Masters Research Project

Time Series Analysis

Supervisor: 清野健

João Gabriel Segato Kruse

Student ID	**29B23068**
Division	**Bioengineering**

Graduate School of Egineering Sciences

May, 2023

Contents

Abstract

This project has as a main goal of studying and analysing time series data from biomedical sources to obtain a better understanding of their underlying properties.

1 Introduction

A time series is a collection of measurements made sequentially through time, and they are present in a variety of fields of study [1, Chapter 5]. Stock markets, electrocardiograms, weather forecasts, seismic activities, etc, are just a few examples of time series like data that appears in different areas. With such a presence, robust methods to analyse and extract information of time series are necessary, and have been active topic in high level research for a long time. Tasks such as forecasting, anomaly detection, classification and comparison methods for time series are of vital importance to many industries and people on a daily basis.

1.1 Time Series

In this section, the properties inherent in time series analysis will be explored, as understanding their properties is essential for effective analysis and modeling. We will delve into the fundamental characteristics that define time series, such as stationarity, seasonality, trend, and autocorrelation. These properties provide valuable insights into the underlying patterns, trends, and dependencies within the data. By grasping these key concepts, we can develop robust methodologies for forecasting, anomaly detection, and decision-making based on time series data.

1.1.1 Properties

Now that we have established the significance of understanding and analyzing time series, let's delve deeper into the characteristics and properties of these series. To begin with, time series can be categorized as either **continuous** or **discrete**, depending on whether the observations are recorded continuously over time or at discrete intervals. For the purposes of our discussion, all the time series considered will be discrete in nature, even if the underlying variable being measured is continuous.

Another crucial differentiation lies in the classification of time series as either **deterministic** or **stochastic** processes. A deterministic time series implies that its future values can be precisely predicted based on its past values, indicating a predictable pattern. On the other hand, a stochastic time series suggests that future values are not solely determined by past observations, rendering exact predictions unattainable. Instead, stochastic time series necessitate the concept of probability distribution to capture the uncertainty associated with future values. Again, all the time series that will be worked with will be stochastic in nature.

One of the main ways of describing a stochastic process is through its moments, in particular the first (called the *mean*, μ) and second (called the autocovariance function, acv.f.) one, that can be defined as:

$$\mu(t) = E[X(t)]$$
$$acv.f. = \gamma(t_1, t_2) = E\{[X(t_1) - \mu(t_1)][X(t_2) - \mu(t_2)]\}$$

with this, we can also define the autocorrelation ($\rho(\tau)$) as the normalized acv.f., .

$$\rho(\tau) + \frac{\gamma(\tau)}{\gamma(0)}$$

Another property of the time series that will be introduced here is **stationarity**. Simply put, a series is stationary if there is no systematic change in the mean value and variance and if purely cyclic variations have been removed. Mathematically, a time series is said to be strictly stationary if the joint distribution of X(t_1), ..., X(t_k) is the same as the one of X($t_1 + \tau$), ..., X($t_k + \tau$) for any k. Because this is such a restrictive definition, in many instances it is better to use the looser **second-order stationarity**, that only requires that the mean is constant and that the autocovariance function only depends on the lag. One way to remove a trend in order to make the series more stationary is by differencing it (usually first order differencing is already enough).

When dealing with a series that has a trend and a seasonal component, it is possible to analyze it by decomposing it in a trend, a periodic and a noise component. An example of such decomposition can be viewed in Figure 1.

