International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report **IR-07-038**

Weather Indicators and Crop Yields Analysis with Wavelets

Bartosz Kozłowski (b.kozlowski@elka.pw.edu.pl)

Approved by

Marek Makowski (marek@iiasa.ac.at)
Senior Research Scholar, Risk, Modeling and Society Program

September 2005

Foreword

This report describes the research which the author advanced during his participation in the 2005 Young Scientists Summer Program (YSSP).

The reported research deals with application of wavelet analysis to the analysis of complex data thus contributes to the IME research agenda, which includes development of methodologies and software tools for analysis and support of complex decision problems. The reported work is interdisciplinary in nature. It not only required combining thorough comprehension of mathematical modeling, optimization techniques, and relational databases with the understanding of complex analysis of agriculture and climatic data in the context of modeling crop yields. It also required cooperation with colleagues of the IIASA Land Use Change and Agriculture (LUC) Program. Moreover, the author was a member of a three-person YSSP team working on researching different approaches to the same problem[1]. Such a team work was very fruitful. We mention here only the main aspects of this successful collaboration:

- different approaches are complementary, i.e., they provide various insights to the same problem,
- team-members having different skills and experiences have learned from each other,
- time-consuming jobs of data collection, verification, and processing (by organizing a dedicated database and the supporting software) were shared among team members.

The Summer-time of the YSSP is only three months short, and this type of research requires substantial amount of time for the initial stage. This initial stage includes a correct specification of the problem, and then data collection, pruning, verification, and processing. The problem is challenging, therefore no standard ready-to-use methods were available. The author therefore had to first advance his research on wavelet analysis, develop or adapt software to support the new approaches to data analysis, and then to verify various approaches with the data before discussing the results with the colleagues from the LUC Program. The constructive criticism by the LUC colleagues had led to the development of better methods and the corresponding software tools.

Marek Makowski

[1] The results of another approach are described by Wojciech Kotłowski in IR-07-034.

Abstract

This paper presents a case study and methods for trend and periodicity identification and for forecasting which are useful for determining the impact of weather factors to crop yields. These methods combine standard approaches with wavelet analysis in order to improve the results. Hence, a brief introduction to wavelet theory is included. All analyses provided in this research are conducted on United States weather and crop yields data. Where applicable, the results obtained with the use of standard methods are compared with those which are enforced by application of the wavelet transform, illustrated with examples, and evaluated.

Keywords: trend analysis, seasonality identification, forecasting, impact, crop yields, weather factors, wavelet transform, Haar basic wavelet function

Acknowledgments

This paper reports the results of the research done during the author's participation in the 3 months Young Scientists Summer Program (YSSP) 2005, held at the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria.

The author would like to thank Dr. Marek Makowski for his supervision during the YSSP 2005 and to Prof. Yurii Ermoliev for the insight into the methodological background. Many thanks go also to the Land Use program, especially to Dr. Günther Fischer, Dr. Tatiana Ermolieva, and Dr. Harrij van Velthuizen, for sharing expert knowledge from the domain of the research. The current shape of the report is also a result of numerous discussions with colleagues from both RMS and LUC programs what is happily acknowledged.

Finally, the author would like to thank the Polish National Member Organization of IIASA for the financial support which made his participation in the YSSP possible.

About the Author

Bartosz Kozlowski received M.Sc. in computer science from the Warsaw University of Technology (WUT), Poland, in 2004. He is currently a Ph.D. student at WUT in the Institute of Control and Computation Engineering. In his doctoral thesis he is dealing with new methods for multicriteria model analysis characterized by a large number of structured criteria. His fields of scientific interest include decision support, forecasting, wavelet analysis, operational research, data mining, and software development methodologies. He has been involved in projects and live implementations concerning computer networks, security, scientific and business applications development. Bartosz has also seven years of experience in teaching in WUT and Polish-Japanese Instutute of Information Technology and in running IT courses for business sector.

Bartosz first came to IIASA in 2005 as a participant of the Young Scientists Summer Program (YSSP). Since 2005 he is IIASA collaborator for the RMS Program and then IME Project participating in the development of the Structured Modeling Technology, multicriteria analysis methodologies and joining other research activities conducted by IME.

Contents

Notation

Alg.	algorithm
Fig.	figure
Tab.	table
p.	page
App.	appendix
Sec.	section
$\mathbf{X}$	time series
$\mathbf{F}$	short-term component of the time series
$\mathbf{G}$	long-term component of the time series
$\mathbf{T}$	ordered set of locations of the time series
T	length of the time series
t	location within the time series
x_t	value of the time series in location t
i_t	value of the impact coefficient in location t
$\overline{x}$	average value of the time series
$\widehat{x_t}$	estimated value of the time series in location t
r_t	smoothness coefficient defined in location t
R_1	value of first kind smoothness measure
R_{WBK}	value of wavelet coefficients series smoothness measure
ϵ	a small positive value
D2DWT	dense discrete wavelet transform
MODWT	maximal overlap discrete wavelet transform
DWT	discrete wavelet transform
WT	wavelet transform
iMODWT	inverse maximal overlap discrete wavelet transform
iDWT	inverse discrete wavelet transform
iWT	inverse wavelet transform
λ	scale, length of each neighboring interval used to calculate a wavelet coefficient
j	wavelet level
J	maximal wavelet level for a time series of a given length
$\mathbf{W}$	a vector of WT coefficients
$\mathbf{V}_j$	a sub vector of scaling coefficients calculated on the level j
$\mathbf{W}_j$	a sub vector of wavelet coefficients calculated on the level j
$V_{j,i}$	an i^{th} scaling coefficient calculated on the level j
$W_{j,i}$	an i^{th} wavelet coefficient calculated on the level j

I_j	a number of wavelet coefficients on the level j
$\overline{\mathbf{W}}$	a vector of DWT coefficients
$\overline{\mathbf{V}}_j$	a sub vector of DWT scaling coefficients calculated on the level j
$\overline{\mathbf{W}}_j$	a sub vector of DWT wavelet coefficients calculated on the level j
$\overline{V}_{j,i}$	an i^{th} DWT scaling coefficient calculated on the level j
$\overline{W}_{j,i}$	an i^{th} DWT wavelet coefficient calculated on the level j
$\widetilde{\mathbf{W}}$	a vector of MODWT coefficients
$\widetilde{\mathbf{V}}_j$	a sub vector of MODWT scaling coefficients calculated on the level j
$\widetilde{\mathbf{W}}_j$	a sub vector of MODWT wavelet coefficients calculated on the level j
$\widetilde{V}_{j,i}$	an i^{th} MODWT scaling coefficient calculated on the level j
$\widetilde{W}_{j,i}$	an i^{th} MODWT wavelet coefficient calculated on the level j
$\breve{\mathbf{W}}$	a vector of D2DWT coefficients
$\breve{\mathbf{V}}_j$	a sub vector of D2DWT scaling coefficients calculated on the level j
$\breve{\mathbf{W}}_j$	a sub vector of D2DWT wavelet coefficients calculated on the level j
$\breve{V}_{j,i}$	an i^{th} D2DWT scaling coefficient calculated on the level j
$\breve{W}_{j,i}$	an i^{th} D2DWT wavelet coefficient calculated on the level j

δ	correction coefficient
κ	threshold value
n	standard deviation multiplication parameter

SSE	sum of square errors
mSSE	SSE divided by multiplication of an average of original time series by the number of locations of this time series
wSSE	sum of weighted (by the value at the location of the error) square errors
MSE	mean square error
r^2	correlation coefficient
df	number of degrees of freedom
SST	total sum of squares
SSR	sum of squares for regression

LR	linear regression
PA	polynomial approximation

avg	average
max	maximum
min	minimum

Weather Indicators and Crop Yields Analysis with Wavelets

Bartosz Kozłowski[*]

1 Introduction

The goal of the reported work was to explore the opportunities provided by wavelet analysis (WA) for supporting analysis of potential impact of weather factors to crop yields and reusing established information to e.g. planning. The three main analysis problems are pertinent to this research:

- trend identification,
- periodicity identification,
- forecasting.

All of these problems were approached with respect to variously profiled time series data of weather factors and crop yields. Such analysis can be done with standard statistical methods, like e.g.:

- autocorrelation,
- auto regression,
- trend extrapolation,
- exponential smoothing,
- curve fitting (regression),
- the Box-Jenkins approach.

On the other hand, as a result of previous research, applications of wavelet analysis occurred to be very useful in many applications. They were applied to economical and finance data and it occurred that WA based approach provides better results than some other approaches to forecasting, see e.g. (Kozlowski 2004). Wavelet domain also reveals additional information about time series what was applied in the telecommunications area (Kozlowski 2005). Hence it was decided to explore applicability of wavelet-based methods in combination with standard approaches to the above defined analysis.

This paper is organized as follows. The next section provides a brief overview of the approaches and justifies their application. It also presents a general discussion of the reasons for using WA and describes the general wavelet-based approach. In the last part of this section case study is brought forward including short information on data issues. Section 3 focuses on the method for the trend identification and its applications. In section 4 applications of a wavelet enforced method for periodicity identification in time series is introduced and investigated. Section 5 describes the wavelet approach to time series forecasting. Section 6 concludes the paper. For convenience and self-containment of the report App. B and C contain excerpts form authors previous work and describe basics of WA and fast polynomial approximation, respectively.

[*]Institute of Control and Computation Engineering, Warsaw University of Technology,
Nowowiejska st 15/19, 00-665 Warsaw, Poland Phone: +48 (22) 660 7124, Fax: +48 (22) 8253719

2 Overview

2.1 Approaches

First of the approaches discussed in this paper is trend identification. As it was assumed that the weather factors influence only a variability of the total characteristics of the crop yields time series, the natural step was to distinguish these variabilities. The "'extraction"' is achieved by estimating a trend contained in the time series. Then the trend is subtracted from the original time series data. After this operation further analysis is performed only on variabilities in the original data (according to the assumption that these are caused by weather factors). This approach eliminated the need to analyze influences by long-term factors of crop variability like, e.g., fertilization or mechanization.

The second problem studied is periodicity identification. Since adding additional information to the analysis process usually implies better results, it is natural to seek for this kind of information. In the case of time series a natural information, which one may find, is seasonality occurring throughout the time series.

Last of the problems approached in this paper is time series forecasting. The idea of forecasting is to assume what will be the state of a phenomenon in the future based on the knowledge of its "behavior" in the past. If we possess good knowledge of this phenomenon then we may apply various model-based or rule-based forecasting approaches. Otherwise, the most commonly used methodologies are based on statistics, see e.g. (Aczel 1989). To all these problems the wavelet transform (WT) based methodology was applied successfully (mostly to improve the accuracy of the results).

The WT provides a representation of a time series in a different domain, e.g. the time series of mean monthly temperature of Junes in Baldwin county in Alabama shown in Fig. 1 is transformed into a set of coefficients presented in Fig. 2 by Discrete Wavelet Transform (DWT).

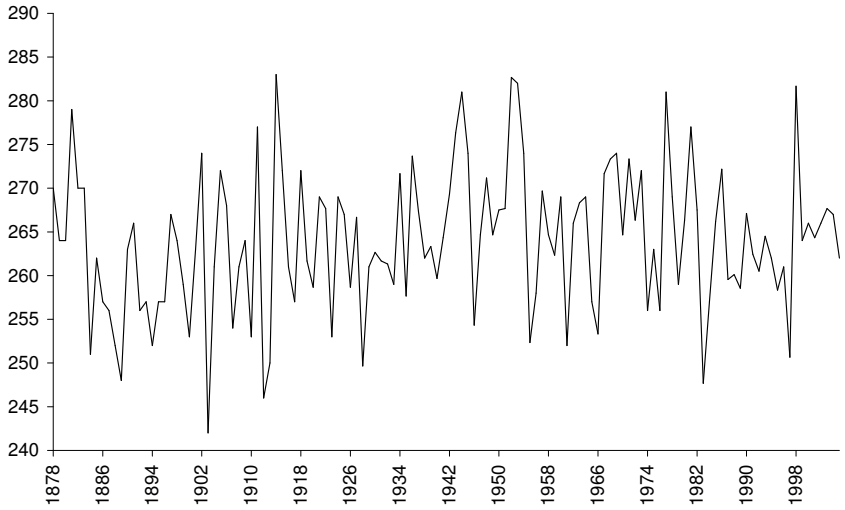

Figure 1: Original time series of mean monthly temperature of Junes in Baldwin county in Alabama.

