Contents lists available at ScienceDirect

Applied Soft Computing

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

Hybrid models in agent-based environmental decision support

Marina V. Sokolova^{a,b,c}, Antonio Fernández-Caballero^{a,b,c,*}

^a Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, 02071 Albacete, Spain

^b Kursk State Technical University, Kursk, ul.50 Let Oktyabrya, 305040, Russia

^c Departamento de Sistemas Informáticos, Escuela de Ingenieros Industriales de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain

ARTICLE INFO

Article history:

Received 13 September 2010

Received in revised form 10 March 2011

Accepted 4 May 2011

Available online xxx

Keywords:

Hybrid models

Human health

Environmental pollution

Decision support systems

Multi-agent systems

Simulation

ABSTRACT

Providing informational support in decision making is one of the priority directions of research in the sphere of public health management. Modern approaches suggest wide usage of intelligent data mining methods and Web-services, but just a few enable to study a complex system from an interdisciplinary point of view. In this paper an agent-based decision support system (ADSS), which embodies the principles of the interdisciplinary approach and facilitates multi-focal view and examination of the "Environment–Public health" system, is introduced. The detailed design of the system with the emphasis on the roles, scenarios and its implementation is presented. The data mining procedures used for data preparation, modeling and simulation, which include statistics, methods of artificial intelligence and hybrid models in form of cascade committee machines, are described. Then, the advantages of the proposed hybrid models against "singular" modeling methods are demonstrated. Finally, a case study for a selected region is presented and its results are discussed.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Health care decision support is faced with the challenges of complex and diverse data and knowledge, the lack of standardized terminology compared to basic sciences, the stringent performance and accuracy requirements and the prevalence of legacy systems [20]. Capturing domain knowledge has proven to be one of the largest challenges for expert systems builders [10]. It has become evident, through research into intelligent decision support systems (DSS), that expert knowledge needs to be supplemented with facts gleaned from machine learning processes to solve more difficult problems. Due to the stochastic and complex nature of most real world systems, simulation models of these systems are themselves difficult to build as well as time consuming to execute. In many cases, decision makers cannot afford to explore a large area of the decision variable space or to conduct a lengthy search for the best set of decision variables [4]. One feasible alternative is to build a meta-model [19]. So, due to the growing complexities and uncertainties in decision making situations, model-driven DSS have become increasingly important to decision makers [25]. Model-driven DSS can assist decision makers in applying quantita-

tive models to support the decision-making process. The simulation model is an important type of quantitative model used in model-driven DSS. A simulation model can imitate the behavior of an actual or anticipated human or physical system. It can capture much more detail about a specific system than algebraic models. It can capture underlying mechanism and dynamics of a system, which enables decision makers to effectively manage daily operations and make long term plans. It provides also a test-bed to assess changes in operations and managerial policies [11]. Decision making involves processing or applying information and knowledge, and the appropriate information/knowledge mix depends on the characteristics of the decision making context [38]. Intelligent data analysis may be defined as "encompassing statistical, pattern recognition, machine learning, data abstraction and visualization tools to support the analysis of data and discovery of principles that are encoded within the data" [18]. The authors state that the principal differences between intelligent data analysis and knowledge discovery in databases is that the techniques used are those of artificial intelligence [7] rather than pure traditional statistical methods. Intelligent data analysis refers to all methods that are devoted to support the transformation of data into information exploiting the knowledge available on the domain. Very recently, a novel framework for the construction of augmented fuzzy cognitive maps based on fuzzy rule-extraction methods for decisions in medical informatics has been presented [24].

Environment is a clear example of a complex domain, composed of numerous self-organized subsystems. If interactions of humans within the environment are studied, the level of

* Corresponding author at: Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, 02071 Albacete, Spain.

Tel: +34 967599200; fax: +34 967599224.

E-mail addresses: marina.v.sokolova@gmail.com (M.V. Sokolova), Antonio.Fdez@uclm.es (A. Fernández-Caballero).

complexity of such a system greatly increases [27,32]. A very recent paper describes the design of a fuzzy decision support system in multi-criteria analysis approach for selecting the best plan alternatives or strategies in environment watershed [5]. It is a fact that environment affects human health. Climate changes together with growing anthropogenic impact intensify interactions within the “environment–human health” system. Humans are affected by this global imbalance, and react with direct and indirect health problems, some example include “excessive heat-related illnesses, vector- and waterborne diseases, increased exposure to environmental toxins, exacerbation of cardiovascular and respiratory diseases due to declining air quality, and mental health stress. Vulnerability to these health risks will increase as elderly and urban populations increase and are less able to adapt to climate change. In addition, the level of vulnerability to certain health problems varies by location. As a result, strategies to address climate change must include health as a strategic component on a regional level. Improving health while addressing climate change will contribute to public health infrastructure today, while reducing the negative consequences of a changing climate for future generations” [36].

This complex system can be decomposed into “Air pollutants”, “Climate change”, “Water”, “Ecological”, and “Social/economic” sub-systems. And, each of the sub-systems affects human health. The strength and dynamics of health outcomes can be measured with statistical indicators: mortality, morbidity, birth defects rate, etc. The “Health” concept represents a complex system, which includes physical, social, mental, spiritual and biological well-being, spanning across all the spheres of human lives. Environmental pollution, as one of the factors with dominant and obvious influence upon human health, causes direct and latent harmful effects, which must be evaluated in order to create a set of preventive health-preserving solutions. That is why linking all the named components in one system and studying of this system leads to the analysis of potential and present health problems, retrieval of the new ones and to working out the in-depth view of situation development, strategies, and activities oriented to situation management and control.

Numerous studies have shown an adverse relationship between environmental hazards and human health. Some research works were aimed to discover detail mechanisms of these relationships [13,31]. For example, a known and aggressive air contaminant is ambient fine particulate matter (PM_{2.5}). It has been demonstrated that it increases cardiovascular risks, with a stronger impact on heart failure [12] and causes premature death if it is locally emitted [23]. Ref. [34] examines and proves the presence of associations between carbon monoxide (CO), nitrogen dioxide (NO₂), ozone (O₃), sulfur dioxide (SO₂), and particulate matter (PM₁₀ and PM_{2.5}), and visits for angina/myocardial infarction, heart failure, dysrhythmia/conduction disturbance, asthma, chronic obstructive pulmonary disease, and respiratory infections. Another paper [35] describes a research in which adverse respiratory health effects on children caused by the petrochemical refinery's emissions are studied. The emissions include SO₂, particles and oxides of nitrogen as well as fugitive emissions consisting of numerous aliphatic and aromatic hydrocarbons were studied. Some studies of prenatal exposure to solvents including tetrachloroethylene have shown increases in the risk of certain congenital anomalies among exposed offspring [1]. Some of the environmental pollutants play the role of carcinogens. For example, there are two types of environmental exposures, which are related to lung cancer: radon in homes and arsenic in drinking water [2]. Also, in [22] a causal effect that leads to increased respiratory illnesses in children due to traffic pollution is studied.

In case of environmental impact assessment (EIA), all the advantages of intelligent agents become crucial. Indeed, the intelligent agent and multi-agent systems (MAS) approach is an excellent

technique that can help to reduce the complexity of a system by creating modular components, which solve private subtasks that together achieve the whole goal. Every agent utilizes the most effective technique for solving the subtask and does not apply the general approach, which is often acceptable for the system in the whole, but not optimal for a concrete subtask [29]. In accordance to a recent paper [6] “the applications of agents and multi-agent systems in the health care and clinical management environments are becoming a reality. Most agent-based applications are related to the use of this technology in patient monitoring, treatment supervision and data mining.” EIA is an indicator, which enables evaluation of the negative impact upon human health caused by environmental pollution. Environmental pollution, a factor with dominant and obvious influence, causes direct and latent harm, which must be evaluated and simulated in order to create a set of preventive health-preserving solutions. Large amounts of raw data describe the “Environment–Human health” system, but not all the information is used. It transforms from the initial “raw” state to the “information” state, which suggests organized data sets, models and dependencies, and, finally, to the “new information” which is represented as a set of recommendations, risk assessment and forecast values.

1.1. Notations

This paper focuses on the description of an agent-based decision support system, and its application to environmental domain. The nomenclature and abbreviations used throughout the paper are provided in Table 1.

2. Description of the agent-based DSS

The proposed framework consists of three phases, which are reflected in the architecture of the agent-based DSS (ADSS). The proposed system is logically and functionally divided into three layers: the first is dedicated to meta-data creation (information fusion), the second is aimed at knowledge discovery (data mining), and the third layer provides real-time generation of alternative scenarios for decision making [30,28]. The levels do not have strongly fixed boundaries, because the agents construct a community, in which the agents' spheres of competence can overlap, and the boundaries are smooth. The ADSS asserts the main points of a traditional decision making process, and includes the following steps:

1. Problem definition.
2. Information gathering.
3. Alternative actions identification.
4. Alternatives evaluation.
5. Selection of decision.
6. Decision implementation.

