IRRIGATION AND DRAINAGE
Irrig. and Drain. (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/ird.1933

ESTIMATING DRAINAGE PERIODS FOR AGRICULTURAL FIELDS FROM MEASURED DATA: DATA-MINING METHODOLOGY AND A CASE STUDY (LA JAILLIÈRE, FRANCE)†

ANETA TRAJANOV[1]*, VLADIMIR KUZMANOVSKI[1,6], FLORENCE LEPRINCE[2], BENOIT REAL[3], ALAIN DUTERTRE[4], JULIE MAILLET-MEZERAY[5], SAŠO DŽEROSKI[1,6,7] AND MARKO DEBELJAK[1,6]

[1]*Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia*
[2]*ARVALIS— Institut du végétal, Montardon, France*
[3]*ARVALIS— Institut du végétal, Peronne, France*
[4]*ARVALIS— Institut du végétal, Station expérimentale de La Jaillière, La Chapelle Saint Sauveur, France*
[5]*ARVALIS— Institut du végétal, Station expérimentale, Boigneville, France*
[6]*Jozef Stefan International Postgraduate School, Ljubljana, Slovenia*
[7]*Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Ljubljana, Slovenia*

ABSTRACT

The identification of intensive drainage periods is important for determining mitigation strategies for protecting water against pollution with plant protection products (PPPs). Most attempts to estimate the start, duration and the end of a drainage period are based either on mechanistic modelling approaches or on empirical knowledge about tile drainage. Mechanistic modelling requires many parameters, while the empirical approach does not allow for making the simulations and predictions needed for proposing reliable mitigation measures. In order to complement these two approaches, we have used a data-mining approach on data from 25 (1987–2011) agricultural seasons (campaigns) from the experimental station La Jaillière, France. The models for estimating the start and the end of the intensive drainage period for a particular campaign have the form of decision trees and tell us which factors influence these dates the most. The start of a drainage period depends mostly on the cumulative drainage and the cumulative rainfall since the beginning of the campaign and the average air temperature of the last 7 days. For estimating the end of a drainage period, the most important variables are the cumulative rainfall of the last 7 days and the average air temperature of the following 7 days. Copyright © 2015 John Wiley & Sons, Ltd.

KEY WORDS: tile drainage; drainage period; empirical data; data mining; decision trees

Received 13 January 2014; Revised 16 January 2015; Accepted 16 January 2015

RÉSUMÉ

L'identification des périodes d'écoulement, et plus particulièrement la saison de drainage intense, est importante pour mettre en oeuvre des stratégies de limitation du risque de transfert des produits phytosanitaires dans les eaux. La plupart des démarches appliquées pour estimer le début, la durée et la fin de la période de drainage sont basées sur des modèles mécanistes ou sur la connaissance empirique du fonctionnement des réseaux de drainage. Les modèles mécanistes requièrent généralement de nombreux paramètres, et l'approche empirique ne permet pas de faire des simulations et des prédictions, indispensables pour la mise en place de mesures efficaces d'atténuation du risque. Dans le but de compléter ces deux approches, nous avons utilisé une méthodologie basée sur la fouille de données (data mining), en valorisant les informations recueillies durant 25 campagnes (1987 à 2011) par ARVALIS Institut du végétal sur le dispositif expérimental de La Jaillière à la Chapelle-Saint-Sauveur (Ouest de la France). Les modèles développés pour simuler les dates de début et fin de drainage prennent la forme d'arbres de décision qui hiérarchisent les facteurs ayant le plus d'influence sur la détermination de ces dates. Ainsi, la date de début du drainage dépend principalement de la quantité de pluies cumulées depuis le début de la campagne (1er septembre), mais aussi du cumul d'eau drainée, ainsi que de la température moyenne durant les sept jours précédents. La simulation de la date

*Correspondence to: Aneta Trajanov, Department of Knowledge Technologies. Jozef Stefan Institute. Jamova cesta 39. 1000 Ljubljana. Slovenia. Tel.: + 386 1 477 3662, Fax: + 386 1 477 3315. E-mail: aneta.trajanov@ijs.si
†Estimation des périodes de drainage sur des parcelles agricoles à partir de données mesurées: méthodologie de data mining et d'une étude de cas (La Jaillière, France).

de fin du drainage est déterminée par le cumul des pluies durant les sept derniers jours, et la température moyenne au cours des sept jours suivants. Copyright © 2015 John Wiley & Sons, Ltd.

MOTS CLÉS: réseau de drainage; période de drainage; données empiriques; fouille de données; arbres de décision

INTRODUCTION

Water pollution by plant protection products (PPPs) can be a source of ecological problems and necessitates careful evaluation and revision of the way we manage and protect surface and groundwaters. The pollution of surface water with PPPs in agricultural areas occurs when PPPs (herbicides and/or pesticides) are discharged directly or indirectly into water bodies due to the absence of adequate treatment that would protect the environment from harmful compounds, while the pollution of groundwater occurs with leaching of PPPs from the soil. A recent overview of the US Geological Survey National Water-Quality Assessment programme and the National Stream Quality Accounting Network concerning pesticide occurrence in US streams and rivers over two decades (1992–2001, 2002–2011) shows that one or more pesticides or pesticide degradates were detected more than 90% of the time in streams across all land uses during both decades (Stone *et al.*, 2014). These losses could occur in a very short time after their release in the environment, either as an agricultural application in the fields or as spillages due to bad management practices.

PPPs enter surface water via point and diffuse sources (Reichenberger *et al.*, 2007). Point source pollution comes from a single, identifiable source, such as a sewage plant, a sewer outflow or bad farm management practices (e.g. a spill during the filling of spraying equipment, cleaning of equipment, and processing of spray waste; Neumann *et al.*, 2002; Beernaerts *et al.*, 2005). Holvoet *et al*. (2008) published a review paper about the occurrence and sources of pesticides in surface waters across Europe. Results of studies conducted in Germany, Switzerland, Sweden, the UK and Belgium show that 20–80% of the load of PPPs in rivers could be attributed to point source pollution.

Diffuse source pollution denotes diffuse contamination that does not originate from a single discrete source. The main pathways of diffuse pollution at the field scale are surface runoff, discharge through subsurface drainage systems, and lateral seepage (lateral hypodermic flow on a nonpermeable soil substratum), while infiltration is identified as a direct pollutant transfer path in groundwater (Holvoet *et al.*, 2008; Brown and van Beinum, 2009). The results of a review study by Brown and van Beinum (2009) have shown that surface runoff and drainage make a significant contribution to the pollution of surface waters with PPPs as well. Drainage has been considered a relevant route for transport of PPPs in 6 out of 10 environmental scenarios representative of agricultural conditions across Europe (Forum for the Coordination of Pesticide Fate Models and their Use (FOCUS), 2001). In our paper, we will focus on diffuse sources of pollution of surface water through drainage outflow.

Drainage outflow denotes the discharge from the tile drainage infrastructure installed to enhance the moisture and aeration conditions of the soil and to lower the groundwater table (Zimmer, 1988). Tile drainage relies on subsurface drains with perforated plastic pipes. It shortens the residence time of water in biologically active root zones and aggravates the diffuse pollution of adjacent surface water with nutrients and PPPs (Tomer *et al.*, 2003). In our study, we focus on tile drainage water discharge and address the problem of estimating the time period when intensive drainage events occur in a tile-drained field.

