Wavelet Analysis of Magnetometer Data

Inga Maslova
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Mathematics Commons

Recommended Citation
WAVELET ANALYSIS OF MAGNETOMETER DATA

by

Inga Maslova

A report submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in Statistics

UTAH STATE UNIVERSITY
Logan, Utah
2005
ABSTRACT

Wavelet analysis of magnetometer data

by

Inga Maslova, Master of Science
Utah State University, 2005

Major Professor: Dr. Piotr Kokoszka
Department: Mathematics and Statistics

The wavelet analysis of the ground-based magnetograms' records is performed in this project. We explore the records from low, medium and high latitude stations during a calm period and a stormy one. Different methods for detecting and estimating the tail index of heavy-tailed distributions are compared. A detailed analysis of the properties of the distributions of the discrete wavelet transform coefficients of magnetometer data is presented. Conclusions on the tail index estimation techniques and the distribution of the discrete wavelet transform coefficients are made.

(124 pages)
Appendix

A. APPENDIX. DWT Plots. Summary Plots of the Wavelet Coefficients .......... 49

B. APPENDIX. Graphs for March 30 – April 2, 2001 .................................. 67

C. APPENDIX. Graphs for April 23 – April 25, 2001 ................................. 85

D. APPENDIX. Hill estimates of the tail index of the DWT coefficients .......... 103
LIST OF TABLES

2.1 Coordinates of the INTERMAGNET observatories ........................................ 3

3.1 Mean and standard deviation (in parentheses) of 1000 Hill estimates of the
tail index $\alpha$ of the stable distribution. .................................................. 30

3.2 Mean and standard deviation (in parentheses) of 1000 Hill estimates of the
tail index $\alpha$ of t-distribution. ................................................................. 30

3.3 Mean and standard deviation (in parentheses) of 1000 average Hill estimates
of the tail index $\alpha$ of the stable distribution. ........................................... 33

3.4 Mean and standard deviation (in parentheses) of 1000 average Hill estimates
of the tail index $\alpha$ of t-distribution. ........................................................ 33

4.1 The LSE estimates of $\hat{\rho}$ for the DWT coefficients fitting AR(1) to the DWT
coefficients of Honolulu Boulder, and College stations for March 29 – 31,
2001, and April 5 – 7, 2001 ........................................................................ 40

4.2 Hill estimates of the tail index, $\hat{\alpha}$ of the residuals of DWT coefficients $d_i$. .. 42

4.3 Average Hill estimates of the tail index, $\hat{\alpha}$ of the residuals of DWT coeffi-
cients $d_i$. ........................................................................................................ 42

4.4 Hill estimates of the tail index, $\hat{\alpha}$ of the residuals of DWT coefficients $d_i,$
using squaring technique. .............................................................................. 43
LIST OF FIGURES

2.1 Horizontal intensity in 0.1 nT, Honolulu, Boulder, College stations March 30 – April 2, 2001 ................................................................. 4

2.2 Horizontal intensity in 0.1 nT, Honolulu, Boulder, College stations, April 23 – April 25, 2001 ................................................................. 5

3.1 AR(1) stable time series with increasing values of \( a \): (from left to right) \( a = 0.5; a = 1; a = 1.25; a = 1.5; a = 1.75; a = 2 \) ............................................. 16

3.2 AR(1) series with \( t \) distribution of \( \varepsilon_t \): (from left to right) \( \nu = 2; \nu = 2.25; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.75; \nu = 4 \) .............................................. 17

3.3 ACF of AR(1) processes with residuals following (a) stable distribution with \( a = 1.5 \); (b) Gaussian distribution; (c) \( t \)-distribution with \( \nu = 3.5 \). Sample size \( n = 2160 \) ................................................................. 20

3.4 Variance plots for AR(1) series with stable \( \varepsilon_t \): (from left to right) \( a = 0.5; a = 1; a = 1.25; a = 1.5; a = 1.75; a = 2 \) ............................................. 21

3.5 Variance plots for AR(1) series with \( t \) distributed of \( \varepsilon_t \): (from left to right) \( \nu = 2; \nu = 2.25; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.75; \nu = 4 \) .............................................. 22

3.6 QQ-plots of (a) stable random variables with \( a = 1.5 \), (b) Gaussian random variables, (c) \( t_{3.5} \)-distributed random variables versus the standard normal distribution. ................................................................. 25

3.7 Hill-plots for \( \alpha \)-stable random variables with different tail indexes: (from left to right) \( a = 0.5; a = 1; a = 1.25; a = 1.5; a = 1.75; a = 2 \) ............................................. 28

3.8 Hill-plot of \( \nu \)-distributed random variables with different degrees of freedom: (from left to right) \( \nu = 2; \nu = 2.25; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.75; \nu = 4 \) .............................................. 29

3.9 Average Hill-plot (dashed) and Hill-plot (solid) of \( \alpha \)-stable random variables with different tail indexes: (from left to right) \( a = 0.5; a = 1; a = 1.25; a = 1.5; a = 1.75; a = 2 \). ............................................. 34

3.10 Average Hill-plot (dashed) and Hill-plot (solid) of \( \nu \)-distributed random variables with different degrees of freedom: (from left to right) \( \nu = 2; \nu = 2.25; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.75; \nu = 4 \) .............................................. 35

A.1 DWT transform of the Horizontal intensity at the Honolulu station, March 30 – April 2, 2001 ................................................................. 49

A.2 DWT of the Horizontal intensity at the Boulder station, March 30 – April 2, 2001 ................................................................. 50

A.3 DWT of the Horizontal intensity at the College station, March 30 – April 2, 2001 ................................................................. 51
A.4 DWT of the Horizontal intensity at the Honolulu station, April 23 – April 25, 2001 ................................................................. 52
A.5 DWT of the Horizontal intensity at the Boulder station, April 23 – April 25, 2001 ................................................................. 53
A.6 DWT of the Horizontal intensity at the College station, April 23 – April 25, 2001 ................................................................. 54
A.7 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001(from left to right) 55
A.8 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001(from left to right) 56
A.9 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001(from left to right) 57
A.10 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001(from left to right) 58
A.11 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001(from left to right) 59
A.12 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001(from left to right) 60
A.13 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001(from left to right) 61
A.14 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001(from left to right) 62
A.15 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, College station, March 30 – April 2, 2001(from left to right) 63
A.16 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, College station, March 30 – April 2, 2001(from left to right) 64
A.17 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, College station, March 30 – April 2, 2001(from left to right) 65
A.18 Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, College station, March 30 – April 2, 2001(from left to right) 66

B.1 Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001 ................................................................. 67
B.2 Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001 ................................................................. 68
B.3 Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001 ................................................................. 69
B.4 ACF of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001 ................................................................. 70
B.5 ACF of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001 ................................................................. 71
B.6 ACF of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001 ................................................................. 72
B.7 Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001 ................................................................. 73
C.12 QQ-plots of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001 .......................... 96

C.13 Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001 .......................... 97

C.14 Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001 .......................... 98

C.15 Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001 .......................... 99

C.16 Average Hill-plot (dashed) vs Hill-plot (solid) of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001 .......................... 100

C.17 Average Hill-plot (dashed) vs Hill-plot (solid) of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001 .......................... 101

C.18 Average Hill-plot (dashed) vs Hill-plot (solid) of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001 .......................... 102

D.1 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, March 30 – April 2, 2001 .......................... 103

D.2 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, March 30 – April 2, 2001 .......................... 104

D.3 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, March 30 – April 2, 2001 .......................... 105

D.4 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, March 30 – April 2, 2001 .......................... 106

D.5 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, March 30 – April 2, 2001 .......................... 107

D.6 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, March 30 – April 2, 2001 .......................... 108

D.7 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, April 23 – April 25, 2001 .......................... 109

D.8 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, April 23 – April 25, 2001 .......................... 110

D.9 Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, April 23 – April 25, 2001 .......................... 111

D.10 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, April 23 – April 25, 2001 .......................... 112

D.11 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, April 23 – April 25, 2001 .......................... 113

D.12 Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, April 23 – April 25, 2001 .......................... 114
1. INTRODUCTION

This thesis was accomplished as a part of initial analysis for the WAMI project (Wavelet Analysis of the Magnetosphere-Ionosphere), in which a wavelet-based statistical methodology for magnetogram data analysis is being developed.

The dynamic variations in a system of currents that flow in the magnetosphere-ionosphere are caused by various electrodynamic processes. Observing and monitoring these variations has been an important issue when trying to understand the electrodynamics in the magnetosphere and ionosphere. The magnetic field of the Earth is measured by ground-based magnetometers. Their records are used to produce the geomagnetic activity indices, such as $Ost$. Traditionally, based on the assumption that the magnetometers in the certain latitude are more sensitive to specific currents, an index that characterizes the variation of a certain current is based on the magnetometer data from a particular region. For example, the $Dst$ index is produced using the records from the equatorial regions that characterize the variations of the ring current.

The horizontal intensity of the magnetometer observations is used in the construction of $Dst$ index. That is why in this paper we explore the properties of the horizontal intensity obtained from three different latitude stations: Honolulu (low-latitude), Boulder (mid-latitude), and College (high-latitude) stations (see Table 2.1). One-minute data from the internationally-operated International Real-time Magnetic Observatory Network (INTERMAGNET) database is used (see http://www.intermagnet.org/).

In order to compare geomagnetically quiet and active events two sequences of the horizontal intensity are selected. The first one correspond to a stormy period, March 29 – April 2, 2001, and locates a storm that took place on March 31, 2001. The second time series selected for analysis represent a calm period, April 23 – April 25, 2001.
The main goal is to identify the class of distributions of the coefficients of the discrete wavelet transform of the horizontal intensity and to compare the results for various latitude stations as well as quiet versus stormy periods.

