Graduate Studies
Mathematical Sciences

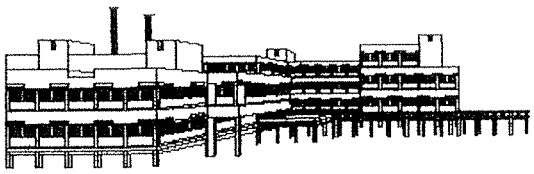
University of Twente

Wavelet Shrinkage for Modeling Time Series Data: a Data Compression Method

Agus Sukmana

Executed at: **Amsterdam Water Supply, Amsterdam**

Gemeentewaterleidingen
Amsterdam



Msc Engineering Mathematics

Final Project Report

Date: 8 June 1999

Wavelet Shrinkage for Modeling Time Series Data: a Data Compression Method

Agus Sukmana

Executed at: **Amsterdam Water Supply, Amsterdam**

Members of the student's Final Project Committee

Company Supervisor	Ing. V. A. W. G. Jansen.
UT- Supervisor	Dr. K. Poortema, Faculty of Mathematical Sciences
Third Member	Prof. Dr. W. Albers, Faculty of Mathematical Sciences
Fourth Member	Dr. J. F. Frankena, Faculty of Mathematical Sciences

Date: 8 June 1999

Preface and A cknowledgements

Even a journey of one thousand lie begins with a single step.
Lao Tze.

Wavelet is not a trendy name for a kind of food (the sound is rather like *omelette*). Wavelet is a mathematical tool with a great variety of possible applications.

I have been studying the application of wavelet since January 1999, and this thesis is my first step result. Since there is much room for error and difference of opinion in my thesis, I will be grateful for comments. (For this purpose, please send an email to asukmana@home.unpar.ac.id).

This thesis and computer program has been improved by the comment of my supervisor who has been kind enough to read and to discuss it. I should like to thank *Vincent Jansen* and *Klaas Poortema* for their invaluable assistance in this respect.

Many people assisted me at various times in preparing this thesis. I wish to thank *Jan Scholte*, *Rinaldo Saimbang*, *Oscar Werner*, *Kees van der Drift*, *Jurek Gorniewicz*, *Piet Visser*, *Dharma Lesmono*, *Arianto Wibowo*, *Gregoria*, *Saladin Uttunggadewa* and his family. This thesis and my study could not have been completed without the financial support of the *Parahyangan University* at Indonesia.

Finally, I am greatly indebted to my wife *Lingga*. The encouragement and understanding of her during my study in the Netherlands is gratefully acknowledged. The time and effort spent on studying and doing my research project must certainly have been reflected in neglect of her, whom I thank for her forbearance.

Amsterdam, June 99.

Summary

This thesis is my final project report that has been executed at Research and Development section, department of Distribution, Amsterdam Water Supply on from January until June1999.

The aim of this research project is designing a data compression scheme that is suitable for *cbd* data in general sense, through a modeling process. The compressed data must be contained small numbers of parameters (compare to the original data) and it can be reconstructed to original data (with high quality of approximation)

The original data is compressed by transforming into wavelet coefficients in the wavelet domain via Discrete Wavelet Transform (DWT). Because the value of most of the wavelet coefficients are close to zero, so a cutting method is done using three thresholding techniques viz., *hard*, *soft*, and *SURE* that yield wavelet shrinkage coefficients. Wavelet shrinkage coefficients will be stored as a representation of the original data. The proportion of the number of non-zero parameters of the wavelet shrinkage and the number of original date is used as a compression quality measure.

The compressed data are reconstructed by transforming the wavelet shrinkage coefficients back to the original domain through Inverse Discrete Wavelet Transform (IDWT). The difference between approximation data and original data is used as approximation quality measure.

The cutting method becomes essential in this problem; that's way the research is focused on how to design the cutting procedure. Three scenarios have been designed viz., scenario that emphasis on the compression, approximation and compromise between both compression and approximation. Six data sets are chosen as representations of cbd data; scenario 3 gives the best results 7.4% parameter and 99.12% for quality.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This thesis is my final project report that has been executed at section Research and Development, Department of Distribution, *Gemeentewaterleidingen Amsterdam* (GWA) from January until June 1999, under supervision of *Dr. K. Poortema* (Twente University) and *Ing. V. A. W. G. Jansen* (GWA).