The first and the second stages are performed during the initial step, when the expert information and initial retrospective data is gathered, stages three, four and five are solved by means of the MAS, and the sixth stage is supposed to be carried out by the decision maker. Although the goals of the ADSS are determined and keep constant within various domains, the goals of the case study are not visible and clear at first sight. That is why the domain of interest has been studied with the creation of a goal tree.

2.1. Goals and scenarios

The principal goals of the proposed multi-agent system follow the logical sequence of the main stages of the DeciMaS framework. For this reason, the process of the system design starts with iden-

Table 1
Nomenclature and abbreviations.

Nomenclature	Abbreviation
AA	ANN agent
ADSS	Agent-based decision support system
BOD	Biochemical oxygen demand
BP	Back-propagation algorithm
CA	Correlation agent
CMA	Committee machine agent
CO	Carbon monoxide
COD	Chemical oxygen demand
CSA	Computer simulation agent
DA	Decomposition agent
DAA	Data Aggregation agent
DOA	Domain Ontology agent
DPA	Data Preprocessing agent
DSA	Data Smoothing agent
DSS	Decision support system
EA	Evaluation agent
EIA	Environmental impact assessment
FAA	Function Approximation agent
FHX	Farfield Human Exposure
GA	Genetic algorithms
GAA	Gaps and Artifacts Check agent
GANN	Neural network trained with genetic algorithms
GMDH	Group Method of Data Handling
GMDHA	GMDH agent
ICD	International Statistical Classification of Diseases and Related Health Problems
ICIDH	International Classification of Functioning and Disability
MAS	Multi-agent system
MFA	Mining Data Fusion agent
MoFA	Morbidity Data Fusion agent
NA	Normalization agent
NO ₂	Nitrogen dioxide
O ₃	Ozone
PFA	Petroleum Data Fusion agent
PM _{2.5}	Ambient fine particulate matter
PM ₁₀	Particulate matter
RA	Regression agent
RPROP	Resilient propagation algorithm
SO ₂	Sulfur dioxide
TFA	Traffic Pollution Fusion agent
WDFA	Waste Data Fusion agent
WFA	Water Data Fusion agent

tification of general goals that are divided into subgoals and then refined.

The final goal is *Create recommendation* and it is achieved as a result of three parallel goals *Make forecast*, *Make sensitivity analysis* and *Check for alarm*. The goals *Make forecast* and *Make sensitivity analysis* use parts of the same knowledge. Both sensitivity analysis and forecasting are based on models received as a result of *Create models* and *Select the best models* goals. The goal *Check environmental impact* is independent from the other goals of the second logical layer, although its outputs are used for making recommendations on the third logical layer. The goal *Preprocess data* is the initial goal for all the data mining procedures. The correlation between goals is shown in Fig. 1.

A scenario represents a purposeful behavioral model of collective activity [33]. A scenario serves to achieve a practical goal and may include sub-scenarios as well. It involves, at least, two agents performing particular roles. A scenario is a composition of agents' scenarios oriented to achieve goals, which in their turn, may be achieved by sets of problem-solving activities that include agent roles. Every role that is played by an agent, includes actions and plans. Tables 2–5 present five scenarios that the system should enact. Each scenario is described as a sequence of steps, where each step has its type, name, role, and type of the data it uses and produces. The type term can be a goal, an action, a percept, or a sub-scenario.

The scenario of achieving a *Retrieve and fuse data* goal is provided in Table 2. It is initiated by the Data Aggregation agent, and the Domain ontology and the data mining agents participate in this scenario. It suggests collaboration with the actor EXPERT. The scenario has seven steps. The first step suppose receiving the perception "Obtain expert knowledge". The remaining steps are actions. The scenario to achieve a *Preprocess data* goal is shown in Table 3 and it describes interaction of the Data Aggregation agent and the Data Preprocessing agent and its team. This scenario has one goal *Prepare data for modeling* and performs five action, namely *Eliminate artifacts*, *Normalize data*, *Smooth data*, *Parametric correlation* and *Non-parametric correlation*.

The Check Environmental Impact scenario has two actions: *Create neural networks* and *Evaluate impact assessment*, and invokes the Function approximation agent and the Artificial Neural Network agent. This scenario is presented in Table 4. The *Create models* scenario has 11 steps, which are actions to be completed by agents from the Function Approximation agent team, as it is shown in Table 5. The *Create recommendation* scenario suggests collaboration with the external actor USER/DECISION MAKER and contains ten steps. These steps are carried out within two roles: *Computer simulation* and *Decision making*. Table 6 provides a detailed view of this scenario.

2.2. Actors

The proposed ADSS supposes communication with two actors. One actor, EXPERT, embodies the external entity which possesses the information about the problem area. In more detail, the EXPERT contains the knowledge of the domain of interest represented as an ontology, and delivers the knowledge to the ADSS. Through the *Simulate Models* scenario, the user interacts with the knowledge base, and gets recommendations if they have been previously simulated and stored before, or creates and simulates the new alternative decisions. As a result of the interaction within the *Retrieve and Fuse data* scenario, the raw information is being read, and is shown as "Heterogeneous Data Sources" data storage, and the "Pollutants" and "Morbidity" data sources are created. The second actor, named USER/DECISION MAKER, is involved in an interactive process of generation of alternative decisions to choose the optimal one/ones. This actor communicates with the agents by passing a message, stating the model, simulating values, predicting periods, levels of variable change, etc. The actor accepts the best alternative in accordance with its beliefs and the MAS.

The flows of works, which are essential for decision making, include three sub-scenarios: the *Simulate models* scenario, the *Create recommendation* scenario and the *Search for the adequate model* scenario. Additionally, there are three goals, which are related to each scenario and have similar names. Each goal has a number of activities, and within each scenario resources in the form of data sources are used, modified or created.

2.3. Roles of the proposed MAS

Scenarios focus on how a multi-agent system achieves goals, interaction models define the agent's outgoing communications, and the actors represent external entities, which interact with the system [33]. The detailed behavior of an agent is represented by roles, plans. The communication between agents is shown in acquaintance models. Roles represent agent's functions, responsibilities and expectations. A role enables pooling together the goals of the system in accordance with different types of behavior that an agent assumes when archiving a goal or a series of goals. Table 7 shows a view of the correspondent logical levels, and the roles, which are played there.

Fig. 1. Goals tree.

Table 2
Scenario for achieving the “Retrieve and fuse data” goal.

Name	Retrieve and fuse data scenario				
Actors	Expert				
Initiator	The Data Aggregation agent				
Trigger	The “Read ontology” message to the Domain Ontology agent				
Step #	Type	Name	Role	Data used	Data created
1	Percept	Obtain expert knowledge	Data Fusion	Domain Ontology agent beliefs	Domain Ontology agent beliefs
2	Action	Read ontology and data sources	Data Fusion	External data sources	Domain Ontology agent beliefs
3	Action	Fuse morbidity data	Data Fusion	External data sources	Domain Ontology agent beliefs
4	Action	Fuse transport data	Data Fusion	External data sources	Domain Ontology agent beliefs
5	Action	Fuse data on mines	Data Fusion	External data sources	Domain Ontology agent beliefs
6	Action	Fuse petroleum usage data	Data Fusion	External data sources	Domain Ontology agent beliefs
7	Action	Fuse data on wastes	Data Fusion	External data sources	Domain Ontology agent beliefs

Table 3
Scenario for achieving the “Preprocess data” goal.

Name	Preprocess data scenario				
Actors	No				
Initiator	The Data Preprocessing agent				
Trigger	The “ready” message from the Data Aggregation agent				
Step #	Type	Name	Role	Data used	Data created
1	Action	Eliminate artifacts	Data Clearing	Domain Ontology agent beliefs	Morbidity, Pollutants
2	Action	Normalize data	Data Clearing	External data sources	Morbidity, Pollutants
3	Action	Smooth data	Data Clearing	Morbidity, Pollutants	Morbidity, Pollutants
4	Goal	Prepare data for modeling	Data Clearing	Morbidity, Pollutants	dataX, dataY
5	Action	Parametric correlation	Data Clearing	dataX, dataY	Correlation table
6	Action	Non-parametric correlation	Data Clearing	dataX, dataY	Correlation table

Table 4
Scenario for achieving the “Check Environmental Impact” goal.

Name	Check Environmental Impact scenario				
Actors	No				
Initiator	The Function Approximation agent				
Trigger	The “Start Data Mining” message from the Function Approximation agent				
Step#	Type	Name	Role	Data used	Data created
1	Action	Create neural networks	Impact Assessment	dataX, dataY, Correlation table	IAResults
2	Action	Evaluate impact assessment	Impact Assessment	IAResults	IAResults

Table 5
Scenario for achieving the “Create models” goal.