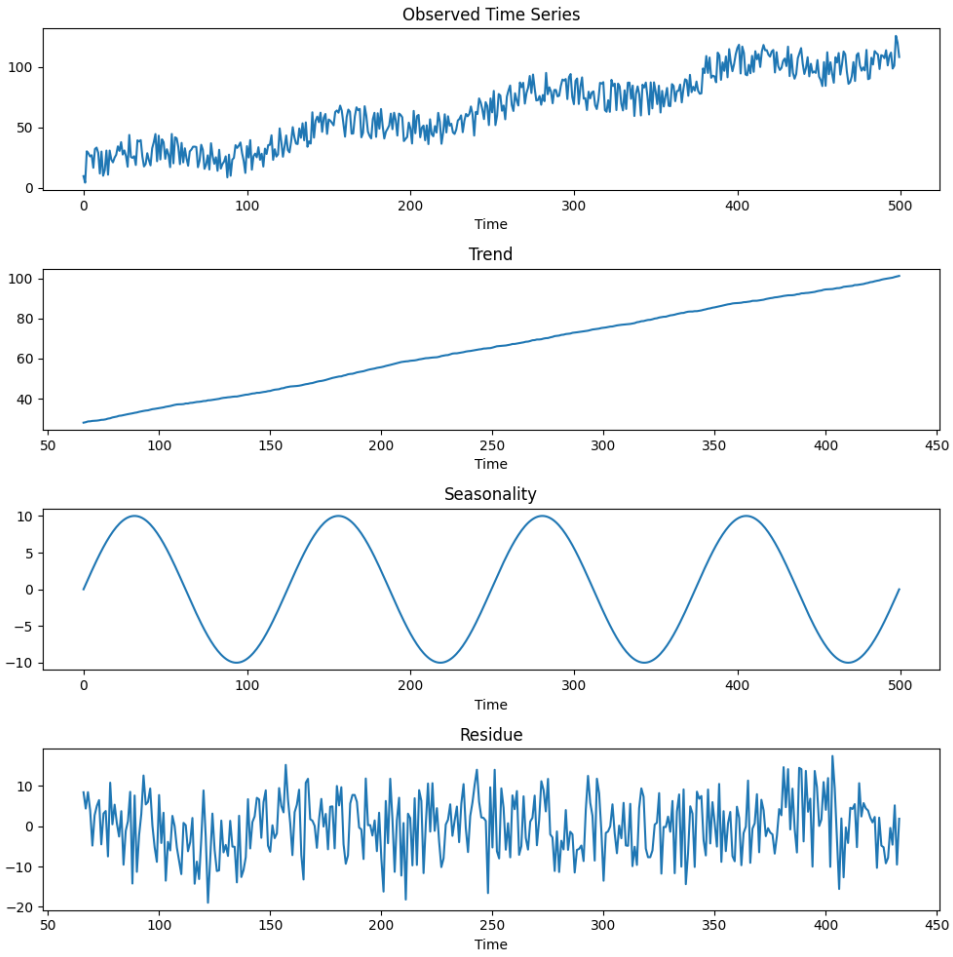

Figure 1: Example of the decomposition of a signal in trend, seasonality and residual noise.

By considering these distinctions, we can gain a more comprehensive understanding of the various types of time series and their inherent properties. The goal of this work is around the description, explanation, prediction and simulation of time series.

1.1.2 Linear Time Series Models

In time series analysis, various models are used to describe the behavior and dynamics of the data. A purely random process is characterized by a sequence of random variables, denoted as Z_t, that are mutually independent and identically distributed (i.i.d.). Building upon this notion, we can define a random walk process, denoted as X_t, where each value is obtained by adding the current random variable, Z_t, to the previous value, X_{t-1}.

$$X_t = X_{t-1} + Z_t$$

Taking a step further, a moving average process of order q, denoted as MA(q), can be formulated. In this model, the value of X_t is determined by a linear combination of the most recent q random variables, Z_t, Z_{t-1}, ..., Z_{t-q}, with respective coefficients β_0, β_1, ..., β_q.

$$X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + ... + \beta_q Z_{t-q}$$

Similarly, an autoregressive process of order p, denoted as AR(p), is defined. In this case, the value of X_t depends on a linear combination of the past p values, X_{t-1}, X_{t-2}, ..., X_{t-p}, with respective coefficients α_0, α_1, ..., α_p, along with the current random variable Z_t.

$$X_t = \alpha_0 X_{t-1} + \alpha_1 X_{t-2} + ... + \alpha_p X_{t-p} + Z_t$$

It is worth noting that AR and MA processes are mathematically equivalent, and they can be combined to create other useful models. For instance, the combination of autoregressive and moving average components gives rise to the autoregressive moving average (ARMA) model. The inclusion of differencing, which involves computing the differences between consecutive observations, allows the use of ARMA models on non-stationary data, leading to the autoregressive integrated moving average (ARIMA) model. Additionally, the ARFIMA model incorporates long-range dependence by including fractional differencing in the time series analysis.

By employing these various models, time series analysts can capture different patterns, trends, and dependencies present in the data, enabling a deeper understanding and facilitating effective forecasting and analysis.

1.2 Time Series in Biomedical Data

In the realm of medicine and health monitoring, time series data of biosignals holds immense value for understanding the dynamic nature of physiological processes and diagnosing various health conditions. In this section, we delve into the fascinating world of time series analysis specifically tailored to biosignals. Time series data derived from biosignals, such as electrocardiograms (ECGs), electroencephalograms (EEGs), and electromyograms (EMGs), offers a unique window into the intricate interplay of physiological signals over time. By unraveling the patterns, rhythms, and anomalies embedded within these temporal sequences, we can unlock invaluable insights for disease detection, monitoring patient health, and personalized treatment strategies. With an emphasis on the unique challenges and specialized techniques associated with biosignal time series analysis, this section aims to equip researchers, clinicians, and data scientists with the tools necessary to harness the full potential of temporal data in the field of medicine.

