Visualisierungsinstitut der Informatik
Universität Stuttgart
Allmandring 19
70569 Stuttgart
Germany

Studienarbeit Nr. 2426

Visualisierung multivariater Zeitreihen mit der „Symbolic Aggregate Approximation“

Wolfgang Hinkel

Studiengang:	Informatik
Prüfer:	Prof. Daniel Weißkopf
Betreuer:	Dipl.- Inf. Julian Heinrich
begonnen am:	08.04.2013
beendet am:	07.10.2013
CR-Nummer:	G.2.2

Zusammenfassung

In dieser Studienarbeit über "Visualisierung multivariater Zeitreihen mit der Symbolic Aggregate Approximation" wurde eine grafische Benutzeroberfläche programmiert, die potentiell lange Zeitserien als klassische Liniendiagramme darstellt, mit Hilfe der Symbolic Aggregate Approximation diskretisiert und anschließend als Chaos Game Representation wiedergibt. Der Fokus dabei wurde auf die gemeinsame Analyse zwischen den Liniendiagrammen und der Chaos Game Representation, dem sogenannten "Brushing & Linking", gelegt. Als Testserien kamen EKG-Daten von der MIT-BIH Arrhythmia Database zum Einsatz. Ziel der Tests war herauszufinden, ob man durch Chaos Game Representation - Bitmaps verschiedene Herzerkrankungen darstellen und anhand markanter Merkmale wiedererkennen kann. Bei der Analyse der Bitmaps stellte sich jedoch heraus, dass keine Aussage über Herzerkrankungen getroffen werden können. Die Hauptgründe dafür liegen in der Definition des Sinusrhythmus mit seiner Frequenz und bei den Darstellungen der Ableitungen. Werden jedoch einzelne Herzschläge von Menschen verglichen, sind Unterschiede in der Struktur bzw. in der Häufigkeit der Verteilungen einzelner Kästchen im Chaos Game Representation - Bitmap zu erkennen. Daraus lässt sich aus der angewandten Methode kein praktischer Nutzen für die Erkennung von Herzerkrankungen ableiten.

Inhaltsverzeichnis

Kapitel I: Einleitung

1.1 Problembeschreibung

Zeitreihen sind allgegenwärtig und stellen eine zeitlich geordnete Folge von Messdaten dar. In den verschiedensten Anwendungsbereichen wie zum Beispiel die täglichen Werte des deutschen Aktienindex (DAX) in der Wirtschaft, Elektrokardiogramm - Daten (EKG) und Elektroenzephalografie - Daten (EEG) in der Medizin oder Temperaturdaten in der Meteorologie, werden solche Messdaten generiert und zur weiteren Verarbeitung gespeichert. Ihre hohe Datendimensionalität und der damit verbundene Speicherplatzbedarf stellen heutige Rechnersysteme noch vor Probleme. Anders sieht es bei der Visualisierung solcher Messdaten und den damit verbundenen Zeitreihen aus. Während die Visualisierung einzelner Zeitserien mit Liniendiagrammen vergleichsweise übersichtlich zu gestalten ist, stellen hunderte oder tausende solcher Zeitverläufe in der Visualisierung ein schwieriges Problem dar. Ein vielversprechender Ansatz besteht darin, die Zeitreihen mit Hilfe des Symbolic Aggregate Approximation - Algorithmus (SAX) zu diskretisieren und mit der Chaos Game Representation (CGR) darzustellen. Bisherige Arbeiten (Kumar, Nishanth, Keogh, Lonardi, Ratanamahatana, Wei, 2005) beschränken sich dabei auf die Visualisierung einzelner Zeitreihen und bieten keine Möglichkeit zur gemeinsamen Analyse mit klassischen Liniendiagrammen, dem sogenannten "Brushing & Linking".

1.2 Zielsetzung

In dieser Arbeit wird die Anwendung von SAX und CGR für die Darstellung vieler und potenziell langer Zeitreihen untersucht. Hierzu soll ein graphisches User Interface (GUI) implementiert werden, die das laden von Zeitreihen und die Einstellung der wichtigsten Parameter für SAX erlaubt und durch CGR dargestellt wird. Um das Verständnis dieser Visualisierung und die darin erkennbaren Muster zu gewährleisten, werden Zeitreihen auch in klassischen Liniendiagrammen dargestellt. Interaktionsmöglichkeiten zwischen CGR und Liniendiagrammen, wie zum Beispiel Zeitpunkt- oder Feldauswahl werden ebenfalls implementiert.

Eine kurze Analyse der Skalierbarkeit des Systems bezüglich der Länge der Zeitreihen und ein Test der Implementierung mit verschiedenen Zeitreihen werden am Ende vorgenommen.

1.3 Aufbau der Arbeit

In Kapitel II werden die Grundlagen für die Studienarbeit erklärt. Anschließend folgt eine ausführliche Programmbeschreibung, die im Kapitel III beschrieben wird. Anwendungstest mit EKG-Daten werden im Kapitel IV dargestellt. Am Ende der Arbeit werden die Ergebnisse noch einmal kurz zusammengefasst und der Ausblick skizziert.

Kapitel II: Grundlagen

Um ein Verständnis für die Darstellungsformen und dessen Strukturen zu erhalten, sind Definitionen für eine genaue Beschreibung von Inhalten unumgänglich. In diesem Abschnitt werden Strukturen erklärt und Definitionen festgelegt, die in dieser Studienarbeit zur Anwendung kommen.

2.1 Zeitreihen und Zeitserien

Eine Zeitreihe T bezeichnet eine zeitlich geordnete Folge von Beobachtungen d einer Größe, wohin gegen eine Zeitserie verschiedene Zeitreihen der gleichen Beobachtung sind. Würde man täglich die Durchschnittstemperatur in Stuttgart über ein ganzes Jahr lang hin messen, so würde man 365 diskrete Temperaturwerte erhalten. Diese Temperaturwerte stellen eine Zeitreihe dar. Liegen mehrere Zeitreihen verschiedener Jahre vor, sie müssen nicht zusammenhängend sein, spricht man von Zeitserien. Reelle Zeitreihen besitzen stets endliche Indexmengen bzw. Beobachtungszeiträume. Die aufgenommenen Messdaten können kontinuierlich oder diskontinuierlich sein. Die Merkmalsausprägungen sind meistens metrisch. Durchaus kommen auch andere Ausprägungen, zum Beispiel als Index oder als qualitative Aussage wie ”gut”, “befriedigend” oder “schlecht”, die in einen Index transformiert wurden, vor. Eine formalere Definition (Springer Gabler Verlag, 2013) für Zeitreihen lautet:

Eine Zeitreihe $T = ((t_1, d_1), ..., (t_n, d_n))$ ist eine Folge von n-Tupeln, bestehend aus einem Zeitstempel t_i und einem dazugehörigen eindeutigen Messwert d_i, mit $i \in 1, 2, ..., n$ und $n \in \mathbb{N}$. Der Zeitstempel t_i kann sich aus verschiedenen Zeitpunkten oder Zeitintervallen zusammensetzen. Zeitpunkte erhält man dann, falls das zu messende Merkmal eine Bestandsmasse ist, Zeitintervalle, wenn es eine Bewegungsmasse ist.

- **Die Bestandsmasse**

  Eine Bestandsmasse (Springer Gabler Verlag, 2013) ist eine Anzahl statistischer Einheiten, die über einen gewissen Zeitraum gemeinsam in einem Bestand

verweilen. Beispielsweise wir die Bevölkerung immer zu einem konkreten Zeitpunkt gemessen oder geschätzt; sie ist ein Bestandsmasse.

- **Die Bewegungsmasse**

  Eine Bewegungsmasse (Springer Gabler Verlag, 2013) stellt zeitpunktbezogene, zustandsändernde Ereignisse dar und kann nur über einen bestimmten Zeitraum hinweg ihren Umfang nach erfasst werden. Beispielsweise ist die Gesamtheit der Personen, die innerhalb eines Jahres in einer Region arbeitslos werden, eine Bewegungsmasse; nicht jedoch die Gesamtheit der Arbeitslosen dieser Region zu einem bestimmten Zeitpunkt.

Zeitreihen werden im klassischen Komponentenmodell analysiert. Das Ziel einer Zeitreihenanalyse ist, einen Trend als Funktion der Zeit oder einen Effekt zu erkennen und eine Prognose über den zukünftigen Verlauf vorherzusagen. Auch die Frage nach wechselseitigen Beziehungen oder Abhängigkeiten zeitlich unterschiedlicher Beobachtungen ist von Bedeutung.

2.1.1 Das klassische Komponentenmodell

Eine Zeitreihe besteht aus Komponenten, die in einem Komponentenmodell zusammengefasst sind. Es gibt die systematischen Komponenten und die irreguläre Komponente. Die systematischen Komponenten sind Trend, Konjunkturkomponente und Saisonkomponente.

1. Der ***Trend*** = T_t erfasst langfristige Entwicklungstendenzen im Mittel einer Zeitreihe. Der Verlauf der Trendkomponente ist bedingt durch die langfristigen Ursachen monoton wachsend oder monoton fallend. Sie wird mittels einer linearen oder nicht linearen Regression mit der Zeit t als unabhängige und y als abhängige Variable berechnet.
2. Die ***Konjunkturkomponente*** = K_t ist eine zyklische, mehrjährige aber nicht notwendigerweise regelmäßige Schwankung. Sie nimmt einen wellenförmigen Verlauf aufgrund sich stetig, aber langsam ändernder Einflüsse an.
3. Die ***Saisonkomponente*** = S_t ist eine jahreszeitlich bedingte Schwankungskomponente, die sich nahezu unverändert jedes Jahr wiederholt. Ihr Verlauf ist

wie die Konjunkturkomponente wellenförmig aufgrund des periodischen Zeiteinflusses auf die Komponente.

4. Die ***irreguläre*** oder auch ***Restkomponente*** = I_t genannt, besteht aus unvorhersehbare, nicht regelmäßig wiederkehrende und in den übrigen Komponenten nicht enthaltene Einflüsse und Störungen.

Der Trend und die Konjunkturkomponente werden häufig in der Literatur als „glatte Komponente“ beschrieben ($G_t = T_t + K_t$). Konjunktur- und Saisonkomponente stellen die „zyklische Komponente“ dar. Die Ermittlung der zyklischen Komponente erfolgt dadurch, dass zuerst die glatte Komponente, also die Überlagerung von Trend und zyklischer Komponente geschätzt wird. Dies geschieht mit Hilfe des gleitenden Mittelwertes (siehe Anhang A.1). Die zyklische Komponente ergibt sich dann durch Subtraktion des Trends von der glatten Komponente.

Das Komponentenmodell wird überwiegend für wirtschaftswissenschaftliche Zwecke genutzt. Es ist aber auch möglich, Zeitreihen aus anderen Bereichen zu analysieren, wie zum Beispiel EKG- oder EEG-Datenreihen aus der Medizin. Mehrjährige oder auch jahreszeitliche Schwankungen wird man vergebens suchen, aber zyklische Schwankungen bzw. periodische Verläufe gibt es auch hier.

2.1.2 Visualisierung von Zeitreihen durch Diagramme

Diagramme sind ein wichtiges Element in der Darstellung von sachlich und wissenschaftlichen Zusammenhängen in schriftlichen Berichten und in Präsentationen. Komplexe Vorgänge, mathematische Zusammenhänge, Statistiken, Messergebnisse, Abläufe und vieles mehr, die nur mit schwer verständlichen Texten zu beschreiben wären, können mit einem Blick erfasst werden.
Zeitreihen können durch verschiedene Diagrammtypen dargestellt werden. Die Zielsetzung, was dargestellt werden soll und kann, schränkt die Benutzung der Diagrammtypen ein. Dabei ist auf die richtigen Ausprägungen des darstellenden Merkmals zu achten. Im Nachfolgenden werden verschiedene Diagrammtypen für Zeitreihen vorgestellt.

2.1.2.1 Das Punktdiagramm

Das Punktdiagramm zeigt Zeitreihen als Mengen von Punkten an. Die Werte werden durch die Position der Punkte im Diagrammraum dargestellt. Jeder einzelne Punkt kann durch eine gerade Linie (Gerade) oder einer geglätteten Linie (Kurve) verbunden werden. Punktediagramme eignen sich hervorragend, um große Mengen verwandter Daten in einem einzelnen Diagramm anzuzeigen. In Punktdiagrammen wird auf der X-Achse ein numerischer Wert und auf der Y-Achse eine andere Kennzahl dargestellt, sodass die Beziehung sich zwischen den beiden Werten für alle Elemente im Diagramm leicht erkennen lässt. Die Abbildungen 1, 2 und 3 zeigen drei verschiedene Darstellungstypen von Punktdiagrammen.

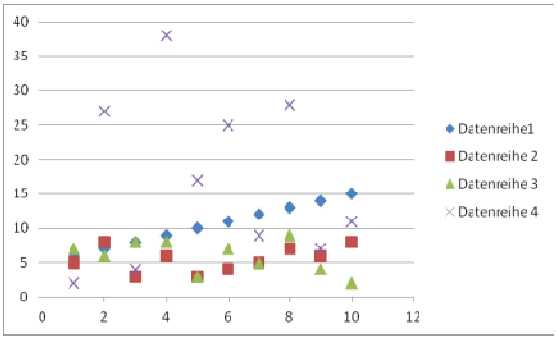


Abbildung 1: Beispiel für ein Punktdiagramm ohne Verbindungslinien von Messwerten

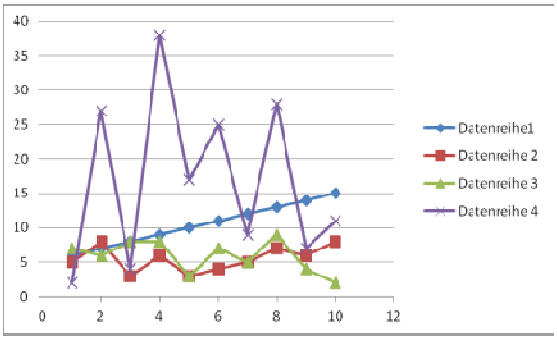


Abbildung 2: Beispiel für ein Punktdiagramm mit geraden Verbindungslinien zwischen den Messwerten

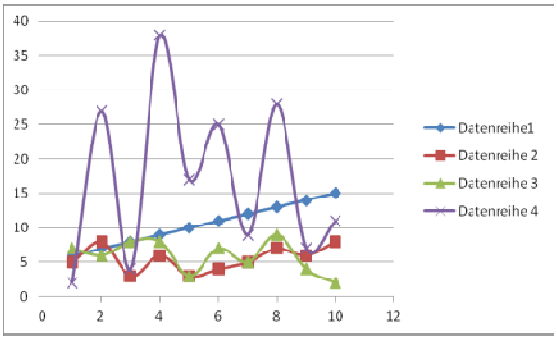


Abbildung 3: : Beispiel für ein Punktdiagramm mit geglätteten Verbindungslinien zwischen den Messwerten

2.1.2.2 Das Liniendiagramm

Ein Liniendiagramm ist ein Diagrammtyp, bei dem einzelne Punkte einer Datenreihe oder Zeitreihe mit Hilfe von Liniensegmenten verbunden werden. Die Werte werden durch die Höhe der Punkte, gemessen an der Y-Achse, dargestellt. Sie ist einem Punktdiagramm sehr ähnlich, unterscheiden sich jedoch in einem wesentlichen Merkmal. Liniendiagramme haben nur eine Werteachse im Gegensatz zum Punktdiagramm mit zwei Werteachsen. Die zweite Achse im Liniendiagramm ist die Rubrikenachse. Liniendiagramme dienen in der Regel zum Vergleich von Werten über die Zeit und eignen sich auch besonders, um Entwicklungen und Trends zu veranschaulichen. Die Abbildung 4 zeigt ein solches Liniendiagramm.

