Computers and Electronics in Agriculture 69 (2009) 59–72

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Analysis of mealybug incidence on the cotton crop using ADSS-OLAP (Online Analytical Processing) tool

Ahsan Abdullah

Dept. of Telecom. Engineering, Foundation University Institute of Engineering & Management Sciences (FUIEMS), Rawalpindi, Pakistan

ARTICLE INFO

Article history:
Received 30 April 2008
Received in revised form 28 May 2009
Accepted 1 July 2009

Keywords:
Decision Support System
Agriculture
Data Warehouse
Mealybug
Pest
Cotton
Pesticide
Data quality
OLAP

ABSTRACT

Traditionally the agriculture expert's knowledge is descriptive and experimental, therefore, it is difficult to describe it mathematically and subsequently build agriculture Decision Support Systems (DSS). Furthermore, the corresponding data may be in such a raw form that it cannot be used in a DSS. The Agriculture Decision Support System (ADSS) is a 26-month project based on the Agro-met data from 2001 to 2006 of Punjab (the bread-basket of Pakistan), its ADSS-OLAP, i.e. Online Analytical Processing tool (www.agroict-olap.org) allows for quick analysis of all possible interesting aggregates of the ADSS data by employing drag-drop and mouse-click and is used in this paper to identify the effective pesticide groups related to the mealybug incidence. Pakistan is the world's fifth-largest producer of cotton, but the emergence of the mealybug as a new cotton pest is likely to reduce the cotton yield by 1.3 million bales. The research work reported in this paper is based on the detailed pest-scouting data of 2300+ farmers of District Multan (cotton hub of Pakistan) for the years 2005 and 2006. This paper will also provide guidelines for the design and development of similar complex systems/tools and discusses the issues of agricultural data-quality management, particularly in the field of insect-pest management.

1. Introduction

Agriculture is the largest sector of Pakistan's economy. Cotton is one of the important fiber crops of Pakistan. Agriculture accounts for 20.9% of the Gross Domestic Product (GDP) and employs 43.4% of the total work force. Pakistan is one of the largest cotton-producing countries of the world. In the fiscal year 2006–2007, the cotton share increased to 7.6% of the total Agriculture GDP. Cotton itself accounts for 8.6% of the value-added product in agriculture and 1.9% in GDP (Anonymous, 2007). The Mealybug is becoming the major impediment to the cotton production-target achievement. The United States Department of Agriculture (USDA) stated that it is improbable for Pakistan to achieve its cotton production target of 14.2 million bales during the fiscal year 2007, due to the mealybug attack on the cotton crop (Razzaq, 2007a,b).

Different Decision Support Systems have been developed to facilitate the agro-industry in the world. A Spatial Decision Support System (SDSS), designated MedCila was developed for controlling the Medfly on citrus grown in Israel. This system provides recommendations and controlled decision-making. These recommendations of MedCila are useful in reducing unnecessary spray application in the absence of a Medfly threat and the number of plots for which the decisions are needed to be made (Cohen et al., 2008). Another DSS was developed using Dutch climatic data for decision support in Northern Europe and the Netherlands. The main focus of this research was to make a decision-support tool for deciding the optimum time for the use of particular climatic control regimes in order to realize optimal gain in sustainability and plant quality. Temperature integration, dynamic humidity control, and negative DIF (difference between average day and average night temperature) regimes were used for the study (Korner and Van Straten, 2008).

1.1. The Agriculture Decision Support System (ADSS)

The Center for Agro-Informatics Research (C@IR) undertook the Agriculture Decision Support System (ADSS) project in July 2006, the project successfully ended in August 2008; the concluding workshop of the project was chaired by the Federal Minister for Agriculture and reported in leading English Daily newspaper www.dawn.com/2008/08/22/nat11.htm This 26-month long, 35-person strong project was supported with approximately US$ 0.5 million by the National ICT R&D Fund. The ADSS project is focused on the Punjab's cotton and rice crops. The main objectives of the ADSS are to use Agro-Meteorological data for Decision Support, Research and Development and Education in an effort to solve agriculture-related problems. The ADSS consists of the following main components:Agriculture Data Warehouse,ADSS-Macro Application (www.agroict.org/adss),ADSS-Micro Application,Online

E-mail address: ahsan1010@yahoo.com.

0168-1699/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2009.07.003

Analytical Processing (OLAP) Tool (www.agroict-olap.org),Data Mining Tool,Yield Loss Forecasting Tool.

ADSS-Macro is a SQL-based application that provides online analysis of 21 key questions to web-based users, which is based on normalized values of data. The tool provides the options to analyze pest-scouting data of past years and help to deduce results that may benefit the agriculture industry. Users are provided with the options to make parameter selections and then view results graphically for each question.

ADSS-Micro is also a SQL-based, in-house application that provides a detailed analysis of data based on the same 21 key questions. ADSS-Micro allows users to perform a detailed analysis and view results based on daily aggregates at the 'Markaz' level (Fig. 1). A Markaz is a central place where farmers from different villages meet, like a central meeting point. ADSS-Micro provides LAN-based access on the C@IR premises. A major enhancement in Micro application over Macro Web application is the level of detail.

ADSS-OLAP (Online Analytical Processing) is a web-based OLAP tool developed by the ADSS team that provides a quick user-driven analysis as opposed to the programmer-driven analysis approach of Macro and Micro. By using the ADSS-OLAP tool, all possible aggregates can be viewed quickly by concurrent users (details in Section 1.3). Based on the detailed end-user feedback (summary in Table 1), and the close involvement of the end-users in ADSS system design (details in Sections 2.8 and 2.10), it can be concluded that the ADSS-OLAP has a user-friendly interface. No installation or configuration is required for analysis through the ADSS-OLAP tool. This tool is easy to use, allowing for the quick transformation of clean data into valuable and useful information.

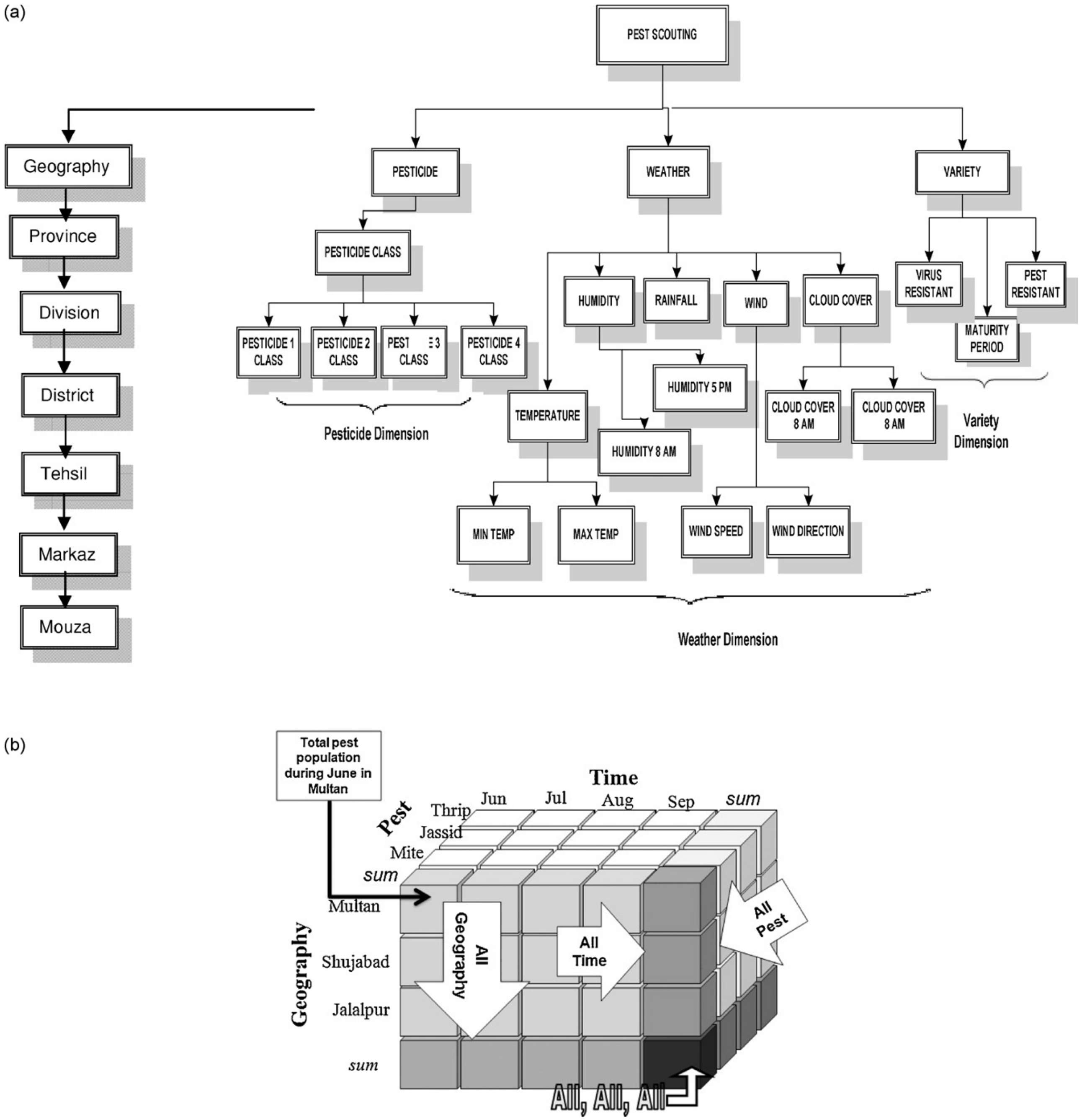


Fig. 1. (a) Geography, pesticide, weather and variety dimensions in the dimensional model of ADSS-OLAP. (b) Different dimensions aggregated as a logical OLAP cube.

