EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

A Mathematical Model for Numerical Simulation of Shallow Water Flow: Description and Practical Application of GUAD|2D

Javier Murillo[1], Pilar Brufau[2], Pilar García-Navarro[3], M. Rodríguez Pallarés[4], Alfonso Andrés-Urrutia[5]

Abstract

GUAD 2D is a finite volume based two-dimensional model for the numerical simulation and analysis of flood waves caused by reasons such as extreme rainfall, gradual or sudden dam break or the release of a big water tank retaining wall. GUAD 2D allows to obtain real results in, critical, subcritical or mixed flow situations over irregular topography. GUAD 2D makes possible different initial conditions, like starting from dry or wet situations, or from the results obtained in a previous simulation.

1. Introduction

Many engineering and environmental problems involve the study of unsteady water flows with solute transport and erosion/sedimentation processes. River flows, in particular, are mostly unsteady and, as they are characterized by the presence of a vertical scale much smaller than the horizontal ones, they can be described by the shallow water model (Cunge et al. 1980) which forms a set of non linear hyperbolic equations.

A great deal of work has been devoted to develop 1D and 2D numerical models for unsteady shallow flows in the last decades and various computational techniques using finite difference, finite element and finite volume methods have been reported (Cunge et al., 1980; Bellos et al., 1991; Alcrudo and García-Navarro, 1993; Sleigh et al., 1998; Bermúdez et al., 1998; García-Navarro and Vázquez-Cendón, 2000). Several numerical difficulties must be adequately treated to obtain an accurate solution without numerical errors. Zhao et al. (1994) provided a good historic revision and the features required for a two dimensional river flow simulation model: it should be able to handle complex topography, dry bed advancing fronts, wetting-drying moving boundaries, high roughness values, steady or unsteady flow and subcritical or supercritical conditions. Natural topography is the main challenge. Dominant source terms and open boundaries are two important difficulties to face when using a conservative method since they both can damage the conservative character of the solution. Bed slope and friction source terms are of special relevance in hydraulic applications based on a shallow flow model. For that reason, a considerable effort has been recently devoted to this topic in a search for the correct source term discretization given a particular numerical scheme with good properties for the homogeneous case (Leveque, 1998; Roe, 1981; Glaister, 1992; Vázquez-Cendón, 1999; Bermúdez and Vázquez-Cendón, 1994; Bermúdez et al., 1998).

Another numerical problem of relevance is the modelling of wet/dry interfaces between internal cells, that have traditionally represented a difficulty for modellers wanting to solve the shallow water equations over a bed of irregular geometry. Flow over dry bed involves a complicated situation that can be analysed as a boundary condition which is dynamically changing in time with the moving front and continuously expanding or reducing the flow domain. The alternative is to include the wet/dry interfaces in the full

[1] Fluid Mechanics, CPS, University of Zaragoza, Spain, e-mail: Javier.Murillo@unizar.es
[2] Fluid Mechanics, CPS, University of Zaragoza, Spain, e-mail: brufau@unizar.es
[3] Fluid Mechanics, CPS, University of Zaragoza, Spain, e-mail: pigar@unizar.es
[4] INCLAM, S.A. Ingeniería del Agua, Samaria, 4. 28009 Madrid, Spain, e-mail: martin@inclam.com
[5] INCLAM, S.A. Ingeniería del Agua, Samaria, 4. 28009 Madrid, Spain, e-mail: a.andres@inclam.com

domain of computation in which there may be wet cells and dry cells at the same time. In this case the numerical scheme chosen for the discretization must be able to cope with them. In general, cells being flooded or dried during the computation tend to introduce numerical instabilities in the solution, resulting for example in negative water depths or unphysical high velocities. Different approaches have been proposed to handle them (Kramer, 2001; Beffa and Connel, 2001; Heniche and Secretan, 2000; Kawahara and Umetsu, 1986; Khan, 2000; Bradford and Sanders, 2002; Brufau et al., 2002 and Brufau et al., 2004).

In this sense, GUAD 2D is a two-dimensional model for the numerical simulation and analysis of flood waves caused by reasons such as extreme rainfall, gradual or sudden dam break or the release of a big water tank retaining wall. GUAD 2D simulates through algorithms based on a finite volume method developed at the University of Zaragoza (Murillo et al. 2006). It allows to obtain real results in, critical, subcritical or mixed patterns, GUAD 2D makes possible certain initial conditions, like starting from dry or wet situations, or from a previous simulation instant. GUAD 2D has been validated using both data from real cases and over examples under laboratory conditions.

GUAD 2D has been specially designed for standard personal computers, therefore is not necessary to have very expensive or mainframes computers. Furthermore GUAD 2D relies on a friendly and intuitive interface that allows every kind of user to create a two-dimensional simulation without high computer or topology knowledge needs. GUAD 2D counts on a post-processing module called GUADView which is fed from the GUAD 2D results. GUADView can analyze those results thanks to certain specific application tools.

GUADView includes the most necessary GIS capabilities to analyse this kind of simulations making possible to visualize raster images and vector layers together in the same frame with the results of the simulation generated by GUAD 2D simplifying the right identification of flood zones.

Based on GUAD 2D outcomes, GUADView, is able to stand the depth, water level or velocity in a water layer, showing the correct and specific value on each selected cell. GUADView also can show time histories of depth at different locations all over simulation time in a selected cell, terrain and water sections and hydrographs.

GUADView makes possible to include all the simulation results in depth, level or velocity video with a standard ".avi" format. GUADView also can start the analysis before the conclusion of the simulation process, only with the help of a local area network.

Nowadays, we are developing and debugging an application named GUADCreator. GUADCreator, with a friendly interface, can create the GUAD 2D configuration simulation files. With GUADCreator the user can introduce, all the necessary data for the simulation. (GUADCreator will be totally developed in the ENVIROINFO2007 date).

