Portfolio Management,” *Sloan Management Review*, vol. 12, pp. 59–75, 1971

[93] J. D. C. Little, “Braindaid, an On-Line Marketing Mix Model, Part 2: Implementation, Calibration and Case Study,” *Operation Research*, vol. 23, pp. 656–673, 1975

[94] P. G. W. Keen, M. S. S. Morton, “Decision Support System: An Organizational Perspective, Addison-Wesley, 1978

[95] J. F. Rockart, “Chief Executives Define Their Own Data needs,” *Harvard Business Review*, vol. 67, pp.81–93, 1979

[96] J.-C. Courbon, J. Grajew, J. Tolovi, “Design and Implementation of Interactive Decision Support Systems: An Evolutionary Approach,” Technical Report, Institute d’Administration des Enterprises, 1978

[97] R. H. Bonczek, C. W. Holsapple, A. B. Whinston, “Foundation of Decision Support Systems,” Academic Press, 1981

[98] R. H. Sprague Jr., “A framework for the Development of Decision Support Systems*”, Management Information System Quarterly*, vol. 4, pp.1–26, 1980

[99] R. H. Sprague Jr., E. D. Carlson, “Building Effective Decision Support System,” Prentice-Hall, 1982

[100] P. Gray, “Guide to IFPS,” McGraw-Hill Book Company, 1983

[101] G. DeSanctis, B. R. Gallupe, “A Foundation for the Study of Group Decision Support Systems,” *Management Science*, vol. 33, pp.589–609, 1987

[102] M. P. Amstrong, P. J. Densham, G. Rushton, “Architecture for a Microcomputer-based Spatial Decision Support System,” Proceedings of Second International Symposium on Spatial Data Handling, pp.120–132, 1986

[103] E. Turban, “Decision Support and Expert Systems: Management Support Systems,” Macmillan Pub Co, pp850, 1988, ISBN: 0024216631

[104] R. Sharda, S. Barr, J. McDonnell, “Decision Support System Effectiveness: A Review and an Empirical Test,” *Management Science*, vol. 34, pp.139–159, 1988

[105] B. A. Devlin, P. T. Murphy, “An Architecture for a Business and Information System,” *IBM Systems Journal*, vol. 27,pp.60–80, 1988

[106] M. D. Crossland, B. E. Wynee, W. C. Perkins, “Spatial Decision Support Systems: An overview of technology and a test of efficacy,” *Decision Support Systems*, vol. 14, pp.219–235, 1995

[107] L. Fuld, “A Recipe for Business Intelligence Success," Journal of Business Strategy, vol. 12, pp.12–17, 1991

[108] W. H. Inmon, “Building the Data Warehouse,” QED Technical Publishing Group, 1992

[109] E. F. Codd, S. B. Codd, C. T. Salley, “Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate,” Technical report, 1993

[110] B. Franks, “Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics,” Wiley, pp336, 2012, ISBN: 1118208781

[111] M. F. Shakun, “Airline Buyout: Evolutionary System Design and Problem Restructuring in GroupDecision and Negotiation,” *Management Science*, vol. 37, pp.1291–1303, 1991

[112] A. P. Sage, “Decision Support Engineering,” Wiley, pp360, 1991, ISBN: 047153000X

[113] M. S. Silver, “System that Support Decision Makers: Description and Analysis,” Wiley, pp272, 1991, ISBN: 0471919683

[114] N. J. Car, E. W. Christen, J. W. Hornbuckle, G. Moore, “Towards a new generation of Irrigation Decision Support Systems – Irrigation Informatics”, Proceedings of the International Congress on Modelling and Simulation, 2007

[115] S. P. Van Gosliga , I. Van De Voorde, “Hypothesis Management Framework: a flexible design pattern for belief networks in decision support systems,” Proc. of the 6th Bayesian Modelling Applications Workshop, 2008

[116] D. J. Power, R. Sharda, “Model-driven decision support systems: Concepts and research directions,” *Decision Support Systems*, vol. 43, pp.1044–1061, 2007

[117] O. Cakir, M. S. Canbolat, “A web-based decision support system for multi-criteria inventory classification using fuzzy AHP methodology,” *Expert Systems with Applications*, vol. 35, pp.1367–1378, 2008

[118] J. Mustajoki, R. P. Hämäläinen, “Web-Hipre: Global Decision Support by Value Tree and AHP Analysis,” *INFOR Journal*, vol. 38, pp. 208–220, 2000