Table 1
End of training user feedback of ADSS-OLAP training.

Questions	Results			
	Adequate	Good	Very good	Excellent
How useful do you find ADSS-OLAP	3	8	14	3
How useful do you find ADSS-Micro?	0	11	14	3
How do you find Scouting Entry Tool-cotton (SET-C)?	0	5	14	8
How clear were the Training Presentations?	0	8	14	5
Did the presentations met your professional objectives?	Yes (70%)	No (30%)		
How will you overall rate the training?	0	11	11	5

1.2. Study area

Among the four provinces of Pakistan, Punjab (land of five rivers) is the country's most populated and culture centric province. Nearly 60% of the nation's total population lives in this province. Rice and cotton are the major crops cultivated in the area; these are the cash crops that contribute substantially to the national exchequer with Punjab's share in cotton production estimated to be approximately 81%. Multan, Lodhran and Rahim Yar Khan are some of the major cotton-growing districts of Punjab (Alam, 2000). The Directorate General of Pest Warning (DGPW) Punjab performs pest scouting of different pests for different crops and vegetables across the study area of Punjab. For decades the pest scouting data collected by the personal of DGPW has been used only for comparing Economic Threshold Level (ETL) crossings of this year vs. last year by considering a subset of pests. ETL is the pest population beyond which it is economical to use the pesticides, as the cost of using pesticides after ETL crossing is much lower than the damage done by the pests if pesticides are not used.

Punjab is divided administratively into 8 divisions. There are a total of 34 districts, 124 teshils and scores of markaz within a tehsil. A markaz is further divided into mouzas with dozens of farmers within a mouza.

1.3. Why does the ADSS uses OLAP?

The process of decision-making is flexible; therefore, the corresponding DSS tool should provide the required flexibility and flow. Because when a decision maker is working (mentally), the answer to a question leads to more questions and this goes on. Being an OLAP application, all possible aggregates are pre-computed and stored in a hypercube.

Observe that in the context of OLAP, a query is an aggregate corresponding to different hierarchies and a combination of dimensions. Decision support is ad hoc, i.e. the decision-maker does not have to follow a fixed sequence of steps in order to analyze the information (like traveling by road and only going where the road goes). Instead the decision-maker should be able to analyze the information without any restrictions and as per his choice (like traveling by helicopter and skipping all roads); such access is not supported by a Management Information System (MIS) system. Using the ADSS-OLAP the user only has to navigate by point-click and drag-drop, since the answers are already loaded in the main memory with random or ad hoc access. Note that the ADSS-OLAP application is fully de-normalized, while traditional relational database systems are fully normalized.

Thus the ADSS-OLAP application is unlike a typical MIS application which is meant to run the business, not to optimize it. Recognize that exploring a hypercube is not ADSS but a small part of ADSS (details in Section 2), because unless the data is extracted, transformed, cleansed and loaded into the Data Warehouse, creating the cubes from raw data will not lead to any useful or meaningful results. Actually in our case, cubes cannot even be created from raw data, which is in hard-copy form. While using the ADSS-OLAP, the user (decision maker) is guided by his domain knowledge to identify the information (dimensions, their hierarchy and measures) of interest and quickly look at the results from different perspectives (or dimensions). Since ADSS-OLAP is not an expert system, it is not designed to give any advice to the user (decision-maker), nor does it find hidden patterns as it not a traditional data-mining system. ADSS-OLAP is also not a multidimensional spread-sheet as it allows for the viewing of all possible aggregates with selectively of dimensions, which are not present in a spread-sheet in an interlinked manner. Furthermore, a spread-sheet does not support drill-down or roll-up or pivoting operations or their combination with filtering. Finally, working only with raw data is counterproductive as per Orr's Laws: raw data is likely to be full of errors and data-quality issues will render the data useless. For the analysis of the mealybug incidence using ADSS-OLAP, the user analyses the information from different perspectives (dimensions) using different measures and looks for anomalies or interesting relationships, which are then further explored.

Observe that since the OLAP cube is stored as a multidimensional array (Fig. 1b), therefore, the access time is $O(1)$ just like Random Access Memory (RAM), resulting in very quick data retrieval, giving an impression of On-Line query response. However, this is also one of the limitations of OLAP cube, i.e. requirement of a large main memory for storing all possible aggregates resulting in under 5 s response time. This shortcoming is resolved using other implementation methodologies such as ROLAP (or Relational OLAP), the discussion of which is beyond the scope of this paper. The other limitations of OLAP being long time for cube generation and rebuilding the cube every time fresh data is loaded.

In a nut-shell, OLAP (Online Analytical Processing) is a methodology to provide end users (decision makers) with access to large volumes of data intuitively and quickly to assist with deductions based on investigative reasoning. OLAP systems are expected to:

- Provide ad hoc access.
- Support the complex analysis requirements of decision-makers.
- Analyze the data from a number of different perspectives (business dimensions).
- Support complex analyses against large input (atomic-level) data sets.

Traditionally, the personal of DGPW have used the pest scouting data for only comparing ETL crossings of this year vs. last year by considering a subset of pests. Thus there is a cycle time of one year between data capture and analysis. Traditionally the limited analysis was done at the DGPW Head Office located in Lahore (5 h drive from the cotton fields of Multan). For decades hand-drawn charts were used for this analysis, during the last couple of years the only automation achieved was charts generated using MS Excel.

However, with the ADSS OLAP system, the decision makers are not restricted to analysis limited to a few attributes at a limited level of detail (only ETL crossings) and independent of the Computer Operator to create the charts locally at the Head Office only where (limited) data was available. ADSS-OLAP based analysis supports drill-down or roll-up or pivoting (or their combinations) at different levels of detail. Using ADSS OLAP the decision maker can go much deeper, i.e. up to the farmer level and even at different stages of the life-cycle of a pest. Last but not the least, being web based no configuration or software installation is required at the user end in order to use ADSS-OLAP, and ad hoc analysis can be done from anywhere with Internet access. The ADSS-OLAP analysis was actually performed right next to the cotton fields in Multan using a dial-up Internet connection with a richer set of options and detail, unlike the traditional limited analysis done away from the fields and only by a selected few.

1.4. *Components of ADSS-OLAP*

1.4.1. *Cube*

A 'cube' is the primary structure used to store and retrieve aggregate data from a multi-dimensional database (Fig. 1(b)). A data cube is required in order to work with a MOLAP (Multi-dimensional OLAP) system. The key benefit of MOLAP is that all possible interesting aggregates are calculated and stored in a multi-dimensional array or cube. The cube is then loaded in the main memory and is accessible in $O(1)$ time through a GUI, giving an impression of real-time aggregate based operation. A logical OLAP cube is shown in Fig. 1(b).

1.4.2. *Facts*

'Facts' are measures or additive numerical data being tracked. For example: cultivated area, amount of pesticide sprayed and pest population.

1.4.3. *Dimensions*

A 'dimension' is a business parameter that defines a transaction (or record) and usually consists of hierarchies. For example: ADSS uses different dimensions for analyses of weather, pesticide and geography, as shown in Fig. 1(a).

1.4.4. *Hierarchy*

Generally, each dimension in a Data Warehouse has one or more defined hierarchies. For example: the "Weather" dimension has multiple levels of possible hierarchies such as "Weather > Humidity > Humidity 8 am" and "Weather > Temperature > Minimum Temperature".

1.4.5. *Level*

A 'level' is a step of aggregation within a hierarchy. One can move up and down into the hierarchy of data by using its levels. The first level represents the highest step in the hierarchy; the second level is the child of the first level, and the third level is the child of the second level and so on. For example consider the Geography dimension for the USA. The United States is divided into four regions (Northeast, Midwest, South, West). These regions are further divided into a total of nine divisions which are further divided into primary governmental divisions called States. The primary division of most of the states is a county. Counties have sub-counties called minor civil divisions, further divided all the way up to the block and, finally, the street address. So at the first level is the USA, second level 'region', third level 'division' and so on.

Fig. 1(b) shows some of the dimensions of Fig. 1(a) aggregated as a logical cube which forms the basis of OLAP.

In Fig. 1(b) the shaded regions are at different levels of aggregation, such as light grey corresponding to second level of aggregates or aggregate of aggregate, while the darkest box indicated by ALL, ALL, ALL corresponds to aggregate across three dimensions, i.e. total pest population for all the pests from June to September in all the three tehsils of District Multan. Observe that the arrows indicate the "direction" of aggregation. Thus the massive information is stored at different levels of aggregates, and the user can select the desired level of aggregate/detail (by point-click) for viewing without being overwhelmed by the volume of information available.

