FACHHOCHSCHULE **TRIER**
Hochschule für Technik, Wirtschaft und Gestaltung
University of Applied Sciences
Informatik

Informatik-Bericht Nr. 2008-7

Schriftenreihe Fachbereich Informatik, Fachhochschule Trier

iEMSs 2008:International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Identifying State Variables in Multivariate Hydrological Time Series Using Time Series Knowledge Mining

O. Gronz [a b], **M. Casper**[a], **P. Gemmar**[b]

[a] *Department of Physical Geography, University of Trier, Behringstrasse, 54286 Trier, Germany (gronz@uni-trier.de, casper@uni-trier.de)*

[b] *University of Applied Sciences (FH) Trier, Institute for Innovative Informatics Applications i3A, Schneidershof, 54293 Trier, Germany (o.gronz@fh-trier.de, p.gemmar@fh-trier.de)*

Abstract: The challenge of modeling rainfall-runoff processes is to define a suitable functional relationship between input variables representing precipitation measurements and the output runoff. Depending on the system's state, the amount of precipitation resulting in direct flow varies strongly. Thus, state variables indicating the system's actual state can enhance the accuracy of rainfall-runoff models significantly. Fuzzy models of Takagi-Sugeno-type are one of the effective approaches to simulate rainfall-runoff processes regarding the different system states. The design of a fuzzy rainfall-runoff model consists of two essential steps: the definition of the structure (input quantities, state variables, rules) and the identification of parameters in the conclusion. The latter one can be solved automatically using data-driven techniques in an optimal way concerning root mean square deviation. However, to solve the task of defining the structure, expert knowledge is mandatory to identify those time series that can effectively be used in a fuzzy model. This expert knowledge is not always available and not necessarily complete or correct. To identify effective state variables semi-automatically, the method Time Series Knowledge Mining (TSKM) has been used. TSKM discovers patterns representing the temporal concepts of duration, coincidence and order. Especially the patterns representing coincidence are valuable as the temporal concept of coincidence is used in fuzzy premises, too: the values of several state variables are evaluated simultaneously to determine the system's state. TSKM was applied to identify state variables with data from a 7 km^2 catchment in the northern Black Forest in Germany. From a set of more than 100 time series that were measured resulting in a huge set of possible state variable configurations, two soil moisture time series were identified. The fuzzy models generated using these two state variables were more efficient than all other models previously generated. Additionally, periods of snowfall and snowmelt could be reliably identified.

Keywords: Flood Prediction; Fuzzy Model; Time Series Knowledge Mining; Process Identification

1 INTRODUCTION

Especially in small catchments, the reaction on precipitation varies strongly depending on the catchment's actual state. In dry situations after long periods without precipitation, only a small part of rainfall will result in direct flow. In contrast, after periods with plenty of precipitation, a large part of an additional intensive rainfall will result almost completely in direct flow, indicating a high runoff coefficient. To identify the state of the catchment, different physical values can be measured. Additionally, these measurements allow for the identification of runoff processes and support the modeling of a catchment. Besides the usual values like discharge, precipitation, air temperature, radiation etc., soil moisture measurements can enhance the determination of the system's state significantly. All those measurements result in a large set of available time series.

Unfortunately, the examination of these time series is time consuming and challenging. On the other hand, the knowledge resulting from thorough examination is mandatory to build efficient models. A possible approach to model the rainfall-runoff processes is the usage of fuzzy logic. But for the creation of an efficient fuzzy model, expert knowledge is required in most cases: Only if a good combination of state variables is used in the premises to represent the system state, an efficient fuzzy model can be developed. Thus, there arise some disadvantages: gaining expert knowledge is time consuming; furthermore, this knowledge is not necessarily complete or correct; for each new catchment, the examination needs to be performed again. To solve this problem, a method needs to be applied that examines the various time series automatically and helps thereby identifying state variables.