[119] J. Lu, G. Zhang, F. Wu, “Web-based Multi-Criteria Group Decision Support System with Linguistic Term Processing Function”, *IEEE Intelligent Informatics Bulletin*, vol. 5, pp35–43, 2005

[120] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, B. Middleton, “Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-based Medicine a Reality,” *Journal of the American Medical Informatics Association*, vol. 10, pp.523–529, 2003

[121] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, P. C. Tang, ”Clinical Decision Support Systems for the Practice of Evidence-based Medicine,” *Journal of the American Medical Informatics Association*, vol. 8, pp.527–534, 2001

[122] E. P Hoffer, M. J Feldman, R. J Kim, K. T Famiglietti, G. O. Barnett, "DXplain: Patterns of Use of a Mature Expert System," Proceedings of AMIA Annual Symposium, pp.321–324, 2005

[123] M. L. Graber, A. Mathew, "Performance of a Web-Based Clinical Diagnosis Support System for Internists," *Journal of General Internal Medicine*, vol. 23, pp.37–40, 2008

[124] A. A. E. Saleh, S. E. Barakat, A. A. E. Awad, "A Fuzzy Decision Support System for Management of Breast Cancer," *International Journal of Advanced Computer Science and Applications*, vol. 2, pp.34–40, 2011

[125] P. Gago, M. F. Santos, Á. Silva, P. Cortez, J. Neves, L. Gomes, "INTCare: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine," *Journal of Decision Systems*, vol. 14, pp.241–259, 2005

[126] D. J Power, "Decision Support Systems: Concepts and Resources for Managers," Praeger, 2002

[127] R. Goodwin, R. Akkiraju, F. Wu, "A Decision-Support System for Quote Generation," Proceedings of American Association for Artificial Intelligence, 2002

[128] Y. Luo, K. Liu, D. N. Davis, "A Multi-Agent Decision Support System for Stock Trading," *IEEE Network*, vol. 16, pp.20–27, 2002

[129] O. Kulak, "A decision support system for fuzzy multi-attribute selection of material handling equipments," *Expert Systems with Applications*, vol. 29, pp310–319, 2005

[130] Q. Wen, Z. Yang, Y. Song, P. Jia, "Automatic stock decision support system based on box theory and SVM algorithm," *Expert Systems with Applications*, vol. 37, pp.1015–1022, 2010

[131] V. Cho, "MISMIS – A comprehensive decision support system for stock market investment," *Knowledge-Based Systems*, vol. 23, pp.626–633, 2010

[132] I. Istudor, L. Duta, "Web-Based Group Decision Support System: an Economic Application," *Informatica Economică*, vol. 14, pp.191–200, 2010

[133] F. Ghasemzadeh, N.P. Archer, "Project portfolio selection through decision support," *Decision Support Systems*, vol. 29, pp.73–88, 2000

[134] D. D. Archabal,, S. H. Mcintyre, S. A. Smith, K. Kalyanam, "A Decision Support System for Vendor Managed Inventory", *Journal of Retailing*, vol. 76, pp.430–454, 2000

[135] G. Wang, S. H. Huangb, J P. Dismukes, "Product-driven supply chain selection using integrated multi-criteria decision-making methodology," *International Journal of Production Economics*, vol. 91, pp.1–15, 2004

[136] S. Biswas, Y. Narahari, "Object oriented modeling and decision support for supply chains," *European Journal of Operational Research*, vol. 153, pp.704–726, 2004

[137] B. Ezzeddine, B. Abdellatif, B. Mounir, "An Agent-based framework for cooperation in Supply Chain," *International Journal of Computer Science*, vol. 9, pp.77–84, 2012

[138] M. Battaglia, P. Sands, D. White, D. Mummery, "CABALA: a linked carbon, water and nitrogen model of forest growth for silvicultural decision support," *Forest Ecology and Management*, vol. 193, pp.251–282, 2004

[139] S. Gilliams, J. Van Orshoven, B. Muys, H. Kros, G. W. Heil, W. Van Deursen, "AFFOREST sDSS: a metamodel based spatial decision support system for afforestation of agricultural land," *New Forests*, vol. 30, pp.33–53, 2005

[140] J. Van Orshoven, V. Kint, A. Wijffels, R. Estrella, G. Bencsik, P. Vanegas, B. Muys, D. Cattrysse, S. Dondeyne, "Upgrading Geographic Information Systems to Spatio-Temporal Decision Support Systems," *Mathematical and Computational Forestry & Natural-Resource Sciences*, vol. 3, pp.36–41, 2011

