Raport Badawczy
Research Report

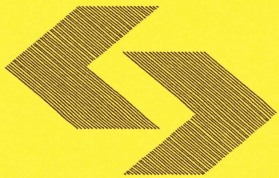
RB/41/2015

Construction of prior model probability distributions for time series

K. Kaczmarek, O. Hryniewicz

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2015

Abstract

Practitioners are very often posed to the dilemma of choice between the wealth of mathematic models for time series forecasting. They choose the forecasting method and its assumptions based on some prior information from the literature and rough guesses. The Bayesian approach enables formalizing this approach and expressing prior knowledge as probability distributions. Unfortunately, proper selection of the prior probability distributions may become very challenging task. Within this research, linguistic summaries are successfully incorporated into the construction of the prior information for forecasting models. Instead of defining probability distributions, the user validates expected trends and the system creates the probability distributions automatically. Linguistic summaries are constructed as linguistically quantified sentences mined from databases, that may be exemplified by *'Among all increasing trends, most are short'*. The proposed approach is evaluated with experiments on real-life time series from the pharmaceutical market and the M3 Competition benchmark datasets. The results confirm that the incorporation and processing of the linguistic summaries increases the interpretability of the forecasting process and may improve its accuracy.

1 Introduction

Recalling the 'No Free Lunch' theorem of Wolpert [Wolpert(1996)] and applying the metaphor for the mathematical forecasting problems, the cost of finding a solution, averaged over all problems, is the same for any method, and there is no method that is best for any mathematical problem. However, for some classes of problems, the selected forecasting methods outperform the others. Choosing the proper forecasting method for the given problem becomes very challenging task.

For a recent review of competitive forecasting models and methods, see e.g., Gooijer and Hyndman [Gooijer and Hyndman(2006)]. Among competitive methods and models for forecasting, the Bayesian approach has been proven successful in various practical applications including industry, financial market, quality control, healthcare and more [Box and Tiao(1973), Geweke(2005), Petridis et al.(2001)Petridis, Kehagias, Petrou, Bakirtzis, Kiartzis, Panagiotou et al.]. The main principles for the Bayesian approach are to express all assumptions using probability statements, and then to design distribution for the future events conditional on the observed values and a certain loss function. However, the main technical obstacles are in the expression of assumptions and the proper selection of the prior proba-

bility distributions for the unknown variables. Usually, the assumptions are expressed by the analysts or experts of the field based mostly on their expertise and intuitions. In practice, the problem arises when experts fail to adequately establish the prior probability distributions. It is still one of the main challenges in the Bayesian forecasting.

Furthermore, the research on psychology of decision-making proves that people may take irrational decisions and be influenced by emotions. Therefore, simple and easily interpretable tools are required to support the selection of the prior knowledge for the forecasting process. Hopefully, there is a wealth of modern data mining techniques that discover interesting imprecise knowledge from large datasets, and the literature on discovery of summaries about time series data is extensive, see e.g., [Nauck and Kruse(2014), Kempe et al.(2008)Kempe, Hipp, Lanquillon and Kruse, Moewes and Kruse(2009)].

The objective of this paper is the introduction of the predictive method for time series with the use of linguistic summaries being human-consistent results of data-mining.

This paper is continuation of our previous works [Hryniewicz and Kaczmarek(2014), Kaczmarek et al.(2015)Kaczmarek, Hryniewicz and Kruse]. Nonetheless, the method presented in this paper is applicable for time series of any length. The proposed method is in line with the granular computing approach for time series forecasting introduced in [Hryniewicz and Kaczmarek(2015)]. However, the information granules considered in this paper are limited to linguistic summaries, and the prior model probability distributions are constructed automatically with their use. The calculations are transparent for experts, nonetheless, no user input is obligatory for generating forecasts.

An illustrative overview of the proposed method is presented in Figure 1.

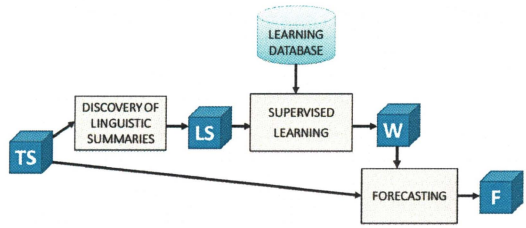


Figure 1: Overview of the proposed method.

The performance of the F-LS method is illustrated with the study for the real-life sales time series from the pharmaceutical market. The proposed

approach is evaluated also with experiments for benchmark time series from the M3-Competition repository. The numerical results of the forecast accuracy show that the proposed approach with linguistic summaries may lead to the increase of the accuracy compared to the benchmark methods. None of the studied benchmark methods outperforms or dominates the proposed F-LS method. F-LS approach delivers forecast on the similar accuracy level as the ForecastX method, which has scored the best results (on average) in the whole competition. At the same time, it is observed that for real-life time series, the approach delivers forecasts with reasonably small error.

The structure of this paper is as follows. Next Chapter briefly explains the discovery of linguistic summaries from the time series datasets. In Chapter 3, the supervised learning module of the proposed approach is explained. Then, in Chapter 4, the Bayesian procedure for forecasting with linguistic summaries is presented. The numerical results of the experiments are described in Chapter 5. This paper concludes with general remarks and further research opportunities in Chapter 6.

2 Discovery of Linguistic Summaries

When analyzing time series and performing the visual inspection, people perceive and process shapes rather than single data points [Attneave(1954)] and easily describe the evolution of time series with adjectives, like *increasing trend, constant, decreasing, long, short, high values, medium, low, interesting feature, strong, weak, slight, etc.* People have the ability to reason in the imprecise and uncertain environment and process such labels that refer to e.g., imprecise values, trends, judgments or features.