GWA is the water supply company for Amsterdam and has been operating since 1850. This company is the Government Company that is responsible for the water supply for people and industries in Amsterdam and surroundings. Continuity of delivery and water quality are most important goals.

GWA supervises the water quality and distribution continuously. A continuous monitoring in distribution is done via measurement channels in water reservoirs, pump stations, etc. For hundreds of measurements in the process computer at the plant and in the city, most of which are available at the control center, the sector R&D can select 120 data channels on a 3 second basis. Nowadays 80 selections have been made, another 40 are in reserve. These channels measure various data such as pressure, flow, conductivity, water pH etc. The measurement is done every 3 seconds, and the result is recorded.

Measuring every 3 seconds enables to detect singularities as soon as possible, so they can take an action to handle it. The possible singularities are the following:
1. Decreasing of water quality
2. Decreasing of water pressure instantaneously.
3. A leaking in distribution pipe
4. A large demand at the same time, i.e. when there is a football match.

For each day one channel produces 28.800 measurements (= 20 observations per minute x 60 minutes per hour x 24 hour per day). One month, it produces 864.000 and about 7 millions measurements for whole active channels (there are 80 active channels). For storing those heaps of measurements (they are called *cbd*[1] data), they need large storage media. There are some techniques available for storing measurements in an efficient way. Those techniques are called data compression techniques.

The principle of data compression techniques is designing a compact representation of data generated by a data source [14]. The source data is converted to *the compressed data* by an *encoder*, and a *decoder* reconstructs the compressed data. There are two varieties of data compression systems; one of them is *lossy compression system*. In a lossy compression system, the decoder is not able to reconstruct the source data perfectly. The 'quantizer' is needed in this system for producing a new vector, which is close enough to the source data, before the data are converted in compressed data by the encoder.

In this thesis, a *lossy compression system* is chosen for compressing the *cbd* data. For applying this technique, a quantizer scheme is needed instead of an encoder. The following quantizer scheme is used for this project:

The source data are modeled to obtain a model with high approximation quality. Parameters of that model will be stored or encoded.

The Modeling process is done by considering the following matters:

1. The reconstruction result has high approximation quality. There is a trade-off between the number of involved parameters and quality of approximation.
2. The number of model's parameters should be smaller than the number of original data.
3. The general process (smoothness of curve process) is recorded.
4. The exceptional measurements are recorded.

Summarizing: The aim of my research is designing a data storage scheme which is suitable for *cbd* data in a general sense, through a modeling process that satisfies the four criteria of above. A note: The *scheme* here means that the design contains the storing process step by step. *In general sense* means that the scheme is useful for every kind of characteristic *cbd* data, i.e. pressure, flow etc. *Suitable* means that the scheme satisfies some criteria of quality (described at chapter 5).

For that purpose, 5 steps will be taken:

1st stage is Quantization process:

Step 1: Original data are transformed into wavelet domain by discrete wavelet transforms for obtaining sparse wavelet coefficients. The wavelet coefficient vector has same size, compared with the original data. (Discrete Wavelet Transform will be discussed in Chapter 3).

Step 2: The wavelet coefficients that are obtained from step 1 are not sparse enough, but most of them are close to zero. For reducing the number of non zero coefficients, a truncation process is done. (Some methods are discussed in Chapter 4).

2nd stage is encoding process:

Step 3: From the previous stage, we obtain a sparse vector, which has the same size as the original data. In this stage, a coding process for storing that sparse vector is done. Only the non-zero coefficients and their indices will be stored.

[1] Centrale Bediening Distributie

3rd stage is reconstruction process:

Step 4: The compressed data is converted into a vector form, to obtain the vector of the same size as obtained in step 2.

Step 5: The vector that is obtained in step 4 is converted by Inverse Discrete wavelet Transform, to obtain the approximation of the original data.

