Application of Data Mining and Mathematical Analysis to the Zeta Function and the Riemann Hypothesis

Isa M. Muqattash

Advisors:
Dr. A. Kontostathis
Dr. M. Yahdi

December 9, 2004

Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science and Mathematics
Acknowledgement

Many thanks to my advisors Dr. M. Yahdi and Dr. A. Kontostathis for the tremendously useful help and support that they provided throughout this project. I would like to also thank Ursinus College for the great opportunity of summer research that I was provided over the past two years as a summer fellow researcher.

Isa M. Muqattash
CONTENTS

1. Abstract and Introduction ................................. 5

2. Poem: Where are the Zeros of Zeta of s? ...................... 7

3. Notations .................................................. 9

4. The Zeta Function ......................................... 10
   4.1 History of Zeta ......................................... 10
   4.2 Importance of Zeta ..................................... 12
   4.3 Theory and Properties of Zeta .......................... 13

5. The Riemann Hypothesis (RH) ................................. 16
   5.1 Introduction to RH ..................................... 16
   5.2 Evidence in Support of RH .............................. 17
   5.3 Consequences of RH ................................... 19
   5.4 Approaches to Prove RH ................................. 20
5.5 New Approach to Prove RH .............................. 21
5.6 Verifying RH .............................................. 23
5.7 Conclusion .................................................. 25

6. Introduction to Data Mining ................................. 26
6.1 What is Data Mining? ...................................... 26
6.2 History of Data Mining .................................... 27
6.3 Measuring the Reliability of Data Mining .................... 27
6.4 Types of Data Mining ...................................... 30
6.5 Conclusion .................................................. 31

7. Combining Data Mining and Mathematical Analysis ............ 32

8. Applying Data Mining and Mathematical Analysis to the Zeta Function and RH .................................................. 34
8.1 Data Mining Study ........................................... 34
8.2 Mathematical Analysis Study ................................ 41

9. Concluding Remarks ......................................... 52

10. Appendix A: Tables of Empirical Results ..................... 54

11. References .................................................. 59
1. ABSTRACT AND INTRODUCTION

We propose a methodology that is based on an iterative process of interaction between mathematical analysis and data mining. This approach gives better results and knowledge about the given data than using only one of the two approaches. The technique proposed may be applied to any set of data, including non-numeric sets, if the relations and patterns are restricted to contain only operations under which the studied set is closed. Hence, the technique proposed may be applied to various fields where random data or apparently chaotic phenomena can arise from functional processes.

To test the power of the proposed method, we study the embedded patterns and relations amongst the imaginary parts of the non-trivial zeros of the Riemann zeta function. In particular, we simplify the study of the Riemann zeta function and the Riemann Hypothesis; an open problem for about 145 years, and one of only seven problems of the millennium. Using data mining techniques, we construct an infinite family of recursively-defined fitting
curves for the imaginary part of the zeros on the critical line. We then apply techniques from analysis to show that those fitting curves are all asymptotic approximations. For the first $10^5$ zeros of the zeta function, the relative errors of the proposed approximations were bounded in $[1.402 \times 10^{-8}\%, 7.007\%]$.

We introduce the idea of reducing the study of the Riemann Hypothesis from the entire complex plane, to a study of a real line. This is achieved by fixing the imaginary part of the non-trivial zeros of Riemann’s Zeta function to work within a horizontal real line. Via this approach, a new equivalence to the Riemann Hypothesis is introduced. The result is that the Riemann Hypothesis needs to be validated only for a real line rather than on the entire complex plane or critical strip.
2. POEM: WHERE ARE THE ZEROS OF ZETA OF S?

*A Song by Tom Apostol*¹

Where are the zeroes of zeta of s?
G.F.B. Riemann has made a good guess;
They’re all on the critical line, said he,
And their density’s one over \( \frac{2}{\pi} \log t \).

This statement of Riemann has been like a trigger,
And many good men, with vim and with vigor,
Have attempted to find, with mathematical rigor,
What happens to zeta as \( \operatorname{mod} t \) gets bigger.

The efforts of Landau and Bohr and Cramer,
Littlewood, Hardy and Titchmarsh are there,
In spite of their effort and skill and finesse,
In locating the zeros there’s been little success.

In 1914 G.H. Hardy did find,
An infinite number do lay on the line,
His theorem, however, won’t rule out the case,
There might be a zero at some other place.

Oh, where are the zeroes of zeta of s?
We must know exactly, we cannot just guess.

¹ Professor Emeritus of Mathematics at Caltech. The song was written and performed at the Caltech Number Theory conference in June of 1955. Refer to [54] and [53] for more and for a sound/video clip.
In order to strengthen the prime number theorem,
The integral’s contour must never go near ‘em.

Let $P$ be the function $\pi$ minus $Li$,
The order of $P$ is not known for $x$ high,
If square root of $x$ times $\log x$ we could show,
Then Riemann’s conjecture would surely be so.

Related to this is another enigma,
Concerning the Lindelöf function $\mu$ sigma.
Which measures the growth in the critical strip,
On the number of zeros it gives us a grip.

But nobody knows how this function behaves,
Convexity tells us it can have no waves,
Lindelöf said that the shape of its graph,
Is constant when sigma is more than one-half.

There’s a moral to draw from this sad tale of woe,
Which every young genius among you should know:
If you tackle a problem and seem to get stuck,
Use the Riemann Mapping Theorem and you’ll have better luck.
3. NOTATIONS

Throughout that paper, we make use of the following notation:

- $N$ denotes the set of all natural numbers $\{1, 2, 3, \ldots\}$.
- $n$ denotes a natural number.
- $x, y$ and $u$ denote real numbers.
- $s$ denotes a complex number $\sigma + it$.
- $\bar{s}$ denotes the complex conjugate of $s$; given by $\sigma - it$.
- $\Re(s)$ denotes the real part of $s$.
- $\Im(s)$ denotes the imaginary part of $s$.
- $p$ denotes a prime number.
- $\{\alpha_n\}$ denotes a countable sequence of real numbers.
- $f \sim g$ denotes that the functions $f$ and $g$ are asymptotic.
4. THE ZETA FUNCTION

4.1 History of Zeta

Through his study of prime numbers early in the 18th century, the Swiss mathematician Leonard Euler (1707-1783) discovered the zeta function \( \zeta(s) \). Euler inquired about the convergence of the sum of the multiplicative reciprocals of all primes, \( \sum \frac{1}{p} \). Starting with the divergent harmonic series \( \sum_{n \in \mathbb{N}} \frac{1}{n} \), Euler thought of splitting this sum into two parts, a sum of all the terms involving the primes, and a sum involving the terms with composite numbers. He wanted to show that the latter sum is convergent, and thus conclude that the sum of the reciprocals of all primes diverges. Yet, since it is infinite, and its two parts do not both converge, Euler was not able to split the harmonic series the way he wanted. That is, an infinite series cannot be split into various parts unless all the parts converge.

An alternate was to study the Dirichlet \( L_{+1} \)-series \( \sum_{n \in \mathbb{N}} \frac{1}{n^s} \) as \( s \) approaches
one from the right (see [11]). This series converges as long as the single complex variable $s$ is strictly larger than one. Hence, it can be split into the two sums the way Euler desired. The Dirichlet $L_{+1}$-series became an essential element in mathematics, and is known as the zeta function

$$\zeta(s) = \sum_{n \in \mathbb{N}} \frac{1}{n^s}. \quad (4.1)$$

In 1859, in his paper “On the Number of Primes Less Than a Given Quantity” [34], Bernard Riemann (1826-1866) relied on analytic continuation to extend Euler’s zeta function to the entire complex plane, with a single pole at $s = 1$. It is worth noting here that Riemann did not talk about analytically continuing Euler’s zeta function beyond the half-plane $\Re(s) > 1$, but rather talked about finding a valid formula that defines the zeta function for all complex numbers, $s$. This differs from the current view of analytic continuation. From here on, by zeta we will mean Riemann’s zeta function given by equation 4.2,

$$2 \sin(\pi s) \Pi(s - 1) \zeta(s) = i \int_{-\infty}^{\infty} \frac{(-x)^{s-1} \, dx}{e^{x} - 1}; \quad (4.2)$$

where $\Pi(s - 1)$ is the Gaussian notation for the extended factorial function for all complex numbers with $\Re(s) > -1$, and satisfies
4. The Zeta Function

\[ \int_0^\infty e^{-nx}x^{s-1} \, dx = \frac{\Pi(s - 1)}{n^s}. \] (4.3)

4.2 Importance of Zeta

Although a long time has passed since Euler discovered the zeta function, one might wonder why mathematicians continue to study the zeta function. There is a tight connection between the zeta function and the pattern of the prime numbers. Euler discovered that the zeta function can be defined as a series of all primes. He showed (see [11]) that if \( p \) ranges over all the primes, then the zeta function can be given by the Euler Product Formula

\[ \zeta(s) = \sum_{n \in \mathbb{N}} \frac{1}{n^s} = \prod \frac{1}{1 - \frac{1}{p^s}}. \] (4.4)

The relation between the zeta function and the primes extends beyond equation 4.4. A hypothesis due to Riemann regarding the zeros of zeta, and hence known as the Riemann Hypothesis (see chapter 5), has many implications about the distribution of the primes, as well as better approximations to various arithmetic functions such as the difference between the prime counting function \( \pi(x) \) and the logarithmic integral \( \text{Li}(x) \).
The zeta function appears in other fields of mathematics such as applied statistics (Zipf’s law and Zipf-Mandelbrot law) and the mathematical theory of music tuning. The zeta function also appears in physics; especially in areas relevant to chaos in classical and quantum mechanics. For instance, one study uses prime numbers to define an abstract numerical gas, and thus uses the zeta function as a (thermodynamic) partition function [21].