2 FUZZY RAINFALL-RUNOFF MODELS

Catchments are complex systems. Modeling these systems using analytical approaches is challenging as most processes involved in discharge generation are hidden, not measurable, nonlinear or too complex. Thus, the effort in developing or computing a conceptual model is large. But elementary dependencies between the different variables precipitation and runoff are known and can be described: Depending on the system's state, a certain amount of precipitation will result in discharge. Additionally, historical time series with measurements of different values are available. These preconditions facilitate the usage of a fuzzy model (e. g. Cox [1994]). In fuzzy systems of Tagkagi-Sugeno type (Takagi and Sugeno [1985]), the description of the system state in the premise is separated from the quantitative influence of the variables used in the consequent. Depending on the system's state, these variables, e. g. precipitation, are differently weighted.

Figure 1 shows a possible design of a fuzzy model. The system uses m state variables in the premise like shown in part a). In this example, a soil moisture time series and a temperature time series is used. For each state variable, different fuzzy sets are defined, assigning each value from the range of values a degree of membership by membership function μ. A linguistic term can be applied to each fuzzy set, e.g. *dry*, *wet*, *medium* or *warm*. Overall, n rules are defined, following the general form IF ... AND ... THEN .... In the premise (IF part), different combinations of the individual fuzzy sets for each state variable are used to determine the state of the system (part b). For it, the degree of membership $\mu_{i,j}, 1 \leq i \leq n, 1 \leq j \leq m$ for each crisp value of the state variables is determined and all degrees are combined by calculating the product (part c) resulting in $\mu_i, 1 \leq i \leq n$. Furthermore, the result of each consequent is calculated. In our example, we assume a fixed base flow Q_B. The precipitation time series is convolved with a previously determined unit hydrograph, scaled by factor p and added to Q_B. The overall result Q_t is a linear combination of the degree of fulfillment of each rule and the individual consequent results $Q_{i,t}, 1 \leq i \leq t$ (part d).

Of course, this approach is extremely simplified. Processes like snowfall or snowmelt are not included. Nevertheless, it showed promising results like shown in Casper et al. [2007]. In designing such a model, one question arises: which configuration of state variables results in the most efficient models? Only soil moisture time series or accumulated precipitation or a combination of both of them? If n different time series are observed within a catchment, $2^n - 1$ different state variable configurations can be used as there are 2^n different subsets.

Optimizing fuzzy models can be subdivided into two different parts: the definition of the structure and identification of parameters. The latter one can automatically be solved data-driven in an optimal way (concerning root mean square deviation, Gemmar et al. [2006]). For the first step, the definition of the structure, expert knowledge is required (e.g. Cox [1994]). Once the state variables have been identified, different approaches can be applied to optimize structure like shown by Vernieuwe et al. [2003]. In our former models, state variables have been identified using expert knowledge resulting in a large set of rules (Casper et al. [2007]). In the following, we will present an alternative method.

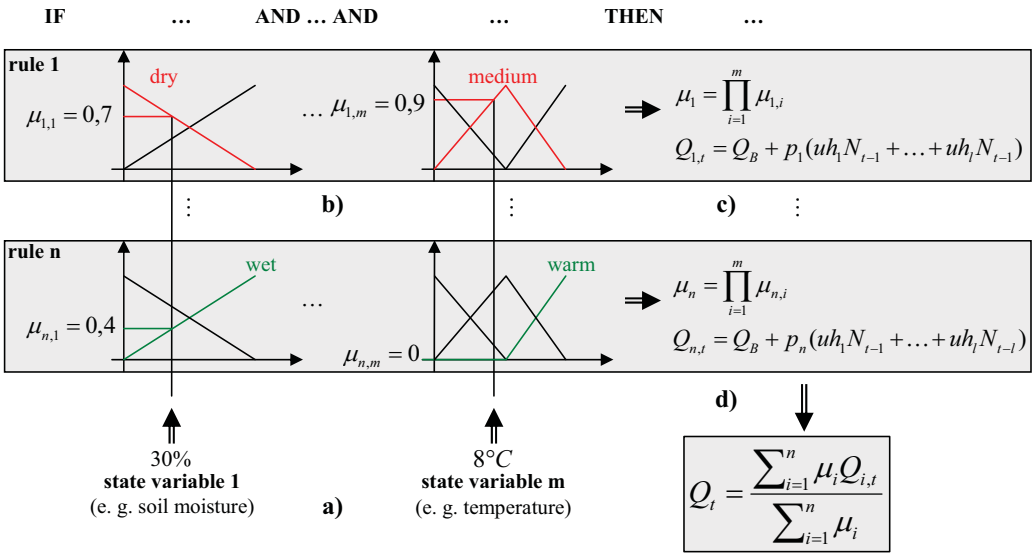

Figure 1: Possible structure of a rainfall-runoff fuzzy model of Takagi-Sugeno-type

