Data Mining a New Pilot Agriculture Extension Data Warehouse

Ahsan Abdullah

Center for Agro-Informatics Research, FAST National University of Computer & Emerging Sciences, Islamabad, Pakistan. www.nu.edu.pk/cair
Department of Computing Sciences & Mathematics, University of Stirling, Stirling, Scotland, UK
Email: ahsan@nu.edu.pk

Amir Hussain

Department of Computing Sciences & Mathematics, University of Stirling, Stirling, Scotland, UK
Email: ahu@cs.stir.ac.uk

Pakistan is the world's fifth largest cotton producer. To monitor cotton growth, different government departments and agencies in Pakistan have been recording pest scouting, agriculture and metrological data for decades. Coarse estimates of just the cotton pest scouting data recorded stands at around 1.5 million records, and growing. The primary agro-met data recorded has never been digitized, integrated or standardized to give a complete picture, and hence cannot support decision making. In this paper, a complete life-cycle implementation of a novel Pilot Agriculture Extension Data Warehouse is discussed, followed by data analysis by querying the Data Warehouse and some interesting findings through data mining using an indigenous technique based on the crossing minimization paradigm. Actual cotton pest scouting data of 1,500+ farmers for years 2001 and 2002 for the Multan district was processed and used in the pilot project.

Key Words: Data Warehouse, Data Mining, Decision Support System, Agriculture, Cotton, Pest Scouting.

ACM Classification: H.4.2

1. INTRODUCTION

Every year different government departments are tasked to monitor dynamic agricultural situations all around the Punjab – the breadbasket of Pakistan. As a result, thousands of digital and non digital data files are generated from hundreds of pest-scouting and yield surveys, metrological data recordings and other such undertakings. The data collected, due to its multivariate nature and disparate origins, is hard to integrate and does not provide a complete picture. Thus the lack of data integration (and standardization) contributes to an under-utilization of valuable and expensive historical data, and inevitably results in a limited capability to provide decision support and analysis.

Traditionally, Agriculture decision making in Pakistan is not data driven, but usually based on expert judgment. The data under consideration, such as pest scouting, pesticide usage and metrological recordings contains huge analytical potential in two major respects. Firstly, short term decision making and day to day tactical handling of issues related to pest management and secondly,

Manuscript received: 14 April 2005
Communicating Editor: John Yearwood

long-term decision making, strategic planning and policy making where one needs to observe the complete history of events, with issues related to nature and amount of pesticides etc.

In this paper, the implementation of a New Pilot Agriculture Extension Data Warehouse (PAE DWH) is discussed, which is followed by knowledge discovery in PAE DWH using an indigenous data mining technique based on the crossing minimization paradigm. Based on literature review, other than the work of the primary author, no such work was found to have been undertaken in the agriculture sector of Pakistan (Ahmed and Nagy, 2001) and in the region. Data warehouses are quite popular in telecommunications, the travel industry, government etc. but an application in agriculture extension is a novel idea. The strength of this novel idea is demonstrated through a pilot implementation and discussion of interesting findings using real data.

This paper is organized as follows: In Section 2 a brief background is given about cotton grown in Pakistan, followed by a short introduction of entomology terminology used in the paper. In Section 3 the need for a pest scouting data warehouse is discussed. In Section 4 comparison is made with other related work. Section 5 covers in detail the development of PAE DWH following the 12-step approach of Atre (2005). In Section 6 query based results of exploring the PAE DWH are discussed. In Section 7 discoveries made by data mining the PAE DWH are discussed. Finally in Section 8 conclusions are given.

2. BACKGROUND

The aim of this section is to give a brief background about cotton grown in Pakistan, followed by a short introduction of entomology terminology used throughout the paper. Pakistan is the fifth largest cotton-growing country of the world. Almost 70% of world cotton is produced in China (Mainland), India, Pakistan, USA and Uzbekistan (Chaudry, 2000). As textile exports comprise more than 60% of Pakistan's total exports, the success or failure of the cotton crop has a direct bearing on the economy. Cotton production is the inherent comparative advantage of the textile sector of Pakistan (Cotton and Ginning, 2003); with total textile industry exports amounting to US$ 7 billion (Economic Survey, 2001/2003) and 68% share in export earnings.

Punjab is the breadbasket of Pakistan, and is administratively divided into eight divisions, including the Multan division. The Multan division is further divided into six districts, including district Multan. District Multan has three *Tehsils*. Within each *Tehsil* are central points or *Markaz*. This work is centred around the Mutlan district (Figure 4). The area under study is shown in Figure 1.

In the context of this paper, a pest is an insect that eats or damages the crop. Since the cotton crop is being considered, so some of the pests considered are jassid, thrips, SBW (spotted ball worm) etc. A predator is an insect that eats the pests. Some of the cotton pest predators are ladybug beetles, spiders, ants, assassin bug etc. A sample of cotton virus, pest and predator are shown in Figure 2. Other than pests, the cotton crop is also effected by viruses, the predominant one being CLCV (cotton leaf curl virus). In this paper, **field** would mean the cultivated land, with certain area and ownership.

ETL A: Economic Threshold Level in agriculture extension is that pest population beyond which the benefit of spraying outweighs its cost. It is highly infeasible and expensive to eradicate all pests, therefore, pest control measures are employed, when pest populations cross a certain threshold. This threshold varies from pest to pest, and from crop to crop. Figure 3 shows the ETL A by a dotted line, and undesired pest populations by "humps" above the said line.

3. NEED FOR AN AGRICULTURE EXTENSION DATA WAREHOUSE (AE DWH)

The aim of this section is to briefly discuss the need for establishing an Agriculture Extension Data Warehouse. Motivation of this work arises from the need to have a better insight into the dynamics

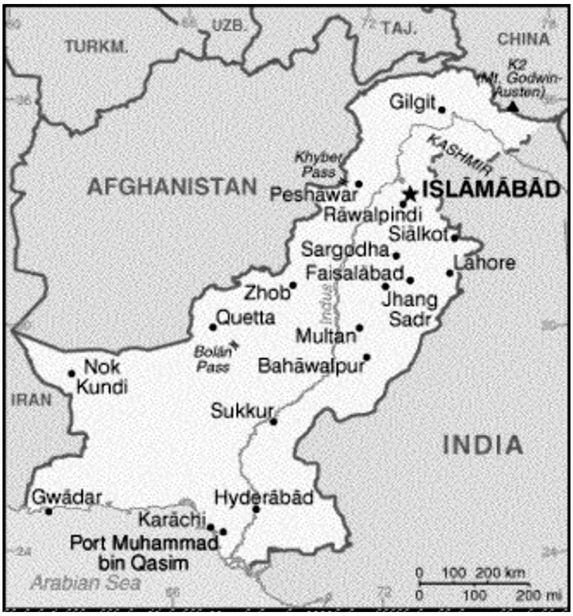
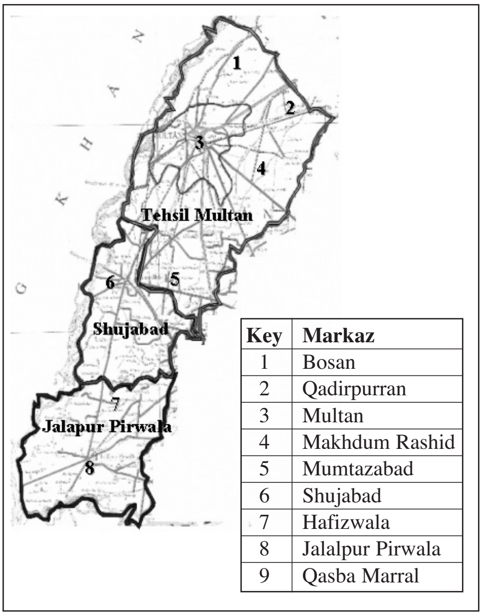
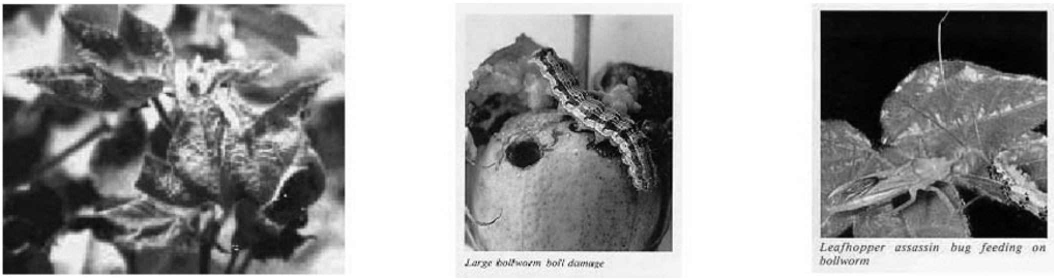


Figure 1(a): Map of Pakistan (www.cia.gov)