The current paper is organized in the following way: in Chapter 2 we introduce the general theory necessary for further discussions as well as the results of the preliminary analysis of the magnetometer data. Next, the basis of the wavelet analysis are presented in Section 2.2. Definitions and the basic properties of stable and Student’s distributions are established in Section 2.3.1 and Section 2.3.2, accordingly. In Section 2.4 heavy-tailed time series are introduced. Further, in Chapter 3 the aspects of practical implementation of the methods for detection of heavy-tailed distributions are examined. Different techniques of estimation of the tail index are compared in Section 3.5 and comments on their performance are provided. In addition to the theoretical validation of the tail index estimators the performance of these procedures is tested using the computer generated data. Finally, in Chapter 4 the methods described before are applied to the discrete wavelet transform of the real data and the conclusions are drawn in Chapter 5.
2. THEORETICAL BACKGROUND

2.1 Data description

Geomagnetic activity is extremely important not only for the space physics community, but for power companies, satellite operators, etc, since the storms that take place in the magnetosphere and ionosphere can cause breakdowns of electronic devices.

During the storm electric currents are flowing in various regions of the magnetosphere and ionosphere and change the magnetic field of the Earth.

A magnetic field is produced by moving electric charges. It is normally defined by the Lorentz-force equation, \( \mathbf{F} = q(\mathbf{v} \times \mathbf{B}) \), for the force on a particle of charge \( q \) with a velocity \( \mathbf{v} \), where \( \mathbf{B} \) is the vector magnetic field. In Figure 13.1, [8], page 403, the relationships of geomagnetic coordinate system to the earth are shown. The magnitude of the field projected in the horizontal plane is called \( H \), or horizontal component (for more detail see [8]). Since it is used to find \( Dst \) index, its statistical properties are discussed in this paper.

A global network of observatories, INTERMAGNET, monitors the magnetic field of the Earth, and provides 1-minute, 0.1 nT resolution data.

<table>
<thead>
<tr>
<th>Station</th>
<th>Geographic Latitude</th>
<th>Geographic Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honolulu</td>
<td>21.32</td>
<td>202.00</td>
</tr>
<tr>
<td>Boulder</td>
<td>40.14</td>
<td>254.76</td>
</tr>
<tr>
<td>College</td>
<td>64.87</td>
<td>212.14</td>
</tr>
</tbody>
</table>

Table 2.1 provides a list of INTERMAGNET stations considered in this thesis, i.e. low
Fig. 2.1: Horizontal intensity in 0.1 nT, Honolulu, Boulder, College stations March 30 – April 2, 2001

latitude (Honolulu), mid-latitude (Boulder), and high latitude (College). Three-day observations of the $H$ component during two periods are considered (Figure 2.1 and Figure 2.2). The first sequence consists from data observed in March 30 – April 2, 2001. It was chosen because of a strong storm that took place on March 31, 2001 (see [16]). In Figure 2.1 both the storm and its recovery phase are clearly seen, e.g. Honolulu station. The second period, chosen for analysis, April 23 – April 25, 2001 is considered to be relatively calm. The changes of the horizontal intensity during the quiet period, observed in Figure 2.2, are due to the daily variations.
The observations from various latitude stations are compared and the statistical properties of the discrete wavelet transform (DWT) of the horizontal intensity are explored. The basis of the wavelet analysis are presented in the following section.

2.2 Wavelet Analysis

This section introduces the main wavelet analysis techniques and properties necessary for the current paper.
Wavelets are mathematical tools used in time series, images analysis, signal processing, etc. The major emphasis of the thesis is made on a discrete wavelet transform (DWT) that is applied to magnetometer data.

The name "wavelet" suggests that we deal with a "small wave" that oscillates in a limited period of time. Due to the fact that wavelets are localized in time they are good building block functions for a variety of signals. Linear combinations of wavelet functions are used to represent signals. In order to stay consistent, the notations of S+Wavelets software are used in this thesis (see [3]).

2.2.1 The Wavelet Approximation

The orthogonal wavelet series approximation to a time signal \( X(t) \) is given by the following formula

\[
X(t) \approx \sum_{k} s_{j,k} \phi_{j,k}(t) + \sum_{k} d_{j,k} \psi_{j,k}(t) + \sum_{k} d_{j-1,k} \psi_{j-1,k}(t) + \ldots + \sum_{k} d_{1,k} \psi_{1,k}(t),
\]

where a number of multiresolution components (scales) is \( J \), \( k \) changes from 1 to the number of coefficients in the specified component. If \( N \) is divisible by \( 2^j \), then \( k_j = N/2^j \), \( j = 1, \ldots, J \).

The functions \( \phi_{j,k}(t) \) and \( \psi_{j,k}(t) \) are given by

\[
\phi_{j,k}(t) = 2^{-\frac{j}{2}} \phi \left( \frac{t - 2^j k}{2^j} \right)
\]

and

\[
\psi_{j,k}(t) = 2^{-\frac{j}{2}} \psi \left( \frac{t - 2^j k}{2^j} \right).
\]

Functions \( \phi(\cdot) \) and \( \psi(\cdot) \) are called father and mother wavelets accordingly, and defined in such way that

\[
\int_{-\infty}^{\infty} \phi(u) du = 1,
\]
Father wavelets (2.2) are good for representation of smooth and low-frequency parts of a signal, whereas mother wavelets (2.3) are good at representing the high-frequency parts.

In (2.1) the coefficients are given by the integrals

\[ s_{j,k} \approx \int \phi_{j,k}(u)X(u)du, \]
\[ d_{j,k} \approx \int \psi_{j,k}(u)X(u)du, \quad j = 1, \ldots, J. \]

Approximation (2.1) is an orthogonal decomposition since the basis functions \( \phi_{j,k}(t) \) and \( \psi_{j,k}(t) \) are by construction orthogonal:

\[ \int_{-\infty}^{\infty} \psi_{j,k}(t)\psi_{j',k'}(t)dt = \delta_{k,k'}, \]
\[ \int_{-\infty}^{\infty} \psi_{j,k}(t)\psi_{j,k'}(t)dt = 0, \]
\[ \int_{-\infty}^{\infty} \psi_{j,k}(t)\psi_{j',k'}(t)dt = \delta_{k,k'}\delta_{j,j'}. \]

where

\[ \delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \]

Families of father and mother wavelets provide the orthogonality of the wavelet approximation. On practice, there are many types of wavelets \( \psi(\cdot) \) and \( \phi(\cdot) \) used. The choice of an appropriate filter is discussed in Section 2.2.2. In this project the least asymmetric wavelets, that have a compact support, are used. A discrete wavelet transform with a least asymmetric filter LA(8) is applied to the magnetogram observations. For a more detailed discussion on orthogonality of the wavelet transform, see [11].

### 2.2.2 The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a basic tool needed for studying discrete time series via wavelets.
The DWT maps the discrete signal vector $X = (X_1, X_2, \ldots, X_N)'$ to a vector of $N = 2^J$ wavelet coefficients $W = (W_1, W_2, \cdots, W_N)'$. The DWT is an orthonormal transform that can be written in the following way:

$$W = \mathcal{W}X,$$

where $\mathcal{W}$ is an orthogonal $N \times N$ real-valued matrix defining the DWT and satisfying $\mathcal{W}^T \mathcal{W} = I_N$.

The vector $W$ consists of the coefficients $s_{j,k}$ and $d_{j,k}$, $j = 1, 2, \cdots, J$ of the wavelet approximation (2.1). The smooth coefficients $s_{j,k}$ represent the smooth behavior of the data at the scale $2^j$. Here, we focus on the detail coefficients $d_{j,k}$ that show the scale deviations from the smooth behavior.

When $N$ is divisible by $2^J$ the number of coefficients $d_{1,k}$ at the finest scale is $N/2$. The next scale contains $N/4$ wavelet coefficients $d_{2,k}$. Similarly, there are $N/2^J$ coefficients each for $d_{j,k}$ and $s_{j,k}$.

To compute the DWT a fast "pyramid" algorithm with complexity $O(n)$ is used (see [11], Section 4.4, 4.5, 4.6).

In this thesis a DWT with LA(8) filter is applied to the magnetogram records. Figure A.1 – Figure A.6 show the DWT plot of the magnetometer records for three stations, i.e. Boulder, Honolulu, and College for two periods of time. The top row shows the original signal recomputed from the wavelet coefficients by means of the inverse discrete wavelet transform (IDWT). That is why the label \texttt{idwt} is used in these plots. The wavelet coefficients, $d_{1,k}, \ldots, d_{J,k}$, are plotted in the following rows and $s_{J,k}$ in the bottom row. The coefficients are plotted as vertical lines extending from zero.

Further, as a part of the preliminary analysis of the DWT of the magnetometer records, refer to Figure A.7 – Figure A.18, where the series of four plots of wavelet coefficients $d_{j,k}$, $j = 1, \cdots, 4$ are shown. The stack plots present the coefficients $d_{j,k}$, and the reconstructed component signals $D_{j,k}$, $j = 1, \cdots, 4$. However, in this paper we do not work with the reconstructed signals. Notice that the wavelet coefficients show significant jumps, which
suggests heavy-tailed distribution of the latter (for further details see Section 3.1).

Next, the autocorrelation functions of $d_{j,k}$ are provided. The ACF's of all plots exhibit significant autocorrelations at lag 1. So, AR(1) model is used as an approximation (Section 3.2).

Quantile-Quantile plots (QQ-plot) of the wavelet coefficients versus the quantiles of a standard normal verify if the data comes from the normal distribution or a heavy-tailed. In case of magnetometer records the QQ-plot curves, which indicates a heavy-tailed distribution of the wavelet coefficients (Section 3.4).

Finally, the histogram and an estimate of the density function provide another method of visual analysis. This plot confirms that the distributions of the wavelet coefficients have longer tails than a normal distribution.

Thus, assume that the DWT coefficients of the magnetometer observations for both stormy and calm periods, for all three different latitude stations, can be approximated by AR(1) series with heavy-tailed innovations. In Chapter 3 the methods for detection the AR(1) series with heavy-tailed noise are discussed. Chapter 4 presents a detailed analysis of the wavelet coefficients of the magnetometer data.

Nevertheless, there are several practical problems one may face performing the wavelet analysis, such as the choice of a particular wavelet filter. As it is stated in [5] in choosing a filter two major issues should be considered. First of all short width wavelet filters ($L = 2, 4, 6$) can sometimes introduce artifacts, such as unrealistic blocks, shark's fins, etc. Second, if the wavelet filters have large $L$ it might better match the characteristic features. However, more boundary coefficients are introduced. So, as it was noted earlier, in this research LA(8) is chosen.