In the next section, we highlight the basic properties and theories related to the zeta function.

4.3 Theory and Properties of Zeta

Riemann pointed out that the zeta function has two types of zeros. First, trivial zeros that consist of all negative even integers. Second, an infinite number of non-trivial zeros which are all complex, and are known to lie in the strip $0 < \Re(s) < 1$. The trivial zeros are well understood, but the study of the non-trivial zeros of zeta is still ongoing.

The Riemann zeta function can be given in various equivalent forms. In the right half-plane $\sigma > 0$, the zeta function can be defined as follows [15]:

$$
\zeta(s) = \frac{s}{s-1} - s \int_1^\infty \frac{u}{u^{s+1}} du.
$$

(4.5)
This formula gives a good approximation for the zeta function in the critical strip, and can thus be used in the study of the Riemann Hypothesis. The uniqueness obtained from the principle of analytic continuation guarantees that this definition is consistent with equation 4.2.

An important reflective property that the zeta function satisfies is the functional equation shown in Theorem 1; which relates \( \zeta(1-s) \) to \( \zeta(s) \).

**Theorem 1.** For all complex numbers \( s \), we have

\[
\pi^{-s/2} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \pi^{-(1-s)/2} \Gamma\left(\frac{1-s}{2}\right) \zeta(1-s).
\]

In his paper [34], Riemann introduces the xi function \( \xi \).

**Definition 1.** For any complex number \( s \), we define

\[
\xi(s) := \frac{1}{2} s(s-1) \pi^{-s/2} \Gamma\left(\frac{s}{2}\right) \zeta(s).
\]

With this definition, the functional equation in Theorem 1 reduces to

\[
\xi(s) = \xi(1-s).
\]

Moreover, since the gamma function does not have any zeros in the critical strip, it follows that the zeta and xi functions have the same zeros in the critical strip.
We conclude this chapter by noting that zeta is a meromorphic function with a simple pole of residue 1 at $s = 1$ [49]. A function is said to be meromorphic if it is a single-valued function that is analytic in all but possibly a discrete subset of its domain. At those singularities, it must go to infinity like a polynomial (i.e., these exceptional points must be poles and not essential singularities) [52]. Theorem 2 can be used to obtain the $n^{th}$ derivative of the zeta function in the complex plane [1].

**Theorem 2.** For each integer $k \geq 1$ and all complex numbers $s$, we have

$$(-1)^k \zeta^{(k)}(1 - s) = 2(2\pi)^{-s} \sum_{m=0}^{k} \sum_{r=0}^{m} \binom{k}{m} \binom{m}{r} \left\{ \Re \left( z^{k-m} \right) \cos \left( \frac{\pi s}{2} \right) \right. + \left. \Im \left( z^{k-m} \right) \sin \left( \frac{\pi s}{2} \right) \right\} \Gamma^{(r)}(s) \zeta^{(m-r)}(s);$$

where $z$ denotes the fixed complex number $z = -\log(2\pi) - \frac{i\pi}{2}$.

This section briefly discussed the basic properties of the Riemann zeta function. Our work is more focused on the zeros of the zeta function and the Riemann Hypothesis, which is discussed in the next chapter.
5. THE RIEMANN HYPOTHESIS (RH)

5.1 Introduction to RH

As mentioned in section 4.3, the zeta function has an infinite number of complex zeros in the strip $0 < \Re(s) < 1$.

**Definition 2.** The strip $0 < \Re(s) < 1$ of the complex plane is called the critical strip.

Riemann calculated several non-trivial zeros of zeta and found that they all have real part $\sigma = \frac{1}{2}$.

**Definition 3.** The vertical line $\Re(s) = \frac{1}{2}$ in the complex plane is called the critical line.

Riemann conjectured that all the non-trivial zeros of zeta lie on this critical line. This is known as the Riemann Hypothesis (RH), considered to be one of the most important open problems in mathematics. Due to
its importance, the Clay Institute of Mathematics is offering a $1 million prize for the first person to prove the hypothesis [47]. Surprisingly, a counter example to the hypothesis does not win the award!

**Riemann Hypothesis.** *The non-trivial zeros of the Riemann zeta function lie on the critical line* $\Re(s) = \frac{1}{2}$.

Riemann first stated this hypothesis in his paper “On the Number of Prime Numbers less than a Given Quantity” that was published in 1859. Mathematicians have since been struggling to prove the hypothesis. This includes Riemann himself, who admits to having given up, at least temporarily, on a proof of RH after several unsuccessful attempts.

### 5.2 Evidence in Support of RH

Though all attempts to prove, or disprove, the Riemann Hypothesis have been unsuccessful thus far, many mathematicians have pursued evidence of its truth. Over the years, there has been increasing evidence in support of RH, both empirically and theoretically. From a theoretical perspective, it has been shown that at least 40% of the zeros of zeta lie on the critical line [42]. Additionally, the zeros are known to be concentrated about $\Re(s) = \frac{1}{2}$; meaning that for any $\delta > 0$, all but an infinitesimal proportion of the zeros
lie in the strip \( \{ s = \sigma + it : \frac{1}{2} - \delta < \sigma < \frac{1}{2} + \delta \} \). Moreover, Hardy proved in 1914 that there is an infinite number of zeros on the critical line [18]. This is in addition to the known fact that a finite number of zeros, at most, lie off of the critical line [16, p 4-5].

Empirically, there is great support for RH. As of the date this paper was completed, over 793 billion zeros have been calculated by zetaGrid.net, an open source and platform independent grid system that uses idle CPU cycles from volunteer computers. All those zeros fall on the critical line! ZetaGrid continues to calculate over one billion zeros of the zeta function on a daily basis. We would here like to invite the reader to support this empirical verification of the Riemann Hypothesis by supporting and taking part in the zetaGrid project. A method for calculating the non-trivial zeros of the zeta function is outlined in [31].

Due to the large evidence in its favor, a great majority of mathematicians believe in the truth of the Riemann Hypothesis. In this paper, we do not search for a disproof or counter example to RH, but rather search for a proof of the hypothesis.
5. The Riemann Hypothesis (RH)

5.3 Consequences of RH

Many mathematicians use RH to establish new “theories” in various fields of mathematics. For example, a proof of RH would lead to new knowledge about the distribution of the primes, as well as better approximations to various arithmetic functions such as the difference between the prime counting function \( \pi(x) \) and the logarithmic integral \( \text{Li}(x) \). Precisely, RH implies that

\[
\pi(x) = \text{Li}(x) + O \left( x^{\frac{1}{2}} \log x \right);
\]

a better approximation than currently known. This is in addition to the fact that RH implies a strong lower bound on \( \pi(x) \). Refer to [15] and [31] for details. Moreover, a proof of the Riemann Hypothesis would validate all the “theories” that were set based on the assumption that the hypothesis is true. An example of such theory is the Lindelöf Hypothesis [12, p 186], which has been shown to be a consequence of RH [37, p 337]. Moreover, the truth of RH implies that the zeros of all derivatives of the function \( \xi \), defined in section 4.3, are on the critical line. The fact that over 99% of the zeros of \( \xi^{(3)} \) are on the critical line is consistent with the Riemann Hypothesis [50].