Key	Markaz
1	Bosan
2	Qadirpurran
3	Multan
4	Makhdum Rashid
5	Mumtazabad
6	Shujabad
7	Hafizwala
8	Jalalpur Pirwala
9	Qasba Marral

Figure 1(b): Area under study District Multan


Large bollworm boll damage
Leafhopper assassin bug feeding on bollworm

Virus: Cotton Leaf Curl Virus (Xiong *et al*, 1997)

Pest: Boll Worm (Bohmfalk *et al*, 1996)

Predator: Assassin Bug (Bohmfalk *et al*, 1996)

Figure 2: Virus, pest and predator present in the cotton fields

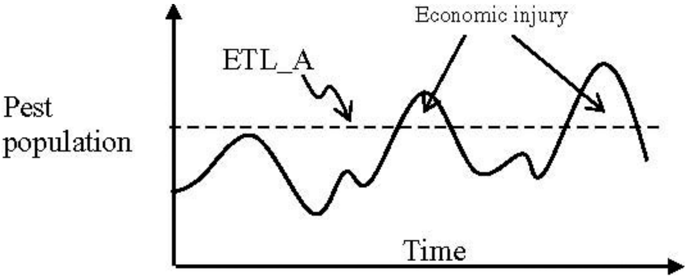


Figure 3: Agriculture Economic Threshold Level (ETL_A) and time.

of weather, pests and pesticide usage, so as to reduce the losses due to pest attacks by analyzing pest scouting data. Pest scouting is a systematic field sampling process that provides field specific information on pest pressure and crop injury (Introduction to Crop Scouting, 2001). The pest scouting data is being constantly recorded by the Directorate General of Pest Warning and Quality Control of Pesticides (DPWQCP), Punjab since 1984. However, despite pest scouting, yield losses have been occurring. The most recent being the Boll Worm attack on the cotton crop during 2003–04, resulting in a loss of nearly 0.5 million bales. This loss cannot be attributed to weather alone, but points to a multitude of factors, requiring efficient and effective data analysis, for better decision making.

The volume of pest scouting data that has been accumulated until now by DPWQCP is enormous both horizontally (scores of factors or attributes) and vertically i.e. number of records. A typical pest scouting sheet consists of 35 variables or attributes. Coarse estimate of pest scouting data recorded for the cotton crop alone stands at more than 1.5 million records, and growing. Tasking the human brain alone, for synthesis of information from this data is not only impractical but is unjust too. The objective of the work discussed in this paper is, complementing knowledge discovery in this massive data set, using proven information management tools and techniques, so as to support decision making.

4. RELATED WORK

The aim of this section is to give an overview of work done by other researchers in the field of Agriculture Data Warehousing. Data warehousing is very popular in domains such as telecommunications, retail sale, manufacturing and scientific research (Poe *et al*, 1998). An agriculture data warehouse is a rather new concept with very few parallels. There have been several references of plans of development of Agriculture Data Warehouses, such as Integrated National Agricultural Resources Information System (INARIS) Sharma *et al* (2000) or Ahmed and Greenaway (2002). Other than the USDA-NASS the world is yet to see a full-blown Agriculture Data Warehouse.

USDA-NASS data warehouse was established in 1997, the basic goal behind its construction was to standardize and integrate survey data generated by NASS (Yost and Nealon, 1999). This work differs in principal with Yost and Nealon (1999) as (i) data is generated by multiple sources, (ii) goal behind the Agri Extension data warehouse is a construction of a foundation on which analytical exploration can take place, and (iii) to demonstrate the viability of data mining of Agro-Met data.

The closest neighbour to USDA-NASS is the work done by the Italian NSI (National Statistics Institute) to set up a general data warehouse on enterprises and farms. It was constituted by the Statistical Information System on Enterprises (Capasso *et al*, 2000) which is a multidimensional and multifunctional structure that integrates all the official information available on enterprises and farms and economic issues in general. Another close runner-up in the agriculture domain is the SyR Data Warehouse developed for Swedish milk producers (Swensson and Sederblad, 1997).

5. PILOT AE DWH (PAE DWH) SYSTEM

The aim of this section is to discuss in detail the complete life cycle implementation of the PAE DWH following the 12-step approach of Shaku Atre, and also discuss the problems, issues encountered and their corresponding solutions. A pilot project strategy is highly recommended in data warehouse construction (Poe *et al*, 1998). As a full blown data warehouse construction requires significant capital investment, effort and resources. Therefore, the same must be attempted only after a thorough analysis, and a valid proof of concept. A small scale project in this regard serves many purposes such as (i) show users the value of Decision Support System (DSS) information, (ii) establish blue print processes for later full-blown project, (iii) identify problem areas and, (iv) reveal true data demographics. Hence doing a pilot project on a small scale seemed to be the best strategy.

Phase 1: Planning & Design	Phase 2: Building & Testing	Phase 3: Roll-Out & Maintenance
1. Determine Users' Needs 2. Determine DBMS Server Platform 3. Determine Hardware Platform 4. Information & Data Modeling	5. Construct Metadata Repository 6. Data Acquisition & Cleansing 7. Data Transform, Transport & Populate 8. Determine Middleware Connectivity 9. Prototyping, Querying & Reporting 10. Data Mining 11. On Line Analytical Processing	12. Deployment & System Management

Table 1: The 12-step implementation approach of a data warehouse of Shaku Atre (Atre, 2005)

As hardly any pilot projects are discarded, PAE DWH was treated with all due respect as a regular project. The other objectives of PAE DWH were (i) demonstrate the utility of such an undertaking (ii) do a dry run of the entire cycle of an AE DWH to get a feel of the "real-thing". For the sake of the pilot project and scarcity of resources, the scope of the work was limited to the pest scouting data of cotton crop for District Multan (Figure 1), for years 2001 and 2002. Actually pilot projects are supposed to work with limited data (Ponniah, 2001). The 12-step approach of Shaku Atre (Atre, 2005) followed is grouped into three main phases as shown in Table 1. The major focus of this paper is on Phase 2, though other phases will also be discussed.

5.1. Step 1: Determine Users' Needs

Agriculture data warehousing, and knowledge discovery using data mining are very new concepts in the agriculture domain of Pakistan. Actually, users are accustomed to decision making based on experience, or comparing the figures of this year versus last year. Therefore, getting proper user needs for a pilot project was not a realistic objective.

5.1.1 Availability of data

The scouts from the Directorate General of Pest Warning and Quality Control of Pesticides weekly sample 50 points in each *Tehsil* of the cotton growing districts of Punjab. Lately 60 *Tehsils* are sampled per week, resulting in sampling of 3,000 points within Punjab, and roughly 1,500 such points in district Multan; this has been going on for the last two decades. Simple estimates place the volume of the cotton pest scouting data recorded to be around 1.5 million records. Table 2 provides the details of the main attributes recorded at each sampling point. Static attributes are those attributes that are recorded on each visit by the scouts. The said attributes are repeated, and usually does not change frequently (slowly changing dimension), such as land ownership or the area of the field.

Pakistan Metrological Department (PMD) has been recording metrological data consisting of 50+ attributes for about five decades. As the cost of data available with PMD is too high, therefore, as a last resort, daily weather estimates (not actual values) for years 2001 and 2002, including minimum, maximum temperatures, % humidity and outlook were downloaded from the website of the newspaper Daily Dawn (www.dawn.com).

Static Attributes		Dynamic Attributes	
1	Farmer Name	1	Date of Visit
2	Farmer Address	2	Pest Population
3	Field Acreage	3	CLCV
4	Variety(ies) Sown	4	Predator Population
5	Sowing date	5	Pesticide Spray Dates
6	Sowing method	6	Pesticide(s) Used

Table 2: Cotton pest scouting attributes recorded by DPWQCP surveyors

5.1.2 Cost/benefit analysis, project estimation and risk assessment

Using the PAE DWH several data driven experiments were performed, the results were confirmed using actual data and also deliberated with an entomologist and extension personnel. In all cases the findings were factual and consistent with the ground realities, pointing to minimum risk in case a full blown AE DWH was developed.

5.2 Steps 2 and 3: Determine DBMS Server aand Hardware Platform

PAE DWH was implemented using NCR Teradata Data Warehousing solution on a server with dual Intel 950 Mhz Xeon processors and 1GB of RAM. Total internal Hard Disk capacity of the server amounted to 36 GB while external RAID control supports 8 additional SCSIs of 18 GB each.

5.3 Step 4: Information and Data Modelling

5.3.1 Dimensional Modelling: Determining the number of dimensions to be effective

Dimensional modelling is a technique used to model databases for analytical applications. It yields a simpler design and hence efficient retrievals using OLAP tools, one of the prime requirements for large data warehouses (Kimball, 1997; Kimball *et al*, 1998; Brobst, 1999). Figure 4 shows the Dimensional Model (DM) for PAE DWH.