Abbildung 4: DAX-Chartverlauf vom 01.01.2011 bis 01.04.2011; mit X-Achse als Zeit und Y-Achse als Preis.

2.1.2.3 Das Säulendiagramm

In einem Säulendiagramm werden Häufigkeitsverteilungen diskreter Werte und Reihengruppen als Sätze dargestellt, indem auf der X-Achse senkrecht stehende, nicht aneinander grenzende Säulen gleicher Breite dargestellt und nach einer Kategorie gruppiert werden. Die Werte werden durch die Höhe der Säulen, gemessen an der Y-Achse, dargestellt. Die Kategorien werden an der X-Achse angezeigt. Säulendiagramme werden normalerweise zum Vergleich von Werten in verschiedenen Kategorien verwendet. Die Abbildung 5 zeigt ein solches Säulendiagramm.

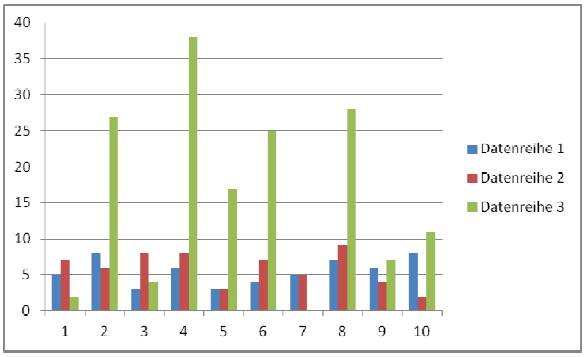


Abbildung 5: Beispiel eines Säulendiagramms

2.1.2.4 Das Balkendiagramm

Das Balkendiagramm ist ein sehr häufig genutzter Diagrammtyp und ist dem Säulendiagramm sehr ähnlich. Es stellt die Datenreihen, im Gegensatz zum Säulendiagramm, durch waagerecht liegende Balken dar. Es eignet sich sehr gut zur Veranschaulichung von Rangfolgen und zum Vergleich von Werten in verschiedenen Kategorien. Die Abbildung 6 zeigt ein solches Balkendiagramm.

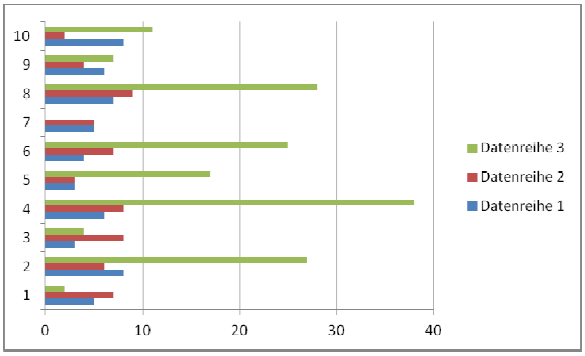


Abbildung 6: Beispiel eines Balkendiagramms

2.1.2.5 Das Kursdiagramm

Ein Kursdiagramm zeigt Reihen als einen Satz von Linien mit Markierungen für den höchsten, den niedrigsten, den Schluss- und den Eröffnungswert an. Die Werte werden durch die Höhe der Markierung, gemessen an der Y-Achse, dargestellt. (Microsoft Press, 2005) Kategorien werden an der X-Achse angezeigt. Kursdiagramme eignen sich besonders gut, um Schwankungen innerhalb eines Zeitraums deutlich zu machen. Die Abbildung 7 zeigt ein solches Kursdiagramm.

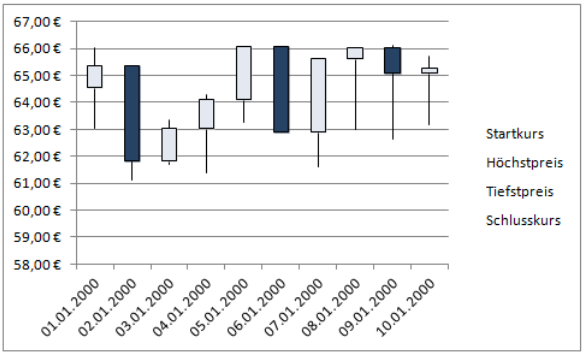


Abbildung 7: Beispiel eines Kursdiagramms einer fiktiven Handelsware

2.1.2.6 Das Netzdiagramm

Ein Netzdiagramm oder auch Spinnennetzdiagramm genannt, wird zur grafische Darstellung von Werten mehrerer, gleichwertiger Kategorien in einer Spinnennetzform genutzt. Es eignet sich besonders gut zum Visualisieren von Evaluationen für zuvor festgelegte Kriterien zweier bzw. mehrerer Serien. Für jede Kategorie gibt es eine Achse, wobei mindestens 3 Kategorien existieren müssen und für alle Achsen die

gleiche Orientierung gilt. Alle Achsen werden kreisförmig innerhalb von 360 Grad gleichmäßig angeordnet. Die Werte jeder Serie werden mit Linien verbunden. Nutzt man mehrere Serien werden diese verschieden farblich dargestellt. Die Abbildung 8 zeigt ein solches Netzdiagramm.

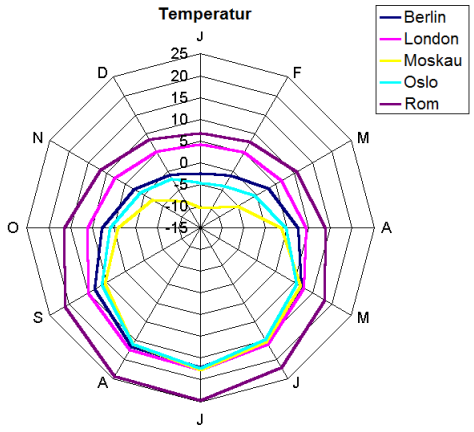


Abbildung 8: Netzdiagramm für die Durchschnittstemperaturen verschiedener Städte. Abbildung aus (ZUM Internet e.V., 1995)

2.1 PAA - Piecewise Aggregate Approximation

Die **P**iecewise **A**ggregate **A**pproximation (PAA), oder auch Piecewise Constant Approximation genannt ist ein sehr einfaches Verfahren zur Datendimensionsreduktion von Zeitreihen. Durch das Einteilen von Zeitreihen in gleichgroße Segmentabschnitte und die Berechnung der einzelnen Mittelwerte für jeden Abschnitt, können Zeitreihen sehr genau approximiert werden. Je kürzer die Segmentabschnitte, desto genauer die Approximierung, aber umso geringer die Datendimensionsreduktion. Die Mittelwerte können schnell berechnet werden und ermöglichen die effiziente Indexierung (Keogh & Pazzan, 2000), (Keogh E. J., 2001) .

- **Der Mittelwert**

  Der Mittelwert oder auch arithmetisches Mittel genannt ist der Quotient aus der Summe aller beobachteten Werte und der Anzahl der Werte. Das arithmetische Mittel einer Menge von n Werten $x_1, x_2, ..., x_n$ ist somit definiert als

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

2.2.1 PAA - Repräsentation

Eine Zeitreihe X der Länge n wird im m-dimensionalen Vektorraum repräsentiert durch den Vektor $T(X) = (\overline{x_1}, \overline{x_2}, ..., \overline{x_{i-1}}, \overline{x_i}, ..., \overline{x_m})$ mit $m \leq n$, $m, n \in \mathbb{N}$ und $\frac{n}{m} \in \mathbb{Z}$. Der Mittelwert $\overline{x_i}$ wird durch die nachfolgende Gleichung berechnet.

$$\overline{x_i} = \frac{m}{n} \sum_{j=\frac{n}{m}(i-1)+1}^{\frac{n}{m}i} x_j$$

PAA minimiert somit die Datendimension n, in dem es die Zeitreihe in m gleich große, nicht überlappende Segmentabschnitte geteilt wird. Über jeden Segmentabschnitt wird ein Mittelwert berechnet, der die Zeitreihe sehr genau approximieren kann. Hierbei ist die Länge der Segmentabschnitte von Bedeutung. Je länger die Segmentabschnitte gewählt werden, desto niedriger die resultierende Approximation. Der neu entstandene Vektor, der aus den Mittelwerten besteht, stellt die reduzierte Repräsentation der Zeitreihe dar und wird als PAA-Signatur, der aus einzelnen Koeffizienten c_i besteht, bezeichnet. Er kann in $O(n)$ berechnet werden. Die Abbildungen 9 - 11 zeigen eine solche Transformation.

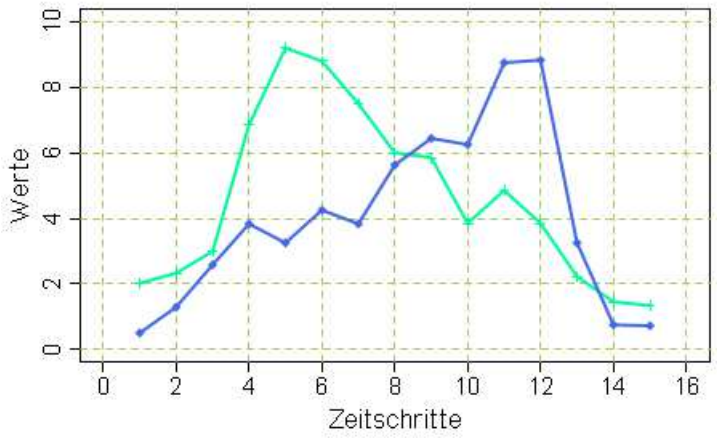


Abbildung 9: Zwei Zeitserien als Liniendiagramme, die jeweils aus 15 Punkten bestehen. Abbildung aus (Google, 2012)

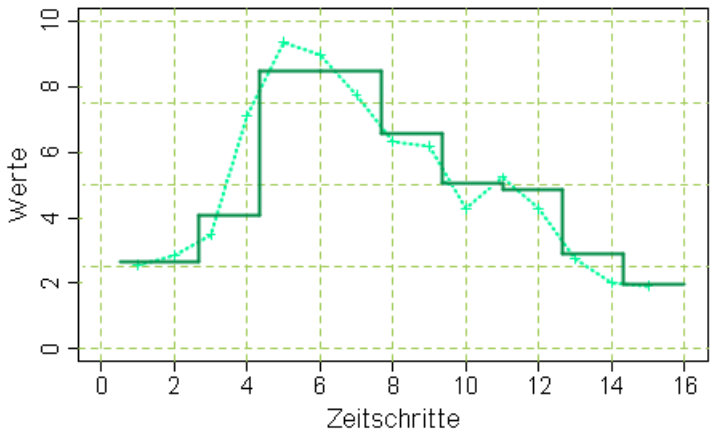


Abbildung 10: PAA Beispiel der ersten Zeitserie aus Abbildung 9 mit m = 9 Segmentstücken

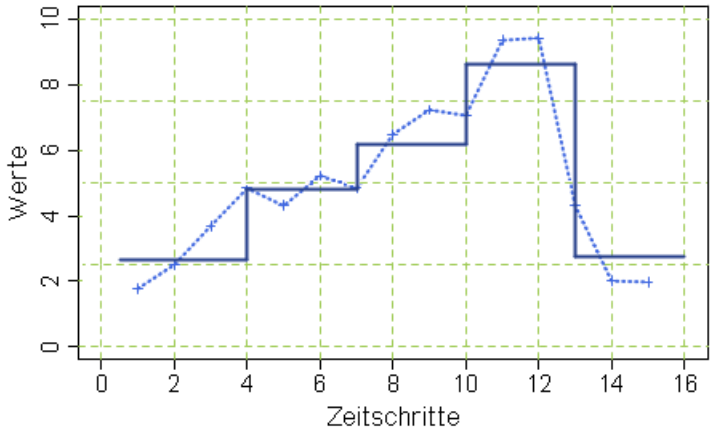


Abbildung 11: PAA Beispiel der zweiten Zeitserie aus Abbildung 9 mit m = 5 Segmentstücken

2.2.2 Distanzmaß zweier Zeitreihen in der PAA-Repräsentation

In vielen Bereichen der Naturwissenschaften interessiert man sich für die Messung der Ähnlichkeit zwischen verschiedenen Objekten und definiert dazu sogenannte Ähnlichkeits- oder auch Distanzmaße. Distanzmaße werden in der Regel für metrisch skalierte Variablen genutzt, während Ähnlichkeitsmaße für nominal oder ordinal skalierte Variablen genutzt werden. Möchte man für zwei Zeitreihen X und Y feststellen, wie ähnlich beide sind, wird das Distanzmaß angewendet.

2.2.2.1 Das Distanzmaß

Sei $I = \{1, 2, ..., N\}$ eine endliche Menge. Eine Funktion $d : I \times I \rightarrow \mathbb{R}$ heißt Distanzmaß (Hartung & Elpelt, 1984) oder Distanzfunktion, falls für alle $i, j \in I$ gilt:

$$d(i,j) = d(j,i)$$
$$d(i,j) \geq 0$$
$$d(i,j) = 0 \Leftrightarrow i = j$$

Die Funktionswerte $d(i,j)$ lassen sich zu einer symmetrischen $N \times N$- Matrix $\left(d(i,j)\right)_{i,j}$ anordnen und heißt Distanzmatrix.

In der Mathematik sind verschiedene Distanzfunktionen bzw. Metriken bekannt.

2.2.2.2 Die Minkowski-Metrik

Ein allgemein gebräuchliches Distanzmaß in der Mathematik ist die Minkowski-Metrik, benannt nach dem deutschen Mathematiker und Physiker Hermann Minkowski, mit

$$d_p(x_i, x_j) = \left(\sum_{i=1}^{n} \left|x_{ik} - x_{jk}\right|^p\right)^{\frac{1}{p}} = \sqrt[p]{\sum_{i=1}^{n} \left|x_{ik} - x_{jk}\right|^p},$$

wobei p für den Metrikparameter der Minkowski-Metrik steht. Der Parameter p ist eine Art „Gewichtungsfaktor“. Mit größer werdenden p kommt es zu einer immer stärkeren Gewichtung großer Distanzen und zu einer schwächeren Gewichtung kleiner Distanzen. x_{ik} und x_{jk} stehen für die konkreten Ausprägungen der Objekte i und j auf der k-ten Variable. Um Abweichungen nach unten und oben zu betrachten, wird der Betrag genommen. Das Endresultat nach eingesetzten Werten ist eine konkrete Kennzahl für den Abstand zwischen zwei Punkten bzw. zwei Objekten. Diese Kennzahl lässt sich paarweise für alle Objekte bestimmen und zu einer, wie bereits erwähnt, Distanzmatrix zusammenfassen.

2.2.2.3 Das Euklidische Distanzmaß

Das Euklidische Distanzmaß ist ein Spezialfall der Minkowski-Metrik, wobei der Metrikparameter $q = 2$ gesetzt wird. Der Euklidischen Abstand ergibt sich dann mit:

$$d_2(x_i, x_j) = \sqrt[2]{\sum_{i=1}^{n} \left|x_{ik} - x_{jk}\right|^2}$$

Als Kennzahl für den Abstand zwischen zwei Objekten auf allen relevanten Variablen erhält man eine Kennzahl als Teilmenge von $\mathbb{R}^+$. Weil die Distanzen quadriert werden, können keine Werte kleiner Null vorkommen.