3 Time Series Knowledge Mining

The purpose of the method Time Series Knowledge Mining (TSKM) as proposed by Mörchen [2006] is to extract temporal knowledge from multivariate time series: Using TSKM, patterns can be found that help to understand the underlying process and therefore allow modeling this process easier. The patterns are described using Time Series Knowledge Representation (TSKR), a hierarchically structured pattern language. The constructs of TSKR are named using terms from musicology, as these terms describe the underlying temporal concept vividly: Tones represent the concept of duration, overlapping parts of Tones, the Chords, stand for coincidence and partial order of Chords is represented by Phrases. The patterns found by TSKM are new, useful, understandable to humans, more compact and more abstract than the original time series. The discovery of TSKR patterns using TSKM is an interactive and iterative process consisting of five steps: pre-processing, defining Aspects, finding Tones, finding Chords and finding Phrases (Figure 2).

The goal of the first step, the pre-processing, is to remove systematic and random errors like noise, outliers, drift etc. As this is a problem regularly occurring in knowledge mining with plenty of different available approaches, there is no need for introducing new methods. Furthermore, the choice of the most suitable method to remove errors is highly application dependent and thus Mörchen [2006] does not suggest specific methods. A suitable well-known method can be chosen after examining the specific characteristics of a time series and the included errors.

In the second step, the dimensionality of the d-dimensional input space is reduced by grouping the multivariate time series to semantic blocks by selecting k subsets of the dimensions $[1, \ldots, d]$. These subsets are called *Aspects*. A descriptive, unique label is assigned to each Aspect allowing for an intuitive interpretation. Multivariate Aspects can profit from further processing like reducing the dimensionality by using e.g. PCA or calculating other derived time series.

Afterwards, in the third step, *Tones* are mined, representing the persistent occurrence of a state. Each Tone pattern contains a unique symbol, a descriptive label allowing for easy interpretation and a characteristic function indicating the occurrence of a state in an Aspect on a given time interval. The definition of the characteristic function is not restricted; various different types are possible, e.g. value-based functions, trend-based, shape-based etc. Using this characteristic function, the numerical Aspect is transformed into a symbolic interval series indicating the occurrence of Tones at each time point. Usually, this symbolic interval series will often be shortly interrupted e.g. due to noise. These gaps will split the potentially longer interval into two parts. To remove the gaps, a filter can be applied resulting in a symbolic time interval series of marginally interrupted