What would the consequences be if the hypothesis turned out to be false? The most important result of the falsehood of RH would be strange irregularities and chaos in the distribution of the prime numbers.
5.4 Approaches to Prove RH

Mathematicians from various fields have attempted proofs of the Riemann Hypothesis. In this section, we provide several theorems and conjectures that are equivalent statements to RH.

Many mathematicians, including Hilbert and Pólya, suggest that the best approach to prove the Riemann Hypothesis is through finding a Hermitian operator whose eigenvalues are the non-trivial zeros of $\zeta\left(\frac{1}{2} + it\right)$. Finding such an operator would imply the truth of the Riemann Hypothesis since the operator has real-valued eigenvalues [50].

Theorems 3, 4, 5, and 6 are equivalent to the truth of RH (see [50]). Nyman studied the Riemann Hypothesis from a functional analysis perspective and has come to the following result.

**Theorem 3.** RH holds if and only if $\text{span}_{L^2(0,1)}\{n_\alpha, 0 < \alpha < 1\} = L^2(0,1)$; where $n_\alpha(t) = \left\{\frac{\alpha}{t}\right\} - \alpha \left\{\frac{1}{t}\right\}$, and $\{x\} = x - \lfloor x \rfloor$ is the fractional part of nonnegative $x$.

Balazard and Saias were able to show that Theorem 3 can be simplified to Theorem 4.
5. The Riemann Hypothesis (RH)

**Theorem 4.** RH holds if and only if

\[
\inf_A \int_{-\infty}^{\infty} \left| 1 - A \left(\frac{1}{2} + it\right) \zeta \left(\frac{1}{2} + it\right) \right| \frac{dt}{\frac{1}{4} + t^2} = 0;
\]

where the infimum is taken over all Dirichlet polynomials \(A\).

Xian-Jin Li found that we can reduce the study of the Riemann Hypothesis to a study of certain series. Theorem 5 is an example of this approach.

**Theorem 5.** RH holds if and only if \(\lambda_n = \sum_{\rho} (1 - (1 - 1/\rho))\) is nonnegative for all positive integer \(n\); where the sum is taken over \(\rho\), the non-trivial zeros of \(\zeta\). \(\lambda_n\) can also be given by

\[
\frac{1}{(n-1)!} \left. \frac{d^n}{ds^n} \left( s^{n-1} \log \xi(s) \right) \right|_{s=1}.
\]

A recent result obtained in 2002 by Lagarias is shown in Theorem 6.

**Theorem 6.** If by \(\sigma(n)\) we denote the sum of the positive divisors of \(n\), then RH holds if and only if \(\sigma(n) \leq H_n + \exp(H_n) \log H_n\); where \(H_n = \sum_{\nu=1}^{n} \frac{1}{\nu} \).

In the next section, we introduce a new approach to prove the Riemann Hypothesis, and derive a new statement equivalent to RH.

### 5.5 New Approach to Prove RH

We see that mathematicians study the zeta function and the Riemann Hypothesis from various perspectives and fields of mathematics such as number
theory, complex analysis, and random matrices. Here, we would like to introduce a new approach \(^1\) which simplifies the study of RH from the complex plane to a real line. This is done via two methods. First, by fixing the imaginary part of arbitrary zeros of the zeta function, and thus studying the real part of those zeros along a horizontal line. Second, we can study the imaginary part of the zeros of zeta on the critical line, and search for embedded patterns and relations amongst them (refer to chapter 8). This provides valuable information on the distribution of the zeros of the zeta function on the critical line. In this section, we give an example on the first approach. We make use of the following fact that appears in [37, p 30], as it is the core of the proof of Theorem 7.

**Lemma 1.** If \( s \) is a zero of zeta, then \( \bar{s}, 1 - s, \) and \( 1 - \bar{s} \) are also zeros of zeta; where \( \bar{s} \) is the complex conjugate of \( s \).

**Theorem 7.** RH is equivalent to the statement that no two distinct zeros of zeta have the same imaginary part.

**Proof.** Suppose no two distinct zeros of zeta have the same imaginary part.

Let \( s = \sigma + it \) be complex such that \( \zeta(s) = 0 \).

By Lemma 1, \( \zeta(1 - \bar{s}) = 0 \).

\(^1\) To our knowledge, there is no explicit literature on this approach.
5. The Riemann Hypothesis (RH)

But $1 - \overline{s} = (1 - \sigma) + it$ has the same complex part as $s$.

Hence, $s = 1 - \overline{s}$, and thus $\sigma = \frac{1}{2}$.

Conversely, if RH is true, then all the zeros lie on the critical line,

and hence no two distinct zeros of zeta have the same imaginary part. \hfill \Box

Evidently, by studying arbitrary zeros of the zeta function with fixed imaginary parts, the study of RH can be tremendously simplified. Furthermore, we can obtain many sufficient conditions for RH to hold.

5.6 Verifying RH

In the previous sections, many equivalent conditions to the truth of RH were introduced. In this section, we provide a method to verify RH in the critical strip up to height $T$. This method enables us to verify the Riemann Hypothesis without calculating the actual values of the non-trivial zeros of zeta. To explain the methodology behind this approach, the following definitions are needed.

**Definition 4.** For real $t$, the Riemann-Siegel formula is defined by

$$Z(t) := e^{i\theta(t)} \zeta \left( \frac{1}{2} + it \right);$$

where $\theta(t) = \arg \left( \pi^{-it/2} \Gamma \left( \frac{1 + 2it}{4} \right) \right) = \Im \left( \log \Gamma \left( \frac{1 + 2it}{4} \right) \right) - \frac{t}{2} \log(\pi)$. 

Definition 5. By $N(T)$, we denote the number of non-trivial zeros of the Riemann zeta function in the critical strip with imaginary part $t$ between zero and $T$.

It is clear that for all real $t$,

$$|Z(t)| = \left| \zeta \left( \frac{1}{2} + it \right) \right|.$$  

The function $Z$ has the same number of zeros as the zeta function. Hence, the Riemann Hypothesis holds up to height $T$ if and only if for $0 < t < T$, the number of zeros of $Z(t)$ is equal to $N(T)$. On the other hand, if an exact match is not obtained, then the Riemann Hypothesis would have been disproved.

The number of zeros of $Z(t)$ can be counted in accordance to Gram’s Law [51]. It is worth noting here that $Z(t)$ is the primary method for counting the zeros of zeta on the critical line.

Finally, for any fixed positive real number $h$, $N(T)$ satisfies

$$N(T) \sim \frac{2\pi}{T} \log T$$

and

$$N(T + h) - N(T) = O(\log T).$$
5.7 Conclusion

The Riemann Hypothesis has been carefully studied from many perspectives. Nonetheless, the hypothesis has resisted all attempts of proof or disproof. Some believe that the mathematical tools and knowledge to date are not sufficient to provide a proof for RH.

Over the years, the evidence in support of the Riemann Hypothesis has been dramatically increasing, and will continue to increase with the development of both mathematics and technology. In particular, the increasing power of computers brings along the ability to conduct empirical investigations of the topic.

Before we continue forward in the study, we give an overview of Data Mining since it is core to our study of RH. The next two chapters introduce Data Mining and its applications in mathematical studies. In section 8.1, we apply data mining to the study of the zeta function and the Riemann Hypothesis.
6. INTRODUCTION TO DATA MINING

6.1 What is Data Mining?

Did you ever take a look at the patient filing shelf at your family doctor’s office? There is no doubt that the number of files shelved has increased to include new patients over the years. Obviously, this increase in the amount of information and data is not only at your doctor’s office, but all over the world. It is estimated that every twenty months, the amount of data stored in databases worldwide increases by a factor of two [44, p 3]! As the amount of data available increases, it becomes harder for humans to detect patterns within the data. Automated tools have been developed to assist in this process.

Data mining involves deep digging into data, to extract valuable information. Formally, data mining can be defined as follows:

Definition 6. Data mining is the automated or semi-automated act of
searching and extracting subtle relationships and information from existing
data (or databases). Consequently, rules and patterns embedded in the data
might be discovered, and future events can be predicted.

6.2 History of Data Mining

The idea of data mining is not new, although the actual term was not intro-
duced until the 1990s [55]. Finding the patterns and rules that are hidden
in data relies heavily on basic concepts from statistics. Moreover, statistical
rules can be used to verify those conjectured rules and patterns.

Data mining also borrows techniques from artificial intelligence (AI). Us-
ing AI, we can discover patterns imbedded within data.

6.3 Measuring the Reliability of Data Mining

Data mining must deal with imperfect data. Common problems include:

- The data might be missing crucial rows of information; known as in-
  stances of data.