2.2.2.4 Die City-Block-Metrik

Ein weiterer Spezialfall der Minkowski-Metrik ist die City-Block-Metrik oder auch Manhattan-Distanz genannt. Der Metrikparameter wird in der City-Block-Metrik auf $q = 1$ gesetzt. Die Distanz zwischen zwei Punkten wird als die Summe der absoluten Differenzen ihrer Einzelkoordinaten definiert, mit

$$d_1\left(x_i, x_j\right) = \sum_{i=1}^{n} \left| x_{ik} - x_{jk} \right| .$$

2.2.2.5 Die Mahalanobis Distanz

Die Mahalanobis Distanz ist ein Distanzmaß zwischen zwei Punkten in einem Vektorraum mit:

$$d_M\left(x_i, x_j\right) = \left(x_i, x_j\right) \cdot \mathrm{F}^{-1} \cdot \left(x_i, x_j\right)^{\mathrm{T}}$$

x_i und x_j sind Vektoren von Koordinaten zweier Punkten und F ist eine Kovarianzmatrix. Die Kovarianz stellt einen monotonen Zusammenhang zweier Zufallsvariablen mit gemeinsamer Wahrscheinlichkeitsverteilung dar. Damit ist eine Kovarianzmatrix eine Matrix aller paarweisen Kovarianzen der Elemente eines Zufallsvektors. Die Kovarianzmatrix enthält Informationen über die Streuung eines Zufallsvektors und über Korrelationen zwischen dessen Komponenten.

Dieses Distanzmaß nutzt man, wenn lineare Korrelationen der Variablen untereinander vorliegen, um diese heraus zu rechnen. Zuvor korrelierte Merkmale werden dabei erst durch Datentransformation modifiziert, so dass Unkorreliertheit entsteht. Danach wird die quadrierte Euklidische Distanz berechnet, die der Mahalonobis Distanz entspricht.

2.2.2.6 Weitere Distanzmaße

Weitere Distanzmaße sind zum Beispiel das Pearson-Distanzmaß, welches ähnlich dem Euklidischen Distanzmaß die Standardabweichung in seine Berechnung mit aufnimmt

oder das Gower-Distanzmaß, ähnlich der City-Block-Metrik, welches die Spannweite (Distanz zwischen dem größten und dem kleinsten Messwert) mit berücksichtigt.

2.2.3 Euklidische Distanz und die PAA-Repräsentation

Das wohl am häufigsten verwendete Distanzmaß ist die Euklidische Distanz. Die Bedeutung ist wohl darauf zurückzuführen, dass sie Inhaltlich dem im Alltag verwendeten Abstandbegriff entspricht. Die geometrische Distanz zweier Objekte im Raum wird hier auf Basis der kürzesten direkten Entfernung zueinander bestimmt. Neben Ihrer Anschaulichkeit hat die Euklidische Distanz den Vorteil, dass die daraus abgeleiteten Konfigurationen jeder Zeit orthogonal um das Koordinatenkreuz rotiert oder an den Achsen gespiegelt werden können. Die Distanzen bleiben dadurch unverändert. (Sturm, Hans-Jörg, Markenfit und Markenwirkung, 2012). Dies machen sich Yi & Faloutsos und Keogh et al. zu Nutze und wenden dies auf Zeitserien an. Sie beweisen für die Euklidische Distanz, dass das Distanzmaß d_{PAA} zweier Zeitreihen X und Y in der PAA-Repräsentation das Korrektheitskriterium "Lower Bounding Lemma" (untere Schranke) einhält, mit

$$d_{PAA}(\overline{X},\overline{Y}) \equiv \sqrt{\frac{n}{m}} \cdot \sqrt{\sum_{i=1}^{m} (\overline{x}_i - \overline{y}_i)^2} \leq d(X,Y)$$

- **Lower Bounding Lemma**

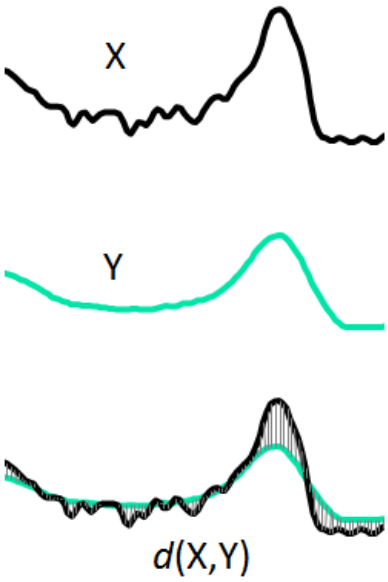
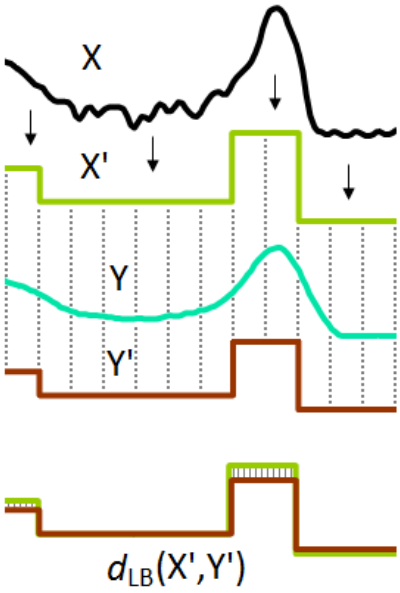
  Seien O_1 und O_2 zwei Objekte und d die Euklidische Distanz. Dann gilt: Wenn der Abstand zweier ähnlicher Objekte kleiner oder gleich ε im Originalraum ist, dann ist er auch kleiner oder gleich ε im Merkmalsraum, mit

  $$d_{Merkmal}\left(M\left(\mathrm{O}_1\right), M\left(\mathrm{O}_2\right)\right) \leq d\left(\mathrm{O}_1, \mathrm{O}_2\right) \leq \varepsilon .$$

Anders ausgedrückt bedeutet das Lower Bounding Lemma, wenn man den Abstand zweier Objekte im Originalraum vergleicht und den Abstand der Approximierten beider Objekte vergleicht (in diesem Fall die PAA-Objekte), so ist der Abstand zwischen den Approximierten kleiner oder gleich dem Abstand der Objekte im Originalraum.

Tabelle 1 zeigt noch einmal anschaulich den Zusammenhang zwischen der Euklidischen Distanz und der Lower Bounding Distanz.

Tabelle 1: Zusammenhang zwischen Euklidischer Distanz und Lower Bounding Distanz

Euklidische Distanz $d(X,Y)$	**Lower Bounding Distanz** $d_{LB}(X',Y')$
	
$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	$d_{PAA}(\bar{X},\bar{Y}) \equiv \sqrt{\frac{n}{m}} \cdot \sqrt{\sum_{i=1}^{m}(\bar{x}_i - \bar{y}_i)^2}$

2.3 SAX - Symbolic Aggregate approXimation

Auf dem Gebiet des Data Mining (Siehe Anhang A.4) wurden viele High-Level-Repräsentationen für Zeitserien vorgeschlagen, die beispielsweise die Fourier-Transformation, Wavelets (Siehe Anhang A.5) oder auch verschiedene polynomielle Modelle nutzen. Die Abbildung 12 illustriert einen hierarchischen Ansatz unterschiedlicher Repräsentationen von Zeitserien. Viele Wissenschaftler haben sich ebenfalls mit Symboldarstellungen von Zeitserien beschäftigt, wobei keiner der Repräsentationen potenziell Forschern erlaubt, die Fülle an Datenstrukturen und Algorithmen bei der Textverarbeitung oder in der Bioinformatik zu nutzen. In den letzten Jahrzehnten sind viele solcher Symboldarstellungen von Zeitreihen veröffentlicht worden, wobei alle Symboldarstellungen mit zwei grundsätzlichen

Schwachpunkten versehen sind. Erstens ist die Datendimensionalität der Symboldarstellung dieselbe wie seine Ursprungsdaten. Dies hat zur Folge, dass alle Data Mining - Algorithmen eine schlechte Skalierbarkeit bei hoher Datendimensionalität aufweisen. Zweitens, obwohl Abstandsmaße auf Symboldarstellungen definiert wurden, haben diese eine geringe Korrelation mit dem Abstandsmaß der ursprünglichen Zeitreihe. Ein neuartiger Ansatz ist die Symbolic Aggregate approXimation. SAX ermöglicht die Datendimensionalität zu reduzieren und erlaubt auch ein Distanzmaß zu definieren, das als Untergrenze das Abstandsmaß der originalen Zeitserie hat (Lower Bounding Lemma). SAX wurde erstmals von Lin at al. vorgestellt und wandelt Daten von Zeitreihen in symbolische Strings um. Damit ist es das erste symbolische Dimensionsreduktionsverfahren für Zeitreihen, welches auf einem Distanzmaß definiert wurde. Basierend auf dem zuvor beschriebenen PAA-Verfahren nutzt SAX die PAA-Signatur, um diese zu diskretisieren. Darüber hinaus eröffnet die Verwendung einer Symboldarstellung die Tür zu bestehenden Datenstrukturen und Stringmanipulationsalgorithmen in der Informatik wie Hashing, reguläre Ausdrücke, Pattern Matching, Suffix-Bäumen und viele andere mehr (Lin, Keogh, Wei, & Lonardi, 2006). SAX hat innerhalb kurzer Zeit auch Spuren in der Industrie und anderen Wissenschaftsbereichen hinterlassen. Beispielsweise analysiert Androulakis komplexe kinetische Mechanismen in Anlehnung an SAX (Androulakis, 2005). Dr. Amy McGovern von der Universität Oklahoma leitet ein Projekt basierend auf einem dynamisch relationalen Modelle für eine verbesserte Vorhersage gefährlicher Wetterbedingungen. Sie nutzt eine diskrete Repräsentation von meteorologischen Realwertdaten. Dabei nutzt sie SAX zur Erstellung von diskreten Daten aus kontinuierlichen (McGovern, Kruger, Rosendahl, & Droegemeier, 2006). Ein weiteres Beispiel ist die Verwendung von SAX und Zufallsprojektionen, um Motive in telemedizinischen Zeitreihen zu finden (Duchene, Garbay, & Rialle, 2004), (Silvent, Carry, & Dojat, 2003), (Silvent, Dojat, & Garbay, 2004). Unter anderen nutzt auch der Telekommunikationskonzern AT&T SAX zur Erkennung von Anomalien in Zeitserien für sehr große Datenmengen (Riehl, 2010).

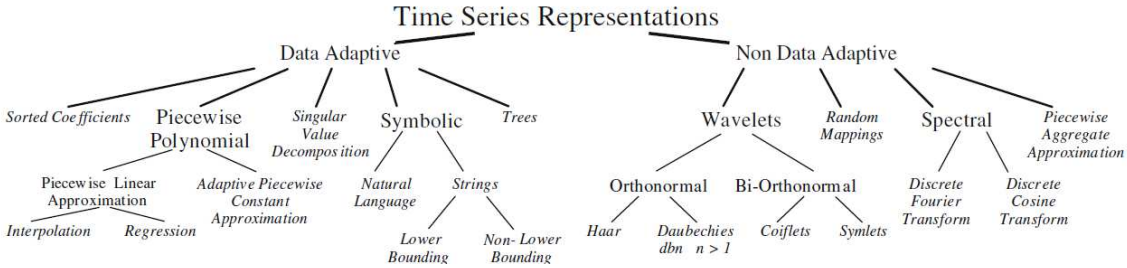


Abbildung 12: Hierarchie verschiedener Zeitreihendarstellungen. Die Blattknoten sind die eigentlichen Darstellungen und die internen Knoten sind die Klassifizierungen der Annäherungen (Lin, Keogh, Wei, & Lonardi, 2006)

2.3.1 SAX Transformation

SAX transformiert eine Zeitreihe X der Länge n in einen String beliebiger Länge w, $w \ll n$ und einem Alphabet $\sum$ der Größe $l > 2$.

Der SAX-Algorithmus setzt sich aus zwei Schritten zusammen. Im ersten Schritt wird die „originale“ Zeitreihe in die PAA-Darstellung überführt und diese Zwischendarstellung wird in eine Zeichenfolge, während des zweiten Schritts, umgewandelt. Die Überführung der PAA-Koeffizienten in Buchstaben ist in der Zeit O(n) effizient durchführbar. Die Verwendung von PAA im ersten Schritt bringt, wie bereits oben erwähnt, den Vorteil einer einfachen und effizienten Datendimensionsreduktion.

Die Diskretisierung der PAA-Signatur einer Zeitreihe in SAX kann mit einer Tonleiter aus der Musik verglichen werden. Die in ein bestimmtes Intervall fallenden Koeffizienten der PAA-Signatur werden einem Buchstaben bzw. einem Ton zugeordnet. Die Intervalle werden mit Hilfe einer gaußschen Normalverteilung bestimmt. Dabei wird die Gaußnormalverteilung so eingeteilt, dass in jedes Intervall die gleiche Fläche fällt. Abbildung 13 verdeutlicht dies. Eine Normierung bei der Nutzung mehrerer Zeitreihen mittels Z-Standardisierung ist vor der Transformation notwendig, da normierte Zeitreihen der Normalverteilung folgen und dies eine der Grundannahmen für die Effizienz des Verfahrens ist. Folgen Zeitreihen nicht der Normalverteilung, würden die PAA-Koeffizienten nicht gleichmäßig auf die Buchstaben verteilt werden. Ein direkter Vergleich zweier Zeitreihen wäre damit ausgeschlossen. [Schäfer Diplom]

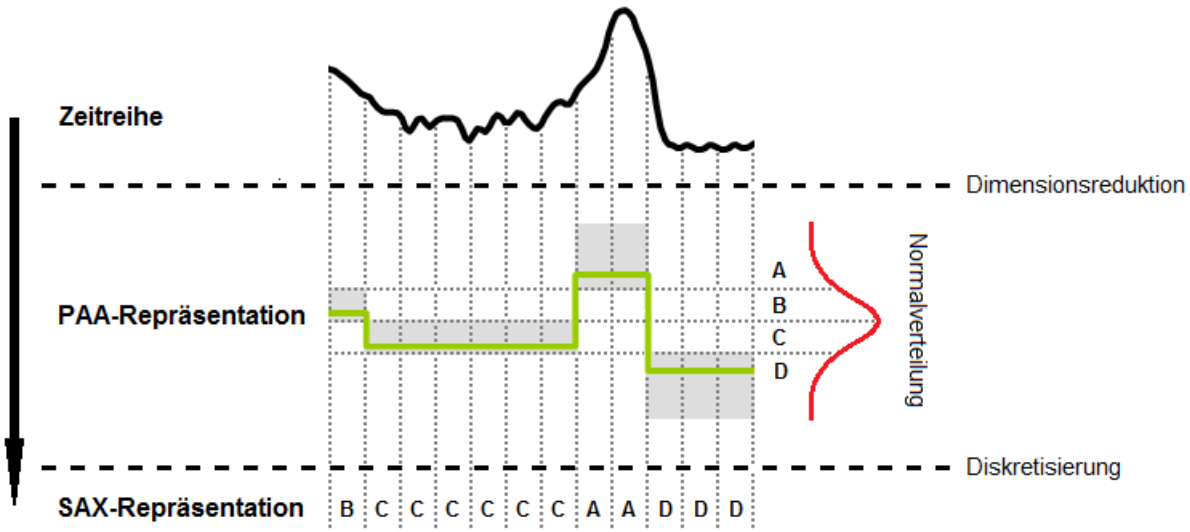


Abbildung 13: SAX Transformation: Diskretisierung der PAA-Koeffizienten mittels Normalverteilung

2.3.1.1 Gaußsche Normalverteilung

Die Gaußsche Normalverteilung ist eine Wahrscheinlichkeitsverteilung, benannt nach Carl Friedrich Gauß. Die Parameter der Normalverteilung sind der Erwartungswert μ und die Varianz σ^2. Die Varianz σ^2 ist die Quadrierte Standartabweichung σ. Mithilfe der Standardtransformation können Normalverteilungen mit beliebiger Parameterlage in die Standardnormalverteilung überführt werden ($\mu = 0$ *und* $\sigma^2 = 1$). Bei einer grafischer Darstellung (Abbildung 6) ergibt die Dichtefunktion einer Normalverteilung eine glockenförmige Kurve, die symmetrisch zur Geraden $x = \mu$ ist. Die Dichtefunktion einer Normalverteilung mit den Parametern μ und $\sigma^2 > 0$ hat die Form:

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \text{ mit } -\infty < x < \infty$$

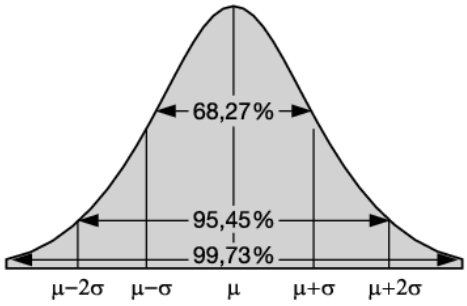


Abbildung 14: Gaußsche Normalverteilung. Abbildung aus (Springer-Verlag GmbH)

2.3.1.2 Z-Standardisierung

Mit Hilfe von Standardisierungsverfahren werden Merkmale in ein gleiches Größenverhältnis transformiert. Die Z-Standardisierung nutzt dazu den Mittelwert μ und die Standardabweichung σ. Die Berechnungsvorschrift für die Z-Standardisierung bzw. Z-Normalisierung von Variablen ist:

$$x_i' = \frac{x_i - \mu}{\sigma}$$, mit μ = Mittelwert und σ = Standardabweichung der Verteilung.