Wavelet domain may be considered as a set of a series of observation points (wavelet coefficients) derived from the original time series based on a special single (but scaled and modified) function. In Fig. 2 the Haar basic wavelet function is used. Each point in each of these series may be interpreted as a difference of weighted (in the Haar based

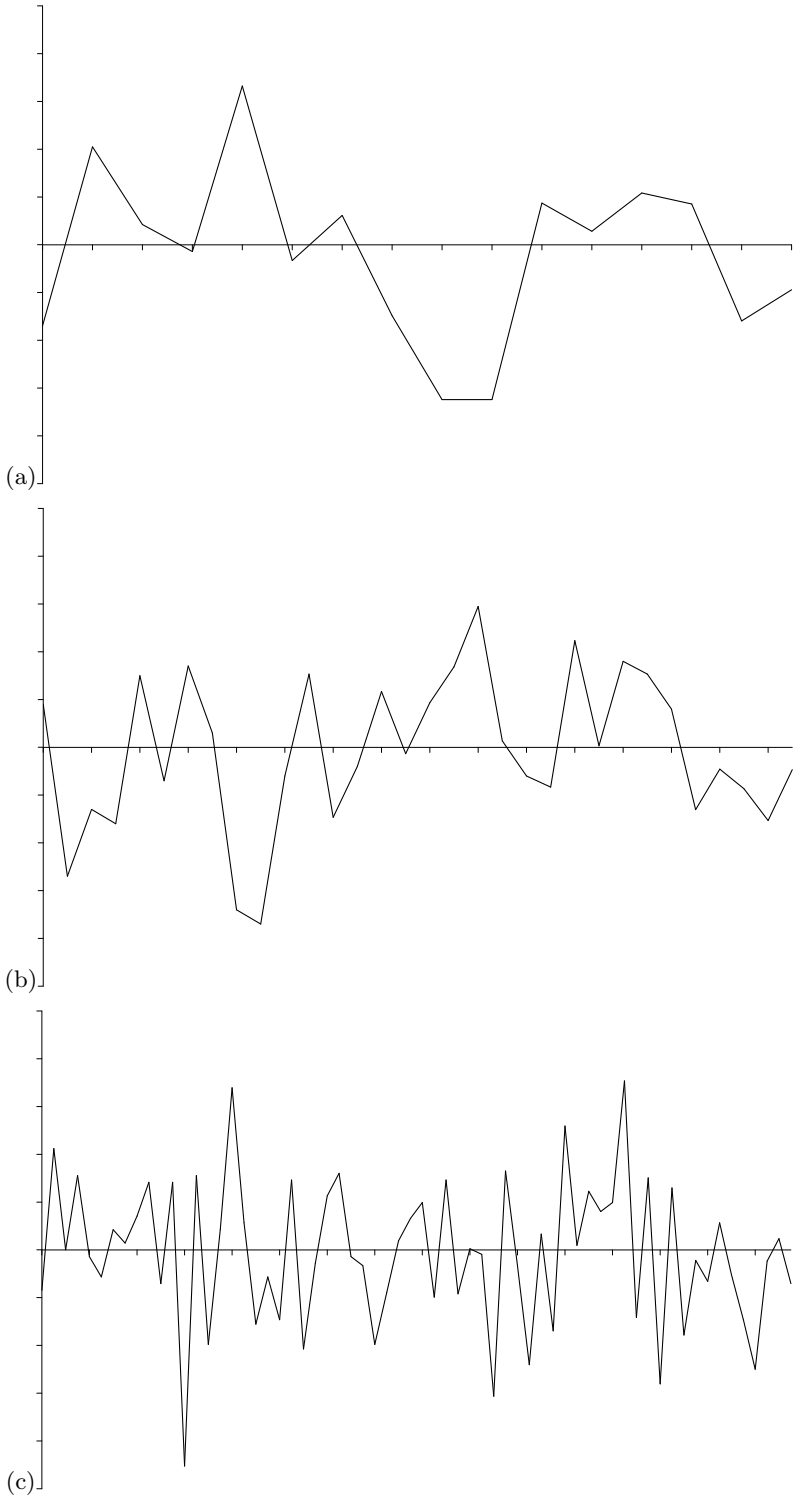

Figure 2: Result of sample wavelet transform of the original time series presented in Fig. 1. Sub figures (c), (b), and (a) present values (vertical axes) of first three wavelet coefficients sets for corresponding time points (horizontal axes).

example weights are equal to 1) averages in neighboring intervals. In Fig. 2c each point (wavelet coefficient) is a difference between two neighboring points. Fig. 2b shows wavelet coefficients where each is a difference between two averages of two points "on each side". In Fig. 2a wavelet coefficients are equal to differences of averages of four points so 8 points are considered to calculate each wavelet coefficient. A very important (although not always used) feature of this transform is inversability, which means that the original data may be reproduced by using wavelet coefficients. As the WT splits time series into short and long term trends and allows to perform a multiresolutional insight into the time series, it is natural to assume that the results enforced with wavelet analysis may outperform the standard analysis approaches. All of the wavelet-based methodologies used here utilize an analysis schema presented in Fig. 3.

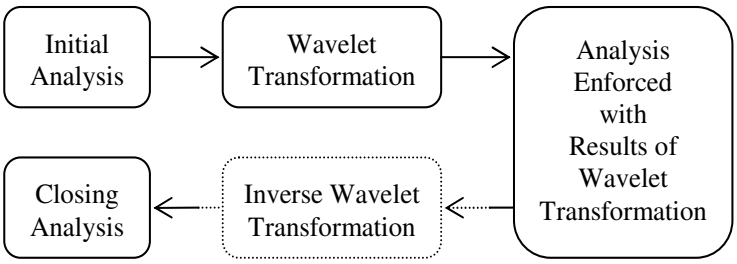

Figure 3: Schema of wavelet-based analysis flow.

After initial analysis (e.g. selecting appropriate length of the time series for wavelet transform) of the time series it is transformed with wavelet transform. Next, an analysis is performed on the resulting wavelet coefficients. Then, optionally, wavelet coefficients are inversely transformed into the original time series space and final analysis is performed (e.g. combining results of two inverse transform procedures). A bit more detailed introduction to wavelet enforced analysis processes may be found in App. B.

2.2 Data

All experimental results presented in this paper are derived from freely available data sets gathered for the territory of the United States of America. Weather data includes weather station level monthly observation series of four following factors:

- precipitation,
- minimum temperature,
- maximum temperature,
- average temperature.

Crop data is available as a set of yearly series of yields (for this research two crops were used, namely corn and wheat). Thus, we had to deal with a large data heterogeneity, both spatial and temporal.

One of the activities performed during the data preprocessing phase was upscaling. Weather station data was upscaled to the county level with simple rules using an administrative map associating weather stations with counties. In most cases there was a one-to-one correspondence between a weather station and a county. In cases where there was more than one weather station in the county an average of measurements of particular weather factors was calculated and assumed to be representative for the county. Analysis was not performed for counties in which there was no weather station data available.

Another preprocessing issue was data quality. Although the quality of data is generally acceptable, there is a certain amount of missing and suspicious values throughout all data

sets. In the reported research only the data, which contained full set of values for a certain (relatively long) time interval was considered. Also a lot of discussions and time was dedicated to analyzing China weather and crop yields data. Unfortunately, it turned out that the data was too imprecise and its amount was not sufficient. Therefore it was decided not to attempt analysis of weather impact on crop yields in China[2].

In case of a trend identification the results established by the wavelet-based methodologies are compared with the results obtained on the basis of the well known, statistical or analytical analyses. Some measures used for these evaluations are presented in Sec. 3.2. Also time series forecasting problem has been similarly approached, that is the results of applications of WA-based methodology have been evaluated with a measure by comparison with other non-WA approaches.

3 Trend Identification

3.1 The Purpose

Changes in time series of crops may be divided into two groups. The first one will be referred to as a trend and the second one as a variability. Factors which influence the first one include, among others, quality and type of soil, fertilization, mechanization, etc. A variability part observed "around" the trend is supposed to be mostly caused by weather factors. As in this research the main consideration are weather factors, there is a need for analyzing only a variability of time series. In time series of crop yields one may observe certain general trends. Hence, a couple of trend identification methods are presented, applied, and evaluated.

A desired trend is characterized by the following features:

- it should fit relatively well the corresponding data,
- it should be smooth enough to distinguish it from the original time series, and to not represent fast changes,
- the impact coefficient (see 3.2.3) defined on top of this trend should have a distribution with a possibly small number of values far from 0.

Considering these desired characteristics it was decided to use three criteria groups. These are fitness and smoothness measures of a trend and a distribution of an impact coefficient. Therefore descriptions of those measures are provided in following subsections.

3.2 Trend Quality Measures

3.2.1 Fitness Measures

Fitness is one of the oldest measures, and a very important factor in trend analysis. Various coefficients (e.g., MSE and r^2) aim to denote how good is the fit. A list of measures along with their short definitions used in this research, for evaluating accuracy of newly derived time series, which target fitting the original time series, is provided below. These are some of the standard coefficients used in various statistical approaches (Aczel 1989).

$$\mathrm{SSE} = \sum_{t=0}^{T-1} (x_t - \widehat{x}_t)^2$$

$$\mathrm{mSSE} = \frac{\mathrm{SSE}}{T \cdot \overline{x}}$$

[2] A more detailed description of available China data can be found in (Kotlowski 2007).

$$\text{wSSE} = \sum_{t=0}^{T-1} \frac{1}{x_t}(x_t - \widehat{x_t})^2$$

All of these measures are applied (apart from the whole data set) to 95%, 90%, and 75% of the best fitting estimated values.

3.2.2 Smoothness Measures

It is very useful to measure smoothness of a trend of a given time series. Generally, the more smooth is the trend, the better it is, because the chance that it includes fast, local changes is lower. Also one of the key elements of the algorithm described in Sec. 4 is to determine how smooth each wavelet coefficients' series is. Herein a measure applied to evaluate the smoothness of the derived trends is proposed.

Let us once again consider a time series $\mathbf{X}$ of T observation points defined in locations $t = 0...T-1$. Now let us define a coefficient r_t as:

$$r_t = \left|\left| x_t - \frac{x_{t-1} + x_{t+1}}{2} \right| - \left| x_{t-1} - \frac{x_{t-2} + x_t}{2} \right|\right|, \; t = 2...(T-2).$$

Consider Fig. 4. In case of a line (Fig. 4a) (which is the smoothest function, of course) values of r_t coefficients are equal to 0. In case of an arc (Fig. 4b) r_t coefficients are also equal to 0 as we assume that an arc function is smooth. For functions not containing line and arc based shapes the values of these coefficients differ from 0.