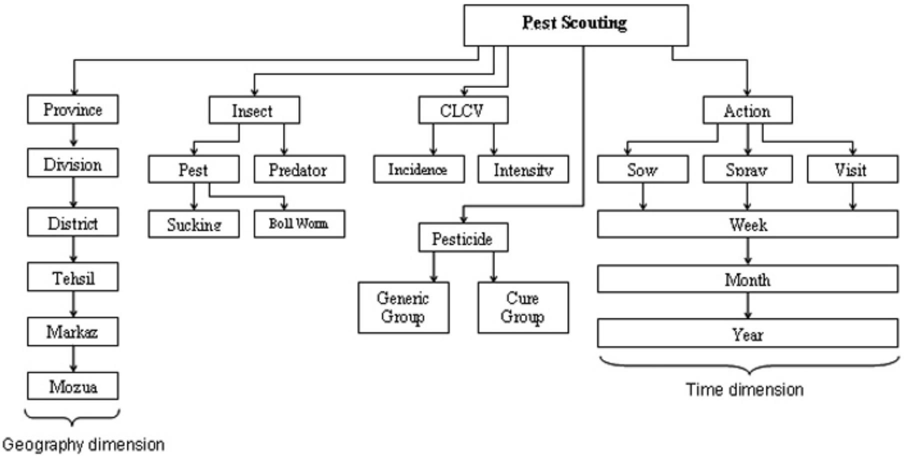


Figure 4: Dimensional Model PAE DWH

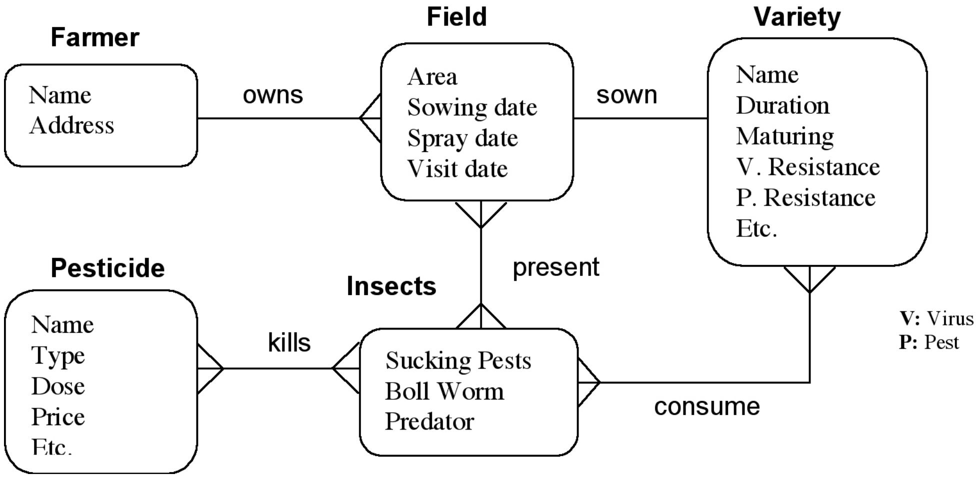


Figure 5: A simplified ERD of the Agri System being considered

Figure 5 shows a simplified ERD (Entity Relationship Diagram) of the Agriculture System being considered. The only field input shown is pesticides, although several other field inputs are there too, such as fertilizer, irrigation, machinery etc. that have an impact on the yield etc. As the data for these field inputs was not available, hence the corresponding entities have not been covered in the ERD

5.3.2 Logical and physical design of the data warehouse

Looking at the whole scouting process, number of data elements, type and frequency of data generated and nature of queries likely to be faced, a modified star schema is proposed for PAE DWH (Figure 6). It has been persistently reported in the literature that star schema best support decision support applications due to its simplified nature (Kimball, 1997; Kimball *et al*, 1998).

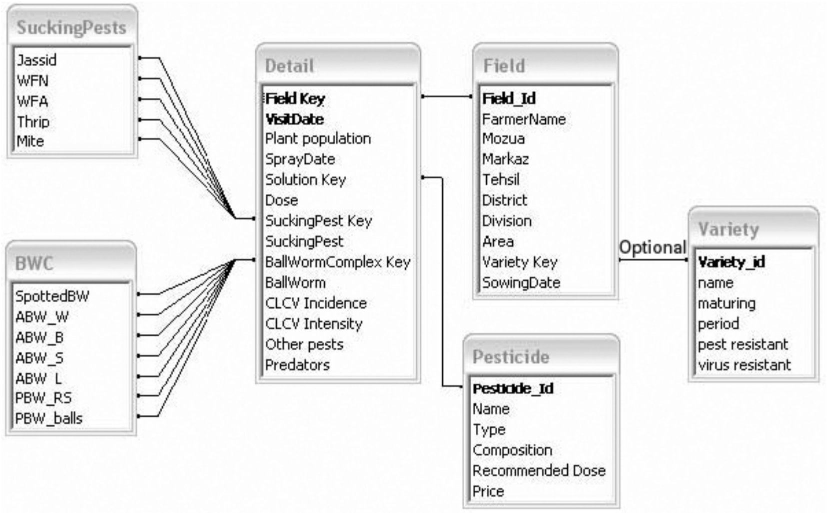


Figure 6: A simplified schema of the PAE DWH

Relevance of star schema to PAE DWH implementation can be seen by the facts that (i) using an RDBMS product that relies heavily on indexing for better performance and, (ii) enabling extraction of results from this data warehouse implementation through OLAP tools.

The goal of a star schema design is to simplify the physical data model so that RDBMS optimizers can exploit advanced indexing and join techniques in a straightforward manner. Moreover, a star schema helps put into place a physical data model capable of very high performance adhering to an OLAP model of data delivery (Poe *et al*, 1998). An added advantage of this schema is simplified SQL generation for front-end tools.

5.4 Step 5: Construct Metadata Repository

At the pilot level of AE DWH the most important aspects were "what mean what". The issues of who can access what and how, was not there, as there were only one or two users. The business metadata issues in the context of business rule will be briefly discussed in this section.

5.4.1 Building a metadata repository

To develop domain knowledge and to know the business rules, other than reading technical literature, meetings were held with the domain experts. The domain experts met were mostly from the Directorate of Pest Warning Multan, National Agriculture Research Centre (NARC), Islamabad and Pakistan Agriculture Research Council (PARC), Islamabad. The main benefit of the methodology was that the real problems were identified, based on which SQL queries were developed that were subsequently used to explore the database (Section 5.8). However, for knowledge discovery using data mining, mostly the domain knowledge was used in conjunction with help from domain experts for interpretation of results. Despite the meeting methodology adopted, metadata issues still cropped up, such as a skewed Pest-Predator relationship (Figure 7).

Pest-Predator relationship is a well-known entomological phenomenon. This relationship is counter-intuitive in the sense that, a high predator population does not mean a low pest population. If pest population is low, predator population will also be low, because there will be less "food" for predators to live on i.e. pests. Hence, in reality population of both the insect classes show similar

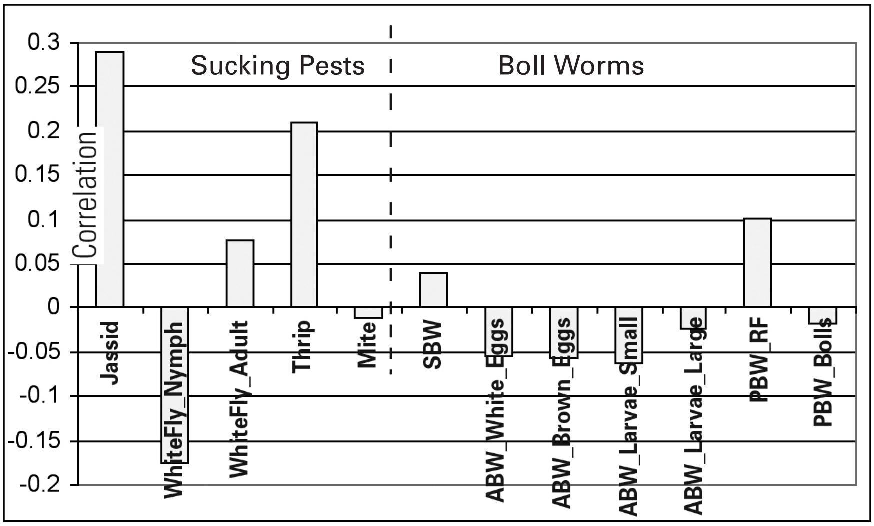


Figure 7: Predators vs. Pests Year 2001 ABW: Army Boll Worm **PBW:** Pink Ball Worm **RF:** Rc settled Flower

growth (and decline) rate at a particular point in time i.e. high positive correlation. The pest-predator correlations for the entire cotton season of 2001, are shown in Figure 7.