Heart rate signals, RR intervals, electrocardiograms (ECGs), and electroencephalograms (EEGs) are pivotal time series data sources in the domain of biosignals and medicine, and will be some of the emphasis of this work. Heart rate signals provide a representation of the cardiac activity over time, reflecting the rhythm and frequency of the heartbeats. Derived from heart rate signals, RR intervals capture the time intervals between consecutive R-peaks in an ECG waveform, providing valuable insights into heart rate variability and cardiac health. ECG time series data consists of voltage measurements that depict the electrical activity of the heart, offering a wealth of information about cardiac function, arrhythmias, and myocardial abnormalities. On the other hand, EEG time series data records the electrical activity of the brain, enabling the study of neural dynamics, sleep patterns, seizures, and cognitive processes. These time series data forms are indispensable in diagnosing cardiovascular disorders, monitoring cardiac health, studying brain activity, and facilitating personalized treatment approaches, making them vital tools in the field of biosignals and medicine.

Time series forecasting of biosignals plays a crucial role in various applications within the field of healthcare and biosignal analysis. Here are some notable applications:

1. Disease Diagnosis and Monitoring: Time series forecasting enables the detection and diagnosis of various medical conditions by analyzing biosignals such as electrocardiograms (ECGs), electroencephalograms (EEGs), and electromyograms (EMGs). By forecasting changes in biosignals, abnormalities and patterns associated with diseases like cardiac arrhythmias, epilepsy, and neuromuscular disorders can be detected, facilitating early intervention and monitoring of patients' health status [2–4].

2. Treatment Optimization: Time series forecasting aids in optimizing treatment strategies for patients. By analyzing biosignals over time, healthcare professionals can predict the efficacy of different treatments and adjust medication dosages, therapy interventions, or stimulation parameters in real-time [5]. This helps in tailoring personalized treatment plans and achieving better patient outcomes.

3. Wearable Devices and Remote Monitoring: The use of wearable devices for continuous biosignal monitoring has gained prominence. Time series forecasting techniques can analyze biosignals collected from wearables, such as heart rate variability (HRV) or sleep patterns, to provide insights into an individual's health and well-being. These forecasts enable proactive health management, early warning systems for critical events, and remote patient monitoring, leading to timely interventions and improved patient care. One very proeminent application is in female workers health assessment, as approximately half of female workers leave their jobs due to childbirth and female-specific health issues such as menopausal disorders and premenstrual syndrome [6].

4. Sports Performance and Fitness Tracking: Time series forecasting techniques applied to biosignals, such as heart rate and lactate levels, enable the prediction of performance and fatigue in athletes. This aids in optimizing training regimens, preventing injuries, and enhancing athletic performance. Additionally, in fitness tracking applications, forecasting models applied to biosignals collected from wearable devices help individuals track their progress, set achievable goals, and optimize their fitness routines [7, 8].

By harnessing the power of time series forecasting in biosignal analysis, healthcare professionals, researchers, and individuals can make informed decisions, enhance patient care, and improve overall well-being.

2 Research Project

2.1 Proposal

2.2 Motivation

2.3 Previous Works

3 Time Series Forecasting

In this comprehensive section, we will delve deeper into the fascinating world of time series forecasting. Our exploration will encompass an extensive review of the various techniques available, allowing us to gain a comprehensive understanding of each approach's particularities and guiding principles. To ensure a well-rounded discussion, we will commence with an overview of traditional forecasting methods, exploring their strengths and limitations. Subsequently, we will progress to more advanced techniques that harness the power of deep learning, leveraging its capabilities for enhanced prediction accuracy.

Throughout this section, we will enrich our understanding of these forecasting techniques by presenting a compelling case study. The case study will focus on data collected from Lawson female workers, comprising a diverse range of variables, including heart rate (HR), movement patterns, and temperature. By utilizing this real-world dataset, we will illustrate how these forecasting methods can be effectively applied in practice, elucidating their practical implications and showcasing their performance on complex biosignal data. The dataset measures activity as the root of the sum squared of the acceleration in the three axis (x, y and z) subtracted by one to disregard the gravitational acceleration component, as shown in equation 1. All the values were actually not measured in equally spaced intervals, but instead at every heartbeat, so an averaging of the values at every minute was taken for the heart rate, and the maximum activity in the interval was taken. Figure 2 shows the variation of the activity and heart rate as a function of time for the dataset at hand.

$$Activity = \sqrt{a_x^2 + a_y^2 + a_z^2} - 1 \tag{1}$$

By integrating theoretical discussions, practical examples, and case studies, this section aims to equip readers with a holistic understanding of time series forecasting techniques. Whether you are a researcher, data scientist, or healthcare professional, this comprehensive exploration will empower you to confidently select and apply the most suitable forecasting approach for your specific problem domain, unlocking the potential to make accurate predictions and informed decisions based on biosignal time series data.

