

Linguistic summarization of time series using a fuzzy quantifier driven aggregation

J. Kacprzyk*, A. Wilbik, S. Zadrozny

Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland

Available online 12 February 2008

Abstract

We propose new types of linguistic summaries of time-series data that extend those proposed in our previous papers. The proposed summaries of time series refer to the summaries of trends identified here with straight line segments of a piecewise linear approximation of time series. We first show how to construct such an approximation. Then we employ a set of features (attributes) to characterize the trends such as the slope of the line segment, the goodness of approximation and the length of the trend. The derivation of a linguistic summary of a time series is then related to a linguistic quantifier driven aggregation of trends. For this purpose we employ the classic Zadeh's calculus of linguistically quantified propositions but, extending our previous works, with different t-norms in addition to the basic minimum. We show an application to the analysis of time-series data on daily quotations of an investment fund over an eight year period, present some interesting linguistic summaries obtained, and show results for different t-norms. The results are very promising.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Time series analysis; Linguistic data summary; Fuzzy logic; Linguistic quantifier; t-norm

1. Introduction

A linguistic data (base) summary is meant as a concise, human-consistent description of a (numerical) data set. This concept has been introduced by Yager [22], and then presented in a more implementable form and further developed by Kacprzyk and Yager [14], and Kacprzyk et al. [15]. In this approach the contents of a database is summarized via a natural language like expression semantics of which is provided in the framework of Zadeh's [26,29,30] calculus of linguistically quantified propositions.

Since data sets in most nontrivial cases are large, if not huge, it is very difficult for a human being to capture and understand their contents. A natural language like description would be very helpful as natural language is the only fully natural means of articulation and communication for a human being. In this paper we consider a specific type of data, namely time series, i.e., some real valued function of time. For a manager, stock exchange players, etc. it might be convenient and useful to obtain a brief, natural language like description of trends present in the data (time series) on a company performance, stock exchange quotations, etc. over a certain period of time.

Though statistical methods exhibit their strength in such cases, and are often used, in our case we attempt to derive (quasi)natural language like descriptions that should be considered to be an additional form of data description of a

* Corresponding author.

E-mail address: kacprzyk@ibspan.waw.pl (J. Kacprzyk).

remarkably high human consistency because—as we have already indicated—for a human being the only fully natural means of articulation and communication is natural language. Hence, our approach is not meant to replace the classic statistical analyses but rather serve as an additional form of data description characterized by its very concise form and an extremely high human consistency.

The summaries of time series we propose refer in fact to the summaries of trends identified here with straight line segments of a piecewise linear approximation of time series. Thus, the first step is the construction of such an approximation. For this purpose we use a modified version of the simple, easy to use Sklansky and Gonzalez algorithm presented in [20].

Then we employ a set of features (attributes) to characterize the trends such as the slope of the line, the goodness of approximation of the original data points by line segments, and the length of a period of time comprising the trend.

Basically the idea of linguistic summaries proposed by Yager boils down to the interpretation in terms of the number or proportion of elements possessing some property. In the setting considered here a linguistic summary might look like: “Most of the trends are short” or, in a more sophisticated form: “Most long trends are increasing”. Such expressions are easily interpreted using Zadeh’s calculus of linguistically quantified propositions [26]. The most important element of this interpretation is a linguistic quantifier exemplified by “most”. In Zadeh’s [26] approach it is interpreted in terms of a proportion of elements possessing a certain property (e.g., long trends) among all the elements considered (e.g., all trends).

In Kacprzyk et al. [9,10] we proposed to use Yager’s linguistic summaries, interpreted and dealt with using Zadeh’s calculus of linguistically quantified propositions, for the summarization of time series. In our further papers (cf. [11–13]) we proposed, first, another type of summaries that does not use the linguistic quantifier based aggregation over the number of trends but over the time instants they take altogether. For example, such a summary can be: “Trends taking most of the time are increasing” or “Increasing trends taking most of the time are of a low variability”. Such summaries do not directly fit the framework of the original Yager’s approach and to overcome this difficulty we have generalized our previous approach by modelling the linguistic quantifier based aggregation both over the number of trends and over the time they take using, first, the Sugeno integral and, then, the Choquet integral. All these approaches have been proposed using a unified perspective given by Kacprzyk and Zadrozny [16] that is based on Zadeh’s [27] concept of a protoform.

In this paper we will basically employ the classic Zadeh’s calculus of linguistically quantified propositions. However, we will extend the idea proposed in our source paper [9] by using various t-norms as opposed to the minimum operation used in that source paper. By employing data on daily quotations of an investment (mutual) fund over an eight year period we will present an implementation of the new method proposed, and an analysis of results for various t-norms used in Zadeh’s calculus of linguistically quantified propositions.

The paper is in line with some modern approaches to a human consistent summarization of time series. First of all, one should cite here the works of Batyrshin and his collaborators [1,2]. Basically, they consider the problem in terms of devising a rule base, and then assume a different approach to linguistic granulation. Chiang et al. [6] approach, though it basically addresses a problem that is similar in spirit, is somehow conceptually different.

To see our approach in a proper perspective it may be expedient to refer to an interesting project coordinated by the University of Aberdeen, UK, SumTime, an EPSRC Funded Project for Generating Summaries of Time Series Data (cf. www.csd.abdn.ac.uk/research/sumtime/). The essence of this project can be summarized by the citation from its Web site: “Our goal is to develop technology for producing English summary descriptions of a time-series data set. Currently there are many visualization tools for time-series data, but techniques for producing textual descriptions of time-series data are much less developed. Some systems have been developed in the natural-language generation (NLG) community for tasks such as producing weather reports from weather simulations, or summaries of stock market fluctuations, but such systems have not used advanced time-series analysis techniques. Our goal is to develop better technology for producing summaries of time-series data by integrating leading-edge time-series and NLG technology”.