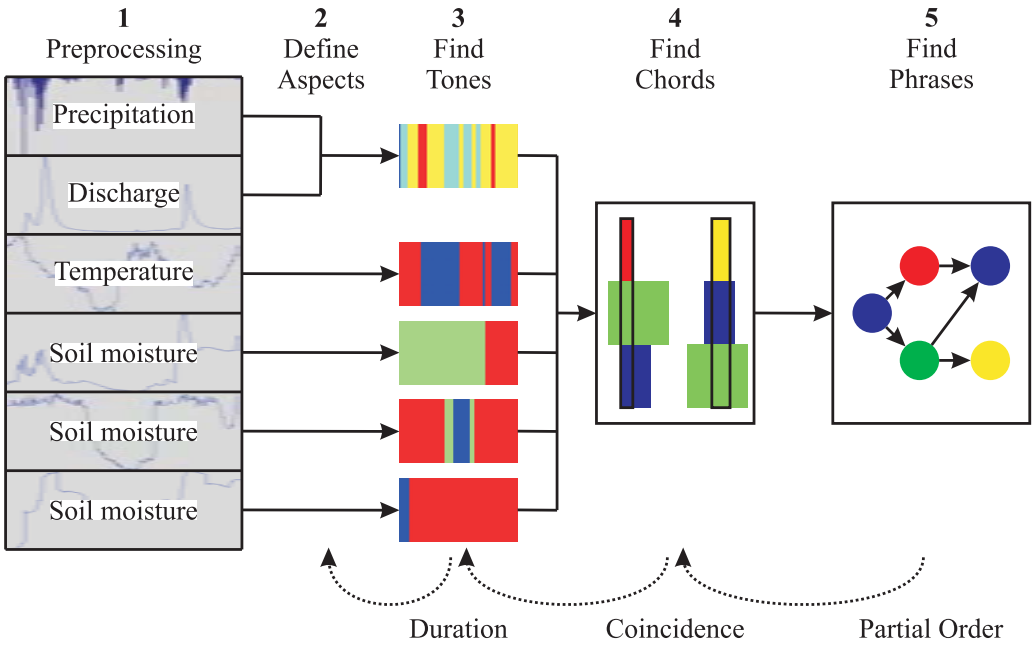

Figure 2: The five steps of Time Series Knowledge Mining (modified from Mörchen [2006])

occurrences. The filter can be parameterized by defining the maximum absolute length of a single gap and the amount of all gaps to be filled to the complete length of the time series. Finally, short Tones can be filtered, too.

Chord patterns also consist of a unique symbol, a label and a characteristic function indicating the simultaneous occurrence of Tones. A partial order is defined to allow for sub- and super-Chord relations between Chords. E.g. the Chord D consisting of Tones a and b is a sub-Chord of the Chord E consisting of Tones a, b and c - E is super-Chord of D. The hierarchy resulting from this relations can be visualized in a diagram and can be useful as we will see in section 4. To allow for tolerance concerning small differences of the support, margin-closed Chords are defined with respect to a threshold α indicating the allowed deviation of the ratio of the compared supports. Hence, the algorithm to mine Chords uses the following parameters: a threshold for the minimum duration of a Chord, the minimum and maximum size of a Chord and the threshold α for mining margin-closed Chords.

Finally, *Phrases* are mined. A Phrase pattern consists of a unique symbol, a descriptive label and a characteristic function indicating the subsequent occurrence of Chords according to a partial order on a given time interval. Again, a partial order is defined to allow for sub- and super-Phrase relations. The algorithm to mine Phrases uses the following parameters (extract): the minimum support of a Phrase and the minimum length of the paths in a Phrase.

4 Application of Time Series Knowledge Mining

In the following, the application of TSKM with data from a 7 km^2 catchment in the Northern Black Forest in Germany is described. The data contains various time series representing discharge, precipitation, air and soil temperature, soil moisture, radiation etc. in 1 h samples covering 2 years. Within this period, an extreme flood with a return period of 150 years was captured. The data has been examined thoroughly in different studies before (e.g. Casper [2002]).

To allow for unique and compact notation, the following nomenclature is used: Numeric, univariate time series are represented by vectors. E.g. $\vec{n}$ denotes a vector containing precipitation measurements. v_t describes the tth discrete sample of $\vec{v}$ corresponding to time point t. The convolution of two vectors $\vec{a}$ and $\vec{b}$ is denoted by $\vec{a} * \vec{b}$.

4.1 Pre-Processing

Although the database containing the different time series stores a quality flag for each value, some errors like outliers, drift or noise do remain. To identify and - if possible - to remove those errors, different techniques were applied. The effectiveness of the methods differed for the different physical values. For the soil moisture time series, outliers were identified reliably by applying the 2-sigma rule, indicating a value as an outlier if the difference to the mean value of the time series is bigger than the standard deviation times two (e.g. Runkler [2000]). For the runoff time series, this method failed as it indicates rare flood peaks as outliers. Another method is to plot the time series in a suitable way and use the human eye for outlier detection. For all time series, single outliers were replaced by linear interpolation of neighboring values. For some soil moisture time series, drift effects were visible, possibly due to changes in the soil matrix surrounding the TDR probes caused by water or animals. This was identified comparing the absolute values of different obviously saturated situations. Usually, these values will increase. Filters to remove noise were not applied, as tolerant algorithms are used in the following steps of TSKM.