- Some instances of the data might be incomplete, and may be missing
  some of its aspects; formally known as attributes.
The data itself might be incorrectly entered.

- The amount of available data might not be large enough.

To achieve the best results possible, problems need to be eliminated prior to the study by “cleaning” and preparing the data for analysis. When possible, missing attributes should be completed, and false data should be removed.

It is important to note that the patterns and rules identified by Data Mining tools are conjectures that need to be tested or proved. The conclusions of a data mining study can be verified via two methods. First, the rules can be empirically tested against a portion of the available data. In this approach, part of the data is used to construct the rules, while another part of the data is used to test the rules. The first type of data is commonly referred to as training data, while the latter is known as the test data. Statistical measures are calculated to determine the accuracy of the model.

Several statistical measures of accuracy can be used. Those used in our study are shown below. The notation is as follows: $\alpha$ is taken to be a natural number, $p_\alpha$ denotes the calculated approximation of $a_\alpha$, $\bar{a}$ and $\bar{p}$ are the mean (average) values of $a$ and $p$ respectively, and $n$ is the size of the data studied.
6. Introduction to Data Mining

- Correlation Coefficient (CC): 
  \[ (n - 1) \sum_{i=1}^{n} (p_i - \bar{p})(a_i - \bar{a}) \]
  \[ \frac{\left( \sum_{i=1}^{n} (p_i - \bar{p})^2 \right) \left( \sum_{i=1}^{n} (a_i - \bar{a})^2 \right)}{n} \].

- Mean Absolute Error (MAE): 
  \[ \frac{\sum_{i=1}^{n} |p_i - a_i|}{n} \].

- Root Mean-Squared Error (RMSE): 
  \[ \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}} \].

- Relative Absolute Error (RAE): 
  \[ \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |a_i - \bar{a}|} \].

- Root Relative Squared Error (RRSE): 
  \[ \frac{\sqrt{\sum_{i=1}^{n} (p_i - a_i)^2}}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2}} \].

Rules generated by data mining tools can also be evaluated by developing an actual proof or disproof. We rely on this second approach for our results that appear in Chapters 7 and 8. We now give an overview of the different types of data mining.
6.4 Types of Data Mining

Generally speaking, four types of patterns and relationships can be obtained via data mining. Reference [46] defines the data mining types as follows:

- **Classes**: Stored data is used to sort data into predetermined groups. For example, a restaurant chain could mine data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped by creating groups with similar attributes. Clustering can be used to identify market segments or consumer affinities.

- **Associations**: Associations between attributes can be identified. For example, a store did a study and realized that men tend to buy beer when they buy diapers.

- **Sequential patterns**: Behavior patterns and trends are identified. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer’s purchase of sleeping bags and hiking shoes.
6.5 Conclusion

In this chapter, we discussed the data mining techniques, as well as means for measuring the accuracy of the rules generated by data mining tools. In the next chapter, we introduce a technique that combines data mining with mathematical analysis.
The previous chapter presented an overview of data mining. We noted that several conditions might lead to wrong or misleading results in the data mining process. One way of validating the rules generated by data mining tools is an actual proof. Although such a proof does not have to be mathematical, we limit the discussion here to proofs that are mathematical in nature.

Picture the following scenario, where a mathematician or a scientist is trying to find certain common mathematical properties within a set of objects, such as sand particles from a beach. Data mining could be used to identify commonalities and patterns in the available data. This information can be used to formulate a hypothesis that can be proved or disproved using traditional techniques.

Interactions between mathematical analysis and data mining can be used to obtain improved results. There are no visible drawbacks to applying a
Combining the two fields. In chapter 8, we use data mining techniques to conjecture patterns and relations within a given sequence. We then use techniques from mathematical analysis to validate the patterns. Conversely, we use mathematical analysis to obtain solid facts that can assist in the data mining process. This creates cycles of interaction between data mining and mathematics, which may result in significant findings not easily obtained using only one of the two approaches.

It is important to note that this methodology can be applied to any set of data, including non-numeric sets, if the relations and patterns contain only operations under which the studied set is closed. By a closed set under an operation we mean:

**Definition 7.** A set $S$ is said to be closed under the operation $\star$ if and only if $a \star b$ is an element of $S$ for all elements $a$ and $b$ of $S$.

An obvious example of possible sets are mathematical groups and rings. A less apparent example of a set with a closed operation is the set of all colors with operation of mixing the colors. The outcome of mixing colors is always a color, and thus the given set is closed with respect to the given operation. In the next chapter, we apply this technique to the zeta function and the Riemann Hypothesis.
8. APPLYING DATA MINING AND MATHEMATICAL ANALYSIS TO THE ZETA FUNCTION AND RH

The idea of combining data mining techniques with mathematical analysis is very appealing. In this chapter, we apply this technique to the Riemann zeta function and the Riemann Hypothesis. We regard the set of non-trivial zeros of the zeta function as our data to be analyzed. The first section of the chapter (8.1) is devoted to the numerical prediction (data mining) process. The second section of the chapter (8.2) discusses the mathematical analysis of the study, as well as the interaction that occurs between the math and data mining processes.

8.1 Data Mining Study

Since all the non-trivial zeros of the zeta function calculated thus far have real part equal to $\frac{1}{2}$, it suffices to only study their imaginary parts. We study the zeros of zeta on the critical line by studying the sequence given by
definition 8. It is important to note that by \( n \), we mean a positive integer, and that we use the notation \( \{\alpha_n\}_n \) for countable sequences of real numbers \( \alpha_n \).

**Definition 8.** For \( n \) covering all positive integers, denote by \( \{t_n\}_n \) the sequence of the positive imaginary parts of all the zeros of the zeta function on the critical line in the upper half of the complex plane, in increasing order.

We wish to find relationships between a given element \( t_n \), and previous elements \( t_1, t_2, t_3, \ldots, t_{n-1} \). The first problem to answer at this point is what type of relationship we are seeking. We first consider a linear formula.

We search for possible formulas for \( t_n \) that are linear in all or some of \( t_1, t_2, t_3, \ldots, t_{n-1} \). Unfortunately, no good model could be found. This is expected; for it is generally believed that the zeros of the zeta function in particular, and any L-function in general, are linearly independent [50].

In order to get past this problem, we chose a different approach for a solution. Instead of studying the sequence \( \{t_n\}_n \), we map the sequence to another sequence \( \{f(t_n)\}_n \), and then study this new sequence for possible linear patterns.

Based on mathematical intuition, accompanied with empirical trial and
error, we identified many possibilities for mapping functions $f$. In the remainder of this section, we give examples of some of the suggested models and compare them. The development of the models was done in Weka; a collection of machine learning algorithms for data mining tasks\footnote{See reference [44] for more details on the Weka tool.}. As a measure of success of any model, we use the statistical measures CC, MAE, RMSE, RAE and RRSE as defined in section 6.3.

We would like to note that there is no specific rationale behind the choice of the statistical measures used; except that they give good pointwise (local)
and overall (global) statistics on the validity and success of the models. The statistical measures Root Mean Square Error and Mean Absolute Error are concerned with the average difference errors. The Relative Absolute Error and Root Relative Squared Error statistics give more intuition about the difference errors relative to the actual values studied. The latter statistics might be misleading. Take for example an error of 1% in a calculation where the actual outcome is known to be 100. The approximate answer given is 100 ± 1. On the other hand, say we approximate a function with smaller relative error .5%, but are approximating a value that is actually $10^6$. The estimated answer would be off by ± $10^6 \times .5\% = ± 5 \times 10^3$. Although the relative error is smaller in the second example, the difference error (which is the mean absolute error for data of size one) is much smaller in the first case. Therefore, to obtain strong local approximations, we attempt to minimize the mean absolute and the root mean square errors.

For practicality in the speed of the study, we calculated all our models and their statistical measures of accuracy using the first $10^4$ elements of $\{t_n\}_n$, split into two subsets: The first $5 \times 10^3$ elements used as training data, while the rest being test data.

For each study, we examined the relationship between the proposed model
<table>
<thead>
<tr>
<th>$n$</th>
<th>$\frac{t_{n+1}}{t_n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.706104125321350</td>
</tr>
<tr>
<td>50</td>
<td>51.009398160818996</td>
</tr>
<tr>
<td>500</td>
<td>500.978044956544980</td>
</tr>
<tr>
<td>2000</td>
<td>2001.021407482380000</td>
</tr>
<tr>
<td>8000</td>
<td>8000.783730290239000</td>
</tr>
<tr>
<td>10000</td>
<td>10000.882909059201000</td>
</tr>
</tbody>
</table>

Tab. 8.1: Sample of input data for the model $n \frac{t_{n+1}}{t_n}$.