Basically, the essence of this project is close in intent and spirit to our works. However, the type of summaries they generate is different, not accounting for an inherent imprecision of natural language. A good example is here the case of weather prediction that is one of the main application areas in that project. For instance, cf. Sripada et al. [21], linguistic summaries related to wind direction and speed can be:

- WSW (west of south west) at 10–15 knots increasing to 17–22 knots early morning, then gradually easing to 9–14 knots by midnight.

- During this period, spikes simultaneously occur around 00:29, 00:54, 01:08, 01:21, and 02:11 (o'clock) in these channels.

Similar linguistic summaries have been obtained for time series data on blood pressure, gas turbines, etc.

Notice that these linguistic description of time series data concerning wind directions and speed do provide a higher human consistency as natural language is used but they capture imprecision of natural language to a very limited extent. In our approach this will be overcome to a considerable extent.

In this paper, first, we describe the way the trends are extracted from time series and characterized using a set of attributes. Then, we present the ideas of some basic characteristics of a dynamic behavior of time series that will be used in our further analyses, i.e., the dynamics of change, duration, and variability. Then, we briefly recall the basics of the original Yager's approach to linguistic summarization and discuss how it may be used to describe a sequence of trends (time series). Next, we show how these linguistic summaries can be derived using the classic Zadeh's calculus of linguistically quantified propositions, and show how various t-norms can be involved. Then, we comment upon the computer implementation, and show some examples of linguistic summaries of time series of daily quotations of an investment (mutual) fund over an eight year period. We analyze the impact of various parameters, notably the choice of a t-norm. We finish with some concluding remarks.

2. Temporal data and trend analysis

We deal with numerical data that vary over time, and a time series is a sequence of data measured at uniformly spaced time moments. We identify trends as linearly increasing, stable or decreasing functions, and therefore represent given time series data as piecewise linear functions. Evidently, the intensity of an increase and decrease (slope) will matter, too. These are clearly partial trends as a global trend in a time series concerns the entire time span of the time series, and there also may be trends that concern parts of the entire time span, but more than a particular window taken into account while extracting partial trends by using the Sklansky and Gonzalez [20] algorithm.

In particular, we use the concept of a uniform partially linear approximation of a time series. Function f is a uniform ε -approximation of a time series, or a set of pairs of points $\{(x_i, y_i)\}$, if for a given, context dependent $\varepsilon > 0$, there holds

$$\forall i: |f(x_i) - y_i| \leq \varepsilon \quad (1)$$

and, clearly, if f is linear, then such an approximation is a linear uniform ε -approximation.

We use a modification of the well known, simple yet quite effective and efficient Sklansky and Gonzalez's [20] algorithm that finds a linear uniform ε -approximation for subsets of points of a time series. The algorithm constructs the intersection of cones starting from point p_i of the time series and including a circle of radius ε around the subsequent data points p_{i+j} , $j = 1, 2, \dots$, until the intersection of all cones starting at p_i is empty. If for p_{i+k} the intersection is empty, then we construct a new cone starting at p_{i+k-1} . Figs. 1(a) and (b) present the idea of the algorithm. The family of possible solutions is indicated as a gray area. Clearly other algorithms can also be used, and there is a lot of them in the literature; in particular, those proposed by Keogh and his collaborators should be mentioned (cf. [18,19]).

To present details of the algorithm, let us first denote:

- p_0 —a point initializing the current cone,
- p_1 —the last point checked in the current cone,
- p_2 —the next point to be checked,
- Alpha_01 —a pair of angles (γ_1, β_1) , meant as an interval, that defines the current cone as shown in Fig. 1(a),
- Alpha_02 —a pair of angles of the cone starting at the point p_0 and inscribing the circle of radius ε around the point p_2 (cf. (γ_2, β_2) in Fig. 1(a)),
- function $\text{read_point}()$ that reads a next point of data series,
- function $\text{find}()$ that finds a pair of angles of the cone starting at the point p_0 and inscribing the circle of radius ε around the point p_2 .

A pseudocode of the algorithm that extracts trends is depicted in Fig. 2.

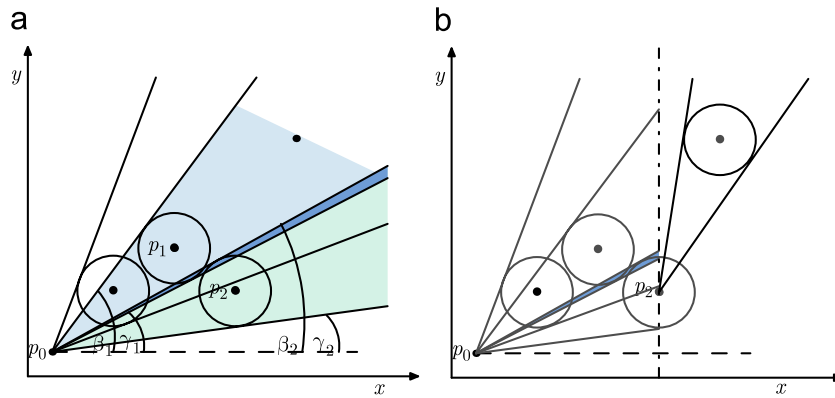


Fig. 1. An illustration of the algorithm for the uniform ε -approximation. (a) The intersection of the cones is indicated by the dark gray area. (b) A new cone starts in point p_2 .

```

read_point(p_0);
read_point(p_1);
while(1)
{
  p_2=p_1;
  Alpha_02=find();
  Alpha_01=Alpha_02;
  do
  {
    Alpha_01 = Alpha_01 ∩ Alpha_02;

    p_1=p_2;
    read_point(p_2);
    Alpha_02=find();
  } while(Alpha_01 ∩ Alpha_02 ≠ ∅);

  save_found_trend();
  p_0=p_1;
  p_1=p_2;
}

```

Fig. 2. Pseudocode of the modified Sklansky and Gonzalez [20] algorithm for extracting trends.