The potential applications are summarization data from sensors e.g., in eldercare to monitor the patients health. The linguistic summary can take the following form: *'On most nights the resident had a high level of restlessness.'* cf. Wilbik and Keller [Wilbik and Keller(2012)].

The knowledge discovery process can be divided into segmentation, clustering, classification of identified meaningful intervals or patterns, detecting anomalies, frequent patterns and discovery of association rules [Batyrshin and Sheremetov(2008)]. The data mining tools enable generation of various semantic and lexico-grammar structures summarizing large datasets. To formally describe the quantified sentences, we adapt the classic calculus of linguistically quantified propositions and the concept of the protoform in the sense of Yager [Yager(1982)] developed by Kacprzyk et al. [Kacprzyk(2008), Kacprzyk and Wilbik(2009), Kacprzyk et al.(2011)Kacprzyk, Wilbik, Partyka and Ziółkowski].

Let us now recall that a discrete time series is a sequence $y = \{y_i\}_{i=1}^m$ of observations measured at successive $\{1, ..., m\}$ moments and at uniform time intervals. Linguistic summaries describe general facts about evolution of time series with quasi natural language. The formal definitions are as follows.

Linguistic summary [Yager(1982)]

Let $O = \{o_1, o_2..., o_b\}$ denote a finite set of objects (e.g., trends of sequence data) in a considered domain. The properties of objects are measured by observables and are called attributes. Let $A = \{a_1, a_2..., a_r\}$ denote a finite set of attributes (e.g., dynamics of change, duration), and $S = \{s_1, s_2..., s_l\}$ is a finite set of imprecise labels for attributes (e.g., *quickly increasing, short*). The protoform-based linguistic summary

$$LS : Q \ o \ are \ P \tag{1}$$

consists of a quantity in agreement Q (quantifier like e.g., *most, among all*), summarizer P (attribute together with an imprecise label) about objects $o \in O$ and a measure T of validity or truth of the summary.

Alternative philosophy about the linguistic summarization is provided by Dubois and Prade [Dubois and Prade(1992)]. The authors introduce and discuss the gradual inference rules of the form *'The more X is F, the more Y is G'* or *'The less X is F, the less Y is G'*, where F, G are gradual properties and X, Y are entities satisfying them to some degree.

The data mining can produce a lot of summaries and it is always important to evaluate their interestingness. There have been proposed different measures to evaluate the quality of linguistic summaries. One of the first measures introduced by Zadeh is the degree of truth (validity) defined as

$$T(LS) = \mu_Q(\frac{\sum_{i=1}^n (\mu_R(y_n) \wedge \mu_P(y_n))}{\sum_{i=1}^n \mu_R(y_n)}) \tag{2}$$

where $\mu_R(y_n), \mu_P(y_n)$ are the membership functions determining the degree to which R, P respectively, are satisfied for the time series y at the given moment n.

Other measures that are commonly used in applications are e.g., the support (covering) defined as follows:

$$d_{sup}(LS) = \frac{1}{n} card\{y : \mu_P(y_n) > 0 \wedge \mu_R(y_n) > 0\} \tag{3}$$

or the degree of imprecision (fuzziness) [Kacprzyk and Zadrożny(2005)]:

$$d_{imp}(LS) = 1 - \sqrt[m]{\prod_{j=1,...,m} \frac{card\{x \in X_j : \mu_{s_j}(x) > 0\}}{card X_j}} \tag{4}$$

where summarizer P is given as a family of fuzzy sets $P = \{s_1, s_2, ..., s_m\}$ and *card* denotes the cardinality of the corresponding (non-fuzzy) set.

Other quality measures (the degree of informativeness, the degree of specificity, the degree of appropriateness, length of the summary) could be also considered. For definitions and the review on alternative quality measures, refer to Yager [Yager(1982)], Kacprzyk and Zadrożny [Kacprzyk and Zadrożny(2005)] and the Chapter 4.3. of the Ph.D. thesis by Wilbik [Wilbik(2010)].

3 Supervised Learning of Probabilistic Models

The algorithm starts from the definition of imprecise concepts that describe the trends and linguistic summaries. Secondly, the preprocessing of the time series data is performed to ensure that they are normalized and without missing values. Next is the supervised learning of the probabilistic models. Its goal is to build the training database and to discover rules enabling the classification of the probabilistic models based on the sets of linguistic summaries describing the evolution of time series.

Then, the mining for the human-consistent prior information is performed. Its goal is to discover and validate with experts the linguistic summaries about the expected evolution of the predicted time series. Next, the prior probability distributions are calculated. Finally, Markov Chain Monte Carlo Posterior Simulation is run to simulate the posterior probability distributions for the vector of interest and calculate the forecast y_{n+1}.

Discrete time series

Let $O = \{o_1, o_2..., o_q\}$ denote a finite set of objects in a considered domain. The properties of objects are measured by observables $P = \{p_1, p_2..., p_r\}$. Discrete time series $\boldsymbol{y} = \{y_t\}_{t=1}^{n} \in \Psi_n$ is a sequence of observations of given object's property (o, p) such that $o \in O$ and $p \in P$ measured at successive $t \in T = \{1, ..., n\}$ moments and at uniform time intervals. For each $t \in T$ the observation y_t is a realization of the random variable Y_t defined on the probability space (Ω, A, P). A sequence of Y_t formulates a stochastic process.