Diese Methode bietet gegenüber anderen Standardisierungsverfahren eine Reihe von Vorteilen. Der Mittelwert der Verteilung ist nach der Standardisierung gleich 0, die Standardabweichung beträgt 1. Am Vorzeichen des standardisierten Wertes ist zu erkennen, ob die Merkmalsausprägung größer oder kleiner als der Mittelwert ist. Fast alle standardisierten Werte liegen in einem Intervall von -2 bis 2. Ist der Betrag nach der Standardisierung größer als 2, so handelt es sich um einen Ausreißer. Allerdings können ursprünglich nichtnegative Variablen negative Werte aufweisen. Die Deutung des standardisierten Wertes im Merkmalskontext ist schwierig. (Universität Greifswald, 2013)

Abbildung 15 zeigt eine solche Z-Normalisierung zweier Zeitserien anhand der Abbildung 9 aus 2.2.1.

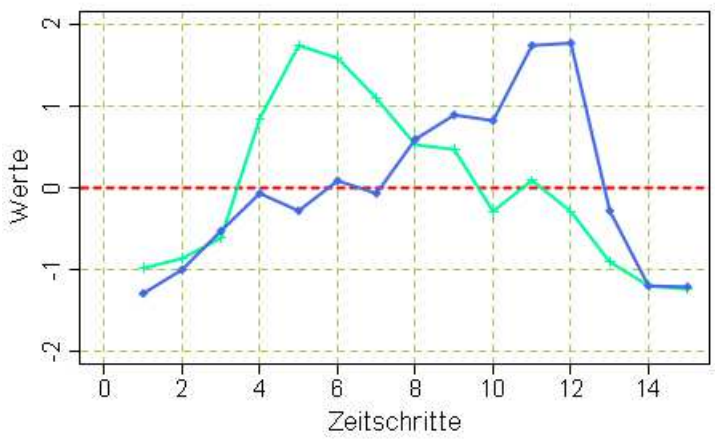


Abbildung 15: PAA zweier Zeitserien mit Z-Normalisierung (Google, 2012)

2.3.2 Breakpoints

Eine sortierte Liste von Intervallgrenzen oder auch Breakpoints B genannt, mit $B = \beta_1, \beta_2, ...\beta_{a-1}$, $\beta_{i-1} < \beta_i$ und $\beta_0 = -\infty, \beta_a = \infty$, teilt die Fläche unter der normierten Gaußkurve N(0,1) für ein Alphabet $\sum$ der Größe $l = |\sum|$ in gleich große Teile. (Google, 2012) Die Breakpoints können für ein gegebenes Alphabet fest im System hinterlegt und mittels einer Lookup Table (siehe Anhang A.2) nachgeschlagen werden. Hat man die PAA-Koeffizienten c_i eines Signals, die Größe eines Alphabets l und $alpha_j$ (j-te Buchstabe des Alphabets) vorliegen, kann die PAA-Signatur in SAX überführt werden. Ein PAA-Koeffizient der durch die Breakpoints in das j-te Intervall fällt, wird auf den j-ten Buchstaben abgebildet, mit $c_i = alpha_j$ und $c_i \in [\beta_{i-1}, \beta_i)$. Abbildung 16 zeigt die Intervallgrenzen (Schnittlinien) zweier SAX-transformierter Zeitreihen für ein Alphabet der Größe 4 anhand einer Normalverteilung und der Z-Normalisierung.

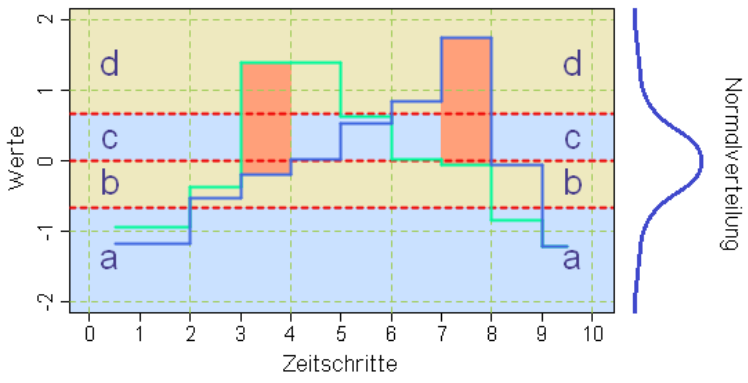


Abbildung 16: Intervallgrenzen zweier SAX-transformierter Zeitreihen für ein Alphabet der Größe 4 (Google, 2012)

2.3.3 Die SAX-Metrik

SAX führt eine neue Metrik zur Messung des Abstands zwischen Strings mit Hilfe des Euklidischen Abstandes und der PAA-Distanz ein. Ein String $\hat{C}$ stellt eine konvertierte PAA-Signatur *C* mit seinen Koeffizienten c_i dar und ist damit ein Vektor. Die Funktion $MINDIST(\hat{X},\hat{Y})$ gibt die minimale Distanz zwischen zwei Stringrepräsentationen der "originalen" Zeitserien X und Y wieder.

$$MINDIST(\hat{X},\hat{Y}) \equiv \sqrt{\frac{n}{m}}\sqrt{\sum_{i=1}^{m}(dist(\hat{x}_i,\hat{y}_i))^2}$$

Die *dist*-Funktion wird unter Verwendung einer Lookup Table für einen spezifischen Satz von Breakpoints ausgeführt. Tabelle 2 zeigt eine solche Tabelle für die Alphabetlänge 4 . Jeder einzelne Wert der Tabelle wird wie folgt berechnet:

Für jede *Zelle(r,s)* gilt, dass

$$Zelle(r,s) = \begin{cases} 0 & , \ falls \ |r-s| \le 1 \\ \beta_{\max(r,s)-1} - \beta_{\min(r,s)}, & sonst \end{cases}$$

ist.

Tabelle 2: Lookup Table - Minimaler Abstand zweier SAX-Koeffizienten bei Alphabetlänge 4 (Google, 2012)

Alphabet	a	b	c	d
a	0	0	0,67	1,34
b	0	0	0	0,67
c	0,67	0	0	0
d	1,34	0,67	0	0

Die Definition des Distanzmaßes ähnelt der Distanz zweier PAA-Signaturen. Der Unterschied liegt darin, dass die Distanz zwischen den beiden PAA-Koeffizienten

($x_i - y_i$) durch die Hilfsmethode zum Nachschlagen der Breakpoints *Zelle(r,s)* ersetzt wird. Die Besonderheit der SAX-Repräsentation ist, dass durch die Diskretisierung der minimale Abstand zwischen zwei benachbarten Buchstaben sehr klein werden kann, falls die diskretisierten PAA-Koeffizienten direkt an einer Diskretisierungs-intervallgrenze, also den Breakpoints, liegen. Dadurch muss der Abstand zwischen den benachbarten Buchstaben als 0 definiert werden. Ansonsten entspricht der Abstand zweier Buchstaben dem minimalen Abstand der dazwischen liegenden Intervalle, die anhand der Breakpoints $\beta_{\max(r,c)-1} - \beta_{\min(r,c)}$ definiert werden. (Schäfer, 2008)

2.4 CGR - Chaos Game Representation

Die Chaos Game Representation ist eine grafische Darstellung einer eindimensionalen Sequenz, wie zum Beispiel eine Gensequenz, ein Text in deutscher Sprache oder wie in diesem Fall eine SAX-Sequenz. Sie wurde erstmals als skalenunabhängige Darstellung für Gensequenzen von Jeffrey 1990 in (Jeffrey, 1990) vorgeschlagen. Jeffrey wurde bei der CGR von einem Algorithmus zur Darstellung von Fraktalen, auch "Chaos Game" genannt, inspiriert (Baransley, 1988).

2.4.1 Chaos Game - Erstellung der Grundstruktur und deren Aufteilung

Das Chaos Game ist mathematisch gesehen ein iteratives Funktionensystem (IFS). Ein IFS ist ein paarweiser Satz von linearen Gleichungen in der Form $x = ax + by + e$, $y = cx + dy + f$ und gibt die Formel zur Berechnung der neuen Werte für die x- und y-Koordinate wieder. Ein neuer Punkt wird im Chaos Game durch die halbe Länge zwischen dem vorhergehenden Punk und dem bestimmten Eckpunkt festgelegt. Beispielsweise braucht man für die Lösung des Chaos Game mit drei Ecken und einer Gleichung für jede Koordinate insgesamt sechs Gleichungen. Fasst man die beiden obigen Koordinatenformeln zusammen, ergibt sich eine kompaktere Schreibweise mit:

$$w(x, y) = (ax + by + e, cx + dy + f).$$

Dadurch erhält man eine Gleichung, manchmal auch als „Abbildung“ oder „Karte“ bezeichnet, wobei jede Abbildung durch 6 Koeffizienten beschrieben wird. Nun kommt noch hinzu, dass jeder Eckpunkt mit einer Nutzungswahrscheinlichkeit versehen ist.

Diese wird im 7. Koeffizienten (*p*) abgelegt. Meistens geht man jedoch von gleichverteilten Wahrscheinlichkeiten aus. Die Tabellen 3 und 4 zeigen für ein gleichseitiges Dreieck und für ein Quadrat die 6 Koeffizienten und die dazugehörigen gleichverteilten Wahrscheinlichkeiten, dem IFS-Code (Jeffrey, 1990).

Tabelle 3: IFS-Code eines gleichseitigen Dreiecks

w	***a***	***b***	***c***	***d***	***e***	***f***	***p***
1	0.5	0	0	0.5	0	0	0.33
2	0.5	0	0	0.5	0	0.5	0.33
3	0.5	0	0	0.5	0.5	0.5	0.33

Tabelle 4: IFS-Code eines Quadrats

w	***a***	***b***	***c***	***d***	***e***	***f***	***p***
1	0.5	0	0	0.5	0	0	0.25
2	0.5	0	0	0.5	0	0.5	0.25
3	0.5	0	0	0.5	0.5	0	0.25
4	0.5	0	0	0.5	0.5	0.5	0.25

Jeffrey nutzt die Kenntnisse über das Chaos Game um Desoxyribonukleinsäuresequenzen (DNA-Sequenzen) besser darstellen zu können. Dabei geht er wie folgt vor. Eine Gensequenz besteht aus vier Buchstaben ‘a‘, ‘c‘, ‘g‘, ‘t‘ bzw. ‘u‘ bei Ribonukleinsäuresequenzen (RNA-Sequenzen), wobei jeder Buchstabe für die Abkürzung der jeweiligen Base ist, mit a für Adinin, c für Cytosin, g für Guanin, t für Thymin und u für Uracil. Vier Buchstaben geben damit die zu nutzende Form vor und zwar ein Quadrat mit seinen vier Ecken. Für den Koeffizienten *p* wird eine Gleichverteilung angenommen und wird nicht weiter betrachtet. Jede Ecke wird mit einer der Basen beschriftet. Der Startpunkt ist der Mittelpunkt des Quadrates. Ist beispielsweise ‘g‘ die nächste Base die dargestellt werden soll, wird der darzustellende Punkt in der Mitte zwischen dem vorhergehenden Punkt und der ‘g‘-Ecke abgetragen. Für die Gensequenz „gaattc“ zeigt die Abbildung 17 die Darstellungen für jeden einzelnen Abtragungsschritt. Der Beginn ist Links oben und endet rechts unten.

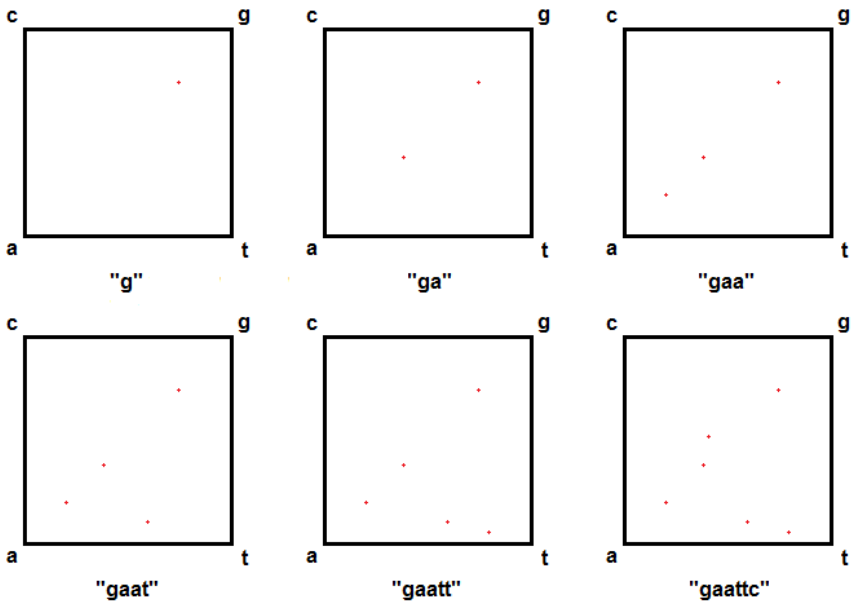


Abbildung 17: CGR Darstellungen für die Gensequenz "gaattc"

CGRs können, wie bereits angedeutet, in verschiedenen Formen repräsentiert werden. Dies gilt nicht nur im zweidimensionalen sondern auch im dreidimensionalen. Während die Anzahl der Ecken im Zweidimensionalen praktisch beliebig gewählt werden können, wobei die Mindestanzahl der Ecken 3 beträgt, gibt es im Dreidimensionalen Beschränkungen, die nicht jede Form zulassen. Ausgangspunkt für die Darstellungen im Zweidimensionalen ist ein regelmäßiges n-gon (Polygon). Die Regelmäßigkeit bei einem n-eder (Polyeder) muss ebenfalls gegeben sein. Die minimale Anzahl an Ecken beträgt 4. Aufgrund der geforderten Regelmäßigkeit lässt sich im Dreidimensionalen nicht jeder Körper nutzen. Die Tabelle 5 zeigt für den zweidimensionalen Raum (2D) und dreidimensionalen Raum (3D) die charakteristischen Polygone bzw. Polyeder mit der dazugehörigen Anzahl an Ecken.