2. Mathematical model

The water movement is governed by the basic laws of mass and momentum conservation under the shallow water hypothtesis. This implies a depth average process in the equations and is associated to the assumption of hydrostatic pressure vertical distribution. The formulation takes the form of a set of non linear hyperbolic equations that, in two dimensions, involve the water depth as well as the two depth averaged components of the velocity:

$$\frac{\partial \boldsymbol{U}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x}(\boldsymbol{U}) + \frac{\partial \boldsymbol{G}}{\partial y}(\boldsymbol{U}) = \boldsymbol{S}(x, y, \boldsymbol{U})$$

$$\boldsymbol{U} = \left(h, q_x, q_y\right)^T$$

$$\boldsymbol{F} = \left(q_x, \frac{q_x^2}{h} + \frac{gh^2}{2}, \frac{q_x q_y}{h} \right)^T \quad , \quad \boldsymbol{G} = \left(q_y, \frac{q_x q_y}{h}, \frac{q_y^2}{h} + \frac{gh^2}{2} \right)^T \tag{1}$$

where $q_x = uh$ and $q_y = vh$ are the unit discharges in the Cartesian directions. The variable h represents the water depth, g is the acceleration of the gravity and (u,v) are the averaged components of the velocity vector $\boldsymbol{u}$ along the x and y coordinates respectively. The source terms in the momentum equations are the bed slopes and the friction losses along the two coordinate directions,

$$\boldsymbol{S} = \left(0, gh(S_{0x} - S_{fx}), gh(S_{0y} - S_{fy})\right)^T \tag{2}$$

where,

$$S_{0x} = -\frac{\partial z}{\partial x} \, , S_{0y} = -\frac{\partial z}{\partial y} \tag{3}$$

and the friction losses in terms of the Manning's roughness coefficient, with

$$S_{fx} = n^2 u \sqrt{u^2 + v^2} \, / h^{4/3} \, , \; S_{fy} = n^2 v \sqrt{u^2 + v^2} \, / h^{4/3} \tag{4}$$

3. Numerical method

The physical domain is subdivided (discretized) in grid cells for the numerical resolution. Our method is able to work over cells of different shapes (triangles, quadrilaterals, hexagons) and the choice of the grid is not a trivial step. Structured cartesian grids are simpler and more efficient whereas unstructured triangular grids offer more geometric adaptivity and better refinement possibilities.

A cell-centred finite volume method has been formulated in GUAD2D where all the dependent variables of the system are represented as piecewise constants (first order). A discrete approximation of equation (1) is applied in every cell at a given time so that the volume integrals represent integrals over the area of the cell and the surface integrals represent the total flux through the cell boundaries. Denoting by $\boldsymbol{U}_i$ the average value of the conservative variables over the volume Ω_i at a given time, from equation (1) the following conservation equation can be written for every cell:

$$\frac{\partial \boldsymbol{U}_i}{\partial t} A_i + \oint_{\partial \Omega} (\boldsymbol{E} \cdot \boldsymbol{n}) \mathrm{d}\, s = \int_{\Omega} \boldsymbol{S} \, \mathrm{d}\, \Omega \tag{5}$$

where A_i is the area of the cell.

A mesh fixed in time is assumed and the contour integral is approached by a sum over the cell edges. In all of them, the normal flux is approximated via an upwind flux difference splitting technique

$$\oint_{\partial \Omega_i} (\boldsymbol{E} \cdot \boldsymbol{n}) \, \mathrm{d}\, s \approx \sum_{k=1}^{NE} (\delta \boldsymbol{E}_k \cdot \boldsymbol{n}_k) s_k \tag{6}$$

where k represents the edge index of the cell, NE is the total number of edges in the cell ($NE = 3$). The vector $\boldsymbol{n}_k$ is the unit outward normal to edge k, s_k is the length of the side, and $\delta \boldsymbol{E}_k \cdot \boldsymbol{n}_k$ is the numerical flux difference. Upwind schemes are based on the idea of discretizing the spatial derivates so that the information is taken from the side it comes. When the source terms are present, it has previously been shown that the flux derivates and the source terms have to be discretized in a similar manner (M.E. Hubbard et al, 2000, and P. Brufau, et al, 2002). The evaluation of fluxes and sources at the same local state is important. The mathematical properties of the hyperbolic system of equations (1) include the existence of a Jacobian matrix, $\boldsymbol{J}_{\boldsymbol{n}}$, of the flux normal to a given direction $(\boldsymbol{E} \cdot \boldsymbol{n})$ defined as

$$\boldsymbol{J}_{\boldsymbol{n}} = \frac{\partial(\boldsymbol{E} \cdot \boldsymbol{n})}{\partial \boldsymbol{U}} = \frac{\partial(\boldsymbol{F})}{\partial \boldsymbol{U}} n_x + \frac{\partial(\boldsymbol{G})}{\partial \boldsymbol{U}} n_y \tag{7}$$

making possible the generation of an approximate matrix $\boldsymbol{J}_{\boldsymbol{n}}^{*}$, whose eigenvalues $\tilde{\lambda}^m$ and eigenvectors $\tilde{\boldsymbol{e}}^m$ (Brufau et al., 2002; Hubbard and Garcia-Navarro, 2000), can be used to define the signals, and also that an upwind treatment can be done over the source terms.
The first order formulation of the upwind scheme is as follows (Murillo et al., 2005a):

$$\boldsymbol{U}_i^{n+1} = \boldsymbol{U}_i^n - \sum_{k=1}^{NE} \sum_m ((\tilde{\lambda}^m - \alpha^m - \beta^m)\tilde{\boldsymbol{e}}^m)_k^n) \frac{l_k}{A_i} \Delta t \tag{8}$$

The discretization of the bottom elevation source terms is successfully constructed when it ensures an exact balance between flux gradients and bed variations (Leveque 1998, Hubbard and García-Navarro, 2000). It has been demonstrated that, in first order finite volume schemes, if the upwind technique is applied to the flux and bottom terms, in the case of still water, equilibrium is maintained for the water level surface (Bermúdez and Vázquez, 1998; Brufau et al, 2002).