4 Forecasting with Linguistic Summaries

We adapt the Bayesian approach. Forecasts for future events ω are generated in the process of the Bayesian inference due to the following posterior density

of the vector of interest:

$$p(\omega|y, M) = \sum_{j=1}^{J} p(M_j|y, M)p(\omega|y, M_j) \tag{5}$$

This posterior density $p(\omega|y, M)$ is a weighted average of the posterior densities of models $\{M_1, M_2, ..., M_J\}$ which are defined as follows:

$$p(M_j|y, M) = \frac{p(M_j)p(y|M_j)}{p(y|M)} = \frac{p(M_j)p(y|M_j)}{\sum_{j=1}^{J} p(M_j)p(y|M_j)} \tag{6}$$

where $p(M_j); M_j \in M$ need to be defined a priori.

Usually, the definitions for the prior probabilities are based on subjective expert's experience, some combination of relevant data, information from the literature and rough guesses [Kass and Raftery(1995)]. Nonetheless, such subjective definitions might be inaccurate, and at the same time, the proper selection of a priori distributions is essential for good performance of the overall Bayesian forecasting process. The theoretical and empirical evidence show that prior assumptions for Bayesian model averaging are critically important [Ley and Steel(2009)].

The input for the algorithm is the discrete time series for prediction y and the set of template probabilistic models M, that need to be defined a priori. Within this research, we focus on supporting forecasting of short time series assuming that $n_{min} = 10, n_{max} = 20$.

Bayesian averaging [Geweke(2005)]

The posterior density of the vector of interest ω *conditional on multiple competitive probabilistic models for forecasting* $M = \{M_1, M_2, ..., M_J\}$ *is defined as follows*

$$p(\omega|y, M) = \sum_{j=1}^{J} p(M_j|y, M)p(\omega|y, M_j) \tag{7}$$

This posterior density $p(\omega|y, M)$ is a weighted average of the posterior densities of models $\{M_1, M_2, ..., M_J\}$ which are defined as follows:

$$p(M_j|y, M) = \frac{p(M_j)p(y|M_j)}{p(y|M)} = \frac{p(M_j)p(y|M_j)}{\sum_{j=1}^{J} p(M_j)p(y|M_j)} \tag{8}$$

We propose the approach to construct the prior model distributions $p(M_i), i \in \{1, ..., J\}$ with the support of the intelligent techniques.

Algorithm 1 Forecasting with Linguistic Summaries (F-LS) provides prediction y_{n+1}

Output: y_{n+1}

Algorithm:

Defining of imprecise concepts:
build_fuzzy_numbers (S)
Data preprocessing:
repeat difference(y) **until** y *is* validated
min-max normalization(y)
Supervised learning for the training database:
while $i \in J$ **do**
    T_m^s, C^s = generate k sample time series (M_i, k, m)
    LI^s = discover_linguistic_summaries (T_m^s)
    V^s = calculate_degree_of_truth (LI^s)
CL = supervised_learning_withSVM (C^s,V^s)
Imprecise knowledge retrieval from humans:
LI^E = create_provisional_linguistic_summaries (y)
v^E = calculate_degree_of_truth (LI^E)
T^E = expert_evaluation (LI^E, v^E) *HUMAN INPUT NEEDED*
while $i \in J$ **do**
    Sc^{M_i} = estimate_classification_scores (T^E, CL)
Posterior simulation and forecasting:
P=construct_prior_prob_distr (M, Sc^M)
y_{n+1}=MCMC_posterior_simulation(P, y)

5 Results and Discussion

The performance of the proposed method is illustrated with the experiments on the artificial and real-life time series datasets.

5.1 About Experiments and Datasets

The goal of the experiments is to provide an illustrative demonstration of the proposed approach and to evaluate its forecasting accuracy. The program has been implemented in Python with the support of NumPy, SciPy and the PyMc extension modules. Linguistic summaries are generated with the Trend Analysis System [Kacprzyk(2008)].

The following forecast accuracy measures are adapted to evaluate the performance:

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{h}\sum_{i=1}^{h}|\frac{100e_{t+i}}{y_{t+i}}| \tag{9}$$

- Symmetric Mean Absolute Percentage Error (sMAPE)

$$sMAPE = \frac{1}{h}\sum_{i=1}^{h}|\frac{200|e_{t+i}|}{(y_{t+i}+F(y)_i)}| \tag{10}$$

where h denotes the forecast horizon, $y_{t+i}, F(y)_i$ and $e_{t+i} = y_{t+i} - F(y)_i$ determine the actual value, the forecasted value and the error for the ith forecast.

Datasets considered for the experiments are as follows:

1. simulated time series;
2. pharmaceutical sales time series;
3. benchmark time series from the M3-Competition by [Makridakis and Hibon(2000)].

First, time series are simulated from the autoregressive processes to verify some theoretical assumptions for the proposed approach.

Secondly, monthly sales time series from pharmaceutical market are evaluated. The exemplary time series from the dataset are presented on Figure 3.

Alternative method has been proposed in [Kaczmarek and Hryniewicz(2013)].

Furthermore, the subset of yearly and monthly real-life time series from the M-3 Competition is considered. The experiments have been performed for the subset of the 10 first yearly time series of medium length (observations from 4 years) from the M3-Competition [Makridakis and Hibon(2000)] dataset repository.