3.1 Moving Averages and Autoregressive Models

In this section, we delve further into the modeling of stochastic processes, particularly through Moving Averages (MA), Autoregressive (AR) models, and their combinations. To

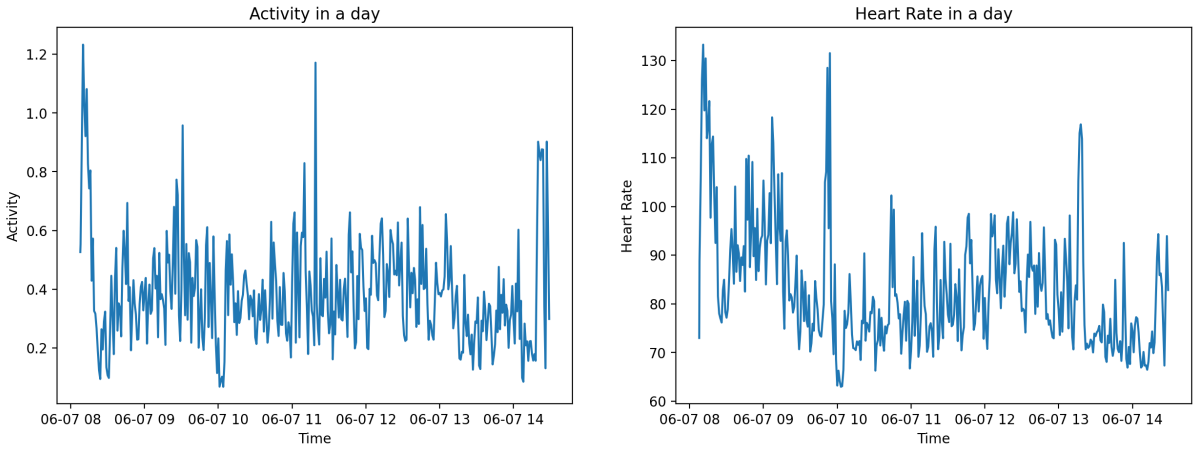

Figure 2: Plot of the measured Heart Rate and Activity for a worker in a given day as a function of time.

make an informed choice among these options, it is crucial to comprehend the correlogram of our variable. Hence, we will now explore the concept of correlograms.

Initially, let's consider a scenario where our time series observations, denoted as x_1, x_2, ..., x_N, are independent and identically distributed random variables. In this case, it can be demonstrated that the expected value for the estimator r_k (representing the autocorrelation coefficient $\rho(k)$) is approximately $-\frac{1}{N}$, while the variance of r_k is approximately $\frac{1}{N}$.

This statistical insight allows us to assess the randomness of a time series by employing the 95% confidence interval of the correlogram coefficients, which can be calculated as $\pm\frac{1.96}{\sqrt{N}}$. Consequently, coefficients falling within this interval suggest a purely random process, while those outside the interval are deemed significant. However, it is important to note that depending on the size of N, approximately 5% of the coefficients may fall outside the interval even in a random process.

By understanding and analyzing the correlogram coefficients, we gain valuable insights into the presence of autocorrelation and the degree of randomness exhibited by the time series. This knowledge aids in selecting appropriate modeling approaches and making meaningful interpretations of the data.

In addition, the correlogram of a time series provides valuable insights into its stationarity and underlying process. When the correlogram coefficients do not decay quickly, it indicates that the series is non-stationary and requires differencing to achieve stationarity. On the other hand, if the correlogram exhibits a sudden cut-off at a specific lag, it suggests a MA(q) process. In contrast, an AR process is characterized by a combination of sinusoids and a damped exponential pattern in the correlogram. Mixed Autoregressive Moving Averages (ARMA) models follow a similar pattern as AR models, having a tendency to display a damped decay and sinusoidal behavior instead of a distinct cut-off. By analyzing the correlogram, we can identify the nature of the underlying process and select an appropriate modeling approach accordingly.