Tabelle 5: Charakteristische Polygone im 2D und charakteristische Polyeder im 3D

2D		3D	
Anzahl Ecken	**Bezeichnung**	**Anzahl Ecken**	**Bezeichnung**
3	Dreieck	4	Tetraeder
4	Quadrat	6	Hexaeder
5	Pentagon	8	Würfel
6	Hexagon	12	Dodekaeder
…	…	…	…
∞	Kreis	∞	Kugel

2.4.2 CGR-Bitmaps und SAX

Basierend auf der Punktdarstellung der CGR von Jeffrey möchte man Bitmaps (Bilder) zur Darstellung von Zeitserien nutzen. Für einen vollständigen Quartärbaum (siehe Anhang A.3) als Bild im Zweidimensionalen wird ein CGR-Bitmap wie folgt erzeugt. Die Grundstruktur ist ein Quadrat. Jede Ecke eines Quadrates erhält aus einem Alphabet $\sum$ einen Buchstaben zugeordnet. Setzt man ein normiertes Quadrat mit der linken unteren Ecke in den Koordinatenursprung eines kartesischen Koordinatensystems, so ergeben sich bei einem Alphabet der Länge 4 mit den Buchstaben A, B, C, D, die Punkte A = (0,0), B = (1,0), C = (0,1) und D = (1,1). Das Quadrat wird anschließend in vier gleichgroße Quadrate geteilt, so dass der Abstand zwischen zwei benachbarten Buchstaben gleich groß ist. Damit ergeben sich vier neue Quadrate. Dieses Schema entspricht einem Rekursionsanfang für CGR mit einem SAX-Suffix der Länge 1. Wiederholt man die Aufteilung der zuvor erzeugten Quadrate entspricht dies einem Rekursionsschritt. Dies wird so lange wiederholt, bis man die gewünschte Aufteilung bzw. die Rekursionstiefe erreicht hat. (Kumar, Lolla, Keogh, Lonardi, & Ratanamahatana, 2005) Die Anzahl der Quadrate in den verschiedenen r Rekursionstiefen beträgt $2^{2(r+1)}$ mit $r \in 0,...,\infty$. Dies entspricht einer Vervierfachung der Quadrate pro Rekursionsschritt. Die Länge l eines SAX-Suffix gibt die Rekursionstiefe, mit $l = r - 1$, vor und damit auch die Anzahl der Aufteilungsschritte. Im 3D Bereich wird analog dem 2D Bereich vorgegangen. Am Beispiel eines Würfels würden 8 Eckpunkte und damit 8 Kuben sich als Grundstruktur ergeben. Die Anzahl der

Kuben in den verschiedenen *r* Rekursionstiefen beträgt $2^{3(r+1)}$, was eine Verachtfachung der Kuben pro Rekursionsschritt entspricht. Im Allgemeinen kann man festhalten, dass ein exponentielles Wachstum aufgrund der Rekursion im 2D- oder 3D-Bereich gegeben ist.

2.4.2.1 Beschriftung der Grundstruktur nach SAX-Suffixe und dessen Häufigkeit

Als zweiter Schritt wird die Beschriftung der einzelnen Quadrate vorgenommen. Ausgehend von der Grundstruktur und seinen vier Quadraten, entspricht das linke untere Quadrat dem SAX-Suffix A, das rechte untere Quadrat dem SAX-Suffix B, das linke obere Quadrat dem SAX-Suffix C und D das letzte Quadrat. Im ersten Rekursionsschritt sind 16 Quadrate mit den dazugehörigen Suffixen AA, AB, AC, AD, BA, ..., DD zu belegen. Die linken unteren 4 Quadrate, die zuvor aus dem Quadrat mit dem Suffix A hervorgegangen sind, haben den Anfangsbuchstaben A. Der zweite Buchstabe ist wie zuvor beschrieben gleichermaßen zu verteilen. Analog werden die restlichen 12 Quadrate beschriftet. Diese Beschriftung ist für jede Rekursionstiefe anzuwenden. Abbildung 9 zeigt eine grafische Repräsentationsform als Grundstruktur eines Quartärbaums, der bis zur zweiten Rekursionstiefe mit den dazugehörigen SAX Suffixe dargestellt ist.

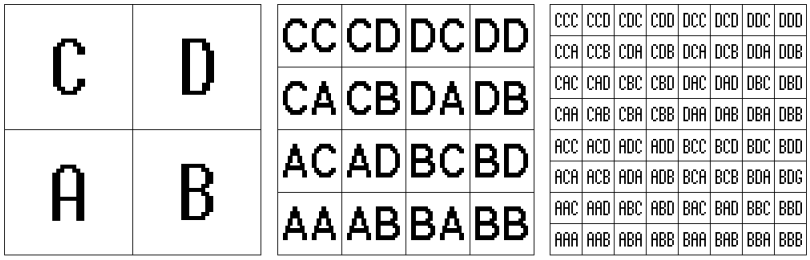


Abbildung 18: Quartärbaum einer Sequenz über dem Alphabet {A, B, C, D} in unterschiedlichen Rekursionstiefen

Zu jedem SAX-Suffix gibt es eine Häufigkeit, wie oft ein Suffix in der SAX-Darstellung vorkommt. Diese Häufigkeit ist ausschlaggebend für die Farbdarstellung der einzelnen Quadrate.

2.4.2.2 Farbdarstellung der CGR-Bitmaps

Der letzte Schritt der CGR ist jedem Quadrat mit der gleichen SAX-Häufigkeit eine Farbe zuzuordnen. Quadrate mit den gleichen SAX-Häufigkeitswerten bekommen die gleiche Farbe. Quadrate unterschiedlicher SAX-Häufigkeitswerte werden beispielsweise durch eine Grauabstufung dargestellt, in dem die Farbe Weiß die geringste Häufigkeitsausprägung darstellt und Schwarz die Höchste. Abbildung 19 verdeutlicht noch einmal die Vorgehensweise der CGR.

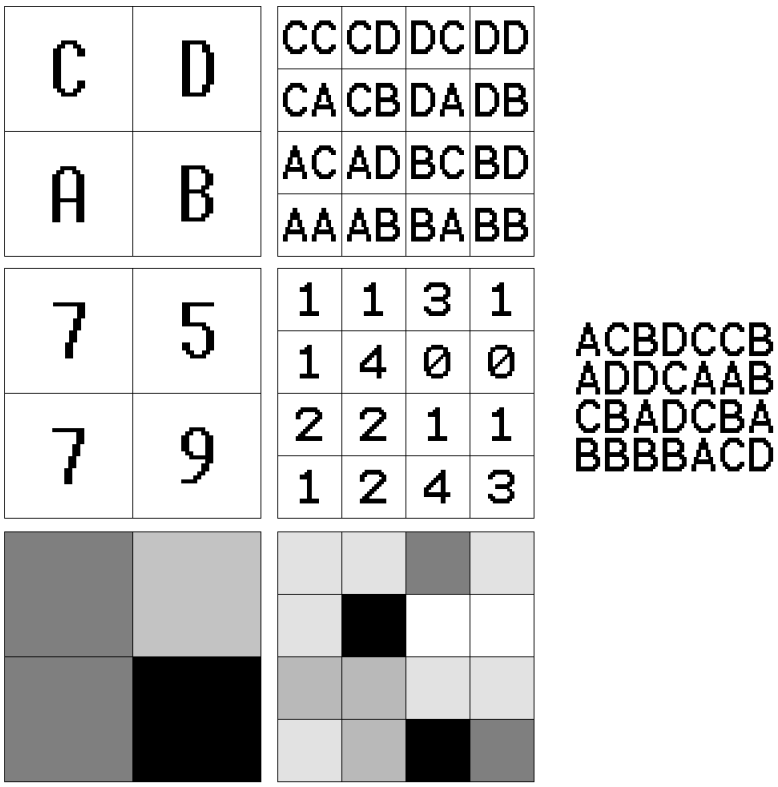


Abbildung 19: *Oben)* Vier mögliche SAX-Symbole werden auf den vier Quadranten eines Quadrates abgebildet, inkl. Rekursionsschritt 1 und die Suffixe der Länge 2. *Mitte)* Eine Sequenz aus 28 Buchstaben, wobei die Anzahl von SAX-Symbolen auf das Raster übertragen wird. *Unten)* Die übertragenen Werte können linear einer Farbpalette zugeordnet werden, wodurch ein CGR entsteht.

2.4.3 Abhängigkeiten in der CGR

Je nach Anwendungsgebiet und Ergiebigkeit der originalen Zeitserie wird die CGR mit der Anzahl seiner darzustellenden Quadrate variieren. Dies hängt im Wesentlichen von zwei Faktoren ab, der PAA-Länge und der SAX-Suffix-Länge. Die PAA-Länge gibt an,

in wie viele Segmente die originale Zeitserie aufgeteilt wurde. Sie sollte nicht zu klein gewählt werden, damit hinreichend verschiedene SAX-Suffixe erstellt werden können, aber auch nicht zu groß, um möglicherweise wichtige Ausschläge in der originalen Zeitserie zu überdecken. Da es für die PAA-Länge kein Patentrezept gibt, muss der Benutzer durch Ausprobieren für sich die beste Lösung finden. Ebenso verhält es sich mit der SAX-Suffix-Länge. Wird die SAX-Suffix-Länge zu groß gewählt, relativieren sich die Häufigkeiten der einzelnen SAX-Suffixe. Die Farbdarstellungen der jeweiligen SAX-Suffixe im CGR würden ähnlich oder gleich aussehen und eine Unterscheidung unbrauchbar machen.

2.4.4 Vergleich zweier CGR mit unterschiedlicher Wortlänge

Der Vergleich zweier CGR mit unterschiedlicher Wortlänger ist ebenfalls möglich. Hierbei wird durch eine Normalisierung der SAX-Häufigkeitswerte, durch den Häufigkeitswert mit dem Maximum der längeren SAX-Sequenz, eine entsprechende Umformung an der kürzeren Sequenz vorgenommen.

2.4.5 Skalierbarkeit der CGR

Betrachtet man die strukturelle Skalierbarkeit von CGRs, stößt man schnell an ihre darstellbaren Grenzen. Limitierender Gegenstand ist der Bildschirm, vor dem der Benutzer sitzt. Nimmt man die darstellbare Auflösung eines Monitors von beispielsweise 1280 x 1024 Bildpunkten (Breite x Höhe) an, so ist nach der obigen Formel $2^{2(r+1)}$ für die Anzahl der darzustellenden Quadrate in den verschiedenen r Rekursionstiefen bezüglich der Höhenangabe nach theoretischen $r = 9$ Schluss. Praktisch muss aber $r = 8$ gewählt werden, da die GUI ebenfalls eine gewisse Rahmenbreite besitzt. In diesem Fall müssen 512 Kästchen pro Seite dargestellt werden, was eine Gesamtzahl von 262144 darzustellenden Kästchen entspricht. Vergleicht man die Zahl von 512 Kästchen mit den Bildpunkten eines Monitors, sollte klar sein, dass die Kästchen Pixelgröße erreicht haben.

Kapitel III: Programmbeschreibung

In diesem Kapitel wird das Konzept des Programms dargestellt. Zunächst werden Anforderungen an das Programm erläutert, welche die Grundlagen des Designprozesses bilden. Danach werden die Aufteilungen des Programms in die einzelnen Module sowie der Datenfluss durch diese Module beschrieben. Im Anschluss wird die Anwenderdokumentation für eine problemlose und sichere Nutzung beschrieben.

3.1 Anforderungsdefinitionen

An das Programm wurde eine Reihe von Anforderungen gestellt, die im Nachfolgenden beschrieben werden.

- **Graphische Benutzeroberfläche (GBO / *englisch* GUI):** Das Programm soll in eine GUI eingebettet werden, was dem Benutzer die Interaktion mit dem Computer über grafische Symbole erlaubt. Die Darstellungen und Elemente sollen unter Verwendung einer Maus als Zeigergerät gesteuert werden können.

- **Einlesen von Zeitserien:** Das Programm soll in der Lage sein, Zeitserien einzulesen und sie zu verarbeiten.

- **Hinzufügen von Zeitserien:** Das Programm soll in der Lage sein, zu vorhandenen eingelesenen Zeitserien weitere hinzuzufügen und zu verarbeiten.

- **Darstellung der Zeitserien durch Liniendiagramme:** Eingelesene Zeitserien sollen mit Hilfe von Linienplots dargestellt werden.

- **Skalierung von Liniendiagrammen:** Linienplots sollen unter Verwendung einer Maus skaliert werden können.

- **Darstellung der Zeitserien durch CGR:** Eingelesene Zeitserien sollen mit Hilfe der CGR dargestellt werden.

- **Leichte Konfigurierbarkeit der CGR:** Die CGR-Darstellung soll leicht konfigurierbar sein.

- **Skalierung der CGR:** Durch die Benutzung der Parameter SAX-Suffix-Länge und PAA-Länge soll eine Skalierung der CGR ermöglicht werden.

- **Skalierung der CGR:** Durch die Benutzung der Parameter SAX-Suffix-Länge und PAA-Länge soll eine Skalierung der CGR ermöglicht werden.

- **Interaktion zwischen Liniendiagramm und CGR:** Mit Hilfe von "Brushing & Linking" sollen zwei verschiedene Darstellungsformen von Zeitserien verknüpft werden. Zeitpunkt- oder Feldauswahl im Liniendiagramm sollen in CGR darstellbar sein. Ebenfalls soll die Feldauswahl im CGR die entsprechenden Felder im Linienplot anzeigen.

- **Modularer Aufbau:** Das Programm sollte modular aufgebaut sein, um eine spätere Erweiterung und/oder Verbesserung zu ermöglichen.

- **Dokumentation des Programms:** Alle Funktionen und Datenschnittstellen sollen dokumentiert werden, so dass eine spätere Erweiterung erleichtert wird.

3.2 Systementwurf

Um die Anforderungen für das Programm erfüllen zu können, wurde es in verschiedene Module zerlegt. Die Modularisierung hat den Vorteil der Austauschbarkeit verschiedener Programmteile. Beispielsweise ist es möglich, den Datenparser (siehe Anhang A.6) durch einen anderen zu ersetzt. Aufgaben, die nicht zum eigentlichen Programmablauf gehörten, wurden in eigene Hilfsprogramme ausgegliedert.
Die wesentlichen Programmmodule sind der Datenparser, die Linienplotdarstellung, die JMotiv Anbindung (siehe Anhang A.7), die PAA und SAX zur Verfügung stellt und die CGR-Darstellung. Prototypisch wurden zwei verschiedene Parser programmiert. Durch eine Neukompilierung des Programms kann der zweite Datenparser verwendet werden.

Die wesentlichen Module und die wichtigsten Datenströme zwischen den Modulen sind noch einmal in Abbildung 20 dargestellt.