The bounding values of Alpha_02 (γ_2, β_2), computed by the function $\text{find}()$ correspond to the slopes of two lines that:

- are tangent to the circle of radius ε around the point $p_2 = (x_2, y_2)$,
- start at the point $p_0 = (x_0, y_0)$.

Thus

$$\gamma_2 = \arctg \left(\frac{\Delta x \cdot \Delta y - \varepsilon \sqrt{(\Delta x)^2 + (\Delta y)^2 - \varepsilon^2}}{(\Delta x)^2 - \varepsilon^2} \right)$$

and

$$\beta_2 = \arctg \left(\frac{\Delta x \cdot \Delta y + \varepsilon \sqrt{(\Delta x)^2 + (\Delta y)^2 - \varepsilon^2}}{(\Delta x)^2 - \varepsilon^2} \right)$$

where $\Delta x = x_0 - x_2$ and $\Delta y = y_0 - y_2$.

The resulting linear ε -approximation of a group of points p_0, \dots, p_1 is either a single segment, chosen as, e.g., a bisector of the cone, or one that minimizes the distance (e.g., the sum of squared errors) from the approximated points, or a whole family of possible solutions, i.e., rays of the cone.

This method is simple, effective and efficient as it requires only a single pass through the data. Now we will identify (partial) *trends* with the line segments of the constructed piecewise linear function. Among some other interesting and promising approaches, works of Keogh and his collaborators [19,18] should be cited.

3. Dynamic characteristics of trends

In our approach, while summarizing trends in time series data, we consider the following three aspects:

- dynamics of change,
- duration, and
- variability,

and it should be noted that by trends we mean here global trends, concerning the entire time series (or some, probably large, part of it), not partial trends concerning a small time span (window) taken into account in the (partial) trend extraction phase via the Sklansky and Gonzales [20] algorithm mentioned above.

These three characteristic features of trends are clearly the most straightforward and intuitively appealing ones as they concern those aspects of what happens with data over time that can easily be understood by domain experts. This has been clearly visible in our case while working with domain experts in the field of finance. It should also be noted that in using well-established statistical tools for time series analysis, these features (mostly the dynamics of change and variability) are also of primordial relevance and many tools for dealing with them are available. However, these three basic features used in this paper are clearly not the only choice, and can be complemented by other suitable aspects of dynamic characteristics of trends, as needed in a specific application. One should be, however, cautious in this respect and choose those which may be acceptable and intuitively appealing to domain experts.

In what follows we will briefly discuss these factors.

3.1. Dynamics of change

Under the term *dynamics of change* we understand the speed of change. It can be described by the slope of a line representing the trend (cf. any angle η from the interval $\langle \gamma, \beta \rangle$ in Fig. 1(a)). Thus, to quantify dynamics of change we may use the interval of possible angles $\eta \in \langle -90^\circ; 90^\circ \rangle$.

However, it might be impractical, and not human consistent, to use such a scale directly while describing trends. Therefore we may use a fuzzy granulation in order to meet the users' needs and task specificity. The user may construct a scale of linguistic terms corresponding to various inclinations of a trend line as, e.g.:

- quickly decreasing,
- decreasing,
- slowly decreasing,
- constant,
- slowly increasing,
- increasing,
- quickly increasing.

Fig. 3 illustrates possible lines corresponding to the particular linguistic terms.

In fact, each term represents a fuzzy granule of directions. In Batyrshin et al. [1,2] there are presented many methods of constructing such a fuzzy granulation. The user may define a membership functions of particular linguistic terms depending on his or her needs.

We map a single value α (or the whole interval of angles corresponding to the gray area in Fig. 1(b)) characterizing the dynamics of change of a trend identified using the algorithm shown as a pseudocode in Fig. 2 into a fuzzy set (linguistic label) best matching a given angle. We can use, for instance, some measure of a distance or similarity, cf. the

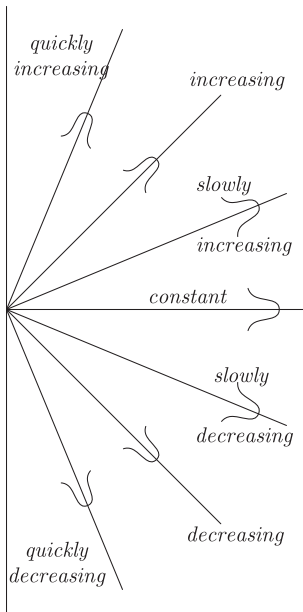


Fig. 3. A visual representation of angle granules defining the dynamics of change.

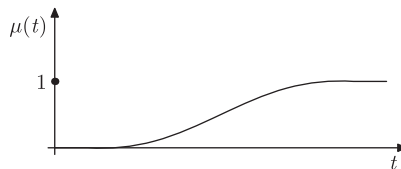


Fig. 4. An example of a membership function describing the term “long” concerning the trend duration.

book by Cross and Sudkamp [5]. Then we say that a given trend is, e.g., “decreasing to a 0.8°”, if $\mu_{\text{decreasing}}(\alpha) = 0.8$, where $\mu_{\text{decreasing}}$ is the membership function of a fuzzy set representing “decreasing” that is a best match for angle α .

3.2. Duration

Duration describes the length of a single trend, meant as a linguistic variable whose linguistic value (label) may be exemplified by a “long trend” defined as a fuzzy set whose membership function may be as in Fig. 4 in which the time axis is divided into appropriate units (time segments).