The friction term dominates over any other influence in many practical situations, in particular, in wetting/drying fronts, characterized by small values of water depth. The explicit upwind discretization of the friction source term in all cases guarantees a perfect discrete balance but imposes a limit over the maximum cell size allowable (Burguete et al., 2006). The reduction of cell size in two-dimensional cases leads to an unaffordable computational cost. On the other hand, the explicit pointwise discretization of the source term also produces numerical oscillations in the solution when the source term becomes of relevance. If the source term R is discretized in an implicit pointwise manner, (Brufau et al., 2004, Murillo et al., 2006) the numerical instabilities and requirement over the size cell disappear.

4. GUAD 2D

4.1 General Working Issues

GUAD 2D is an aplication that solves the most popular problems in two-dimensional trade models. Those models have limited simulation capabilities or are developed for flat terrain, therefore they can´t simulate switching flow regime over leeves and embankments correctly

Nowadays GUAD 2D may simulate 2.000.000 cells grid, managing the transitions from weir flow to supercritical flow regime and the opposite.

All the boundary conditions are handled by GUAD 2D like input hydrographs, constant water levels, dumps etc. allowing the user to initiate the simulation in dry/wet condition terrain. The user will be able to insert input data trough friendly interface, and also importing standard GIS format files.

To create a simulation, the model requires at least a digital model terrain, single or global rub value for each cell terrain, one or more input/output boundary condition, and the final requirement are simulation parameters, like desired length of simulation and results path to allocate the number of images that the user should previously have defined on those configuration parameters.

4.2 A Simulation Generation

Starting up GUAD 2D based on a simulation over a modelled terrain, users should follow the steps above explained.

4.2.1 Enter a Terrain

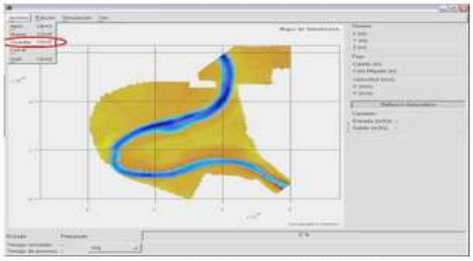

Fig. 4.2.1.1. Simulation Terrain in GUAD 2D

4.2.2 Boundary Condition

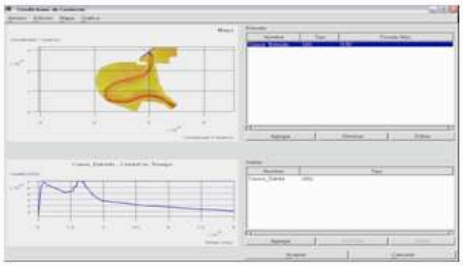

Fig. 4.2.2.1. Boundary Conditions List

4.2.3 Friction

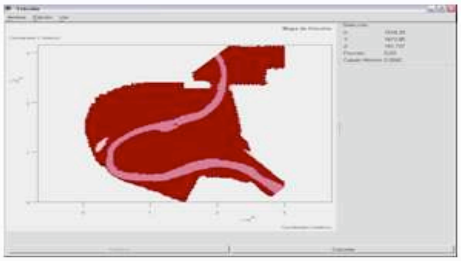

Fig. 4.2.3.1. Rub Grid for each cell

4.2.4 Simulation Parameter

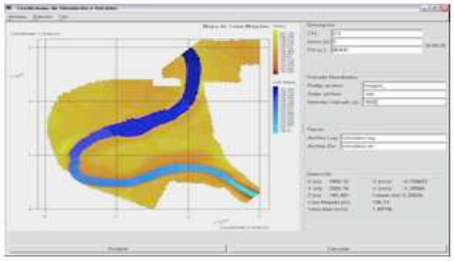

Fig. 4.2.4.1. Configuration of simulation parameter

4.2.5 Validate and Simulation Startup

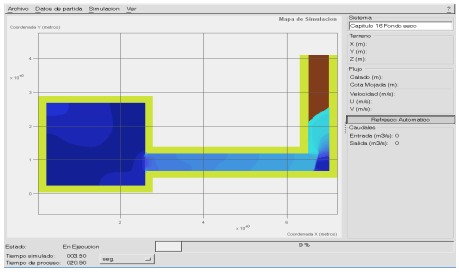

Fig. 2.2.5.1. Simulation Start up

4.3 Simulation Outcomes

GUAD 2D will generate the results in a single extension and proper format file (*.gds).
Those GUADView readable files will content all the generic information from the GUAD 2D simulation i.e.: modelled simulation terrain, results represented such as water layers showing depth, level, and U & V component of velocity. The user will be able to get this information in every moment among the simulation. Therefore this data will be identified by component value and its effective period of time.

5. GUADView

5.1 General Working Issues

GUADView is an analysis application for the generated by GUAD 2D results. GUADView could be considered as a GIS itself with specific analysis capabilities added for this kind of outcomes. Given the generated results file, GUADView will get the terrain grid simulation and all water layers associated at any modelled moment.
GUADView is able to show in different layers data contained on each grid cell. Data is composed by depth, level and module/direction component of velocity. Thanks to those GUADView capabilities, the analysis of hydraulic information layer is very fast and simple, it can show level sondes for each mesh cell in a configurable graphic from the primary to last instant simulation. Moreover GUADView allow the user to visualize terrain, and loaded water layer sections. It can also generate hydrographs in selected terrain sections.
GUADView can also create simulation videos, envelope layers, offers layer control tools and vector and raster layer managing, and exports GUAD 2D layer to commercial GIS format.