European field studies of the transport of PPPs via tile drainage have identified four important factors that influence the concentration of PPPs in drainage: (i) the time interval between the application of PPPs and the occurrence of the first subsequent drainage event; (ii) the strength of sorption of PPPs to soil; (iii) the clay content of the soil; (iv) the degradation half-life of the PPPs in soil (Brown and van Beinum, 2009). According to Brown and van Beinum (2009), the regulatory assessment of the pollution of surface waters with PPPs via drainage could rely on two mitigation measures: (i) decreasing the permitted application rate, resulting in a proportional decrease of the pollution of drained water with PPPs, which decreases the exposure of surface waters; (ii) restricting the time period during which applications of PPPs may be made. Since losses of PPPs to drains are closely controlled by the time between the application and the initiation of drainage (Jones *et al.*, 2000; Renaud *et al.*, 2004), limiting the applications of PPPs to times with no drainage (e.g. early autumn or late spring) can be an effective mitigation option.

The restriction of the timing of application of PPPs to time periods with no drainage, however, can encounter two problems: (i) it could be undermined by unforeseen weather conditions; (ii) the efficiency of PPPs, which require moist soil, could be reduced (e.g. herbicides applied to the soil require moist soil for transfer to the root system of the target plants). Field data demonstrate large losses of PPPs when application is made to very wet soil, because drainage is likely to happen soon after application, or to dry clay soils with extensive cracking, because the transfer via cracks following very intensive rainfall conditions (e.g. storms) can

be very rapid. However, despite the large amount of research on mitigation measures, there are few reliable mitigation options other than to change the rate and timing of application according to weather and soil moisture conditions or to restrict the use of PPPs in the most vulnerable situations (Brown and van Beinum, 2009).

Time periods with intensive drainage events present a very high risk for fast transfer of PPPs to surface waters through the tile drainage discharge: according to timing mitigation measures, except for some foliar herbicides, any application of PPPs should be avoided during such time periods. An intensive drainage event occurs under conditions where the amount of water in the soil greatly exceeds the soil water-holding capacity. The excess soil water flows into the drainage system that takes water from the field. The amount and temporal dynamics of outflow from a drainage system depends on the intensity and persistence of water inputs and the amount of excess water over 100% saturation of the soil with water. If the duration and the amount of outflow from a drainage system differ significantly from the previous time period, then we consider that an intensive drainage event has started. The time period of intensive drainage lasts until the amount of water in the soil falls to or below 100% saturation of the soil with water. The amount of drained water is then very low or even non-existent. If, according to the temporal dynamics and the amount of drained water, this period differs significantly from the previous period, than we consider that a period of intense drainage has ended. Therefore, in this paper, we concentrate on the estimation of the start and the end of the time periods with intensive events of drainage through the subsurface tile drainage system. The time period of intensive drainage events (from the start to the end) will be hereafter referred to as 'a drainage period'.

RELATED WORK

Most of the attempts to estimate the start, duration and end of a drainage period are based on modelling water flow and the transport of PPPs from tile-drained fields. The state of the art of modelling water flow comprises two different approaches. The first is a mechanistic modelling approach and the second is based on empirical knowledge about the drainage.

The first approach is based on a combination of a theoretical description of different water flows in the soil and data from field experiments. Field data or high-quality lab data are used for calibration and validation of such models, while the models are structured based on theoretical knowledge. This (mechanistic) modelling approach is very informative, but requires many parameters, some of which are difficult to set or estimate. It also requires many different kinds of data that are not always trivial to obtain. Due to the large number of parameters that have to be fitted in these models, such mechanistic models can be unstable, which might also lower their predictive power. The verification and validation of the mechanistic models are in most cases performed on data collected from tile drainage discharges. Despite successful verification of the models, their predictions are often overestimates, mostly due to water bypassing the tile drains and discharging into deeper underground water (Gärdenäs *et al.*, 2006), which presents a fuzzy/non-deterministic element that is not sufficiently taken into account in mechanistic models.

The second approach is based on empirical knowledge about tile drainage, obtained mostly from recorded data, observations and experience from tile-drained fields and experimental stations, where different agricultural practices and mitigation measures are used and evaluated (Réal *et al.*, 2013; TOPPS Prowadis, 2014). In such a (empirical) modelling approach, experts make *ex-post* estimations of the drainage periods and their empirical expertise does not allow for making the simulations and predictions needed for proposing reliable mitigation measures. The verification in most cases shows that the beginning of drainage periods is estimated to happen earlier than in reality, which leads to too restrictive advice on the use of PPPs for farmers.

The objectives of our study are to overcome the shortcomings of the state-of-the-art approaches by developing a new methodological approach for estimating drainage periods, which will not be based on mechanistic water flow models.

The first goal of our study is validation of the experts' criteria for the cumulative drainage threshold that determines a drainage period. For the purposes of this goal, we considered the expert judgements for defining the drainage period, by *ex-post* analysis. Currently, the criteria used for estimating a drainage period are based on the quantity of cumulative drained water per campaign (1 September–31 August). In general, the expert judgement for the start of a drainage period is based on graphs of cumulative drainage and is not supported by a model of any kind. The amount of cumulative drainage when the experts decide that the drainage period starts is not precisely specified, but is in the range of 5–10 mm. On the other hand, according to expert judgement, the drainage period ends when the weekly cumulative drainage is below 1 mm and does not change in the next period. In order to quantify the expert criteria for the start and end of drainage period we apply appropriate statistical methods.

The second goal of our research is to find other conditions (e.g. past and forecasted meteorological conditions, like temperature, precipitation and their cumulative values for different time periods, as well as applied agricultural practices) that also affect the start and end of a drainage period.

These conditions, however, are neither defined in advance nor deterministic. Therefore, by exploring available data on water flows, applied agricultural practices and meteorological data, collected in the period from 1987 to 2011, we tried to identify these conditions and find regularities/rules for estimating a drainage period.

To address the second goal, we propose the use of data-mining techniques (Cortet *et al.*, 2011; Debeljak and Džeroski, 2011), which have proved useful for this kind of problem in the field of agricultural and environmental sciences (Debeljak *et al.*, 2007, 2008; Trajanov, 2011).

AREA DESCRIPTION

Our case study was one of 10 representative agricultural experimental sites in the EU chosen for the purposes of assessment of the Predicted Environmental Concentration in Surface Water of Active Substances under Directive 91/414/EEC and Regulation (EC) No. 1107/2009 (FOCUS, 2001). It is located in La Jaillière, France, and run by the technical institute ARVALIS—Institut du végétal. The site is situated at the southern end of the Armorican massif in western France. The site has been dedicated to the study of the influence of agricultural management practices on water quality since 1987.

The La Jaillière site is also considered a representative of the agricultural regions in Europe with shallow silt clay soils. The hydraulic pathways of La Jaillière soils are consistent with more than 90% of hydromorphic soils in France, which are tile drained on 3.1 million ha (AGRESTE, 2000). The soil at the experimental site is gleyic cambisoils lying on an alterite of sandstony schist, which is widely observed in the west of France. The arable layer, whose soil texture class is 'sandy silty loam', surmounts a layer with clay accumulation of soil texture class 'clay loam'. The alterite between 50 and 100 cm depth represents a natural obstacle to water infiltration. This impermeable layer is responsible for a surface water table during winter, the amplitude and duration of which are related to the winter rainfall excess. Due to clay illuviation, the clay content of the upper layer is less important than that of the lower layers. Organic matter content is about 2% in the upper layer but falls in the lower layers. Water content at field capacity varies from 20 to 23% according to layers and fields.