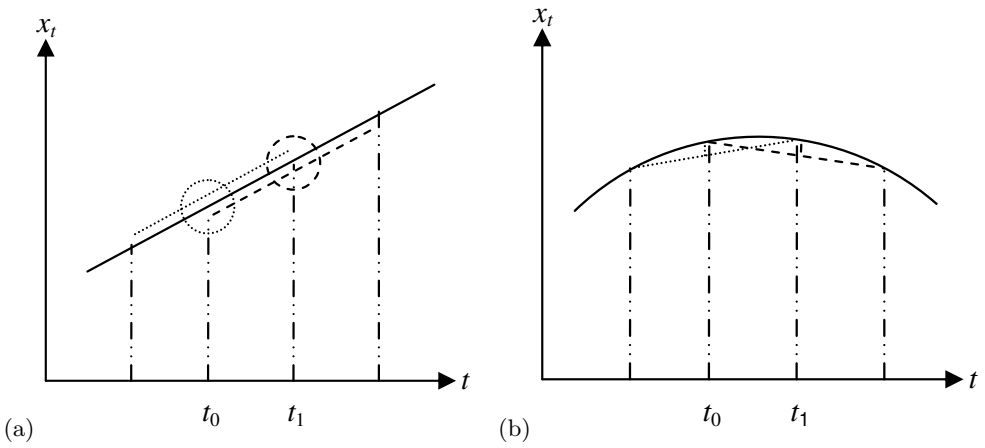

Figure 4: Smoothness measure applied to (a) line and (b) arc.

Having the measure R_1 defined as:

$$R_1 = \text{avg}_t\ (r_t),\ t = 2...(T-2),$$

and assuming that the smaller R_1 gets, the smoother a time series is, both the line and the arc are smooth with respect to R_1,.

For analyzing the smoothness of the wavelet coefficients another measure may be derived:

$$R_{WBK} = \frac{\max_t\ (r_t) - \min_t\ (r_t)}{\sqrt{2^j}},\ t = 2...(T-2),$$

and applied respectively on each wavelet level. In this case a spread (divided by the power of two to the level j) of r_t coefficients calculated from wavelet coefficients within each level j is analyzed.

Both of these measures may be interpreted as an analysis similar to the analysis of second derivatives.

3.2.3 Impact Coefficient

For a measure of impact of weather to crop yields an impact coefficient i_t

$$i_t = \frac{x_t - \widehat{x_t}}{\widehat{x_t}}, \; t = 1...T,$$

is defined. Values of this coefficient may be divided into three groups, for which:

- $i_t = 0$, indicates no impact of weather to crop yields x_t,
- $i_t > 0$, indicates a relative increase of yield due to impact of weather factors,
- $i_t < 0$, indicates a relative decrease of yield due to impact of weather factors.

This measure can also be used for evaluating trend quality: the smaller is the number of values of i_t far from 0 the better trend is (as it explains the phenomenon better).

3.3 Standard Trend Identification Methods

In this subsection an example of trend identification is discussed on crop yields data. Fig. 5 shows an example of the original time series of corn in Alexander county in Illinois.

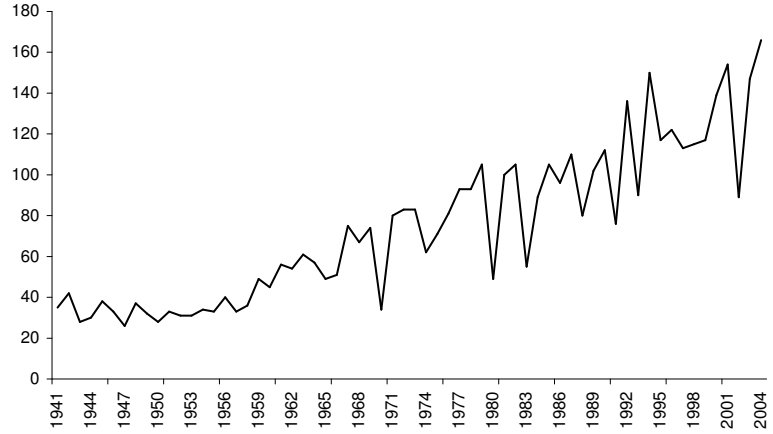

Figure 5: Original time series of corn yields in county Alexander, IL

A linear regression (LR) [3] performed on this data is shown in Fig. 6.
Fig. 7 the results of a polynomial approximation (PA) applied to the same data set are presented.

Other widely used trend identification methods include moving average, moving median, etc.

3.4 Wavelet Example

For fitting a time series, a direct gain is obtained by taking advantage of WT's multiresolutional analysis. By analyzing wavelet coefficients the analysis is performed on different levels of specific aggregation. The results of applications of LR and PA to wavelet coefficients instead of to the original time series are illustrated in Fig. 8 and 9.

[3] Exhaustive description of linear regression and other statistical methods may be found in (Aczel 1989).

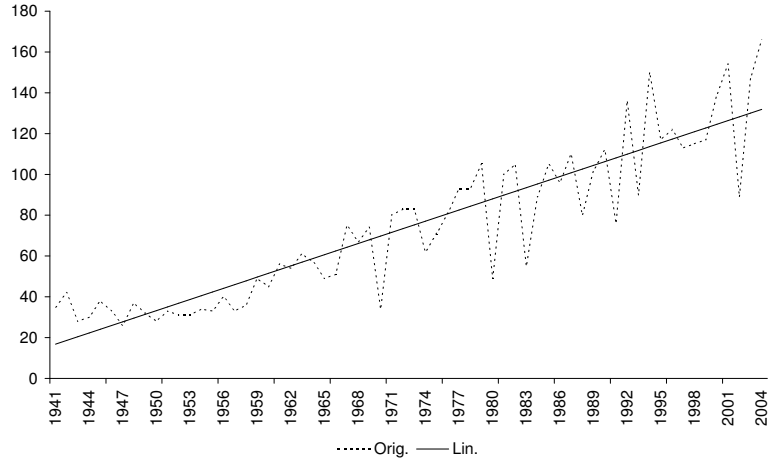

Figure 6: Result of application of LR to time series presented in Fig. 5.

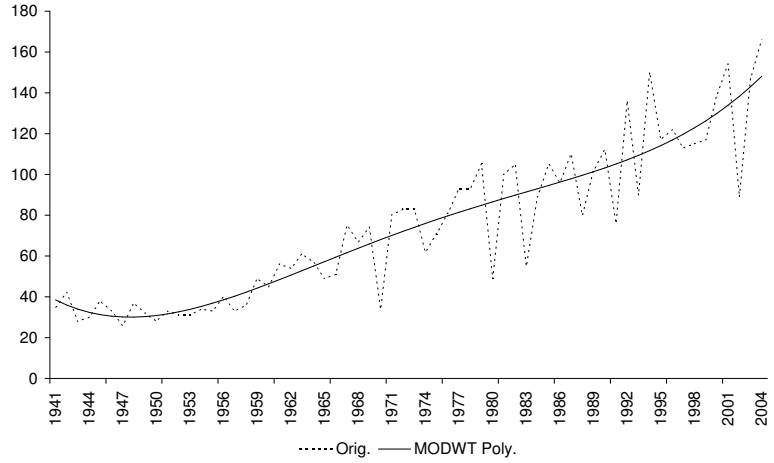

Figure 7: Result of application of PA to time series presented in Fig. 5.

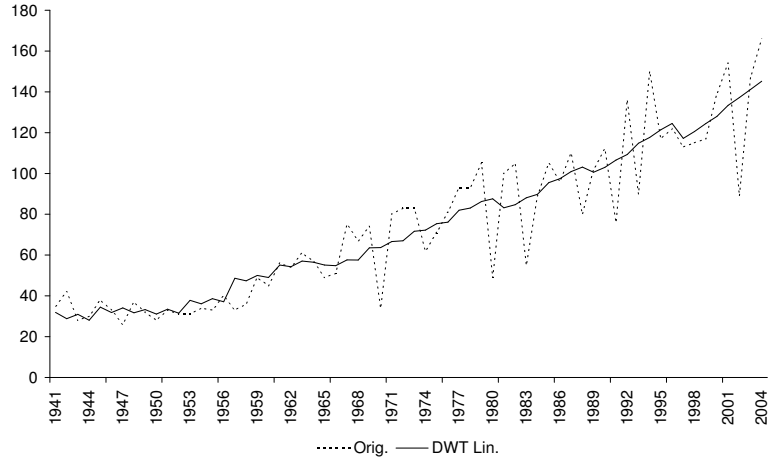

Figure 8: Result of application of DWT enforced LR to time series presented in Fig. 5.

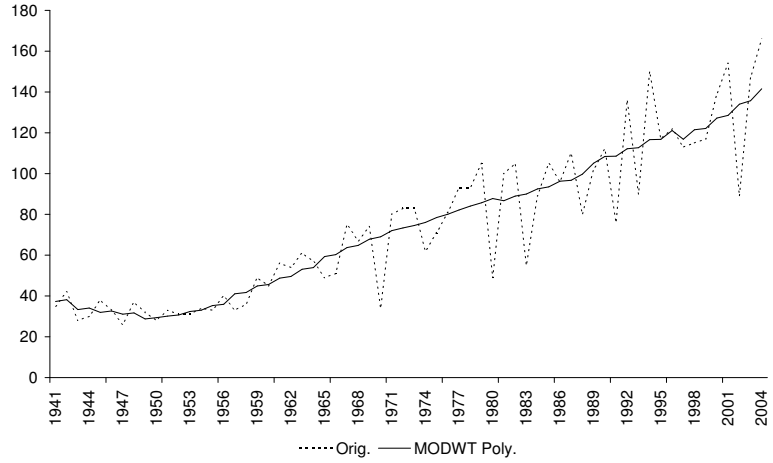

Figure 9: Result of application of MODWT enforced PA to time series presented in Fig. 5.

3.5 Evaluations

Tables 1a and 1b summarize the results of evaluations of fitness of trends for the corn and wheat data, respectively.
As it comes to the averages of fitness measures MODWT PA method provides the best results.

Tables 2a and 2b show results of evaluations of the smoothness of the trends for the data.
In case of smoothness the best-worse order of the methods follows: LR, PA, MODWT enforced PA, DWT enforced LR.

Let us now look at the distributions of the impact coefficient. Plots of this factor for different trends for corn and wheat are presented in Figures 10 and 11, respectively.
In most cases the distribution of the value of the i_t coefficient is around zero with a little shift toward positive values. The number of values close to -1 or 1 is very small. The PA method has more values concentrated around zero than in the case of linear regression. Also MODWT based approach is better than the DWT based one. What is most important that both MODWT and DWT based approaches outperform their non-wavelet equivalents. From all of the methods MODWT PA has the best result concerning the impact coefficient distribution.

3.6 Summary

In this Section methods pertinent to trend identification problem were discussed. Three quality measures of a trend were presented. Theoretical part was followed by some examples which evaluated derived trends with proposed measures. In case of averages of two measures, namely fitness and distribution of an impact coefficient MODWT PA approach provided the best results. In case of a smoothness a linear trend is, of course the best. If averages of all measures are considered DWT LR approach gets dominated by the MODWT based one. All in all a general conclusion is that WA enforcement improves results of trend fitting.

Moreover, three groups of approaches to evaluating trends are established. As a summarizing remark we state that the proposed measures are good for our research needs. However, in other applications as some of these measures present a contrary view to the trend quality problem (e.g. in case of most real-life phenomena the more smooth the trend

Table 1: Results of measuring fitness of derived trends for (a) corn and (b) wheat yields.