The other problem one faces applying the DWT are the so called boundary wavelets coefficients. The DWT uses the circular filtering that treats the time series $X(t)$ near the beginning or the end as a portion of a periodic sequence. Since the periodicity is a questionable assumption, the boundary coefficients should be dealt with caution. Normally, the boundary wavelet coefficients are excluded from further analysis. From [11], Table
137a, we get that for LA(8) filter $d_1$ has one boundary coefficient at the beginning of series, and two at the end; $d_2$ – three at the beginning, and two at the end; $d_3$ and $d_4$ – three boundary coefficients at the beginning, and three at the end. Hence, we remove them to eliminate their affect on further analysis.

2.3 Random Variables with heavy-tailed distributions

Before introducing the methods of detection of heavy-tailed distributions and tail index estimation (Chapter 3), and analyzing the magnetometer data (Chapter 4), define the distributions with heavy tails.

Definition 2.3.1: Independent identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$ come from a heavy-tailed distribution with tail index $\alpha$ if the following holds

\[ P(X > x) = L(x)x^{-\alpha}, \quad x > 0, \]

where $\alpha > 0$ and $L(x)$ is a slowly varying function satisfying

\[ \lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1. \]

It means that the tails of such distribution approach zero slower than in the case of the normal distribution. Different from the normal distribution, there is a large probability that a heavy-tailed random variable takes a value far from the center of the distribution.

The preliminary analysis of the wavelet coefficients (Section 2.2) showed that the distributions of the DWT coefficients belong to the class of symmetric heavy-tailed distributions. Hence, in the following part the stable and Student’s random variables are defined.

2.3.1 Stable Random Variables

There are several equivalent definitions of a stable distribution, see [15], Chapter 1, Section 1.1. Next, one of the possible definitions is given.
Definition 2.3.2: A real-valued random variable $X$ is said to have a stable distribution if for any $n \geq 2$ there are $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$a_n \sum_{i=1}^{n} X_i + b_n \overset{d}{=} X,$$

where $X_i$ are i.i.d. copies of $X$.

$X$ is said to be strictly stable if $b_n = 0$.

In other words, a random variable comes from a stable distribution, if a certain linear combination of its copies belongs to the same distribution.

The following definition, [15], specifies the characteristic function of a stable random variable and it is equivalent to Definition 2.3.2.

Definition 2.3.3: A random variable $X$ is said to have a stable distribution, written as $X \sim S_{\alpha}(\sigma, \beta, \mu)$, if there are parameters $0 < \alpha \leq 2$, $\sigma \geq 0$, $-1 \leq \beta \leq 1$, and $\mu$ real such that its characteristic function is:

$$E^{itX} = \begin{cases} \exp \left\{ -\sigma^\alpha |t|^\alpha (1 - i\beta (\text{sign}(t)) \tan(\pi\alpha / 2)) + i\mu t \right\}, & \text{if } \alpha \neq 1, \\ \exp \left\{ -\sigma |t|^\alpha (1 + i\beta^2 \frac{2}{\pi} (\text{sign}(t)) \ln(t) + i\mu t) \right\}, & \text{if } \alpha = 1. \end{cases}$$

Here

$$\text{sign} (t) = \begin{cases} -1, & \text{if } t < 0, \\ 0, & \text{if } t = 0, \\ 1, & \text{if } t > 0. \end{cases}$$

A univariate stable distribution is characterized by four parameters: the index of stability $\alpha$, the scale parameter $\sigma$, the skewness parameter $\beta$ and the shift parameter $\mu$. Since there is no closed form neither for the cumulative distribution nor for the density function, the characteristic function (2.9) is used to explore the properties of this distribution. For example, when $\alpha = 2$, the characteristic function (2.9) takes a simple form:

$$E^{itX} = \exp \{ -\sigma^2 t^2 + i\mu t \},$$

which is the characteristic function of a Gaussian distribution with mean $\mu$ and variance $2\sigma^2$. 
Next, denote some properties of the stable distributions, which are important for our further exploration. For more properties of stable distributions and their proofs refer to [15] (Chapter 1, Section 1.2) and [7].

Property 2.3.1: Let \( X \sim \mathcal{S}_\alpha(\sigma, \beta, \mu) \) with \( \alpha \neq 1 \). Then \( X \) is strictly stable if and only if \( \mu = 0 \).

Property 2.3.2: Let \( X \sim \mathcal{S}_\alpha(\sigma, \beta, \mu) \) with \( 0 < \alpha < 2 \). Then
\[
E|X|^p < \infty, \text{ for any } 0 < p < \alpha,
\]
\[
E|X|^p = \infty, \text{ for any } p \geq \alpha.
\]

For \( 0 < \alpha < 2 \) the \( \alpha \)-stable random variables have an infinite second moment, which means that the techniques valid for Gaussian case do not apply.

Property 2.3.3: \( X \sim \mathcal{S}_\alpha(\sigma, \beta, \mu) \) is symmetric if and only if \( \beta = 0 \) and \( \mu = 0 \).

If \( X \) is a symmetric \( \alpha \)-stable (sas) then by (2.9) its characteristic function is
\[
(2.10) \quad Ee^{-\sigma|t|^\alpha}, \quad \sigma > 0.
\]

The sas random variables are used to model infinite variance series. The following section deals with the heavy-tailed distributions, with a finite variance.

### 2.3.2 Student’s random variables

Definition 2.3.4: Let \( X_1, X_2, \ldots, X_n \) be a random sample from a \( \mathcal{N}(\mu, \sigma^2) \) distribution. The quantity \( \frac{X - \mu}{S/\sqrt{n}} \) has Student’s t distribution with \( n - 1 \) degrees of freedom. Equivalently, a random variable \( T \) has Student’s t distribution with \( \mu \) degrees of freedom, and we write \( T \sim t_\mu \) if it has pdf
\[
(2.11) \quad f_T(t) = \frac{\Gamma \left( \frac{\mu+1}{2} \right)}{\Gamma \left( \frac{\mu}{2} \right)} (\pi \mu)^{-1/2} (1 + t^2 / \mu)^{-\frac{\mu+1}{2}}, \quad -\infty < t < \infty.
\]
Note that t-distribution has finite variance, \( \text{Var}(T) = \frac{\mu}{\mu-2} \), for \( \mu > 2 \).

It is clear from the definition of t-distribution that it is close to the normal distribution, however, it has heavier tails, i.e.

\[
P(T > x) = \int_x^\infty f_T(t) dt \sim L(x)x^{-\mu},
\]

where \( L(x) \) is a slowly varying function, (2.7).

So the tail index \( \alpha \) for t-distribution is equal to degrees of freedom \( \mu \). Further, consider t-distributed random variables with \( \alpha \in (2,4] \) and, hence, finite variance.

### 2.4 Heavy-tailed Time Series

In this paper we show that most of the standard results of classical time series are applicable to the heavy-tailed case.

Classical time series deal with linear processes

\[
X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \quad t \in \mathbb{Z},
\]

where \( \psi_0 = 1 \) and noise variables \( \varepsilon_t \) are i.i.d. real-valued random variables with mean zero and finite variance \( \sigma^2 \). The process \( \{X_t\} \) is strictly stationary, i.e. the finite-dimensional distributions of the process are invariant under shifts of the time index.

Next, consider a strictly stationary linear process (2.13) where i.i.d. innovations \( \varepsilon_t \) belong to a heavy-tailed distribution. If the noise \( \varepsilon_t \) is t-distributed with \( 2 < \mu < 4 \), it has the finite variance, and methods of classical time series apply. So, assume that the noise \( \varepsilon_t \) belongs to stable distribution with \( \alpha \in (0,2) \). In other words, the decay of the tails of such distribution is slower than in the case of normal random variables. From Section 2.3.1 recall that \( \varepsilon_t \) has an infinite variance.

Furthermore, from the properties of stable distributions we get

\[
X_t \overset{d}{=} \left( \sum_{j=-\infty}^{\infty} |\psi_j|^\alpha \right)^{1/\alpha} \varepsilon_t, \text{ for each } t,
\]

which implies that \( X_t \) is a sas.
The following condition guarantees the a.s. existence of the series in (2.13),

\[ \sum_{j=-\infty}^{\infty} |\psi_j|^\alpha < \infty. \]

For further results and properties of heavy-tailed time series refer to [5].

Here we focus on the tail index \( \alpha \leq 4 \). As noted before, when \( 0 < \alpha < 2 \) we deal with the infinite variance distribution. For \( 2 \leq \alpha \leq 4 \) the distribution has a finite variance. As it appears in Chapter 4 the DWT coefficients of the magnetometer data belong to both types of distributions.
3. METHODS FOR DETECTING AND ESTIMATING HEAVY TAILS

Given a set of data to be analyzed, one usually starts with graphical data exploration.

The following section deals with different methods of visual analysis of heavy tailed time series. Preliminary research described in Section 2.2 indicates that the coefficients of the discrete wavelet transform (DWT) of the magnetometer observations can be approximated by an AR(1) model:

\[ X_t = \varphi X_{t-1} + \varepsilon_t, \]

where the errors \( \{ \varepsilon_t \} \) follow a distribution with heavy tails (see Section 2.3). Hence, the methods given below are designed either for AR(1) time series with heavy-tailed errors or for i.i.d. random variables with heavy-tailed distributions.

The following sections discuss methods of detection of heavy-tails in time series, estimation of the tail index \( \alpha \), and evaluate their performance using time series with known heavy-tailed distributions. First, we analyze graphs of heavy-tailed time series, then the autocorrelation functions of AR(1) series are presented. After we have discussed the methods of detecting AR(1) series with heavy tailed residuals, we focus on the "converging variance" test, which is used to determine if the data have finite variance. Further, another way of detection the heavy-tailed distribution using QQ-plots is given. Finally, the methods of estimating the tail index \( \alpha \) are presented.

All these methods are used to examine properties of the coefficients of the DWT of magnetometer records in Chapter 4.
3.1 Plots of Time Series With Heavy-Tailed Residuals

In Figure 3.1, plots of several $\alpha$-stable time series are presented. Each of them follows the AR(1) model

\[ X_t = -0.5X_{t-1} + \epsilon_t, \]

where $\epsilon_t$ are symmetric i.i.d. $\alpha$-stable random variables with different tail indexes $\alpha$. Distributions of such random variables, as it is described in Section 2.3.1, are bell-shaped but have longer tails than the normal distribution.