In this thesis, the main emphasis is on stage 1 and stage 3, the quantizer arrangement. We developed three scenarios.

Scenario 1: Maximize Quality of Approximation.

The focus of this scenario is the quality of the model's approximation. We would like to obtain a number of parameters that is smaller than the number of original measurements, but the main thing is to maintain the quality of approximation as high as possible. If all original data are stored, of course we retain 100% information. But maybe it is possible to store less numbers and to retain the high quality approximation.

In this scenario, reducing the numbers of parameters is done by throwing away coefficients that are related to the noise. This process, for recovering data from noise, is called de-noising. Soft thresholding and SURE thresholding techniques in [5] and [9] will be used in this process. These two techniques are called Hybrid thresholding. (These techniques will be discussed at Chapter 4).

The following diagram should make a scenario become clear:

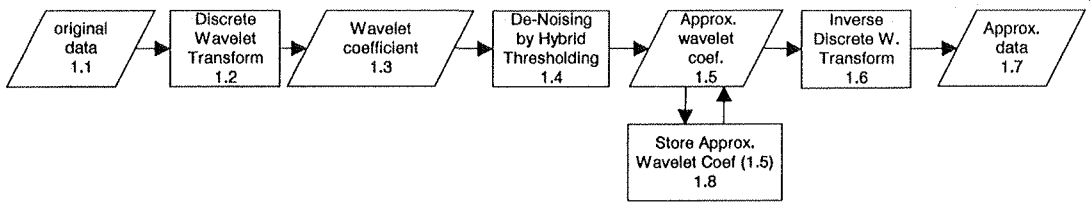


Figure 1 Scenario to maximize quality of approximation

The original data (1.1) are transformed into the same size wavelet coefficients vector (1.3) by discrete wavelet transform (1.2) with certain cutoff level and bases. There are some information losses in this process, and the quantity of the loss of information depends on the chosen cut-off level.

De-noising process (1.4) is applied to the wavelet coefficient vector. This can be done because discrete wavelet transform transforms the white noise from the original domain into white noise in the wavelet domain (This property will be discussed in page 42).

For obtaining the approximation of original data (1.7), inverse discrete wavelet transform is applied (1.6). We expect that quality of approximation is still high (because we only throw away white noise) and the number of parameters is less than the number of original measurements (because there was a truncating process involved).

Scenario 2: Minimize the Number of Non Zero Coefficient

The main focus of this scenario, is the reduction of data to lower the number of parameters as much as possible. In the extreme case, we can only store all the data in one parameter, but of course approximation quality is not guaranteed to be good.

In this scenario, parameter reduction is done by cutting non-dominant coefficients up to a certain level. The following diagram describe compression process in scenario 2 step by step:

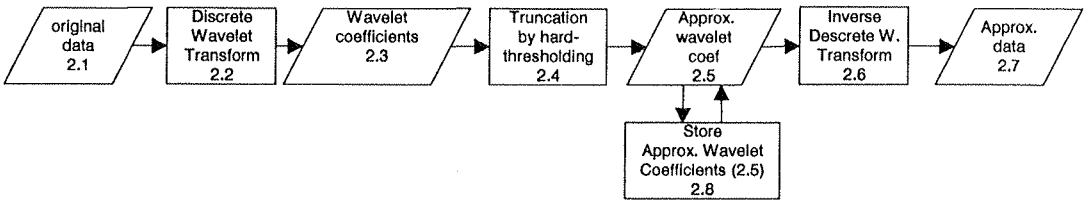


Figure 2 Scenario to minimized the number of non-zero parameters.

The original data (2.1) are transformed into wavelet coefficient (2.3) by Discrete Wavelet Transform (2.2) with certain cut-off level and basis. The truncation process is done in (2.4). The criterion, whether the wavelet coefficients are cut or not, does not depend on the data but depends on the desire of users, how many coefficients they want to keep. For example, if they want to keep only 1% of all coefficients then the 99^{th} *percentile* of coefficients is chosen as the cut-off level. This truncation method is called *hard thresholding* (discussed in Chapter 4). The approximation wavelet coefficients (2.5), which are produced by that process, are stored as a representative of the original data.