The climate at the site is of oceanic type with an average temperature of 12.3 °C per campaign (from 1 September to 31 August). The mean precipitation for the study period (1987–2012) per campaign is 717 mm and the mean reference evapotranspiration for study period per campaign is 712 mm (as measured at the meteorological station of la Jaillière, 1987–2012). The variability of precipitation and temperature for the study period is given in Figure 1.

The site is divided into a north and a south part. Each part contains blocks of fields. In our study, we have included data from 11 fields, where 10 fields are used to collect drainage and/or runoff water, while the 11th field (T2) is used as a reference field.

The surface area of the fields ranges from 0.34 to 1.08 ha and is cultivated following a traditional winter wheat/corn crop rotation. Each field is equipped with an independent tile drainage system (for drainage water collection) and/or runoff traps consisting of metal cuttings placed at the field edges to collect runoff water. A methodological precaution has been taken by establishing drainage trenches between the fields to avoid water passing from one field to another. The tile drains are located at a depth of d = 0.9 m below the soil surface, with a spacing of 10 m (Branger *et al.*, 2006).

The water is collected from drainage and runoff separately for each field with sampling proportional to the water flows. Meteorological data were collected from two meteorological stations in La Jaillière (ARVALIS—Institut du vègètal, 2010):

- the old station from 1982, situated 1.4 km from the fields;
- a new station from 2005, located at the experimental site, operating since 1 January 2006.

All the data about the agricultural practices (tillage, sowing, fertilizing and application dates of PPPs), the amount of water flows and the concentration of the water solution (mineral and active substances) in the surface waters (drainage and runoff) are collected in the PCQE (Pratiques Culturales et Qualité des Eaux) database at ARVALIS. The data are stored for each field separately for 25 campaigns (1987–2011) and can be directly linked to the weather database containing data from the meteorological stations. The data collected during drainage periods about the transfer of PPPs (13 264 data points based on 76 active ingredients and 4 metabolites) make the La Jaillière site a unique experimental site in Europe.

METHODS

Statistical methods

For the first goal of our study, i.e. validating the expert criteria for determining drainage periods, we considered expert judgments for defining the drainage period. The criteria used for estimating a drainage period are based on the quantity of cumulative drained water per campaign. In general, the expert judgement for the start of a drainage period considers a threshold of cumulative drainage, which is not precisely specified but is in the range of 5–10 mm. In

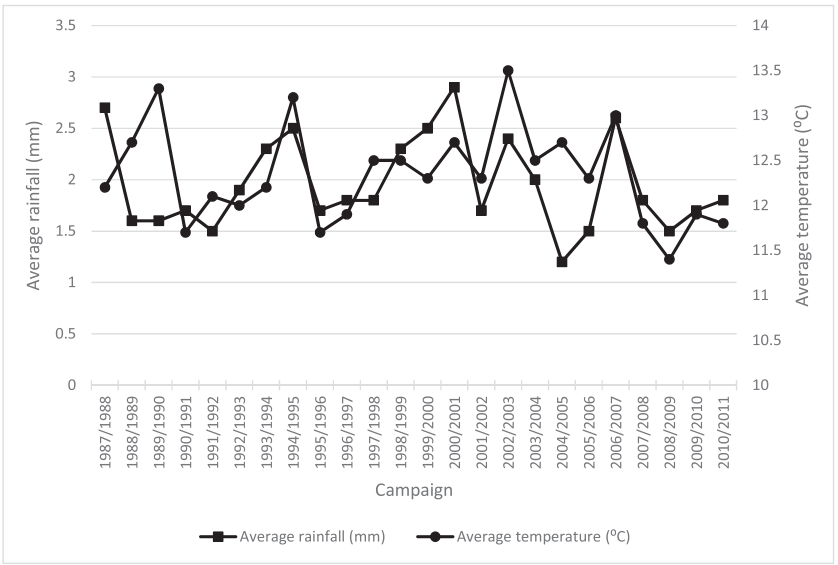

Figure 1. Variability of the average rainfall and average temperature per campaign (from 1st September to 31st August) for the period 1987-2011.

addition to the absolute cumulative drainage of water, they also take into account the temporal trend of drainage water, which must be distinguished from the trend of the previous period. Therefore, we defined five different hypothetical thresholds of cumulative drainage: 5, 6, 7, 8, 9 and 10 mm and used statistical measures to discover a rule from the available data regarding the possible thresholds of cumulative drainage since the start of a campaign (which influences the start of a drainage period). On the other hand, according to expert judgement, the drainage period ends when the weekly cumulative drainage is below 1 mm and does not change in the next period.

In order to make a decision about the start of a drainage period, we attempted to quantify the threshold of cumulative drainage that can be considered a condition for the start of a drainage period. The predefined thresholds were used as hypotheses. We used the mean value and standard deviation of the difference between the dates when cumulative drainage achieved the hypothetical thresholds, on one hand, and the provided starting dates for drainage periods, on the other. It can be seen as a statistical measure used to perform the task defined in the first goal of our study. The statistical significance of the differences between hypothetical and provided drainage period starting dates was assessed by the *t*-test.

Data mining

To build predictive models for the start and end of a drainage period, defined as our second goal, we used data mining. Data mining is a process that attempts to find patterns and new knowledge in data. Its overall goal is to extract information from a data set and transform it into an understandable structure for further use (Witten and Frank, 2011). Since data mining is concerned with finding patterns in data, the notions of most direct relevance here are the notions of data and patterns or models. Another key notion is that of a data-mining algorithm, which is applied to data to find patterns valid in the data. Different data-mining algorithms address different data mining tasks, i.e. have different intended use for the discovered patterns.

The data input to data mining is most commonly a single flat table (such as a spreadsheet table) comprising cases or examples (rows) and attributes or features (columns). In our study, cases are the daily records of drainage and runoff outflow from the fields and attributes are the meteorological conditions (like rainfall and temperature), the number of the field, the year, the cumulative drainage over different time periods, etc. The agricultural practices applied at the fields were also used as attributes in our data set, but our preliminary analyses showed that they do not have any influence on the start and end of a drainage period in our case study, so they were later left out of the analyses.

The output of a data-mining algorithm is typically a pattern (model) or a set of patterns that are valid in the given data. A pattern is defined as a statement (expression) in a given language that describes the facts or relationships in a subset of the given data and is (in some sense) simpler than the enumeration of all facts in the subset (Frawley *et al.*, 1991; Fayyad *et al.*, 1996). Different classes of pattern languages are considered in data mining and they depend on the data-mining task at hand. Typical representatives are decision (classification and regression) trees; association, classification and regression rules; and equations.

Many data-mining algorithms come from the fields of machine learning and statistics. Machine learning is a subfield of computer science and artificial intelligence that deals with the construction and study of algorithms that can learn from data. A common view in machine learning is that machine-learning algorithms perform a search (typically heuristic) through a space of hypotheses (patterns or models) that explain (are valid in) the data at hand. Similarly, we can view data-mining algorithms as searching, exhaustively or heuristically, a space of patterns in order to find interesting patterns that are valid in the given data. In our case, we used decision trees (Quinlan, 1986) and more specifically classification trees (Quinlan, 1993) to build predictive models for the start and end of a drainage period.