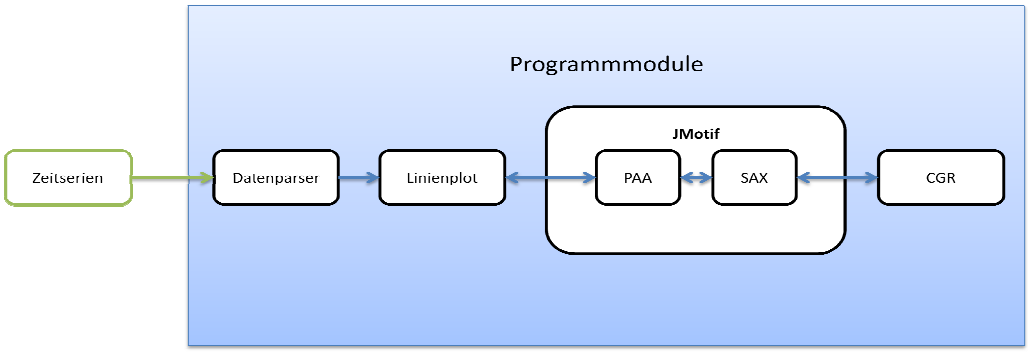


Abbildung 20: Vereinfachte Darstellung der wichtigsten Datenströme und der Programmmodule

3.3 Datenfluss

Ein wichtiger Designpunkte für die Erstellung eines Programms ist die Festlegung des Datenflusses vom Parsen bis zur Darstellung der CGR. Dabei wird der Datenfluss von verschiedenen Datenpuffern unterteilt. Datenpuffer sorgen beim Lesen und Schreiben für einen kontinuierlichen Datenfluss. Der Datenparser liest die Zeitserien ein und verarbeitet sie, so dass diese als Linienplot dargestellt werden können. Gleichzeitig übergibt der Parser die Daten an das JMotiv-Modul. Hier werden die Daten im PAA-Modul weiterverarbeitet. Ein neuer Datenpuffer ist für die PAA-Darstellung notwendig. SAX erhält den Datenfluss der PAA-Darstellung, der wiederum einen eigenen Datenpuffer anlegt. Dieser Datenpuffer enthält die SAX-Darstellung der originalen Zeitserie. Der SAX-Datenpuffer wird zur Darstellung der CGR (CGR-Bitmaps) genutzt. Eine Rückrechnung der CGR-Daten auf die Linienplots ist ebenfalls möglich.

3.4 Anwenderdokumentation

Die Anwenderdokumentation wird während des Produkteinsatzes benutzt und dient dem Zweck, das Produkt problemlos und sicher einsetzen zu können.

3.4.1 Voraussetzungen an Soft- und Hardware, Standardeinstellungen

Der CGR Viewer ist ein in Java programmiertes Programm und ist damit plattformunabhängig. Prinzipiell wird jedes Betriebssystem (Windows, Linux, OS-Apple, ...) unterstützt, wobei ein lauffähiges Java installiert sein muss. Hardwareseitig

kann keine Mindestanforderung definiert werden, da die Testmöglichkeiten fehlen. Als Mindestauflösung bei der Bildschirmeinstellung empfehle ich 1024 x 786 Pixel.

3.4.2 Datenparser

Der Datenparser hat die Aufgabe, die eingelesenen Daten für die Weiterverarbeitung umzuwandeln. Nachfolgende Struktur (Tabelle 6) muss die eingelesene Datei aufweisen, damit der CGR Viewer sie nutzen kann: Die erste Spalte beinhaltet eine fortlaufende Nummerierung für die aufgenommenen Zeitdaten, welche bei 0 beginnt. Jede weitere Spalte enthält eine fortlaufende Zeitreihe. Alle Spalten sind durch Tabs oder Lehrzeichen getrennt.

Tabelle 6: Strukturelle Darstellung der einzulesenden Daten

Nummerierung	Zeitreihe 1	Zeitreihe 2	Zeitreihe ...
0	4	3	...
1	12	8	...
2	13	4	...
3	15	9	...
4	6	23	...
...	...	...	...

3.4.3 Programmstart

Der CGR Viewer muss nicht installiert werden. Er kann von jedem beliebigen Datenträgermedium gestartet werden. Beim Start des Programms wird automatisch ein Koordinatensystem erzeugt, welches interaktiv benutz werden kann (siehe Mouseevents ab 3.4.6.1 bis 2.4.6.3). Dazu ist eine Mouse oder ein Touchpad zwingend erforderlich.

3.4.4 Menüleiste und Navigation

Das „Main – Window“, in dem ein Koordinatensystem erzeugt wird, besitzt eine Menüleiste. Diese Menüleiste hat zwei Einträge, "File" und "Display".

3.4.4.1 Das File-Menü

Im File-Menü befinden sich die Einträge "Open File ...", "Add File ...", "Clear All" und "Exit".

- "Open File ..." kann Zeitserien öffnen.
- "Add File ..." kann zu bestehenden eingelesenen Zeitserien weitere Zeitserien hinzufügen
- "Clear All" löscht den gesamten Speicher (Datenpuffer) mit den eingelesenen Zeitserien
- "Exit" beendet das Programm

3.4.4.2 Das Display-Menü

Im Display-Menü befinden sich die Einträge "Time Series" mit den Unterkategorien "Select Time Series ..." sowie "Clear Marker" und "Chaos Game Representation" mit der Unterkategorie "Select Chaos Game Representation ...".

- "Select Time Series ..." öffnet ein neues Fenster, in dem einzelne Zeitserien zur Darstellung ausgesucht werden können.
- "Clear Marker" löscht die Sektionen im Linienplot, die bei der Interaktion zwischen CGR-Bitmaps und Linienplots angezeigt werden können. (siehe auch unter Interaktionen und Mouseevents)
- "Select Chaos Game Representation ..." öffnet ein neues Fenster, in dem die Parameter für die CGR-Bitmaps eingestellt werden können.

3.4.5 Fensterdarstellung

3.4.5.1 Main - Window

Dieses Fenster biete den Zugriff auf die Menüleiste. Des Weiteren können geladene Zeitserien als Linienplot angezeigt werden.

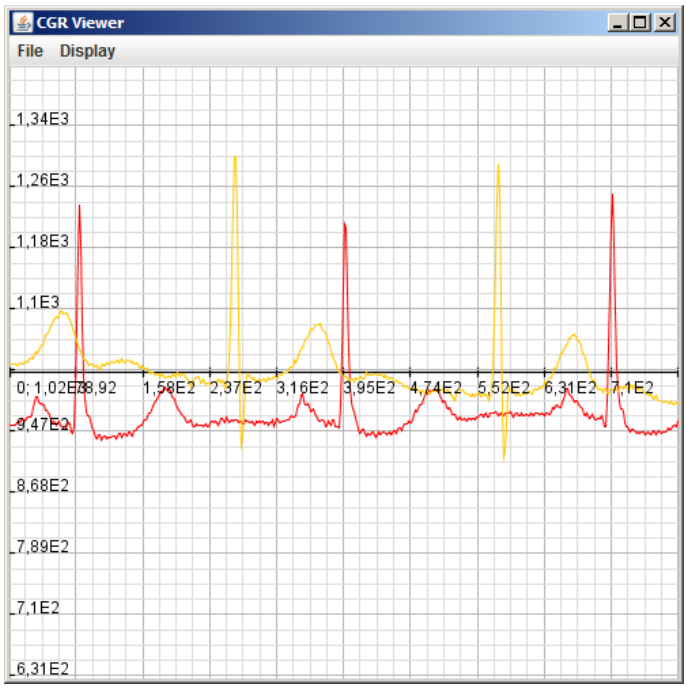


Abbildung 21: Main - Window mit geladener Zeitserie

3.4.5.2 Open/Add File - Window

Um Zeitserien öffnen/hinzufügen zu können, gibt es ein eigenes Fenster. Im oberen Bereich wählt man sich den Ort (Ordner) aus, in dem die Daten liegen. Im mittleren Bereich werden die Dateien angezeigt, die der Ordner enthält. Das Öffnen mehreren Zeitserien ist möglich, in dem man die betreffenden Dateien markiert oder den Namen der zu öffnenden Dateien im unteren Bereich unter "Dateiname" mit Anführungszeichen eingibt.

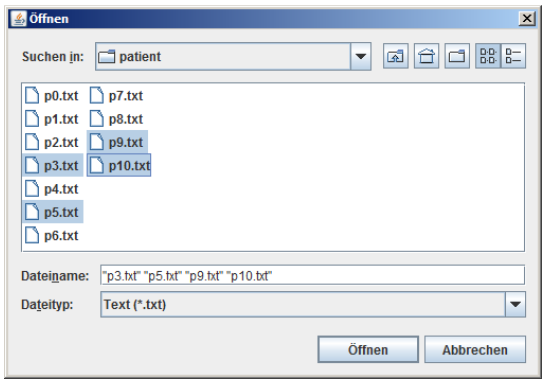


Abbildung 22: Open/Add File - Window zum Selektieren der öffnenden Daten

3.4.5.3 Define Time Series – Window

Dieses Fenster gibt dem Anwender die Möglichkeit, Zeitserien umzubenennen und eine andere Farbdarstellung zu wählen. Durch klicken in das Name-Textfeld ist es möglich, die ausgewählte Zeitserie umzubenennen. Klickt man auf den Color-Button, der die aktuelle Farbe der Zeitserie anzeigt, öffnet sich ein neues Fenster (Select Line Color), worin man eine gewünschte Farbe einstellen kann.

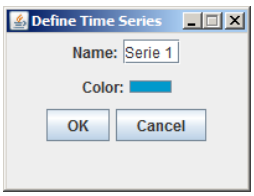


Abbildung 23: Define Time Series -Window zur Umbenennung der Zeitserie

3.4.5.4 Select Line Color – Window

Dieses Fenster gibt dem Anwender die Möglichkeit, die Farbendarstellung der Zeitserien in den Linienplots zu ändern.

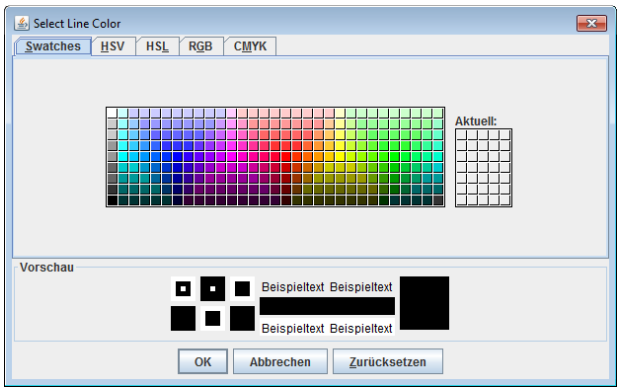


Abbildung 24: Select Line Color - Window zur farblichen Gestaltung der Zeitserien in Linienplots

3.4.5.5 Select Time Series – Window

Dieses Fenster bietet die Möglichkeit, aus den geladenen Zeitserien einzelne zu selektieren, um diese im Main - Window anzeigen zu lassen. Die Namensgebung der Zeitserien resultiert aus den Dateinamen und der Anzahl an Zeitserien, die eine Datei enthält. Die Farben für die Zeitserien sowie die Beschriftung sind vorgegeben, können

aber beliebig abgeändert werden. Dazu ist es nötig, auf die Farbdarstellung der Zeitserie zu klicken.

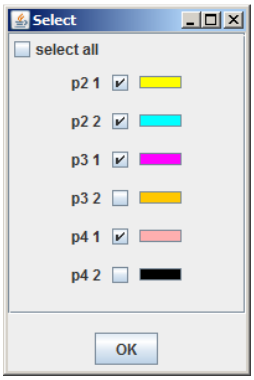


Abbildung 25: Select Time Series - Window zur Auswahl der anzuzeigenden Zeitserien

3.4.5.6 Select Chaos Game Representation – Window

Dieses Fenster bietet die Möglichkeit, aus den dargestellten Zeitserien im Main Window einzelne zu selektieren, um diese in einem neuen Fenster, dem CGR Graph, anzeigen zu lassen. Für jede Zeitserie wird ein eigenes CGR Graph - Window erzeugt. Die Parameter "length suffix" und "segments PAA" sind voreingestellt und können abgeändert werden. Der voreingestellte Wert im "segments PAA" - Feld gibt die maximal mögliche Anzahl an PAA-Segmenten an. Diese entspricht der Gesamtzahl an Datenpunkten in einer Zeitserie. Die Einstellmöglichkeiten bezüglich Name und Farbe der Zeitserien ist analog dem Select Time Series - Window.

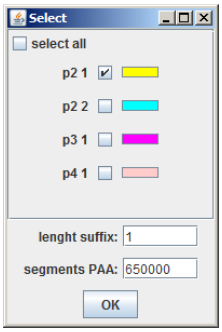


Abbildung 26: Select Chaos Game Representation - Window zur Anzeige und Parameterwahl von CGR-Bitmaps

3.4.5.7 CGR Graph – Window

In der Abbildung 27 wird ein CGR-Bitmap mit den zuvor eingestellten Werten aus dem Select Chaos Game Representation - Window dargestellt. Die Skala auf der rechten Seite des CGR-Bitmap repräsentiert die Häufigkeiten der einzelnen Farben im CGR-Bitmap, wobei die Farben im CGR-Bitmap von blau über grün nach gelb und rot und dessen Schattierungen die Häufigkeiten des jeweiligen SAX-Suffix angeben. Bei einer linearen Skalierung ist blau die Farbe mit der geringsten Anzahl an Übereinstimmungen und Rot die Farbe mit den meisten. Ein weiteres Highlight ist die Umstellung von "linear" auf die "logarithmic" (siehe Anhang A.8) Skalierung. Die logarithmische Skalierung wird nur in einer Farbe und deren Schattierungen dargestellt. Abbildung 28 zeigt ein solches Bild.

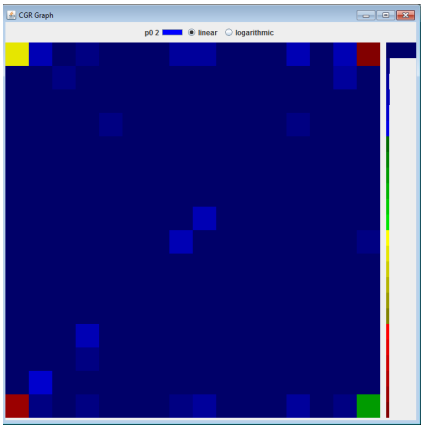


Abbildung 27: CGR Graph - Window zeigt ein Bitmap einer Zeitserie mit linearer Farbdarstellung

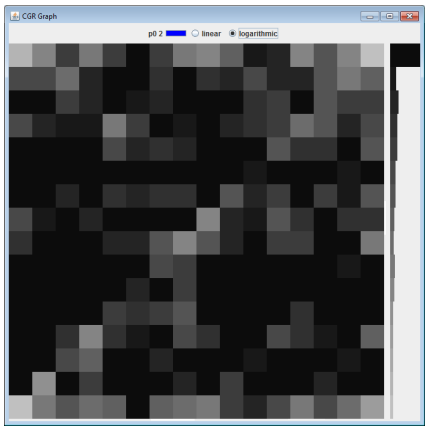


Abbildung 28: CGR Graph - Window zeigt ein Bitmap der gleichen Zeitserie wie aus Abbildung 27 mit logarithmischer Farbdarstellung

3.4.6 Interaktionen und Mouseevents

In diesem Abschnitt werden verschiedene Mouseevents und deren Aktionen im Main - Window und CGR-Graph - Window erklärt. Um alle Events ausführen zu können, ist eine geladene Zeitserie und das dazugehörige CGR-Bitmap notwendig.

3.4.6.1 Mouseevents im Main - Window

Das Main – Window besitzt ein Koordinatensystem, das durch das gedrückt halten der linken Maustaste verschoben werden kann. Mit dem Scrollrad der Maus kann man das Koordinatensystem vergrößern oder verkleinern (zooming), um detailliertere oder nicht detailliertere Darstellungen zu erhalten.
Der erste Einfachklick der linken Maustaste setzt an der Stelle des Mauszeigers eine vertikale Linie als Markierung in das Koordinatensystem. Der zweite Einfachklick der linken Maustaste setzt eine weitere Markierung als vertikale Linie in das Koordinatensystem. Beide Linien spannen nun einen Abschnitt auf, der für spätere Interaktionen und Darstellungen zwischen dem Main – Window und CGR-Graph – Window notwendig sind. Abbildung 31 zeigt einen solchen markierten Abschnitt auf der linken Seite. Ist ein Abschnitt gesetzt kann er auch mit Hilfe der linken Maustaste vergrößert oder verkleinert werden. Durch einen Doppelklick der linken Maustaste entfernt man den markierten Abschnitt.