5.2 Supervised Learning of Probabilistic Models

We consider following 6 imprecise labels in the experiment: *low, medium, high, increasing, constant, decreasing.* Values for imprecise labeled sequences referring to *increasing, constant, decreasing* labels are defined based on experts' subjective beliefs. For labels: *low, medium, high* triangular fuzzy

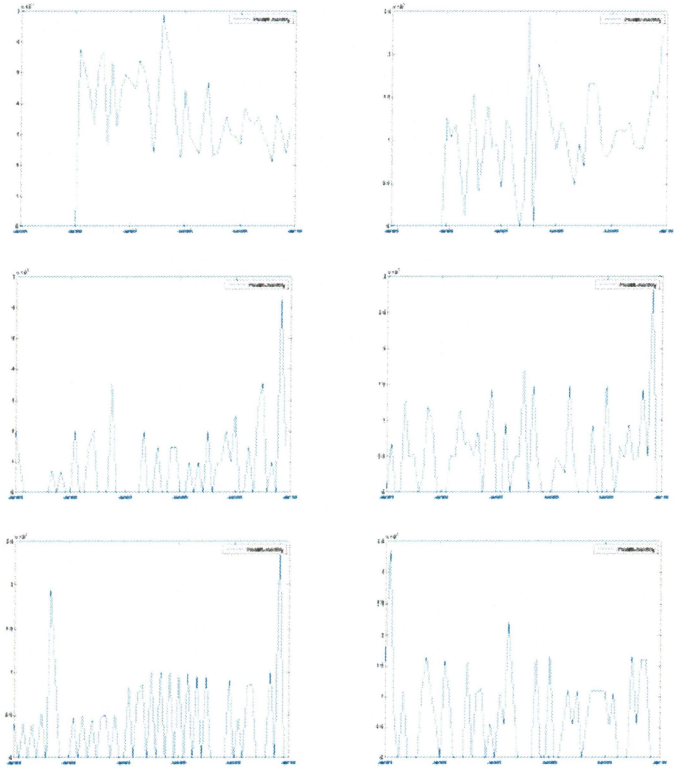


Figure 2: Sales time series representing monthly sales of Product 1-6 from Jan'05 to Dec'09.

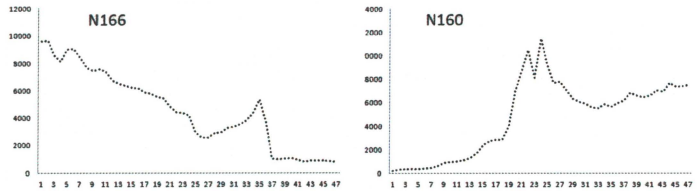


Figure 3: Exemplary medium long time series from M3 dataset: N160 and N166.

numbers are constructed based on the minimum, average and maximum values calculated from the time series. Then, imprecise labeled sequences are calculated from appropriate membership functions.

Types of linguistic summaries produced from different time series within the same model.

Table 1: Degree of truth for sample time series from process AR with $\phi_1 = 0.9$

Autoregressive coefficient	0.9										
Sample No.	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**	**Avg**
Among all const y, most are med	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Among all const y, most are med and mod	1.0	0.9	0.9	1.0	1.0	1.0	0.8	1.0	1.0	0.8	0.9
Among all const y, most are mod	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0
Among all const y, most are short	0.1	0.4	0.1	0.3	0.0	0.0	0.2	0.1	0.2	0.1	0.2
Among all const y, most are short and mod	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.2	0.0	0.1
Among all decr y, most are low	0.0	0.4	0.5	0.2	0.0	0.9	1.0	1.0	0.7	0.0	0.7
Among all decr y, most are med	1.0	0.3	0.7	0.6	1.0	0.6	0.9	0.2	0.6	1.0	0.7
Among all decr y, most are med and low	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.4
Among all decr y, most are med and mod	1.0	0.2	0.4	0.4	1.0	0.1	0.0	0.0	0.3	1.0	0.5
Among all decr y, most are mod	1.0	0.6	0.5	0.8	1.0	0.2	0.0	0.0	0.5	1.0	0.7
Among all decr y, most are short	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0
Among all decr y, most are short and low	0.0	0.4	0.5	0.2	0.0	0.8	0.8	1.0	0.7	0.0	0.6
Among all decr y, most are short and mod	0.9	0.5	0.0	0.1	1.0	0.0	0.0	0.0	0.2	0.0	0.6
Among all incr y, most are low	0.9	0.3	0.2	1.0	0.2	0.7	0.8	0.7	1.0	0.4	0.6
Among all incr y, most are med	0.6	0.9	0.6	0.6	0.5	0.4	0.4	0.3	0.4	0.9	0.6
Among all incr y, most are med and low	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.2
Among all incr y, most are med and mod	0.0	0.5	0.2	0.0	0.2	0.1	0.0	0.0	0.0	0.5	0.2
Among all incr y, most are mod	0.1	0.9	0.8	0.0	0.8	0.3	0.3	0.3	0.0	0.7	0.5
Among all incr y, most are short	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Among all incr y, most are short and low	0.8	0.2	0.2	1.0	0.1	0.7	0.8	0.7	1.0	0.3	0.6
Among all incr y, most are short and mod	0.1	0.8	0.4	0.0	0.6	0.0	0.0	0.1	0.0	0.0	0.3
Among all y, most are const	0.7	0.7	0.7	1.0	1.0	0.5	0.5	0.8	0.7	1.0	0.8
Among all y, most are const and mod	0.3	0.4	0.3	0.7	0.9	0.5	0.1	0.6	0.4	0.6	0.5
Among all y, most are decr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are incr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are incr and low	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are low	0.1	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.1	0.0	0.1
Among all y, most are med	1.0	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0	1.0
Among all y, most are med and const	0.5	0.5	0.6	0.7	1.0	0.5	0.3	0.6	0.6	0.9	0.6
Among all y, most are med and const and mod	0.2	0.2	0.2	0.5	0.8	0.4	0.0	0.4	0.3	0.5	0.4
Among all y, most are med and mod	0.7	0.6	0.6	0.7	1.0	0.8	0.3	0.7	0.4	0.8	0.6
Among all y, most are mod	0.8	1.0	0.8	1.0	1.0	0.9	0.5	0.9	0.7	1.0	0.9
Among all y, most are short	0.5	0.7	0.5	0.5	0.2	0.4	0.7	0.6	0.6	0.2	0.5
Among all y, most are short and const	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are short and decr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are short and incr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are short and incr and low	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Among all y, most are short and low	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
Among all y, most are short and mod	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1