The definitions of linguistic terms describing the duration depend clearly on the perspective or purpose assumed by the user.

3.3. Variability

Variability refers to how “spread out” (“vertically”, in the sense of values taken on) a group of data is. The following five statistical measures of variability are widely used in traditional analyses:

- The range (maximum–minimum). Although the range is computationally the easiest measure of variability, it is not widely used, as it is based on only two data points that are extreme. This makes it very vulnerable to outliers and therefore may not adequately describe the true variability.
- The interquartile range (IQR) calculated as the third quartile (the third quartile is the 75th percentile) minus the first quartile (the first quartile is the 25th percentile) that may be interpreted as representing the middle 50% of the data. It is resistant to outliers and is computationally as easy as the range.

- The variance is calculated as

$$\frac{\sum_i (x_i - \bar{x})^2}{n},$$

where \bar{x} is the mean value.

- The standard deviation, i.e., the square root of the variance. Both the variance and the standard deviation are affected by extreme values.
- The mean absolute deviation (MAD), calculated as

$$\frac{\sum_i |x_i - \bar{x}|}{n}.$$

It is not frequently encountered in mathematical statistics. This is essentially because while the mean deviation has a natural intuitive definition as the “mean deviation from the mean”, the introduction of the absolute value makes analytic calculations using this statistic more complicated.

We propose to measure the variability of a trend as the distance of the data points covered by this trend from a linear uniform ε -approximation (cf. Section 2) that represents a given trend. For this purpose we propose to employ a distance between a point and a family of possible solutions, indicated as a gray cone in Fig. 1. Eq. (1) assures that the distance is definitely smaller than ε . We may use this information for the normalization. The normalized distance equals 0 if the point lays in the gray area. In the opposite case it is equal to the distance to the nearest point belonging to the cone, divided by ε . Alternatively, we may bisect the cone and then compute the distance between the point and this ray.

Similarly as in the case of dynamics of change, we find for a given value of variability obtained as above a best matching fuzzy set (linguistic label) using, e.g., some measure of a distance or similarity, cf. the book by Cross and Sudkamp [5]. Again, the measure of variability is treated as a linguistic variable and expressed using linguistic terms (labels) modelled by fuzzy sets defined by the user.

4. Linguistic data summaries

A linguistic summary is meant as a (usually short) natural language like sentence (or some sentences) that subsumes the very essence of a set of data (cf. [16,17]). This data set is numeric and usually large, not comprehensible in its original form by the human being. In Yager’s [22] basic approach (cf. [14,15] for an extended, more implementable exposition) the following perspective for linguistic data summaries is assumed:

- $Y = \{y_1, \dots, y_n\}$ is a set of objects (records) in a database, e.g., the set of workers;
- $A = \{A_1, \dots, A_m\}$ is a set of attributes characterizing objects from Y , e.g., salary, age, etc. in a database of workers, and $A_j(y_i)$ denotes a value of attribute A_j for object y_i .

A linguistic summary of a data set consists of:

- a *summarizer* P , i.e., an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_j (e.g., “low salary” for attribute “salary”);
- a *quantity in agreement* Q , i.e., a linguistic quantifier (e.g., most);
- *truth* (validity) \mathcal{T} of the summary, i.e., a number from the interval $[0, 1]$ assessing the truth (validity) of the summary (e.g., 0.7); usually, only summaries with a high value of \mathcal{T} are interesting;
- optionally, a *qualifier* R , i.e., another attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_k determining a (fuzzy subset) of Y (e.g., “young” for attribute “age”).

Thus, a linguistic summary may be exemplified by

$$\mathcal{T}(\text{most of employees earn low salary}) = 0.7 \tag{2}$$

or, in a richer (extended) form, including a qualifier (e.g., young), by

$$\mathcal{T}(\text{most of young employees earn low salary}) = 0.9. \tag{3}$$

Thus, basically, the core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh [26] which, for (2), may be written as

$$Qy's \text{ are } P \tag{4}$$

and for (3), may be written as

$$QRy's \text{ are } P. \tag{5}$$

Then, \mathcal{T} , i.e., the truth (validity) of a linguistic summary, directly corresponds to the truth value of (4) or (5). This may be calculated by using either the original Zadeh's calculus of linguistically quantified propositions (cf. [26]) or other tools for dealing with linguistic quantifiers. In the former case, the truth values (from [0, 1]) of (4) and (5) are calculated, respectively, as

$$\mathcal{T}(Qy's \text{ are } P) = \mu_Q \left(\frac{1}{n} \sum_{i=1}^n \mu_P(y_i) \right), \tag{6}$$

$$\mathcal{T}(QRy's \text{ are } P) = \mu_Q \left(\frac{\sum_{i=1}^n (\mu_R(y_i) \wedge \mu_P(y_i))}{\sum_{i=1}^n \mu_R(y_i)} \right), \tag{7}$$

where \wedge is the minimum operation (more generally it can be another appropriate operation, notably a t-norm), and Q is a fuzzy set representing the linguistic quantifier in the sense of Zadeh [26], i.e., $\mu_Q: [0, 1] \rightarrow [0, 1], \mu_Q(x) \in [0, 1]$. We consider *regular nondecreasing monotone* quantifiers such that

$$\mu(0) = 0, \quad \mu(1) = 1, \tag{8}$$

$$x_1 \leq x_2 \Rightarrow \mu_Q(x_1) \leq \mu_Q(x_2). \tag{9}$$

They can be exemplified by “most” given as in (10):

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8, \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8, \\ 0 & \text{for } x \leq 0.3. \end{cases} \tag{10}$$

5. Protoforms of linguistic trend summaries

It was shown by Kacprzyk and Zadrozny [16] that Zadeh's [27] concept of a protoform is convenient for dealing with linguistic summaries. This approach is also employed here.