A decision tree is a hierarchical model, where each internal node contains a test on a descriptive feature of an example and each branch leaving this node corresponds to an outcome of this test (Figure 3. Terminal nodes (leaves) of a tree contain models defining the values of the target feature (dependent variable) for all examples falling in a given leaf. Given a new example, for which the value of the target feature should be predicted, the tree is interpreted from the root. In each inner node, the prescribed test is performed, and according to the result, the corresponding sub-tree is selected. When the selected node is a leaf, the value of the target feature for the new example is predicted according to the model in this leaf. If the target feature has nominal values (in our case, we want to predict whether the drainage period has started/ended or not), the decision tree is called a classification tree.

Measures of predictive performance

For classification problems, it is natural to measure a classifier's performance in terms of accuracy. The classifier (classification tree) predicts the class of each example: if the prediction is correct, that is counted as a success and if not as an error. The accuracy is the proportion of successful predictions made over the whole set of instances and measures the overall performance of the classifier (Witten and Frank, 2011).

The data set on which we build a predictive model (classification tree) is called a training data set. Normally, we are interested in the future performance of the model on new data. Thus, to evaluate the performance of a classifier, we have to assess its accuracy on a set of data that was not included in the formation of the model. This independent data set is called a test data set.

A commonly used technique for estimating the predictive performance of a classifier on test data is 10-fold cross-validation (Witten and Frank, 2011). In cross-validation, we decide on a fixed number of folds, or partitions of the data. In our case, we used the number 10. Then, the data are split into 10 approximately equal partitions, each of which (in turn) is used for testing, while the remainder of the data is used for training. This procedure is repeated 10 times so that, in the end, each partition has been used exactly once for testing. At the end, the obtained accuracies on the different iterations are averaged to yield an overall accuracy. This standard technique of 10-fold cross-validation is used in our study.

Other measures for predicting the performance of a classifier are the true positive rate (TPR) and the false positive rate (FPR) (Davis and Goadrich, 2006). The TPR measures the fraction of positive examples that are correctly classified, while the FPR measures the fraction of negative examples that are misclassified as positive. TPR and FPR can take up values from 0 to 1. The closer TPR is to 1, the better the predictive performance of the classifier.

The data-mining analyses consisted of two parts: estimating the start and estimating the end of a drainage period. We used the data-mining suite Weka (Witten and Frank, 2011), a collection of many machine-learning algorithms for different data-mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. For our analyses, we employed the classification tree algorithm J48 (Quinlan, 1993).

J48 produces classification trees. Besides the overall performance of such a tree, we can also consider individual parts of the tree. Their quality can be assessed from the information about the total number of all examples classified and the number of incorrectly classified examples in each leaf.

DATA PREPROCESSING

In order to test the predefined hypothetical thresholds, we used the available measured data and calculated the cumulative drainage since the start of a campaign, separately for each field and each year (1987–2011). Furthermore, we used the cumulative drainage to calculate the dates when the drainage period starts, according to the predefined hypothetical thresholds. These dates were later used for a statistical comparison with the dates provided by the experts.

For the second goal (data mining, generating predictive models), we first preprocessed the daily meteorological data, described above, for the period of 25 years (1987–2011). From these data, we calculated several new aggregated attributes and finally obtained a data set containing information for 9 fields (fields 1 and 2 were excluded because they were not drained) over 25 years, resulting in 78 894 daily records (i.e. examples).

The attributes used for estimating the start and end of drainage period are given in Table I. They include average temperature and total rainfall, measured for each of the last

Table I. Attributes (A) and dependent variable (DV) included in the analysis of: S—the start of a drainage period, E—the end of a drainage period. S1 and S2 denote analysis using total drainage and meteorological attributes, and only meteorological attributes, respectively

Attribute name	Description	S1	S2	E
Avg_temp_past_1-7_days	For each day, the average air temperature in the past 7 days	A	A	A
Avg_temp_past_8-14_days	For each day, the average air temperature in the week before the last one	A	A	A
Avg_temp_next_1-7_days	For each day, the average air temperature forecast for the next 7 days	A	A	A
Rainfall_cumul	For each day, the cumulative rainfall from the beginning of the campaign	A	A	A
Tot_rainfall_past_1-7_days	For each day, the total rainfall in the past 7 days	A	A	A
Tot_rainfall_past_8-14_days	For each day, the total rainfall in the week before the last one	A	A	A
Tot_rainfall_next_1-7_days	For each day, the total rainfall forecast for the next 7 days	A	A	A
Drainage_cumul_total	For each day, the total cumulative drainage since the beginning of the campaign	A		
Provided_start_of_drainage	The dates of the start of a drainage period provided by the experts, estimated *ex-post*	DV	DV	
Start_drainage_5mm	The dates of the start of a drainage period estimated by using 5 mm total cumulative drainage threshold		DV	
Start_drainage_10mm	The dates of the start of a drainage period estimated by using 10 mm total cumulative drainage threshold		DV	
End_of_drainage	The dates of the end of a drainage period provided by the experts, estimated *ex-post*			DV

two weeks and forecasted for the following week. Additionally, cumulative drainage and rainfall are considered.

RESULTS

Validation of the expert criteria for determining drainage periods

ARVALIS experts distinguish two drainage periods during a single campaign. The first one starts in late autumn–early winter when the water balance/status (rainfall − evapotranspiration) is in excess and the soil is water-saturated. The drainage system then removes excessive rainfall until the end of winter, when the water balance goes into deficit, thanks to the development of winter crops and warmer weather, both of which draw water from the soil. A second drainage period can occur almost exclusively after spring sowing. It is much shorter and it happens after significant spring rainfall or storms. When spring crops cover the ground, plants take up a lot of water, the water in the soil tends to decrease, and rainfall no longer causes drainage events. Experts approximate the dates of the beginning and end of a drainage period *ex-post*.

The dates of the start and end of a drainage period are available for each field separately for 25 campaigns (1987–2011). The variability of the start and the end, and the duration of drainage periods, is evident from Figure 2.

With the statistical analyses, we investigate the correlation between the actual (*ex-post*) and estimated start of drainage period for the estimates based on the hypothetical thresholds of cumulative drainage. The results are averaged for the 25 campaigns and the level of statistical significance is presented per field in Table II. The last two lines in the table present the overall averages (and standard deviations) of the differences between the two types of starting days for the drainage period. We test the significance of the differences per field by using the *t*-test, but do not perform the *t*-test on the overall averages, due to the differences in the characteristics of the fields.

The results of the *t*-test (Table II) show that the difference between the expert-provided dates for the start of drainage and the dates calculated by a 5 mm cumulative drainage threshold was not significantly different for any field, which means that the threshold of cumulative drainage of 5 mm matches most closely the dates defined by the experts. However, some results seem unexpected (e.g. fields T5 and T10 have significant differences of dates only at 9 and 10 mm threshold). They can be explained by the very fast growth of the cumulative value of measured drainage, which could achieve 10 mm in less than 2 days.