[141] A. De Meyer, R. Estrella, P. Jacxsens, J. Deckers, A. Van Rompaey, J. Van Orshoven, "A conceptual framework and its software implementation to generate spatial decision support systems for land use planning," *Land Use Policy*, vol. 35, pp.271–282, 2013

[142] F. Dalemans, P. Jacxsens, J. Van Orshoven, V. Kint, P. Moonen, B. Muys, “Assisting Sustainable Forest Management and Forest Policy Planning with the Sim4Tree Decision Support System,” *Forests*, vol. 6, pp.859–878, 2015

[143] V. Redsven, H. Hirvelä, K. Härkönen, O. Salminen, M. Siitonen, ”MELA2012 Reference Manual,” Finnish Forest Research Institute, 2012

[144] M. J. Twery, P. D. Knopp, S. A. Thomasma, H. M. Rauscher, D. E. Nute, W. D. Potter, F. Maier, J. Wang, M. Dass, H. Uchiyama, A. Glende, R. E. Hoffman, “NED-2: A decision support system for integrated forest ecosystem management,” *Computers and Electronics in Agriculture*, vol. 49, pp.24–43, 2005

[145] M. Bonazountas, D. Kallidromitou, P. Kassomenos, N. Passas, “A decision support system for managing forest fire casualties,” *Journal of Environmental Management*, vol. 84, pp412–418, 2007

[146] J.-L. de Kok, S. Kofalk, J. Berlekamp, B. Hahn, H. Wind, “From Design to Application of a Decision-support System for Integrated River-basin Management,” *Water Resour Manage*, vol. 23, pp.1781–1811, 2009

[147] J. D. C. Little, “Models and Managers:The Concept of a Decision Calculus,” *Management Science*, vol. 16, pp466–485, 1970

[148] M. S. S. Morton, “Management Decision System: Computer-Based Support Of Decision Making, Harvard University Press, 1971

[149] R. I. Mann, H. J. Watson, "A contingency model for user involvement in DSS development," *MIS Quarterly*, vol. 8,pp27–38, 1984

[150] H. Bidgoli, “Decision Support Systems: Principles and Practice,” West Publishing Company, pp365, 1989, ISBN: 031446560X

[151] P. N. Finlay, “Introducing decision support systems,” NCC Blackwell, pp240, 1994, ISBN: 185543141

[152] E. Turban, “Decision support and expert systems: management support systems,” 4th edition, Prentice Hall, pp976, 1995, ISBN: 0024217018

[153] E. Turban, R. K. Rainer, R. E. Potter, “Introduction to Information Technology,” 3rd edition, John Wiley & Sons, pp544, 2004, ISBN: 0471347809

[154] R. H. Sprague, H. J. Watson, “Decision Support for Management,” Prentice Hall, pp490, 1995, ISBN: 01339626

[155] V. L. Sauter, “Decision Support Systems: An Applied Managerial Approach,” John Wiley & Sons, pp432, 1996, ISBN: 0471311340

[156] P. G. W. Keen, “Decision support systems: a research perspective,” Proceedings of an International Task Force Meeting, pp23–44, 1980

[157] A. Schroff, “An approach to user oriented decision support systems,” Doctoral Dissertation, Universität Freiburg, 1998

[158] J. J. Donovan, S. E. Madnick, “Institutional and Ad Hoc DSS and Their Effective Use,” *ACM SIGMIS Database* – Proceedings of a conference on Decision Support Systems, vol. 8, pp79–88, 1977

[159] S. L. Alter, “Decision Support Systems: Current Practice and Continuing Challenge,” Addison-Wesley Publishing Co., pp316, 1979, ISBN: 0201001934

[160] R. D. Hackathorn, P. G. W. Keen, “Organizational Strategies for Personal Computing in Decision Support Systems, *MIS Quarterly*, vol. 5, pp21–26, 1981

[161] C. W. Holsapple, A. B. Whinston, “Decision support systems: A knowledge-based approach,” West Publishing Co., pp9660, 1996, ISBN: 0314065105

[162] P. Hättenschwiler, ” Neues anwenderfreundliches Konzept der Entscheidungsunterstützung,” in Absturz im freien Fall – Anlauf zu neuen Höhenflügen: Gutes Entscheiden in Wirtschaft, Politik und Gesellschaft, VDF Hochschulverlag AG., pp189–208, 2001

[163] D. J. Power, “What is DSS?,” *The On-Line Executive Journal for Data-Intensive Decision Support*, vol. 1, 1997, <http://www.tgc.com/dsstar/971021/100015.html>, Last visited: 18.10.2015.