The dotted line divides the pests into two categories i.e. sucking pests and Boll Worm (BW) complex. From the pest scouting sheets it was found that unlike pests, the scouts do not note the predator details i.e. do not identify the specific predators present; just record the number of predators present.

However, note the surprising finding of negative or low positive correlations for most of the BW complex with respect to predator populations. It was hard to believe that BW predators were absent! On enquiring about this from DPWQCP personnel, it was found that only the presence of predators of sucking pests was noted. The exception of weak positive correlation in BW Complex is due to noting the presence of the Assassin Bug predator, for reasons best known to the scouts.

5.4.2 Business user's view of metadata

After several presentations and discussions with the end users from the Public sector organizations, the common apprehensions encountered were about the (i) existence of the primary scouting data (ii) quality of the data itself. These apprehensions were minimized by showing the actual scouting sheets and in one case confirmation by a senior provincial government official about the existence of the scouting data. Some soft measures were also used to demonstrate the data quality, such as demonstrating data validation (predator-pesticide relationship Figure 9).

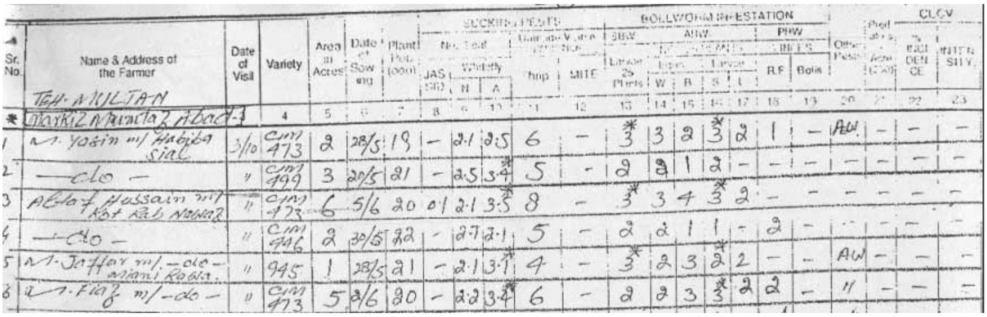
5.4.3 Steps to develop an effective metadata repository

It is not possible to discuss all the steps required to develop an effective metadata repository for different types of metadata. However, some of the steps that are intended to be followed in a full-blown DWH are:

1. Develop a statistical metadata repository for each column of the pest scouting sheet. This will be done for the metrological data too.
2. Writing the business rules such as the fractional population of sucking pests that is actually the average per leaf for 20 leafs from four plants.
3. Maintaining detailed records of data entry i.e. by whom, when, errors found, removed, reconciliation dates, etc.

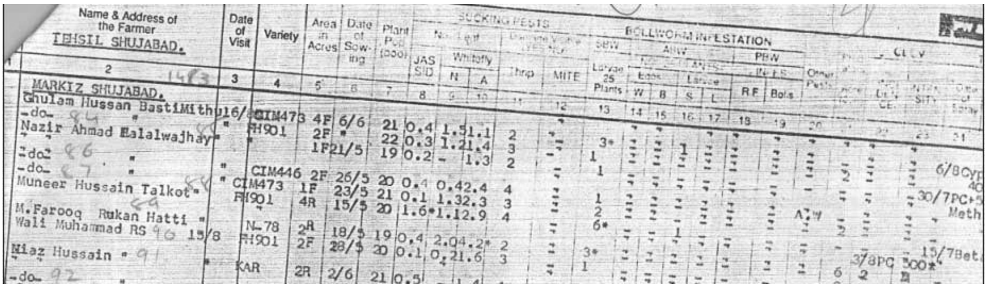
5.5 Step 6: Data Acquisition and Cleansing

Trained scouts from DPWQCP periodically visit randomly selected points and manually note 35 attributes, with some given in Table 2. These hand-written sheets are subsequently filed. For the last 10 years, the data collected was recorded by typing the hand-filled pest scouting sheets. Copy of a hand filled pest scouting sheet is shown in Figure 8(a):



Sr. No.	Name & Address of the Farmer	Date of Visit	Variety	Area in Acres	Date of Sowing	Plant Pop. (000)	SUCKING PESTS					BOLLWORM INFESTATION							Other Pests	Pred ators Ave	CLCV	
							JAS SID	Whitefly N	Whitefly A	Thrip	MITE	SBW Larvae 25 Plants	ABW Eggs W	ABW Eggs B	ABW Larvae S	ABW Larvae L	PBW RF	PBW Bolls			% INCI DEN CE	INTEN SITY
	TEH- MULTAN																					
*	Markiz Mumtaz Abad-1		4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	M. Yasin m/ Habiba Sial	3/10	CIM 473	2	28/5	19	–	2.1	2.5	6	–	3*	3	2	3*	2	1	–	AW.	–	–	–
2	—do—	"	CIM 499	3	20/5	21	–	2.5	3.4*	5	–	2	2	1	2	–	–	–	–	–	–	–
3	Altaf Hussain m/ Kot Rab Nawaz	"	CIM 473	6	5/6	20	0.1	2.1	3.5*	8	–	3*	3	4	3*	2	–	–	–	–	–	–
4	—do—	"	CIM 446	2	30/5	22	–	2.7	2.1	5	–	2	2	1	1	–	2	–	–	–	–	–
5	M. Jaffar m/ —do— Miani Rabia.	"	945	1	28/5	21	–	2.1	3.1*	4	–	3*	2	3	2*	2	–	–	AW	–	–	–
6	M. Fiaz m/ —do—	"	CIM 473	5	2/6	20	–	2.2	3.4*	6	–	2	2	3	3*	2	2	–	"	–	–	–

Figure 8(a): Hand filled Pest Scouting sheet



Name & Address of the Farmer TEHSIL SHUJABAD.	Date of Visit	Variety	Area in Acres	Date of Sowing	Plant Pop (000)	JAS S/D	Whitefly N	Whitefly A	Thrip	MITE	SBW Larvae 25 Plants	W	B	S	L	RF	Bols	Other Pests				
2 1483	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MARKIZ SHUJABAD.																						
Ghulam Hussan BastiMithu	16/8	CIM473	4F	6/6	21	0.4	1.5	1.1	2	-	3*	-	-	1	-	-	-	-	-	-	-	
-do- 84	"	FH901	2F	"	22	0.3	1.2	1.4	3	-	1	-	-	-	-	-	-	-	-	-	-	
Nazir Ahmad Balalwajhay 85	"	"	1F	21/5	19	0.2	-	1.3	2	-	1	-	-	-	-	-	-	-	2	-	-	6/8Cyp 40
-do- 86	"									-		-	-	-	-	-	-	-	-	-	-	
-do- 87	"	CIM446	2F	26/5	20	0.4	0.4	2.4	4	-		-	-	-	-	-	-	-	-	-	-	30/7PC+5 Meth
Muneer Hussain Talkot 88	"	CIM473	1F	23/5	21	0.1	1.3	2.3	3	-	1	-	-	-	-	-	-	-	-	-	-	
M.Farooq Rukan Hatti 89	"	FH901	4R	15/5	20	1.6*	1.1	2.9	4	-	2	-	-	-	-	-	-	A.W	2	-	-	
Wali Muhammad RS 90	15/8	N-78	2R	18/5	19	0.4	2.0	4.2*	2	-	6*	-	-	1	-	-	-	-	-	-	-	15/7Bet.
Niaz Hussain 91	"	FH901	2F	28/5	20	0.1	0.2	1.6	3	-	3*	-	-	-	-	-	-	-	-	3/8PC	500*	
-do- 92	"	KAR	2R	2/6	21	0.5	-	1.4		-	1	-	-	-	-	-	-	-	6	2		

Figure 8(b): Typed Pest Scouting sheet

The * in Figure 8 corresponds to pest hot spot or flare-up or ETL A.

The pest scouting sheets are larger than A4 size (8.5” x 11”), hence the right end was cropped when scanned on a flat-bed A4 size scanner. The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.

As a first step, OCR (Optical Character Reader) based image to text transformation of the pest scouting sheets was attempted. But it did not work even for relatively clean sheets with very high scanning resolutions, such as 600 dpi. Subsequently DEOs (Data Entry Operators) were employed to digitize the scouting sheets by typing. To reduce spelling errors in pesticide names and addresses, drop down menu or combo boxes with standard and correct names were created and used.

Data from other sources was collected and integrated into the data warehouse, such as lists of standard pesticide names obtained from National Agriculture Research Centre (Irshad *et al*, 2001) and standard crop varieties list obtained from CCRI (Central Cotton Research Institute) Multan. Details about cotton varieties such as leaf colour, ball size etc. was obtained from Seed Certification Authority, Islamabad.

5.5.1 Data cleansing and reconciliation

To maintain full compliance between the data sheets and their digitized copies, a double entry strategy was adopted. Every row of data was entered by two data entry operators separately. Using SQL both copies of the scouting database were compared row by row, and column by column. Subsequently a report of conflicts was generated, and then the corresponding data sheets were consulted for final reconciliation. Corresponding to around 200 typed scouting sheets, 4,400 records were generated with 35 columns.