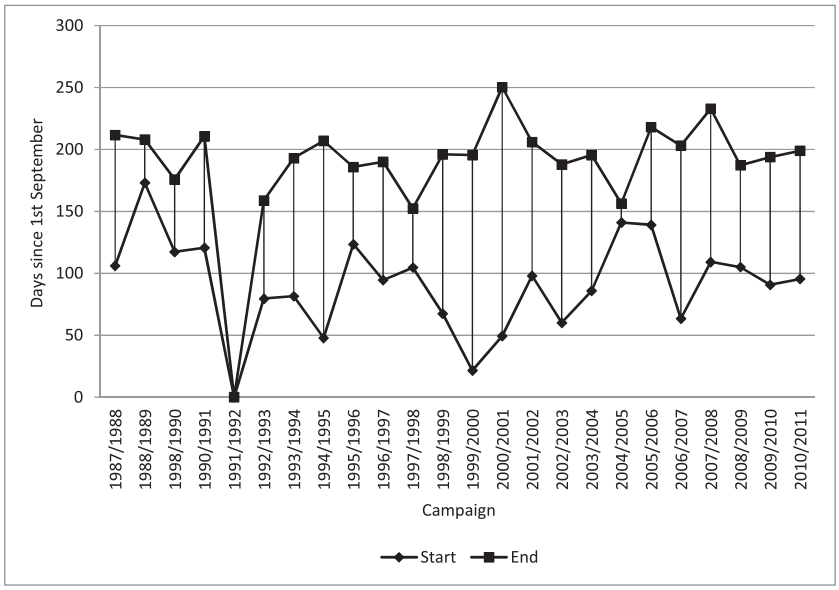

Figure 2. Variability of the start and the end of drainage period. The length of the line linking the start and the end of drainage period marks the duration of drainage period. Campaign 1991/1992 was without drainage period because of the draught.

Table II. The average and the standard deviation of the difference in days between the dates for the start of a drainage period provided by the experts (actual, *ex-post*) and the dates calculated (estimated) by using the proposed hypothetical thresholds (5, 6, 7, 8, 9 and 10 mm). These are given separately for each field where a drainage network is installed. Avg. Diff. stands for the average difference (in number of days), while Std Dev. stands for the standard deviation of the difference. The level of statistical significance is marked with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$

Field	Statistics	5 mm	6 mm	7 mm	8 mm	9 mm	10 mm
T3	Avg. Diff.	0.0	0.7***	1.7***	2.2***	3.7**	7.5*
	Std Dev.	0.0	0.9	2.3	2.5	5.8	15.2
T4	Avg. Diff.	0.0	3.8*	4.4*	5.0**	5.7**	6.3***
	Std Dev.	0.2	8.6	9.1	9.0	9.2	9.1
T5	Avg. Diff.	0.0	1.2	6.0	6.4	9.0*	10.6**
	Std Dev.	0.0	4.4	16.1	16.1	20.2	20.2
T6	Avg. Diff.	0.0	1.3**	3.7**	4.5**	5.2***	6.1***
	Std Dev.	0.2	1.9	6.4	7.4	7.7	7.9
T7	Avg. Diff.	0.0	2.3	8.3*	8.9*	9.7**	11.3**
	Std Dev.	0.0	6.4	16.5	17.0	17.8	18.6
T8	Avg. Diff.	0.0	1.1	2.6*	7.0*	9.3**	11.4**
	Std Dev.	0.0	3.2	5.6	16.2	18.1	18.6
T9	Avg. Diff.	0.0	1.2*	4.1**	6.5**	8.3***	10.6***
	Std Dev.	0.0	2.5	7.1	9.0	12.4	14.9
T10	Avg. Diff.	0.1	4.3	5.0	6.0	6.5*	6.7*
	Std Dev.	0.4	14.9	14.9	15.4	15.4	15.4
T11	Avg. Diff.	0.1	1.0***	1.9***	2.6**	3.9**	4.6**
	Std Dev.	0.5	1.4	2.6	4.7	8.8	9.0
	Total Avg. Diff.	0.0	1.9	4.2	5.4	6.8	8.4
	Total Std Dev.	0.1	4.9	9.0	10.8	12.8	14.3

Estimation of the start of a drainage period

The data set described above consisted of daily data for each field and each year in the period 1987–2011 (Table I). To estimate the start of a drainage period, we filtered the data set and chose only the data calculated before and during the drainage period. All the data collected after the end of the drainage period in an agricultural campaign were excluded from the data set. To estimate the start of the drainage period, we used meteorological data (rainfall and

temperature) and the total cumulative drainage since the beginning of a campaign.

We used three dependent (target) variables (attributes) for the estimation of the start of a drainage period: (i) the dates of the start of a drainage period provided by the experts, approximated *ex-post* (*Provided_start_of_drainage*); (ii) the dates of the start of a drainage period at 5 mm total cumulative drainage threshold; (iii) the dates of the start of a drainage period at 10 mm total cumulative drainage threshold. The dependent variables have two possible values: *no_drainage*, if the drainage period has not started yet, and *start_drainage*, if the drainage period has already started.

The analyses were divided into two parts. In the first part, we used only meteorological data and supposed that the cumulative drainage for the fields is unknown. In the second part of the analyses, beside the meteorological data, we also used the cumulative drainage since the start of a campaign as an attribute, but only when the target attribute was the provided date for the start of a drainage period. Namely, the calculated dates for the start of a drainage period were calculated from the cumulative drainage since the start of a campaign and because of this, the cumulative drainage could not be used as an independent attribute in the prediction of the calculated start of a drainage period.

From the meteorological data, we used the average temperature and rainfall for the 2 previous weeks and the following week.

We carried out three sets of experiments for each of the target attributes, using three sets of attributes (listed in the first column of Table III):

- only past data about temperature and cumulative rainfall (plus one additional experiment including cumulative drainage for the prediction of the provided date for the start of a drainage period);
- past and future data about temperature and rainfall (plus one additional experiment including cumulative drainage for the prediction of the provided date for the start of a drainage period);
- past and future data for temperature and cumulative rainfall (plus one additional experiment including cumulative drainage for the prediction of the provided date for the start of a drainage period).

In total, we obtained 12 models with the different sets of attributes and for each of the target attributes.

In Table III, we present the accuracy of the induced predictive data-mining models for the start of a drainage period. *Drainage_cumul_total* was used only for predicting the date of the start of a drainage period provided by the experts (*Provided_start_of_drainage*). These results are given in brackets. In Figure 3 we present the structures of three selected predictive models. In Figure 3(a), the model built for the provided date for the start of a drainage season, using past and future meteorological data, as well as cumulative drainage since the start of the campaign, is presented. In Figure 3(b), we present the model using a 5 mm threshold for the start of a drainage period as a target attribute and only past temperature data and cumulative rainfall as attributes. In Figure 3(c), we present the model using a 10 mm threshold for the start of a drainage period as a target attribute and past and future temperature data, and cumulative rainfall as attributes.