5.2 Outcomes Analysis

5.1.1 Layer Control

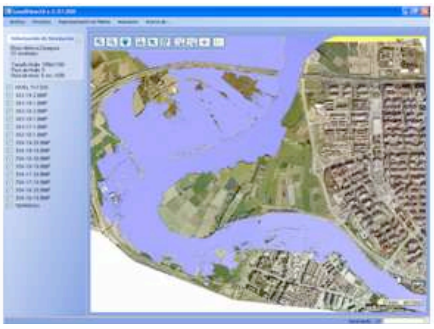

Fig. 5.1.1.1 Water Layer Visualization ovner raster

5.1.3 Hidrographs

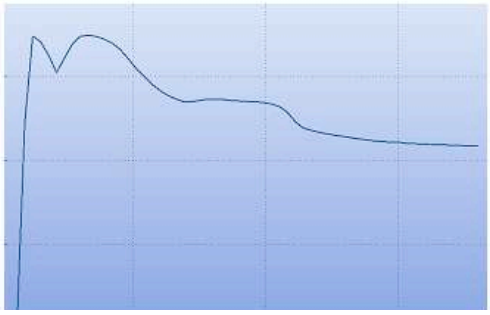

Fig. 5.1.3.1 Terrain section hidrographs

5.1.2 Sonders, Sections

Fig. 5.1.2.1 Visualize section details

5.1.4 Video Generation

Fig. 5.1.4.1 Video with GUAD 2D water leyer

6. Conclusions

As we have seen in the above mentioned paper, we can create hydraulic simulation and analyze the results through both applications easily and fast. Furthermore we could see that results obtained have an optimal and friendly analysis in most commercial GIS application. Nowadays GUAD 2D is been used in real cases and tested in several laboratory cases, thereby the results obtained have been contrasted with present hydraulic situation, offering to hydraulic community a reliable, fast and functional solution.

Bibliography

Alcrudo, F. and García-Navarro, P. (1993) A high resolution Godunov-type scheme in finite volumes for the 2D shallow water equations. Int. Journal for Num. Meth. in Fluids, 16(6): 489-505.

Beffa, C., Connel, R. (2001): Two-dimensional flood plain flow. I: Model description, J. of Hydraulic Engineering, 6(5): 397-405.

Begnudelli, L., Sanders, F. (2006): Unstructured Grid Finite-Volume Algorithm for Shallow-Water Flow and Scalar Transport with Wetting and Drying. Journal of Hydraulic Engineering, 132(4): 371-384.

Bellon, C.V., Soulis, J.V., J.G. (1991): Computation of two-dimensional dam-break induced flows. Adv. Water Res., 14(1): 31-41.

Bradford, S.F., Sanders B.F.(2002): Finite-volume model for shallow water flooding of arbitrary topography, Journal of Hydraulic Engineering, 128(3): 289-298.

Brufau, P., Vázquez-Cedón, M.E., García-Navarro, P. (2002): A numerical model for the flooding and drying of irregular domains. International Journal for Numerical Methods in Fluids. 39: 247-275.

Brufau, P., García-Navarro P. and Vázquez-Cendón M.E. (2004): Zero mass error using unsteady wetting/drying conditions in shallow flows over dry irregular topography. Int. Journal for Num. Meth. in Fluids ,45: 1047-1082.

Burguete, J., and García-Navarro, P. (2001): Efficient construction of high-resolution TVD conservative shemes for equations with source terms: application to shallow water flows. International Journal for Numerical Methods in Fluids. 37: 209-248.

Burguete J., García-Navarro P. and Murillo J. Analysis of the friction term in one-dimensional shallow water model: application to open channel and river flow. Journal of Hydraulic Engineering, (under revision).

Chow, V.T. (1959): Open channel flow. MacGraw-Hill Book Co. Inc.

Cunge, J.A., Holly, F.M. and Vervey A. (1980): Practical aspects of computational river hydraulics. Pitman.

Heniche, M., Secretan, Y., Boudreau, P. and Leclerc, M. (2000): A two-dimensional finite element drying-wetting shallow water model for rivers and estuaries, Adv. in Water Research, 23: 359-372.

Hubbard, M. E. and García-Navarro, P. (2000): Flux Difference Splitting and the Balancing of Source Terms and Flux Gradients. Journal of Computational Physics. 165: 89-125.

Kawahara, M., Umetsu, T. (1986): Finite element method for moving boundary problems in river flow, Int. Journal for Num. Meth. in Fluids, 6: 365-386.

Khan, A.A. (2000): Modeling flow over an initially dry bed, Journal of Hydraulic Research, 38(5): 383-389.

Murillo, J., García-Navarro, P., Burguete, J. and Brufau, P. (2006): A conservative 2d model of inundation flow with solute transport over dry bed. Int. Journal for Numerical Methods in Fluids, Published Online in www.interscience.wiley.com.

Sleigh, P.A., Berzins, M., Gaskell, P.H. and Wright, N.G. (1998): An unstructured finite volume algorithm for predicting flow in rivers and stuaries. Computers and Fluids, 27(4): 479-508.

Studying the Presence of Genetically Modified Variants in Organic Oilseed Rape by Using Relational Data Mining

Aneta Ivanovska[1], Celine Vens[2], Sašo Džeroski[1], Nathalie Colbach[3]

Abstract

The production of genetically-modified (GM) crops has increased rapidly over the last 10 years. The possibility of GM crops mixing with conventional or organic crops is becoming a problem and estimating the adventitious presence of GM seeds into conventional crop harvests presents a challenge. In this study we used outputs from a previously developed computer model for gene flow between GM and conventional oilseed rape to construct relational classification trees that predict the adventitious presence of GM seeds in the central field of a large-risk field pattern as a function of cultivation practices. Unlike propositional data mining methods, relational methods (relational classification trees) enable us to examine the relations among fields, for example, the influence of the neighbouring fields on the adventitious presence of GM seeds in a given field. For that purpose we used the relational data mining system TILDE.