Name	Create models scenario				
Actors	No				
Initiator	The Function Approximation agent				
Trigger	The “ready” message from the Data Preprocessing agent				
Step #	Type	Name	Role	Data used	Data created
1	Action	Decomposition	Decomposition	dataX, dataY, Correlation table, ranges	Groupings table
2	Action	Create univariate regression models	Function Approximation	dataX, dataY, Correlation table	Models table
3	Action	Create multiple regression models	Function Approximation	dataX, dataY, Correlation table	Models table
4	Action	Create neural network models	Function Approximation	dataX, dataY, Correlation table	Models table
5	Action	Create GMDH-models	Function Approximation	Models table, dataX, dataY	Models table
6	Action	Evaluate univariate regression models	Function Approximation	Models table	Models table
7	Action	Evaluate multiple regression models	Function Approximation	Models table	Models table
8	Action	Evaluate neural network models	Function Approximation	Models table	Models table
9	Action	Accept models	Function Approximation	Models table	Models table
10	Action	Create committee machines	Function Approximation, Impact Assessment	Models table	Final models
11	Action	Creation of reports	Function Approximation, Impact Assessment, Decomposition	IAResults, Final models, CS results, Grouping table	

The distribution of roles for agents determines the agents' specialization and knowledge. One of the intentions for the system design was to assign one role to each agent or agent team. That requirement was met for the roles *Data Fusion* and *Data Clearing* where the teams of Data Fusion agent and of the Data Preprocessing agent carry out these roles. Moreover, the Function Approximation agent manages three data mining roles: *Impact Assessment*, *Decomposition* and *Function Approximation*, and the Computer Simulation agent takes on *Computer Simulation*, *Decision Making* and *Data Distribution* roles.

2.4. Description of the agents

Once the multi-agent system notions have been defined and its logical architecture has been determined, with the set of goals, scenarios, interaction models and data usage, a global view of the system's layers and description of agent teams may be provided.

With regard to the proposed multi-agent architecture and in order to gain time of the recommendation generation process and optimize interactions between agents, local agent teams were used. The teams coordinate and supervise task execution and utilization of resources. Agent teams synchronize the work of the system, execute plans in a concurrent mode, and strengthen the internal management by local decision making. There are four agent teams defined within the system: two within the first level, and one team on the second and third level. Each “main” agent plays several roles.

2.4.1. The Data Aggregation agent and its team

The Data Aggregation agent (DAA) is the principal agent, which acts within the first logical layer. One of the agents is oriented to read the Domain Ontology, and the others have to retrieve information from the identified data sources. There are a number of subordinate agents under its control. These are:

Table 6
Scenario for achieving the “Create recommendation” goal.

Name	Create recommendation scenario				
Actors	User/decision maker				
Initiator	The Computer Simulation agent				
Trigger	The “ready” message from the Function Approximation agent				
Step #	Type	Name	Role	Data used	Data created
1	Percept	Obtain preferences for simulation	Computer Simulation		Data for simulation
2	Goal	Make forecast			
3	Action	Forecasting	Computer Simulation	Models tables, dataX, dataY	CS Results table
4	Goal	Make sensitivity analysis	Computer Simulation	dataX, dataY, Correlation table	CS Results table
5	Action	Models simulation	Computer Simulation	Models table, dataX, dataY	CS Results table
6	Percept	Preferences for decision	Decision Making		Data for decision
7	Action	Criteria application	Decision Making	CS Results table, IAResults	Final models
8	Goal	Check for alarm	Decision Making		
9	Action	Alarm generation	Decision Making	Alarm levels	Final models
10	Action	Creation of reports	Decision Making	IAResults, Final models, CS results	

Table 7
Roles played in the multi-agent system.

Logical level	Main agent	Subordinate agent	Role
Data Fusion	Data Aggregation agent	Domain Ontology agent Traffic Pollution Fusion agent Water Data Fusion agent Petroleum Data Fusion agent Mining Data Fusion agent Morbidity Data Fusion agent Waste Data Fusion agent	Data Fusion
	Data Preprocessing agent	Normalization agent Correlation agent Data Smoothing agent Gaps and Artifacts Check agent	Data Clearing
Data Mining	Function Approximation agent	Regression agent ANN agent GMDH agent Committee Machine agent Decomposition agent Evaluation agent	Impact Assessment Decomposition Function Approximation
Decision Making	Computer Simulation agent	Forecasting agent View agent Alarm agent	Computer Simulation Decision Making Data Distribution

1. The Domain Ontology agent (DOA).

2. The fusion agents:

- Water Data Fusion agent (WFA),
- Petroleum Data Fusion agent (PFA),
- Mining Data Fusion agent (MFA),
- Traffic Pollution Fusion agent (TFA),
- Waste Data Fusion agent (WDFA)
- Morbidity Data Fusion agent (MoFA).

The process of information fusion requires working with multiple data sources. Some of them can vary significantly in their format and internal data structure. These are the reasons why the Data Aggregation agent receives several subordinate agents in its disposition. They facilitate data retrieval because each of them is specified in a particular type of pollutant. The Data Aggregation agent must achieve the following goals:

1. Obtain information from the ontology of the domain.
2. Search for information sources, which may contain information of interest stored in the ontology of the domain.
3. Retrieve information from the found sources.
4. Transform the retrieved information in order to avoid heterogeneity.
5. Fuse information.

2.4.2. The Data Preprocessing agent and its team

The Data Preprocessing agent (DPA) aims to prepare the initial data for further modeling. It manages a number of subordinate agents, which make up its team. Each subordinate agent specializes in a different data clearing technique:

- Gaps and Artifacts Check agent (GAA) clears fused raw information from missing and inconsistent values and fill the gaps.
- Data Smoothing agent (DSA) carries out exponential and moving average smoothing procedures.
- Normalization agent (NA) normalizes data sets.
- Correlation agent (CA) calculates correlation matrices.

2.4.3. The Function Approximation agent and its team

The Function Approximation agent (FAA) has a hierarchical team of subordinate agents, which serve to carry out the roles:

“Impact Assessment”, “Decomposition” and “Function Approximation”. FAA has under its control a number of subordinate agents:

- Data mining agents incorporate well-known techniques already applied to more or less recent decision support systems (e.g. [17,3,9,37], among others). In our case they work in a concurrent mode and create models of the following types:

- The Regression agent (RA) creates regression models for given independent and dependent variables. It reads information from belief data sets for independent variables X and dependent variables Y . The Regression agent disposes two plans:

- * *simpleRegression* used for the generation of univariate regression models,
- * *multipleRegression* used for multiple regression models generation.

Each plan permits the creation of linear and non-linear models: hyperbolic, exponential and power models.

- The Artificial Neural Network agent (AA) creates neural network models. The ANN agent creates different types of models based on neural networks. For this reason, it executes several plans:

- * *evaluateImpactAssessment*, aimed to calculate environmental impact assessment and to select the most influential factors X for every dependent variable Y ,
- * *neuralNetwork*, used for the generation of approximation and autoregression models of $Y = F(X)$, $Y = F(t)$ and $X = F(t)$.

Feed-forward neural networks trained with backpropagation algorithm are calculated within the *evaluateImpactAssessment* plan. When an agent chooses the *neuralNetwork* plan, it creates feed-forward neural network models trained with:

- * backpropagation algorithm (BP),
- * genetic algorithms (GA),
- * resilient backpropagation algorithm (RPROP).

- The GMDH agent (GMDHA) creates polynomial models with the group method of data handling.

- The Evaluation agent (EA) calculates evaluation criteria for the models. The Evaluation agent evaluates the received models and checks the adequacy of the model against the experimental data, and returns the list of the accepted models, while the others are banned and deleted. The Evaluation agent has four plans for model evaluation. These plans are:

extractCSV.plan

Input: Data Sources *.csv.

Output: Retrieved information **C** and **P**.

- (1) Read file line to line and analyze tokens. Search for elements **C** and **P**. Extract found concepts/properties and “remember” their positions in the file.
- (2) Search for concepts/properties as intersections of found concepts.
- (3) Change found properties if their scales are different from the standard ones.
- (4) Write retrieved concepts **C** and **P** to internal file/DB and send it to **DAA**.

extractDOC_XLC.plan

Input: Data Sources *.doc, *.xls.

Output: Retrieved information **C** and **P**.

- (1) Search for elements from **C** and **P** in rows or/and columns. Mark found concepts/properties.
- (2) Search for concepts/properties as intersections of found rows and columns.
- (3) Change found properties if their scales are different from the standard ones.
- (4) Send retrieved concepts **C** and **P** to **DAA**.

Fig. 2. Plans for data extraction from the CSV, XLS and DOC files.

- *evaluateANN*, which is used when the Evaluation agent receives the *StartEvaluation* message from the Artificial Neural Network agent;
- *evaluateSimpleRegression*, which is used when the Evaluation agent receives the *StartEvaluation* message indicating the “univariate” parameter as a regression type from the Regression agent;
- *evaluateMultipleRegression*, which is used when the Evaluation agent receives the *StartEvaluation* message indicating the “multiple” parameter as a regression type from the Regression agent;
- *acceptModels*, which is used for each data mining agent in order to check the adequacy of a model from the initial data.
- The Committee Machine agent (CMA) creates hybrid models, using models created by the RA and the AA, and evaluated by the EA.
- The Decomposition agent (DA) carries out the decomposition procedure. It provides a relevant plan *Decomposition* to range the inputs of the neural model by their importance, and illustrates how the model output may change in response to variation of an input.