Table III. The accuracy of the models predicting the start of a drainage period for different combinations of attributes using different thresholds for the start of a drainage period (provided dates, 5 mm and 10 mm of cumulative drainage respectively), estimated with 10-fold cross-validation. The accuracy of the models with the cumulative drainage since the beginning of campaign as an additional attribute is given in brackets

Attributes	Provided dates	5 mm	10 mm
Avg_temp_past_1-7_days Rainfall_cumul *(Drainage_cumul_total)*	89.9% (94.5%)	90.8%	91.2%
Avg_temp_past_1-7_days Avg_temp_past_8-14_days Avg_temp_next_1-7_days Tot_rainfall_past_1-7_days Tot_rainfall_past_8-14_days Tot_rainfall_next_1-7_days *(Drainage_cumul_total)*	74.6% (94.6%)	75.8%	75.1%
Avg_temp_past_1-7_days Avg_temp_past_8-14_days Avg_temp_next_1-7_days Rainfall_cumul *(Drainage_cumul_total)*	89.8% (94.1%)	90.7%	90.8%

The predictive model built with the total cumulative drainage in the set of attributes (Figure 3a) has the total cumulative drainage (*Drainage_cumul_total*) at the topmost position in the model structure and its crucial value that determines the starting date of a drainage period is 4.99 mm: this is in line with the hypothesis that 5 mm presents the threshold for assessing the beginning of a drainage period.

The best model (Figure 3a) uses meteorological data and cumulative drainage. The models that do not use the cumulative drainage from the beginning of a campaign as an input attribute, but only use meteorological data, can still successfully estimate the beginning of a drainage period (Figures 3b, 3c). This approach leads to a solution that can be easily applied to an arbitrary field, where drainage is not measured, but meteorological data are available. The performance figures for all models for

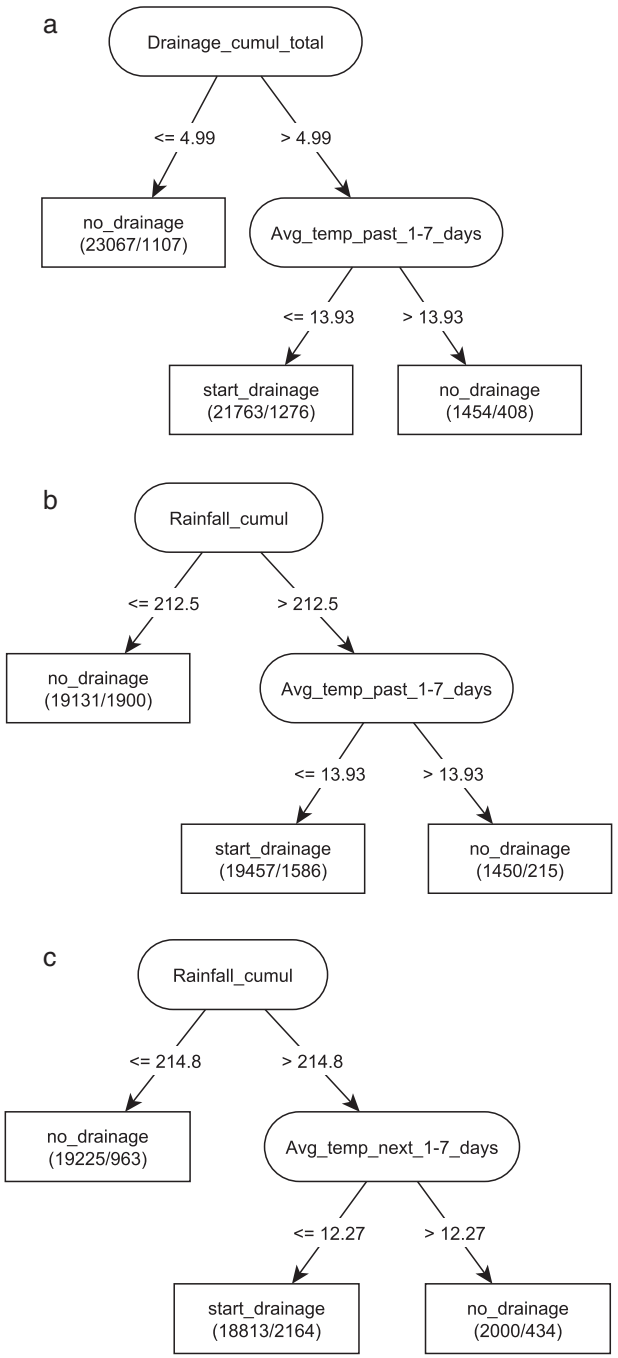

Figure 3. The predictive models for the start of a drainage period: (3a) using dates provided by the expert for the start of a drainage period as dependent variable, and past and future meteorological data and cumulative drainage as attributes (accuracy 94.56%, TPR 0.946, FPR 0.06); (3b) using a 5 mm threshold for the start of a drainage period as dependent variable and only past meteorological data and cumulative rainfall as attributes (accuracy 90.83%,TPR 0.908, FPR 0.092); (3c) using a 10 mm threshold for the start of a drainage period as dependent variable, past and future temperature data and cumulative rainfall as attributes (accuracy 90.82%,TPR 0.908, FPR 0.091). The numbers in parenthesis at the end of each leaf show the number of all examples classified in this leaf and the number of incorrectly classified examples.

estimating the start of a drainage period constructed with different combinations of attributes are given in Table III.

Estimation of the end of a drainage period

To estimate the end of a drainage period, we filtered the data set and used only the data for the drainage period and after the drainage period. The data collected before the start of a drainage period were excluded from these analyses.

To estimate the end of a drainage period, we used the list of attributes described above where the attribute *End_of_drainage* denotes the dependent variable (the end of drainage) and has only two possible values: *drainage*, if the drainage period is still continuing, and *end_drainage*, if the drainage period is finished (Table I).

We obtained several predictive models for the end of a drainage period, using two different combinations of attributes. The first uses only past meteorological data and the second one, besides past, also uses future meteorological data. The attributes used for estimating the end of a drainage period and the accuracies obtained with the predictive models are presented in Table IV, while example models are presented in Figures 4(a) and (b).

DISCUSSION

The statistical and data-mining analyses described above have provided us with interesting insights about the influence of the meteorological conditions and cumulative drainage on the start and end of a drainage period.

First, we tried to find a rule, based on the cumulative drainage since the start of a campaign, which can explain the start of a drainage period. The statistical analyses (above) showed that the starting dates determined by a cumulative drainage threshold of 5 mm are the closest to the dates determined by the experts. Therefore, for the data-mining analyses (above), we generated and compared pre-

Table IV. The accuracies of the models predicting the end of a drainage period for different combinations of attributes, estimated with 10-fold cross-validation

Dependent variable	Attributes	Accuracy
End_of_drainage	Avg_temp_past_1-7_days Avg_temp_past_8-14_days Tot_rainfall_past_1-7_days Tot_rainfall_past_8-14_days	85.9%
End_of_drainage	Avg_temp_past_1-7_days Avg_temp_past_8-14_days Avg_temp_next_1-7_days Tot_rainfall_past_1-7_days Tot_rainfall_past_8-14_days Tot_rainfall_next_1-7_days	88.3%