(a)

		Fitness Measure											
		SSE				wSSE				mSSE			
	%	MAX	AVG	MIN	SD	MAX	AVG	MIN	SD	MAX	AVG	MIN	SD
MODWT Polynomial Approximation	100	27015,7	14042,8	6961,3	3366,8	489,9	187,9	73,0	78,3	5,046	2,458	1,145	0,758
	95	17452,2	8499,7	4044,5	2347,4	273,5	98,7	40,6	40,1	3,742	1,560	0,661	0,547
	90	13246,2	6114,5	2766,3	1767,7	206,1	71,6	24,1	29,6	3,073	1,195	0,456	0,436
	75	5153,6	2424,1	1097,4	759,7	89,5	31,2	12,2	12,8	1,495	0,600	0,228	0,231
DWT Linear Regression	100	27491,9	14543,6	6877,5	3483,4	497,0	195,3	76,8	78,2	5,151	2,537	1,168	0,750
	95	17581,5	8975,1	3706,5	2461,9	261,1	105,8	43,6	39,0	3,792	1,645	0,609	0,559
	90	13561,9	6734,0	2796,4	1923,8	195,0	80,7	30,6	30,0	3,105	1,313	0,482	0,458
	75	5915,1	2971,5	1240,9	907,9	96,1	39,3	13,8	14,9	1,573	0,730	0,261	0,256
Polynomial Approximation	100	31222,3	15172,1	7158,4	4140,8	590,3	208,4	78,7	95,0	5,872	2,649	1,248	0,870
	95	18705,2	9102,2	4404,2	2661,2	298,9	106,8	44,6	44,7	4,052	1,667	0,738	0,604
	90	14528,2	6638,4	3001,3	2008,9	203,6	79,1	29,8	34,4	3,239	1,293	0,544	0,482
	75	5805,9	2767,8	998,9	948,1	114,1	36,2	10,9	15,9	1,708	0,681	0,252	0,272
Linear Regression	100	31343,7	16467,7	7549,0	3969,5	600,4	228,4	88,9	93,8	5,925	2,869	1,282	0,836
	95	18808,4	10153,2	4821,8	2565,0	290,3	123,1	51,2	45,5	4,057	1,857	0,911	0,590
	90	14422,9	7591,6	3418,0	1934,5	227,9	93,6	40,1	33,3	3,215	1,476	0,651	0,464
	75	7214,6	3415,1	1408,1	961,8	110,5	45,6	19,4	15,2	1,644	0,833	0,363	0,263

(b)

		Fitness Measure											
		SSE				wSSE				mSSE			
	%	MAX	AVG	MIN	SD	MAX	AVG	MIN	SD	MAX	AVG	MIN	SD
MODWT Polynomial Approximation	100	2249,2	728,6	0,0	491,2	160,1	20,0	0,0	16,3	1,486	0,627	0,000	0,368
	95	1246,3	476,8	0,0	320,0	36,3	12,4	0,0	8,6	1,100	0,447	0,000	0,264
	90	1035,5	392,3	0,0	276,3	31,3	10,2	0,0	7,5	0,960	0,377	0,000	0,235
	75	504,6	173,4	0,0	127,1	17,8	4,6	0,0	3,5	0,720	0,204	0,000	0,142
DWT Linear Regression	100	2024,6	770,1	0,2	512,3	96,4	20,7	0,0	14,2	1,479	0,667	0,002	0,348
	95	1567,3	523,3	0,2	342,7	35,0	13,5	0,0	8,5	1,212	0,494	0,002	0,247
	90	1387,2	437,5	0,1	296,6	30,3	11,3	0,0	7,4	1,106	0,426	0,001	0,221
	75	758,5	208,9	0,1	145,2	18,1	5,5	0,0	3,7	0,771	0,254	0,001	0,138
Polynomial Approximation	100	2436,2	838,9	0,0	558,0	289,9	23,9	0,0	23,5	1,971	0,734	0,000	0,433
	95	1599,7	542,6	0,0	365,0	40,4	14,1	0,0	9,8	1,757	0,516	0,000	0,313
	90	1356,3	445,4	0,0	316,1	33,4	11,5	0,0	8,4	1,571	0,430	0,000	0,278
	75	556,0	198,0	0,0	147,4	18,0	5,2	0,0	4,0	1,293	0,235	0,000	0,171
Linear Regression	100	2654,0	1060,2	12,5	635,4	340,5	30,0	0,6	26,5	2,177	0,969	0,068	0,428
	95	1697,6	707,0	12,5	410,0	42,3	18,5	0,6	10,3	1,789	0,713	0,068	0,306
	90	1490,4	584,3	4,4	362,4	38,5	15,3	0,2	9,4	1,533	0,595	0,027	0,289
	75	704,0	277,4	2,6	173,7	20,4	7,3	0,1	4,6	1,193	0,352	0,019	0,189

Table 2: Results of measuring smoothness of derived trends for corn yields.

(a)

	Smoothness Measure			
	BKSM			
	MAX	AVG	MIN	SD
Original	13,7934	9,6378	6,3803	1,3908
MODWT Polynomial Approximation	1,8009	0,7645	0,3302	0,2583
DWT Linear Regression	3,4123	1,3536	0,4115	0,4940
Polynomial Approximation	0,1701	0,0171	0,0000	0,0287
Linear Regression	0,0000	0,0000	0,0000	0,0000

(b)

	Smoothness Measure			
	BKSM			
	MAX	AVG	MIN	SD
Original	6,7308	3,8718	0,9500	1,0866
MODWT Polynomial Approximation	5,7500	1,0361	0,2048	0,8080
DWT Linear Regression	5,0641	1,2319	0,2080	0,6877
Polynomial Approximation	5,7500	0,4192	0,0000	0,9370
Linear Regression	0,0000	0,0000	0,0000	0,0000

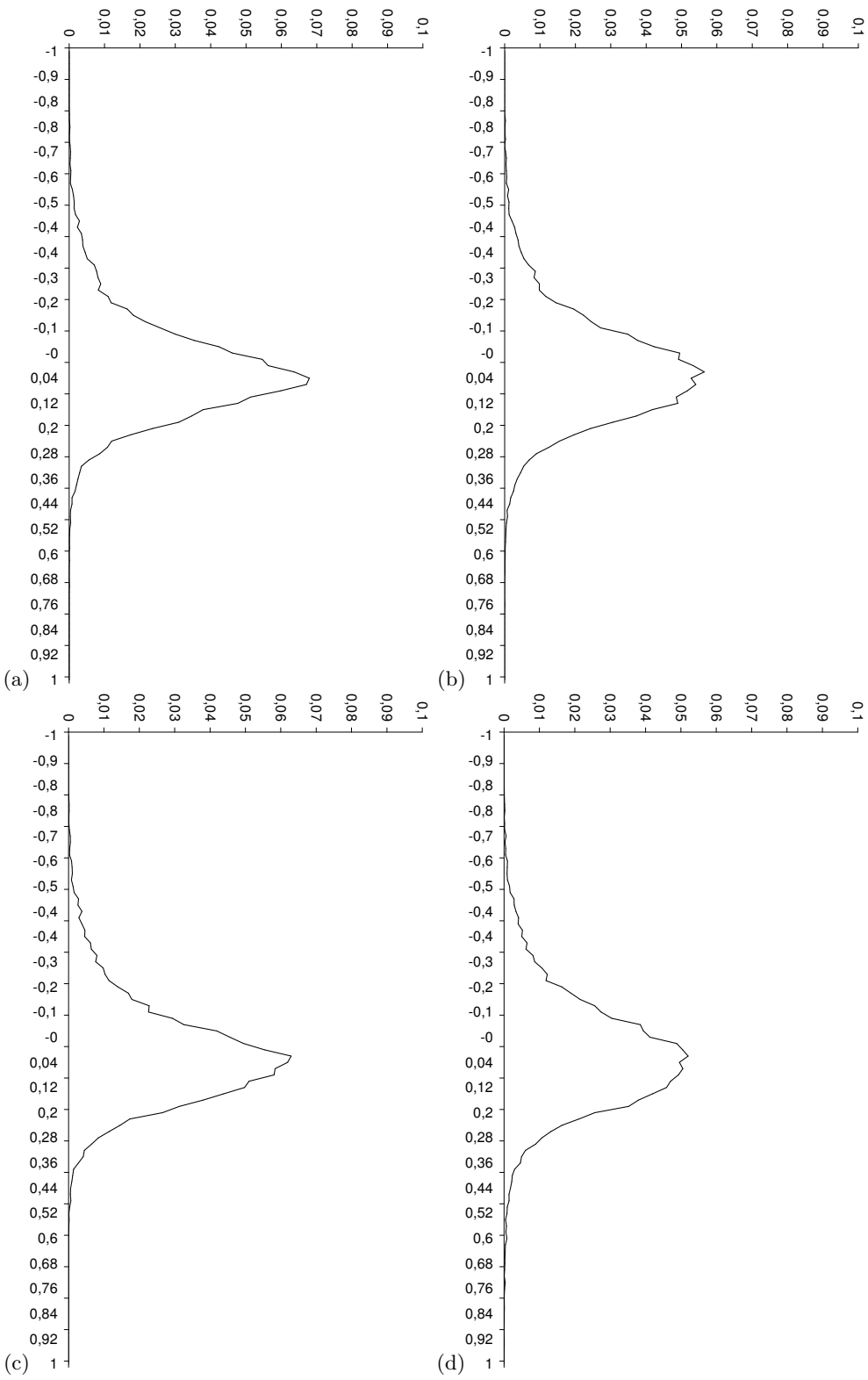

Figure 10: Distribution of values of i_t for corn, for (a) MODWT PA, (b) DWT LR, (c) PA, and (d) LR, respectively.

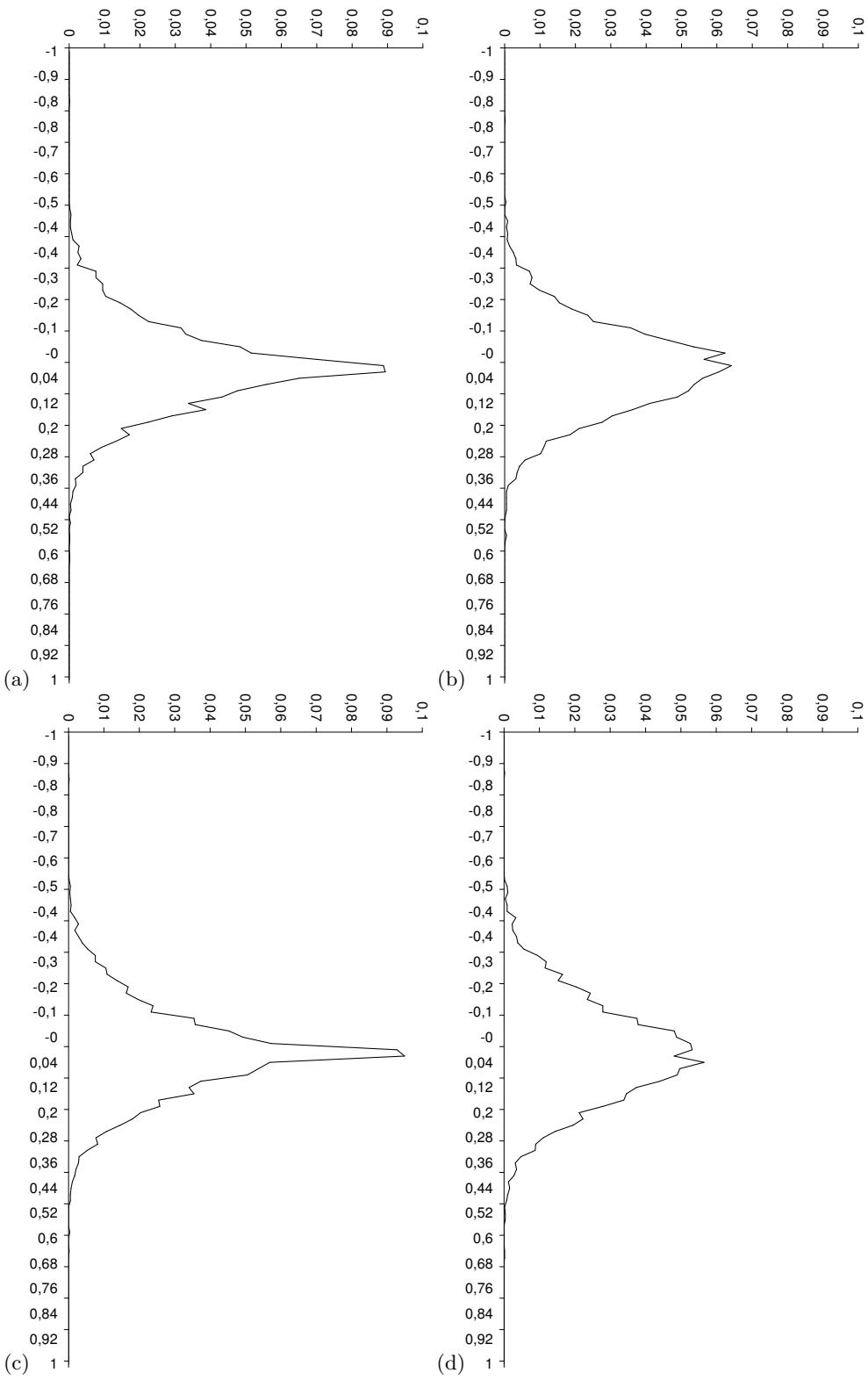

Figure 11: Distribution of values of i_t for wheat, for (a) MODWT PA, (b) DWT LR, (c) PA, and (d) LR, respectively.

will be the worse fit to the original time series it will have) some extensions could be considered. E.g. the "final" choice of the best trend may be made with a multi-criteria approach (Wierzbicki 1980).