2.4.4. The Computer Simulation agent

The Computer Simulation agent interacts with the user and performs a set of tasks within the *Computer Simulation*, *Decision Making* and *Data Distribution* roles. Its subordinate agents are the following:

- The Forecasting agent, which is used to create forecasts and predictions of dependent and independent variables.
- The Alarm agent, which is used to identify the values that are received by the Forecasting agent, which exceed permissible level.
- The View agent, which is used to organize the computer-user interaction and create textual, graphical, and other types of documents.

The Computer Simulation agent asks for the user’s preferences, and, to be more precise, for the information of the disease and pollutants of interest, the period of the forecast, and the ranges of their value changes. Once the information from the user is

received, The Computer Simulation agent sends a message to the Forecasting agent, which reasons and executes one of the plans, which include *Forecasting*, *ModelSimulation*, and *CriterionApplication*. When the alternative is created, the Computer Simulation agent sends another message to the Alarm agent. The Alarm agent compares the simulation and forecast data from the Forecasting agent with the permitted and alarm levels for the correspondent indicators. If they exceed the levels, the Alarm agent generates alarm alerts.

3. Data for experiment

In order to evaluate the proposed agent-based DSS, retrospective data dated from 1989 until 2007 from Castilla-La Mancha (a Spanish region) is used. Indeed, resources offered by the Instituto Nacional de Estadística (Spanish Statistics Institute) [16] and by the Instituto de Estadística de Castilla-La Mancha (Castilla-La Mancha Statistics Institute) are used for the research. The factors that describe the *Environmental pollution–Human health* system are used as indicators of human health and influencing factors of environment. These can cause negative effects upon the above noted indicators of human health. The factors used in the experiment are presented in Table 8.

Morbidity, classified by sex and age, is accepted as an indicator to evaluate human health. Table 9 provides a list of diseases examined in this case study. The diseases included in the research are chosen in accordance with the International Statistical Classification of Diseases and Related Health Problems [14]. The sex groups include “males”, “females” and “total”; and the age groups consist of “all the ages”, “under 1 year”, “1–4 years”, “5–14 years”, “15–24 years”, “25–34 years”, “35–44 years”, “45–54 years”, “55–64 years”, “65–74 years”, “75–84 years” and “85 years and over”.

4. Results of the experiment

4.1. Data retrieval

In the data for experiment described there are 148 data files that contain information of interest. These files are used to extract the

Table 8
Pollutants studied in the research.

	Pollutant class	Factors
1	Transport	Number of lorries, buses, autos, tractors, motorcycles, others
2	Usage of petroleum products	Petroleum liquid gases; petroleum autos; petroleum; kerosene; gasohol; fuel-oil
3	Water characteristics	Chemical oxygen demand (COD); biochemical oxygen demand (BOD5); solids in suspension; nitrites
4	Wastes	Non-dangerous chemical wastes; other non-dangerous chemical wastes; non-dangerous metal wastes; wastes from paper industry; dangerous wastes of glass; dangerous wastes of rubber; dangerous solid wastes; dangerous vitrified wastes; wastes from used equipment; metallic and phosphorus wastes
5	Principal miner products	Hull; mercury; kaolin; salt; thenardite; diatomite; gypsum; rock; others

information. After extraction, data is placed in data arrays in accordance with the ontology. The fusing agents, which are supervised by the Data Aggregation agent, work with CSV, XLS and DOC-files. The algorithm of data extraction for these types of files is shown in Fig. 2. The Domain Ontology agent reads the OWL-file that contains the ontology of the system. Firstly, the agent creates the hierarchy of classes:

```
Class :Pollution
  Class :Water_pollution
  Class :Solar_radiation
  Class :Transport
  Class :Dangerous_wastes
    Class :Urban_waste_products
    Class :Industrial_waste_products
  Class :Industry
    Class :Minery_products
Class :individuals_pollution
Class :Morbidity
  Class :Exogeneous
  Class :Endogeneous
Class :Ontology_of_Environment
Class :Data
Class :Region
} Class :Pollution is a super-class of Class :Water_pollution
is a super-class of Class :Transport
is a super-class of Class :Dangerous_wastes
is a super-class of Class :Industry
Class :Urban_waste_products is a sub-class of Class :Dangerous_wastes
```

The part of the output, given above, shows the names of the classes from the OWL-file. Next, the hierarchical links between the classes, shown with “sub-class” and “super-class” properties, are also retrieved. For example, the line `Pollution is a super-class of Class :Water_pollution` shows one such relation. The restrictions and properties of each class, generated by the Domain Ontology agent, are shown next:

Table 9
Diseases studied in the research.

	Disease class	Factors
1	Endogenous diseases	Certain conditions originating in the perinatal period Congenital malformations, deformations and chromosomal abnormalities
2	Exogenous diseases	Certain infectious and parasitic diseases Neoplasm, diseases of the blood and blood-forming Organs and certain disorders involving the immune mechanism Endocrine, nutritional and metabolic diseases Mental and behavioral disorders, diseases of the nervous system Diseases of the eye and adnexa, diseases of the ear and mastoid process Diseases of the circulatory system, diseases of the respiratory system Diseases of the digestive system, diseases of the skin and subcutaneous tissue Diseases of the musculoskeletal system and connective tissue Diseases of the genitourinary system, pregnancy, childbirth and the puerperium Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified External causes of morbidity and mortality

```
Class :Ontology_of_Interactions
is a sub-class of owl:Thing
  Restriction with ID a-5
    on property :has_initiator
    some values from Class :Ontology_of_Agents
is a sub-class of Restriction with ID a-6
  on property :has_receiver
  some values from Class :Ontology_of_Agents
is a sub-class of Class owl:Thing
Class :Ontology_of_Agents is a sub-class of is
a sub-class of owl:Thing
  on property :has_believes
  some values from Class :Data
is a sub-class of Restriction with ID a-8
  on property :has_desires
  some values from Class :Methods
is a sub-class of Restriction with ID a-9
  on property :has_intentions
  some values from Class :Methods
is a sub-class of Class owl:Thing
Class :Ontology_of_TASKS is a sub-class of is a
sub-class of owl:Thing
  on property :has_method
  some values from Class :Methods
is a sub-class of Class owl:Thing
is a super-class of Class :Methods
```

The ontology data is converted into agents' beliefs, which are used for further data retrieval.

4.2. Information fusion and preprocessing

4.2.1. Detection and elimination of artifacts

Data is checked for the presence of missing values and outliers. These can be caused by registration errors or misprints. First, the data sets are checked for the presence of missing values. It is discovered that 13 out of 65 factors have more than 50% of gaps, as these

Table 10
Outcomes of missing values and outliers detection.

Factor	Description	Missing values (%)	Outliers (%)	Factor	Description	Missing values (%)	Outliers (%)
X ₀	Petroleum for general use	38	0	X ₁	Petroleum liquid gases	78	0
X ₂	Petroleum for cars	38	0	X ₃	Petroleum for other purposes	40	0
X ₄	Kerosene	87	0	X ₅	Gasohol	87	0
X ₆	Fuel oil	60	0	X ₇	Production of asphalts	0	0
X ₈	Chemical demand of oxygen in water	0	0	X ₉	Biochemical demand of oxygen in water	0	0
X ₁₀	Concentration of solids in suspension	0	0	X ₁₁	Nitrites in water	0	0
X ₁₂	Hull	30	0	X ₁₃	Mercury	0	6
X ₁₄	Kaolin	0	6	X ₁₅	Salt	0	0
...	...	...	...	...	...	...	...
X ₄₈	Non-dangerous chemical wastes	6	6	X ₄₉	Dangerous chemical wastes	18	0
X ₅₀	Non-dangerous metal wastes	18	0	X ₅₁	Dangerous metal wastes	0	0
X ₅₂	Non-dangerous waste from chemical substances	12	0	X ₅₃	Dangerous waste from chemical substances	0	0
X ₅₄	Dangerous metallic and phosphorus wastes	0	0	X ₅₅	Non-dangerous metallic and phosphorus wastes	37	0
X ₅₆	Dangerous wastes of paper, glass and rubber	25	0	X ₅₇	Non-dangerous wastes of paper, glass and rubber	18	0
X ₅₈	Dangerous solid and vitrified wastes	18	0	X ₅₉	Non-dangerous solid and vitrified wastes	25	0
X ₆₀	Other non-dangerous chemical wastes	25	0	X ₆₁	Number of tractors	25	0
X ₆₂	Number of lorries	25	0	X ₆₃	Number of buses	25	0
X ₆₄	Number of cars	25	0	X ₆₅	Number of motorcycles	25	0
...	...	...	...	...	...	...	...
Y ₀	Total number of diseases	12	0	Y ₁	Certain infectious and parasitic diseases	12	0
Y ₂	Neoplasm	0	6	Y ₃	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	0	6
Y ₄	Endocrine, nutritional and metabolic diseases	8	0	Y ₅	Mental and behavioral disorders	0	0
...	...	...	...	...	...	...	...

factors have not been registered until several years ago. For example, the data for pollutants “Solids in suspension” and “Nitrites” are only available since 1996, and data for some types of wastes such as “Dangerous wastes of paper and carton” and “Dangerous chemical wastes” have no records from 1989 to 1998. As a result, the number of valid pollutants for further processing decreases from 65 to 52. These data sets are excluded from the analysis. Next, pollution indicators are checked for the presence of outliers. The results are provided in Table 10. The human health indicators appear to be more homogeneous, and there are more data sources containing information of interest. These data do not contain missing values.