1. Introduction

Genetically-modified (GM) crops in Europe were first introduced for commercial production in 1996. Since then the planted area has increased rapidly. The first GM varieties were engineered to tolerate herbicides and/or resist pests. Crops carrying genes coding for herbicide tolerance were developed so that farmers could spray their fields with non-selective herbicides to eliminate weeds irrespective of species and stage without damaging the crop. Pest-resistant crops have been engineered to contain a gene for a protein from the soil bacterium, *Bacillus thurigiensis*, which is toxic to certain pests. This protein, referred to as Bt, is produced by the plant, thereby making it resistant to insect pests like the European Corn Borer (*Ostrinia nubilalis*) or Cotton Boll Worm (*Helicoverpa zea*). So the main purpose of growing genetically modified crops is not to achieve higher yields, but to reduce producers' inputs and operating costs.

However, GM crops were not primarily developed with environmental benefit in mind and the introduction of transgenic crops and foods into the existing food production system has generated a number of questions about possible negative consequences. The main concern about GM crops today is the coexistence issue, i.e., the possibility of GM seeds mixing with conventional or organic crops of the same species. GM crops can contaminate other crops simply by pollen being blown by wind from one field to another. In addition, seed persistence in time and, to a lesser degree, seed immigration from neighbour fields, road margins or agricultural tools also contribute to gene flow.

To estimate the rate of adventitious presence of GM varieties in non-GM crops and to compare the effects of changing farming practises, a computer simulation model GENESYS was developed (Colbach, 2001a, 2001b). GENESYS's purpose is to rank cropping systems according to their probability of gene flow between genetically modified and non-genetically modified oilseed rape, via pollen dispersal and volunteers. GENESYS predicts the level of harvest contamination of conventional oilseed rape crops by genetically modified rape seeds.

[1] Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
e-mails: aneta.ivanovska@ijs.si, saso.dzeroski@ijs.si
Tel: +386 1 477 3144 (Aneta Ivanovska)
[2] Department of Computer Science, K.U. Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium
e-mail: celine.vens@cs.kuleuven.be
[3] UMR1210, Biologie et Gestion des Adventices, INRA, 21000 Dijon, France
e-mail : colbach@dijon.inra.fr

The dataset produced by GENESYS was previously analysed using propositional data mining techniques (Ivanovska et al., 2006), where it was shown that the cropping techniques of a field (sowing date and density) have the biggest influence on its contamination with GM content. However, our assumption was that the contamination of a field with GM seeds depends on the cropping techniques and crops grown in the surrounding fields (e.g., the level of contamination of a field may be influenced by the crop grown in or the level of contamination of its neighbouring fields). So it seems worthwhile to exploit neighbourhood relations in the predictive model and create a relational representation of the problem. Therefore, in this study we investigate the use of relational data mining to analyze a dataset produced by GENESYS. For that purpose we used the relational data mining system TILDE (Blockeel, 1998).

The remainder of this paper is structured as follows: Section 2 presents the GENESYS simulation model. Section 3 discusses the dataset used in our study, as well as the relational data mining methods used for the analysis. The experimental setup and results are presented in Section 4. Section 5 concludes and gives directions for further work.

2. The GENESYS simulation model

The computer model GENESYS is used to asses the probable effects of changing farming practices on contamination rates of oilseed rape crops by extraneous genes. GENESYS (Colbach, 2001a, 2001b) was developed by INRA (French National Institute for Agronomy Research) to rank cropping systems according to their probability of gene flow from herbicide-tolerant winter oilseed rape to rape volunteers and neighbour crops, both in time via seeds and in space via pollen and seeds. The model works for seed as well as crop production. GENESYS integrates various input variables:

- The field plan of the region, comprising cultivated fields as well as uncultivated field- and road-margins (hence "borders"). Borders consist of strips of spontaneous vegetation where rape volunteers can appear, produce pollen and seeds that are dispersed to fields and other borders,
- The crop rotation of each field,
- The cultivation techniques applied to each crop (summer tillage, primary tillage and tillage for seed bed preparation, sowing date and density, herbicide applications, cutting dates and seed loss at rape harvest) as well as the management of the borders (herbicides and/or cutting), and
- The type of the simulated gene (dominant A or recessive a), the genotype of the rapeseed varieties and the self-pollination rates, pollen emission and yield productions.

The model is based on the life-cycle of oilseed rape, which concerns both cropped and volunteer rape plants, starting with the seed bank at harvest and determining seedling emergence. Some of these seedlings become adults, flower and produce new seeds, part of which replenish the seed bank at the end of the season. The relationships between the various life-stages depend on the crop grown in the field and on the cultivation techniques used to manage the crops. The model calculates for each stage of the annual rapeseed life-cycle and for each field or border the number of individuals per m^2 (number of seeds in the seed bank, of seedlings etc.) and the proportions of these individuals with and without transgenes (i.e., the contamination with GM seeds). During flowering and seed production, pollen and seeds are exchanged between fields and/or borders, depending on field areas, shapes and distances.

3. Materials and methods

3.1 Dataset

The aim of our analysis was to estimate how the properties of the farming region and the cropping system influence the rate of contamination on non-GM crops with GM seeds. The focus of the analysis was on

predicting the rate of adventitious presence of GM seeds in the central field of a large-risk field pattern, showed in Figure 1 (small and rectangular central field - number 14 - surrounded by large neighbour fields).