Data cleansing and standardization is probably the largest part in an ETL exercise. For PAE DWH major issues of data cleansing had arisen due to data processing and handling at four levels by different groups of people i.e. (i) Hand recordings by the scouts at the field level (ii) typing hand recordings into data sheets at the DPWQCP office (iii) photocopying of the scouting sheets by DPWQCP personnel and finally (iv) data entry or digitization by hired data entry operators.

After achieving an acceptable level of data quality, the data was loaded into Teradata data warehouse; subsequently each column was probed using SQL for erroneous entries. Some of the errors found were correct data in wrong columns, nonstandard or invalid variety names etc. There were some intrinsic errors, such as variety type “999” or spray date “12:00:00 AM” inserted by the system against missing values. Variations found in pesticide names and cotton variety names were removed by comparing them with standard names.

5.6 Step 7: Data Transform, Transport and Populate

Among the different types of transformations performed in the implementation, only the more complex i.e. multiple M:1 transformations for field individualization will be discussed in this section.

5.6.1 Motivation

The data recorded consists of two parts i.e. static and dynamic (Table 2). On each visit, the static, as well as the dynamic data is recorded by the scouts, thus resulting in static values getting recorded repeatedly. Since no mechanism is used to uniquely identify each and every farmer, therefore, trivial queries, such as total area scouted, distribution of varieties sown etc. gives wrong results. For example, while aggregating area, the area of the farmer with multiple visits during the season is counted multiple times, giving incorrect results, the same is true for varieties sown. Therefore, to do any reasonable analysis after data cleansing, the most important step of data transformation being individualization of the cultivated fields, not farmers. The reason being, a farmer can have multiple fields, but a field is associated or owned by a single farmer.

5.6.2 Method

Field individualization turned out to be a very laborious process. It was attempted by first uniquely identifying the farmers. This was achieved by collectivity sorting farmer name, *Mozua* and *Markaz* using a variant of the Basic Sorted Neighbourhood (BSN) method. The grouping of farmer names was scrutinized to fix the spelling errors in the farmer names and unique farmer ID was assigned to each farmer. Subsequently based on the farmer ID, sowing date, area and variety, cultivated fields were uniquely identified and field ID assigned to each field. Figure 9 shows the two level mechanism of field individualization.

5.6.3 Results

To demonstrate the amount of error removed because of field individualization, consider the case of scouted area and unique farmers.

Without field individualization, the cotton scouted area for 2001 and 2002 added to **23,293** and **26,088** acres, respectively. After field individualization, the correct scouted area turned out to be **14,187** and **13,693** acres respectively i.e. a correction of about 50%. Similarly unique farmers reduced from **2,696** to **1,567**. The method of field individualization is in no way perfect, there were few cases of farmers with the same geography, sowing date, same variety and same area. Such cases were dropped.

5.6.4 Transporting the data

Once the data entry was complete, double checked and reconciled the corresponding files were compressed and moved from the premises of the DEO (Data Entry Operator) to the University,

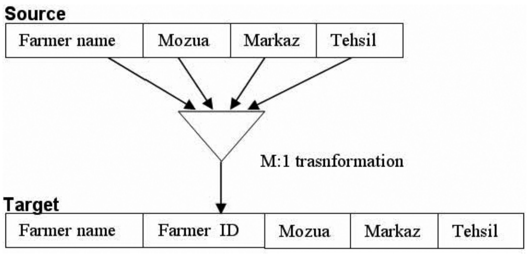


Figure 9(a): Level-I of field individualization

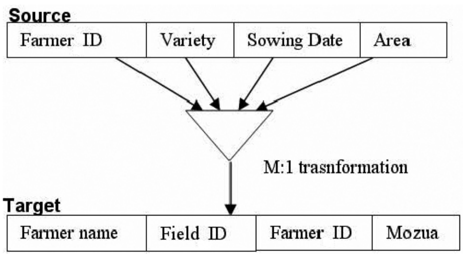


Figure 9(b): Level-II of field individualization

where sample printouts of data entered were taken and a final random quality check was performed. Subsequently minor errors, if any were fixed and data was loaded into the PAE DWH.

5.6.5 Populating or loading the data warehouse

To begin with the PAE DWH was populated using full refresh strategy using block slamming. Subsequent refreshes are likely to take less time because of the static component of the data, which gets loaded only once. Hence for future loading, incremental refresh strategy is going to work with lesser overhead. The only pitfall is the change in farmer and field (land) demographics i.e. field changing hands, or getting divided due to family inheritance or multiple fields getting combined into a single field. These subtle changes can only be recorded and communicated with close interaction between the scouts and the farmers.

5.6.6 Data Validation

Quality and validity of the underlying data is the key to meaningful and authentic analysis. After ensuring a satisfactory level of data quality (based on cost-benefit trade-off) it is extremely important to somehow judge the validity of data that a data warehouse constitutes. Some very natural checks were employed for this purpose. Relationship between the pesticide spraying and predator (beneficial insects) population is a fact well understood by agriculturists (James and Price, 2002). Predator population decreases as pesticide spray increases and then continually decreases till the end of season, as shown in Figure 10. In Figure 10 the y-axis shows the relative frequency of pesticide sprays in multiples of 100 ml, and average predator populations greater than zero.

5.7 Step 8: Determine Middleware Connectivity

Since the source data is maintained in a non digital format, connectivity with the data warehouse was irrelevant. Once digitized, it was rather trivial to load the data into the warehouse. Furthermore, in the foreseeable future, it was not anticipated that the scouting sheets were going to be maintained in a digitized form.

5.8 Step 9: Prototyping, Querying and Reporting

The PAE DWH was implemented with the involvement of the potential end users. The implementation was centred around numerous meetings with the potential end users, discussion of results, and also explicit sets of questions provided by them. Despite small numbers of rows i.e. 4,000+, the PAE DWH was implemented using Teradata for the sake of completion of the entire cycle. The following SQL query was used to generate Figure 10.

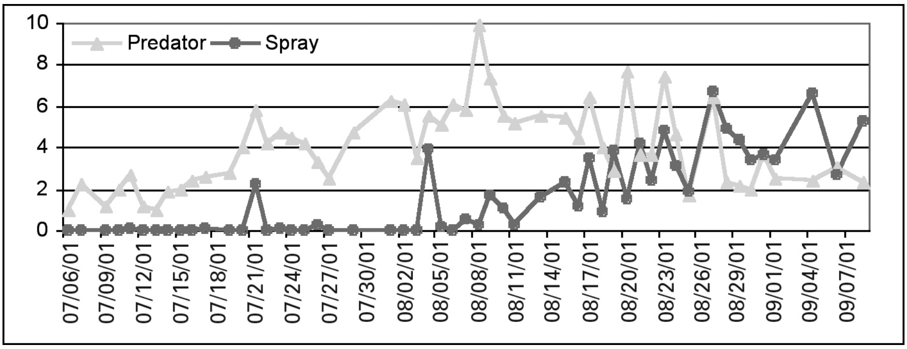


Figure 10: Year 2001 Frequency of spray vs. Predators population

```
SELECT Date_of_Visit, AVG(Predators), AVG(Dose1+Dose2+Dose3+Dose4)
FROM Scouting_Data WHERE Date_of_Visit < #12/31/2001# and predators > 0
GROUP BY Date_of_Visit;
```

5.9 Steps 10 and 11: OLAP and Data Mining

A low-cost OLAP tool was indigenously developed; actually it was a multi dimensional OLAP or MOLAP. Using the MOLAP tool, agriculture extension data was analyzed. The details are not given here, interested readers are referred to Abdullah *et al* (2004) for details. A data mining tool was also developed based on an indigenous technique (Abdullah and Hussain, 2006) that used the crossing minimization paradigm for unsupervised two-way clustering. The results of using the said tool for one-way clustering are discussed in detail in Section 7.

5.10 Step 12: Deployment and System Management

Since PAE DWH was a pilot project, therefore, the traditional deployment methodologies and system management techniques were not followed to the word, and are not discussed here.

6. DECISION SUPPORT USING PAE DWH

The aim of this section is to demonstrate using real data how a PAE DWH can be used to support decision making. For this purpose the results of querying the spray dates and the sowing dates will be analyzed.

6.1 Working Behaviours at Field Level: Spray dates

As expected, the results of querying for spray dates and spray frequency for 2001 and 2002 do not display any well defined patterns; as it is dependent on pest populations (Figure 3), availability of pesticides etc. To study the relationship between sprays and time, moving average of sprays for five days, and a moving correlation of sprays for five days were calculated. For the sake of uniformity, the moving average of spray was normalized using the maximum spray frequency. The results are shown in Figure 11.