$\{f(t_n)\}_n$ and the number of the instance $n$. Tables 8.1 and 8.2 show samples of the training and test instances for two of the proposed models.

Table 8.3 lists the models that were proposed, their approximate linear models, and their statistical measures of validity. The models were trained and tested through Weka [44].

It is clear from Table 8.3 that the mapping model $n \frac{\ln(t_{n+1})}{\ln(t_n)}$ has the best overall statistical measures, and thus has the best linear approximation. Table 8.4 shows a sample of terms of the sequence $\{t_n\}_n$ along with their corresponding approximations, $an + b$, calculated using this model, with the
8. Applying Data Mining and Mathematical Analysis to the Zeta Function and RH

\[ n \ln(t_{n+1}) \]

\[ \ln(t_n) \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \frac{\ln(t_{n+1})}{\ln(t_n)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.189007239929700</td>
</tr>
<tr>
<td>50</td>
<td>50.20133522284500</td>
</tr>
<tr>
<td>500</td>
<td>500.145867031175040</td>
</tr>
<tr>
<td>2000</td>
<td>2000.130412290700200</td>
</tr>
<tr>
<td>8000</td>
<td>8000.087023205120200</td>
</tr>
<tr>
<td>10000</td>
<td>10000.095984553700000</td>
</tr>
</tbody>
</table>

Tab. 8.2: Sample of input data for the model \( n \frac{\ln(t_{n+1})}{\ln(t_n)} \).

<table>
<thead>
<tr>
<th>The Model</th>
<th>Generated Model ((an + b))</th>
<th>CC</th>
<th>MAE</th>
<th>RMSE</th>
<th>RAE(%)</th>
<th>RRSE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n \frac{t_{n+1}}{t_n} )</td>
<td>( n + 0.79 )</td>
<td>1</td>
<td>0.279</td>
<td>0.344</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td>( ne^{\left(\frac{t_{n+1}}{t_n}\right)} )</td>
<td>2.72n + 2.15</td>
<td>1</td>
<td>0.757</td>
<td>0.935</td>
<td>0.006</td>
<td>0.007</td>
</tr>
<tr>
<td>( n\pi^{\left(\frac{t_{n+1}}{t_n}\right)} )</td>
<td>3.14n + 2.84</td>
<td>1</td>
<td>0.001</td>
<td>1.237</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>( n \frac{\ln(t_{n+1})}{\ln(t_n)} )</td>
<td>( n + 0.12 )</td>
<td>1</td>
<td>0.032</td>
<td>0.041</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>( n \ln\left(\frac{\ln(t_{n+1})}{\ln(t_n)}\right) )</td>
<td>0.12</td>
<td>0.04</td>
<td>0.032</td>
<td>0.041</td>
<td>98.492</td>
<td>103.552</td>
</tr>
<tr>
<td>( n \ln\left(\frac{t_{n+1}}{t_n}\right) )</td>
<td>0.79</td>
<td>0.01</td>
<td>0.28</td>
<td>0.344</td>
<td>102.886</td>
<td>100.918</td>
</tr>
</tbody>
</table>

Tab. 8.3: Various model’s and statistical measures of their accuracy.


<table>
<thead>
<tr>
<th>Actual Value</th>
<th>Approximation</th>
<th>Difference Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>21.02204</td>
<td>20.75236</td>
<td>+0.26968</td>
</tr>
<tr>
<td>30.42488</td>
<td>29.22134</td>
<td>+1.20354</td>
</tr>
<tr>
<td>37.58618</td>
<td>36.44748</td>
<td>+1.13870</td>
</tr>
<tr>
<td>4819.028</td>
<td>4820.397</td>
<td>−1.36900</td>
</tr>
<tr>
<td>4824.678</td>
<td>4826.047</td>
<td>−1.36900</td>
</tr>
<tr>
<td>45186.26</td>
<td>45187.50</td>
<td>−1.24000</td>
</tr>
<tr>
<td>51317.06</td>
<td>51318.29</td>
<td>−1.23000</td>
</tr>
<tr>
<td>51321.28</td>
<td>51322.51</td>
<td>−1.23000</td>
</tr>
</tbody>
</table>

Tab. 8.4: Random actual values from \( \{t_n\}_n \) and their approximations given by the model \( n \frac{\ln(t_{n+1})}{\ln(t_n)} \).

choice of \( b = .12244 \). A larger table is given in Appendix A. The remainder of the study is devoted to this model since it has the best performance.

Having identified a satisfactory model for the sequence \( \{t_n\}_n \), we turn to the mathematical analysis process of our study. The next section of this chapter validates the model mathematically.
8.2 Mathematical Analysis Study

Using data mining techniques, we studied the first $10^5$ elements of $\{t_n\}$, and have determined that the function $n \frac{\log(t_{n+1})}{\log(t_n)}$ is “almost” linear in $n$. That is, for certain real constants $a$ and $b$,

$$n \frac{\log(t_{n+1})}{\log(t_n)} \approx an + b. \quad (8.1)$$

The concern that immediately arises is the best value(s) for the constants $a$ and $b$. We make use of the following lemma [37, p 214] and theorem.

**Definition 9.** Let $f$ and $g$ be two functions depending on the same single variable and well-defined near $+\infty$. We say that $f$ is asymptotic to $g$ and write $f \sim g$ if and only if $\lim_{+\infty} \frac{f}{g} = 1$.

**Lemma 2.** The following statements hold:

1. $\log t_n \sim \log n$.
2. $t_n \sim \frac{2\pi n}{\log n}$.

**Theorem 8.** Let $f$, $g$, $f^*$ and $g^*$ be single variable functions such that $f^*$ and $g^*$ are without zeros. Suppose that $f \sim f^*$ and $g \sim g^*$, then:

- $\lim_{+\infty} \frac{f}{g} = \lim_{+\infty} \frac{f^*}{g^*}$ if either limit exists.
8. Applying Data Mining and Mathematical Analysis to the Zeta Function and RH

- \( \lim_{+\infty} f g = \lim_{+\infty} f^* g^* \) if either limit exists.

**Proof.** Let \( f, g, f^* \) and \( g^* \) be single variable functions such that \( f^* \) and \( g^* \) are without zeros.

Now suppose that \( f \sim f^* \) and \( g \sim g^* \), then \( \lim_{+\infty} \frac{f}{f^*} = \lim_{+\infty} \frac{g}{g^*} = \lim_{+\infty} \frac{g^*}{g} = 1 \).

Hence,

\[
\lim_{+\infty} \frac{f}{g} = \lim_{+\infty} \frac{f f^* g^*}{g f^* g^*} = \lim_{+\infty} \frac{f}{f^*} \lim_{+\infty} \frac{g}{g^*} \lim_{+\infty} \frac{f^*}{g^*} = \lim_{+\infty} \frac{f^*}{g^*}
\]

and

\[
\lim_{+\infty} f g = \lim_{+\infty} \frac{f g f^* g^*}{f^* g^*} = \lim_{+\infty} \frac{f}{f^*} \lim_{+\infty} \frac{g}{g^*} \lim_{+\infty} f^* g^* = \lim_{+\infty} f^* g^*.
\]

Using Lemma 2 and Theorem 8, we see that

\[
\lim_{n \to \infty} \frac{\log(t_{n+1})}{\log(t_n)} = \lim_{n \to \infty} \frac{\log(n+1)}{\log(n)} = 1.
\]

But from equation 8.1,

\[
\frac{\log(t_{n+1})}{\log(t_n)} \approx \left( a + \frac{b}{n} \right) \to a \quad (n \to \infty).
\]

Hence, we deduce that the value of \( a \) ought to be 1.
Leaving the “best” value(s) of $b$ aside momentarily, we can conjecture from equation 8.1 that

$$t_{n+1} \approx t_n^{1 + \frac{b}{n}}.$$ 

We now define the following recursive estimations for $t_n$. As we will see later, these are special cases of a general form of estimations of $t_n$.

**Definition 10.** For any integer $n \geq 2$ and for a fixed positive real $b$, let $I_{b}(n)$ denote the approximation of $t_n$ given by

$$I_{b}(n) = t_{n-1}^{1 + \frac{b}{n}}.$$ 

How “good” are $I_{b}(n)$ in approximating $t_n$? We now show that $I_{b}(n)$ are asymptotic to $t_n$ independently of the value of $b$. First, the following lemmas are needed.