4.2.2. Filling of missing values

Data sets are checked for the presence of artifacts. The method of outlier detection, based on using the interquartile range is used to identify artifacts. This method uses the interquartile range (IQR), which is a measure of variability and is calculated as $IQR = Q3 - Q1$, and represents the spread of the middle 50% of the data. A value is an artifact if:

- It is lower than $(Q1 - 1.5 (IQR))$.
- It is higher than $(Q3 + 1.5 (IQR))$.

where $Q1$ is a 25th percentile, $Q3$ is a 75th percentile. If a value is an artifact, it is eliminated from the data set. As a result, data sets contain missing values after artifact detection. The presence of missing values skews the data and may lead to incorrect or unreliable conclusions. In the current study, some data sets have a lot of gaps. In order to fill missing values, the “golden ratio” method is used. The golden ratio is 0.62% of the previous value and 0.38% of the next value to the gap. The bar chart provided in Fig. 3 visualizes

some data sets before (in red) and after (in blue) the filling gaps procedure.

4.2.3. Smoothing the results

The reason to apply smoothing is to homogenize the data after the management of missing values. Exponential smoothing with

Fig. 3. Bar chart exemplifying the filling gaps procedure: data before (in red) and after (in blue) filling the gaps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

coefficient α equal to 0.15 is used, as this value of α provides a “light” smoothing.

4.2.4. Normalizing the results

Data is normalized using two normalization methods: Z-score standardization and *min–max* normalization. The example from Fig. 4 shows the results of the normalization. The column “Normalized X” contains values normalized within the interval [0, 1]. The extreme values of the real data are shown below as the minimum and the maximum values of the data set.

4.2.5. The results of the correlation analysis

Unfortunately, the data sets fused from the data sources are short and only contain 32 values. Owing to this fact non-parametric correlation coefficients are calculated using Kendall's τ statistic. For the given data sets the critical value of Kendall's τ is equal to 0.6226. The correlation analysis has shown the following results. In the same data pool the variables correlating significantly with morbidity from “Neoplasm” for the age group “under 1 year” are “Water characteristics” and “Wastes from used equipment”; for the age group “more than 85 years”, apart from the same factors, are “Wastes: non-dangerous and dangerous chemical waste” (Kendall's τ –0.726 and –0.983); for the age group “more than 1 and less than 4”, significant correlation is found with “Usage of petroleum products: gases” (Kendall's τ –0.650), with “Wastes” (Kendall's τ –0.850 and –1.0).

In relation to endogenous diseases, the outcomes of the “Certain conditions originating in the perinatal period” correlation demonstrate relations with “Water characteristics” (Kendall's τ –0.650), “Principal miner products” (Kendall's τ –0.750), “Non-dangerous waste from chemical substances” (Kendall's τ –0.733), and “Metallic and phosphorus wastes” (Kendall's τ –0.750). The data from the class “Congenital malformations, deformations and chromosomal abnormalities” correlates with “Principal miner products” (Kendall's τ –1.0), “Dangerous wastes of paper, glass and rubber” (Kendall's τ –1.0), and with “Dangerous solid and vitrified wastes” (Kendall's τ –0.767).

The closer the Kendall's τ value to 1, the stronger the agreement between the two rankings of the analyzed variables. The closer the Kendall's τ value to –1, the stronger the disagreement is between the two rankings of the analyzed variables. For these reasons, as their the Kendall's τ coefficients' values are greater than the critical value 0.6226, the calculated values of non-parametric correlation for the above mentioned variables proves the existence of statistical similarities between them. During the testing, an interface window is created, which contains the final results of data cleaning. Fig. 4 shows the changes in data after the detection of outliers, filling gaps, and normalization (above). The bar chart shows the data set before and after filling the gaps.

In conclusion, knowledge about the domain of interest is successfully retrieved from the OWL-file. Next, data is successfully retrieved from various containers, and, as a result, agents' beliefs are filled with data. With regards to data quality and their initial preprocessing, several data sets related to pollutants appear to have many missing values and outliers (between 40 and 87%). Therefore, the factors that correspond to those data sets are eliminated from the study. Elimination of these factors, which are the components of the *Environmental pollution–Human health* system, does not significantly affect the description of the system. The factors eliminated are represented by so scarce data, which cannot be interpolated, and, hence, the presence of these factors in the study can only impede correct data treatment.

Thus, the data sets that contain less number of gaps are preprocessed. Firstly, the outliers are eliminated. Second, missing values are filled by using the “gold ratio” method. Third, data sets are smoothed with exponential smoothing, and, finally, normalized. As

a result of these procedures, the initial information is transformed and prepared for knowledge discovery.

4.2.6. Decomposition results

The decomposition of the studied complex system, *Environmental pollution–Human health*, is carried out by means of correlation analysis. As correlation between variables can impede the correct execution of data mining procedures and lead to false results, a set of non-correlated independent variables **X** for each dependent variable **Y** are created. The independent variables (pollutants) that offer insignificant correlation with the dependent variable (disease), are also included into the set. The mutual correlation between the variables of a model is also studied. The variables that have a correlation coefficient greater than 0.7 are marked for exclusion from the model. This procedure is applied for regressions, artificial neural networks, and so on. The results are offered in Table 11.

4.3. Knowledge discovery results

The ADSS recovers data from plain files which contain the information about the factors of interest and pollutants. The files are fused in agreement with the ontology of the problem area. Some necessary changes of data properties (scalability, etc.) and their preprocessing are assumed. The ADSS has a wide range of methods and tools for modeling, including regression, neural networks, GMDH, and hybrid models. The Function Approximation agent selects the best models. These models include simple regression—43 models; multiple regression—24 models; neural networks—4098 models; GMDH—1409 models. The selected models are included into the committee machines. Next, the values for diseases and pollutants are extrapolated for the period of 10 years with a six month step. This extrapolation allows visualizing the dynamics of the factors and detecting if their values overcome the critical levels. A control under the “significant” factors, which impact health indicators, could decrease some types of diseases.

As a result, the MAS provides all the necessary steps for the standard decision making procedure by using intelligent agents. The levels of the system architecture, logically and functionally connected, have been presented. Real-time interaction with the user provides a range of possibilities in choosing one course of action from several alternatives, which are generated by the system through guided data mining and computer simulation. The system is designed for regular usage to achieve adequate and effective management by responsible municipal and state government authorities.

Also, both traditional data mining techniques and other hybrid and specific methods, with respect to data nature (incomplete data, short data sets, etc.) were used. The combination of different tools enabled us to gain quality and precision of the reached models, and, hence, in making recommendations, which are based on these models. The received dependencies of interconnections and associations between the factors and dependent variables help to correct recommendations and avoid errors.

4.3.1. Regression models

For every class of disease, plotting morbidity value against one or several pollutants, simple and multiple, linear and non-linear, regressions are performed. As a result, regression models are created of least-squared, power, exponential and hyperbolic types. Each model is evaluated with Fisher *F*-value. The models that do not satisfy the *F*-test are eliminated from the list of accepted models. The critical *F*-value for $(m-1)=2-1=1$ and $2(n-1)=2(16-1)=30^\circ$ of freedom is 4.35.

Generally, the number of accepted regression models is low. Indeed, the predictability of the best performing univariate regression models range from 0.48 to 0.82 for the discrimination

Fig. 4. Outcomes of data clearing: elimination of outliers, filling gaps and normalization, where X1 is the data before and X2 is the data after filling the gaps.

Table 11
Decomposition results.