When using a AR(p) model, the fitting of the data can be done as a simple regression model to find the coefficients $\{\hat{\alpha}_1, ..., \hat{\alpha}_p\}$ that best fit equation 2. To determine the order of such process, looking at the ac.f. might not be so insightful as in the MA case. Because of that, we can use what is called the **partial autocorrelation function**, that measures the excess correlation at lag p that is not accounted for in lag $p-1$. In the

partial autocorrelation function, the properties of the MA and AR are inverted when compared to the normal ac.f., meaning that now the AR process will have a cut-off at lag p, while the MA one will tend to attenuate slowly.

$$(X_t - \bar{x}) = \hat{\alpha}_1(X_{t-1} - \bar{x}) + ... + \hat{\alpha}_p(X_{t-p} - \bar{x}) \tag{2}$$

Fitting a MA process is a little harder because efficient explicit estimators cannot be found. Assuming a process of the from

$$X_t = \mu + Z_t + \beta_1 Z_{t-1} + ... + +\beta_q Z_{t-q} \tag{3}$$

we would start with an initial guess for the values of μ, β_1, ..., β_q and then iteratively improve upon it by means of optimization procedures that minimize the residual sum of squares. Feating ARMA and ARIMA models follow similar paths, typically involving iteratively optimizing the residue. Because of this added complexity for fitting MA models when compared to AR, many times AR ones are prefered, even at the expense of more parameters.

Applying this concepts to our study case, we can see in figure 3 the autocorrelation and partial autocorrelation functions for our Heart Rate data. As we can see, the autocorrelation shows a very slow decay, meaning that the series is most likely not stationary and should be differentiated. Doing that, we get the plots on the right, that show a different picture. With that, we can see that mostly only one lag is relevant both for the AR and the MA (the other values that lie outside the 95% confidence interval in farther lags are probably only due to chance, as was mentioned before that approximately 5% of the coefficients would lie outside the bounds even in completely random series). With these informations, we know that a ARIMA model should be considered (as the data is not stationary) taking into account only a single lag.

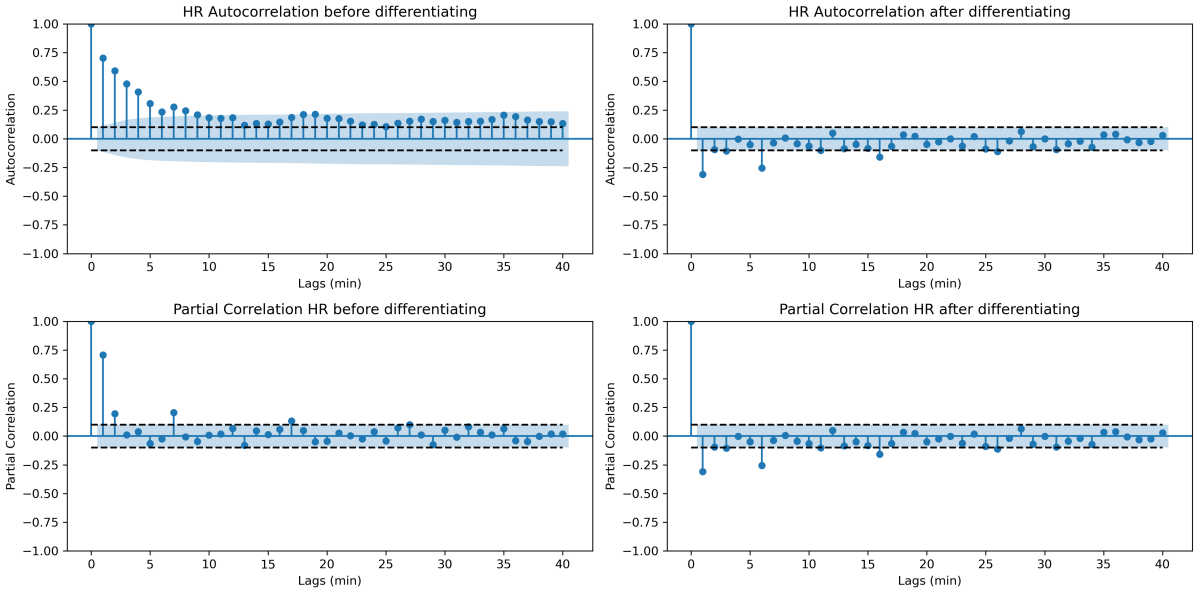

Figure 3: Plot of the autocorrelation and partial autocorrelation for the raw and differentiated HR time series data. The dashed lines mark the 95% confidence interval for the coefficients.

With this information, we can fit an ARIMA model using 1 lag for the MA and one lag for the AR. We can see from Figure 4 that the model tends to fit pretty well to

the data, although it tended not to follow the data on the peaks. Considering the non normalized data, it achieved an MSE of 71.3.