4 Periodicity Identification

4.1 The Purpose

This Section focuses on the weather data and seeks for the methodology to identify potential periodicities which then may be used in further analyses. For example in various forecasting algorithms one needs to specify the length of the season. If this information is not known or not easily derivable it is necessary to provide a tool for periodicity identification. Herein a description of such a method (based on wavelets) is provided and experimentally verified.

4.2 The Methodology

DWT and MODWT consider only scales which are powers of 2 but for periodicity identification it is desired to detect any potential lengths of season, e.g. 3, 4, 5, 6, 7, 9, 10, 11, **1**2, 13, 14, ... As shown below especially loss of a length of 12 is not desired. Therefore, D2DWT is used as WT. D2DWT transform calculates wavelet coefficients on all levels (i.e. not necessarily of power of 2) and as a result any possible period length may be discovered with this methodology.

Six main steps of this approach are:

- D2DWT transform of a time series,
- calculate smoothness measures for all wavelet coefficients series,
- determine local minimums of smoothness,
- select number of wavelet level for which the local minimum was also found for wavelet coefficients series of multiplied level,
- repeat above steps for all available series of the phenomenon,
- select most frequently occurring number of wavelet level (in case of a very equally distributed values it is considered that there is no seasonality).

This may be expressed by the following algorithm (for notation see p. *viii*):

Algorithm 4.1: WAVELETPERIODICITYIDENTIFICATION($\mathbf{X}, J$)

for each time series $\mathbf{X}$ of a given phenomenon
do
- $\breve{\mathbf{W}} \leftarrow$ D2DWT of $\mathbf{X}$
- **for each** $\breve{\mathbf{W}}_j \subset \breve{\mathbf{W}}$
  - **do** $R^j \leftarrow$ smoothness of $\breve{\mathbf{W}}_j$ according to a given smoothness measure (in this case R_{WBK})
- $PeriodLength_{\mathbf{X}} \leftarrow j$ of the best of R^j, according to the criterion provided together with the measure

$Period \leftarrow$ most frequent value among $PeriodLength_{\mathbf{X}}$ or return 0 if the frequency distribution is flat

return ($Period$)

Level number *Period* returned by an above described algorithm is a detected length of a period for a given phenomenon.

4.3 Illustrative Example

The first example shows that the proposed methodology is verified on the seasonal time series of monthly precipitation observations. Fig. 12 shows the original time series of these observations.

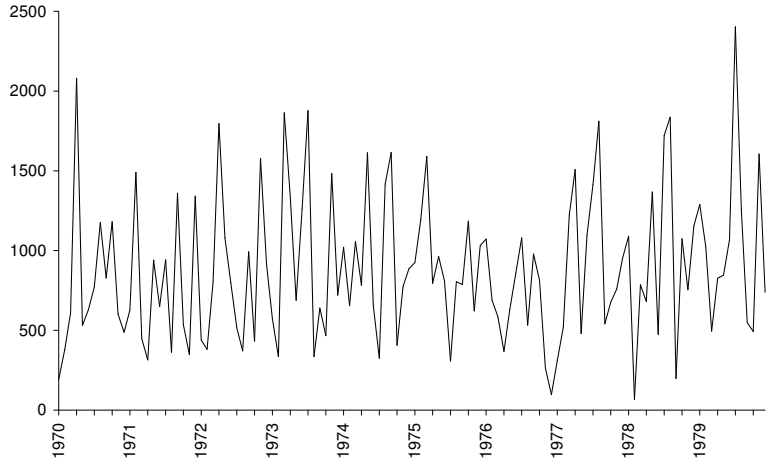

Figure 12: Original time series of precipitation in Bartholomew county in Indiana.

Fig. 13 presents several D2DWT wavelet coefficients series and Tab. 3 contains results of smoothness measuring for this time series but more calculations is done for all the available data.

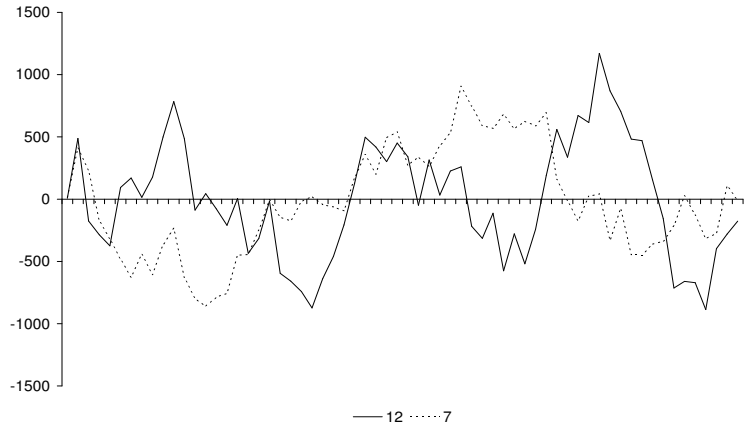

Figure 13: D2DWT coefficients series calculated from time series from Fig. 12. Length of season equals 7 and 12, respectively.

The results show that the smallest repeatable local minimum in values of the smoothness measure is for wavelet coefficients series of length 12. Fig. 14 shows a seasonal decomposition of a sample time series for which seasonality seems not very obvious. The result of 12 month periodicity was observed for a majority of time series so the final conclusion is that there is a yearly seasonality in precipitation data.

This experiment has been performed for all series of weather parameters on the whole available database. As a result it occurred that the most frequent value of a season length is 12. This is consistent with the general knowledge.

Additional experiments were performed on various profiles and aggregations of the weather parameters' data. Values of season length were equally distributed for all weather

Table 3: Results of measuring smoothness of wavelet coefficient series for time series from Fig. 13

Length	Smth.	Length	Smth.	Length	Smth.	Length	Smth.	Length	Smth.
1	358,714	16	87,524	31	65,644	46	52,591	61	43,605
2	272,139	17	86,042	32	62,453	47	52,017	62	46,439
3	209,213	18	86,519	33	59,493	48	48,503	63	45,321
4	176,392	19	81,353	34	62,224	49	52,204	64	42,624
5	153,477	20	77,963	35	59,503	50	50,216	65	45,543
6	146,354	21	75,439	36	59,663	51	50,464	66	43,936
7	139,688	22	76,351	37	58,869	52	48,912	67	44,553
8	125,069	23	77,98	38	58,122	53	48,659	68	41,303
9	119,962	24	68,23	39	56,214	54	51,26	69	45,067
10	113,659	25	69,661	40	54,978	55	49,492	70	43,164
11	107,548	26	69,882	41	56,591	56	49,962	71	42,29
12	97,091	27	69,967	42	56,899	57	47,495	72	41,562
13	98,036	28	65,968	43	55,911	58	45,451	73	41,104
14	94,33	29	64,132	44	52,65	59	46,334	74	41,398
15	92,612	30	66,44	45	52,627	60	43,644	75	40,976

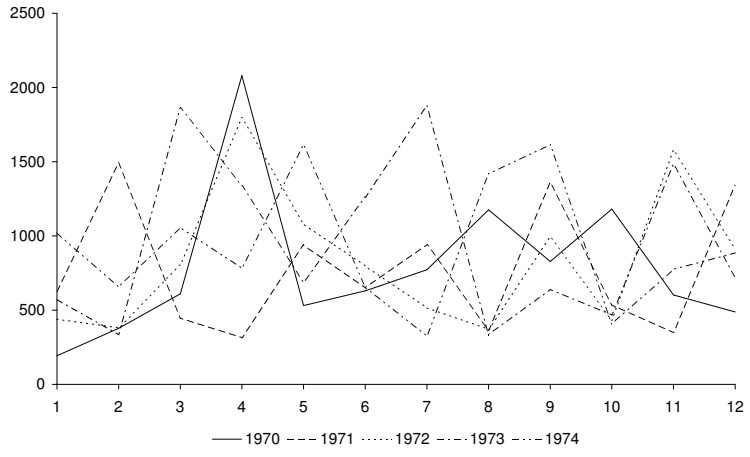

Figure 14: Decomposition of precipitation time series from Fig. 12 into seasons (first five).

factors so the conclusion is that there was no other periodicity information discovered.

4.4 Summary

The goal of this analysis was to discover various lengths of periodicity (season) for all considered weather factors, i.e. precipitation and maximum, mean, and minimum monthly temperatures. A periodicity identification method based on wavelet transform was introduced. It was applied to the U.S. weather data upscaled to the county level. The only periodicity detected is 12 months long; it was found for all considered weather factors, and obviously correspond to the common knowledge. The lack of longer (multi-year) periodicities provides arguments for disputes on multi-year weather seasonality.

5 Forecasting

5.1 The Purpose

This Section elaborates on possibilities of forecasting crop yields. Knowing the future yields could vastly improve planning in various areas, e.g. consumer market, food processing industry, trade.

Forecasting has been fascinating humanity since a very long time. First noticeable attempts we performed in Japan in 14th century (Murphy 1986). Currently a very popular approach to forecasting derives from Charles Dow's theory (Rhea 1993). History based forecasting applications in the stock market area are nowadays called technical analysis.

Herein a wavelet-based methodology of forecasting is presented. It is also compared with two simple standard approaches. Evaluation is performed based on the errors of forecasts.

5.2 Standard Forecasting Methods

The are various forecasting methodologies used in many areas of applications. Let us consider an example data shown in Fig. 15.

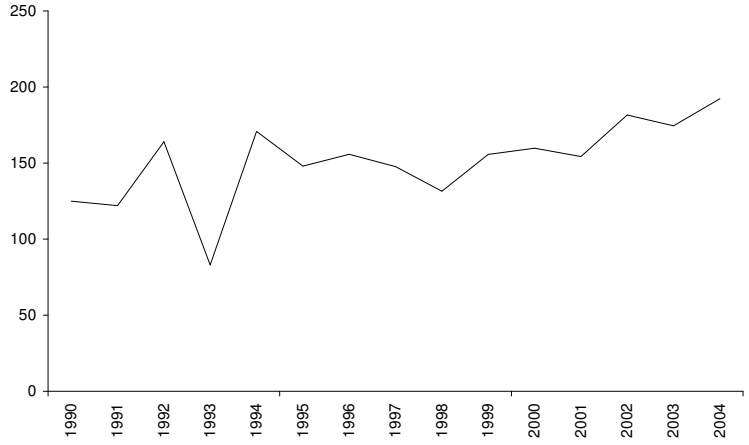

Figure 15: Original time series of corn yields in Webster, Iowa.

Below there are the series of results of methods using linear (Fig. 16) and polynomial (Fig. 17) extrapolations.
Please keep in mind that we consider a series of forecasts and therefore, the forecasts' plot presented in Fig. 16 not necessarily has to be a straight line.