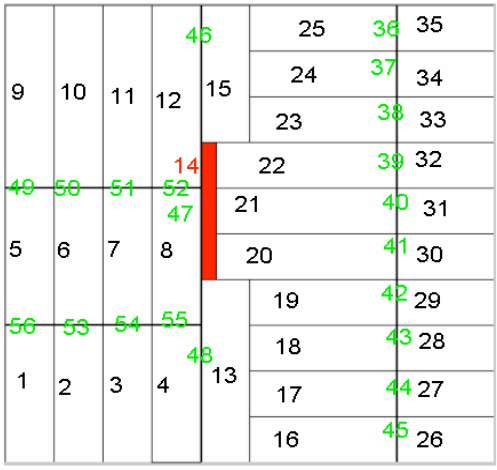

Figure 1: Large-risk field plan. Outcrossing rate for the central field (dark-shadowed field with number 14) was predicted. Neighbour fields are numbered from 1 to 13 and 15 to 35; borders are numbered from 36 to 56 and are small grass strips between cultivated fields (Colbach et al., 2005).

The dataset used in this analysis was based on simulations used for sensitivity analysis of GENESYS to field pattern characteristics (Colbach et al., 2005) and maximises pollen and seed input into the central field. Each simulation starts with an empty soil seedbank and lasts for 25 years. Each year, the crops and the management techniques for crops were chosen randomly, as well as the genetic variables describing the oilseed rape varieties. The only exception was the crop grown during the 25th year in the central field which was always non-GM oilseed rape. Our target variable was the rate of harvest contamination (adventitious presence of GM seeds) in this crop. 100 000 simulations were performed. Of the 25 simulated years of each simulation, full details were kept only for the last 4 years.

According to previous analyses of factors for presence and abundance of GM oilseed rape (Colbach et al., 2005; Debeljak et al., 2006), GM oilseed rape was most likely to be present if a previous GM oilseed rape had been grown in the same field. Having this in mind, we filtered the dataset originally consisting of 100,000 examples, excluding the examples in which there was GM oilseed rape grown on the target field in the last four years. The reason to do this was to avoid generating very obvious rules (for example: if there was GM oilseed rape on the target field in the last four years, the probability that it will now be contaminated is almost 100%) and try to see what is the interaction between the neighbouring fields. At the end the dataset consisted of 64,877 examples formulated as described previously.

3.2 Relational Data Mining: Relational Classification Trees

Most data mining algorithms assume that the input data is stored in a single table where each example is represented by a fixed number of attributes. These techniques are called attribute-value or propositional techniques because the patterns found can be expressed in propositional logic. Propositional data mining techniques (such as classification or regression decision trees) are popular, mainly because they are effi-

cient, easy to use and are widely accessible. In practice, however, the single table assumption turns out to be a limiting factor for many data mining tasks that involve data residing in multiple related tables. An example of such a problem is the analysis of co-existence of GM and non-GM crops in a region with many fields, where there is a need to examine the relations among the fields, and thus a need to use relational data mining techniques.

Data scattered over multiple relations (or tables) can be analyzed by conventional data mining techniques, by transforming it into a propositional table (attribute-value representation). This process is called *propositionalization* and it allows a wide choice of robust and well known algorithms. On the other hand, the multi-relational approach takes into account the structure of the original data by providing functionalities to navigate relational structure in its original format and generate potentially new forms of evidence not readily available in a flattened single table representation.

Since decision tree induction is one of the major approaches to data mining, upgrading this approach to a relational setting has been of great importance. Like in the propositional case, a table or relation is given, which contains at least two columns where the IDs of the examples and the values of the target variable are stored. An example of such a relation is *target(Field1, contaminated)*, which means that the field whose ID is 1, is labelled as contaminated. In addition, a set of background knowledge predicates/relations, stored in other tables, may be given.

Relational classification trees have much the same structure as propositional classification trees. Like classification trees, they predict the value of a dependent variable (class) from the values of a set of independent variables (attributes). Both types of classification trees have a test in each inner node that tests the value of a certain attribute and compares it with a constant, while leaf nodes give a classification that applies to all instances that reach the leaf. To classify an unknown instance, it is routed down the tree according to the outcomes of tests in successive nodes, and when a leaf is reached the instance is classified according to the class assigned to the leaf. On Figure 2 we show an example of the structure of a propositional classification tree.

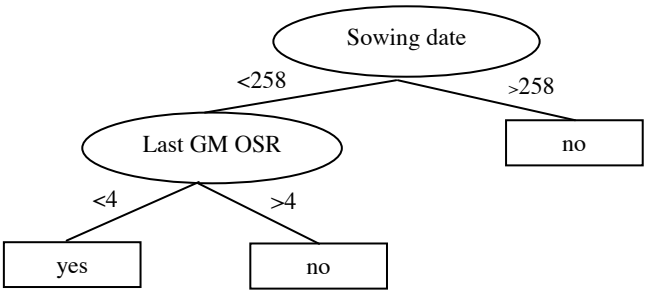

Figure 2: A simplified example of propositional classification tree predicting whether the central field in a large-risk field plan is contaminated with GM crops. Sooner sowing date and GM oilseed rape grown in the field not more than 4 years ago lead to contamination with GM crops. If the sowing date of the field is later or there are more than 4 years since the last GM crop grown, the field is predicted to be not contaminated

The major difference between propositional and relational classification trees is in the tests that can appear in the internal nodes. In the propositional case, the tests compare the value of a variable (property of the object) to a value, while in the relational case, the test can also refer to background knowledge relations or tables. For example, if *Field1* already occurs in the tree, the test *neighbour*(*Field1, Field2, noborder*) and *lastOSR*(*Field2, GM, NonGM*), *GM<2* examines whether there exists a field *Field2,* which

is a neighbour of *Field1* with no border between (they are contiguous and have no field margin in between) and which had its last GM oilseed rape crop grown two years ago.