4.2 Defining Aspects

For the following processing steps, time series representing discharge, precipitation, air temperature and soil moisture were selected from the available hydrometeorological data. The time series containing discharge measurements of the catchment's main outlet was used as Aspect *discharge*. The two available precipitation time series measured at two different sites within the catchment were combined by calculating the mean values element-by-element and used as Aspect *precipitation 1 h*. As the air temperature was measured at the same two sites, the two time series were combined similarly and used as Aspect *temperature*. Soil moisture was measured at four different sites in different depths resulting in 18 different time series. For each of the four sites, one probe was selected manually by an expert. As most of the probes of one site showed similar behavior, the probe showing least noise and drift was selected. Furthermore, the selected probe has to represent the typical behavior of the site from a hydrologic point of view. Thus, four Aspects were defined: *soil moisture 1* to *soil moisture 4*.

Besides those 7 Aspects, new Aspects were defined using derived times series. The precipitation time series contains samples representing the accumulated precipitation during one hour. Usually, those values will differ in consecutive samples resulting in a time series without persistent values. Even in periods with plenty of precipitation, the single samples can strongly vary. Hence, searching for persistent states within this time series will fail. Then again, the amount of precipitation within a certain time appears to be a good state variable as the system's state will change with the amount of precipitation. Unfortunately, the most appropriate duration of such a period is unknown. Regarding all these assumptions, new Aspects were derived to estimate the optimal size and to damp the high frequency properties. New Aspects *precipitation 2 h*, *precipitation 4 h*, *precipitation 8 h*, *precipitation 16 h . . . precipitation 128 h* were derived by convolving the original time series: e.g. *precipitation 2 h* is $\vec{n} * (1\ 1)^T$, *precipitation 4 h* is $\vec{n} * (1\ 1\ 1\ 1)^T$ etc.

In the premises of the fuzzy system, the amount of precipitation resulting in discharge is scaled depending on the system state. This parameter, which might be compared to the runoff coefficient of the corresponding time point, is not yet incorporated in the previously defined Aspects. The exclusive usage of the amount of discharge is not adequate as a great amount of discharge resulting from a great amount of precipitation does not indicate a critical system state. On the other hand, flood waters resulting from a small amount of precipitation only occur in critical system states. Thus, a time series containing the runoff coefficient of each hour would be beneficial, but is not available. To approximate a comparable time series, the system approach of the fuzzy system is used in a modified version. Besides the amount of precipitation resulting in discharge, the constant base flow is scaled by the same factor:

$$Q_t = Q_B + Q_B \cdot p_t + p_t \cdot F_t, \tag{1}$$

where $\vec{F} = \vec{n} * \vec{uh}$. This approach can be transformed as follows:

$$p_t = \frac{Q_t - Q_B}{F_t}. \tag{2}$$

Using this approach, a time series containing the continuous values of p_t can be calculated. For the application of TSKM to determine state variables, this time series is useful as it indicates the proper value of p_t in each time step. In the following, this time series is used as Aspect *discharge disposition*.

4.3 Finding Tones

In the third step of TSKM, Tones are mined. Each Tone represents a persistent symbolic state of an Aspect. As described in section 3, various different types of characteristic functions can be defined indicating the occurrence of a Tone in a certain time interval. In the premises of fuzzy systems, the value of a sample is used to determine the degree of membership for each fuzzy set. Thus, value-based characteristic functions are defined for the 15 Aspects described in section 4.2.

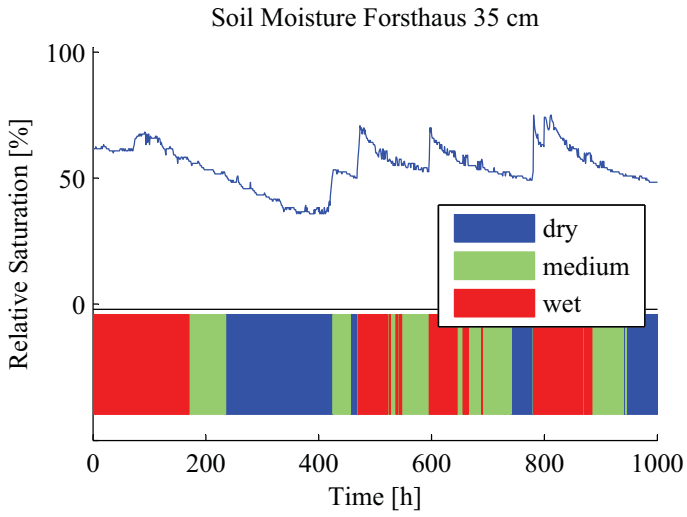

Figure 3: Transforming the numerical soil moisture time series to value-based Tones