No relationship should have existed for the two years. But note the surprising finding that most sprays occurring on and around 12 August in BOTH years with a high correlation, appearing as a spike. Also note the dip in sprays around 11 September! Sowing at a predetermined time makes

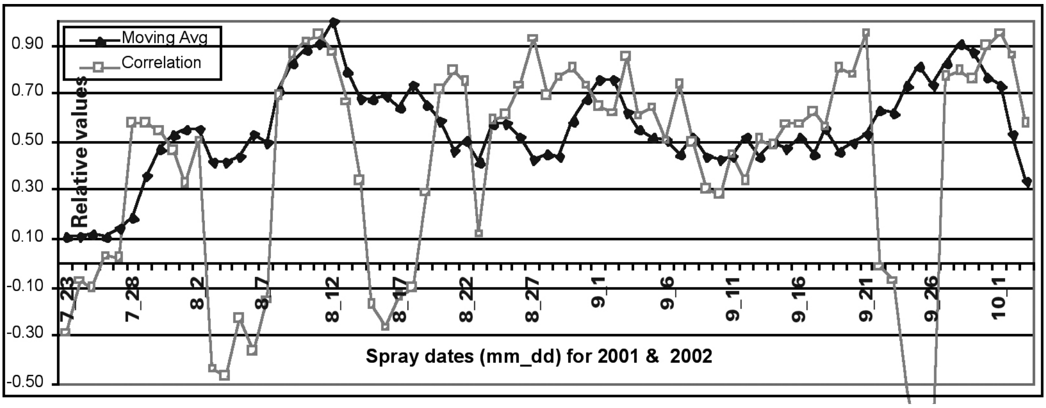


Figure 11: Spray frequency vs. day of year for Year 2001

sense, as it is under the control of the farmer, but that is not true for spraying. Pests don't follow calendars; therefore, whenever, ETL A is crossed pesticides are sprayed.

The independence day of Pakistan is 14 August and a national holiday. In Pakistan, people are in a habit of sandwiching gazetted holidays with casual leaves; consequently offices are closed for a longer period, including that of pesticide suppliers. The 14 August occurred on Tuesday and Wednesday in 2001 and 2002, respectively, thus making it ideal to stretch the weekend. During August/September, humidity is also high, with correspondingly high chances of ball worm infestations. Therefore, apparently the farmers decided not to take any chances, and started spraying around 11 August; evidently even when it was not required. Unfortunately, the weather forecast for 13 August 2001 and 2002 was showers and cloudy, respectively. Therefore, most likely the pesticide sprayed was washed-off. Unfortunately the decline in sprays around 9/11 could not be explained.

6.2 Working Behaviours at Field Level: Sowing dates

PAE DWH contains integrated data and hence can be probed along any dimension. Generally an iterative analysis technique proves most beneficial. Analysts then capitalize on these records and asks a more specific question. The process of iteratively building up on the previous question continues until a result of significant importance is achieved. Results of querying for sowing dates in Year 2001 and Year 2002 based on day of the month are shown in Figure 12.

The sowing dates seem to correctly follow the prescribed guidelines i.e. (i) it is illegal to sow before 1 May and (ii) as per recommendation of CCRI (Central Cotton Research Institute) Multan, yield drastically reduces for sowing after 15 June. Checking the peak sowing dates for both years

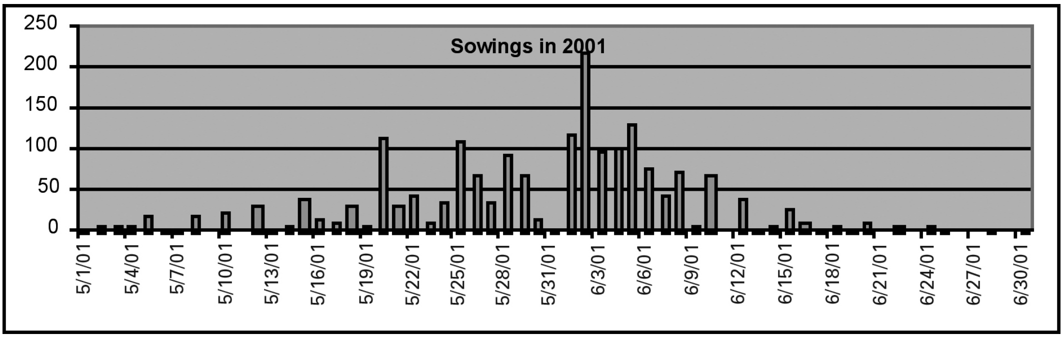


Figure 12(a): Sowing frequency vs. day of year for Year 2001

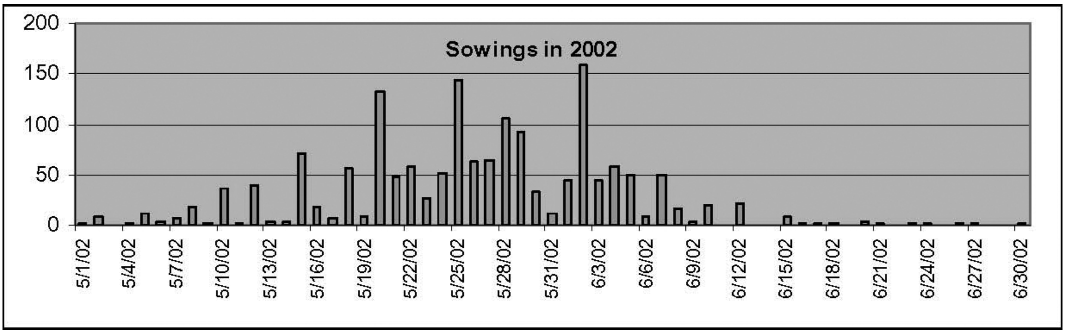


Figure 12(b): Sowing frequency vs. day of year for Year 2002

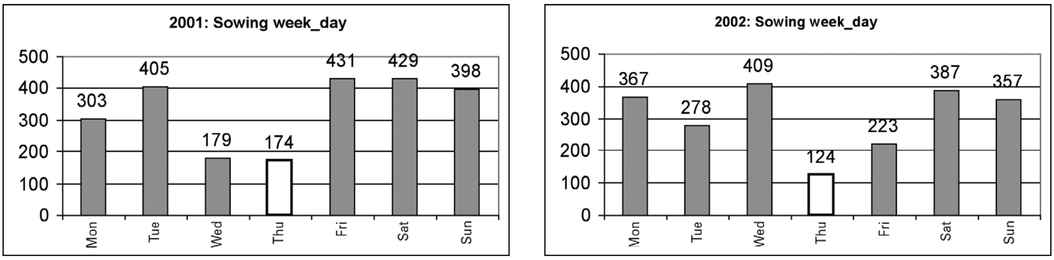


Figure 13: Number of sowings against week days

revealed no relationship using different types of calendars (solar, lunar, seasonal). So it was felt pertinent to check the detailed static data instead i.e. farmer address, area and variety sown, however, no significant relationship was found. The results of querying the sowing date based on the day of the week are shown in Figure 13.

Observe least number of sowings done on Thursdays, in each year. This finding was later confirmed by extension personnel. Multan is famous for its shrines. Thursdays are usually related with religious festivals and activities, a mix of devotion and recreation, and usually held at shrines, hence a tendency of doing less work on Thursdays. Similar behaviour was observed for spraying too.

6.3 Sowing date vs. Weather

Sowing dates are also influenced by rainfall and release of water in canals. Consulting the CCRI records showed rainfall in first week of June but not in May 2001. However, closer scrutiny of records for the third week of May 2001 showed the highest humidity and lowest maximum temperature as compared to the remaining month. Probably the farmers in anticipation of a rainfall, sowed most during the third week of May 2001. Similar weather patterns were found to be true for 2002 as shown in Table 3.

7. DATA MINING PAE DWH

The aim of this section is to discuss the results of performing data mining on the PAE DWH, in an attempt to provide a possible explanation about an apparent anomaly between pesticide consumption and cotton yield.