As explained in Section 2.3 heavy-tailed distributions are parameterized by the tail index $\alpha$. The smaller the $\alpha$, the "heavier" the tails. In case of $\alpha$-stable series, if $\alpha \in (0, 2)$, then $\text{Var}(X_t) = \infty$ and if $\alpha = 2$, then $\text{Var}(X_t) < \infty$. As can be seen from the characteristic function (2.10), the case $\alpha = 2$ corresponds to a Gaussian distribution (Figure 3.1, $\alpha = 2$).

However, in general $\alpha = 2$ does not imply the finite variance. Compare Figure 3.1, $\alpha = 2$ and Figure 3.2, $\nu = 2$, where the latter is a distribution with a heavy tail.

To investigate the properties of distributions with $\alpha > 2$ we use the Student's $t$-distribution. The tail index of the $t$-distribution is equal to the number of the degrees of freedom.
Fig. 3.2: AR(1) series with t distribution of $\varepsilon_i$: (from left to right) $\nu = 2; \nu = 2.25; \nu = 2.5;\nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.5; \nu = 3.75; \nu = 4$

(see Section 2.3.2).

Figure 3.2 shows AR(1) series defined in (3.1), where the errors $\varepsilon_i$ have a t-distribution with degrees of freedom $\nu \in [2, 4]$. When $\nu \in (1, 2)$ the plots are similar to stable series.

It is obvious that in the case of $\alpha$-stable processes we observe longer tails, i.e. higher jumps (Figure 3.1), than in case of t-distribution with $\nu \geq 2$ (Figure 3.2).

As a part of a preliminary data analysis, one can refer to the data plots. Compare the plots of AR(1) series with $\alpha$ - stable noise, Figure 3.1, $\alpha \in (0, 2)$, and t-distributed noise, Figure 3.2, to the Gaussian case, Figure 3.1, $\alpha = 2$. Presence of sudden big jumps suggests a heavy-tailed distribution of the data.
In the following section, the description of the autocorrelation function (ACF) of AR(1) time series is provided and the ACF graph is discussed.

### 3.2 ACF of AR(1) Time Series With Heavy-Tailed Residuals

In classical time series analysis one often considers a stationary linear process

\[(3.2)\]
\[X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t \in \mathbb{Z},\]

with i.i.d. real-valued noise variables \(\varepsilon_t\) which are normally distributed with mean zero and with square summable coefficients \(\psi_j; \sum_{j=0}^{\infty} \psi_j^2 < \infty\).

The autocorrelation function of \(\{X_t\}\) is defined as follows:

\[(3.3)\]
\[\rho(h) = \frac{\gamma(h)}{\gamma(0)},\]

where \(\gamma(\cdot)\) is the autocovariance function (ACVF) defined by

\[(3.4)\]
\[\gamma(h) = \text{cov}(X_{t+h}, X_t) = E(X_{t+h}X_t) = E(X_hX_0), \quad h \in \mathbb{Z}.\]

Natural estimators for \(\gamma(h)\) and \(\rho(h)\) are given by the sample autocovariance \(\tilde{\gamma}(h)\) and the sample autocorrelation \(\tilde{\rho}(h)\):

\[(3.5)\]
\[\tilde{\gamma}_n(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} X_t X_{t+|h|}, \quad h \in \mathbb{Z},\]

\[(3.6)\]
\[\tilde{\rho}_n(h) = \frac{\tilde{\gamma}_n(h)}{\tilde{\gamma}_n(0)} = \frac{\sum_{t=1}^{n-|h|} X_t X_{t+|h|}}{\sum_{t=1}^{n} X_t^2}, \quad h \in \mathbb{Z},\]

assuming that \(\tilde{\gamma}_n(h) = \tilde{\rho}_n(h) = 0\), for \(|h| \geq n\). For normal errors \(\varepsilon_t\), \(\tilde{\gamma}_n(h)\) and \(\tilde{\rho}_n(h)\) are consistent and asymptotically normal estimators (see [5], Theorem 7.3.1, page 381).

Now consider a process where the sequence \(\varepsilon_t\) is i.i.d. symmetric \(\alpha\)-stable (sas) noise, with \(\alpha < 2\), i.e. \(\text{Var}(\varepsilon_t) = \infty\). In this case the autocovariance (3.4) does not exist, nevertheless, the corresponding sample analogues, (3.5) and (3.6), are well defined.
If the coefficients $\psi_j$ satisfy the condition
\begin{equation}
\sum_{j=-\infty}^{\infty} |\psi_j|^\delta j < \infty,
\end{equation}
where $\delta > 0$ is a constant, such that $\delta = 1$, if $\alpha < 1$ and $\delta < \alpha$, if $\alpha \geq 1$, then
\begin{equation}
\rho(h) = \frac{\sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|}}{\sum_{j=-\infty}^{\infty} \psi_j^2}, \quad h \in \mathbb{Z}
\end{equation}
are finite numbers. Although, the quantity (3.8) cannot be interpreted as autocorrelations of the sas process $\{X_t\}$, the same notation will be used. It is known that the sample autocorrelations are consistent estimators of the quantities $\rho(h)$, just as in the classical case: the sample autocorrelation $\tilde{p}_n(h)$ converges in probability to the quantity $\rho(h)$ defined in (3.8) (see [5], Theorem 7.3.2, page 382).

So, in both infinite and finite variance cases we use sample autocorrelation as a consistent estimator.

Next, consider the AR(1) model with autoregressive parameter $\varphi$:
\begin{equation}
X_t = \varphi X_{t-1} + \varepsilon_t,
\end{equation}
where $\varepsilon_t$ are i.i.d. random variables following a heavy-tailed distribution with mean zero.

Equation (3.9) can be written in the following way:
\begin{align*}
X_t &= \varepsilon_t + \varphi X_{t-1} \\
&= \varepsilon_t + \varphi \varepsilon_{t-1} + \varphi^2 X_{t-2} \\
&= \varepsilon_t + \varphi \varepsilon_{t-1} + \varphi^2 \varepsilon_{t-2} + \ldots \\
&= \sum_{j=0}^{\infty} \varphi^j \varepsilon_{t-j},
\end{align*}
where $|\varphi| < 1$.

So, in case of the AR(1) model $\psi_j = \varphi^j$, and the sample autocorrelation function is a consistent estimator of
\begin{equation}
\rho(h) = \frac{\sum_{j=-\infty}^{\infty} \varphi^j \varphi^{j+|h|}}{\sum_{j=-\infty}^{\infty} \varphi^{2j}}
\end{equation}
Fig. 3.3: ACF of AR(1) processes with residuals following (a) stable distribution with $\alpha = 1.5$; (b) Gaussian distribution; (c) t-distribution with $\nu = 3.5$. Sample size $n = 2160$

in both infinite and finite variance cases.

Figure 3.3 presents ACF’s of AR(1) series defined in (3.9) with stable, Gaussian and Student’s errors. The dashed lines indicate the 95% confidence intervals for sample autocorrelations of the i.i.d. $\mathcal{N}(0, 1)$ random variables. For non-normal time series, especially those with heavy tails, these confidence intervals must be treated with caution. For example, in case of the AR(1) model, the autocorrelations at lags 1 and 2 are significantly different from zero, if compared to the dashed lines on the plots. These are however simulated AR(1) series. Note that the lag 1 autocorrelation is a very good estimator of the autoregressive parameter $\varphi$.

Analysis of the ACF of time series can be used as a powerful tool to determine if the observations are uncorrelated. Moreover, if one gets an ACF plot that looks similar to Figure 3.3, i.e. the correlation at lag 1 is the highest and for lags greater than 1 it is close to zero, then it is an AR(1) model.

This method is applied to DWT coefficients of magnetometer records in Section 4.5 in order to verify that they follow an AR(1) model.
3.3 The “Converging Variance” Test

The “converging variance” test is an informal visual data analysis technique used to determine whether data have infinite variance.

The main idea of the test consists of plotting sample variance $s_m^2$ of the first $m$ observations. If the data comes from the distribution with infinite variance, $s_m^2$ will show large jumps. Otherwise, it will converge to a finite value. Despite the fact that this test was originally designed for i.i.d. random variables, it also works well for dependent data, as long as the order of the observations is first randomized. ([1], page 137)

Figure 3.4 shows variance plots of AR(1) series, where the errors $\varepsilon_t$ have $\alpha$-stable distribution with $\alpha = 0.5, \alpha = 1, \alpha = 1.25, \alpha = 1.5, \alpha = 1.75, \alpha = 2$. As one can notice the smaller the $\alpha$, the bigger the jumps on the variance plots. When $\alpha \in (0, 2)$, the variance is infinite. A similar picture is produced for AR(1) process with $\varepsilon_t$ having t-distribution and degrees of freedom from the interval $(0, 2)$. So, the conclusion is that if the variance plot of an AR(1) series has significant jumps, the process has infinite variance, marginal distribution with heavy tails.
Fig. 3.5: Variance plots for AR(1) series with t distributed of $\epsilon_t$: (from left to right) $\nu = 2; \nu = 2.25; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.5; \nu = 3.75; \nu = 4$

In Figure 3.5 variance plots of AR(1) process with t-distributed $\epsilon_t$ with $2 \leq \nu \leq 4$ are presented. In this case we deal with finite variance data, with tail index $2 \leq \alpha \leq 4$.

However, one can notice that in case of sas noise with $\alpha = 1$ (Figure 3.1) the variance plot converges, which suggests the finite variance, which is not the case here. Plots based on different realizations do not show this “anomaly”.

The similarities between AR(1) series with sas with $\alpha = 1.75$ noise (Figure 3.1) and t-distributed noise with $\nu = 2.25$ and $\nu = 2.75$ (Figure 3.2) are due to the chance error and the fact that the tail index $\alpha$ is close to 2.

Variance plots work well for identifying extremely heavy-tailed series. When the tail
index approaches 2 and gets bigger the jumps on the variance plots are not significant, but still visible.

This method is used to find out if the DWT coefficients of the magnetometer records have infinite variance.