Table 2 presents the degree of truth T for the considered time series.

For example, the degree of truth for the summary *Among all trends, most are short* amounts to 0.2, 0.8, 0.9, 0.7, 0.8 and 0.3 for time series of Products No. 1 to 6, respectively.

Table 2: Evaluated linguistic summaries for sales time series of Prod 1- Prod 6

	Prod1	Prod2	Prod3	Prod4	Prod5	Prod6
Among all decr trends, most are medium	0.5	0.2	0.2	0.4	0.1	0.5
Among all decr trends, most are moderate	0.5	0.3	0.5	0.4	0.3	0.7
Among all trends, most are constant	0.2	0.4	0.1	0.1	0.1	0.2
Among all trends, most are decr	0.4	0.2	0.3	0.2	0.3	0.3
Among all trends, most are incr	0.4	0.2	0.3	0.2	0.3	0.3
Among all trends, most are low	0.2	0.3	0.4	0.2	0.3	0.3
Among all trends, most are medium	0.8	0.2	0.2	0.4	0.1	0.5
Among all trends, most are moderate	0.7	0.2	0.6	0.5	0.4	0.8
Among all trends, most are short	**0.2**	**0.8**	**0.9**	**0.7**	**0.8**	**0.3**

The next step of the proposed approach is the classification of the time series by probabilistic models based on the sets of the linguistic granules. Again, for the clarity reasons, we group the template probabilistic models into 3 classes as follows: *AR with weak positive autocorrelation* $C_1 = \{M_1, M_2, M_3\}$, *AR with medium positive autocorrelation* $C_2 = \{M_4, M_5, M_6, M_7\}$ and *AR with strong positive autocorrelation* $C_3 = \{M_8, M_9, M_{10}\}$, and these classes are processed in the remaining of this experiment. The classification scores based on the linguistic summaries are presented in Table 3. For example, for TS representing sales of Product 1, the classification scores assigned to models C_1, C_2 and C_3 are 0.02, 0.39 and 0.59, respectively. The weights for the Bayesian averaging are created as a consequence of these classification scores.

Table 3: The classification scores in the 3-class problem based on the sets of linguistic summaries for the considered time series

Scores	C1	C2	C3
Prod 1	**0.02**	**0.39**	**0.59**
Prod 2	0.86	0.10	0.04
Prod 3	0.92	0.06	0.02
Prod 4	0.58	0.23	0.19
Prod 5	0.94	0.04	0.02
Prod 6	0.08	0.46	0.45

5.3 Forecasting performance

To illustrate the performance of the forecasting approach, we use the subset of monthly time series that have length around 100 observations (94-100 ob-

servations) from the M3-Competition Datasets Repository [Makridakis and Hibon(2000)]. They are categorized as industry, macro economy, finance, demographic data or other.

The final step of the approach is the posterior simulation and the verification of forecasting accuracy. The summary of the forecasting accuracy is demonstrated in Table 4.

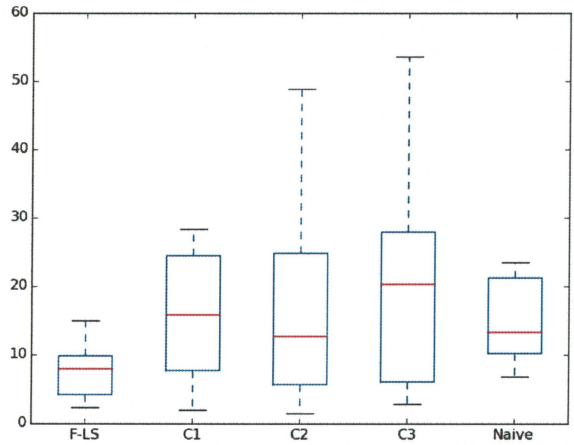


Figure 4: Box plot showing sMAPE forecasting accuracy on sales dataset by F-LS method, predefined 3 classes of autoregressive processes and naive method (last observed is first predicted).

As demonstrated in Table 4, the sMAPE forecasting error for the proposed F-LS method amounts to 7.8, and is around 5 times smaller than for the naive method. More importantly, for the considered time series, the approach also outperforms the traditional autoregressive models. The demonstrated forecasting accuracy results prove that the proposed method outperforms the naive method, however the error values seem to be significant.

It is easily observed that this approach does not include the dependencies between different time series, and as confirmed by experts of the fields, the sales time series seem to be correlated.