The analysis covered in this paper was performed on the data of 2300+ farmers for years 2005–2006. Around 2.8% records of mealybug incidence were observed within the pest-scouting sheets for the year 2005. Therefore, the level of mealybug incidence of that year has not been shown or discussed in detail in the current research work. The natural multi-dimensional hierarchy in the data led to the idea of using the ADSS-OLAP tool for the analysis of mealybug incidence. The organization of the rest of the paper is as follows.

Section 2 describes the details regarding sources of data, cleansing, quality-check processes, steps involved in the ADSS-OLAP to generate graphs and their analysis. Section 3 details the significant findings reported through the use of the ADSS-OLAP tool. Section 4 presents a cause-effect analysis of the results presented in the previous section. Section 5 summarizes the main outcome of the work.

2. Materials and methods

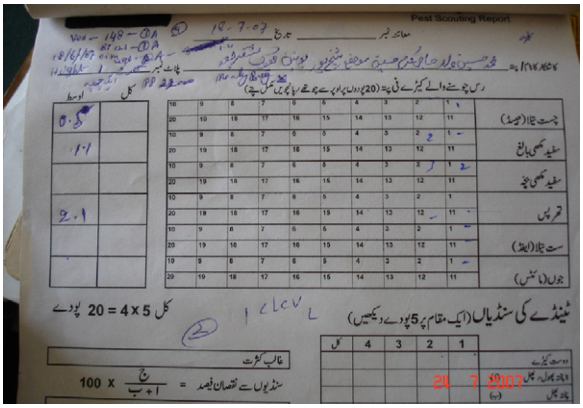
The major steps of the indigenous ADSS workflow are briefly discussed in the following sections.

2.1. *Data collection*

Pest-scouting data is collected seasonally in Punjab by the Directorate General of Pest Warning and Quality Control of Pesticides (DGPWQCP) for the districts under consideration. The scouts from the DGPWQCP weekly sample 50 points in each tehsil of the cotton-growing districts of Punjab. Presently, 60 tehsils are sampled on a weekly basis, resulting in the sampling of 3000 points within Punjab, with approximately 150 such points situated in District Multan. This sampling has been in practice for more than two decades and it is estimated that the pest-scouting data for cotton has thus accumulated to be around 1.5 million records, and growing.

While interviewing the local farmers during their field visits, the scouting teams record the corresponding data of individual farmers on hand-held writing pads as shown in Fig. 2.

Subsequently, the data of individual farmers is transferred on pest scouting sheets as per the scouting area and the scouting time period. A typical hand-filled pest scouting data sheet is shown in Fig. 3. From these sheets ETL (Economic Threshold Level) crossings (represented by *) are noted at tehsil levels, these serve the purpose of informing policy makers as to where hot spots for a particular pest are developing this year with reference to last year. (ETL is the pest population beyond which it is economical to use pesticides.)



Pest Scouting Report

Var – 148 – ①A 18-7-07 تاریخ ______ معائنہ نمبر

رس چوسنے والے کیڑے فی پتہ (20 پودوں پر اوپر سے چوتھے یا پانچویں مکمل پتے)

اوسط	کل											
0.5		10	9	8	7	6	5	4	3	2	1	چست تیلا (جیسڈ)
		20	19	18	17	16	15	14	13	12	11	
1.1		10	9	8	7	6	5	4	3	2	1	سفید مکھی بالغ
		20	19	18	17	16	15	14	13	12	11	
		10	9	8	7	6	5	4	3	2	1	سفید مکھی بچہ
		20	19	18	17	16	15	14	13	12	11	
2.1		10	9	8	7	6	5	4	3	2	1	تھرپس
		20	19	18	17	16	15	14	13	12	11	
		10	9	8	7	6	5	4	3	2	1	سست تیلا (ایفڈ)
		20	19	18	17	16	15	14	13	12	11	
		10	9	8	7	6	5	4	3	2	1	جوں (مائٹس)
		20	19	18	17	16	15	14	13	12	11	

کل 5 × 4 = 20 پودے

ٹینڈے کی سنڈیاں (ایک مقام پر 5 پودے دیکھیں)

1 CLCV L

	1	2	3	4	کل
دوست کیڑے					
(ا)					
(ب)					

غالب کثرت

سنڈیوں سے نقصان فیصد = $\frac{ج}{ا + ب}$ × 100

24 7 2007

Fig. 2. Pest-scouting report of an individual farmer.

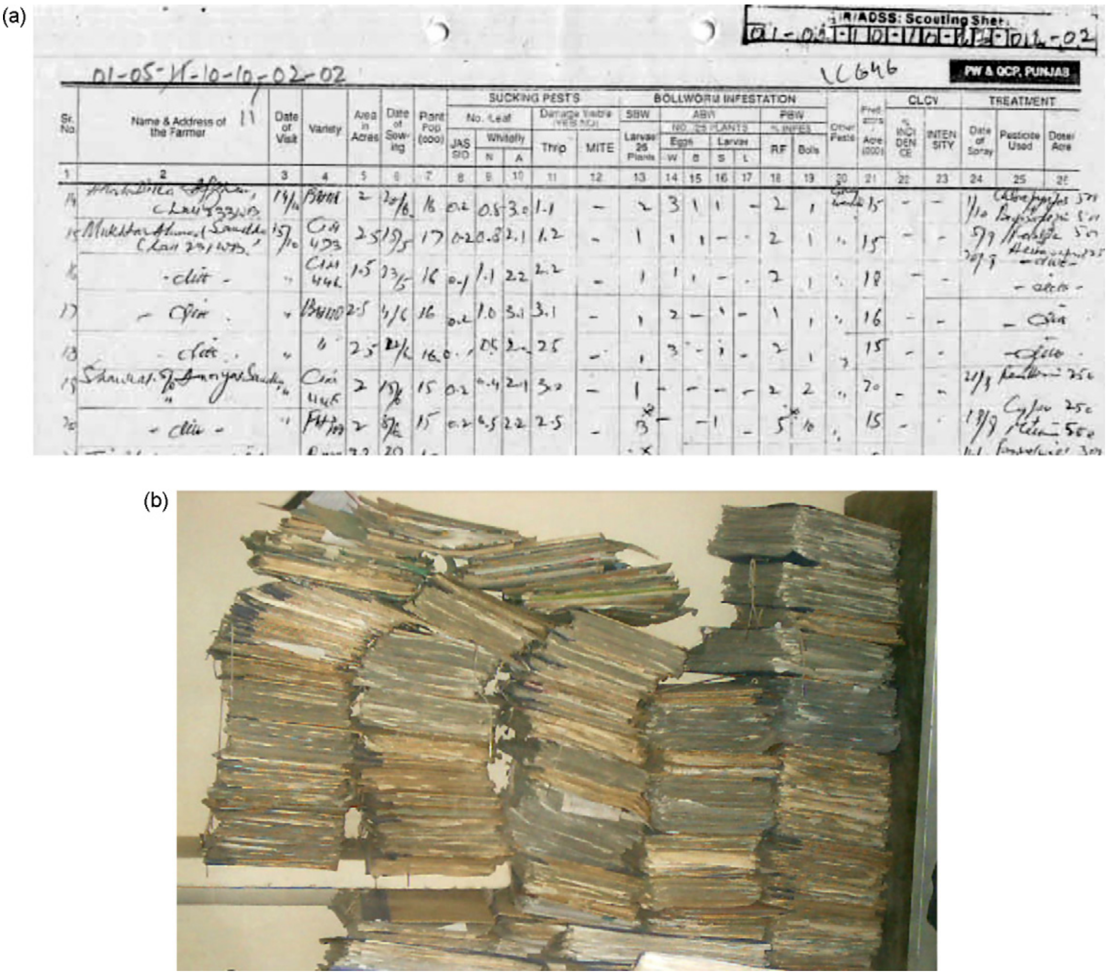


Fig. 3. (a) A typical hand-filled pest-scouting data sheet for several farmers. (b) Some of the pest-scouting sheets stored for decades.

2.2. Data validation

The pest-scouting data collected by the AO/FO (Agriculture or Field Officers) of DG Pest warning is randomly checked by the AD (Assistant Director), which is further randomly checked by the DD (Deputy Director) at the district level. Only after this authentication, the reports/results are issued under the signature of the Director General (DG) Pest Warning. To further validate the data-gathering process, the ADSS team conducted a three-day tour of four cotton tehsils of Punjab in July 2007. Three separate meetings with different officers of the DGPWQCP were conducted at Lahore, Multan and Shujabad in which data-quality procedures were discussed. The C@IR team noted relevant information after interviewing randomly selected farmers from the scouting sheets of four tehsils; Multan, Shujabad, Jalalpur Pirwala and Rahim Yar Khan. The team recorded information regarding farmer demographics including the farmer's name, his level of education, land area owned, mobility, access to TV, radio, computer, Internet facility, etc as well as crop variety sown. A total of 36 randomly selected farmers were visited by the team. Table 2 shows the visited farmers' average age, area owned and other attributes.

2.3. Data acquisition

Manually filled or typed pest-scouting sheets are filed without any order and stored in rooms in different cities by the staff of DG Pest Warning. It is therefore a challenging task, if not impossible, to locate a particular sheet corresponding to a particular geographic location and time for analysis of a particular pest, thus preventing any analysis to date that is similar to what is presented in this paper using ADSS-OLAP. Fig. 3(b) shows how the pest-scouting sheets are stored for decades.