For each of the different precipitation Aspects, three Tones were defined with labels *little*, *medium* and *plenty* to allow for an intuitive interpretation. The bin for each Tone has been determined manually in a way resulting in many Tones of type *little*, a couple of type *medium* and few of type *plenty*. The Aspect *discharge* has been subdivided into three bins similarly with labels *low*, *medium* and *high*. For *temperature*, only two Tones *cold* and *warm* were defined using the freezing point as threshold. The range of *discharge disposition* has been subdivided into 5 equidistant bins to allow for a more detailed resolution. To determine suitable bins for the soil moisture Aspects, the PERSIST algorithm has been used (Mörchen and Ultsch [2005]), which performs discretization observing the temporal order of samples. The algorithm produces bins that are similar to bins defined manually by an expert but result in fewer and more persistent occurrences of Tones. Unfortunately, this algorithm did not produce reasonable bins for the other Aspects. An example of descretization for Aspect *soil moisture 1* (site Forsthaus, depth 35cm) is shown in figure 3.

The symbolic interval series of Tones shown at the bottom of figure 3 in red, green and blue still contains many short Tones, e.g. at $t \approx 550$ and $t \approx 700$. These short Tones may be due to noise and do not represent a separated symbolic state and should be removed as a large set of short Tones will result in a huge set of Chords. Thus, this symbolic interval series is filtered resulting in a set of marginally interrupted occurrences. The first parameter of the filter, the maximum absolute length of any interruption, has been determined using the value of time to peak of the catchment. The second parameter, the maximum relative total length of all interruptions, has been limited to 10 %. The result of filter application is shown in figure 4 in the middle. As a last step, remaining short Tones are removed as shown at the bottom.

4.4 Finding Chords

In the next step, Chords are mined in the symbolic interval sequence containing Tones from the previous step. The algorithm that mines Chords uses different parameters. The first two parameters limit the size of a Chord. A Chord should at least consist of three Tones: besides Tones of Aspect *discharge* or *discharge disposition*, two further Tones should be contained as at least two time series should be used to represent system state accurately. The maximum size of a Chord is limited by the number of Aspects, that is to say 15. Again, the value of time to peak has been used to limit the minimum duration of a Chord. The threshold that limits the maximum deviation of support is 10 %. The flag whether to mine closed Chords has been set to *false* as the hierarchy of sub- and super-Chords will be used later to reduce the number of state variables. Using theses parameters, 2492 Chords were found.

4.5 Finding Phrases

To identify state variables, Chords are useful as they represent the simultaneous occurrence of Tones. For this procedure, Phrases are not necessary. However, they can be used to identify hydrologic processes like snowfall and melt within the multivariate time series.

For it, the set of Aspects has been reduced to *discharge*, *precipitation*, *temperature* and *discharge disposition*. The set of symbolic interval sequences for those Aspects has been reused without modifications. From the resulting set of Chords, specific ones were labeled manually: e.g. the Chord consisting of Tones *discharge is low*, *temperature is low* and *precipitation is medium* has been labeled *potential snowfall*. Using these Chords, Phrases were mined that indicate the different sequences of snowfall and following flood waters without precipitation; snowfall, a break were nothing happens (cold, no precipitation, low discharge) and subsequent flood water; snowfall and following precipitation during snowmelt etc. Those Phrases and especially the interval series indicating their occurrence are helpful in identifying flood waters influenced by snow as they are not included in our approach yet.

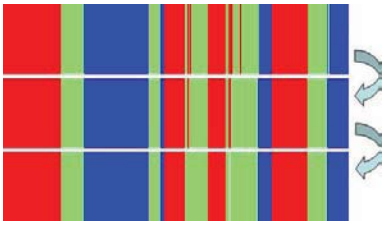

Figure 4: Filtering the symbolic interval series of Tones: value-based Tones at the top, marginally interrupted occurrences in the middle and after filtering short Tones at the bottom