	Dependent variable	Independent variables
1	Y_1	$X_{27}, X_{35}, X_{39}, X_{40}, X_{42}, X_{54}, X_{59}, X_{60}, X_{61}, X_{62}, X_{63}, X_{64}$
2	Y_7	$X_{28}, X_{29}, X_{30}, X_{32}, X_{33}, X_{36}, X_{37}, X_{40}, X_{42}, X_{48}, X_{54}, X_{55}, X_{57}, X_{60}, X_{61}, X_{62}, X_{63}, X_{64}$
3	Y_{20}	$X_{21}, X_{24}, X_{26}, X_{27}, X_{28}, X_{29}, X_{30}, X_{31}, X_{33}, X_{35}, X_{37}, X_{38}, X_{39}, X_{40}, X_{42}, X_{44}, X_{54}, X_{55}, X_{50}, X_{60}, X_{61}, X_{62}$
4	Y_{35}	$X_8, X_9, X_{12}, X_{60}, X_{61}, X_{62}, X_{63}, X_{64}$
5	Y_{96}	$X_{592}, X_{60}, X_{61}, X_{62}, X_{63}, X_{64}$
6	Y_{100}	$X_{26}, X_{27}, X_{28}, X_{29}, X_{30}, X_{31}, X_{33}, X_{35}, X_{37}, X_{38}, X_{39}, X_{40}, X_{42}, X_{49}, X_{54}, X_{55}, X_{59}, X_{60}, X_{61}, X_{62}$
7	Y_{181}	X_6, X_{61}, X_{64}

coefficient. Figs. 5 and 6 show examples of regression models and approximation to real data. The model shown in Fig. 5 is a univariate regression model that constructs the function $Y_0 = f(X_1)$ and is equal to $Y_0 = 6.42X_1 - 0.068$. The red line represents the initial data, and the blue line represents the data approximated with the model. A visual analysis shows that the model does not fit well with the data set. The same conclusion is drawn by analyzing its statistical criteria: the correlation coefficient $R = 0.48$, the determination coefficient $D = 0.23$ and the F-criterion $F = 4.4$. The regression model shown in Fig. 6, which models the dependency $Y_{14} = f(X_{44})$ and has the form $Y_{14} = 4.43X_{44} - 0.144$, shows better results in fitting the initial line, as well as proving the statistical criteria: the correlation coefficient $R = 0.82$, the determination coefficient $D = 0.68$ and the F-criterion $F = 30.09$. In general, univariate regression models for the current case study are characterized with low values of statistical

indicators and cannot be used for modeling. Nonetheless, multiple regression models show better performance results. For example, the multiple regression model for the Y_{15} is given in Fig. 7.

The model is written as $Y_{15} = 0.022X_{14} + 0.001X_4 + 0.012$ and its statistical criteria for this model are: the correlation coefficient $R = 0.77$, the determination coefficient $D = 0.59$ and the F-criterion $F = 20.69$. The meaning the explanatory variables, X_4 and X_{14} , explains the dependent variable Y_9 in 59% of cases. In other words, in 59% of cases the model would give a correct result, and in 41% of cases the model would give an incorrect result.

4.3.2. Neural network models

Neural network-based models, calculated for the experimental data sets, have shown high performance results. The Encog library [15] is used to create and to train neural network models. The Encog

Fig. 5. Univariate linear regression to model $Y_0 = f(X_1)$.

Fig. 6. Univariate regression to model $Y_{14} = f(X_{44})$.

Fig. 7. Multiple regression to model $Y_{15} = f(X_4, X_{14})$.

Fig. 8. An example of BP model for Y_{35} .

library contains classes which allow creating a wide variety of networks (feedforward neural networks, Hopfield neural network, Radial Basis Function network, etc.) and training them with various algorithms (backpropagation, resilient propagation, genetic algorithms, etc.). The library also includes support classes to normalize and process data for the neural networks. Networks trained with resilient propagation and with backpropagation algorithms have similar architectures, and the training and testing procedures are equivalent. Before modeling, some previous experiments are carried out. The training parameters are varied and the outputs of the neural network models are evaluated. These experiments help to determine the optimal values of the parameters. The best results are obtained from the networks with a limited number of hidden layers and neurons. In fact, the neural network with one hidden layers appears to be the optimal architecture for working with short data sets, which is the case in our experiments. For feedforward networks trained with the backpropagation algorithm the values of the learning rate and the momentum are varied within the interval $[0, 0.99]$. The best results are obtained for the learning rate within the interval $[0.85, 0.99]$ and the momentum within range $[0.3, 0.4]$. An example is shown in Fig. 8. Feedforward neural networks trained with the resilient propagation training algorithm show a high performance with the zero tolerance equal to 10^{-15} , the initial update value within the range $[0.05, 0.15]$ and the maximum step equal to 50.

Cross-validation is used to evaluate the performance of the neural networks. The first sample is used to fit a model and the second one is used to estimate the expected discrepancy of the model. With respect to the selection of the number of observations in the validation data set m , there is a practical recommendation that supposes 75 and 25% for training and validation data sets, respectively [8]. Figs. 9–12 represent training error functions for some

Fig. 10. Neural network trained with backpropagation algorithm, learning rate 0.9, momentum 0.3, number of epochs 5500.

Fig. 11. Neural network trained with resilient propagation algorithm, zero tolerance 10^{-15} , initial update 0.1, and maximum step 50.

neural network models for the variable Y_0 . The models are neural networks trained with backpropagation algorithm and neural networks trained with resilient propagation algorithm. The learning parameters of learning are provided in the legends and the activation function is the sigmoid function. The stopping criteria are double: the training error should not exceed 0.1 and the number of epochs should not be less than 5500.

Also neural networks are trained with genetic algorithms using training sets with the following parameters:

- population size—the size of population, used for training,
- mutation percent—the percent of the population to which the mutation operator is applied,
- percent to mate—the part of the population to which the crossover operator is applied.

In fact, weight optimization occurs after several populations are created; so, the training error curves have a step-like form. The error curves help to set the parameters for GANN training optimal. The determination coefficient is taken as a fitness function. Fig. 13 shows how this form of error training curve changes depending on the values of the parameters mentioned above.

The analysis of the evaluation criteria shows that GANN provides better results for the given type of data series. The variation of the number of iterations (900 and 1500 population created, see Fig. 13(a) and (b)) influences on the value of the training error (≈ 0.30 versus ≈ 0.25). Various values for the population size used for training and mutation percent are experimented. The charts (c) and (d) on Fig. 13 show that the combination of a high percent of population used for training (0.7) and for mutation (0.7) does not

Fig. 9. Neural network trained with backpropagation algorithm, learning rate 0.95, momentum 0.4, number of epochs 5500.

Fig. 12. Neural network trained with resilient propagation algorithm, zero tolerance 10^{-15} , initial update 0.1, and maximum step 50.

Fig. 13. Error functions for neural networks training with genetic algorithms with different parameters. The chart training parameters are the population size, the mutation percentage, and the percent to mate, respectively. The parameters for each chart are: (a) 900, 0.1, 0.3; (b) 1500, 0.1, 0.3; (c) 1500, 0.7, 0.7; (d) 1500, 0.3, 0.2.

significantly affect the training error. However, the combination of a relatively low percent of population used for training (0.3) and for mutation (0.2) offers better performance indicators for the models. In fact, the training error is reduced to the value of ≈ 0.105 .

4.3.3. Models received with the group method of data handling

An important feature of the iterative GMDH algorithm is its ability to identify both linear and non-linear polynomial models by using the same approach. The results of the GMDH-modeling and the best approximation models obtained are provided in Table 12. The best results are obtained by models number 3 and 1. In general, GMDH-based models obtain high performance results and efficiency when working with short data sets. The models are received with a combinatorial algorithm, where the combination of the following polynomials are used: X , X^2 , X_1X_2 , $X_1X_2^2$, $X_1^2X_2$, $X_1^2X_2^2$, $1/X$, $1/(X_1X_2)$. The choice of these polynomial terms are used in agreement with well-known recommendations [21]. More GMDH algorithms and polynomial terms of higher order will be used in future enhancements of the system. The selection of the model is stopped when the regulation criterion starts to reduce.

The GMDH model provided in Fig. 14 is created for variable Y_{23} Disease: Mental behavior and disorders, age group: under 1 year. and has the form $Y_{23} = 4.153X_{42}^2 + 1.156X_2^2 - 2.014$. The statistical features of this model are the correlation coefficient $R=0.95$, the determination coefficient $D=0.91$ and the F-criterion $F=50.825$. This model fits well the experimen-

Fig. 14. The GMDH model for Y_{23} .

tal data and has high values of statistical parameters. The model $Y_{14} = 5.292X_{64}^2 + 0.161X_{60}^2 - 0.813$ for the variable Y_{14} . Disease : Certain conditions originating in the prenatal period.

Age group : all the ages, which is shown in Fig. 15, obtains very high statistical performance results (see Table 12).

The visual analysis shows that this model fits the initial data almost perfectly. Models with similar characteristics as the one shown in Fig. 15 are candidates to be included into the hybrid committee for the given variable.

4.3.4. Model selection

After the modeling, several models for each variable of interest are created, and the best evaluated models are selected to be included into a committee machine. Table 13 shows the results of the accepted models for variable Y_0 “Total morbidity from respiratory diseases”. The columns are: R —correlation coefficient, D —determination coefficient, and F —Fisher criterion, where RM stands for regression models, GMDH stands for group method of data handling models, RPROP, BP and GANN stand for neural networks models, trained with resilient propagation, backpropagation and genetic algorithms, respectively. In accordance with Table 13, neural network models and GMDH-models get a better performance than regression models. Consider the values for determination coefficient D for regression models 0.46 and 0.30, and for neural networks trained with RPROP algorithm 0.91 and 0.81, for BP-networks 0.77 and 0.82, and for GANN 0.76 and 0.79, respectively.

In general, neural networks models and models created with GMDH show better results for all the variables. Thus, 32 models are created for the variable Y_0 = “Total morbidity from respiratory diseases”, where 28 are neural networks, 1 GMDH model and 3 regression models, and for the variable Y_{35} = “External causes of death” the number of models is 37, where 25 are neural networks, 8 multiple regression models, 1 GMDH and 3 simple regressions.