To acquire these sheets, teams of Data Collection agents of the ADSS project routinely traveled for 5–10 h (one-way) visited the data sources in different cities and sat for hours in dusty, spider-infested rooms separating the cotton and rice scouting sheets from other sheets. The sheets were then taken to the hotel and numbered, so that they could be returned in the same order. Subsequently the

Table 2
Farmer statistics collected by the ADSS team in July 2007.

	Multan	Shujabad	Bahawalnagar	Rahim Yar Khan
No. of farmers	13	12	5	6
Avg age (years)	44.30	43.83	47	46.5
Avg area (acres)	36.96	12.95	20.6	28.91
Most cultivated variety	CIM 496	CIM 496	CIM 496, CIM 473	CIM 496
Mode of mobility	Motorcycle	Motorcycle	Car	Car

sheets were photo-copied and ordered as per initial numbering, packs of three hundred sheets were made with labels and those packs placed in card-board boxes and labeled. These boxes were then sent on a daily basis to the project via courier.

2.4. Folder preparation

After the boxes of sheets were received at the ADSS project premises, they were recorded, opened and the sheets were punched using heavy-duty paper punchers by another team of data collection agents. The contents of each pack were then transferred to a box folder and labeled. A methodology was developed for assigning unique numbering, i.e. C@IR ID to each pest-scouting sheet based on time, geography and the crop sown. A group of data collection agents first numbered each pest-scouting sheet by pencil which, after verification by random checking was stamped and the C@IR ID was permanently written by ink. The sheets were then sorted and separated according to the C@IR ID and organized into separate, labeled folders and stored in the data room or data bank. Preparing a single folder typically takes 22 man-hours or about three working days. The ADSS has 150 such folders.

2.5. Folder maintenance

The folders are maintained in the data bank, for easy retrieval the folders are labeled based on the combination of time and geography. The data bank is accessible to relevant personnel who need the pest-scouting sheets for scanning and data-entry purposes. A desired sheet can be easily located in the corresponding folder as it has separators corresponding to the scouting dates.

2.6. Scanning of sheets

Every pest-scouting sheet is scanned on a high speed scanner with document feeder and saved in the computer in *pdf* format. A pre-determined scanning plan is followed to ensure that all scanned sheets are stored and are just as accessible as the original pest-scouting sheets organized in the physical folders. The main purpose of scanning the data sheets is to protect them from wear and tear during data entry, by providing the printout of the scanned sheet to the Data Entry Operator (DEO) rather than the actual data sheet. Due to the quality of the scanned sheets, and some of them being hand-filled, OCR could not be used for electronically extracting the data, thus requiring manual data entry as described in the next section. Note that scanning also ensures the safety of the data sheets from the hazards of moisture, termites and fire.

2.7. Data entry

A team of seven Data Entry Operators (DEOs) transforms manually filled or typed data sheets into digitized form. First, the sheets are categorized according to their quality which is basically based on their legibility; first, the latest sheets are selected, followed by older-dated sheets. The DEO Lead first assigns the sheets to the DEO, and subsequently the DEOs are given the printouts of the sheets. The sheet assignment is done electronically and the progress of the DEOs is also monitored electronically through a web-based data-entry application called as SET-C (Scouting data-Entry Tool for Cotton). Two different teams of DEOs enter the same data. After data entry is done by each team, the entered data is compared programmatically for each row and column and an error report is generated. The error report identifies those values (in the two sets of data entered) which are different, i.e. one of the two is incorrect. Subsequently, the data sheet corresponding to these values is retrieved, checked against the entered value and the data is corrected. The flow-chart covering most of the process described is shown in Fig. 4(a).

The latest statistics of the acquired, scanned and entered data-sheets are shown in Fig. 4(b) (*y*-axis represents the number of sheets).

2.8. Data profiling and cleansing

Digitized data is then cleansed automatically, by using business rules through the in-house prototype Data Profiling Tool. Note that here the Business Rules have nothing to do with trade or commerce; instead in our context this means the rules covering the domain/process being considered. Fig. 5(a) shows the screen-shot of the Business Rule module of the Data Profiler. Some of the typical business rules (out of the 22 used) are listed as follows:

Business Rule 1: Visit Date: May–30th October. Must be later/greater than the sowing date and the spray date.
Business Rule 2: Sowing Dates: April–July. Must be earlier than visit and spray date.
Business Rule 3: Plant Population: Cannot be zero. Values can be from 12,000 to 75,000.
Business Rule 4: Plant Height: As high as 60 inches/180 cm.
Business Rule 5: Jassid Population: 0.1–3 (ETL = 1).
Business Rule 6: Whitefly Nymph: 1–10 (ETL = 5).

The Data Profiling Tool can also perform data-quality assessment using metrics such as (i) completeness, (ii) free of error, (iii) consistency and (iv) believability. Data profiling can also perform summarized profiling, detailed profiling and the finding of duplicate records. Fig. 5(b) shows the screen-shot of the tool for detailed profiling, which shows the presence of non-standard values, distribution of values, etc. for the cotton variety attribute. Since the said attribute is textual, the ‘Mean’, ‘Average’, etc. are not meaningful, and are computed for numeric attributes, such as area and pest population. In Fig. 5(b), *Variety* is on the *x*-axis and *Frequency* is on the *y*-axis.

The data quality and cleansing work is done under the supervision of the Data Quality Manager who subsequently generates a data-quality report providing an over-all assessment of the data quality for each attribute. This report is generated for each District and for each year covered under ADSS. The summary of the data-quality report of District Multan for 2006 is shown in Table 3.

The data entry errors are quantified in terms of *believability*. To quantify this metric, the errors found and corrected by each DEO are noted. The maximum of the data entry errors is used to quantify this dimension. Note that the error reports of all DEOs were compiled on weekly basis and publicly displayed, consequently the overall data entry errors (corresponding to believability) went down over the period of time, and stayed low, i.e. under 5%. In case of Multan data for 2006, believability was 95%. These and other metrics (Pipino et al., 2002) such as (i) Free-of-Error, (ii) Completeness and (iii) Consistency are also computed, and used to create the Data Quality Assessment Graph for each attribute (Fig. 5d).

The difference between the metrics of *Free-of-Error* and *Believability* is explained with an example. Some variety names were not written clearly or not written at all in the data sheets (Figs. 2 and 3a) or only numbers were written (corresponding to the variety names) thus resulting in erroneous data. The error occurs, because for some of the cotton varieties the prefix in the variety name is different, while the remaining part of the name is same. For example, 115 might represent variety CIM 115 as well as BH 115; therefore, just based on the number the prefix variety name cannot be ascertained. This type of problem is noted by entering the asterisk sign (*) as a