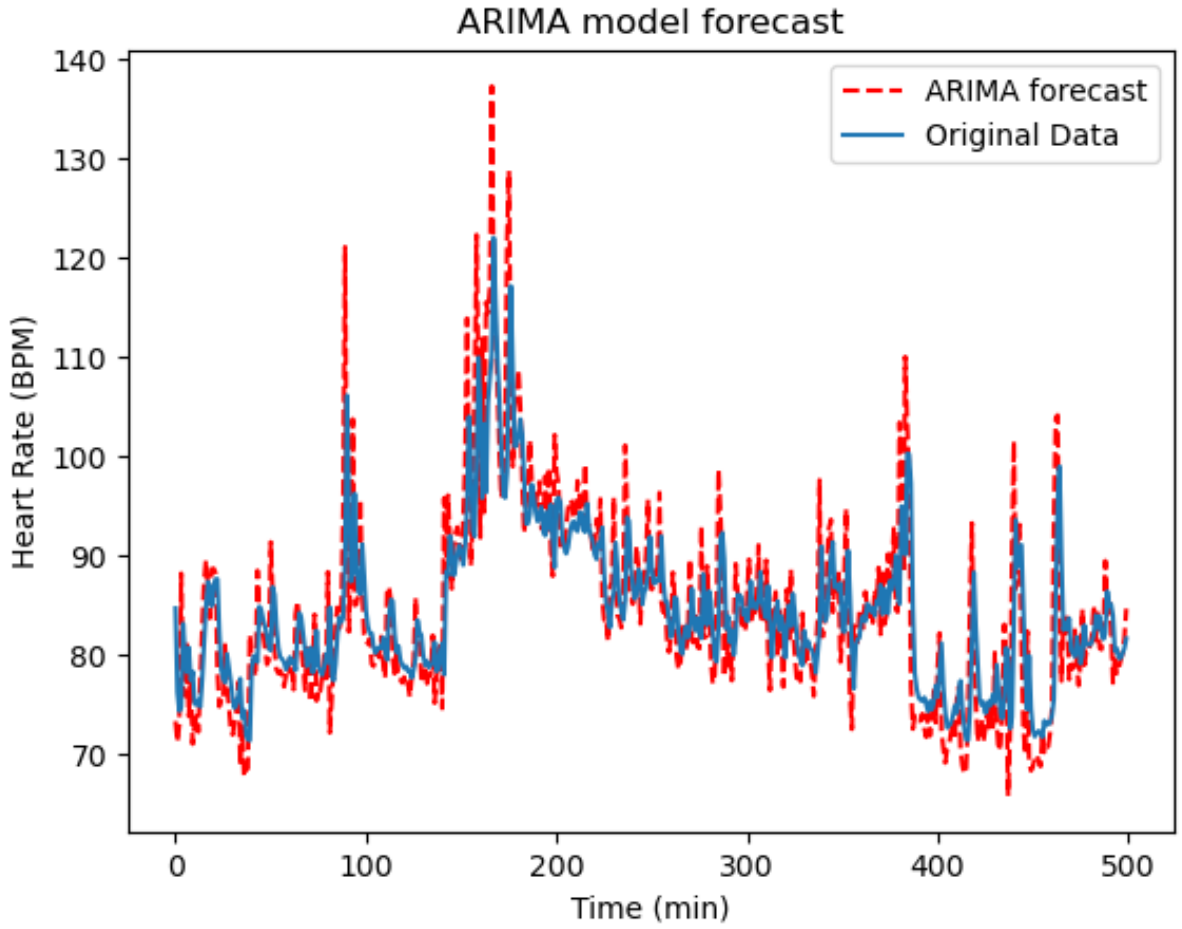

Figure 4: ARIMA model forecast compared to original time series data for the HR of the Lawson dataset using a period not used during the parameter estimation.

From the ac.f. and pac.f. of the residuals of the ARIMA model (the difference between the original series and the predicted one) shown in Figure 5, it is possible to see that no correlation appears to be left, meaning the model fitted was with the right specifications (remember that even for a completely random series some coefficients are expected to be above the threshold by pure chance).