Pesticides are used as a means for increasing yield by controlling pest populations, thus a positive correlation is believed to exist between yield and pesticide consumption. However,

Month	Week	Air Temp °C				Relative Humidity % at 0800 hrs		Rain fall (mm)		Sunshine Hours	
		Maximum		Minimum							
		2001	2002	2001	2002	2001	2002	2001	2002	2001	2002
May	I	43.2	41.0	25.5	25.2	41.2	41.0	-	-	9.25	8.72
	II	43.3	43.5	28.8	26.1	47.2	49.9	-	-	8.19	9.63
	III	38.7	43.5	27.2	25.8	62.2	46.0	-	-	7.82	10.75
	IV	42.2	39.2	29.3	26.5	53.5	62.0	-	11.0	9.72	8.35
June	I	39.5	41.5	28.5	29.3	64.0	54.6	1.3	-	7.57	9.11

Table 3: Comparative Metrological Data (Weekly Average) for 2001 and 2002 recorded at CCRI, Multan

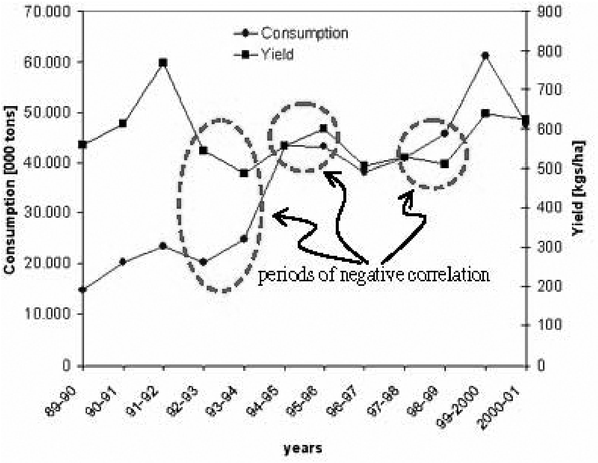


Figure 14: Yield and Pesticide Usage in Pakistan: Source FAO (2001)

existence of an undesirable, sometime even negative correlation between pesticide consumption and yield has been observed in Pakistan (Food and Agriculture Organization 2001). Figure 14 shows periods of marked decreases in yield while the pesticide consumption is on the rise, and also its converse, thus creating a complex situation and an apparent anomaly.

Excessive use of pesticides is harmful in multiple ways. Farmers have to bear the additional cost of buying and applying pesticides, while increased pesticide usage develops immunity in pests, thus making them more harmful to the crops. Excessive usage of many pesticides is also harmful for the environment and hazardous to humans, such as those who spray the pesticides, and the women who pick cotton.

Reasons for the apparent anomaly of pesticide consumption and yield could be many. Because of the size and complexity of the data sets, automatically discovering such reasons by clustering seems to be the only viable way. In this work an attempt has been made to do exactly the same i.e. using an approach never tried for this purpose (other than the work of the authors) in the agriculture pest scouting domain, i.e. data mining.

7.1 Data considered

For 2001 the available pest scouting data covered spray dates from 1 June 2001 to 29 October 2001, there were 948 spray records, using SQL Structured Query Language (SQL) with GROUP BY clause the records grouped into 94 groups of unique dates along with spray frequency for each date. For 2002 the scouting data available covered spray dates from 14 July 2002 to 12 October 2002. There were 1,014 spray records, using SQL the records grouped into 74 groups of unique dates along with spray frequency. For the remaining records there were no sprays. Using the spray dates as reference, T_{min} (minimum temperature of the day), T_{max} (maximum temperature of the day), % Humidity along with spray frequency were used to compute pair wise correlation similarity matrices for both years. The matrices were subsequently symmetrically discretized using the median value of each column. The discretized similarity matrices for both years but not in any particular order of rows (or columns) are shown in Figures 15 (a, c).

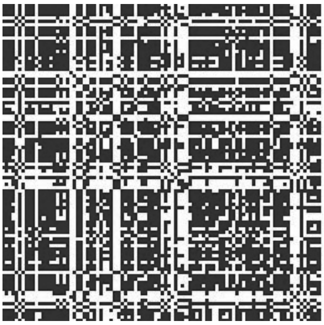


Figure 15(a): Input for 2001 (94 groups)

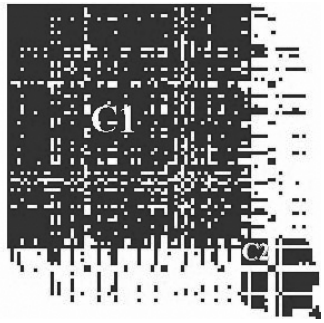


Figure 15(b): Clustered Output for 2001

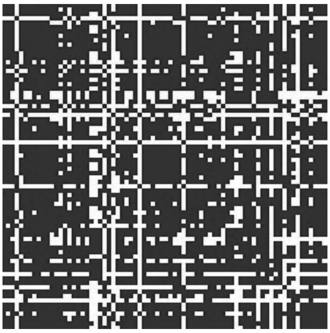


Figure 15(c): Input for 2002 (74 groups)

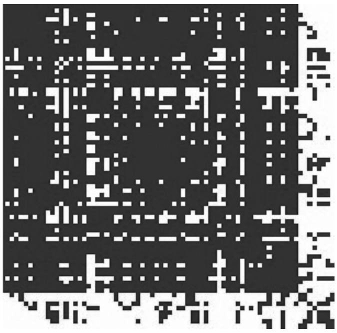


Figure 15(d): Clustered Output for 2002

7.2 Clustering of Agro-Met data

Figure15(a) shows the input similarity matrix for year 2001. The biclustering technique using the crossing minimization paradigm (Abdullah and Hussain, 2006) was modified to perform one-way clustering, resulting in Figure 15(b) showing two clusters i.e. C1 and C2.

After extracting clusters C1 and C2 for 2001, detailed scrutiny of the smaller cluster i.e. C2 turns out to be roughly the first unique 30 dates of records, for this date range the average sprays per day throughout the district Multan are five, while for the larger cluster i.e. C1 average sprays per day are 13, which is as per the prevalent practice. As shown in Figure 15(c, d) no clustering occurred for year 2002.

7.3 Discussion

Pesticides are either sprayed in the early morning or by the end of the day i.e. avoiding high temperatures to prevent breakdown and evaporation; hence spraying at low temperatures are recommended i.e. negative correlation between temperature and frequency of spray. Similarly, sprays are advised in high humidity to prevent evaporation i.e. positive correlation between frequency of spray and % humidity.

Analysis of detailed data corresponding to C2 as shown in Figure 15(b) revealed **negative** correlation with humidity (i.e. -0.17), and **positive** correlation with minimum temperature (i.e. 0.32) i.e. opposite to the recommendations. For the other large cluster i.e. C1 the correlation between frequency of sprays was found to be zero against both met elements i.e. humidity and minimum temperature Tmin. Analysis of detailed data corresponding to Figure 15(d) i.e. year 2002 showed no distinct clustering, a desirable positive correlation with humidity i.e. 0.24 but an undesirable positive correlation with T_{min} i.e. 0.064. Again this does not guarantee maximum efficacy of pesticide, yet it is better than 2001.

After individualization of the cultivated fields the total cotton area scouted was established for 2001 and 2002. Based on this area, the average pesticide used per acre for the entire cotton season was calculated to be **23.96** ml/acre and **22.46** ml/acre for year 2001 and 2002, respectively. Thus in 2001 around 6% more pesticide was used as compared to 2002.

On further querying the data for specific pesticides used in 2001 and 2002 during roughly the same time period covered in C2, it was discovered that in 2001 the pesticide *Cypermethrin* was used during the early season at three times the dose as compared to the dosage used in 2002. *Cypermethrin* was also the top pesticide of choice, but during 2002 it was the fifth pesticide of choice. *Cypermethrin* is a systemic pesticide i.e. absorbed by foliage and translocated throughout the plant, and is recommended against sucking and chewing cotton pests i.e. bollworms. There are about a dozen brands of pesticides based on *Cypermethrin* in the market, some more focused towards bollworms (such as *Ripcord*), while others covering both sucking pests and bollworms (such as *Arrivo*). Unfortunately the scouting data did not specify which brand of *Cypermethrin* was used. However, as per the pesticides data available at <www.nationalpak.com> collected from the Central Cotton Research Institute, *Cypermethrin* is mainly recommended to control the population of the bollworm complex.

The next obvious, and interesting question is, “the effect of these malpractices on pest populations”. The findings are given in Figure 16 that shows the number of records when ETL A (Economic Threshold Level) was crossed.

It can be seen that in year 2001 there are higher incidences of ETL A crossings for sucking pests as compared to 2002. Because of the attack of sucking pests, the plants lose their vigour, their growth is stunted and they are more susceptible to attack by bollworms at the later stage of growth. This is also evident by the higher number of ETL A crossings for Spotted Boll Worm (SBW) in 2001 as compared to 2002.

In a nut-shell, in 2001 higher doses of *Cypermethrin* were used against metrological recommendations basically for controlling sucking pests, which is also not the best recommendation. As the most suitable pesticide was not used under lethal conditions, this could have resulted in developing resistance in the subsequent generations of the surviving pests, evident from comparatively many instances of ETL A crossings in 2001, thus resulting in weak plants with low yields. This could be one of the possible reasons for a low cotton yield during the 2000–2001 cotton season.