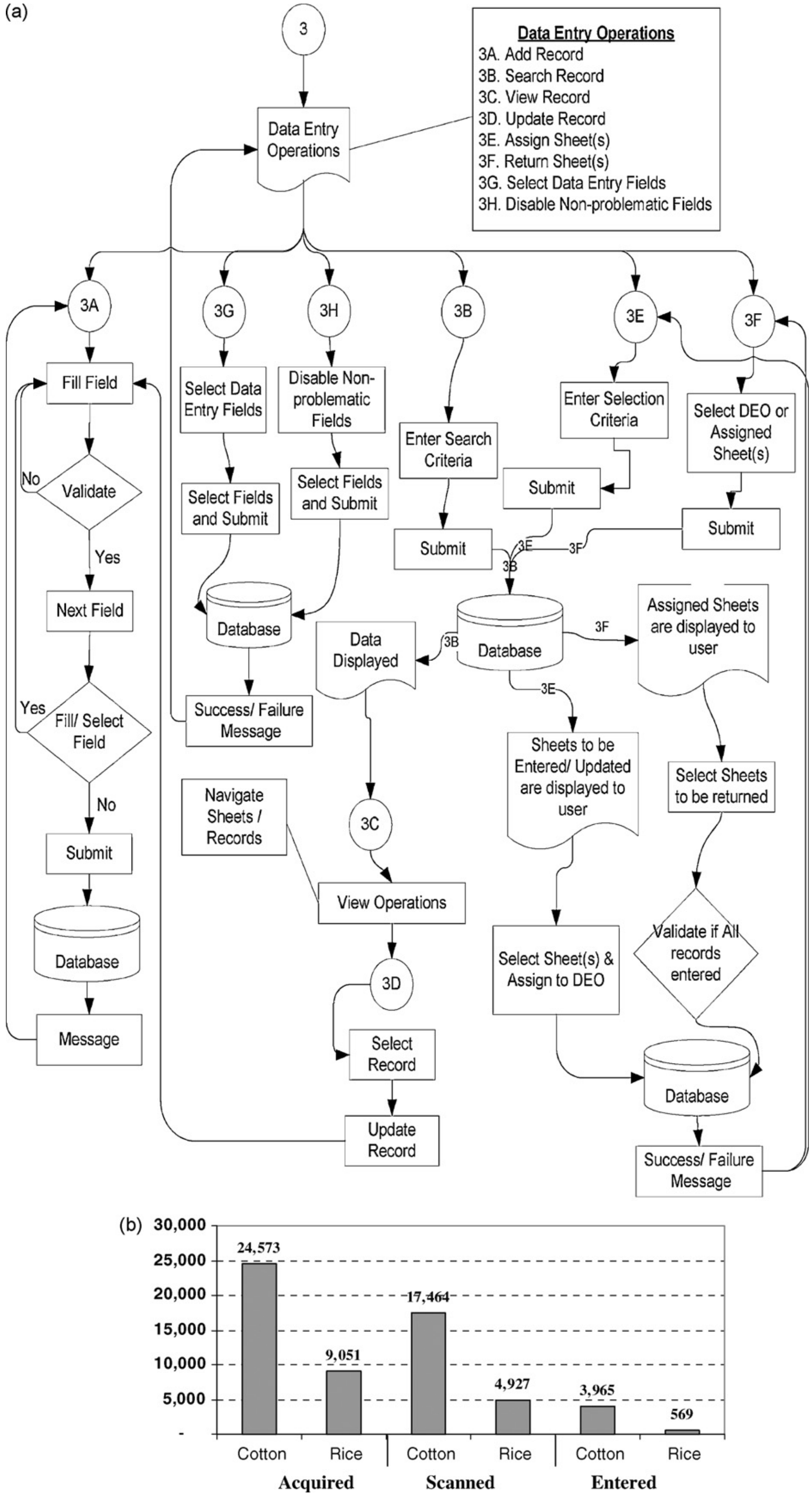
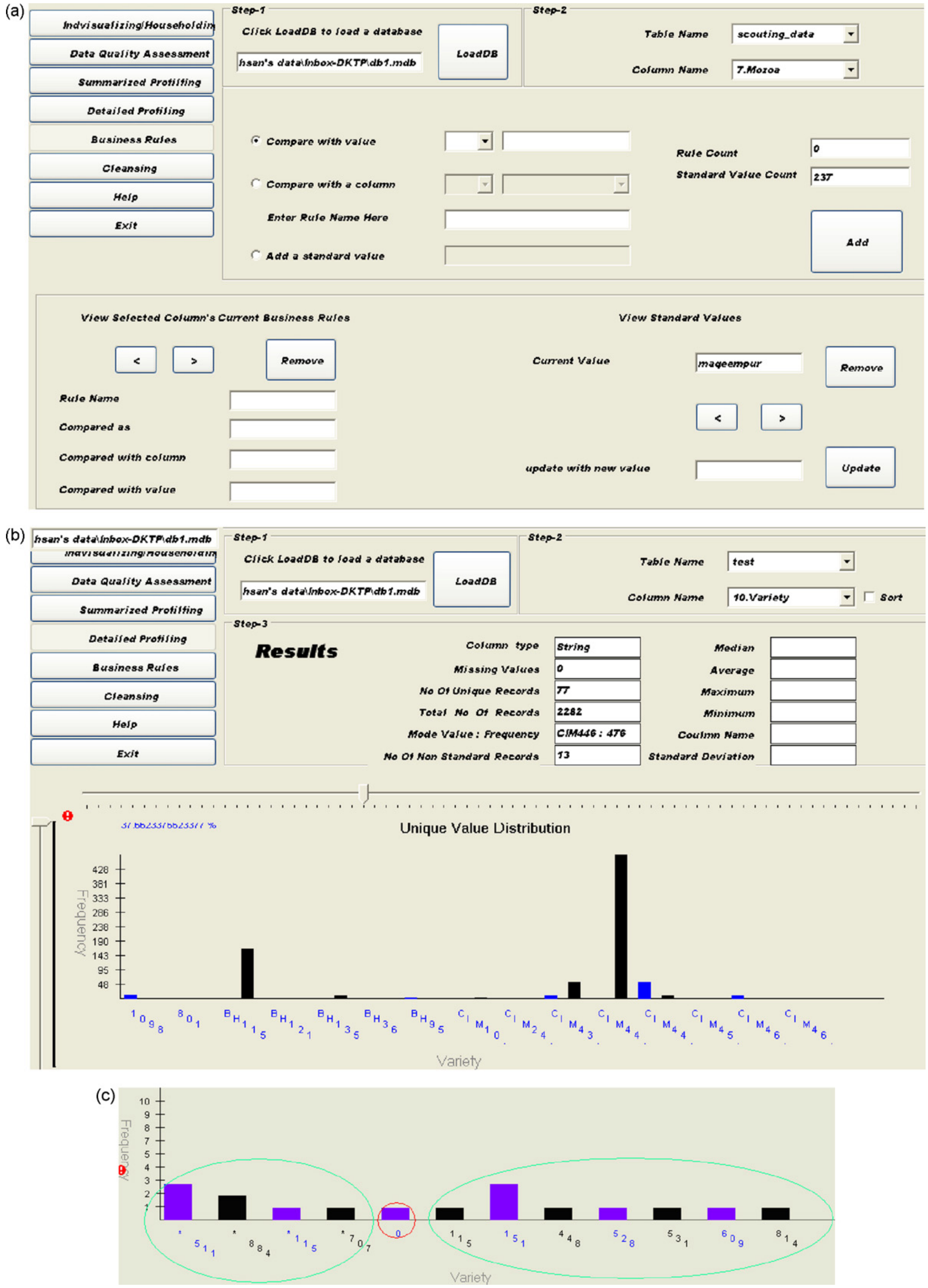


Fig. 4. (a) The process of ADSS data entry. (b) Statistics of acquired, scanned and entered data sheets.



(a)

hsan's data\Inbox-DKTP\db1.mdb

Table Name: scouting_data

Column Name: 7.Mozoa

● Compare with value

○ Compare with a column

Enter Rule Name Here

Rule Count: 0

Standard Value Count: 237

Rule Name:

Compared as:

Compared with column:

Compared with value:

Current Value: mageempur

update with new value:

(b)

hsan's data\Inbox-DKTP\db1.mdb

hsan's data\Inbox-DKTP\db1.mdb

Table Name: test

Column Name: 10.Variety

□ Sort

Column type	String	Median	
Missing Values	0	Average	
No Of Unique Records	77	Maximum	
Total No Of Records	2282	Minimum	
Mode Value : Frequency	CIM446 : 476	Coulmn Name	
No Of Non Standard Records	13	Standard Deviation	

Fig. 5. (a) Data Profiling Tool: module for business rule processing. (b) Data Profiling Tool: module for detailed profiling. (c) Unique value distribution of erroneous variety names. (d) Data Quality Assessment Graph of the variety attribute—Multan 2006.

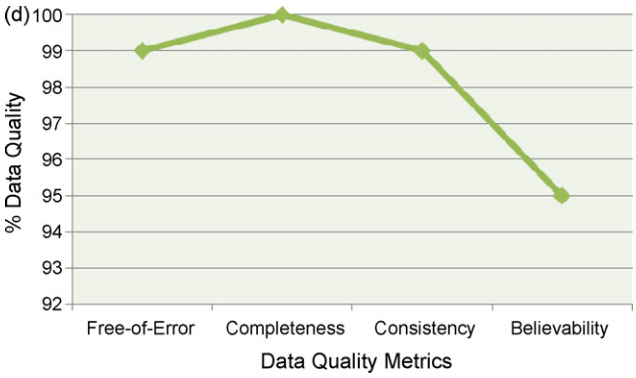


Fig. 5. (*Continued*).

prefix at the time of data entry. The distribution of such erroneous variety names is shown in Fig. 5(c), here *Variety* is on *x*-axis and *Frequency* is on *y*-axis.

The 16 records shown in Fig. 5(c) are out of a total of 2500+ records. One record shows a variety name as ‘zero’ which is highly improbable. The entry 115 or 151 or 448, etc. should have been entered with * prefix.

The overall data-quality assessment graph for the *variety* attribute for District Multan for the year 2006 is shown in Fig. 5(d). Note that graphs similar to Fig. 5c and d are made for 22 attributes of the Pest Scouting Sheet, and form part of the complete Data Quality Assessment Report. Note that the Data Quality Assessment Report (which consists of 40–50 pages) is used to determine if the data quality is as per acceptable standards, and if acceptable, then its impact on the quality of the results after data loading/cube-generation and subsequent analysis can be ascertained.

Table 3
Summary of Data Quality Assessment Report of Multan for 2006.

Attribute	Peaks	No. of errors/business rule violations	Null values
Visit date	1st June to 30th October	7	None
Variety	CIM variety	16	None
Area	1–6, 10 and 25 acres	1	None
Sowing dates	1st of May to 15th June	7	None
Plant population	20–28 thousand plants	34	34
Plant height	2–36 cm	11	2024
Jassid (per leaf)	0.1–0.8	0	332
Whitefly N (per leaf)	0.2–2.8	0	535
Whitefly A (per leaf)	0.9–3.7	1	129
Thrips	2–7	0	324
Mite	1–5	0	2592
SBWLarvae (per 25 plants)	1–4	0	1486
ABWEggsW	1–2	0	2498
ABWEggsB	1	0	2498
ABWLarvaeS	1	0	2131
ABWLarvaeL	1	0	2362
PBWRF	1	0	2142
PBWBolls	1–2	0	2424
Predators (per acre)	2–5	0	341
CLCVIncidence	1–6%	0	513
Spray date	15th July to 15th September.	86	1507
Tehsil	Shujabad Tehsil	5	0

2.9. *Meetings with stake-holders*

For the ADSS requirements gathering, a workshop titled “First Workshop on Agro-Informatics” was held after the launching of the project in September 2006, in which participants from seven public sector organizations participated. This workshop led to the signing of an MoU with the Pakistan Agriculture Research Council (PARC). Based on the participant’s feedback of the first workshop, ADSS system specifications were developed, which were sent to all the stake-holders. The replies received were compiled, finalized, and based on them the results were obtained using the 2001–2002 pest-scouting data. The second workshop on Agro-Informatics was held in April 2007 in which the results of using data for the years 2001–2002 were discussed, and the ADSS system requirements were approved and signed. This workshop was attended by participants from eight public sector organizations. Both workshops were chaired by the Chairman PARC. Other than these two workshops, internal and external meetings are periodically arranged with stakeholders regarding ADSS system requirements gathering and the analysis of results and findings.