Next, the performance of the proposed approach assuming human input is compared to best 13 benchmark methods studied in [Makridakis and Hibon(2000)]. These methods are briefly presented in Table 5. Methods marked

Table 4: Symmetric Mean absolute percentage error (sMAPE) for 6- step-ahead forecast of the F-LS method, alternative autoregressive processes and naive method (last observed is first predicted)

	sMAPE				
	C1	**C2**	**C3**	**F-LS**	**Naive**
Prod 1	2.0	1.5	3.6	3.4	6.9
Prod 2	28.4	28.6	28.5	9.4	23.5
Prod 3	6.4	3.8	2.8	2.3	12.2
Prod 4	26.0	48.9	53.6	10.0	9.6
Prod 5	20.0	11.7	26.8	15.0	14.6
Prod 6	11.8	13.7	14.0	6.6	1 47.7
Total	**15.8**	**18.0**	**21.5**	**7.8**	**35.7**

with * are commercially available in forecasting packages.

For real-life monthly time series (100 observations), the proposed method has scored sMAPE result of 11.03.

Secondly, we test the proposed approach for the yearly time series.

As demonstrated, none of the benchmark methods outperforms the proposed F-LS method for all time series. On average, the proposed method scored a very good sMAPE result of 4.16, which is the best among the considered competitive methods. At the same time, it is observed that for some time series there exist methods that deliver more accurate forecast, so there is still potential for improvement. For example, when considering the time series N158, the sMAPE amounts to 3.41, 2.02, 3.04 for the proposed F-LS, ForecastX and the ForecastPRO method, respectively. The numerical results confirm that the proposed F-LS approach delivers very competitive results in terms of the forecasting accuracy.

We have identified why the difference is significant. In general, the Bayesian approach is most successful for short time series.

To sum up, the numerical results of the forecast accuracy show that the proposed approach of combining human input about linguistic summaries and various Box-Jenkins models through the Bayesian averaging may lead to the increase of the accuracy compared to the benchmark methods. None of the studied benchmark methods outperforms or dominates the proposed F-LS method for all time series. In almost half cases for the short time series, the proposed F-LS approach delivers more accurate forecast than the ForecastX method, which has scored the best results (on average) in the competition. F-LS provides forecasts which are similarly accurate to the ones provided by Comb S-H-D, Robust-Trend, Theta and RBF methods.

Table 5: Best methods from the M3-Competition. Methods marked with * are commercially available in forecasting packages. Source: Table 2, The 24 methods included in the M3-Competition classified into six categories in [Makridakis and Hibon(2000)]

Method	Name	Author	Description
F1	ForecastX*	J. Galt	Expert System - selects from among several methods: Exponential Smoothing/Box Jenkins /Poisson and negative binomial models/ Croston's Method/Simple Moving Average
F2	Theta [Assimakopoulos and Nikolopoulos(2000)]	V. Assimakopoulos	Decomposition technique - projection and combination of the individual components
F3	ARARMA	N. Meade	ARIMA models - Automated Parzen's methodology with Auto regressive filter
F4	ForecastPro*	R. Goodrich, E. Stellwagen	Expert System - selects from among several methods: Exponential Smoothing/Box Jenkins /Poisson and negative binomial models/ Croston's Method/Simple Moving Average
F5	Comb S-H-D	M. Hibon	Trend model - combining three methods: Single / Holt/ Dampen
F6	B-J Auto	M. Hibon	ARIMA models - Box-Jenkins methodology of 'Business Forecast System'
F7	Auto-ANN	K. Ord, S. Balkin	Automated Artificial Neural Networks
F8	RBF	M. Adya, S. Armstrong, F. Collopy, M. Kennedy	Rule-based forecasting: using random walk, linear regression and Holt's to estimate level and trend, involving corrections, simplification, automatic feature identification and re-calibration
F9	Robust-Trend	N. Meade	Trend model - Non-parametric version of Holt's linear model with median based estimate of trend
F10	Naive2	M. Hibon	Deseasonalized Naive (Random Walk)

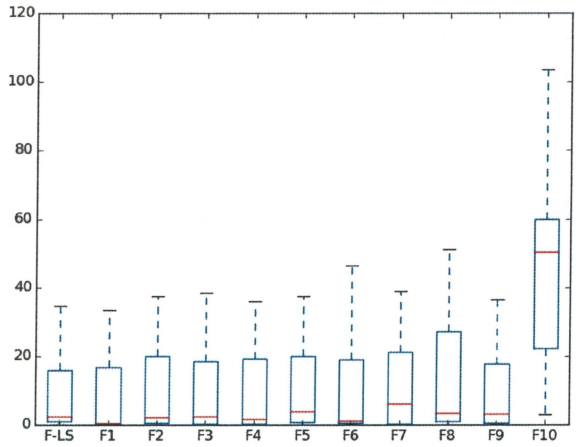


Figure 5: Box plot showing sMAPE forecasting accuracy on monthly time series (N 2011 - N 2781) dataset from M-3 Competition by F-LS method, predefined 3 classes of autoregressive processes and naive method (last observed is first predicted).