3.2 Multivariate Forecasting

3.3 Deep Learning Approaches

4 Anomaly Detection

4.1 Problem Outline

Anomaly detection consists on the task of identifying patterns in data that are not consistent with its expected behaviour [9]. As such, it is applied in many different areas,

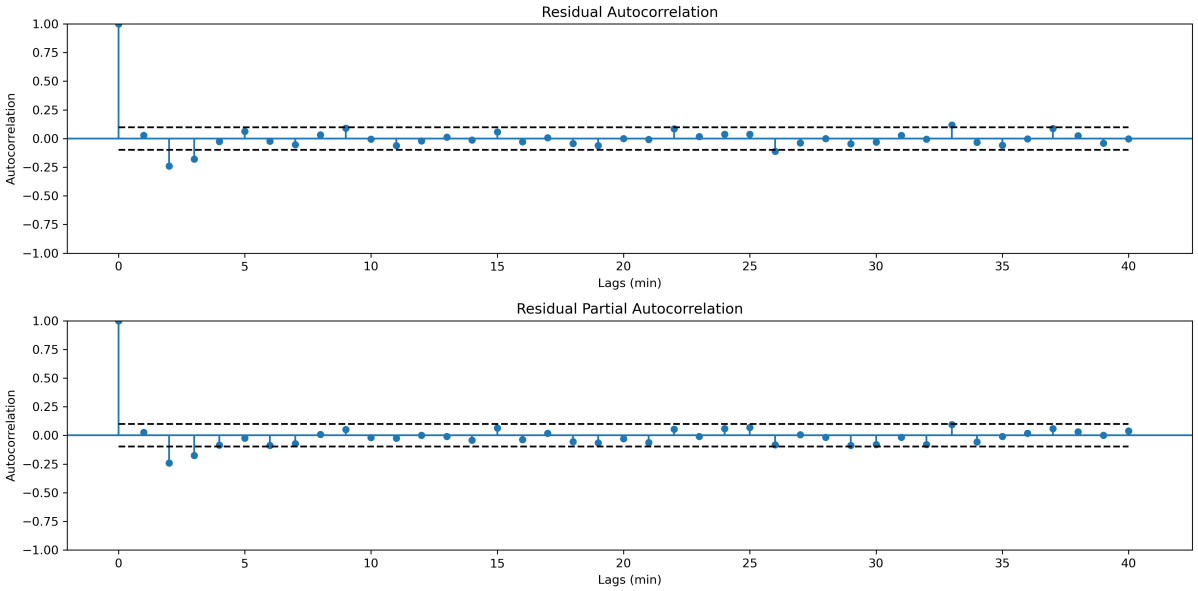

Figure 5: Autocorrelation and Partial Autocorrelation for the residuals from the ARIMA model.

from fraud detection in bank transactions all the way to health monitoring systems, as will be shown later in more details. It might be important first to distinguish between anomalies and novelties in the data, with the difference being that the later has patterns that are incorporated into the data after their first appearance. Although anomalies can intuitively be fairly easy to understand (such as the example shown in Figure 6) there are many challenges related to their actual detection, with some of the most relevant ones being:

- Defining a normal region that actually encompasses every possible normal behaviour;
- Anomalous patterns can change with time;
- Small quantities of labeled data for training;
- Noise in the samples may have similar properties to anomalies;
- The actual description of the anomaly is strictly related to the application at hand.

Anomalies can come in a few different forms. When there is a single instance of the data that has an anomalous behaviour when compared to the rest, it is called a **point anomaly**. When the anomalous behaviour is only classified as so in a specific context, being considered normal in others, it is called **contextual** or **conditional anomaly**. When a collection of related instances is anomalous with respect to the rest of the data (even if those individual instances are not anomalies by themselves) is called a **collective anomaly**.

Another important distinction to make is in the nature of the available data, that can be either labeled or unlabeled, depending on whether the dataset contains the information on the anomalies locations or not. When choosing an anomaly detection algorithm,

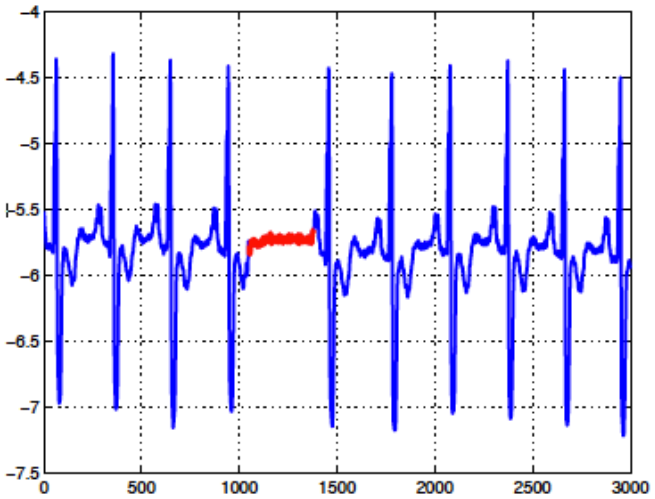

Figure 6: ECG signal showing the data deemed normal in blue and the anomaly marked in red. In this case, the anomaly correspond to an Atrial Premature Contraction(Image adapted from [9])