2.10. *Transformation to the ADSS*

The end-user requirements are transferred to the ADSS applications, i.e. ADSS-Macro, ADSS-Micro, ADSS-OLAP, the Data Mining Tool and the Yield Loss Forecasting Tool.

2.10.1. *Information and data modeling*

A Data Warehouse is needed to maintain huge volumes of data for quick access; thus, the efficiency and response time are crucial, particularly while retrieving data from the warehouse. In order to improve the efficiency and response time of the Data Warehouse, the preferred structure is the Star Schema. Star Schema is a database structure in which data is maintained in a single fact table located at the center of the schema with additional dimension data stored in dimensional tables, with all hierarchies collapsed.

Each dimension table is usually joined to the fact table by a key column. In a Star Schema structure, the data is stored in de-normalized form to improve performance by reducing the number of joins and run-time aggregates. The new dimensions added in the model (Abdullah and Hussain, 2006) are shown in Fig. 1(a). The Star Schema database structure for the ADSS dimensional model is given in Fig. 6. Attributes and their respective data types have been differentiated using different signs, i.e. Integer type attributes are shown with ‘#’, real with ‘$’, text with ‘+’ and date/time with ‘∼’. The dimensional model (Fig. 1a) and the star-schema (Fig. 6) are the core ADSS system designs around which the entire system is developed. To the best of our knowledge, other than the work of the author, no

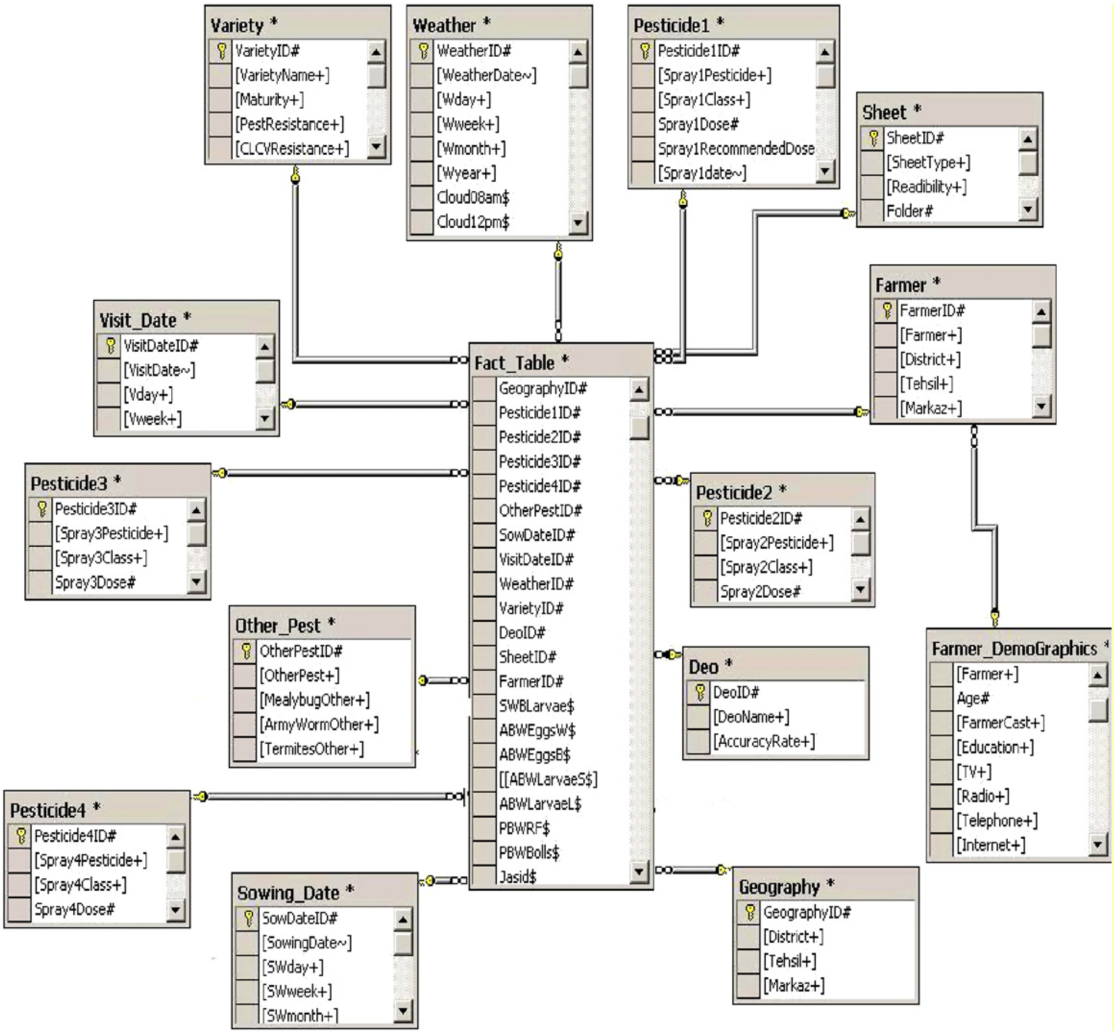


Fig. 6. ADSS-Micro star-schema.

such designs are in use that reflects the pest-pesticide relationship in the Agriculture sector.

The ADSS-OLAP application consists of 13 classes; it is not possible to legibly show the complete class diagram, therefore, a partial class diagram is shown in Fig. 7. The main class is OLAP.DAT.MANAGER which represents the OLAP Manager, and is the key middle component of ADSS-OLAP. The OLAP Manager is the middleware component that acts as a link between the user interface controls and the underlying data. Using this control, the developer can add all available functionality provided in ADSS-OLAP.

2.10.2. Data analysis using ADSS-OLAP

Fig. 8(a) shows the view of the ADSS-OLAP's main screen (www.agroict-olap.org) which provides a selection of different measures versus different dimensions. The tool also facilitates its users by allowing them to select the parameter (or measure) and then drill-down and roll-up into the dimensions at different levels thereby allowing for a quick viewing of different trends and patterns in the data. Drilling down and rolling-up into the data is accessed by simply clicking on the positive (+) and negative (−) symbols present in the ADSS-OLAP tool. For example, within the 'time' hierarchy, one may drill into the data to view various parameters such as year, month, week and/or day and also pivot interchange the axis and the corresponding grouping.

Note that ADSS-OLAP is currently not open source, though interested users can provide their data with related dimensional information for creating the corresponding cubes at our end, and subsequently loading them on the ADSS web (www.agroict-olap.org) with controlled access.

Fig. 9 shows the ADSS-OLAP grid, which is an important part of the ADSS-OLAP tool and is a valuable comparison feature. The grid represents sums and averages, which are used to enhance the analysis of data. The ADSS-OLAP grid generates its results based on the dimensions selected and the results can be copy-pasted into other applications such as MS Excel. Graphical representations of different relationships are viewed and analyzed by the agriculture specialist to extract valuable findings.

2.11. Training

Day-long National Training Workshops were held for imparting hands-on training of ADSS tools to the end users. A total of four training workshops were held, with the workshop on ADSS-OLAP attracting 25+ participants from all over the country, including four PhD scholars. The participants were given a 128-page training folder of lab exercises and other related material and underwent supervised training. The participants were also given an end-of-training questionnaire; the summary of the replies received is given in Table 1.