**Lemma 3.** The quantity $\left(\frac{2\pi n}{\log n}\right)^{\frac{b}{n}}$ tends to 1 as $n$ tends to infinity.

**Proof.** For any real $x \geq 2$, let the function $f$ be defined by $f(x) := \left(\frac{2\pi x}{\log x}\right)^{\frac{b}{x}}$.

Then

$$\log(f(x)) = \frac{b}{x} \log \left(\frac{2\pi x}{\log x}\right).$$
Using L’Hospital’s rule, we deduce that
\[
\lim_{x \to \infty} \left[ \frac{b}{x} \log \left( \frac{2\pi x}{\log x} \right) \right] = \lim_{x \to \infty} \frac{\log(2\pi x) - \log(\log(x))}{x/b} \\
= \lim_{x \to \infty} b \left( \frac{1}{x} + \frac{1}{x \log x} \right) = 0.
\]

Hence,
\[
\lim_{x \to \infty} \log f(x) = 0;
\]
from which the result follows. \(\square\)

**Lemma 4.** For any fixed real number \(b\), \(t_n^{\frac{b}{n}} \sim \left( \frac{2\pi n}{\log n} \right)^{\frac{b}{n}}\).

**Proof.** From Lemma 2,
\[
t_n \sim \frac{2\pi n}{\log n}.
\]
Equivalently,
\[
\lim_{n \to \infty} \frac{2\pi n}{t_n \log n} = 1.
\]
Therefore,
\[
\lim_{n \to \infty} \log \left( \frac{2\pi n}{t_n \log n} \right)^{\frac{b}{n}} = \lim_{n \to \infty} \left[ \frac{b}{n} \log \left( \frac{2\pi n}{t_n \log n} \right) \right] \\
= \lim_{n \to \infty} \frac{b}{n} \times \log \left( \lim_{n \to \infty} \frac{2\pi n}{t_n \log n} \right) \\
= 0 \times \log(1) = 0.
\]
Thus,
\[
\lim_{n \to \infty} \left( \frac{2\pi n}{t_n \log n} \right)^{\frac{b}{n}} = 1.
\]

Equivalently,
\[
t_n^{\frac{b}{n}} \sim \left( \frac{2\pi n}{\log n} \right)^{\frac{b}{n}}.
\]

\begin{lemma}
For any positive real \( b \), \( I_b(n + 1) \) is asymptotic to \( t_n \).
\end{lemma}

\begin{proof}
We have
\[
\frac{I_b(n + 1)}{t_n} = \frac{t_n^{1 + \frac{b}{n}}}{t_n} = t_n^{\frac{b}{n}}.
\]

From Lemma 4,
\[
t_n^{\frac{b}{n}} \sim \left( \frac{2\pi n}{\log n} \right)^{\frac{b}{n}}.
\]

Using Theorem 8 (with \( g = g^* = 1 \)) and Lemma 3 respectively, we deduce that
\[
\lim_{n \to \infty} t_n^{\frac{b}{n}} = \lim_{n \to \infty} \left( \frac{2\pi n}{\log n} \right)^{\frac{b}{n}} = 1.
\]
\end{proof}

\begin{lemma}
The ratio \( \frac{t_n}{t_{n+1}} \) tends to 1 as \( n \) tends to infinity.
\end{lemma}

\begin{proof}
The result immediately follows from Theorem 8 and Lemma 2.
\end{proof}
Theorem 9. For any positive real $b$, the approximation $I_b(n)$ is asymptotic to $t_n$.

Proof. This consequently follows from Lemmas 5 and 6. \qed

We have seen that, regardless of the value of $b$, $I_b(n)$ is asymptotic to $t_n$. Yet, certain values of $b$ give “better” local approximations. Numerical calculations on the first $10^5$ terms of $\{t_n\}$ show that $b = 0.12244$ is a possible value\(^2\). Using this value of $b$, the difference error $|I_b(n) - t_n|$ was found to be bounded in the interval $[8.564 \times 10^{-6}, 2.106]$. Moreover, the relative errors $\left|\frac{I_b(n) - t_n}{t_n}\right|$ were bounded in the interval $[1.402 \times 10^{-8\%}, 7.007\%]$. The optimal value of $b$ is not known, but is positive since $t_n$ is monotone increasing by definition.

We now extend the approximations given by $I_b$ to a more generic structure. First, by $x$, we will denote an arbitrary real number.

Definition 11. Let $c$ be any constant, and let $f$ be a differentiable function on $(c, \infty)$ and $f'$ its first derivative, such that $f(n) \geq 1$ for all positive integer $n$, $\lim_{x \to \infty} f(x) = 1$ and $f'(x) \neq 0$ for all $x \in (c, \infty)$. Moreover, suppose that

\(^2\) Note that this value differs slightly from that calculated in the previous section (8.1). This is due to a change in the size of the data analyzed.
8. Applying Data Mining and Mathematical Analysis to the Zeta Function and RH

\[
\lim_{x \to \infty} \frac{(f(x) - 1)^2}{f'(x)} \text{ exists, we then define for any positive integer } n:
\]

\[J(n + 1) = t_{n}^{f(n)}.\]

Note that the function \(1 + \frac{b}{x}\) in the approximations \(I_{n}(n)\) satisfies the conditions of the function \(f\) in definition 11. Therefore, such functions exist.

For \(J(n + 1)\) to be a reasonable approximation of \(t_{n+1}\), we need \(f(n)\) to be greater than 1 for all \(n \geq 1\) because \(\{t_{n}\}\) is monotone increasing. We will now show that \(J(n + 1)\) is a good approximation to \(t_{n+1}\). The proof follows a similar procedure to that of Theorem 9.

**Theorem 10.** Any sequence \(\{J(n)\}_n\) satisfying the conditions of Definition 11 is asymptotic to the sequence \(\{t_{n}\}_n\).

**Proof.** Consider

\[
\log \frac{J(n + 1)}{t_{n}} = (f(n) - 1) \log t_{n}.
\]

Since \(f(n) - 1\) is asymptotic to itself, it follows from Lemma 2 and Theorem 8 that, if either limit exists, then

\[
\lim_{n \to \infty} ((f(n) - 1) \log t_{n}) = \lim_{n \to \infty} ((f(n) - 1) \log n).
\]
The last limit is of indeterminate form. By L'Hospital's rule we have

$$\lim_{x \to \infty} \left( (f(x) - 1) \log x \right) = \lim_{x \to \infty} \left( \frac{1}{x} \frac{(f(x) - 1)^2}{f'(x)} \right);$$

which converges to zero due to the restrictions on the function $f$ given in Definition 11.

Hence,

$$\lim_{n \to \infty} \frac{J(n+1)}{t_n} = 1.$$

From this and Lemma 6, we conclude that

$$\lim_{n \to \infty} \frac{J(n)}{t_n} = \lim_{n \to \infty} \frac{J(n+1)}{t_n} \times \lim_{n \to \infty} \frac{t_n}{t_{n+1}} = 1.$$

Though they were found to be asymptotic to $t_n$, there are unanswered questions about the approximations $J$ in general, and $I_b$ in particular. Are the difference errors of the approximations uniformly bounded? What conditions need to be met for those errors to converge? What are the orders of convergence of the functions $J$? Work to answer these questions is still in progress.
The previous form of approximations seems promising. Yet, a recursive sequence of approximations would be easier to deal with and to evaluate. Fortunately, the approximations previously introduced can, in fact, be modified to recursive forms; where $t_{n+1}$ can be approximated in terms of $t_1$ rather than in terms of $t_n$. It is natural then to consider the following approximations, where $t_1 \approx 14.134725142$.

\[
t_{n+1} \approx I_b(n + 1) = t_n + \frac{b}{n} \prod_{\nu=1}^{n} \left(1 + \frac{b}{\nu}\right) := II(n+1). \tag{8.2}
\]

\[
t_{n+1} \approx J(n + 1) = t^{f(n)}_n \prod_{\nu=1}^{n} f(\nu) := JJ(n+1). \tag{8.3}
\]

Dealing with the approximations $II_b$ and $JJ$ has several advantages over dealing with the approximations $I_b$ and $J$. In particular, less needs to be known about the sequence $\{t_n\}_n$, and an approximation of the value of $t_{n+1}$ may be given without knowing the value of the precedent term $t_n$.