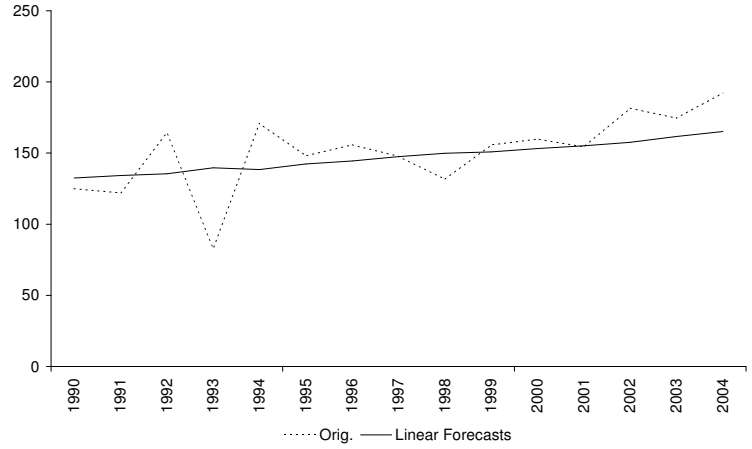

Figure 16: Result of forecasts of time series presented in Fig. 15 by linear extrapolations.

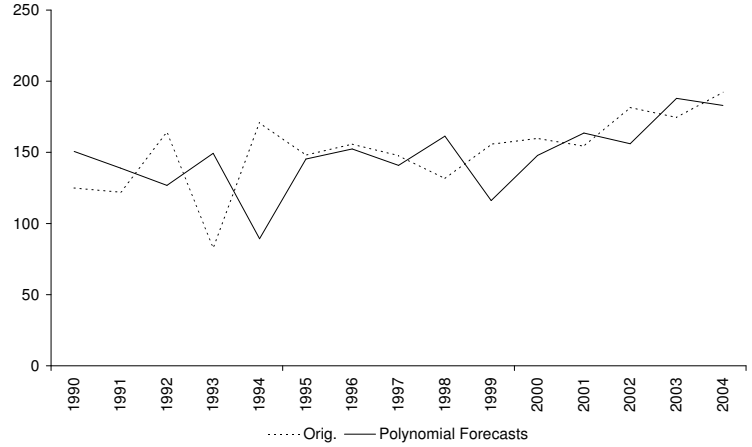

Figure 17: Result of forecasts of time series presented in Fig. 15 by polynomial extrapolations.

5.3 Direct Wavelet-based Approach

The essence of wavelet based forecasting is that approximation and extrapolation of time series are applied for each level in a wavelet space. Inversely transforming wavelets together with their forecasts back to the time series space provides a forecast of the original time series. This procedure is organized by Alg. 5.1 (see p. *viii* for the notation).

Algorithm 5.1: waveletForecast($\mathbf{X}, J$)

$\widetilde{\mathbf{W}} \leftarrow$ MODWT of $\mathbf{X}$
for each $\widetilde{\mathbf{W}}_j \subset \widetilde{\mathbf{W}}$
do $\begin{cases} \widetilde{E}_j \leftarrow \text{ extrapolation of } \widetilde{\mathbf{W}}_j \\ \widetilde{\mathbf{W}}'_j \leftarrow \widetilde{\mathbf{W}}_j \cup \left\{\widetilde{E}_j\right\} \\ \mathbf{W}_j \leftarrow \text{ an orthonormal (for level } j\text{) subset of } \widetilde{\mathbf{W}}_j \text{ containing } \widetilde{E}_j \end{cases}$
$\mathbf{X}' \leftarrow$ iDWT from all $\mathbf{W}_j$
$E \leftarrow \mathbf{X}' \setminus \mathbf{X}$
return (E)

In order to forecast the wavelet space coefficients similar methods to those used in direct time series forecasting can be applied. Because these methods are applied many times it is important to be aware of their computational complexity.

In this case a forecast of the mean temperature is attempted. Fig. 15 presents the original time series which is going to be forecasted.
Fig. 18 shows the results of series of forecasts performed using the described methodology.

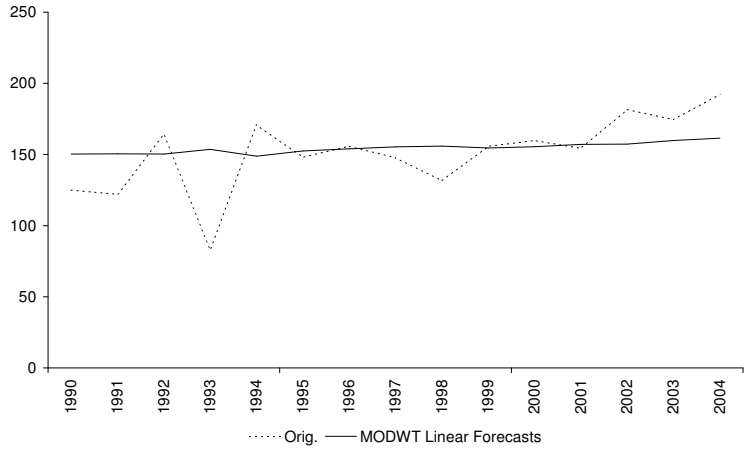

Figure 18: Result of forecasts of time series presented in Fig. 15 by MODWT-based method with linear extrapolation.

5.4 Evaluations

Tables 4a and 4b summarize the results of the forecasting methods for corn and wheat respectively based on the errors of forecasts. The following notation is used:

- MAX - maximum error,
- AVG - average error,
- MIN - minimum error,
- SD - standard deviation.

Table 4: Results of evaluations of MODWT linear, polynomial, and simple linear forecasts for (a) corn and (b) wheat.

(a)

	MAX	AVG	MIN	SD
MODWT Linear Forecasts	1,572	0,133	0,000	0,147
Linear Forecasts	1,821	0,140	0,000	0,166
Polynomial Forecasts	2,428	0,176	0,000	0,202

(b)

	MAX	AVG	MIN	SD
MODWT Linear Forecasts	1,058	0,148	0,000	0,139
Linear Forecasts	1,123	0,162	0,000	0,146
Polynomial Forecasts	11,399	0,789	0,004	1,493

It may be easily seen (as it is a case for all of above mentioned measures of error) that the best accuracy of the forecast is established using wavelet based methodology. The worst results are coming from the polynomial extrapolation. The reason for this is that the polynomial approximation may fit very well at the ends of approximation intervals, and as a result extrapolation may be misled for values far from the maximum and minimum values of the time series.

5.5 Summary

This Section provided a description of enforcing forecasting methods with WT. Two traditional approaches and a wavelet-based method based on one of them (the linear) are explored and evaluated. WT enforced methodology provided better results in terms of the used measures of forecasts' quality.

6 Conclusions and Further Studies

6.1 Conclusions

This paper describes several methods utilizing wavelets and their application to analysis of time series of crop yields and weather factors. Possibilities of application of wavelets were explored and positively verified to the following problems:

- trend identification,
- periodicity identification,
- forecasting.

For the trend analysis wavelet-based methods provided a better fit to the original data than the other methods used for this purpose. Also the distribution of an impact coefficient defined on top of WA-based trends had better characteristics.

Periodicity identification method found a yearly seasonality which is commonly known. It did not find any other periodicities what according to the experts is a positive result.

The forecasting method resulted with better quality of forecast with respect to applied measures when using wavelet-based approach.

It is worth mentioning that a special Data Analysis Framework was prepared for the purpose of this research. This framework consists of almost 100 classes implemented in Java language. Classes are designed based on modified JavaBeans specification and comply with a high reusability paradigm. Each of these classes is designed for a special purpose, e.g. to perform a Haar DWT, MODWT forecast, polynomial approximation, and many others.

Also current status of the research has some limitations which are a direct impulse for further work on this topic. Those limitations include, but are not limited to: lack of consideration of spatial variability of precipitation and landscape, soil buffering, or irrigation. Also the water deficit should be used directly instead of precipitation as the latter only indirectly and imprecisely captures the deficit.

Finally during the reported research it was found that the weather data aggregation to the monthly level may be too coarse. Monthly level aggregation looses too much information about quick changes in the weather factors. Especially in case of WA, which performs its own aggregations of a specific type, this loss is even bigger. E.g. in hydrology for weather data a 10 day aggregation is used.

6.2 Possible Further Studies

The reported research has shown the potential of wavelet-based methods for data analysis and provides a good basis for further investigations. Below some possible activities are summarized.

Particularly analysis of spatial aspects of quality distributions of trend approximations and forecasts seems to be very interesting. However interpretation of results of this analysis would require either an insight into the explored domain or an interaction with domain experts.

Another forecasting method could be also applied and tested for quality. This method is described in (Kozlowski 2004) and uses information about seasonality (e.g. derived from the periodicity identification approach). It is highly probable that its application to the seasonal weather data would provide good results.

Moreover, there is an issue with data correctness. To address this problem one should explore methods of outliers identification and handling missing data. Hence, in the preprocessing phase of the research two wavelet-based outliers identification methods described in (Donoho and Johnstone 1995) and (Kozlowski 2005) could be applied and missing values treatment with wavelets could be investigated.

As the research was performed parallely and in cooperation with two other YSSP 2005 participants (Hai Nguyen and Wojciech Kotłowski) it would be good to join the results and explore possibilities of common applications of all the approaches. Hai has been using sliding windows technique and Wojciech applied rough sets and decision rules. The dependences and possible future collaboration areas between these researches are illustrated in “Joint Research Overview Diagram”, in Fig. 19.

Considering the last concluding remark summarizing the results (weather data aggregation) another natural step would be to perform analysis with data of lower level aggregation. Most preferable, in case of WA enforced methodologies, would be a daily collected data.

Another possibility of extending this research is a big extension of evaluations of presented methodologies. Especially a comparison with a spectrum of standard methodologies should be performed.

Finally, a continuation of this study could exploit legacy of a quite sophisticated software developed during the Summer 2005.

References

Aczel, A. D.: 1989, *Complete Business Statistics*, Irwin, Boston, MA.

Donoho, D. L. and Johnstone, I. M.: 1995, Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.* **90**, 1200–1224.

Haar, A.: 1910, Zur Theorie der orthogonalen Funktionensysteme, *Mathematische Annalen* **LXIX**, 331–371.

Kotlowski, W.: 2007, Qualitative models of climate variations impact on crop yields.

Kozlowski, B.: 2004, On time series forecasting methods of linear complexity utilizing wavelets, *Advances in Intelligent Systems - Theory and Applications, In Cooperation with the IEEE Computer Society*, Kirchberg – Luxembourg.

Kozlowski, B.: 2005, Time series denoising with wavelet transform, *Journal of Telecommunications and Information Technology* **3**, 91–95.

Li, T., Li, Q., Zhu, S. and Ogihara, M.: 2003, Survey on wavelet applications in data mining, *SIGKDD EXplorations* **4**, 49–68.

Mallat, S. G.: 1989, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7), 674–693.

Murphy, J. J.: 1986, *Technical Analysis of Futures Markets: A Comprehensive Guide to Trading Methods and Applications*, New York Inst. of Finance, New York, NY.

Percival, D. B. and Walden, A. T.: 2000, *Wavelet Methods for Time Series Analysis*, Cambridge University Press, Cambridge, UK.

Prasad, L. and Iyengar, S. S.: 1997, *Wavelet Analysis with Applications to Image Processing*, CRC Press, USA.

Rhea, R.: 1993, *The Dow Theory*, Fraser Publishing, USA.

Wierzbicki, A. P.: 1980, *The Use of Reference Objectives in Multiobjective Optimization*, Springer-Verlag, New York, NY.