Basically, a protoform is defined as a more or less abstract prototype (template) of a linguistically quantified proposition. Then, the summaries mentioned above might be represented by the two types of protoforms:

- Frequency based summaries:
 - a protoform of a short form of linguistic summaries:

$$Q \text{ trends are } P \tag{11}$$

and exemplified by:

Most of trends are of a large variability

- a protoform of an extended form of linguistic summaries:

$$QR \text{ trends are } P \tag{12}$$

and exemplified by:

Most of slowly decreasing trends are of a large variability

- Duration based summaries:
 - a protoform of a short form of linguistic summaries:

$$\text{Trends that took } Q \text{ time are } P \tag{13}$$

and exemplified by:

Trends that took *most* of the time are of a *large variability*

- a protoform of an extended form of linguistic summaries:

$$R \text{ trends that took } Q \text{ time are } P \tag{14}$$

and exemplified by:

Slowly decreasing trends that took *most* of the time are of a *large variability*.

It should be noted that the latter summaries should be properly understood as, e.g., for (14), that the slowly decreasing (partial) trends that altogether took most of the time have a large variability.

The truth values of the above types and forms of linguistic summaries will be found using the classic Zadeh’s calculus of linguistically quantified propositions as it is effective and efficient, and provides the best conceptual framework within which to consider a linguistic quantifier driven aggregation of partial trends that is the crucial element of our approach.

It should be noted that the protoforms of linguistic summaries given above are not the only possible ones though they are intuitively appealing to domain experts in many fields, including finance we will be dealing with, and they can be represented and formally processed in quite a straightforward way using our means for dealing with linguistically quantified propositions.

6. The use of Zadeh’s calculus of linguistically quantified propositions

Using Zadeh’s [26] fuzzy logic based calculus of linguistically quantified propositions, a (proportional, nondecreasing) linguistic quantifier Q is assumed to be a fuzzy set defined in the unit interval $[0, 1]$ as, e.g., (10).

The truth values (from $[0,1]$) of (11) and (12) are calculated, respectively, as

$$\mathcal{T}(Qy's \text{ are } P) = \mu_Q \left(\frac{1}{n} \sum_{i=1}^n \mu_P(y_i) \right), \tag{15}$$

$$\mathcal{T}(QRy's \text{ are } P) = \mu_Q \left(\frac{\sum_{i=1}^n (\mu_R(y_i) \wedge \mu_P(y_i))}{\sum_{i=1}^n \mu_R(y_i)} \right), \tag{16}$$

where \wedge is the minimum operation.

The computation of truth values of duration based summaries is more complicated and requires a different approach. While analyzing a summary “the trends that took Q time are P ” we should compute the time which is taken by those trends for which “trend is P ” is valid. It is obvious that when “trend is P ” is to degree 1, then we can use the whole time span taken by this trend. However, what should we do if “trend is P ” is to some degree? We propose to take only a part of the time span defined by the degree to which “trend is P ”. In other words we compute this time as $\mu(y_i)t_{y_i}$, where t_{y_i} is the duration of trend y_i . The obtained value (duration of those trends for which “trend is P ”) is then normalized by dividing it by the overall time T . Finally, we may compute to which degree the time taken by those trends which “trend is P ” is Q . A similar line of thought might be followed for the extended form of linguistic summaries.

The truth value of the short form of duration based summaries (13) is calculated as

$$\mathcal{T}(y \text{ that took } Q \text{ time are } P) = \mu_Q \left(\frac{1}{T} \sum_{i=1}^n \mu_P(y_i)t_{y_i} \right), \tag{17}$$

where T is the total time of the trends summarized and t_{y_i} is the duration of the i th trend.

The truth value of the extended form of duration based summaries (14) is calculated as

$$\mathcal{T}(\text{Ry that took } Q \text{ time are } P) = \mu_Q \left(\frac{\sum_{i=1}^n (\mu_R(y_i) \wedge \mu_P(y_i)) t_{y_i}}{\sum_{i=1}^n \mu_R(y_i) t_{y_i}} \right), \tag{18}$$

where t_{y_i} is the duration of the i th trend.

Both the fuzzy predicates P and R are assumed above to be of a rather simplified, atomic form referring to just one attribute. They can be extended to cover more sophisticated summaries involving some confluence of various, multiple attribute values as, e.g., “slowly decreasing and short”.

Alternatively, we may obtain the truth values of (13) and (14) if we divide each trend which takes t_{y_i} time units into t_{y_i} trends, each lasting one time unit. For this new set of trends we use frequency based summaries with the truth values defined in (15) and (16).

It may readily be noticed that though in the source Zadeh’s fuzzy logic based calculus of linguistically quantified propositions the “ \wedge ” (minimum) operation is used, which is well founded and intuitively appealing, other appropriate operations can also be used, notably the t-norms.

A t-norm is defined as

$$t: [0, 1] \times [0, 1] \longrightarrow [0, 1] \tag{19}$$

such that, for each $a, b, c \in [0, 1]$:

1. it has 1 as the unit element, i.e.,

$$t(a, 1) = a,$$

2. it is monotone, i.e.,

$$a \leq b \implies t(a, c) \leq t(b, c),$$

3. it is commutative, i.e.,

$$t(a, b) = t(b, a),$$

4. it is associative, i.e.,

$$t[a, t(b, c)] = t[t(a, b), c].$$

Evidently, a t-norm is monotone non-decreasing in both arguments, and $t(a, 0) = 0$.