3.4 QQ-Plots

A graphical technique called the quantile-quantile plot (QQ-plot) is used to solve a wide variety of problems, such as

- Exploring the distribution data come from. If the observed data set comes from the sample of the reference distribution the plot should be roughly linear.

- Isolating the outliers, which are easily identified in these plots.

- Identifying a distribution with heavy tails. If the observed data come from distribution with tails heavier than the reference distribution, then the plot will curve down at the left, and / or up on the right.

In this thesis the distributions with heavy tails are of the main interest. So, we use the QQ-plots to verify whether the coefficients of the DWT of the magnetometer records follow the distribution with heavy tails.

In order to define QQ-plots we use a lemma stated on page 188 of [5], which for easy reference is stated here as Lemma 3.4.1. Its proof follows immediately from the definition of the uniform distribution.

Lemma 3.4.1: Let $X_1, \ldots, X_n$ be i.i.d. with density function $F$. Let $U_1, \ldots, U_n$ be i.i.d. random variables uniformly distributed on $(0, 1)$ and denote by $U_{n,n} < \ldots < U_{1,n}$ the corresponding order statistics. Then the following result hold:

(a) $F^{-1}(U_1) \overset{d}{=} X_1$. 

(b) For every $n \in \mathbb{N}$,

$$\left(X_{1,n}, \ldots, X_{n,n}\right) \overset{d}{=} (F^{-1}(U_{1,n}), \ldots, F^{-1}(U_{n,n})).$$

(c) The random variable $F(X_1)$ has a uniform distribution on $(0,1)$ if and only if $F$ is a continuous function.

Let $X_{n,n} \leq \ldots \leq X_{1,n}$ be an ordered sample, with the empirical density function $F_n$.

Using Lemma 3.4.1 we get that for a continuous $F$, the rv’s $U_i = F(X_i)$, for $i = 1, \ldots, n$, are i.i.d. uniform on $(0,1)$.

Hence,

$$E[F(X_{k,n})] = \frac{n - k + 1}{n + 1}, \quad k = 1, \ldots, n,$$

and

$$F_n(X_{k,n}) = \frac{n - k + 1}{n + 1},$$

where $F_n$ is the empirical density function.

The QQ-plot gives the points

$$\left(X_{k,n}, F^{-1}\left(\frac{n - k + 1}{n + 1}\right)\right), \quad k = 1, \ldots, n.$$

By Glivenko - Canteli theorem, which says that the empirical density converges almost surely to the theoretical one, the plots should be roughly linear (see [2], Theorem 20.6). This is illustrated in panel (b) of Figure 3.6 which shows QQ-plots obtained by simulating 2160 standard normal observations with $F$ being standard normal cdf. By contrast, panels (a) and (c) large deviations from the straight line are seen which are attributable to the heavy tails of the simulated distribution. The heavier the tails, the more curved the QQ-plot becomes.

So, we conclude that if the QQ-plot of some sequence curves then this data belongs to the class of heavy-tailed distributions.
3.5 Hill Estimator

In the previous section various methods of detection of heavy-tailed residuals in AR(1) series were described. In this part the Hill estimator of the tail index \( \alpha \) is introduced and its practical application is discussed.

Recall from Section 2.4, that i.i.d. random variables \( X_1, \ldots, X_n \) have the tail index \( \alpha \) if

\[
P(X > x) = L(x)x^{-\alpha}, \quad x > 0,
\]

where \( L(x) \) is a slowly varying function satisfying

\[
\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1.
\]

The Hill estimator of \( \alpha \) is based on the \( k \) upper order statistics \( X_{k,n} \leq \ldots \leq X_{1,n} \) and takes the form:

\[
\hat{\alpha}_{k,n} = \left( \frac{1}{k} \sum_{j=1}^{k} \ln X_{j,n} - \ln X_{k,n} \right)^{-1},
\]

where \( k = k(n) \to \infty \).
As it is stated in [5] the Hill estimator is very natural, since different asymptotically equivalent versions of $\hat{a}$ can be derived by various methods.

If $X$ is a random variable with some density function $F$ such that for $\alpha > 0$ equation (3.11) holds, provided that $x \geq 1$. Then it follows that the density function of $Y = \ln X$ is

$$P(Y > y) = e^{-\alpha y}, \quad y \geq 0.$$ 

Notice that $Y \sim \text{Exp}(\alpha)$ and the MLE of $\alpha$ is given by

$$\hat{\alpha} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_j \right)^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_{j,n} \right)^{-1}.$$

Hence,

$$P(X > x) = Cx^{-\alpha}, \quad x \geq u \geq 0,$$

with known $u$ and $C = u^\alpha$.

Then we obtain

$$\hat{\alpha} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln \left( \frac{X_{j,n}}{u} \right) \right)^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_{j,n} - \ln u \right)^{-1}.$$

Let

$$K = \text{card}\{i : X_{i,n} > u, i = 1, \ldots, n\}.$$

So the joint density of $k$ upper order statistics (Theorem 4.1.3, [5]) is

$$f_{x_{k,n}, \ldots, x_{1,n}}(x_k, \ldots, x_1) = \frac{n!}{(n-k)!} \left( 1 - Cx_k^{-\alpha} \right)^{n-k} C^k \alpha^k \prod_{i=1}^{k} x_i^{-(\alpha+1)} - u < x_k < \cdots < x_1.$$

And the conditional MLEs are

$$\hat{\alpha}_{k,n} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln \left( \frac{X_{j,n}}{X_{k,n}} \right) \right)^{-1} = \left( \frac{1}{n} \sum_{j=1}^{n} \ln X_{j,n} - \ln X_{k,n} \right)^{-1}.$$

Hence, as stated before, the Hill estimator has the same form as the MLE.

The following theorem summarizes the asymptotic properties of the Hill estimator.
Theorem 3.5.1: Assume that $X_i$ are i.i.d random variables with the marginal distribution $F$ such that for some $\alpha > 0$ equations (3.11) and (2.7) hold. If $\hat{\alpha}$ be the Hill estimator (3.12), then the following statements hold:

1. (Weak consistency, [10])
   
   If $k \to \infty$, $\frac{k}{n} \to 0$, as $n \to \infty$, then
   
   $\hat{\alpha}_{k,n} \overset{p}{\to} \alpha$.

2. (Strong consistency, [4])
   
   If, in addition, $\frac{k}{\ln n} \to \infty$ as $n \to \infty$, then
   
   $\hat{\alpha}_{k,n} \overset{a.s.}{\to} \alpha$.

3. (Asymptotic normality, [5])

   If conditions (3.16) and (3.11) on $k$ and $F$ are satisfied, then
   
   $\sqrt{k} (\hat{\alpha}_{k,n} - \alpha) \overset{d}{\to} \mathcal{N}(0, \alpha^2)$.

Note: Similar properties hold in case of dependent data (see [13]).

Theorem 3.5.1 says that generally for the Hill estimator the standard statistical properties hold. However, a certain set of conditions on $F$ and $k(n)$ is needed.

In order to find the estimate $\hat{\alpha}$ one has to choose appropriate $k$ value, which is the main issue of this section. The mathematical conditions for $k$ are the following:

(3.16) \hspace{1cm} k(n) \to \infty, \quad \frac{n}{k(n)} \to \infty, \quad \frac{n}{k(n)} \to \infty,

i.e. a sufficiently large number of order statistics should be used, however, this number should be asymptotically negligible relative to the sample size.

On practice, the Hill estimator is difficult to use due to its sensitivity to the choice of the number of the upper order statistics $k$. Unfortunately, there is no algorithm for selecting an
appropriate $k$. So, before estimating the tail index $\alpha$, we have to find a reasonable number of upper order statistics $k$ for a given length of series $n$.

First, Hill-plots that provide the tail index estimate for different $k$ values are used to find an approximate $k$ value that gives a reasonable estimate of the tail index for a heavy tailed distribution with $\alpha \in (0, 4]$. Hill-plots tend to have a noticeable horizontal stretch across different $k$ values. It is preferable that $k$ is chosen from this region. We start from experimenting on simulated data in order to get a feeling for what is going on. Since we want to apply this method to the coefficients of the DWT of the magnetometer records, we generate series of $\alpha$-stable random variable with $0.5 \leq \alpha \leq 2$, and $t$-distributed series with $2 \leq \nu \leq 4$, of length $n = 2160$, $n = 1080$, $n = 540$, and $n = 270$, which are equal to the number of coefficients of the DWT at different scales.

Figure 3.7 and Figure 3.8 show the Hill-plots of $\alpha$-stable and $t$-distribution with the length of series $n = 2160$ and $k \in [5, 400]$. Notice that for most distributions plotted here the Hill estimates of $\hat{\alpha}$ are close to the actual value of the tail index when $k = 100$. Similarly, analyzing the Hill-plots of the sequences of different distributions we get approximate $k$ value for the remaining $n$. So, when the length of series $n = 2160$, the corresponding
number of upper order statistics $k = 100$; for $n = 1080, k = 125$; for $n = 540, k = 50$; and finally, for $n = 270, k = 40$.

However, these values of $k$ are found by visual analysis of several realizations of random variables. The tail index estimate will vary slightly each time one generates new sequences of random variables.

We therefore check the accuracy of the Hill estimates of the tail index $\alpha$ using the $k$ found before. In Table 3.1 and Table 3.2 the mean and the standard deviation of 1000 Hill estimates $\hat{\alpha}$ of the tail index of stable and $t$-distributions, correspondingly, are given.

The estimates $\hat{\alpha}$ are closer to the true value for $\alpha$-stable distributions with small $\alpha$. When
Tab. 3.1: Mean and standard deviation (in parentheses) of 1000 Hill estimates of the tail index $\alpha$ of the stable distribution.