Appendices

A Joint Research Overview Diagram

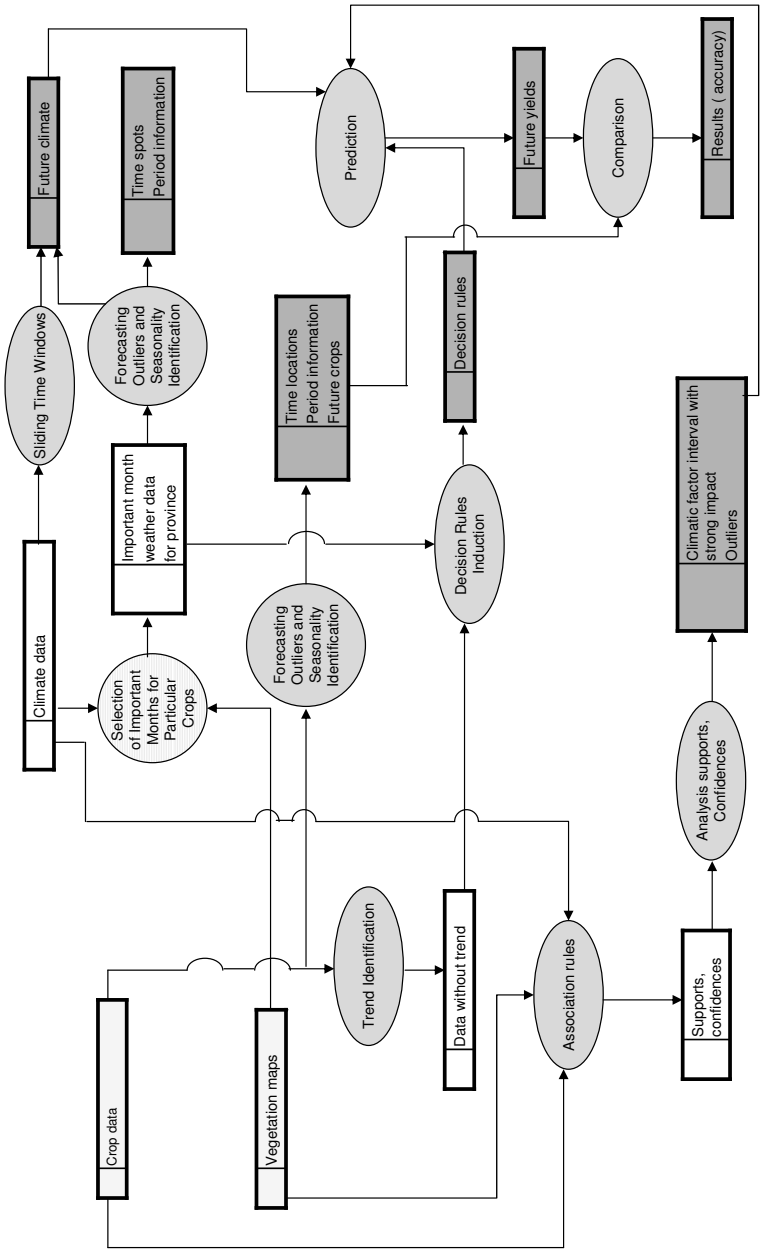

Figure 19: Joined research overview diagram showing dependences and areas of possible future collaboration with two other YSSPers, namely Hai Nguyen and Wojciech Kotłowski.

B Wavelet Analysis

B.1 Wavelet Background

As a beginning of wavelet methodologies we consider year 1909 when Hungarian mathematician Alfred Haar introduced a two-state function in appendix to his doctoral thesis published later on (Haar 1910). Today a slightly modified version of this function is regarded as the first basic wavelet function. Wavelet founded methods began to develop very quickly in the 1990s and turned out to be very useful for analysis of many problems, e.g. including analysis and synthesis of time series (Percival and Walden 2000) (in acoustics, geology, meteorology, and economics), effective data storage, especially images (Prasad and Iyengar 1997) (computer graphics, movie industry). Lately a very fast development of wavelet-based data mining (see e.g. (Li, Li, Zhu and Ogihara 2003)) techniques may be observed.

B.2 Wavelet Functions

Wavelets are functions having nonzero values in a relatively short interval. In this regard they differ from "normal", long waves (such as sinusoids) which are determined on a whole time domain $(-\infty, \infty)$. Let ψ be a real function of a real variable u, which satisfies two conditions:

$$\int_{-\infty}^{\infty} \psi(u)du = 0 \tag{1}$$

and

$$\int_{-\infty}^{\infty} \psi^2(u)du = 1. \tag{2}$$

Condition (2) means that for any ϵ from an interval $(0, 1)$ there is an interval $(-T, T)$ such that:

$$\int_{-T}^{T} \psi^2(u)du = 1 - \varepsilon.$$

If ϵ is close to 0, it may be seen that only in an interval $(-T, T)$ corresponding to this ϵ values $\psi(u)$ are different than 0. Outside of this interval they must equal 0. Interval $(-T, T)$ is small compared to an interval $(-\infty, \infty)$, on which a whole function is determined. Condition 1 implies that if $\psi(u)$ has some positive values, it also has to have some negative ones (a function "waves"). Therefore Eqs. (1) and (2) introduce a concept of a small wave, conventionally called wavelet. If Haar function ϕ, which is a two-state function of real variable (Fig. 1a):

$$\phi(u) = \begin{cases} -1 & \text{for} \quad -1 < u \leq 0 \\ 1 & \text{for} \quad 0 < u \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

would be transformed into:

$$\psi^{(H)}(u) = \begin{cases} -\frac{1}{\sqrt{2}} & \text{for} \quad -1 < u \leq 0 \\ \frac{1}{\sqrt{2}} & \text{for} \quad 0 < u \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

then the resulting function $\psi^{(H)}$ satisfies conditions (1) and (2), and is called Haar basic wavelet function (Fig. 1b).

B.3 Wavelet Transform

In this report only discrete wavelet transforms are considered. Having an original time series $\mathbf{X}$ and a function ψ on the input, wavelet transform (WT) results in a vector $\mathbf{W}$ of wavelet coefficients and scaling coefficient(s). These coefficients are calculated at different wavelet levels j (using different scales λ) and in a wide range of locations t. Wavelet coefficients are grouped into level corresponding vectors $\mathbf{W}_j \subset \mathbf{W}$. For each level a vector $\mathbf{W}_j$ consists of I_j $W_{j,i}$ wavelet coefficients ($i = 0, ..., (I_j - 1)$). The last element of a vector $\mathbf{W}$ is a vector $\mathbf{V}_J$

There are many different WTs. Two most often applied are (orthonormal) discrete wavelet transform (DWT) (results are shown in Fig. 2) and its slightly modified version which preserves scales but calculates wavelet coefficients in more densely chosen locations - maximal overlap discrete wavelet transform (MODWT) (results are shown in Fig. 20). For periodicity identification a dense level discrete wavelet transform (D2DWT) is used (results are shown in Fig. 21), which is a modification of MODWT that calculates wavelet coefficients for all possible wavelet levels without respect to the power of 2 constraint.
First wavelet coefficients set of D2DWT is the same as the first one of the MODWT. Sets (b) and (c) do not appear in MODWT. Set number four (not shown in Fig. 21) of D2DWT is the same as set (b) of MODWT (shown in Fig. 20).

For a DWT Mallat (Mallat 1989) proposed a very fast algorithm for calculating wavelet coefficients. With minor modifications it may be applied to other wavelet transforms.

An important feature of DWT is that it may be reversed by an inverse Wavelet Transform (iWT). Original time series can be calculated from wavelet coefficients. This possibility is fundamental for almost all relevant current wavelet applications. Of course each WT has its own implementation of iWT so with transforms DWT, MODWT, and D2DWT inverse transforms iDWT, iMODWT, and iD2DWT are applied, respectively.

Most of the currently applied methodologies using wavelets base on a following schema:

- transformation of an original signal to the wavelet space is done,
- desired analysis is performed on wavelet coefficients resulting from the first step,
- newly established wavelet coefficients are transformed into the starting point space.

This schema (illustrated in Fig. 22) applies to all methodologies used throughout the research presented in this report except for periodicity identification (see Sec. 4) methods, which take an advantage of a multiresolutional analysis provided by wavelet transform to derive results in the original space, but do not require a inverse transformation (Fig. 23). This second approach uses information revealed by wavelet domain to directly draw conclusions about the time series in the original space.
The research described herein uses a Haar DWT, Haar MODWT, and Haar D2DWT as WTs. Haar iDWT and Haar iMODWT were applied as iWTs, respectively.

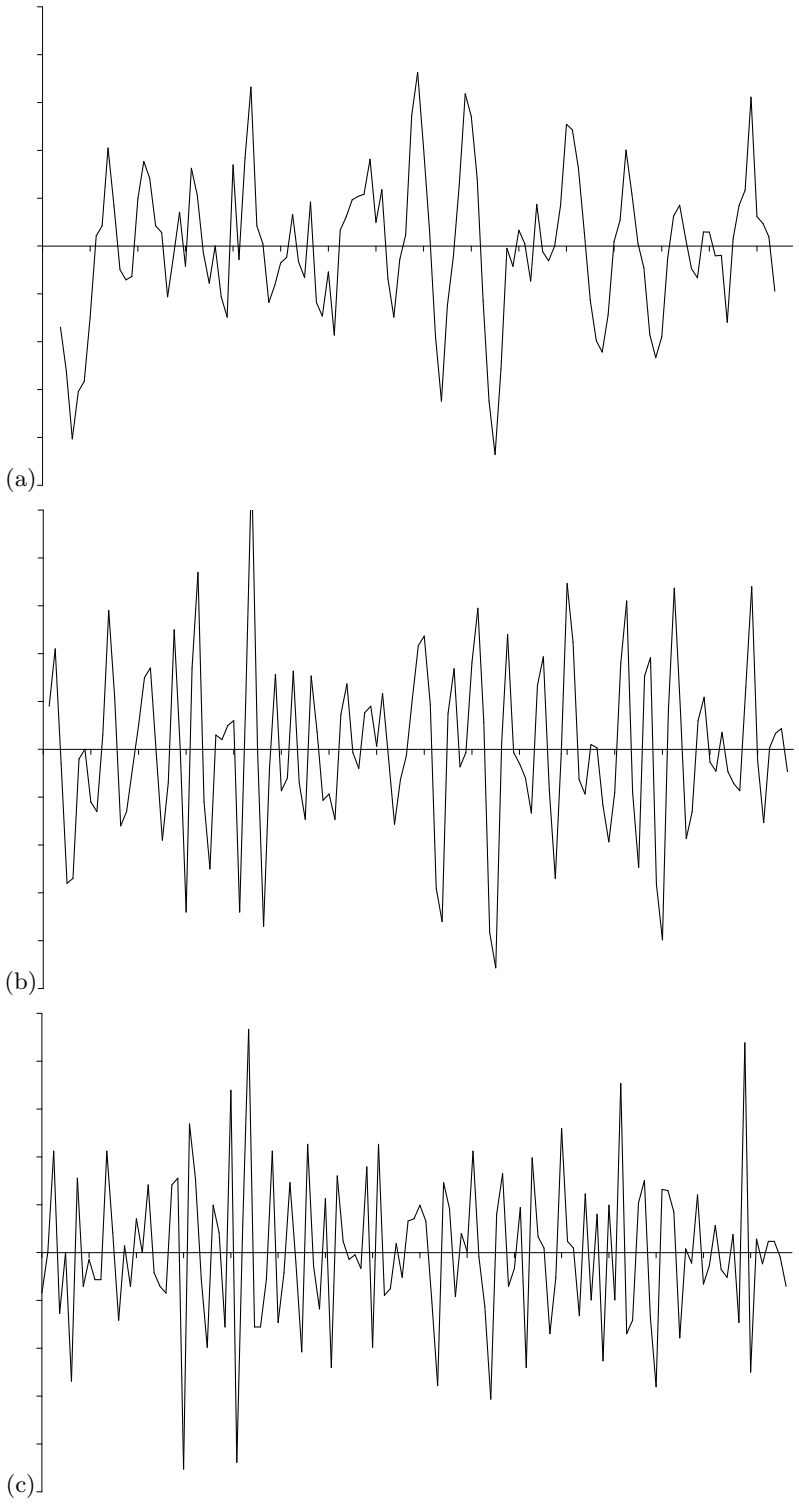

Figure 20: Wavelet coefficient sets of MODWT of time series presented in Fig. 1. Sub figures (c), (b), and (a) present values (vertical axes) of first three wavelet coefficients sets localized in time (horizontal axes).