With respect to the execution time of the different data mining methods, there is no doubt that the most rapid methods are regression methods. The time for generation of GMDH-models greatly depends on the data samples and the order of the polynomials used for model creation. On the other hand, neural network models, that is BP and RPROP models, learn in a fast way. Finally, GANN-based

Fig. 15. The GMDH model for Y_{14} .

Table 12
GMDH-models: several examples (best results are provided in bold).

	Model	R	D
1	$Y_{23} = 4.153X_{42}^2 + 1.156X_{22}^2 - 2.014$	0.95	0.92
2	$Y_{14} = 5.44X_{31} + 0.171X_{34} + 0.26X_{61} - 0.727$	0.67	0.44
3	$Y_{14} = 5.292X_{64}^2 + 0.161X_{60}^2 - 0.813$	0.98	0.98
4	$Y_7 = 3.704X_{28} + 0.301X_{36} + 0.29X_{38} - 1.197$	0.54	0.29
5	$Y_9 = 4.67X_{28} - 0.051X_{30} - 0.222X_{31} - 0.54X_{36} - 0.259$	0.77	0.60
6	$Y_{24} = 0.01X_1X_0 - 69.01$	0.54	0.30

Table 13
Models for disease class Y_0 = “Total morbidity from respiratory diseases” and for age groups “All ages” and “Less than 1 year.”

Model type	Model	R	D	F
Age group “All ages”				
RM	$Y_0 = 4.12X_5 - 0.81$	0.68	0.46	12.2
GMDH	$Y_0 = 4.18X_5 - 0.019X_2 - 0.051X_6 - 0.025$	0.76	0.57	5.30
RPROP	$Y_0 = f(X_1, X_2, X_4, X_7)$	0.95	0.91	27.9
BP	$Y_0 = f(X_3, X_5, X_6)$	0.88	0.77	13.73
GANN	$Y_0 = f(X_0, X_2, X_6)$	0.87	0.76	12.45
Age group “Less than 1 year”				
RM	$Y_0 = 4.11X_5 - 0.52$	0.55	0.30	5.98
GMDH	$Y_0 = -0.11X_0 - 0.08X_6^2 - 0.05X_5^2 + \frac{6.67}{X_4} + 7.78X_5 - 1.73$	0.85	0.72	5.41
RPROP	$Y_0 = f(X_0, X_1, X_2, X_3)$	0.90	0.81	11.72
BP	$Y_0 = f(X_4, X_5, X_6)$	0.91	0.82	18.22
GANN	$Y_0 = f(X_1, X_3, X_4)$	0.89	0.79	15.05

models require more execution time. Evidently, the simultaneous execution of these methods in concurrent mode forces a significant reduction of execution time.

4.3.5. Committee machines

A committee machine is tested as a final model for every variable. Just as an example of a committee machine, the outputs of modeling for the variable of interest Y_{35} = “Disease: External causes of death. Age group: all ages” are discussed. First, after the decomposition of the number of variables (pollutants) that could be included, the models of interest for Y_{35} include the following factors: $X_8, X_9, X_{12}, X_{60}, X_{61}, X_{62}, X_{63}, X_{64}$. Several models that include these factors are created for the variable, Y_{35} , and are then evaluated. The models with higher values of the correlation coefficient R , the determination coefficient D are selected. These are the best models gotten:

1. Multiple regression model $Y_{35} = f_1(X_9, X_{61})$.
2. Neural network trained with backpropagation algorithm $Y_{35} = f_2(X_8, X_{63}, X_9)$ (see Fig. 16).
3. Neural network trained with RPROP algorithm $Y_{35} = f_3(X_{60}, X_{62}, X_{12})$ (see Fig. 16).
4. Neural network trained with genetic algorithms $Y_{35} = f_4(X_{64}, X_{12})$.

The final model generated by the committee machine is:

$$Y_{35} = \frac{f_1(X_9, X_{61})R_{f_1} + f_2(X_8, X_{63}, X_9)R_{f_2} + f_3(X_{60}, X_{62}, X_{12})R_{f_3} + f_4(X_{64}, X_{12})R_{f_4}}{R_{f_1} + R_{f_2} + R_{f_3} + R_{f_4}} \quad (1)$$

where f_i is a model, included into the committee machine, and R_{f_i} is the correlation coefficient for the i -th model, $i \in [0, \dots, n]$, where n is the number of models.

Fig. 16 provides a graphical representation of the models. The factual information covers 28 years, which are given with a six months step. It starts at “0” and finishes at “27.5”. The forecast is made for 10 years, and includes the marks starting from “28” and finishing with “37.5”. The autoregressive neural networks models for all the factors from the formula of the committee machine (see (1)) are calculated to perform the forecast. The autoregressive model is a formula that predicts an output $y(n)$ of a system based on the previous outputs, $y(n-1), y(n-2) \dots$, and inputs, $x(n), x(n-1),$

$x(n-2) \dots$. For the current case, each autoregressive model is calculated as $x(t) = f(x(t-1), x(t-2), \dots, x(t-4))$, where t represents time, and has values $(1, 2, \dots, n)$. n is the length of the data set and $x(t)$ is the value of the factor at step t . Furthermore, each autoregressive neural network model belongs to the feedforward type, and is trained with the RPROP algorithm. Its structure includes an input layer with five input neurons, a hidden layer with three or four neurons, and an output layer with one neuron. When the predictions for the factors from the formula of the committee machine are obtained, they are used to calculate the forecast for Y_{35} . The models show similar results that do not vary too much. In accordance with the forecast, the morbidity from external causes has a tendency to decrease (Fig. 17).

For the period of prediction, all the models throw similar forecasts that which are not strongly dispersed. This similarity in predictions by different models proves the tendency of the situation development. The outputs obtained by the committee machine are marked with red and, in accordance with Eq. (1), the response is a composition of the best models.

Fig. 16. Accepted models for the variable Y_{35} “External causes of death”, age group “under 1 year”. Approximation of real data by BP-trained (above) and RPROP-trained (below) neural networks.

Fig. 17. Models for variable Y_{35} and prognosis for the determined period. Dependent variables are $X_8, X_9, X_{12}, X_{60}, X_{61}, X_{62}, X_{63}$ and X_{64} . The data received by the committee machine is in blue, the data received by the neural network trained with resilient propagation algorithm is in red, the data received by the neural network trained with backpropagation algorithm is in green, the data received by the neural network trained with genetic algorithms in yellow, and the data received by the multiple regression model is in magenta. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 14

Table with the results of impact assessment for selected diseases.

	Disease class	Pollutant, which influence upon the disease
1	Neoplasm	Nitrites in water; miner products; BOD5; asphalts; dangerous chemical wastes; fuel-oil; petroleum liquid gases; water: solids in suspension; non-dangerous chemical wastes
2	Diseases of the blood and blood-forming organs, the immune mechanism	BOD5; miner products; fuel-oil; nitrites in water; dangerous wastes of paper industry; water: solids in suspension; dangerous metallic wastes
3	Pregnancy, childbirth and the puerperium	Kerosene; petroleum; petroleum autos; petroleum liquid gases; gasohol; fuel oil; asphalts; water: COD; BOD5; solids in suspension; nitrites in water
4	Certain conditions originating in the prenatal period	Non-dangerous wastes: general wastes; mineral, construction, textile, organic, metal wastes. Dangerous oil wastes
5	Congenital malformations, deformations and chromosomal abnormalities	Gasohol; fuel-oil; COD in water; producing asphalts; petroleum; petroleum autos; kerosene; petroleum liquid gases; water: BOD5, nitrites, solids in suspension

4.3.6. Environmental impact assessment results

The impact assessment shows the dependencies between water characteristics and neoplasm, complications of pregnancy, childbirth and congenital malformations, and deformations and chromosomal abnormalities. Table 14 shows the output of impact assessment for several variables of interest (classes of diseases), which proves that within the most important factors, apart from water pollutants, there are indicators of petroleum usage, miner output products and some types of wastes.

The results are offered in Table 14. The results are compared to other studies. Thus, in [22] it is reported about the causal effect of traffic pollution upon respiratory illness. In our research we have come to similar conclusions, as we have detected a presence of links between such indirect indicators (lorries, motorcycles, cars, etc.) and diseases of the respiratory system. As it is reported in [2], water pollution may lead to cancer, and we have discovered similar relations through impact assessment evaluation (see Table 14).

5. Conclusions

The proposed ADSS for the study of the “Environment–Public health” complex system includes all the necessary stages of the decision making process, leaving to a user the possibility to simulate possible outputs and to take a weighted decision based on previous analysis and expert knowledge. The ADSS has a strictly

organized structure which, however, can be modified and amplified by user requirements. The fundamental scenarios retrieve and fuse data, preprocess data, check environmental impact, create models, and create recommendation provide the execution of all the data mining procedures carried out by the system and facilitate reaching the aims. In other words, the proposed system may be used as a “black box” system for non-specialists in data processing, or as a “white-box” system for specialists who would modify it in agreement with their needs.