this characteristic of the data determines whether a supervised, semi-supervised or unsupervised approach will be used. In supervised learning techniques, we have the data labeled between the normal and anomalous classes and usually train predictive models that classify the new data between the labels. One problem with this approach is that, by the nature of anomalies, datasets are highly biased with normal instances, which might affect the final predictions if not taken into account. In semi-supervised techniques, the dataset contains only one of the classes (usually the normal one), and then a model for this classes behaviour and apply it to see whether the instance follows it or not. The last type possible is unsupervised learning, in which the data available has no labels indicating what samples are normal and what are anomalous, and the model itself has to learn that in its training. The unsupervised models are the most widely applicable precisely because of this property of not requiring training data.

4.2 Traditional Approaches

4.3 Chaotic Time Series Anomalies

4.4 Deep Learning Approaches

4.5 Applications in Biosignal Analysis

When dealing with anomaly detection in medical and health monitoring applications, most of the times we are dealing with patient records. In this context, the anomalies present in the data can be due to several reasons, from indications of abnormal patient condition to instrumentation errors. As this project is related to time series analysis, anomaly detection in bio signals within this context will be the focus of analysis, but readers interested in research on other types of data can refer to [10] for a review on deep learning approaches for general medical data and [11] for a review on anomaly detection in medical images.

As was shown already, many datasets, especially when dealing with biosignals in smart wearables and such, are time series in nature.

5 Data Augmentation

5.1 Problem Outline

5.2 Traditional Approaches

5.3 Deep Learning Approaches

6 Methods

6.1 Datasets

6.2 Algorithm Description

6.3 Implementation

7 Discussion

7.1 Results

7.2 Comparison with previous methods

7.3 Future Works

8 Conclusion

References

1. Chatfield, C. & Xing, H. *The Analysis of Time Series: An Introduction with R* ISBN: 9781138066137. https://books.google.co.jp/books?id=9tPOwQEACAAJ (CRC Press, Taylor & Francis Group, 2019).
2. Čepulionis, P. & Lukoševičiūtė, K. Electrocardiogram time series forecasting and optimization using ant colony optimization algorithm. *Mathematical Models in Engineering* **2,** 69–77. ISSN: 2351-5279. https://www.extrica.com/article/17229 (2016).
3. Fan, X., Zhao, Y., Wang, H. & Tsui, K. L. Forecasting one-day-forward wellness conditions for community-dwelling elderly with single lead short electrocardiogram signals. *BMC Medical Informatics and Decision Making* **19,** 285. ISSN: 1472-6947. https://doi.org/10.1186/s12911-019-1012-8 (2019).
4. Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine* **26,** 886–891. ISSN: 1546-170X. https://doi.org/10.1038/s41591-020-0870-z (2020).
5. Nutini, A., Zhang, J., Sohail, A., Arif, R. & Nofal, T. A. Forecasting of the efficiency of monoclonal therapy in the treatment of CoViD-19 induced by the Omicron variant of SARS-CoV2. *Results in Physics* **35,** 105300. ISSN: 2211-3797. https://www.sciencedirect.com/science/article/pii/S2211379722000894 (2022).
6. Schoep, M. E., Nieboer, T. E., van der Zanden, M., Braat, D. D. & Nap, A. W. The impact of menstrual symptoms on everyday life: a survey among 42,879 women. *American Journal of Obstetrics and Gynecology* **220,** 569.e1–569.e7. ISSN: 0002-9378. https://www.sciencedirect.com/science/article/pii/S0002937819304272 (2019).
7. Stuart, T., Hanna, J. & Gutruf, P. Wearable devices for continuous monitoring of biosignals: Challenges and opportunities. en. *APL Bioeng.* **6,** 021502 (June 2022).
8. ZhaoriGetu, H. Prediction of Sports Performance Combined with Deep Learning Model and Analysis of Influencing Factors. *Scientific Programming* **2022,** 4082906. ISSN: 1058-9244. https://doi.org/10.1155/2022/4082906 (2022).
9. Chandola, V., Banerjee, A. & Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv.* **41.** ISSN: 0360-0300. https://doi.org/10.1145/1541880.1541882 (2009).
10. Fernando, T., Gammulle, H., Denman, S., Sridharan, S. & Fookes, C. *Deep Learning for Medical Anomaly Detection – A Survey* 2021. arXiv: 2012.02364 [cs.LG].
11. Taboada-Crispi, A., Sahli, H., Hernandez-Pacheco, D. & Falcon-Ruiz, A. in *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications* 426–446 (IGI Global, 2009).