Table 6: sMAPE forecasting accuracy for monthly time series (N 2011 - N 2781) from M-3 Competition by: the proposed F-LS method and the competitive methods from the M3 Competition

	F-LS	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
N 2011	11.8	7.5	7.4	2.5	7.3	7.2	7.4	7.2	5.8	3.3	19.9
N 2215	16.2	33.6	37.4	38.5	33.8	37.4	33.5	27.0	35.9	36.4	27.0
N 2216	1.1	0.2	0.5	0.2	0.0	0.4	0.7	0.2	0.8	0.3	51.4
N 2217	2.4	0.6	0.4	0.5	0.7	0.6	0.6	0.1	0.5	0.9	3.0
N 2218	0.9	0.0	0.0	0.1	0.1	0.0	0.1	0.0	3.1	0.6	6.3
N 2219	0.1	0.2	0.5	0.2	0.2	0.7	0.4	0.3	2.0	0.5	22.2
N 2220	1.0	0.3	0.3	0.2	0.3	0.1	0.4	0.3	1.2	0.4	54.3
N 2640	1.7	0.6	0.5	0.4	0.9	0.7	0.6	0.5	0.1	0.6	60.1
N 2647	2.0	0.4	0.5	0.2	0.5	0.8	0.2	0.3	0.9	0.9	59.0
N 2651	21.2	16.7	20.6	18.4	21.2	20.3	19.0	21.2	51.4	13.3	75.3
N 2753	1.6	0.3	0.6	0.3	0.3	0.4	0.3	0.3	0.2	0.3	135.5
N 2772	16.0	0.8	3.0	3.3	1.7	4.0	1.1	6.2	3.6	5.4	103.4
N 2778	53.7	71.4	68.6	63.4	65.4	69.2	63.4	63.5	81.9	95.1	8.6
N 2779	34.7	32.2	32.0	36.1	36.1	33.1	46.5	39.0	27.2	12.8	43.1
N 2781	0.9	0.3	2.3	16.8	13.3	7.2	8.4	27.6	31.2	96.8	89.5
Avg	**11.0**	**11.0**	**11.7**	**12.1**	**12.1**	**12.1**	**12.2**	**12.9**	**16.4**	**17.9**	**50.6**
Std dev	**15.6**	**20.3**	**20.0**	**19.4**	**19.3**	**20.0**	**19.9**	**19.0**	**24.5**	**33.1**	**38.4**

Table 7: sMAPE forecasting accuracy for yearly time series (N 156 - N168) from M-3 Competition by: the proposed F-LS method and the competitive methods from the M3 Competition

	F-LS	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
N 156	0.6	3.9	0.0	4.2	4.4	1.1	3.3	1.8	3.9	3.1	4.5
N 157	3.1	4.6	1.2	5.3	5.6	2.8	3.8	3.5	5.1	0.2	1.4
N 158	1.3	0.1	2.5	1.0	0.7	1.7	1.5	1.4	0.7	2.8	4.0
N 159	5.8	34.7	13.0	9.4	11.3	12.2	16.3	10.2	33.1	14.9	16.3
N 160	1.3	4.4	2.7	5.9	6.0	5.2	5.9	1.0	3.9	4.9	5.9
N 164	4.6	3.4	4.4	2.5	5.0	4.3	5.0	2.8	4.6	3.3	5.0
N 165	4.4	3.2	4.8	2.2	4.6	4.3	5.1	3.6	5.4	3.4	5.1
N 166	12.2	21.5	1.7	0.6	12.2	3.2	14.1	14.6	6.1	0.6	12.2
N 167	2.1	4.1	5.0	3.7	4.1	5.1	5.6	0.3	4.5	3.2	4.1
N 168	1.3	3.3	0.5	8.6	6.0	1.2	7.5	1.3	5.2	2.0	3.3
Avg	**3.67**	**8.32**	**3.59**	**4.33**	**5.99**	**4.10**	**6.79**	**4.06**	**7.23**	**3.85**	**6.18**
Std dev	**3.46**	**10.96**	**3.75**	**2.98**	**3.38**	**3.21**	**4.73**	**4.65**	**9.20**	**4.11**	**4.51**

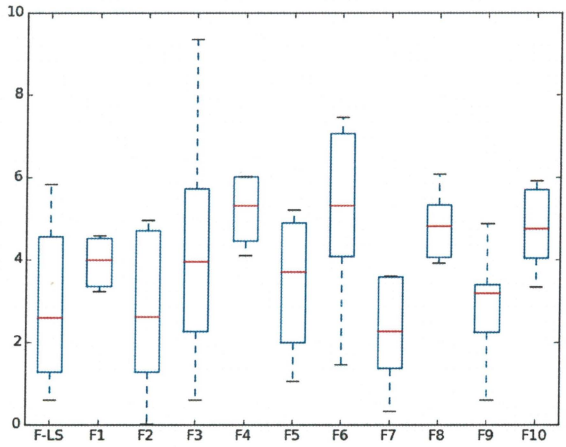


Figure 6: Box plot showing sMAPE forecasting accuracy on yearly time series (N 156 - N168) dataset from M-3 Competition by F-LS method, predefined 3 classes of autoregressive processes and naive method (last observed is first predicted).

6 Conclusion and Future Work

In this paper we presented a novel approach of incorporating linguistic summaries into the construction of the prior information for forecasting models. Instead of defining probability distributions, the user validates expected trends and the system creates the probability distributions automatically. The main advantages of the proposed solution are its human-consistency and accuracy.

Recalling the 'No Free Lunch' theorem of Wolpert, there is no single method or model being the best solution for any problem, but some methods perform well in specific situations. The proposed approach has been evaluated with experiments on real-life time series from the pharmaceutical market and the M3 Competition benchmark datasets. The results confirm that the incorporation and processing of the linguistic summaries increases the interpretability of the forecasting process and may improve its accuracy. As demonstrated in experiments, the proposed approach may outperform the benchmark methods. Furthermore, the considered imprecise labels and attributes enable to describe extremely well the autoregressive time series. The simulation study reveals the correlations between the probabilistic models and linguistic summaries which seems a very appealing idea within the interdisciplinary research on statistics and data mining. At the same time, the accuracy for the considered real-life dataset is still quite average among other methods. Nonetheless, the proposed approach is very plausible due to the human-centricity of the processed features.