Some more relevant examples of t-norms are:

- the minimum

$$t(a, b) = a \wedge b = \min(a, b), \tag{20}$$

which is the most widely used, also in our context,

- the algebraic product

$$t(a, b) = a \cdot b, \tag{21}$$

- the Łukasiewicz t-norm

$$t(a, b) = \max(0, a + b - 1), \tag{22}$$

- the drastic t-norm

$$t(a, b) = \begin{cases} b & a = 1, \\ a & b = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

These operations can be in principle used in Zadeh's calculus but, clearly, their use may result in different results of the linguistic quantifier driven aggregation. One should, however, be cautious in real world applications because the very meaning (semantics) of t-norms maybe not be obvious to domain experts, maybe except for the minimum. This will be discussed later on, and some examples will be shown in the next section.

7. Numerical experiments

The application area in which our work has been applied is related to investment (mutual) funds, and the domain experts have been fund analysts (and, to some extent, fund managers). We have had to follow to a large extent their line of thought, interests, and also some established practices and pragmatics in that area.

Among a number of results obtained, we will propose here an example of using the method proposed to data coming from quotations of an investment (mutual) fund that invests at most 50% of assets in shares. Data shown in Fig. 5 were collected from April 1998 until December 2006 with the value of one share equal to PLN 10.00 in the beginning of the period to PLN 45.10 at the end (PLN stands for the Polish Zloty). The minimal value recorded was PLN 6.88 while the maximal one during this period was PLN 45.15. The biggest daily increase was equal to PLN 0.91, while the biggest daily decrease was equal to PLN 2.41. Needless to say that in the period covered by our analyses there was a good situation at stock exchanges around the world, and this clearly had an impact on the results obtained.

It should be noted that the example shown below is meant to illustrate the method proposed by analyzing the absolute performance of a given investment fund. We do not deal here with a presumably more common way of analyzing an investment fund by relating its performance to a benchmark (or benchmarks) exemplified by an average performance of a group of (similar) funds, a stock market index or a synthetic index reflecting, for instance, the bond versus stock allocation.

Using the Sklansky and Gonzalez algorithm and $\varepsilon = 0.25$ we obtained 255 extracted trends. The shortest trend took two time units only, while the longest one—71. The histogram of duration of trends is presented in Fig. 6.

Fig. 7 shows the histogram of angles which characterize the dynamics of change.

The histogram of variability of trends (in percents) is presented in Fig. 8. Some interesting short form summaries obtained by using the method proposed, employing the classic Zadeh's calculus of linguistically quantified propositions, and for different granulations of the dynamics of change, duration and variability, are:

- for seven labels for the dynamics of change (*quickly increasing, increasing, slowly increasing, constant, slowly decreasing, decreasing and quickly decreasing*), five labels for the duration (*very long, long, medium, short,*

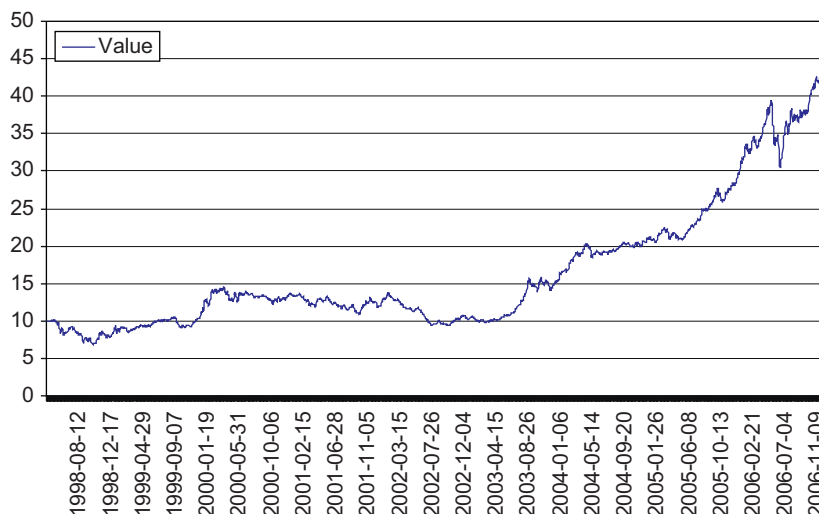


Fig. 5. A view of the original data (daily quotations).

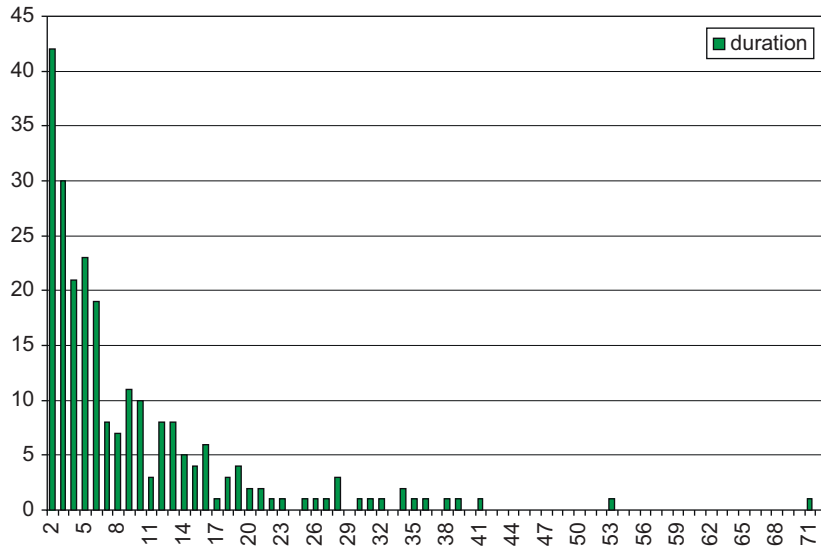


Fig. 6. Histogram of the duration of trends (in the number of days).