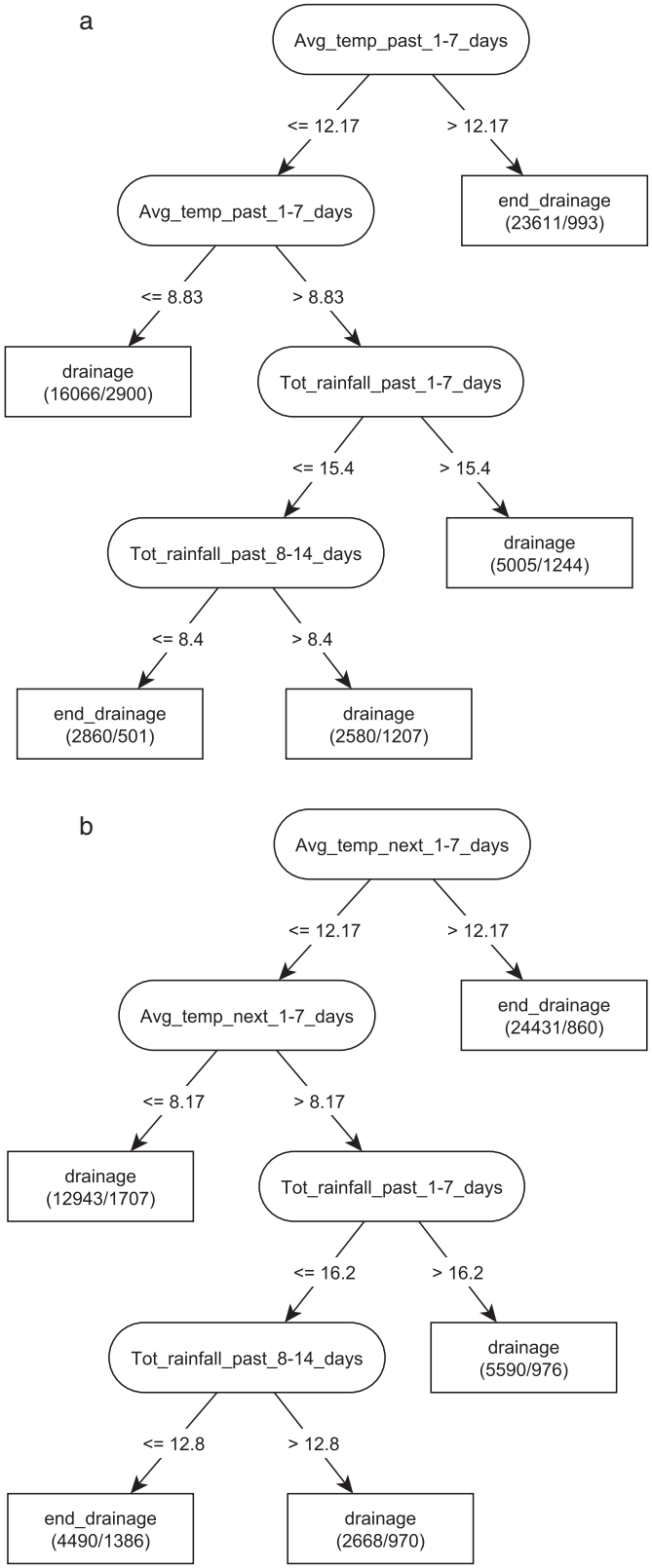

Figure 4. The predictive models for the end of a drainage period: (4a) using only past meteorological data (accuracy 85.94%, TPR 0.859, FPR 0.135); (4b) using past and future meteorological data (accuracy 88.28%, TPR 0.883, FPR 0.133) as attributes. The numbers in parenthesis at the end of each leaf show the number of all examples classified in this leaf and the number of incorrectly classified examples.

dictive models for the start of a drainage period determined by using 5 and 10 mm of cumulative drainage threshold, as well as the dates provided by the experts.

The predictive performance (accuracy) of the models is very high (Table III). The models built on dates determined by a 10 mm threshold of drainage have a slightly higher accuracy than the models built on dates determined by a 5 mm threshold and dates provided by the experts. However, if we consider the complete analyses, i.e. both the statistical and the data-mining analyses, the models obtained with a 5 mm threshold of drainage can be considered the most relevant. This is also supported by the models (exemplified by the model in Figure 3a) that include cumulative drainage since the start of a campaign as an attribute, which place this attribute at the very top (root) of the tree. This proves that cumulative drainage is the most important attribute, and the further splitting of the instances is based on a 4.99 mm threshold of cumulative drainage, which additionally supports the hypothesis with an accuracy of about 94%.

If we compare the models for estimating the start and end of a drainage period, we can note that the most important attributes for estimating the start of a drainage period are the past meteorological data. These are the cumulative drainage from the start of a campaign and the average air temperature in the last 7 days. The future meteorological data were not chosen as an important source of information because they do not provide further insights or contribute to more precise estimation of the drainage period.

In the case of estimating the end of a drainage period, better accuracy is achieved when (beside the past meteorological data) future meteorological data are included as well (average air temperature and rainfall in the next 7 days). Furthermore, they appear at the top of the classification tree, which indicates that they are more important for estimating the end of a drainage period than the past meteorological data. Finally, the high accuracy of the models confirms their reliability and is encouraging regarding their use as a first step in studying the pollution of surface waters with PPPs in arable areas.

In 2012, ARVALIS—Institut du vègètal carried out a synthesis about PPP transfers based on active substances with more than 200 available values. A study performed on isoproturon and diflufenican (1181 and 1176 data records, respectively) shows a significant effect of application time on transfer amounts. When they are applied before the beginning of a drainage period, both herbicides show low transfer through the drainage network, while transfers are important when application takes place during the drainage period. The development of a data-mining model for estimating the start and end of a drainage period and combined with the transfer properties of PPPs, obtained with statistical analysis, would ensure that ARVALIS—Institut du vègètal could give relevant recommendations and suggest appropriate mitigation measures for farmers very early during the agricultural season.

The models obtained for estimating the start and end of a drainage period could be used not just to estimate the daily status of the drainage regime on a particular field (e.g. presence or absence of drainage), but they can also be used to predict the drainage status of the field for a time period covered with reliable weather forecasts. Using information from weather forecasts to run simulations on models for the beginning of a drainage period (or the end, depending on the decision at hand) would make it easier for farmers and advisors to take into account the drainage period when deciding to apply PPPs in the field. Thus, our data-mining models, built from measured data, bring decision-making flexibility to their users, because they can be used either for *ex-ante* or *ex-post* analysis. The combination of both types of analysis presents a very simple decision support system, which significantly increases the reliability and flexibility of management decisions taken by advisors and farmers in the La Jaillière area (ARVALIS) or in other places with the same field and crop management properties.

The models obtained can also be successfully applied in the contents of a decision system for management of outflows from tile outlets. Artificial wetlands for natural purification of outflows from tile outlets have limited water retention capacity (Tournebize *et al.*, 2012). Therefore, managers must decide when to direct the water from the fields to the purification wetland and when they may discharge outflows directly to the surface water. Using models for estimating drainage periods would help farmers to optimize such decisions in order to achieve the best protection of surface water from pollution with PPPs.

CONCLUSIONS

The models that estimate the start and end of a drainage period are of high quality and can be used to give reliable estimation about the drainage regime for the La Jaillière region. They use data that can be obtained from weather forecasts, so that they can estimate whether the drainage will begin soon (sometime in the next week) or whether it will stop in the following week. While the data from the experimental station have so far been used only for *ex-post* analysis, the models we induced give us the possibility to make also *ex-ante* analysis, which could provide practical and relevant information for planning agricultural practices, whose application depends on the drainage period (e.g. the use of PPPs in the fields).

Extrapolation of the models to other locations is possible. Namely, we can use only meteorological data to estimate the start and end of a drainage period. However, in order to assess the accuracy of the models in agricultural regions other than La Jaillière, we need data from these regions.