Some open questions remain, however: Do the errors from the term-by-term recursive formulas accumulate? If the errors do accumulate, then some modifications need to be done so that the difference error does not diverge.
One approach to solve this problem is to replace $b$ in the formula of $I_b(n)$ by a sequence $\{\epsilon_n\}_n$ of positive real numbers; where each term of the sequence is defined such that the approximation of $I_b(n)$ to $t_n$ is turned into equality. In other words, we can set

$$t_{n+1} = t_n^{1+ \frac{\epsilon_n}{n}}.$$

Equivalently, the terms of the sequence $\{\epsilon_n\}_n$ may be given by

$$\epsilon_n = n \left[ \log t_{n+1} - \log t_n - 1 \right].$$

The first $10^5$ values of $\{\epsilon_n\}$ were calculated, and the calculations show that the terms are bounded in the interval $[1.745 \times 10^{-3}, 2.675 \times 10^{-1}]$. Table 8.5 shows the approximate values of the first ten terms of the sequence $\{\epsilon_n\}$. A larger table appears in Appendix A.

The sequence $\{\epsilon_n\}_n$ is still to be studied for possible topological properties, structures, and embedded patterns. Here, we see how the use of mathematical analysis gives insight into directions for pursuing this problem using data mining techniques.
Fig. 8.2: The first few terms of \( \{\epsilon_n\}_n \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \epsilon_n )</th>
<th>( n )</th>
<th>( \epsilon_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.149864710057059</td>
<td>6</td>
<td>0.140545529783504</td>
</tr>
<tr>
<td>2</td>
<td>0.114092639649147</td>
<td>7</td>
<td>0.107859603005814</td>
</tr>
<tr>
<td>3</td>
<td>0.182601735994910</td>
<td>8</td>
<td>0.217642185424379</td>
</tr>
<tr>
<td>4</td>
<td>0.092850572425210</td>
<td>9</td>
<td>0.084113607105365</td>
</tr>
<tr>
<td>5</td>
<td>0.189007239929707</td>
<td>10</td>
<td>0.159289980449841</td>
</tr>
</tbody>
</table>

Tab. 8.5: The first ten elements of the sequence \( \{\epsilon_n\}_n \).
9. CONCLUDING REMARKS

It is evident from this study that approaches from mathematical analysis can be combined with data mining techniques to achieve greater results than can be obtained in either of the two fields separately.

The study of the Riemann Hypothesis (RH) can be reduced from a study of the entire complex plane to a study of a real line. We studied the imaginary parts of the zeros of the zeta function on the critical line, and introduced several asymptotic fitting approximations for those zeros. On the other hand, by fixing the imaginary parts of the zeros of the zeta function, we introduced an equivalence to the Riemann Hypothesis that reduces the study to a real line.

The techniques we used, which combine data mining and various fields of mathematics, should not be restricted only to the study of mathematical functions. The proposed technique can be applied to any set of data, including non-numeric sets, if the relations and patterns contain only operations
under which the studied set of data is closed.
10. APPENDIX A: TABLES OF EMPIRICAL RESULTS

<table>
<thead>
<tr>
<th>n</th>
<th>$t_n$</th>
<th>$I_{0.12244}(n)$</th>
<th>$I_{0.12244}(n)$</th>
<th>$\epsilon_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>21.022</td>
<td>19.5492</td>
<td>19.5492</td>
<td>0.114093</td>
</tr>
<tr>
<td>3</td>
<td>25.0109</td>
<td>25.3308</td>
<td>28.1329</td>
<td>0.182602</td>
</tr>
<tr>
<td>4</td>
<td>30.4249</td>
<td>28.5227</td>
<td>34.5092</td>
<td>0.0928506</td>
</tr>
<tr>
<td>5</td>
<td>32.9351</td>
<td>33.7777</td>
<td>39.8753</td>
<td>0.189007</td>
</tr>
<tr>
<td>6</td>
<td>37.5862</td>
<td>35.8776</td>
<td>44.6376</td>
<td>0.140546</td>
</tr>
<tr>
<td>7</td>
<td>40.9187</td>
<td>40.4734</td>
<td>48.989</td>
<td>0.10786</td>
</tr>
<tr>
<td>8</td>
<td>43.3271</td>
<td>43.6633</td>
<td>53.0381</td>
<td>0.217642</td>
</tr>
<tr>
<td>9</td>
<td>48.0052</td>
<td>45.8997</td>
<td>56.853</td>
<td>0.0841136</td>
</tr>
<tr>
<td>10</td>
<td>49.7738</td>
<td>50.6012</td>
<td>60.4798</td>
<td>0.15929</td>
</tr>
<tr>
<td>11</td>
<td>52.9703</td>
<td>52.2131</td>
<td>63.9511</td>
<td>0.176115</td>
</tr>
<tr>
<td>12</td>
<td>56.4462</td>
<td>55.3634</td>
<td>67.2913</td>
<td>0.1491</td>
</tr>
<tr>
<td>13</td>
<td>59.347</td>
<td>58.8176</td>
<td>70.5189</td>
<td>0.0786673</td>
</tr>
<tr>
<td>n</td>
<td>( t_n )</td>
<td>( I_{0.12244}(n) )</td>
<td>( II_{0.12244}(n) )</td>
<td>( \epsilon_n )</td>
</tr>
<tr>
<td>-----</td>
<td>-----------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>14</td>
<td>60.8318</td>
<td>61.674</td>
<td>73.6486</td>
<td>0.231753</td>
</tr>
<tr>
<td>15</td>
<td>65.1125</td>
<td>63.0571</td>
<td>76.6921</td>
<td>0.106915</td>
</tr>
<tr>
<td>16</td>
<td>67.0798</td>
<td>67.3704</td>
<td>79.6588</td>
<td>0.137374</td>
</tr>
<tr>
<td>17</td>
<td>69.5464</td>
<td>69.2739</td>
<td>82.5568</td>
<td>0.142686</td>
</tr>
<tr>
<td>18</td>
<td>72.0672</td>
<td>71.704</td>
<td>85.3927</td>
<td>0.207207</td>
</tr>
<tr>
<td>19</td>
<td>75.7047</td>
<td>74.1949</td>
<td>88.1722</td>
<td>0.0827502</td>
</tr>
<tr>
<td>20</td>
<td>77.1448</td>
<td>77.8453</td>
<td>90.9001</td>
<td>0.128977</td>
</tr>
<tr>
<td>21</td>
<td>79.3374</td>
<td>79.2248</td>
<td>93.5806</td>
<td>0.211507</td>
</tr>
<tr>
<td>22</td>
<td>82.9104</td>
<td>81.3866</td>
<td>96.2173</td>
<td>0.108434</td>
</tr>
<tr>
<td>23</td>
<td>84.7355</td>
<td>84.9742</td>
<td>98.8136</td>
<td>0.161897</td>
</tr>
<tr>
<td>24</td>
<td>87.4253</td>
<td>86.762</td>
<td>101.372</td>
<td>0.0843064</td>
</tr>
<tr>
<td>25</td>
<td>88.8091</td>
<td>89.4422</td>
<td>103.896</td>
<td>0.226412</td>
</tr>
<tr>
<td>50</td>
<td>143.112</td>
<td>142.88</td>
<td>159.698</td>
<td>0.201334</td>
</tr>
<tr>
<td>51</td>
<td>146.001</td>
<td>144.862</td>
<td>161.735</td>
<td>0.0991739</td>
</tr>
<tr>
<td>52</td>
<td>147.423</td>
<td>147.758</td>
<td>163.762</td>
<td>0.184198</td>
</tr>
<tr>
<td>53</td>
<td>150.054</td>
<td>149.166</td>
<td>165.779</td>
<td>0.0612679</td>
</tr>
<tr>
<td>54</td>
<td>150.925</td>
<td>151.801</td>
<td>167.786</td>
<td>0.148698</td>
</tr>
<tr>
<td>55</td>
<td>153.025</td>
<td>152.652</td>
<td>169.784</td>
<td>0.218446</td>
</tr>
</tbody>
</table>
## 10. Appendix A: Tables of Empirical Results