<table>
<thead>
<tr>
<th>Theoretical $\alpha$</th>
<th>Estimated $\hat{\alpha}_{k,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=2160, k=100</td>
</tr>
<tr>
<td>0.500</td>
<td>0.4997 (0.0494)</td>
</tr>
<tr>
<td>1.000</td>
<td>1.0142 (0.1035)</td>
</tr>
<tr>
<td>1.125</td>
<td>1.1599 (0.1132)</td>
</tr>
<tr>
<td>1.250</td>
<td>1.3264 (0.1352)</td>
</tr>
<tr>
<td>1.375</td>
<td>1.5052 (0.1548)</td>
</tr>
<tr>
<td>1.500</td>
<td>1.748 (0.1871)</td>
</tr>
<tr>
<td>1.625</td>
<td>2.0723 (0.226)</td>
</tr>
<tr>
<td>1.750</td>
<td>2.5831 (0.2977)</td>
</tr>
<tr>
<td>1.875</td>
<td>3.4435 (0.3922)</td>
</tr>
<tr>
<td>2.000</td>
<td>4.9465 (0.4360)</td>
</tr>
</tbody>
</table>

Tab. 3.2: Mean and standard deviation (in parentheses) of 1000 Hill estimates of the tail index $\alpha$ of t-distribution.

<table>
<thead>
<tr>
<th>Theoretical $\alpha$</th>
<th>Estimated $\hat{\alpha}_{k,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=2160, k=100</td>
</tr>
<tr>
<td>2.25</td>
<td>2.0805 (0.2131)</td>
</tr>
<tr>
<td>2.50</td>
<td>2.2497 (0.2171)</td>
</tr>
<tr>
<td>2.75</td>
<td>2.3956 (0.2371)</td>
</tr>
<tr>
<td>3.00</td>
<td>2.5203 (0.2367)</td>
</tr>
<tr>
<td>3.25</td>
<td>2.652 (0.2451)</td>
</tr>
<tr>
<td>3.50</td>
<td>2.7621 (0.2562)</td>
</tr>
<tr>
<td>3.75</td>
<td>2.8890 (0.2771)</td>
</tr>
<tr>
<td>4.00</td>
<td>2.9673 (0.2721)</td>
</tr>
</tbody>
</table>
the tail index increases, the Hill estimator overestimates the tail index in the $\alpha$-stable case. An estimated value of $\alpha$ will likely be above 2, if the true value of $\alpha$ is between 1.75 and 2. In case of t-distribution it gives the estimates the average of which is more than one standard deviation away from the actual value of the tail index $\alpha$. One of the possible solutions is transforming the data so that it would have heavier tail, which gives more accurate estimate of the tail index $\alpha$.

If i.i.d. random variables $X_1,\ldots,X_n$ belong to the class of heavy-tailed distributions then (3.11) holds, and

\[(3.17) \quad P(X^2 > x) = P(|X| > x^{\frac{1}{2}}) = 2L(x^{\frac{1}{2}})x^{-\frac{\alpha}{2}},\]

where function $L(x)$ is defined as in (2.7). It is easy to see from (2.7) that $L(x^{1/2})$ is also slowly varying. Hence, if the distribution of $\{X_t\}$ has the tail index $\alpha$, then the tail index of $\{X_t^2\}$ is $\frac{\alpha}{2}$.

As noted at the beginning of this section, the heavy-tailed distributions with $\alpha \in (0, 4]$ are considered in this paper. After transforming the data according to (3.17), we get new series with $\alpha \in (0, 2)$. In this case the method described above overestimates the tail index, but it is still close to the actual values (Table 3.1).

Nevertheless, Hill estimator is difficult to use in practise due to its dependence on the number of the upper order statistics. Another way to improve it is described in the following section.

3.5.1 **Smoothing the Hill Estimator**

As discussed above, to estimate the tail index $\alpha$, one uses the Hill-plot to search for a stable region. The purpose of this section is to present a smoothing technique that potentially helps to improve the estimate.

The smoothing method proposed in [14] is a simple averaging technique designed to help to minimize the sensitivity of the estimator to the number of the upper order statistics and reduce the high variability of the Hill-plot. It consists in the averaging the Hill
estimator values for different numbers of order statistics, i.e.

\[ \tilde{\alpha}_{k,n} = \frac{1}{(u - 1)k} \sum_{p=k+1}^{\mu} \hat{\alpha}_{p,n}, \]

where \( u > 1 \). As introduced in [14], on practice, we use \( u \in (n^{0.1}, n^{0.2}) \), where \( n \) is the sample size. Here, we used \( u = 4 \) for series of length \( n = 2160, n = 1080 \), and \( u = 3 \) for \( n = 540, n = 270 \).

Due to the fact that the range of the smoothed plot is reduced in comparison to the classical Hill-plot, the method is less sensitive to the choice of \( k \).

Further, find the optimal \( k \) visually analyzing the smoothed Hill-plots. Figure 3.9 and Figure 3.10 depict the classical Hill-plots with the average Hill-plot. It is obvious that the averaging reduces the variability of the Hill-plot, and it is easier to choose the number of upper order statistics \( k \). For the sample size \( n = 2160 \) and \( k = 100 \) the average Hill-estimator (3.18) is close to the actual value for most of the distributions (Figure 3.9, Figure 3.10). So, we choose the same \( k \) values as in the classical Hill-plot case.

To investigate the performance, we have drawn \( n = 2160, 1080, 540 \) and 270 i.i.d. random variables from \( \alpha \)-stable distribution with \( 0.5 \leq \alpha \leq 2 \) and t-distribution with degrees of freedom \( 2 \leq \nu \leq 4 \). The results of the average and standard deviation of 1000 average Hill estimated are summarized in Table 3.3 and Table 3.4.

Comparing these results to the ones we got using the classical Hill estimator (Table 3.1, Table 3.2) we see that the mean and the standard deviation of the average Hill estimates got smaller for both \( \alpha \)-stable and t-distributed random variables. Notice that the average Hill estimator clearly underestimates the tail index of t-distributed random variables. One should keep it in mind when applying this method to the real data.

So, as it was stated before the averaging method really helps to reduce the variability of the Hill-plot and stabilize its behavior. However, the practical methods of estimating the tail index \( \alpha \) need further research.
Tab. 3.3: Mean and standard deviation (in parentheses) of 1000 average Hill estimates of the tail index $\alpha$ of the stable distribution.

<table>
<thead>
<tr>
<th>Theoretical $\alpha$</th>
<th>Estimated $\hat{\alpha}_{k,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n=2160, k=100$</td>
</tr>
<tr>
<td>0.500</td>
<td>0.4903 (0.0366)</td>
</tr>
<tr>
<td>1.000</td>
<td>1.0048 (0.0803)</td>
</tr>
<tr>
<td>1.125</td>
<td>1.1456 (0.0923)</td>
</tr>
<tr>
<td>1.250</td>
<td>1.3179 (0.1073)</td>
</tr>
<tr>
<td>1.375</td>
<td>1.5155 (0.1235)</td>
</tr>
<tr>
<td>1.500</td>
<td>1.7666 (0.1471)</td>
</tr>
<tr>
<td>1.625</td>
<td>2.0923 (0.181)</td>
</tr>
<tr>
<td>1.750</td>
<td>2.5487 (0.2199)</td>
</tr>
<tr>
<td>1.875</td>
<td>3.2045 (0.2593)</td>
</tr>
<tr>
<td>2.000</td>
<td>4.1951 (0.2696)</td>
</tr>
</tbody>
</table>

Tab. 3.4: Mean and standard deviation (in parentheses) of 1000 average Hill estimates of the tail index $\alpha$ of t-distribution.

<table>
<thead>
<tr>
<th>Theoretical $\alpha$</th>
<th>Estimated $\hat{\alpha}_{k,n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n=2160, k=100$</td>
</tr>
<tr>
<td>2.25</td>
<td>1.9615 (0.1462)</td>
</tr>
<tr>
<td>2.5</td>
<td>2.1199 (0.1517)</td>
</tr>
<tr>
<td>2.75</td>
<td>2.2325 (0.1632)</td>
</tr>
<tr>
<td>3</td>
<td>2.3454 (0.1655)</td>
</tr>
<tr>
<td>3.25</td>
<td>2.4542 (0.1735)</td>
</tr>
<tr>
<td>3.5</td>
<td>2.5468 (0.1773)</td>
</tr>
<tr>
<td>3.75</td>
<td>2.6315 (0.1807)</td>
</tr>
<tr>
<td>4</td>
<td>2.7044 (0.1891)</td>
</tr>
</tbody>
</table>
Fig. 3.9: Average Hill-plot (dashed) and Hill-plot (solid) of α-stable random variables with different tail indexes: (from left to right) $\alpha = 0.5; \alpha = 1; \alpha = 1.25; \alpha = 1.5; \alpha = 1.75; \alpha = 2.$
Fig. 3.10: Average Hill-plot (dashed) and Hill-plot (solid) of t-distributed random variables with different degrees of freedom: (from left to right) $\nu = 2; \nu = 2.25; \nu = 2.5; \nu = 2.75; \nu = 3; \nu = 3.25; \nu = 3.5; \nu = 3.75; \nu = 4.$
4. ANALYSIS OF THE MAGNETOMETER RECORDS

This chapter presents the analysis of the magnetometer records. Three different latitude stations are under consideration: Honolulu (low latitude), Boulder (medium latitude), and College (high latitude). Two sets of three-day observations, \( n = 4320 \), are taken, i.e. half of March 30 – half of April 2, 2001 and April 23 – 25, 2001. The first sequence was selected to locate the strong storm that took place on March 31, 2001. The period in April is considered to be relatively calm. So here we compare the stormy period with a calm one, and analyze the differences between high, medium and low latitude stations.

As it has already been mentioned in Chapter 2, a DWT with \( J=4 \), using LA(8) filter is applied to both data sets. After the preliminary analysis presented in Chapter 2, we assume that the DWT coefficients belong to the class of heavy-tailed distributions. This part provides verification of this fact and estimation of the tail index \( \alpha \). The methods described in Chapter 3 are applied to twenty four time series, which are the DWT coefficients of the horizontal intensity of three stations during a stormy and a calm period. We start from the visual analysis of the DWT coefficients.

4.1 Visual Analysis of the Plots of the DWT Coefficients of the Magnetometer Records

Perform, the individual analysis of the DWT coefficients at different frequencies. First, we analyze the plots to find if there are any big jumps that indicate a heavy-tailed distribution. All plots are given in the sets of four, corresponding four levels of the DWT, \( d_1, d_2, d_3 \) and \( d_4 \). As it was mentioned in Section 3.1 if a graph shows big jumps that denotes the heavy-tailed distribution of the data.