3.4.6.2 Mouseevents im CGR-Graph - Window

Im CGR-Graph – Window gibt es zwei Mouseevents. Fährt man mit der Maus über ein Kästchen der CGR-Bitmaps (mouseover), so wird nach kurzer Zeit das dazugehörige Suffix mit der Anzahl an Vorkommen eingeblendet. Der zweite Mouseevent ist das Anklicken (Einfachklick) eines Kästchens. Dies hat keine Auswirkungen auf das CGR-Graph – Window wird aber für die Interaktion zwischen CGR-Graph – Window und Main – Window benutzt.

3.4.6.3 Mouseevents im Überblick

Für eine Kurzübersicht werden alle Mouseevents für das Main - Window und das CGR-Graph - Window in den Tabellen 7 und 8 dargestellt.

Tabelle 7: Mouseevents für das Main -Window

Main – Window	
Mouseevent	**Beschreibung**
Linksklick gedrückt halten	Verschieben des Koordinatensystems
scrolling	Zoomen des Linienplots
erster einfacher Linksklick	Setzen der Anfangsmarkierung eines Abschnitts
zweiter einfacher Linksklick	Setzen der Endmarkierung eines Abschnitts
einfacher Rechtsklick	Ändern der Größe des markierten Abschnitts
doppelter Linksklick	Entfernen des markierten Abschnitts

Tabelle 8: Mouseevents für das CGR-Graph - Window

CGR-Graph - Window	
Mouseevent	**Beschreibung**
mouseover	Suffixanzeige
einfacher Linksklick	Stellt den dazugehörigen Abschnitt im Linienplot dar

3.4.6.4 Brushing & Linking zwischen Main - Window und CGR-Graph – Window

Durch das anklicken eines Kästchens im CGR-Graph – Window werden die dazugehörigen Abschnitte im Linienplot des Main – Window farblich angezeigt. Der Farbton des Abschnitts entspricht dem Farbton der Zeitserie. Allerdings wurde eine geringere Sättigung gewählt, um den Abschnitt transparenter zu gestalten. Abbildung 29 verdeutlicht dies anschaulich.

Wie bereits beschrieben ist es möglich, innerhalb des Main – Window ein Feld zu markieren. Ist ein solches Feld markiert und wurde zuvor das dazugehörige CGR-Bitmap erstellt, werden die zum Abschnitt gehörenden Suffixe im CGR-Bitmap markiert (highlighting). Dies stellt noch einmal anschaulich die Abbildung 30 dar. Ebenfalls ist es möglich den markierten Abschnitt als CGR-Bitmap neu berechnen zu

lassen. In Abbildung 31 kann man das Brushing & Linking für mehrere Zeitserien sehen.

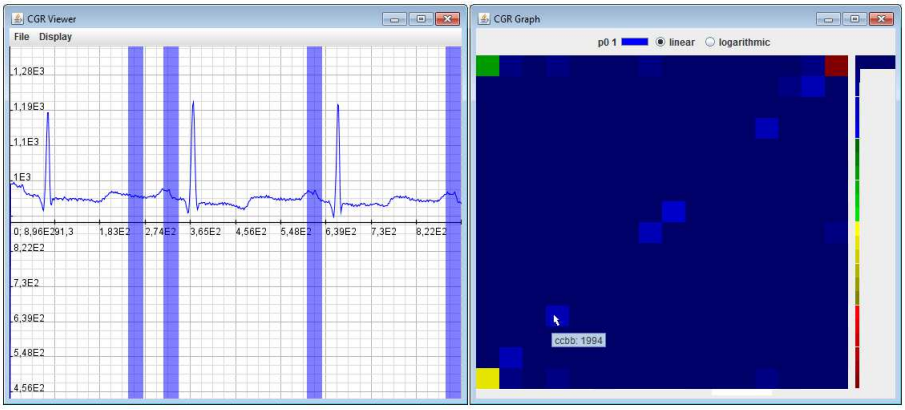


Abbildung 29: Brushing & Linking zwischen CGR-Graph – Window und Main – Window

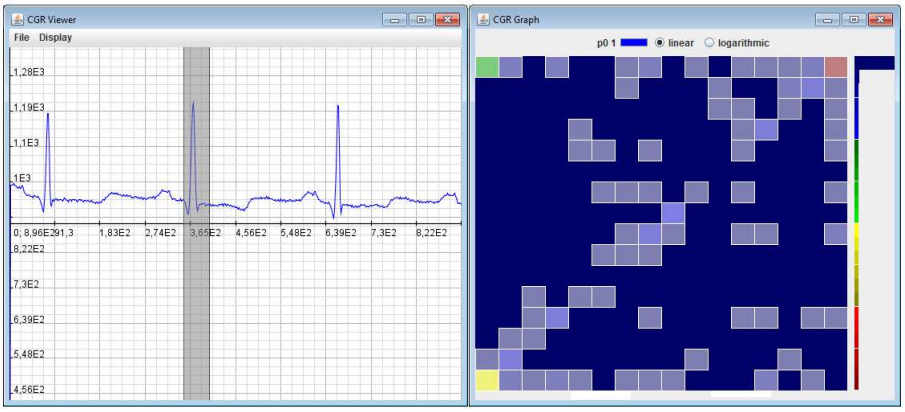


Abbildung 30: Brushing & Linking zwischen Main – Window und CGR-Graph – Window

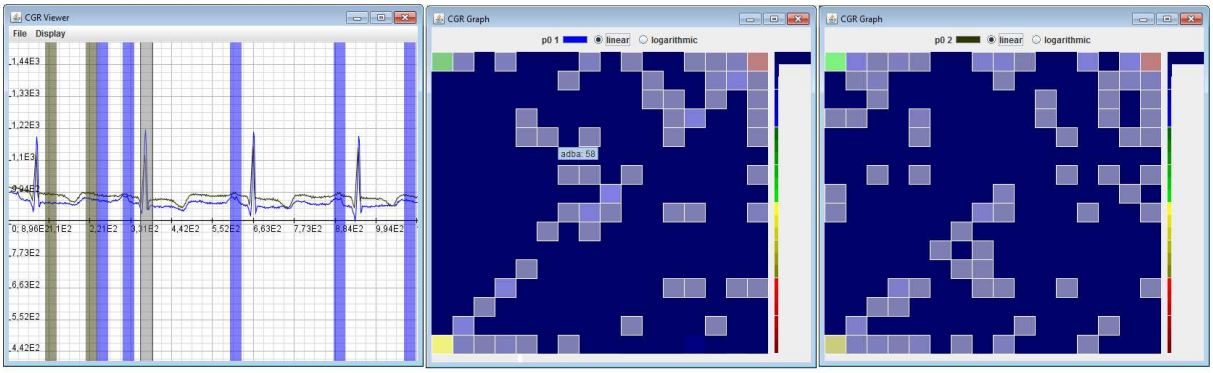


Abbildung 31: Brushing & Linking auf mehreren dargestellten Zeitserien

Kapitel IV: Anwendungstests

In diesem Kapitel werden Anwendungstests auf der Basis von EKG-Aufzeichnungen mit Hilfe des CGR Viewers beschrieben. Um ein besseres Verständnis für die medizinische Seite zu bekommen, werden vorab die dazu nötigen Fachinformationen erklärt. Die bei den Tests entstandenen CGR-Bitmaps wurden medizinischen Spezialisten zur Beurteilung vorgelegt. Dessen Verständnis und Meinungen fließen in die Beurteilung der Tests mit ein.

4.1 Das EKG

Die Elektrokardiografie ist in der Inneren Medizin eine der wichtigsten Untersuchungsmethoden, um elektrische Vorgänge im Herzmuskel grafisch darzustellen und erlaubt damit vielfältige Rückschlüsse auf die Herzfunktion. Jede Kontraktion des Herzmuskels bei einem gesunden Menschen folgt einem bestimmten Muster und wiederholt sich bei jedem Herzschlag. Sie geht mit einer elektrischen Erregung einher, die über genau definierte Ableitungselektroden am Körper abgenommen werden. Die elektrische Erregung entsteht durch die elektrischen Aktivitäten der Herzmuskelzellen und werden bis zur Körperoberfläche weitergeleitet. Die sehr schwachen Signale werden von einem EKG-Gerät verstärkt und als Kurve auf einem Monitor oder ausgedruckten auf Papier dargestellt. Für ein vollständiges 12-Kanal-EKG werden 12 Ableitungen aufgezeichnet: Drei bipolare Extremitätenableitungen nach Einthoven, drei unipolaren Extremitätenableitungen nach Goldberger sowie sechs unipolaren Brustwandableitungen nach Wilson (Kleindienst, 2009).

4.2 Erregung des Herzens und dessen Ableitung

Die Erregung - auch Aktionspotential genannt - geht bei einem gesunden Menschen vom Sinusknoten aus. Er befindet sich im Bereich des rechten Vorhofes direkt unterhalb der Einmündungsstelle der oberen Hohlvene in der Herzwand. Der Sinusknoten bestimmt die Frequenz, mit der das Herz schlägt und wird deshalb auch oft als "Schrittmacher des Herzens" bezeichnet. Ausgehend vom Sinusknoten setzt sich die Erregung über die Vorhofmuskulatur fort bis zum nächsten zentralen Bereich der Erregungsleitung, dem Atrio-Ventrikular-Knoten, kurz AV-Knoten. Der AV-Knoten,

befindet sich im Grenzbereich zwischen Vorhof (Atrium) und Kammer (Ventrikel). Er nimmt die Signale aus dem Sinusknoten auf und leitet sie weiter an das His'sche-Bündel, das am Grund des rechten Vorhofs in Richtung Kammerscheidewand verläuft. Im Bereich der Scheidewand teilt sich dann die Erregungsleitung in einen rechten und einen linken Kammerschenkel auf. Die Kammerschenkel verlaufen entlang der Scheidewand in Richtung Herzspitze und verzweigen sich dann weiter. Die feinen Strukturen der Endabzweigungen des Reizleitungssystems werden Purkinje-Fasern genannt. Sie enden im Herzmuskel (Myokard) und erregen dort die Herzmuskelzellen. Abbildung 33 stellt noch einmal schematisch die Erregungsweiterleitung im Herzen dar (Wehner, 2011).

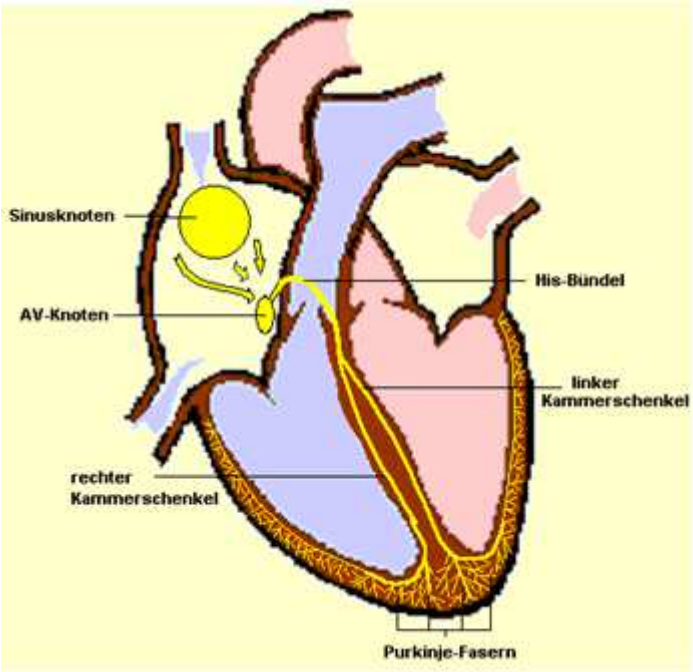


Abbildung 32: Erregungsleitung des Herzens. Bild aus (Wehner, 2011)

Die Ableitung des Aktionspotentials am Herzen ist durch die verschiedenen Stationen der Erregungsweiterleitung definiert und ergibt für einen gesund Menschen einen typische Kurvenverlauf. Abbildung 34 zeigt eine solche Kurve. Zu erkennen sind 5 Abschnitte P, Q, R, S, und T, wobei jede ihr eigenes Kurvenverhalten hat. Die P-Welle stellt die Erregungsausbreitung in den Vorhöfen dar, gefolgt von der P-Q-Strecke, die eine Überleitung auf das His'sche-Bündel bzw. auf die Herzkammern repräsentiert. Der QRS-Komplex ist die vollständige Erregungsausbreitung in den Herzkammern, wobei Q die Erregungsausbreitung in der Kammerscheidewand, R die Erregungsausbreitung

der Kammerwände und S die Erregungsausbreitung der Kammerwände in Richtung Herzbasis darstellen. Die S-T-Strecke stellt die Erregungsrückbildung in den Kammern dar wobei die T-Welle die Spätphase der Erregungsrückbildung entspricht. Sind Erkrankungen am Herzen vorhanden, kann dies mit Veränderung der Teilkurven einhergehen.

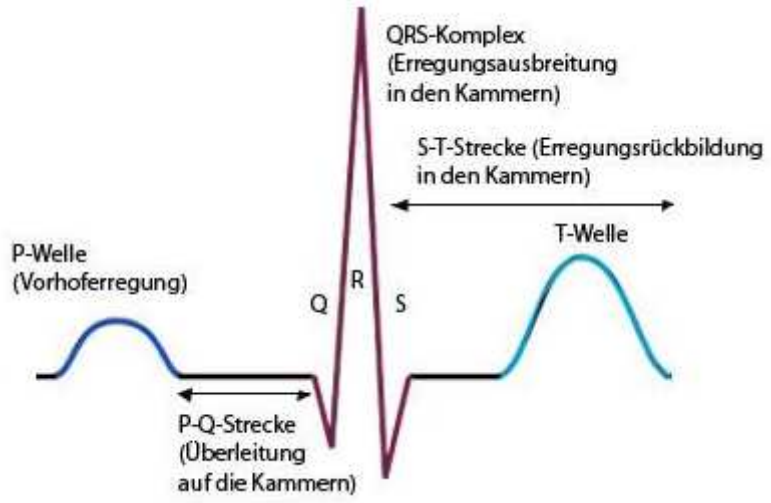


Abbildung 33: Elektrokardiogramm. Bild aus (NetDoktor.de GmbH, 2013)

4.3 Der Sinusrhythmus

Der normale, regelmäßige Herzschlag des Menschen wird als Sinusrhythmus bezeichnet und hat eine Frequenz von 60 bis 100 Schläge pro Minute. Verlangsamt sich der Herzschlag und fällt unter 60 Schläge, spricht man von einer Sinusbradykardie. Ist der Herzschlag beschleunigt und oberhalb von 100 Schlägen, nennt man dies Sinustachykardie. Krankhafte Abweichungen in der Entstehung oder Weiterleitung der Herzerregung werden als Herzrhythmusstörungen bezeichnet.