<table>
<thead>
<tr>
<th>n</th>
<th>$t_n$</th>
<th>$I_{0.12244}(n)$</th>
<th>$II_{0.12244}(n)$</th>
<th>$\epsilon_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>156.113</td>
<td>154.748</td>
<td>171.772</td>
<td>0.104951</td>
</tr>
<tr>
<td>57</td>
<td>157.598</td>
<td>157.846</td>
<td>173.751</td>
<td>0.0891647</td>
</tr>
<tr>
<td>58</td>
<td>158.85</td>
<td>159.32</td>
<td>175.721</td>
<td>0.167284</td>
</tr>
<tr>
<td>59</td>
<td>161.189</td>
<td>160.559</td>
<td>177.683</td>
<td>0.131884</td>
</tr>
<tr>
<td>100</td>
<td>236.524</td>
<td>235.275</td>
<td>252.798</td>
<td>0.0960915</td>
</tr>
<tr>
<td>101</td>
<td>237.77</td>
<td>238.113</td>
<td>254.534</td>
<td>0.138117</td>
</tr>
<tr>
<td>102</td>
<td>239.555</td>
<td>239.352</td>
<td>256.266</td>
<td>0.115722</td>
</tr>
<tr>
<td>103</td>
<td>241.049</td>
<td>241.136</td>
<td>257.995</td>
<td>0.137703</td>
</tr>
<tr>
<td>104</td>
<td>242.823</td>
<td>242.626</td>
<td>259.72</td>
<td>0.0970415</td>
</tr>
<tr>
<td>105</td>
<td>244.071</td>
<td>244.398</td>
<td>261.443</td>
<td>0.238442</td>
</tr>
<tr>
<td>106</td>
<td>247.137</td>
<td>245.641</td>
<td>263.161</td>
<td>0.0749725</td>
</tr>
<tr>
<td>107</td>
<td>248.102</td>
<td>248.715</td>
<td>264.877</td>
<td>0.114771</td>
</tr>
<tr>
<td>108</td>
<td>249.574</td>
<td>249.672</td>
<td>266.59</td>
<td>0.112667</td>
</tr>
<tr>
<td>109</td>
<td>251.015</td>
<td>251.14</td>
<td>268.299</td>
<td>0.160843</td>
</tr>
<tr>
<td>200</td>
<td>396.382</td>
<td>397.041</td>
<td>414.885</td>
<td>0.129373</td>
</tr>
<tr>
<td>201</td>
<td>397.919</td>
<td>397.836</td>
<td>416.426</td>
<td>0.173913</td>
</tr>
<tr>
<td>202</td>
<td>399.985</td>
<td>399.372</td>
<td>417.967</td>
<td>0.155922</td>
</tr>
<tr>
<td>203</td>
<td>401.839</td>
<td>401.44</td>
<td>419.506</td>
<td>0.0860538</td>
</tr>
</tbody>
</table>
10. Appendix A: Tables of Empirical Results

<table>
<thead>
<tr>
<th>n</th>
<th>$t_n$</th>
<th>$I_{0.12244}(n)$</th>
<th>$II_{0.12244}(n)$</th>
<th>$\epsilon_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>402</td>
<td>682.603</td>
<td>683.255</td>
<td>712.555</td>
<td>0.127186</td>
</tr>
<tr>
<td>403</td>
<td>684.014</td>
<td>683.961</td>
<td>713.986</td>
<td>0.0864992</td>
</tr>
<tr>
<td>404</td>
<td>684.973</td>
<td>685.372</td>
<td>715.416</td>
<td>0.107454</td>
</tr>
<tr>
<td>405</td>
<td>686.163</td>
<td>686.329</td>
<td>716.846</td>
<td>0.162308</td>
</tr>
<tr>
<td>406</td>
<td>687.962</td>
<td>687.519</td>
<td>718.276</td>
<td>0.126991</td>
</tr>
<tr>
<td>407</td>
<td>689.369</td>
<td>689.318</td>
<td>719.705</td>
<td>0.0998094</td>
</tr>
<tr>
<td>408</td>
<td>690.475</td>
<td>690.726</td>
<td>721.135</td>
<td>0.178436</td>
</tr>
<tr>
<td>409</td>
<td>692.452</td>
<td>691.831</td>
<td>722.564</td>
<td>0.0654669</td>
</tr>
<tr>
<td>1000</td>
<td>1419.42</td>
<td>1419.96</td>
<td>1550.29</td>
<td>0.0964551</td>
</tr>
<tr>
<td>1001</td>
<td>1420.42</td>
<td>1420.68</td>
<td>1551.68</td>
<td>0.139156</td>
</tr>
<tr>
<td>1002</td>
<td>1421.85</td>
<td>1421.68</td>
<td>1553.08</td>
<td>0.0592734</td>
</tr>
<tr>
<td>1003</td>
<td>1422.46</td>
<td>1423.11</td>
<td>1554.48</td>
<td>0.194275</td>
</tr>
<tr>
<td>1004</td>
<td>1424.46</td>
<td>1423.72</td>
<td>1555.87</td>
<td>0.136832</td>
</tr>
<tr>
<td>2000</td>
<td>2515.29</td>
<td>2515.47</td>
<td>2973.77</td>
<td>0.130412</td>
</tr>
<tr>
<td>2001</td>
<td>2516.57</td>
<td>2516.49</td>
<td>2975.22</td>
<td>0.0584688</td>
</tr>
<tr>
<td>2002</td>
<td>2517.15</td>
<td>2517.78</td>
<td>2976.68</td>
<td>0.142753</td>
</tr>
<tr>
<td>2003</td>
<td>2518.55</td>
<td>2518.35</td>
<td>2978.14</td>
<td>0.109795</td>
</tr>
<tr>
<td>2004</td>
<td>2519.63</td>
<td>2519.76</td>
<td>2979.6</td>
<td>0.0779884</td>
</tr>
</tbody>
</table>
### 10. Appendix A: Tables of Empirical Results

<table>
<thead>
<tr>
<th>n</th>
<th>$t_n$</th>
<th>$I_{0.12244}(n)$</th>
<th>$II_{0.12244}(n)$</th>
<th>$\epsilon_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>40004</td>
<td>33192.5</td>
<td>33193.2</td>
<td>102998</td>
<td>0.109962</td>
</tr>
<tr>
<td>50000</td>
<td>40433.7</td>
<td>40433.6</td>
<td>141781</td>
<td>0.0631335</td>
</tr>
<tr>
<td>50001</td>
<td>40434.2</td>
<td>40434.7</td>
<td>141785</td>
<td>0.0875826</td>
</tr>
<tr>
<td>50002</td>
<td>40435</td>
<td>40435.3</td>
<td>141789</td>
<td>0.0734205</td>
</tr>
<tr>
<td>50003</td>
<td>40435.6</td>
<td>40436</td>
<td>141793</td>
<td>0.0955404</td>
</tr>
<tr>
<td>50004</td>
<td>40436.4</td>
<td>40436.7</td>
<td>141798</td>
<td>0.0860694</td>
</tr>
<tr>
<td>70000</td>
<td>54511.7</td>
<td>54511.9</td>
<td>233502</td>
<td>0.0686237</td>
</tr>
<tr>
<td>70001</td>
<td>54512.2</td>
<td>54512.7</td>
<td>233507</td>
<td>0.0377572</td>
</tr>
<tr>
<td>70002</td>
<td>54512.6</td>
<td>54513.3</td>
<td>233512</td>
<td>0.108149</td>
</tr>
<tr>
<td>70003</td>
<td>54513.5</td>
<td>54513.6</td>
<td>233517</td>
<td>0.0251233</td>
</tr>
<tr>
<td>70004</td>
<td>54513.7</td>
<td>54514.5</td>
<td>233522</td>
<td>0.12435</td>
</tr>
<tr>
<td>99993</td>
<td>74916.3</td>
<td>74916.3</td>
<td>405398</td>
<td>0.0384912</td>
</tr>
<tr>
<td>99994</td>
<td>74916.6</td>
<td>74917.3</td>
<td>405404</td>
<td>0.1331</td>
</tr>
<tr>
<td>99995</td>
<td>74917.7</td>
<td>74917.6</td>
<td>405411</td>
<td>0.0774336</td>
</tr>
<tr>
<td>99996</td>
<td>74918.4</td>
<td>74918.7</td>
<td>405417</td>
<td>0.0381545</td>
</tr>
<tr>
<td>99997</td>
<td>74918.7</td>
<td>74919.4</td>
<td>405424</td>
<td>0.0456315</td>
</tr>
<tr>
<td>99998</td>
<td>74919.1</td>
<td>74919.7</td>
<td>405430</td>
<td>0.140872</td>
</tr>
<tr>
<td>99999</td>
<td>74920.3</td>
<td>74920.1</td>
<td>405436</td>
<td>0.0675093</td>
</tr>
</tbody>
</table>
11. REFERENCES


5. Bohr, H., Jessen, B. On the Distribution of the values of the Riemann


11. Devlin, K. *How Euler discovered the zeta function.*


36. Stopple, J. Euler, the Symmetric Group and the Riemann Zeta Function.


45. Zalcman, L. *Real Proofs of Complex Theorems (and Vice Versa)*. The