The DWT coefficients of a stormy period, March 30 – April 2, from three stations: Hono-
lulu (Figure B.1), Boulder (Figure B.2), and College (Figure B.3) show big jumps, which suggest a heavy-tailed distribution of all series. The spikes on these plots indicate a strong storm that took place on March 31, 2001. The plots of the wavelet coefficient of a quiet period, April 23– April 25, 2001, also present behavior typical for heavy-tailed data (see Figure C.2 – Figure C.3). To avoid an impression that there is more activity during this period note that the range is smaller than in case of the “stormy” data.  

The visual analysis has verified that the distributions of the DWT coefficients of the horizontal intensity of stormy and calm periods of Honolulu, Boulder, College stations belong to the class of heavy-tailed distributions.

Next, the ACF are used to find out if the wavelet coefficients are correlated.

4.2 ACF of the DWT Coefficients

The sample ACF’s are used to identify if the wavelet coefficients are correlated. Figure B.5 – Figure B.6 provide the ACF’s of DWT coefficients for three stations of the stormy period. The autocorrelations are the highest at lag 1 in most cases. However, in Figure B.6 the autocorrelations for bigger lags do not fall into the 95% confidence interval plotted with dashed lines. Recall from Section 3.2 that the dashed lines show the 95% confidence intervals for normal random variables. Here, the heavy-tailed series are under consideration, so the plotted confidence intervals should not be referred to. Note that AR(1) model is chosen as a first approximation. Since the goal of the thesis is to identify heavy-tailed distribution and to estimate the tail index the choice of the model is not that critical.

Similar results are obtained for DWT coefficients of the calm period. Their ACF’s (Figure C.5 – Figure C.6) show the highest jumps at lag 1. Hence, conclude that the AR(1) model with heavy-tailed errors can be fitted to the series.

The ACF’s of the wavelet coefficients of both periods for three stations suggest that we deal with correlated data. Since the autocorrelation coefficient at lag 1 is the biggest we conclude that the autoregressive model, AR(1), with heavy-tailed residuals can be fitted to
4.3 The "Converging Variance" Test of DWT Coefficients

The "converging variance" test is used to identify whether wavelet coefficients have finite or infinite variances. Compare the variance plots of the wavelet coefficients to the variance plots of the known distributions introduced in Section 3.3 (Figure 3.4, Figure 3.5).

Each set of plots discussed further consists of four variance plots that correspond to the four levels, \( d_i, i = 1 \ldots 4 \), of the DWT. In Figure B.7 the variance plots of wavelet coefficients of the horizontal intensity recorded on March 30 – April 2 at Honolulu station are given. In case of \( d_1 \) the jumps of the variance are observed, however, in general it seems to converge to some value, which denotes either the finite variance of the distribution of the data or infinite variance distribution with \( \alpha \in (1.75, 2) \). The variance plots of \( d_2, d_3 \) and \( d_4 \) show bigger jumps than in \( d_1 \) case, which means that data might have an infinite variance.

During the same period at Boulder station only sequence \( d_2 \) converges to a finite value (Figure B.8). For the other three cases, \( d_1, d_3, d_4 \), one observes big jumps which indicates an infinite variance distribution.

The DWT coefficients of College station records, Figure B.9, suggest that all series come from an infinite variance distributions.

So, in case of a stormy day the wavelet coefficients in some cases belong to a finite variance distribution when \( j = 1 \) or \( j = 2 \). However, in most cases they come from an infinite variance distribution. When \( j = 4 \) the variance plots clearly show that the coefficients have infinite variance. Note that the range in all cases is big, that fact is explained by the storm that took place on March 31, 2001.

The range of the variance plots of the wavelet coefficients of the data recorded during the calm period, April 23 – April 25, 2001, is the smaller for Honolulu, Boulder than for College station records. Honolulu station records show that \( \text{Var}(d_{j,k}) < \infty \) for \( j = 1, 2, \ldots, 4 \).
with \( \alpha \in (2,3) \). When \( j = 3,4 \) the observed jumps of the variance plot denote an infinite variance distribution (see Figure C.7).

In case of Boulder station, Figure C.8, the first two levels seem to have a finite variance. The variance plots of \( d_{3,k} \) and \( d_{4,k} \) denote that the tail index of these distributions is around 2, and the variance is infinite.

The variance plots of the wavelet coefficients of the College station, Figure C.9, clearly show that \( \text{Var}(d_{1,k}) < \infty \). For \( d_{2,k} \) we conclude that the distribution has finite variance, and the tail index close to 2, that is why the jumps are observed. The other two sequences, \( d_{3,k} \) and \( d_{4,k} \) present big jumps and no convergence can be observed, that suggests the infinite variance distributions.

Conclude that the coefficients of the DWT belong to a heavy-tailed distribution either with finite variance and the tail index \( \alpha \in (2,3) \), or with infinite variance, with \( \alpha \in (1.5,2) \). The tail index gets smaller for lower frequency wavelet coefficients.

These results are used in Section 4.6. We refer to the variance plots to see if the tail index estimate is consistent with the results presented in this section.

### 4.4 Fitting the AR(1) model

As it was shown in the previous sections, the AR(1) model can be used as an approximation of the wavelet coefficients of the horizontal intensity at different stations. Since methods presented in Section 3.5 are designed for i.i.d. series, the AR(1) model is fitted to the data and the residuals are estimated.

Consider an AR(1) model with one parameter \( \varphi \):

\[
X_t = \varphi X_{t-1} + \varepsilon_t,
\]

where \( \varepsilon_t \) are i.i.d. with mean zero and variance \( \sigma^2 \).

For \( |\varphi| < 1 \) the Ordinary Least Square estimator (OLS) of the autoregressive parameter
is

\[ \phi_n = \frac{\sum_{j=2}^{n} X_{j-1}X_j}{\sum_{j=2}^{n} X_{j-1}^2}. \]  

(4.2)

Table 4.1 provides the estimates \( \phi_n \) of the wavelet coefficients of Honolulu, Boulder, and College stations for stormy and quiet periods.

Tab. 4.1: The LSE estimates of \( \hat{\phi} \) for the DWT coefficients fitting AR(1) to the DWT coefficients of Honolulu Boulder, and College stations for March 29 – 31, 2001, and April 5 – 7, 2001

<table>
<thead>
<tr>
<th>Least Square Estimate of ( \phi_n )</th>
<th>Discrete wavelet transform coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( d_1 )</td>
</tr>
<tr>
<td>March 30 – April 2, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
<tr>
<td>April 23 – 25, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
</tbody>
</table>

Next the estimates of the residuals are found using the following equation,

\[ \hat{\epsilon}_t = X_t - \hat{\phi}_nX_{t-1}, \]  

(4.3)

where \( \hat{\phi}_n \) is an estimate defined by (4.2).

Applying the same technique, discussed in Section 3.1, notice that the residuals of the wavelet coefficients belong to the class of heavy-tailed distributions. In the next section the QQ-plots are used to verify heavy-tailed distribution of the residuals of the DWT coefficients.

4.5 QQ-Plots of the Residuals of the DWT Coefficients

Up to now, we have checked the validity of AR(1) model by means of visual tools. As it is stated in [12], page 62, for larger sample sizes, visual diagnostic tools can be preferable
to goodness-to-fit-tests. In this section we verify that the distribution of the residuals are heavy-tailed, using the technique described in Section 3.4.

The QQ-plots of the residuals of wavelet coefficients are presented in Figure B.4 – Figure B.6, for March 30 – April 2, and in Figure C.10 – Figure C.12, for April 23 – April 25. Here, the empirical distributions of the wavelet coefficients are compared to the standard normal distribution. As described in Section 3.4, if the points in the QQ-plot follow the line, one can infer the normal distribution. On the contrary, if they curve – the distribution is heavy-tailed. Here, the QQ-plots curve for all data sets, which denotes that the distributions are not normal, rather heavy-tailed. Hence, we verified that AR(1) series, where $\varepsilon_i$ are heavy-tailed are under consideration.

Further, the tail indexes of the residuals are estimated applying the Hill estimator introduced in Section 3.5.

### 4.6 The Hill Estimates of The Tail Index

Finally, after investigating the properties of the DWT coefficients, we can move to the main task – tail index estimation.

As it was discussed in Section 3.5 the classical Hill estimator is extremely sensitive to the choice of the number of upper order statistics $k$. We found which $k$ gives reasonable tail index estimates for different sample sizes $n$. These $k$ values are used in this section to estimate the tail index of wavelet coefficients. In Chapter 3 the classical Hill estimator and the methods of its improvement are presented. Simulating the series with $\alpha$-stable and t-distributions we showed that the Hill estimates are closer to the nominal value for smaller values of $\alpha$. The squaring method presented in Section 3.5 overestimates the tail index. The smoothing method, given in the Subsection 3.5.1, reduces the variability of the Hill plot, however, it underestimates the $\alpha$.

In Table 4.2 the classical Hill estimates of the wavelet coefficients are given. The average Hill estimates are presented in the Table 4.3. Finally, the Hill estimates found using the
The difference of these methods discussed before can be traced here as well. To illustrate this fact refer to Figure D.1 - Figure D.3 and Figure D.7 - Figure D.9, where tail index estimates found using three methods, described above, are given. The numbers on the x-axis correspond to the wavelet coefficients, \( d_{j,k}, j = 1, \ldots, 4 \). Notice that the average Hill estimator provides the smallest estimates in all cases. The so called "squared" Hill estimator is the Hill estimator found using the squaring method (see Section 3.5). Due to

### Tab. 4.2: Hill estimates of the tail index, \( \hat{\alpha} \) of the residuals of DWT coefficients \( d_i \).

<table>
<thead>
<tr>
<th>Hill estimates</th>
<th>Residuals of DWT coefficients, ( d_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>( d_1 )</td>
</tr>
<tr>
<td>March 30 – April 2, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
<tr>
<td>April 23 – 25, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
</tbody>
</table>

### Tab. 4.3: Average Hill estimates of the tail index, \( \hat{\alpha} \) of the residuals of DWT coefficients \( d_i \).