Hierarchical agents which have their teams of subordinate agents are used. This decision facilitates distributing the control all over the system. Moreover, in case there is a need to add some additional functionality to any role, often it can be solved by adding a plan or a capability to the subordinate or principal agent. The implementation is performed within the JACK Development Environment. Some plug-ins are used to code data fusion procedures when the ontology is read from the OWL-files, and neural networks models and visualized results are generated.

Practical results and some theoretical outcomes have been described from an experiment for the region of Castilla-La Mancha (Spain). Retrospective information has been fused and preprocessed. The results of data cleaning has been provided and explained in detail with graphics and numerical results. Regression, neural networks, and GMDH-models have been created and then the best model for each variable of interest has been included into

committee machines. The example described in the article shows a committee machine for the “External causes of death” variable together with the simulation results obtained. The procedure of the impact assessment has helped to discover the hidden dependencies between health outcomes and environmental pollutants which were represented by indirect indicators.

The obtained results can help in better understanding the possible hazards for regional public health, and could be used to correct and to improve politics of public health institutions.

Acknowledgements

This work was partially supported by Spanish Ministerio de Ciencia e Innovación TIN2010-20845-C03-01 grant, and by Junta de Comunidades de Castilla-La Mancha PII2I09-0069-0994 and PEII09-0054-9581 grants.

References

- [1] A. Aschengrau, J. Weinberg, P. Janulewicz, L. Gallagher, M. Winter, V. Vieira, T. Webster, D. Ozonoff, Prenatal exposure to tetrachloroethylene-contaminated drinking water and the risk of congenital anomalies: a retrospective cohort study, *Environmental Health* 8 (1) (2009) 814–830.
- [2] I. Bröske-Hohlfeld, Environmental and occupational risk factors for lung cancer, *Methods in Molecular Biology* 472 (2008) 3–23.
- [3] Q. Cao, M.E. Parry, Neural network earnings per share forecasting models: a comparison of backward propagation and the genetic algorithm, *Decision Support Systems* 47 (1) (2009) 32–41.
- [4] F. Castro, A. Nebot, F. Mugica, On the extraction of decision support rules from fuzzy predictive models, *Applied Soft Computing* (2011).
- [5] V.Y.C. Chen, H.P. Lien, C.H. Liu, J.J.H. Liou, G.H. Tzeng, L.S. Yang, Fuzzy MCDM approach for selecting the best environment-watershed plan, *Applied Soft Computing* 11 (1) (2011) 265–275.
- [6] J.M. Corchado, J. Bajo, Y. de Paz, D.I. Tapia, Intelligent environment for monitoring Alzheimer patients, agent technology for health care, *Decision Support Systems* 44 (2) (2008) 382–396.
- [7] O. Córdón, A. Fernández-Caballero, J.A. Gámez, F. Hoffmann, The impact of soft computing for the progress of artificial intelligence, *Applied Soft Computing* 11 (2) (2011) 1491–1492.
- [8] P. Giudici, S. Figini, *Applied Data Mining for Business and Industry*, John Wiley & Sons, 2009.
- [9] W. Fan, P. Pathak, M. Zhou, Genetic-based approaches in ranking function discovery and optimization in information retrieval—a framework, *Decision Support Systems* 47 (4) (2009) 398–407.
- [10] E. Feigenbaum, P. McCorduck, *The Fifth Generation*, Pan Books, London, 1984.
- [11] A.G. Greenwood, S. Vanguri, B. Eksioğlu, P. Jain, T.W. Hill, J.W. Miller, C.T. Walden, Simulation optimization decision support system for ship panel shop operations, in: *Proceedings of the 37th Winter Simulation Conference (WSC)*, vol. 1, 2005, pp. 2078–2086.
- [12] V. Haley, T. Talbot, H. Felton, Surveillance of the short-term impact of fine particle air pollution on cardiovascular disease hospitalizations in New York State, *Environmental Health* 8 (1) (2009) 42–52.
- [13] I. Holman, M. Rounsevell, S. Shackley, P. Harrison, R. Nicholls, P. Berry, E. Audsley, A regional, multi-sectoral and integrated assessment of the impacts of climate and socio-economic change in the UK, *Climatic Change* 71 (1) (2005) 9–41.
- [14] International Classification of Diseases, 2008, <http://www.who.int/classifications/icd/en/> [accessed on July 19, 2008].
- [15] Encog home page. Version from 20 of November, 2009, <http://www.heatonresearch.com/encog> [accessed on June, 2009].
- [16] Instituto de Estadística de Castilla-La Mancha, 2008, <http://www.ine.es/en/welcome.en.htm> [accessed on July 19, 2008].
- [17] R.J. Kuo, L.M. Lin, Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering, *Decision Support Systems* 49 (4) (2010) 451–462.
- [18] N. Lavrac, E. Keravnou, B. Zupan, Intelligent data analysis in medicine, *Encyclopedia of Computer Science and Technology* 42 (2000) 113–157.
- [19] Y.F. Li, S.H. Ng, M. Xie, T.N. Goh, A systematic comparison of metamodeling techniques for simulation optimization in decision support systems, *Applied Soft Computing* 10 (4) (2010) 1257–1273.
- [20] O.R. Liu-Sheng, Decision support for healthcare in a new information age, *Decision Support Systems* 30 (2) (2000) 101–103.
- [21] H.R. Madala, A.G. Ivakhnenko, *Inductive Learning Algorithms for Complex Systems Modeling*, CRC Press Inc., 1994.
- [22] E. Migliore, G. Berti, C. Galassi, N. Pearce, F. Forastiere, R. Calabrese, L. Armenio, A. Biggieri, L. Bisanti, M. Bugiani, E. Cadum, E. Chellini, V. Dell’Orco, G. Giannella, P. Sestini, G. Corbo, R. Pistelli, G. Viegi, G. Ciccone, Respiratory symptoms in children living near busy roads and their relationship to vehicular traffic: results of an Italian multicenter study SIDRIA 2, *Environmental Health* 8 (1) (2009) 27–42.
- [23] H. Orru, E. Teinmaa, T. Lai, T. Tamm, M. Kaasik, V. Kimmel, K. Kangur, E. Merisalu, B. Forsberg, Health impact assessment of particulate pollution in Tallinn using fine spatial resolution and modeling techniques, *Environmental Health* 8 (1) (2009) 98–105.
- [24] E.I. Papageorgiou, A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques, *Applied Soft Computing* 11 (1) (2011) 500–513.
- [25] D.J. Power, R. Sharda, Model-driven decision support systems: concepts and research directions, *Decision Support Systems* 43 (3) (2007) 1044–1061.
- [26] M.V. Sokolova, A. Fernández-Caballero, F.J. Gómez, Agent-based interdisciplinary framework for decision making in complex systems, in: *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2, 2010, pp. 96–103.
- [27] M.V. Sokolova, A. Fernández-Caballero, Modeling and implementing an agent-based environmental health impact decision support system, *Expert Systems with Applications* 36 (2) (2009) 2603–2614.
- [28] M.V. Sokolova, A. Fernández-Caballero, Multi-agent-based system technologies in environmental issues, *Information Technologies in Environmental Engineering* (2009) 549–562.
- [29] M.V. Sokolova, A. Fernández-Caballero, A multi-agent architecture for environmental impact assessment: information fusion, data mining and decision making, in: *Proceedings of the Ninth International Conference on Enterprise Information Systems*, vol. DISI, 2007, pp. 219–224.
- [30] M.V. Sokolova, A. Fernández-Caballero, An agent-based decision support system for ecological-medical situation analysis, *Nature Inspired Problem-Solving Methods in Knowledge Engineering* (2007) 511–520.
- [31] M.V. Sokolova, A. Fernández-Caballero, A meta-ontological framework for multi-agent systems design, *Nature Inspired Problem-Solving Methods in Knowledge Engineering* (2007) 521–530.
- [32] L. Sterling, K. Taveter, *The Art of Agent-Oriented Modeling*, The MIT Press, 2009.
- [33] D. Stieb, M. Szyszkowicz, B. Rowe, J. Leech, Air pollution and emergency department visits for cardiac and respiratory conditions: a multi-city time-series analysis, *Environmental Health* 8 (1) (2009) 75–90.
- [34] N. White, J. teWaterNaude, A. van der Walt, G. Ravenscroft, W. Roberts, R. Ehrlich, Meteorologically estimated exposure but not distance predicts asthma symptoms in schoolchildren in the environs of a petrochemical refinery: a cross-sectional study, *Environmental Health* 8 (1) (2009) 45–59.
- [35] World Health Organization, *Protecting Health from Climate Change: Global Research Priorities*, WHO Press, 2009.
- [36] K. Yada, E. Ip, N. Katoh, Is this brand ephemeral? A multivariate tree-based decision analysis of new product sustainability, *Decision Support Systems* 44 (1) (2007) 223–234.
- [37] M.H. Zack, The role of decision support systems in an indeterminate world, *Decision Support Systems* 43 (4) (2007) 1664–1674.