We identify the following major challenges for future research related to this topic:

- selection of attributes, labels and types of linguistic summaries;
- improving the computational complexity of the summarization;
- search for other linguistic forms of summarization results valuable as features for the construction of prior probability distributions;

Future research assumes also next experimental studies to verify in which contexts linguistic summaries are outstanding to represent the real-life time series data.

References

[Wolpert(1996)] Wolpert D. The lack of a priori distinctions between learning algorithms. Neural Computation 1996;:1341–90.

[Gooijer and Hyndman(2006)] Gooijer JD, Hyndman RJ. 25 years of time series forecasting. International Journal of Forecasting 2006;22:443–73.

[Box and Tiao(1973)] Box G, Tiao G. Bayesian Inference in Statistical Analysis. Wiley; 1973.

[Geweke(2005)] Geweke J. Contemporary bayesian econometrics and statistics. Wiley series in probability and statistics 2005;.

[Petridis et al.(2001)Petridis, Kehagias, Petrou, Bakirtzis, Kiartzis, Panagiotou et al.] Petridis V, Kehagias A, Petrou L, Bakirtzis A, Kiartzis S, Panagiotou H, et al. A bayesian multiple models combination method for time series prediction. J Intell Robotics Syst 2001;31(1-3):69–89.

[Nauck and Kruse(2014)] Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine 2014;16(2):149–69.

[Kempe et al.(2008)Kempe, Hipp, Lanquillon and Kruse] Kempe S, Hipp J, Lanquillon C, Kruse R. Mining frequent temporal patterns in interval sequences. Fuzziness and Knowledge-Based Systems in International Journal of Uncertainty 2008;16 (5):645–61.

[Moewes and Kruse(2009)] Moewes C, Kruse R. Zuordnen von linguistischen ausdrücken zu motiven in zeitreihen (matching of labeled terms to time series motifs). Automatisierungstechnik 2009;:146–54.

[Hryniewicz and Kaczmarek(2014)] Hryniewicz O, Kaczmarek K. Forecasting short time series with the bayesian autoregression and the soft computing prior information. In: Strengthening Links Between Data Analysis and Soft Computing; vol. 315. Springer; 2014, p. 79–86.

[Kaczmarek et al.(2015)Kaczmarek, Hryniewicz and Kruse] Kaczmarek K, Hryniewicz O, Kruse R. Human input about linguistic summaries in time series forecasting. In: Proc. of The Eighth International Conference on Advances in Computer-Human Interactions ACHI 2015. 2015, p. 9–13.

[Hryniewicz and Kaczmarek(2015)] Hryniewicz O, Kaczmarek K. Bayesian analysis of time series using granular computing approach. Applied Soft Computing 2015;.

[Attneave(1954)] Attneave F. Some informational aspects of visual inspection. Psychological Review 1954;61 (2).

[Wilbik and Keller(2012)] Wilbik A, Keller J. A distance metric for a space of linguistic summaries. Fuzzy Sets and Systems 2012;208:79–94.

[Batyrshin and Sheremetov(2008)] Batyrshin IZ, Sheremetov LB. Perception-based approach to time series data mining. Applied Soft Computing 2008;8(3):1211–21.

[Yager(1982)] Yager R. A new approach to the summarization of data. Information Science 1982;28 (1):69–86.

[Kacprzyk(2008)] Kacprzyk J. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets Syst 2008;159 (12):1485–99.

[Kacprzyk and Wilbik(2009)] Kacprzyk J, Wilbik A. Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In: IFSA/EUSFLAT Conf. 2009. 2009, p. 1321 –6.

[Kacprzyk et al.(2011)Kacprzyk, Wilbik, Partyka and Ziółkowski] Kacprzyk J, Wilbik A, Partyka A, Ziółkowski A. Trend Analysis System. Systems Research Institute, Polish Academy of Sciences, Warsaw; 2011.

[Dubois and Prade(1992)] Dubois D, Prade H. Gradual rules in approximate reasoning. Information Sciences 1992;:103–22.

[Kacprzyk and Zadrożny(2005)] Kacprzyk J, Zadrożny S. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. Information Sciences 2005;173:281–304.

[Wilbik(2010)] Wilbik A. Linguistic summaries of time series using fuzzy sets and their application for performance analysis of mutual funds. Ph.D. thesis; Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland; 2010.

[Kass and Raftery(1995)] Kass R, Raftery A. Bayes factors. Journal of the American Statistical Association 1995;90:773–95.

[Ley and Steel(2009)] Ley E, Steel M. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. Journal of Applied Econometrics 2009;24:651–74.

[Makridakis and Hibon(2000)] Makridakis S, Hibon M. The m3-competition: results, conclusions and implications. International Journal of Forecasting 2000;:451–76.

[Kaczmarek and Hryniewicz(2013)] Kaczmarek K, Hryniewicz O. Linguistic knowledge about temporal data in bayesian linear regression model to support forecasting of time series. In: Proc. of Federated Conference on Computer Science and Information Systems. 2013, p. 655 –8.

[Assimakopoulos and Nikolopoulos(2000)] Assimakopoulos V, Nikolopoulos K. The theta model: a decomposition approach to forecasting. International Journal of Forecasting 2000;16:521–530.