<table>
<thead>
<tr>
<th>Average Hill estimates</th>
<th>Residuals of DWT coefficients, ( d_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>( d_1 )</td>
</tr>
<tr>
<td>March 30 – 29, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
<tr>
<td>April 23 – 25, 2001</td>
<td>Honolulu</td>
</tr>
<tr>
<td></td>
<td>Boulder</td>
</tr>
<tr>
<td></td>
<td>College</td>
</tr>
</tbody>
</table>
Tab. 4.4: Hill estimates of the tail index, \( \hat{\alpha} \) of the residuals of DWT coefficients \( d_j \), using squaring technique.

<table>
<thead>
<tr>
<th>“Squared” Hill estimates</th>
<th>Residuals of DWT coefficients, ( d_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\alpha} )</td>
<td>( d_1 )</td>
</tr>
<tr>
<td>March 30 – April 2, 2001</td>
<td>Honolulu: 1.969567</td>
</tr>
<tr>
<td></td>
<td>Boulder: 2.004209</td>
</tr>
<tr>
<td></td>
<td>College: 1.961554</td>
</tr>
<tr>
<td>April 23 – 25, 2001</td>
<td>Honolulu: 3.2284</td>
</tr>
<tr>
<td></td>
<td>Boulder: 3.0932</td>
</tr>
<tr>
<td></td>
<td>College: 2.1753</td>
</tr>
</tbody>
</table>

the fact that it overestimates the \( \alpha \), here, the highest tail index estimates for \( d_j, j = 1, 2, 3, 4 \) are obtained. These results are considered to be the most accurate ones.

However, these are the point estimates. In Figure D.4 – Figure D.6 and Figure D.10 – Figure D.12 the Hill estimates and their asymptotic 95% confidence intervals (see Theorem 3.5.1) are presented. Since the confidence intervals of different techniques overlap, none of the methods can be singled out as the best. Hence, further research in this area is needed.
5. CONCLUSIONS AND FUTURE WORK

The goal of this thesis was to explore and implement the methods for detecting the heavy-tailed distributions and estimating the tail index of the wavelet coefficients of the magnetometer data.

The preliminary analysis of the wavelet coefficients of the magnetometer records suggests a heavy-tailed distribution of the latter. In Chapter 3 these assumptions are validated and the practical aspects of the techniques of visual analysis are presented.

Heavy-tailed distributions are parameterized by the tail index $\alpha$. The Hill estimator is a consistent estimator of $\alpha$ for which the standard statistical properties hold. However, its implementation is complicated due to its dependence on the choice of the number of upper order statistics $k$. In addition, none of the known methods provide exact procedure for the choice of $k$.

First, the classical Hill estimator was applied to the $sas$ and $t$-distributed i.i.d. random variables. As a result, for a $sas$ distribution the Hill estimates of $\alpha$ are close to the nominal value for $\alpha \in (0, 1.5)$. However, the Hill estimator overestimates the true value for fixed $k$ and it gets worse for $1.5 < \alpha < 2$. When the tail index $2 < \alpha < 4$, the Hill estimator underestimates the nominal value of the tail index. This fact was illustrated using the replications of $t$-distributed random variables.

To improve the performance of the estimator we propose a squaring transformation of the original data. It reduces the tail index in half, thus $0 < \alpha < 2$. The results of the experiments demonstrated that this method slightly overestimates the tail index.

The last technique explored in this work is the average Hill estimator. This method is designed to reduce the variability of the Hill-plots and Hill estimator sensitivity on the choice of the upper order statistics. Since it consists of a simple averaging of the classical
Hill estimates for various $k$ it underestimates the tail index keeping a pattern similar to the classical Hill estimator.

The methods of tail index estimation described in the thesis give the estimates that are close to the real value of the tail index. However, further investigations are needed. There are several papers that suggest ways of improvement of the classical Hill-plots and Hill estimator.

As a part of further research, the performance of the alternative Hill plot suggested by S. Resnick and C. Stârică in [14], has to be checked. Another method of the tail index estimation that uses the slope of the QQ-plot is introduced in [9]. The QQ-estimator of the tail index is presented, however, its performance should be analyzed using the simulated data.

Applying the techniques explored in the paper to the DWT of the magnetometer records the following results were obtained: in case of both stormy and calm periods the wavelet coefficients of the Horizontal intensity of all three stations can be approximated by AR(1) model with heavy-tailed errors. However, one must consider the “pitfalls of the fitting autoregressive models for heavy-tailed time series” discussed in [6].

The wavelet coefficients of quiet and stormy periods belong to the heavy-tailed distribution class with different $\alpha$. However, the distributions of $d_{jk}$ have heavier tails, i.e. smaller tail index $\alpha$, for bigger $j$.

In general the major methods for detecting heavy-tailed distribution are based on empirical analysis of the data and have little theory behind. In fact, the estimates of the tail index of the wavelet coefficients fall into the interval $(1, 3)$, which is consistent with the results of the “converging variance” test performed in Section 4.3. However, it is a “boundary” case that must be dealt with caution. As a result, an algorithm for tail index estimation that would be easy to implement should be developed.


Fig. A.1: DWT transform of the Horizontal intensity at the Honolulu station, March 30 – April 2, 2001
Fig. A.2: DWT of the Horizontal intensity at the Boulder station, March 30 – April 2, 2001
Fig. A.3: DWT of the Horizontal intensity at the College station, March 30 – April 2, 2001
Fig. A.4: DWT of the Horizontal intensity at the Honolulu station, April 23 – April 25, 2001
Fig. A.5: DWT of the Horizontal intensity at the Boulder station, April 23 – April 25, 2001
Fig. A.6: DWT of the Horizontal intensity at the College station, April 23 – April 25, 2001
Fig. A.7: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001 (from left to right)
Fig. A.8: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001 (from left to right)
Fig. A.9: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001 (from left to right)
Fig. A.10: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, Honolulu station, March 30 – April 2, 2001 (from left to right)
Fig. A.11: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001 (from left to right)
Fig. A.12: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001 (from left to right)
Fig. A.13: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001 (from left to right)
Fig. A.14: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, Boulder station, March 30 – April 2, 2001 (from left to right)
Fig. A.15: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_1$ wavelet coefficient vector, College station, March 30 – April 2, 2001(from left to right)
Fig. A.16: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_2$ wavelet coefficient vector, College station, March 30 – April 2, 2001 (from left to right)
Fig. A.17: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_3$ wavelet coefficient vector, College station, March 30 – April 2, 2001 (from left to right)
Fig. A.18: Stack plot; ACF plot; QQ-plot; Histogram and density plot for the $d_4$ wavelet coefficient vector, College station, March 30 – April 2, 2001 (from left to right)
B. APPENDIX. GRAPHS FOR MARCH 30 – APRIL 2, 2001

Fig. B.1: Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.2: Plots of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.3: Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001
Fig. B.4: ACF of the DWT coefficients $d_j, j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.5: ACF of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.6: ACF of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001
Fig. B.7: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.8: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.9: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 - April 2, 2001
Fig. B.10: QQ-plots of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.11: QQ-plots of the residuals of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.12: QQ-plots of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001
Fig. B.13: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.14: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.15: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001
Fig. B.16: Average Hill-plot (dashed) and Hill-plot (solid) of the residuals of the DWT coefficients $d_j, j = 1, \ldots, 4$, Honolulu station, March 30 – April 2, 2001
Fig. B.17: Average Hill-plot (dashed) and Hill-plot (solid) of the residuals of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, March 30 – April 2, 2001
Fig. B.18: Average Hill-plot (dashed) and Hill-plot (solid) of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, March 30 – April 2, 2001
Fig. C.1: Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.2: Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001
Fig. C.3: Plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001
Fig. C.4: ACF of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.5: ACF of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001
Fig. C.6: ACF of the DWT coefficients $d_j, j = 1, \ldots, 4$, College station, April 23 – April 25, 2001
Fig. C.7: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.8: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, April 23 - April 25, 2001
Fig. C.9: Variance plots of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001
Fig. C.10: QQ-plots of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.11: QQ-plots of the residuals of the DWT coefficients $d_j, j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001
Fig. C.12: QQ-plots of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001
Fig. C.13: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.14: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, Boulder station, April 23 – April 25, 2001
Fig. C.15: Hill-plot of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001.
Fig. C.16: Average Hill-plot (dashed) visa Hill-plot (solid) of the residuals of the DWT coefficients $d_j, j = 1, \ldots, 4$, Honolulu station, April 23 – April 25, 2001
Fig. C.17: Average Hill-plot (dashed) via Hill-plot (solid) of the residuals of the DWT coefficients \(d_j, j = 1, \ldots, 4\), Boulder station, April 23 – April 25, 2001
Fig. C.18: Average Hill-plot (dashed) visa Hill-plot (solid) of the residuals of the DWT coefficients $d_j$, $j = 1, \ldots, 4$, College station, April 23 – April 25, 2001
D. APPENDIX. HILL ESTIMATES OF THE TAIL INDEX OF THE DWT COEFFICIENTS

Honolulu. March 30 – April 2

Fig. D.1: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 – $d_2$, 3 – $d_3$, 4 – $d_4$. Honolulu station, March 30 – April 2, 2001
Fig. D.2: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, $2 - d_2$, $3 - d_3$, $4 - d_4$. Boulder station, March 30 – April 2, 2001
Fig. D.3: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, March 30 – April 2, 2001
Fig. D.4: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, March 30 – April 2, 2001
Boulder. March 30 – April 2

Fig. D.5: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, March 30 – April 2, 2001
Fig. D.6: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of \(d_1\), 2 - \(d_2\), 3 - \(d_3\), 4 - \(d_4\). College station, March 30 – April 2, 2001
Honolulu. April 23 – April 25

Fig. D.7: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Honolulu station, April 23 – April 25, 2001.
Fig. D.8: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, April 23 – April 25, 2001
College. April 23 – April 25

Fig. D.9: Hill estimates found using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, April 23 – April 25, 2001
Honolulu. April 23 – April 25

Fig. D.10: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, $2 - d_2$, $3 - d_3$, $4 - d_4$. Honolulu station, April 23 – April 25, 2001
Fig. D.11: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. Boulder station, April 23 – April 25, 2001
Fig. D.12: Hill estimates and their confidence intervals using different methods: 1 - Hill estimates of $d_1$, 2 - $d_2$, 3 - $d_3$, 4 - $d_4$. College station, April 23 – April 25, 2001