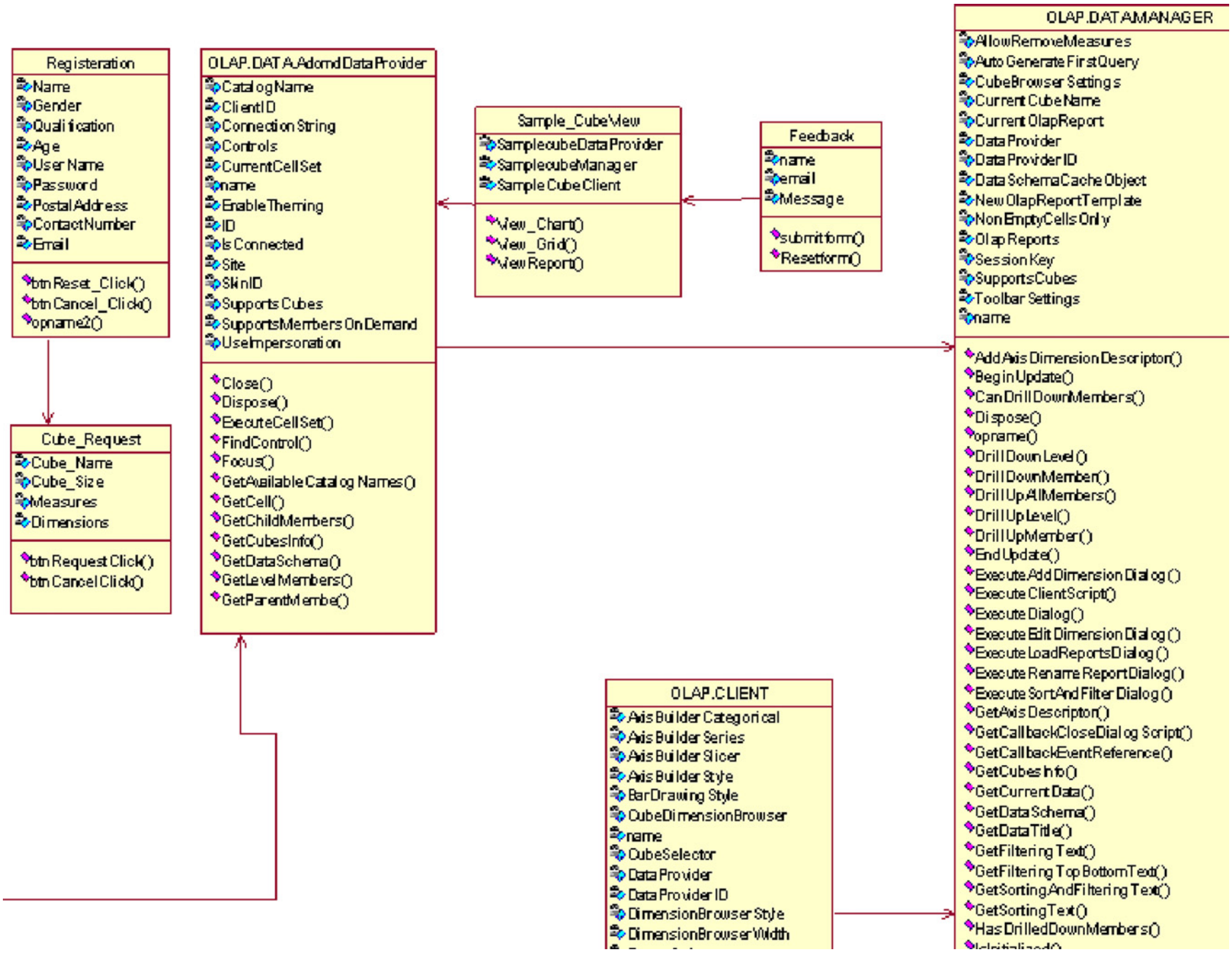


Fig. 7. Partial class diagram of ADSS-OLAP.

3. Results

We proceed by considering the fundamental factors one-by-one which could have contributed to the variation of the mealybug population during the 2006 cotton season. In the greenhouse the mealybug has shown a life cycle of 30–70 days, but usually it is 60–90 days, with some species also producing live young instead of eggs (Pasian and Behnke, 1997). During the winter the mealybug hibernates, therefore, the mealybug incidence seen in May (Fig. 10) probably corresponds to the first generation which left hibernation sometime in March. Thus the dip seen in the population of the mealybug in September is unlikely to be due to termination of the life cycle of the mealybug, as it continues until the onset of winter.

Now let's consider Predators. Although predator population is also recorded on the pest-scouting sheets, these are recorded without identification. The mealybug incidence was recorded as "other pest" because the mealybug was hardly found before 2006, hence it is unlikely that the predators were recorded specific to the mealybug; this was also confirmed from the pest-scouting staff and observed during the field visits. Now consider weather-comparing the mealybug population using the Time dimension and different weather parameters as a measure (Cloud Cover, Min. Temp, Max. Temp, Sunshine hours, % Humidity) did not show any drastic weather changes during September 2006 when the mealybug population dipped. Thus the only other remaining significant parameter that could affect the mealybug incidence is pesticides.

Note that, at any given time, there may be different pests present in the cotton fields in their different stages of life cycle or development. Hence it cannot be assumed that the pesticides sprayed were only targeting the mealybug. Therefore, those pesticide spray results are used when the mealybug incidence was recorded; however, other pests may also have been present during that time. To statistically ensure the targeted pests, the measure of pesticide spray was compared with measures of other sucking pest population with respect to the Time Dimension using the ADSS-OLAP. Subsequently the scatter plot of populations of Thrips, Jassid, mealybug, etc. were made with respect to the Pesticide sprayed and the R^2 value noted. The R^2 value for mealybug was only found to be significant, i.e. 0.56.

4. Discussion

The ADSS project had the services of a full-time Agriculture Specialist with several years of field experience with cotton farmers and pesticide dealers. For further expert opinion, the project also had the service of a cotton consultant. The analysis of the results discussed in this section is based on the feedback/consultation of these experts. Furthermore, the results were cross-checked with the field

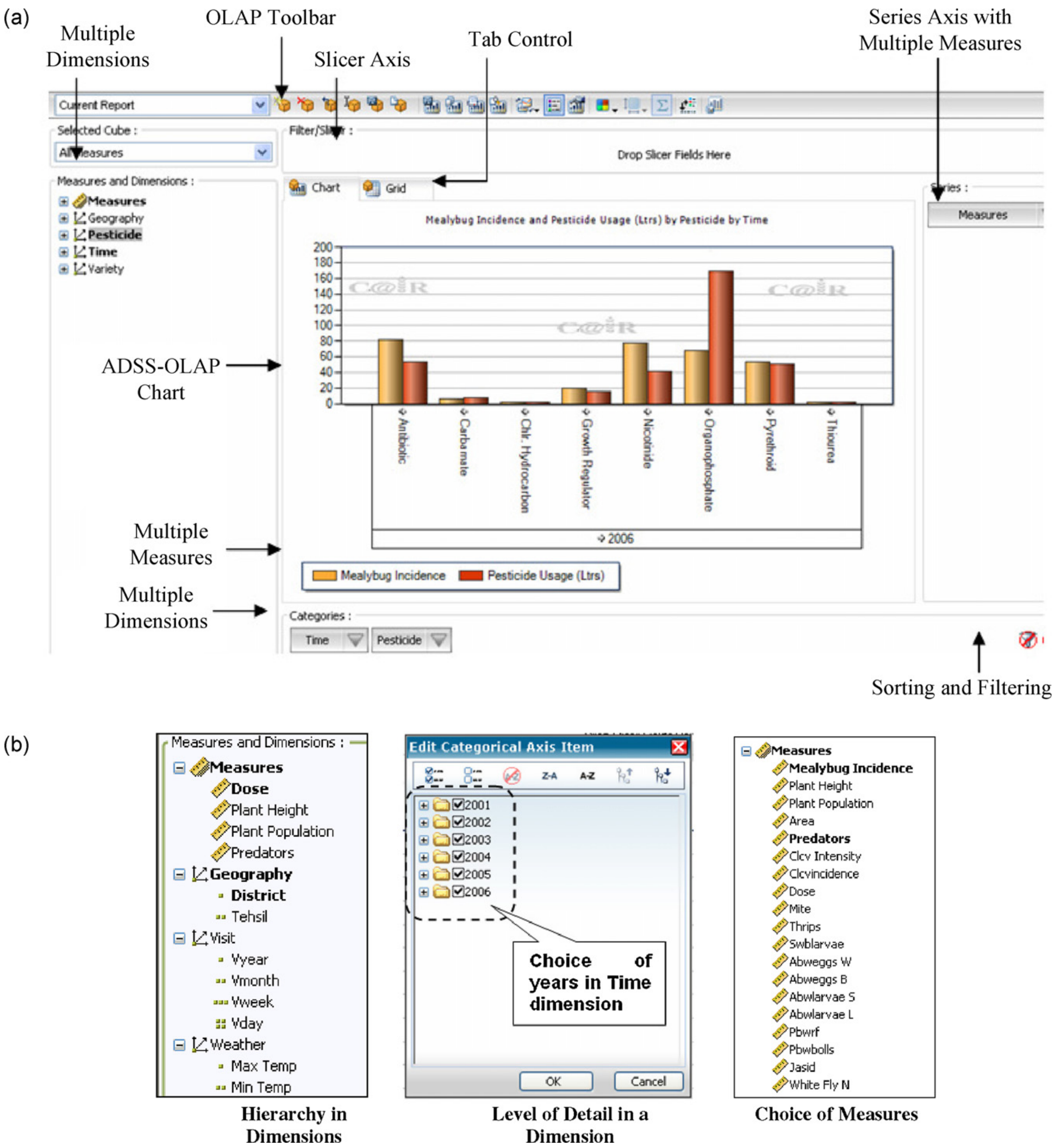
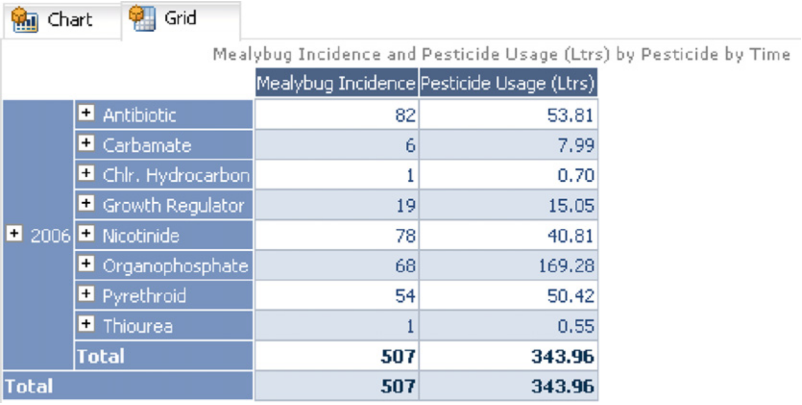


Fig. 8. (a) View of the ADSS-OLAP's main screen. (b) View of the ADSS-OLAP's selection menus.

staff of the Directorate General of Pest Warning from whom the Pest Scouting Data was obtained.