The reconstruction, from the approximation wavelet coefficients into original data, is done by Inverse Discrete Wavelet Transform (2.6). The result is not guaranteed give the quality as good as scenario 1, but under the specified parameter reduction this process gives the optimal quality. This is reasonable because we keep a certain number (could be less or more than that of (1.5)) of dominant parameters.

Scenario 3: Compromise.

This scenario tries to make a compromise between the number of parameters and the quality of approximation. We already discussed that scenario 1 puts the quality approximation as a priority and scenario 2 puts the number of parameters as a priority. In case the number of parameters of (1.5) is less than the numbers parameters in (2.5), we prefer to store (1.5) than (2.5) because we store smaller number parameters while the quality is still high. In case the number of parameters in (1.5) is large, for example more that 25% of total number of measurement, storing the parameter of (1.5) is not preferable. Storing all original data maybe better or reducing the number of parameters (by scenario 2) up to certain quality approximation.

The following diagram should make the scenario become clear:

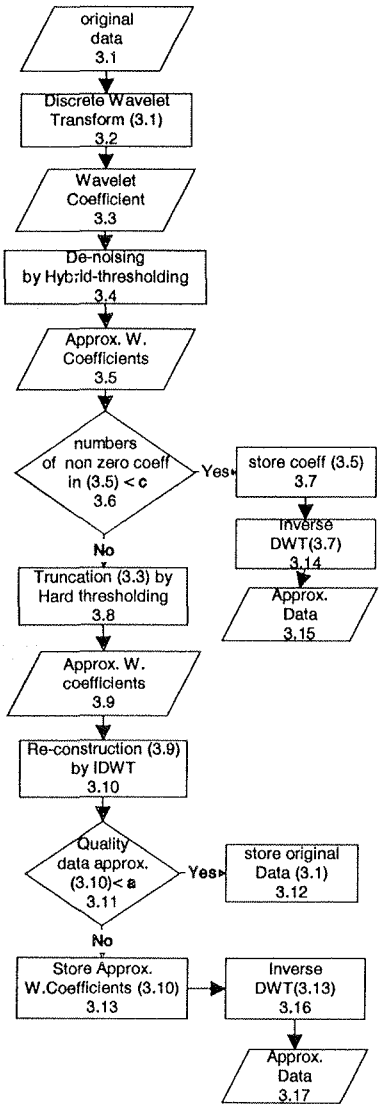


Figure 3 Scenario compromise

We here, suppose that the user decides that the lower bound of approximation quality is ***a*** and the upper bound of the numbers of non-zero coefficients is ***c***.

The first 3 steps are the same for scenario 1 and scenario 2. First we consider maximizing the quality of approximation, so de-noising process is done. After that, the number of parameters is compared to **c**. If the number of parameters is less than **c** then coefficients (3.5) are stored as a representative of original data, but if it is greater than **c**, we try to reduce the number of coefficients (3.3) to ***c*** by (3.8). The quality of approximation from that result is compared to ***a***. If it greater than ***a*** then (3.9) is stored as representative of original data, otherwise the original data (3.1) is stored.

It is reasonable to say that this compromise scenario is better than the two previous scenarios, because this scenario combines the advantages of those scenarios in one scenario.

Outline of thesis

The analysis in the thesis is based on the following outline:

- Chapter 1 gives some motivation and direction what will be done in my thesis.
- The information about the company and department, where this research project has been executed, are discussed in Chapter 2 briefly.
- The important wavelet transform background such as: what is wavelet transform, what is the difference and similarity between wavelet transform and Fourier transform, how to approximate a function with linear combination of wavelet basis, and the fast version algorithm for computing wavelet coefficients, will be discuss in chapter 3.
- Chapter 4 deals with some thresholding techniques that are used in this thesis.
- The modeling of time series data will discuss in chapter 5. Here, we use 6 data set for testing the model.
- The last chapter deals with a conclusion of my thesis and some comment about further research.