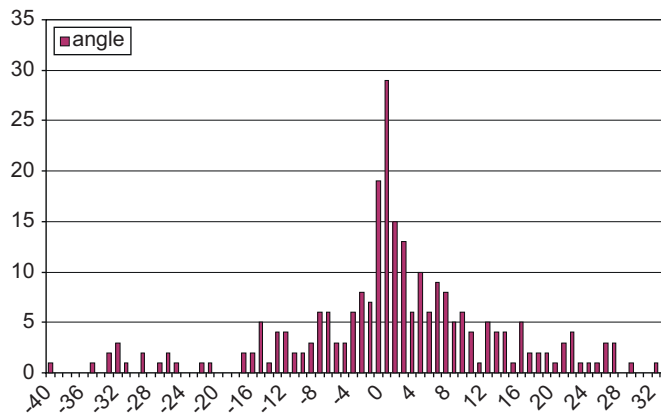


Fig. 7. Histogram of angles (in degrees) describing the dynamic of change.

very short) and five for the variability (very high, high, medium, low, very low):

- Most trends are very short, $\mathcal{T} = 0.78$.
- Trends that took almost all of the time are constant, $\mathcal{T} = 0.639$.
- Trends that took at least a half of the time are of a low variability, $\mathcal{T} = 0.873$.

Some extended form summaries for this granulation and for different t-norms (minimum, product, Łukasiewicz and drastic) are shown in Table 1.

- Five labels for the dynamics of change (increasing, slowly increasing, constant, slowly decreasing, decreasing), three labels for the duration (short, medium, long) and five labels for the variability (very high, high, medium, low, very low):

- Trends that took most of the time are constant, $\mathcal{T} = 0.692$.
- Trends that took most of the time are of a medium length, $\mathcal{T} = 0.506$.

Some extended form summaries for this granulation and different t-norms (minimum, product, Łukasiewicz and drastic) are shown in Table 2.

As it can be seen, the results obtained, that is the particular linguistic summaries and their associated truth values, are intuitively appealing while looking at the time series under consideration. In addition, these summaries have been

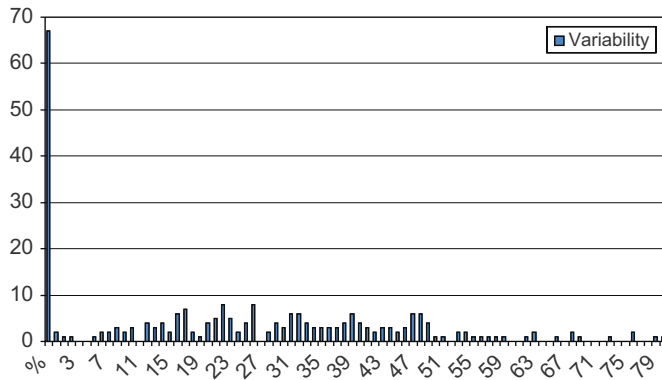


Fig. 8. Histogram of the variability (IQR) of trends.

Table 1

Truth values obtained for extended form summaries with different t-norms used for the first granulation (seven labels for the dynamics of change, five labels for the duration and five labels for the variability)

Summary	Minimum	Product	Łukasiewicz	Drastic
Most trends with a low variability are constant	0.974	0.944	0.911	0.85
Most slowly decreasing trends are of a very low variability	0.636	0.631	0.63	0.589
Almost all short trends are constant	1	1	1	1
Decreasing trends that took most of the time are of a very low variability	0.989	0.989	0.989	0.989
Trends with a low variability that took almost all of the time are constant	1	1	0.994	0.868
Trends with a very high variability that took most of the time are constant	0.94	0.94	0.94	0.94

Table 2

Truth values obtained for extended form summaries with different t-norms used for the second granulation (five labels for the dynamics of change, three labels for the duration and five labels for the variability)

Summary	Minimum	Product	Łukasiewicz	Drastic
Almost all decreasing trends are short	1	1	1	1
Almost all increasing trends are short	0.58	0.514	0.448	0.448
At least a half of medium length trends are constant	0.891	0.877	0.863	0.863
Most of slowly increasing trends are of a medium length	0.798	0.773	0.748	0.748
Most of trends with a low variability are constant	0.567	0.517	0.466	0.466
Most of trends with a very low variability are short	0.909	0.9	0.891	0.891
Most trends with a high variability are of a medium length	0.801	0.754	0.707	0.707
None of trends with a very high variability is long	1	1	1	1
None of decreasing trends is long	1	1	1	1
None of increasing trends is long	1	1	1	1
Decreasing trends that took most of the time are of a very low variability	0.798	0.796	0.788	0.788
Constant trends that took most of the time are of a low variability	0.5	0.466	0.431	0.324
Trends with a low variability that took most of the time are constant	0.898	0.851	0.804	0.658

found interesting by domain experts but a detailed analysis from the point of view of financial analyses is beyond the scope of this paper. The results obtained for different t-norms are similar and, of course, the truth value for the case of the minimum is the highest.

One should, however, bear in mind that the analysis of the impact of choosing a particular t-norm is somehow tricky. First, there are many excellent theoretical analyses of how various t-norms behave but in any practical application like ours the main problem is completely different. Namely, though the very essence of a t-norm as an aggregation operator may be clear, for domain experts the only t-norm that has a clear semantics is presumably the minimum as

it can be related to a well known (e.g., from decision analysis) and intuitively appealing attitude of a decision maker of the pessimistic or safety first type. Domain experts can eventually understand the effect of choosing the algebraic product as something “milder”, less pessimistic or safety-first, than the minimum. Unfortunately, the semantics of other t-norms, that exhibit very interesting formal properties, is unclear to the users.

Of course, this should be properly understood. Namely, though the semantics of those other t-norms may be unclear and not comprehensible, the use of the method proposed by employing various t-norms and evaluating the results obtained may persuade the user to even adopt a t-norm that may have given consistently good results even if its semantics may be unclear. This is, however, beyond the scope of this paper as it is related to some “learning” and, above all, to a practical use (implementation).