4.6 From Chord to state variable

There is one critical point in the usage of TSKM to identify state variables: The occurrence of Chords does not imply that the included time series are adequate state variables; Chords state that the Tones of which they are made up occur simultaneously. Thus, the applicability of the included time series to represent system state has to be derived manually. For this, we propose the following approach: The set of Chords should not be limited by any parameter in the first four steps even if a huge set of Chords is produced. From this set of Chords, all Chords are automatically discarded that do not contain the Tone *discharge disposition is extreme* resulting in a small set of Chords that can be inspected manually. These Chords and their hierarchy are inspected manually by an expert and a Chord is selected that contains promising potential state variables besides the *discharge disposition*. Again, the initial huge set of Chords is inspected discarding automatically all Chords that are not made up of the *discharge disposition* and the identified potential state variables. The remaining set of Chords is again inspected manually by an expert. If the Chords cover the complete range of system states and low Tones of potential state variables coincide with low *discharge disposition* and high Tones of potential state variables coincide with high *discharge disposition*, the potential state variables can be used to generate a fuzzy model. If this fuzzy model shows poor results or if the set of Chords remaining after the last step contains conflicting Chords, the search has to be performed again, selecting different potential state variables.

5 RESULTS AND OUTLOOK

The set of possible state variable configurations resulting from the 15 Aspects defined in section 4.2 contains $32,767$ different configurations. From this large set, two soil moisture time series were identified as applicable state variables using TSKM and the approach described in section 4.6. The Nash-Sutcliffe efficiency of the generated fuzzy models using these two state variables is 0.38. During the identification, an expert assisted the knowledge mining process but no further explicit expert knowledge was used. In former approaches, the state variables were selected manually by an expert (Casper et al. [2007]). The Nash-Sutcliffe efficiency of the resulting model was -0.25. The expert knowledge was gathered during several years of intensive work within the catchment and two additional years of modeling. Comparing the two different approaches, the application of TSKM enhances the model's efficiency significantly. Furthermore, the amount of time that needs to be spent in gathering knowledge is decreased considerably.

The efficiencies mentioned above seam to be low compared to other rainfall-runoff models of other catchments but the small test catchment in the Black Forest shows extreme dynamics which complicates modeling. Additionally, the approach used in the fuzzy model does not yet include snowfall and snowmelt.

Up to now, TSKM has only been used with data of a single, small catchment. For evaluation purposes, the application of TSKM in other catchments is necessary. This evaluation includes the influence of parameters used in the different steps of TSKM, their general applicability and the approach to extract state variables from a set of Chords.

TSKM allows for mining temporal patterns within multivariate time series automatically. The interpretation of these patterns to identify state variables is a manual step. Thus, a huge set of possible state variable configurations can be mined but the identified state variables configuration is not necessarily the most efficient one.

REFERENCES

Casper, M. *Die Identifikation hydrologischer Prozesse im Einzugsgebiet des Dürreychbaches (Nordschwarzwald)*. PhD thesis, Universität Karlsruhe, Germany, 2002.

Casper, M., P. Gemmar, O. Gronz, M. Johst, and M. Stüber. Fuzzy logic-based rainfall-runoff modelling using soil moisture measurements to represent system state. *Hydrological Sciences Journal*, 52(3):478–490, 2007.

Cox, E. *The Fuzzy Systems Handbook*. Academic Press, Inc., 1994.

Gemmar, P., O. Gronz, and M. Stüber. Effiziente Erstellung und praktischer Einsatz von NA-Modellen mittels Fuzzy-Logik und automatisierter Entwicklungsverfahren. In Casper, M. and Herbst, M., editors, *Forum für Hydrologie und Wasserbewirtschaftung*, volume 16.06, pages 27–40, Trier, Germany, 2006.

Mörchen, F. *Time Series Knowledge Mining*. PhD thesis, Philipps-Universität Marburg, Germany, 2006.

Mörchen, F. and A. Ultsch. Optimizing time series discretization for knowledge discovery. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 660–665, New York, NY, USA, 2005. ACM.

Runkler, T. A. *Information Mining*. Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 2000.

Takagi, T. and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, Cybern.*, 15, 1985.

Vernieuwe, H., O. Georgieva, B. D. Baets, V. R. N. Pauwels, and N. E. C. Verhoest. Fuzzy models of rainfall-discharge dynamics. *International Fuzzy Systems Association World Congress*, 2003.