An example of a relational classification tree for predicting the contamination of the central field of a large-risk field plan is given in Figure 3. The top node of the tree calls the target field *fieldA* (*targetField(FieldA)*) and checks whether the sowing date of that field in the present year (year 0) is before the 252th day of the year (*fieldDataYear(FieldA,0,Crop,SowingDate), SowingDate<252*). If not, then the field is predicted not to be contaminated. If yes, there is another test that checks if the sowing date of FieldA in the present year is before the 233th day of the year. If it is the case, then the field is predicted to be contaminated. If not, then the contamination depends on whether the target field has a neighbouring field with which it does not have a border (*neighbour(FieldA,FieldB,noborder)*), that had GM oilseed rape in the previous year (*fieldDataYear(FieldB,1,gm-OSR,SowingDate)*).

Relational decision trees can be easily transformed into first-order decision lists, i.e., ordered lists of relational rules. When applying a decision list to an example, we always take the first rule that applies and return the answer produced. A decision list is produced by traversing the relational decision tree in a depth-first fashion, going down left branches first. At each leaf, a rule is output that contains the prediction of the leaf and all the conditions along the left (yes) branches leading to that leaf.

For the purpose of our analysis, we used the system TILDE for building relational classification trees. TILDE (Blockeel et al., 1998) is a relational top-down induction of decision trees (TDIDT) algorithm, and outputs a first order decision tree, i.e., a decision tree that contains first order (relational) tests in the internal nodes.

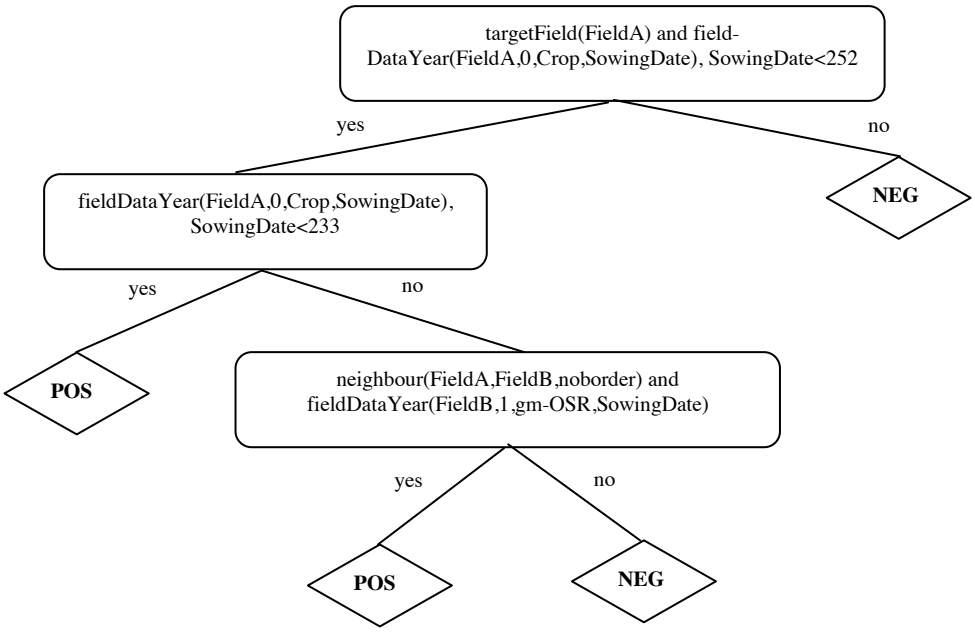

Figure 3: An example of relational classification tree predicting whether the central field in a large-risk field plan is contaminated with GM crops

4. Experiments and results

In the analysis we used the following representation of the data. The target relation (data label) was *rateOfAdvPresenceInField(SimulationID, FieldID, RateAdvPres)*, where *RateAdvPres* is the target/class variable, denoting the rate of adventitious presence of GM varieties of the non-GM field *FieldID*. *SimulationID* is the number (from 1 to 100 000) of the simulation and *FieldID* is always 14: in this analysis we were interested in the rate of adventitious presence of GM varieties of the central field of the field plan (Figure 1).

The background relations were related to the cultivation techniques, the year that oilseed rape was last planted at a given field, and the geometry of the field plan. In the relation *fieldDataYear(SimulationID, FieldID, Year, CultivationTechniques)*, *CultivationTechniques* is a list of variables describing the cropping techniques. Since in our previous analysis (Ivanovska et al., 2006) it was shown that the sowing date and the crop are the most important factors that influence the GM/non-GM cross-pollination, we used only information about crop and sowing date and ignored the other cropping techniques, like tillage, sowing density, efficiency for herbicides on non-GM/GM volunteers, $1^{st}/2^{nd}$ cutting, harvest loss and grazing. *Year* takes values from [22, 23, 24, 25]. In the relation *lastOSR(SimulationID, FieldID, LastGM, LastNonGM)*, *LastGM* is the last year [1..25] in which GM oilseed rape was grown on *FieldID*, and *LastNonGM* is the last year in which non-GM oilseed rape was grown on *FieldID*.

The relation *neighbour(Field1ID, Field2ID, NeighType)* holds if the minimum distance between Field1 and Field2 is zero. If they have a common edge of non-zero length, *NeighType* is *noborder*, and if they have only one point in common (touching with only one corner), then *NeighType* is *corner*. Additional information on the area of fields, their mutual distances (average and minimal), and length of the common edges was available, but was not used in our analyses.

For the experiments, we discretized the target attribute in order to obtain a classification problem. If the rate of harvest contamination exceeds 0.9%, a threshold value mentioned in EU regulations, the target field is considered contaminated, otherwise not. Given the size of the dataset we used a sampling strategy to build the tree: at each node only 10 000 examples are used to evaluate the tests and select the best test. The minimum number of examples a leaf has to cover was set to 600 and a random proportion of 20% of the data was set aside as a validation set for pruning.