8. Concluding remarks

We proposed new types of linguistic summaries of time series. The derivation of a linguistic summary of a time series was related to a linguistic quantifier driven aggregation of trends, and we employed the classic Zadeh’s calculus of linguistically quantified propositions with different t-norms, not only the classic minimum. We showed an application to the analysis of time series data on daily quotations of an investment (mutual) fund over an eight year period, presented some interesting linguistic summaries obtained, and showed results for different t-norms. They suggest that various t-norms exhibit slightly different behavior and their choice may be relevant for a particular application. However, as already mentioned, while working with domain experts—which is crucial in our application—one should take into account that except for the minimum, the semantics of other t-norms may be unclear to them. The results obtained seem to be very promising.

References

- [1] I. Batyrshin, On granular derivatives and the solution of a granular initial value problem, *Internat. J. Appl. Math. Comput. Sci.* 12 (3) (2002) 403–410.
- [2] I. Batyrshin, L. Sheremetov, Perception based functions in qualitative forecasting, in: I. Batyrshin, J. Kacprzyk, L. Sheremetov, L.A. Zadeh (Eds.), *Perception-based Data Mining and Decision Making in Economics and Finance*, Springer, Berlin, Heidelberg, 2006.
- [5] V. Cross, T. Sudkamp, *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications*, Springer, Heidelberg, New York, 2002.
- [6] D.-A. Chiang, L.R. Chow, Y.-F. Wang, Mining time series data by a fuzzy linguistic summary system, *Fuzzy Sets and Systems* 112 (2000) 419–432.
- [9] J. Kacprzyk, A. Wilbik, S. Zadrożny, Linguistic summarization of trends: a fuzzy logic based approach, in: *Proc. 11th Internat. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, Paris, France, July 2–7, 2006, pp. 2166–2172.
- [10] J. Kacprzyk, A. Wilbik, S. Zadrożny, Linguistic summaries of time series via a quantifier based aggregation using the Sugeno integral, in: *Proc. of 2006 IEEE World Congress on Computational Intelligence*, Vancouver, BC, Canada, IEEE Press, New York, July 16–21, 2006, pp. 3610–3616.
- [11] J. Kacprzyk, A. Wilbik, S. Zadrożny, On some types of linguistic summaries of time series, in: *Proc. of the Third International IEEE Conf. on Intelligent Systems*, IEEE Press, New York, London, UK, 2006, pp. 373–378.
- [12] J. Kacprzyk, A. Wilbik, S. Zadrożny, A linguistic quantifier based aggregation for a human consistent summarization of time series, in: J. Lawry, E. Miranda, A. Bugarin, S. Li, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), *Soft Methods for Integrated Uncertainty Modelling*, Springer, Berlin, Heidelberg, 2006, pp. 186–190.
- [13] J. Kacprzyk, A. Wilbik, S. Zadrożny, Capturing the essence of a dynamic behavior of sequences of numerical data using elements of a quasi-natural language, in: *Proc. 2006 IEEE Internat. Conf. on Systems, Man, and Cybernetics*, Taipei, Taiwan, IEEE Press, New York, 2006, pp. 3365–3370.
- [14] J. Kacprzyk, R.R. Yager, Linguistic summaries of data using fuzzy logic, *Internat. J. General Systems* 30 (2001) 33–154.
- [15] J. Kacprzyk, R.R. Yager, S. Zadrożny, A fuzzy logic based approach to linguistic summaries of databases, *Internat. J. Appl. Math. Comput. Sci.* 10 (2000) 813–834.
- [16] J. Kacprzyk, S. Zadrożny, Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools, *Inform. Sci.* 173 (2005) 281–304.
- [17] J. Kacprzyk, S. Zadrożny, Fuzzy linguistic data summaries as a human consistent, in: B. Gabrys, K. Leiviska, J. Strackeljan (Eds.), *Do Smart Adaptive Systems Exist?*, Springer, Berlin, Heidelberg, New York, 2005, pp. 321–339.
- [18] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, Locally adaptive dimensionality reduction for indexing large time series databases, in: *Proc. of ACM SIGMOD Conf. on Management of Data*, Santa Barbara, CA, 2001, pp. 151–162.
- [19] E. Keogh, M. Pazzani, An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback, in: *Proc. Fourth Internat. Conf. on Knowledge Discovery and Data Mining*, New York, NY, 1998, pp. 239–241.
- [20] J. Sklansky, V. Gonzalez, Fast polygonal approximation of digitized curves, *Pattern Recognition* 12 (5) (1980) 327–331.
- [21] S. Sripada, E. Reiter, I. Davy, SumTime-mousam: configurable marine weather forecast generator, *Expert Update* 6 (3) (2003) 4–10.
- [22] R.R. Yager, A new approach to the summarization of data, *Inform. Sci.* 28 (1982) 69–86.

- [26] L.A. Zadeh, A computational approach to fuzzy quantifiers in natural languages, *Comput. Math. Appl.* 9 (1983) 149–184.
- [27] L.A. Zadeh, A prototype-centered approach to adding deduction capabilities to search engines—the concept of a protoform, in: *Proc. of the Annu. Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2002)*, 2002, pp. 523–525.
- [29] in: L.A. Zadeh, J. Kacprzyk (Eds.), *Computing with Words in Information/Intelligent Systems: 1. Foundations*, Physica-Verlag, Heidelberg, New York, 1999.
- [30] in: L.A. Zadeh, J. Kacprzyk (Eds.), *Computing with Words in Information/Intelligent Systems: 2. Applications*, Physica-Verlag, Heidelberg, New York, 1999.