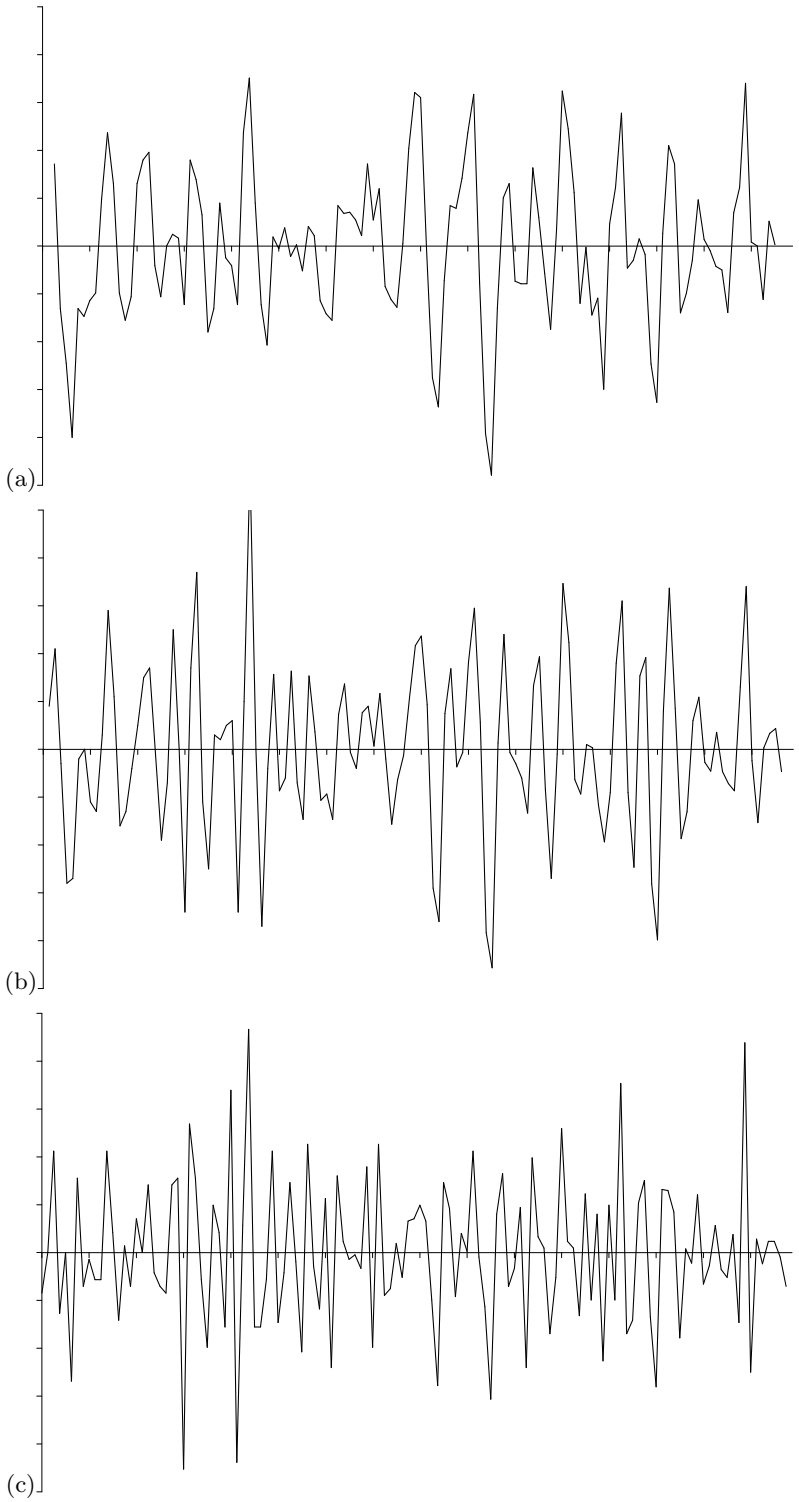

Figure 21: Wavelet coefficient sets of D2DWT of time series presented in Fig. 1. Sub figures (c), (b), and (a) present values (vertical axes) of first three wavelet coefficients sets localized in time (horizontal axes).

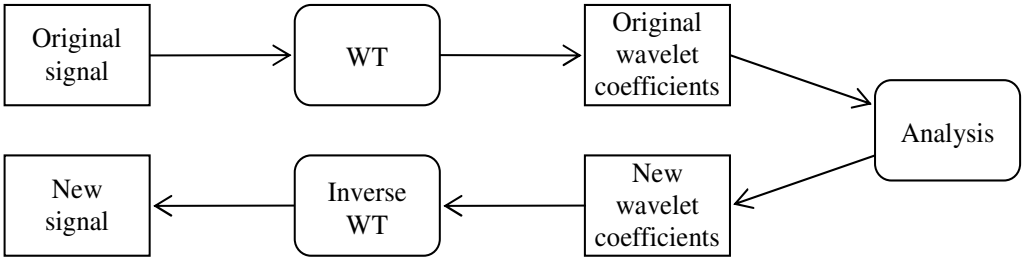

Figure 22: General diagram describing the idea of most wavelet-based analysis methods.

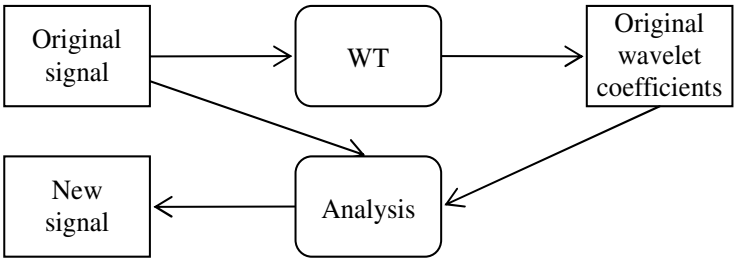

Figure 23: General diagram describing the idea of wavelet-based analysis methods, which do not take an advantage of the wavelet transform's inversability.

C WT Dedicated Polynomial Approximation

When using WT for e.g. trend identification or forecasting number of calculations of trends increases due to the characteristics of a WT. Therefore the computational complexity of used methods is very important. Herein a method of linear order of complexity for polynomial approximation is presented, that was developed for the use inside a WT enforced algorithms (Kozlowski 2004).

We look for approximation $y_i = \sum_{n=0}^{N} a_n x_i^n$ of $y = f(x)$. The approximation's accuracy criterion is usually defined as:

$$K = \sum_{i=1}^{I} (\sum_{n=0}^{N} a_n x_i^n - y_i)^2$$

Its value should be minimized, therefore:

$$\frac{\partial K}{\partial a_j} = 2 \sum_{i=1}^{I} x_i^j (\sum_{n=0}^{N} a_n x_i^n - y_i) = 0, \quad j = 0, 1, ..., N.$$

Hence:

$$\sum_{i=1}^{I} x_i^j (\sum_{n=0}^{N} a_n x_i^n - y_i) =$$

$$= \sum_{n=0}^{N} a_n \sum_{i=1}^{I} x_i^{n+j} - \sum_{i=1}^{I} y_i x_i^j = 0, \quad j = 0, 1, ..., N,$$

what may be presented as follows:

$$\begin{bmatrix} \sum_{i=1}^{I} x_i^0 & \cdots & \sum_{i=1}^{I} x_i^N \\ \sum_{i=1}^{I} x_i^1 & \cdots & \sum_{i=1}^{I} x_i^{N+1} \\ \vdots & \vdots\vdots\vdots & \vdots \\ \sum_{i=1}^{I} x_i^N & \cdots & \sum_{i=1}^{I} x_i^{2N} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{I} y_i x_i^0 \\ \sum_{i=1}^{I} y_i x_i^1 \\ \vdots \\ \sum_{i=1}^{I} y_i x_i^N \end{bmatrix}.$$

Having:

$$\bar{A}_k = \sum_{i=1}^{I} x_i^k, \quad k = 0, 1, 2, ..., N, ..., 2N-1, 2N \tag{3}$$

and:

$$B_n = \sum_{i=1}^{I} y_i x_i^n n = 0, 1, ..., N, \tag{4}$$

a following system of equations:

$$\begin{bmatrix} \bar{A}_0 & \bar{A}_1 & \cdots & \bar{A}_N \\ \bar{A}_1 & \bar{A}_2 & \cdots & \bar{A}_{N+1} \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ \bar{A}_N & \bar{A}_{N+1} & \cdots & \bar{A}_{2N} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_N \end{bmatrix},$$

is derived, where we want to evaluate elements a_n, $n = 0, 1, ..., N$.

Having:

$$A = \begin{bmatrix} \bar{A}_0 & \bar{A}_1 & \cdots & \bar{A}_{N-2} & \bar{A}_{N-1} & \bar{A}_N \\ \bar{A}_1 & \bar{A}_2 & \cdots & \bar{A}_{N-1} & \bar{A}_N & \bar{A}_{N+1} \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots & \vdots \\ \bar{A}_N & \bar{A}_{N+1} & \cdots & \bar{A}_{2N-2} & \bar{A}_{2N-1} & \bar{A}_{2N} \end{bmatrix},$$

A may be noted as $A = [A_{i,j}]_{(N+1)x(N+1)}$, where:

$$A_{i,j} = \bar{A}_{iN+j} \quad i, j \in \{0, 1, ..., N\}.$$

By replacing an n-th column of a matrix A with a column of (single columned) matrix B:

$$B = \begin{bmatrix} B_0 & B_1 & \cdots & B_N \end{bmatrix}^T,$$

matrix C_n is established:

$$C_n = \begin{bmatrix} A_{0,0} & \cdots & B_0 & \cdots & A_{0,N-1} & A_{0,N} \\ A_{1,0} & \cdots & B_1 & \cdots & A_{1,N-1} & A_{1,N} \\ \vdots & \vdots\vdots\vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots \\ A_{N,0} & \cdots & B_N & \cdots & A_{N,N-1} & A_{N,N} \end{bmatrix}.$$

Value a_n may be derived from a formula:

$$a_n = |C_n|/|A| \quad n = 0, 1, ..., N,$$

where $|A|$ and $|C_n|$ are principal determinants of matrices A and C_n respectively.

Expanding a determinant $|C_n|$ by an n-th column, results in:

$$|C_n| = \sum_{j=0}^{N} (-1)^j B_j \, |D_{jn}| \, ,$$

where $|D_{jn}|$ are determinants of a matrix D_{jn} derived from C_n (or from a matrix A) by removing its j-th row and n-th column, therefore:

$$a_n = \frac{1}{|A|} \sum_{j=0}^{N} (-1)^j B_j \, |D_{jn}| \quad n = 0, 1, ..., N.$$

Having $x_i = i$, $i = 1, 2, ..., i, ..., I$ approximation problem of any discrete time series can be replaced by an equivalent problem of approximation of a time series with the same values but on a set $I = \{1, 2, ..., i, ..., I\}$ (subsequently one can go back to the original time series). Thus (3) and (4) may be replaced by:

$$\bar{A}_k = \sum_{i=1}^{I} i_i^k, \quad k = 0, 1, 2, ..., N, ..., 2N - 1, 2N,$$

and:

$$B_n = \sum_{i=1}^{I} y_i i^n, \quad n = 0, 1, ..., N,$$

respectively.

As a result values $|A|$ and $|D_{jn}|$, $i, j \in \{0, 1, ..., N\}$ become equal (for all time series transformed in an above described way). Those values could therefore be calculated only once and stored. Likewise when calculating values B_n dependent on y, $n = 0, 1, ..., N$ subsequent natural numbers' powers may be obtained from an adequate table.

The described method allows for a substantial decrease of computational time because it provides a polynomial approximation algorithm of discrete signal with linear, considering a degree of a polynomial and length of a time series, order of complexity. It may be applied within any WT driven process without increasing its own (in most cases linear) order of computational complexity.