From Fig. 10 it can be observed that there is a strong relationship between pesticides sprayed and mealybug incidence, more precisely the correlation was found to be 0.75. It can also be observed that from May to August the mealybug incidence was almost increasing exponentially, but with the drastic increase of spraying during August, the mealybug incidence was contained. Actually from July to August the pesticide usage increased by roughly 600%, while the mealybug incidence only increased by 32%. For the period of August–September the pesticide used decreased by 66% and at the same time, the mealybug incidence decreased by more than 80%. From September to October the Pesticide spray increased by 300% and at the same time the mealybug population increased by 360%. Now the interesting question is, which group of pesticides contributed to the control of the mealybug and how?

To answer this question using the ADSS-OLAP the Pesticide dimension was used along with the Time dimension in two parts, i.e. first only for the months of August and September and then for the months of September and October. The corresponding result with percentage reduction (change) in pesticides sprayed during August–September is shown in Fig. 10. Note that the actual amount



Mealybug Incidence and Pesticide Usage (Ltrs) by Pesticide by Time

		Mealybug Incidence	Pesticide Usage (Ltrs)
+ 2006	+ Antibiotic	82	53.81
	+ Carbamate	6	7.99
	+ Chlr. Hydrocarbon	1	0.70
	+ Growth Regulator	19	15.05
	+ Nicotinide	78	40.81
	+ Organophosphate	68	169.28
	+ Pyrethroid	54	50.42
	+ Thiourea	1	0.55
	Total	**507**	**343.96**
Total		**507**	**343.96**

Fig. 9. Grid View of the graph shown in Fig. 8(a).

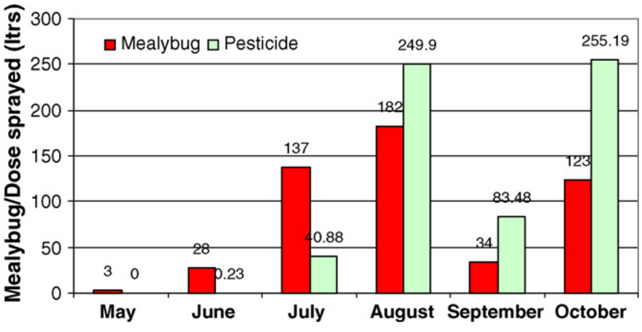


Fig. 10. Comparison of mealybug and pesticide usage.

sprayed is not a good measure for a comparison of pesticides used since pesticides such as Organophosphate (OP) have an intrinsic higher-recommended dose, with the dose being relative to the LD_{50} value of the pesticide (LD_{50} being a measure of toxicity of the pesticide, the lesser the LD_{50} value the more toxic is the pesticide).

From Fig. 11 it can be observed that, although for the major pesticide groups (treated as Dimension), the amount of pesticide used has reduced (in line with the overall reduction), but surprisingly, the usage of the Carbamate group of pesticide has increased by 150%. Carbamate was not found to be sprayed during July.

Similar to Fig. 11 an analysis was done using ADSS-OLAP for the September–October period, during which there was an overall trend of increase of pesticide usage as shown in Fig. 12.

From Fig. 12 it can be seen that except for Insect Growth Regulator (IGR) for which the usage has actually gone down significantly, the usage of all pesticide groups increased during the September–October period, with the largest increase again being

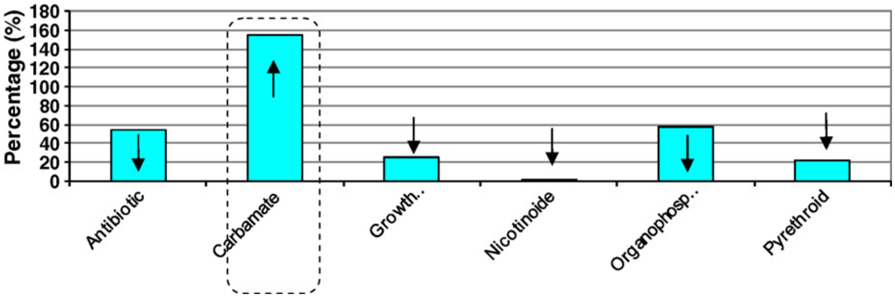


Fig. 11. Percentage “Reduction” in pesticide usage during August–September.

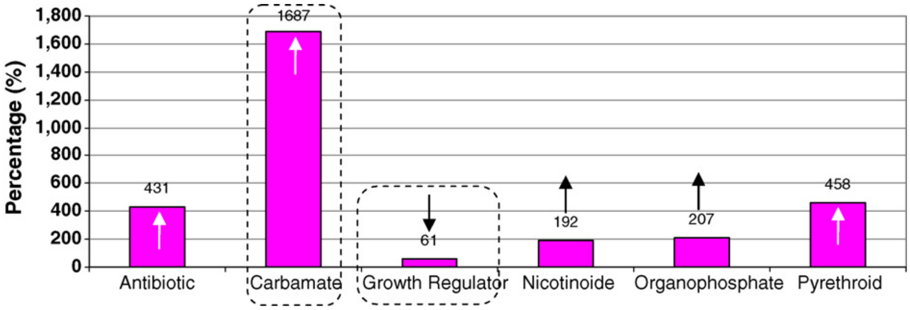


Fig. 12. Percentage increase in pesticide usage during September–October.

for the Carbamate group. Further drill-down of the Carbamate group showed that, among the three brands used, only one was consistently being used from August to October.

5. Conclusions

OLAP has traditionally been used for the analysis of business data captured electronically and mostly used by computer literate professionals. ADSS-OLAP is an effort to move the tool from the desktop to the Internet and to use it for agriculture, specifically pest-scouting data captured manually and used by field-oriented professionals. Based on the feedback of end-users who are accustomed to using spread-sheets, web-based ADSS-OLAP can be used for performing multi-dimensional analysis of agricultural data at different levels of details, supporting ad hoc analysis.

Significant resources and effort went into data-quality management because of the complexity of the data (22 business rules required for cleansing) and availability of manually recorded data in hard copy. The complexity and scope of data cleansing and data entry time could be significantly reduced, if data logging at the field-level is done using PDAs or rugged mobile phones with a back-end database accessible via Internet. This could be viable because of up to 50% tele-density in Pakistan, and farmers found to be using mobile phones even in those areas that are not accessible by roads.

Based on the ADSS-OLAP analysis, it can be concluded that the Carbamate group of systemic pesticides is effective against controlling the mealybug incidence; this was also confirmed by the staff of the Directorate General of Pest Warning.

The cotton farmers need to be educated about the ill effects of using highly toxic and hazardous pesticides, and the benefits of using less toxic, slow acting yet effective pesticides.

Acknowledgements

This work was made possible by the funding of the National ICT R&D Fund and due to the support of Dr. Ijaz Perviaz, the guidance of Dr. Amir Muhammed and the hard work of the entire ADSS team, especially Muhammad Usman, Shahid Saleem-Agriculture Specialist, Zahid Majeed and Shela Rafique. The author is obliged to his late father Ismail Siddiqi for his support and advice from day-one to day-end. The author is also indebted to the anonymous referees whose guidance has significantly improved the quality and presentation of the paper.

References

Abdullah, A., Hussain, A., 2006. Data mining a new pilot agriculture extension data warehouse. Journal of Research and Practice in Information Technology 38 (3).

Alam, 2000. Cotton: An Important Cash Crop. Alam, Tando Jam.

Anonymous, 2007. Pakistan Economic Survey (2006–07). Government of Pakistan, Ministry of Food & Agriculture, Islamabad, p. 16.

Cohen, Y., Cohen, A., Cohen, Hetzroni, A., Alchanatis, V., Broday, D., Gazit, Y., Timar, D., 2008. Spatial decision support system for Medfly control in citrus. Journal of Computers and Electronics in Agriculture 62 (2), 107–117.

Korner, O., Van Straten, G., 2008. Decision support for dynamic greenhouse climate control strategies. Published by Computers and Electronics in Agriculture 60 (1), 18–30.

Pasian, C., Behnke, C., 1997. Mealybugs., http://floriculture.osu.edu.

Pipino, L., Yang, L., Wang, R., 2002. Data quality assessment. Communications of the ACM 45 (April (4ve)), 211–218.

Razzaq, A., 2007a. 0.2 million cotton bales destroyed by mealybug, Islamabad. Business Recorder (29th August), www.brecorder.com.

Razzaq, A., 2007b. Mealybug serious threat to crops, vegetables, Islamabad. Daily Business Recorder (19th May), www.brecorder.com.