Our work has achieved two important goals. First, the conditions for the start and end of a drainage period are now formalized. So far, the drainage periods were manually determined by experts, using visualizations of the quantities of rainfall and drainage in the fields. The experts were only able to do this *ex-post* and were not able to estimate the drainage periods in advance. Also, they were limited only to fields with records about drainage outflows and surface runoff water. The models proposed in this study formalize the conditions that determine the start and end of a drainage period. Even more, the models are general and can also be used on fields with tile drainage but without an installed system to record the amount of drained water. Second, with learning models that can estimate the start and end of a drainage period, we are providing a tool that can help advisors and farmers plan crop management activities in order to avoid or minimize the pollution of surface water with PPPs.

Further work will include applying these models on data from other representative experimental sites in the EU and other agricultural regions in France. This study is an important complementary approach towards assessing the pollution of surface water with PPPs. The next step is to combine the obtained models with models for predicting the concentration of PPPs in drained water and use them for estimating the ecological consequences of the application of certain PPPs and providing alternatives to a decision support system in order to select the most appropriate one (mitigation measures that would minimize pollution).

ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the project EVADIFF (Evaluation of existing models and development of new decision-making tools to prevent diffuse pollution caused by phytopharmaceutical products).

REFERENCES

AGRESTE. 2000. *Agricultural census*. Ministry of Agriculture, Food and Forestry; Montreuil-sous-bois Cedex, France.

ARVALIS—Institut du vègètal. 2010. *EOLE, French Climatic Database*. Boigneville, France.

Beernaerts S, Debongnie P, Gérard M, Barthelemy JP, Copin A, Guns M, Pussemier L. 2005. Evaluation of crop-protection-product losses into surface waters with the SEPTWA system. *International Journal of Environmental Analytical Chemistry* **85**: 41–50.

Branger F, Debionne S, Viallet P, Braud I, Vauclin M. 2006. Using the LIQUID framework to build an agricultural subsurface drainage model. In *Proceedings of the 7th International Conference on Hydroinformatics*, Nice; 2024–2031. 4-8 September 2006.

Brown CD, van Beinum W. 2009. Pesticide transport via sub-surface drains in Europe. *Environmental Pollution* **157**: 3314–3324.

Cortet J, Kocev D, Ducobu C, Džeroski S, Debeljak M, Schwartz C. 2011. Using data mining to predict soil quality after application of biosolids in agriculture. *Journal of Environmental Quality* **40**: 1972–1982.

Davis J, Goadrich M. 2006. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*; ACM: New York; 233–240.

Debeljak M, Cortet J, Demšar D, Krogh PH, Džeroski S. 2007. Hierarchical clasiffication of environmental factors and agricultural practices affecting soil fauna under cropping systems using Bt maize. *Pedobiologia* **51**: 229–238.

Debeljak M, Squire G, Demšar D, Young MW, Džeroski S. 2008. Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the UK. *Ecological Modelling* **212**: 138–146.

Debeljak M, Džeroski S. 2011. Decision trees in ecological modelling. In: *Modelling Complex Ecological Dynamics* Jopp F, Reuter H, Breckling B (eds). Springer: Berlin; 1–15.

Fayyad UM, Piatetsky-Shapiro G, Smyth P. 1996. From data mining to knowledge discovery: an overview. In: *Advances in Knowledge Discovery and Data Mining*; Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds). American Association for Artificial Intelligence: Menlo Park, Calif; 1–34.

Forum for the Coordination of Pesticide Fate Models and their Use (FOCUS). 2001. *FOCUS Surface Water Scenarios in the EU Evaluation Process under 91/414/EEC*. Report of the FOCUS working group on surface water scenarios (SANCO/4802/2001-rev2).

Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. 1991. Knowledge discovery in databases: an overview. In: *Knowledge Discovery in Databases* Piatetsky-Shapiro G, Frawley WJ (eds). MIT Press: Cambridge, Mass; 1–27.

Gärdenäs AI, Šimunek J, Jarvis N, van Genuchten MT. 2006. Two-dimensional modelling of preferential water flow and pesticide transport from a tile-drained field. *Journal of Hydrology* **329**: 647–660.

Holvoet KMA, Seuntjens P, Vanrolleghem PA. 2008. Monitoring and modeling pesticide fate in surface waters at the catchment scale. *Ecological Modelling* **209**: 53–64.

Jones RL, Arnold DJS, Harris GL, Bailey SW, Pepper TJ, Mason DJ, Brown CD, Leeds-Harrison PB, Walker A, Bromilow RH, Brockie D, Nicholls PH, Craven ACC, Lythgo CM. 2000. Processes affecting movement of pesticides to drainage in cracking clay soils. *Pesticide Outlook* **11**: 174–177.

Neumann M, Schulz R, Schäfer K, Müller W, Mannheller W, Liess M. 2002. The significance of entry routes as point and non-point sources of pesticides in small streams. *Water Research* **36**: 835–842.

Prowadis TOPPS. 2014. *Best Management Practices to Reduce Water Pollution with Plant Protection Products from Run-Off and Erosion*. European Crop Protection Association, Brussels, Belgium.

Quinlan J. 1986. Induction of decision trees. *Machine Learning* **1**: 81–106.

Quinlan J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, Calif.

Réal B, Leprince F, Maillet-Mezeray J. 2013. *Pesticide Non-Point Source Pollution Risks—Aquavallée®: a GIS-Based Diagnostic Tool*. Future IPM in Europe, 19–21 March, Riva de Garda, Italy.

Reichenberger S, Bach M, Skitschak A, Frede HG. 2007. Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness: A review. *Science of the Total Environment* **384**: 1–35.

Renaud FG, Brown CD, Fryer CJ, Walker A. 2004. Lysimeter experiment to investigate changes with time in availability of pesticide for leaching. *Environmental Pollution* **131**: 81–91.

Stone WW, Gilliom RJ, Martin JD. 2014. *An Overview Comparing Results from Two Decades of Monitoring for Pesticides in the Nation's Streams and Rivers, 1992–2001 and 2002–2011*. US Geological Survey Scientific Investigations Report 2014–5154. Reston, Virginia.

Tomer MD, Meek DW, Jaynes DB, Hatfield JL. 2003. Evaluation of nitrate nitrogen fluxes from a tile-drained watershed in Central Iowa. *Journal of Environmental Quality* **32**: 642–653.

TOPPS Prowadis. 2014. *Best Management Practices to Reduce Water Pollution with Plant Protection Products from Run-Off and Erosion*. European Crop Protection Association, Brussels, Belgium.

Tournebize J, Passeport E, Chaumont C, Fesneau C, Guenne A, Vincent B. 2012. Pesticide de-contamination of surface waters as a wetland ecosystem service in agricultural landscapes. *Ecological Engineering* **56**: 51–59.

Trajanov A. 2011. *Machine Learning in Agroecology: from Simulation Models to Co-existence Rules*. Lambert Academic Publishing (LAP): Saarbrücken, Germany.

Witten IH, Frank E. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* 3rd edn. Morgan Kaufmann: San Francisco, Calif.

Zimmer D. 1988. Transferts hydriques en sols drainés par tuyaux enterrés. PhD thesis, Université Pierre et Marie Curie, Paris VI; 326 pp.