We tried the following experimental settings:

- *Propositional*: besides the target relation *rateOfAdvPresenceInField(SimulationID, FieldID, RateAdvPres)*, only (propositional) data for the target field is included (not using any relations among the fields), i.e., the following predicates are used:
  - *fieldDataYear(FieldID,Year,Crop,SowingDate)*, for the target field,
  - *lastOSR(FieldID,LastGM,LastNonGM)*, for the target field,
- *Neighbour*: the same relations were used as in the *Propositional* setting, but now other fields are introduced via the *neighbour* relation, starting at the target field:
- *neighbour(Field1ID, Field2ID, NeighType)*

For each of these settings, we report the tree size (number of nodes) and the predictive performance, measured by a three-fold cross-validation in Table 1.

Table 1:
TILDE's experimental results

	PROPOSITIONAL	NEIGHBOUR
TREE SIZE	15	13
ACCURACY	78.35%	79.66%

The predictive performance of the two experimental settings was measured by a three-fold crossvalidation, resulting in accuracy of 78.35% for the *Propositional* setting and 79.66% for the *Neighbour* setting.

The relational classification trees were transformed into rules (for easier interpretation). Since there is one rule for each leaf node in the tree, from the *Propositional* setting there are 15 rules, and from the *Neighbour* setting there were 13 rules extracted. In addition we present an example of a rule extracted from the relational classification tree obtained in the *Propositional* setting:

contamination([neg]):-targetfield(T), fieldDataYear(T, 25, Crop, SowingDate), SowingDate<252, lastOSR(T, Gm, NonGm), Gm<20

The interpretation of the rule is as follows: target field will be predicted as not contaminated, if the sowing date in the present year is before the 252nd day of the year (9 September) and the last GM-oilseed rape grown on it was before year 20, i.e., there are more than 5 years since there was GM-oilseed rape on that field.

The next rule is an example obtained from the *Neighbour* setting:

contamination([pos]):-targetField(T), fieldDataYear(T, 25, Crop, SowingDate), SowingDate<252 neighbour(T, FieldA, noborder), fieldDataYear(FieldA, 24, gm-OSR, SowingDate)

The above rule states that if the sowing date of the target field in the present year is before the 252nd day of the year (9 September) and it has a neighbouring field (FieldA) with which it does not have border (they are contiguous), and the neighbouring field had GM-OSR last year, then the target field is predicted to be contaminated.

The results from the analysis in every experimental setting showed that the most important attribute for determining the contamination of the target field is the sowing date (measured in numbers of days since 1st of January) as was also shown in (Ivanovska, 2006) and (Colbach, 2005). The presented relational rule can be interpreted as follows: if the sowing date of the target field in the present year is before the 252nd day of the year and it has a neighbouring field (FieldA) with whom it does not have border (they are next to each other), and the neighbouring field had GM-OSR last year, then the target field is predicted to be contaminated.

From the results and the comparison of the accuracies of the relational to the propositional experiments we have noticed that the relational models provided small improvement over the propositional models of 1%. The main reason for this is the nature of the simulation data we used which did not use the advantages of the relational learning in their full. We believe that varying the target field and field plan will be more effective and will yield larger improvements over the propositional model.

5. Conclusion

In this paper we presented the use of relational data mining for analysing an output of a complex simulation model GENESYS. For the purpose of our analysis we used a large-risk field plan and learned to predict the contamination of the central field with GM seeds from the cropping systems and farming practices for the central fields and its neighbours. To build relational classification trees, we used the first-order decision tree learning system TILDE.

We tried different experimental settings constructing classical (propositional) trees, using information about the target field only, as well as relational trees, exploiting the relations among fields. The classification accuracies obtained were around 80%, with the relational approach achieving a higher (although only by 1%) accuracy. The learned model also clearly made use of the relational aspects, referring to the properties of and farming practices applied to the neighbouring fields of the target (central) field.

While data analysis and data mining methods had previously been used to analyze the output of simulation models for studying the co-existence of GM and non-GM crops, the use of relational learning methods is a novelty and a unique contribution of our study. The relational learning methods allow us to use the relational aspects, both spatial and temporal, of the information concerning the field plan and farming practices applied to the field in it.

Further work in improving the relational over propositional models should include performing more experiments with GENESYS using different field plans, as well as different target fields within each field plan, in order to exploit the advantages of the relational learning. Besides that, another direction for further work would be to use the same approach of relational learning, to analyse the simulation results of other models designed to study the co-existence of GM and non-GM crops.

Bibliography

Blockeel, H., De Raedt, L. (1998): Top-down induction of first order logical decision trees. Artificial Intelligence, 1011-2): 285-297

Debeljak, M., Squire, G., Demšar, D., Young, M., Džeroski, S. (2006): Data mining methods reveal soil-related and community-dependent factors in the presence and abundance of weedy oilseed rape before GM crop trials. Ecological modeling (in press).

Colbach, N., Clermont-Dauphin, C., Meynard, J.M. (2001a): GENESYS: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. I. Temporal evolution of a population of rapeseed volunteers in a field. Agriculture, Ecosystems and Environment, 83: 235-253

Colbach, N., Clermont-Dauphin, C., Meynard, J.M. (2001b): GENESYS: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. II. Genetic exchanges among volunteer and cropped populations in a small region. Agriculture, Ecosystems and Environment, 83: 255-270

Colbach N., Molinari N., Meynard J.M., Messéan A. (2005): Spatial aspects of gene flow between rapeseed varieties and volunteers: an application of the GENESYS model based on a spatio-temporal sensitivity analysis. Agronomy for Sustainable Development, 25: 355–368

Ivanovska, A., Panov, P., Colbach, N., Debeljak, M., Džeroski, S., Messean, A. (2006): Using simulation models and data mining to study co-existence of GM/non-GM crops at regional level. In Proceedings of 20th International Conference on Informatics for Environmental Protection (EnviroInfo), pp. 489-500, Graz, Austria