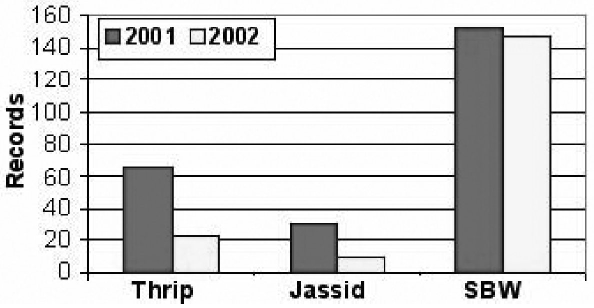


Figure 16: Comparison of pest populations above ETL A for 2001 and 2002

8. CONCLUSIONS

As per the published literature, there has been very little research in the domain of agriculture data warehousing and data mining, especially using the pest scouting data. This may be due to a number of reasons, such as volume and complexity of the data, the complexity of Extract Transform Load (ETL) and dual domain knowledge i.e. IT and Agriculture i.e. Agro-Informatics. Unlike OLTP systems used in business or e-commerce etc., in the agriculture domain there are hardly any digitized operational databases of primary data, so one has to resort to data available in typed (or hand written) pest scouting sheets. Data entry of these sheets is very expensive, slow and prone to errors.

Another problem is data duplication or more precisely repetition without any unique keys. Particular to the pest scouting data, each farmer is repeatedly visited by agriculture extension people. This results in repetition of information, about land, sowing date, variety etc. (as in Table 2). Hence, farmer and land individualization are critical, so that repetition may not impair aggregate queries. Such an individualization task is hard to implement for multiple reasons. There is also skewness in the scouting data. Public extension personnel (scouts) are more likely to visit educated or progressive farmers, as it makes their job of data collection easy. Furthermore, large land owners and influential farmers are also more frequently visited by the scouts. Thus the data may not give a true statistical picture of the farmer demographics.

Although the PAE DWH consists of less than 5,000 records, but the sheer number of attributes i.e. 50+ and their alpha-numeric nature makes data analysis very complex. Whereas some questions, such as farmer behaviour, were successfully answered using SQL, but hidden knowledge cannot be extracted using queries. For handling such complex scenarios, data mining is shown to be successful in addressing even apparent anomalies.

Despite all the problems and complexities of creating the PAE DWH, the results obtained are very encouraging, interesting and useful, thus making a convincing argument for a full blown agriculture DSS based on an Agriculture Data Warehouse, which employs OLAP tools for analysis, and knowledge discovery using data mining. However, embarking on such a task is not only expensive, but challenging too.

ACKNOWLEDGEMENT

This work would not have been possible without the support of Dr Ijaz Pervaiz and guidance of Dr Amir Muhammed.

REFERENCES

ABDULLAH, A. and HUSSAIN, A. (2006): A new biclustering technique based on crossing minimization, to appear in the *Neurocomputing Journal*.

ABDULLAH, A., MALIK, D., BASIT, A., KARIM, F. and UMER, M. (2004): Experiences in development and use of a low cost OLAP tool for analyzing agricultural data, 8th World Multiconference on Systemics, *Cybernetics and Informatics (SCI 2004)* Orlando, USA, July 18–21.

AHMED, A. and GREENAWAY, F. (2002): GIS application for land planning and management in Montserrat, West Indies. *Proceedings of the Open Source GIS – GRASS Users Conference 2002* – Trento, Italy 11–13 September.

AHMED, M. and NAGY, J. G. (2001): Private investment in agriculture research: Pakistan: Economic Research Services, *U.S. Department of Agriculture,* January.

ATRE, S. (2005): 12-step approach: http://www.atre.com/navigator/ viewed on 9 April.

BOHMFALK, G. T., FRISBIE, R. E., STERLING, W. L., METZER, R. B. and KNUTSON, A. E. (1996): Identification, Biology And Sampling Of Cotton Insects, Texas A&M University System.

BROBST, S. (1999): Perfect dimensions: *Intelligent Enterprise*, June.

CAPASSO, G., DEL MONDO, G. and VIGNOLA, L. (2000): Un sistema informativo per la produzione el'integrazione delle statistiche sulle imprese (progetto SISSI). In GIUSTI A. (eds.) *Ingegnerizzazione del processo di produzione dei dati statistici*, 181–191. Padova: Cleup.

CHAUDHRY, M. R. (2000): New frontiers in cotton production: International Cotton Advisory Committee, USA.

COTTON AND GINNING (2003): Small and medium enterprise development agency, Pakistan, http://smeda.org/bopp/cotton-ginning.pdf viewed on 20 September.

ECONOMIC SURVEY OF PAKISTAN (2001/2003): Govt. of Pakistan

INTRODUCTION TO CROP SCOUTING (2001): Plant protection program, College of Agriculture, Food and Natural Resources, MU Extension University of Missouri-Columbia.

IRSHAD, M., HAQ, E. and IQBAL, J. (2001): Catalogue of insecticides for agricultural pests of Pakistan: Integrated Pest Management Institute, *National Agriculture Research Center (NARC)*, Islamabad.

JAMES, D.G. and PRICE, T.S. (2002): Imidacloprid boosts TSSM egg production, *Agriculture and Environment News*, Issue No. 189, Washington State University, USA, July.

KIMBALL, R. (1997): A dimensional modelling manifesto, *DBMS and Internet Systems*, August.

KIMBALL, R., REEVES, L., ROSS, M. and THORNTHWAITE, W. (1998): The data warehouse lifecycle toolkit: Expert methods for designing, developing and deploying data warehouses. Pub. John Wiley & Sons, Chichester.

POE, V., KLAUER, P. and BROBST, S. (1998): Building a data warehouse for decision support: 2nd Edition, Pub. Prentice Hall.

PONNIAH, P. (2001): Data warehousing fundamentals: A comprehensive guide for IT professionals, Pub. Wiley-Interscience.

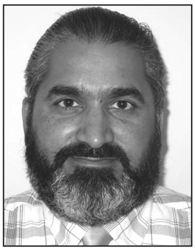
SHARMA, S. D., SINGH, R. and RAI, A. (2000): Integrated National Agricultural Resources Information System (INARIS), *Indian Agricultural Statistics Research Institute*, New Delhi.

SWENSSON, C. and SEDERBLAD, B. (1997): Data warehouse - A new tool in extension service to Swedish Milkproducers. *First European Conference for Information Technology in Agriculture*, Copenhagen, Denmark 15–18 June.

XIONG, Z., NADEEM, A., WENG, Z. and NELSON, M. R. (1997): Cotton leaf curl virus is distinct from cotton leaf crumple virus. *Department of Plant Pathology*, University of Arizona.

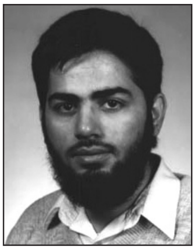
YOST, M. and NEALON, J. (1999): Using a dimensional data warehouse to standardize survey and census metadata: National Agricultural Statistics Service, *U.S. Department of Agriculture*, Fall.

BIOGRAPHICAL NOTES



Ahsan Abdullah

Ahsan Abdullah did his BSc Electrical Engineering with honours from the University of Engineering and Technology, Lahore, MSc Computer Engineering from the University of Southern California, USA, and MSc Computer Sciences also from the University of Southern California. He has authored two books on Internet; and was first author of more than 25 publications in refereed international journals and conference proceedings. In recognition of his achievements in the field of Agro-Informatics he has been given the IT Excellence award for 2004 by NCR Pakistan (NYSE: NCR) and Dr. M. N. Azam Prize in Computer Science for 2005 by the Pakistan Academy of Sciences.



Amir Hussain

Amir Hussain was born in 1972. He obtained his BEng (with First Class Honours) and his PhD in Electronic and Electrical Engineering in 1992 and 1996 respectively, both from the University of Strathclyde in Glasgow, Scotland UK. From 1996–98, he held a post-doctoral research fellowship at the University of Paisley in Scotland, UK. From 1998–2000, he held a Research Lectureship in Applied Computing Science at the University of Dundee in Scotland. Since 2000, he has been with the University of Stirling in Scotland, where he is currently a Senior Lecturer in Computing Science. His research interests include novel interdisciplinary research for modelling and control of complex systems, adaptive non-linear speech signal processing and computational intelligence techniques and applications. His research activities have been funded by, amongst others, the UK Engineering & Physical Sciences Research Council (EPSRC), the UK Royal Society, the European Commission and industry. These have led to one international patent in neural computation and more than 80 publications to-date in *various international journals, books and refereed international conference and workshop proceedings. He is IEEE Chapter Chair for the IEEE UK and RI Industry Applications Society Chapter. He currently serves on the editorial board of a number of international Journals and has acted as an invited guest editor for various journals, special issues, including the (Elsevier) Neurocomputing Journal Special Issue on BICS2004. He serves as an independent Expert for the European Commission's 6th Framework Program for RTD, and as a Consultant for the Pakistan Higher Education Commission, Islamabad. He is an Adjunct Professor and founding co-Director of the Centre for Intelligent Systems Engineering at Muhammad Ali Jinnah University, Islamabad, Pakistan.*