[164] D. J. Power, “Decision Support Systems: A Historical Overview,” in Handbook on Decision Support Systems I, Springer, pp121–140, 2008

[165] N. J. Car, E. Christen, J. Hornbuckle, G. Moore, “Towards a new generation of Irrigation Decision Support Systems – Irrigation Informatics?,” proceedings of International Congress on Modelling and Simulation, pp135–141, 2007

[166] K. P. Tripathi, "Decision support system is a tool for making better decisions in the organization," *Indian Journal of Computer Science and Engineering*, vol. 2, pp112–117, 2011

[167] V. Raheja, S. Mahajan, “Decision Support System, its components, model and types of managerial decisions, “ *International Journal of Innovative Research & Studies*, vol. 2, pp412–418

[168] G. K. Yeo, F. H. Nah, "A Participants' DSS for a Management Game with a DSS Generator," Simulation & Gaming, vol. 23, pp341–353, 1992

[169] C.-S. J.-Dong, “Flexible web-based decision support system generator (FWDSSG) utilising software agents,” IEEE, proceedings of 12th International Workshop on Database and Expert Systems Applications, pp892–897, 2001

[170] P. Keenan, "Using a GIS as a DSS Generator," Geographic Information System, ICFAI University Press, pp97–113, 2007

[171] D.A. Savić, J. Bicik, M.S. Morley, “GANETXL: A DSS Generator for Multiobjective Optimisation of SpreadsheetBased Models,” *Environmental Modelling and Software*, vol. 26, pp551–561, 2011

[172] C. Adamson, “Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance,” Wiley, pp384, 2012, ISBN: 1118429184

[173] J. S. Dodgson, M. Spackman, A. Pearman, L. D. Phillips, “Multi-criteria analysis: a manual,” Communities and Local Government, pp165, 2009, ISBN: 9781409810230

[174] G.-H. Tzeng, C.-W. Lin, S. Opricovic, “Multi-criteria analysis of alternative-fuel buses for public transportation,” *Energy Policy*, vol. 33, pp1373–1383, 2005

[175] S. D. Pohekar, M. Ramachandran, “Application of multi-criteria decision making to sustainable energy planning—A review,” *Renewable and Sustainable Energy Reviews*, vol. 8, pp365–381, 2004

[176] W. Ho, X. Xu, P. K. Dey, “Multi-criteria decision making approaches for supplier evaluation and selection: A literature review, “ *European Journal of Operational Research*, vol. 202, pp16–24, 2010

[177] U. Baizyldayeva, O. Vlasov, A. A. Kuandykov, T. B. Akhmetov, “Multi-Criteria Decision Support Systems. Comparative Analysis,” *Middle-East Journal of Scientific Research*, vol 16, pp1725–1730, 2013

[178] Community of NoSQL, <http://nosql-database.org/>, last seen: 03.03.2016

[179] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, vol. 17, pp37–54, 1996

[180] M. Pezzopane, C. Scott, Ł. Tomasik, I. Krasheninnikov, "Autoscala: an Aid for Different Ionosondes," *Acta Geophysica*, vol. 58, pp513–526, 2009

[181] C. Scotto, M. Pezzopane, “Removing multiple reflections from the F2 layer to improve Autoscala performance,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 70, pp1929–1934, 2008

[182] T. Bullett, A. Malagnini, M. Pezzopane, C. Scotto, “Application of Autoscala to ionograms recorded by the VIPIR ionosonde,” Advances in Space Research, vol. 45, pp1156–1172, 2010

[183] M. Pezzopane, C. Scotto, “Highlighting the F2 trace on an ionogram to improve Autoscala performance,” *Computers & Geosciences*, vol. 36, pp1168–1177, 2010

[184] C. Cesaroni, C. Scotto, A. Ippolito, “An automatic quality factor for Autoscala f_oF2 values,” *Advances in Space Research*, vol. 51, pp2316–2321, 2013

[185] M. Pezzopane, “Interpre: a Windows software for semiautomatic scaling of ionospheric parameters from ionogram,” *Computers & Geosciences*, vol. 30, pp125–130, 2004

[186] World Data Center A for Solar-Terrestr Physicsial, U.R.S.I. Handbook of Ionogram Interpretation and Reduction, second edition, 1972