4.4 Datenmaterial

Die Daten stammen aus der Onlinedatenbank der Website www.physionet.org. PhysioNet bietet einen kostenlosen Web-Zugriff auf eine große Sammlungen von aufgezeichneten physiologischen Signalen. Für die Tests habe ich die MIT-BIH Arrhythmia Database gewählt, da sie mir von Prof. Dr. med. Dipl.-Inform. (FH) Thomas Hilbel für meine Forschungszwecke empfohlen wurde. Die Daten stammen vom Bostoner Beth Isarel Deaconess Medical Center und deren Forschungseinrichtung am MIT (Massachusetts Institute of Technology). Sie enthält 48 halbstündige

Ausschnitte einer ambulanten Zweikanal-EKG-Aufzeichnung. Dreiundzwanzig Aufnahmen wurden nach einem Zufallsprinzip aus einer Menge von 4000 ambulanten 24-Stunden-EKG-Aufnahmen einer gemischten Population gewählt, wobei ca. 60% stationäre Patienten und ca. 40% ambulante Patienten vom Bostoner Beth Israel Deaconess Medical Center beteiligt waren. Die restlichen 25 Aufnahmen wurden aus dem gleichen Datensatz entnommen. Das Augenmerk hierbei wurde auf klinisch signifikante Arrhythmien gelegt, da diese zu selten oder gar nicht in der kleinen Stichprobe vorkamen (Moody & Mark, 2001), (Goldberger, et al., 2000).

Alle Aufnahmen wurde mit 360 Samples pro Sekunde für jeden Kanal mit 11-Bit Auflösung über einen 10 mV-Bereich digitalisiert. Dies entspricht bei einer halbstündigen Aufnahme 648.000 Samples pro Kanal. Zwei oder mehrere Kardiologen kommentierten unabhängig voneinander jeden Datensatz (Moody & Mark, 2001), (Goldberger, et al., 2000).

4.5 Kurze oder lange Zeitserien

Es ist sehr schwierig, eine formale Definition für "kurze" oder "lange" Zeitserien zu finden. Intuitiv würde man bei "kurzen" Zeitserien Genexpressionsprofile aus der Genanalyse nennen können, die 10 - 40 Datenpunkte enthalten. Auch individuelle EKG-Aufnahmen mit 100 - 1000 Datenpunkte wären ein mögliches Beispiel dafür. Im Gegensatz dazu kann man dreiminütige EKG-Aufnahmen oder eine fünftägige Telemetrieaufnahme eines Sensors als "lange" Zeitserien bezeichnen.

4.6 Testbeschreibung

Für das Herausarbeiten von verschiedenen CGR-Bitmaps für unterschiedliche Herzrhythmen bzw. Herzrhythmusstörungen werden kurze und lange Zeitserien mit einer Länge von 2s, 10s, 1min und 3min genutzt. Im Anschluss daran werden CGR-Bitmaps über die Gesamtlänge der Aufzeichnungen (halbstündige Aufnahmen) erstellt und mit den Kurzzeitserien verglichen.

Alle Bitmaps werden mit einem SAX-Suffix der Länge 3 und mit einer PAA-Segmentlänge 2 erstellt. Der Grund für die kleinen Werte sind eine gute Vergleichbarkeit zwischen kurzen und langen Zeitserien zu erreichen. Wählt man eine zu große SAX-Suffix-Länge, beispielsweise 7, sinkt die Anzahl der übereinstimmenden SAX-Suffixe in der Zeitserie. Eine Farbunterscheidung einzelner Kästchen im Bitmap ist dann nicht mehr möglich. Ähnlich verhält es sich mit der PAA-Segmentlänge. Je

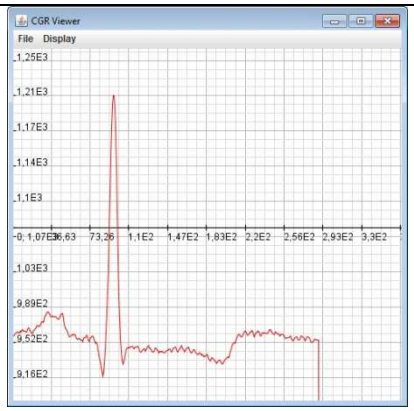
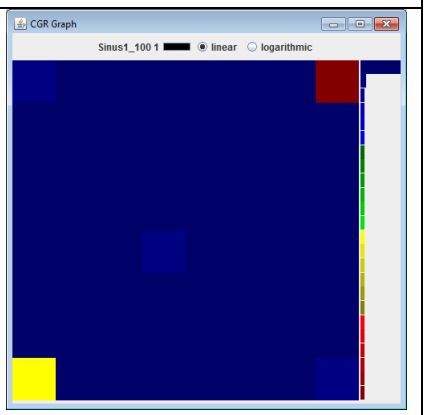
länger die PAA-Segmentlänge gewählt wird, desto schlechter ist die Approximierung der Zeitserie. Dies wirkt sich wiederum negativ auf die Länge der SAX-Sequenz aus. Je kürzer die SAX-Sequenz ist, desto weniger SAX-Suffixe gibt es, und desto homogener ist die Farbverteilung der einzelnen Kästchen im Bitmap.

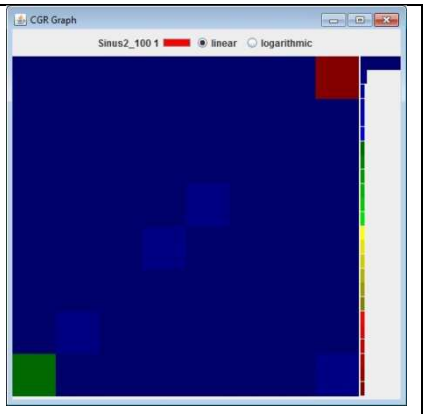
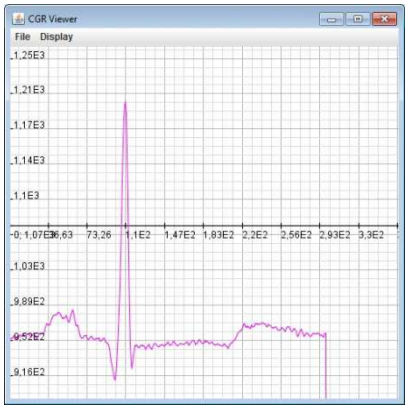
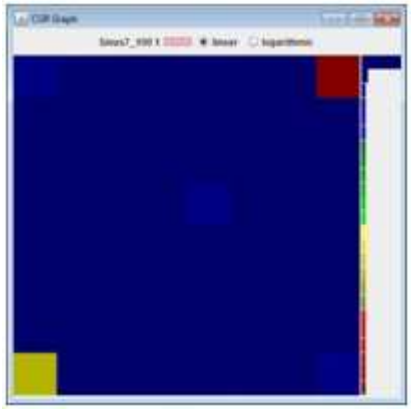
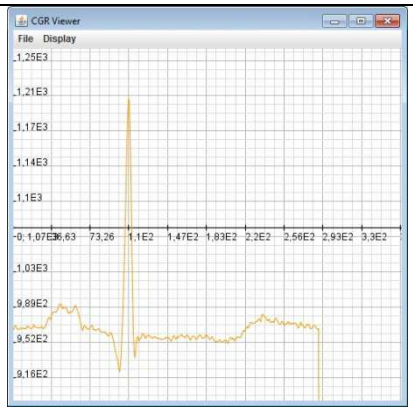
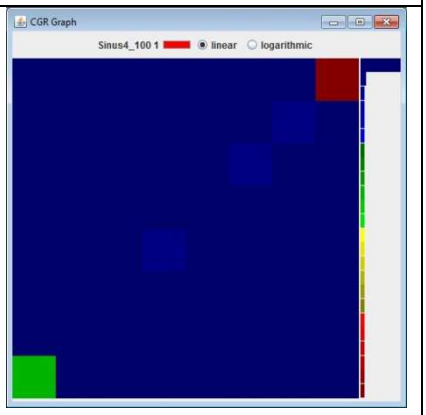
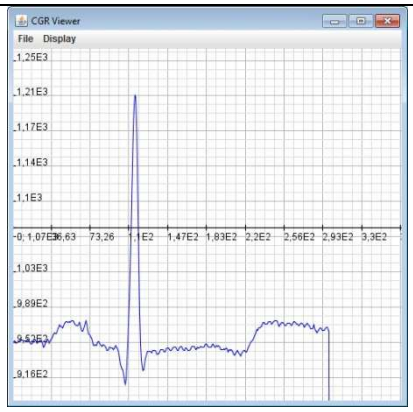
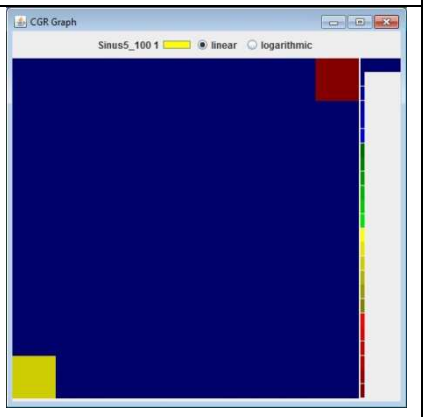
Die Tests wurden mit der ersten und zweiten EKG-Ableitung durchgeführt. Exemplarisch wird die erste Ableitung zur Darstellung genutzt.

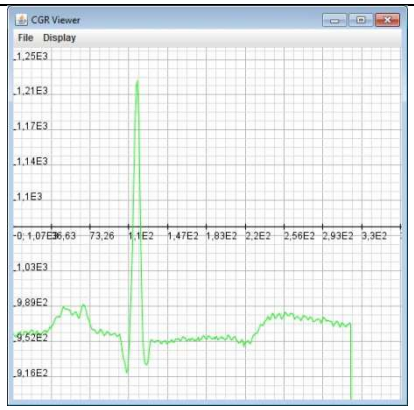
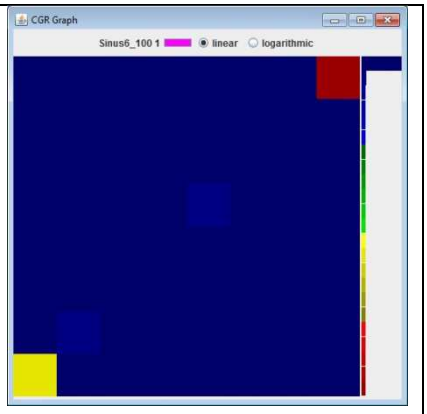
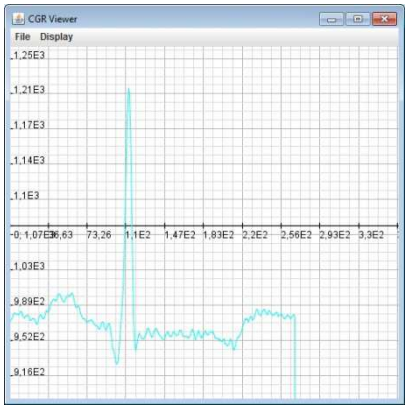
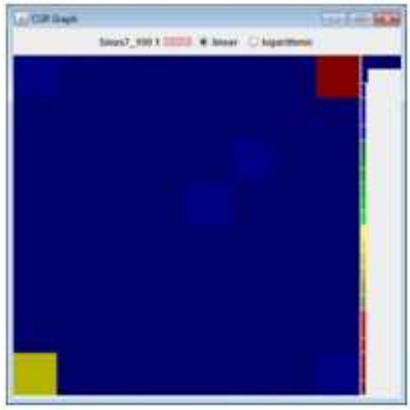
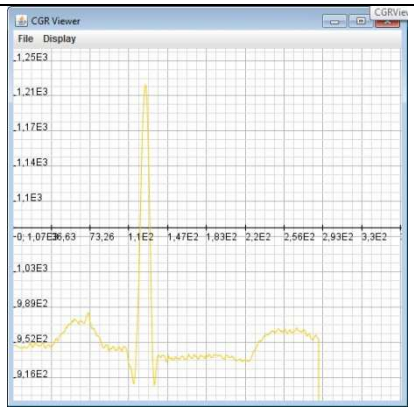
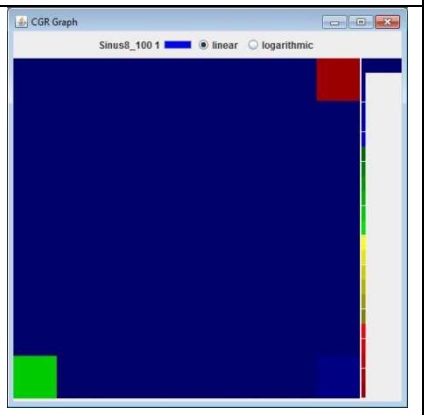
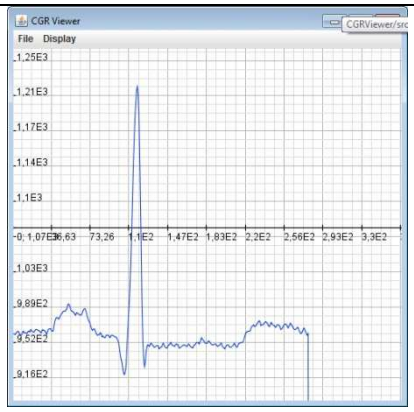
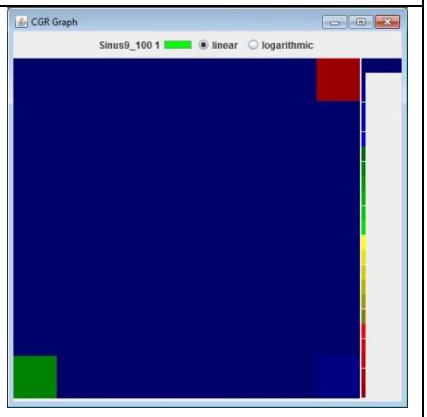
4.6.1 Testreihe 1

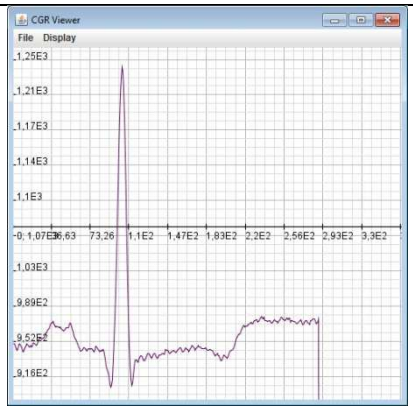
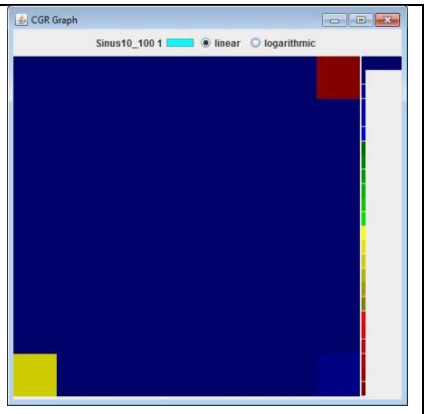
Tabelle 9 zeigt zehn einzelne Herzschläge eines gesunden Herzens. Im CGR-Bitmap sind auf allen vier Eckpunkten Treffer zu erkennen. Deutlich fallen die rechte obere und die linke untere Ecke mit den Farben rot sehr häufig und grün bzw. gelb weniger häufig ins Auge. Diese entsprechen den mittleren Höhen und Tiefen einer Kurve. Im CGR-Bitmap treten weniger deutlich das Maximum und Minimum einer Kurve hervor, die vom rechten unteren und linken oberen Kästchen in blau angezeigt werden. Vereinzelte Treffer sind auf der Diagonalen zwischen linken unteren und rechten oberen Kästchen zu erkennen.

Tabelle 9: Zehn einzelne Herzschläge eines gesunden Herzens als Linienplots und den dazugehörigen CGR-Bitmaps in linearer Farbdarstellung

Anzahl		Linienplot	CGR-Bitmap
Herz-schlag	Samples		
1	290		

1	290		
1	270		
1	290		
1	300		

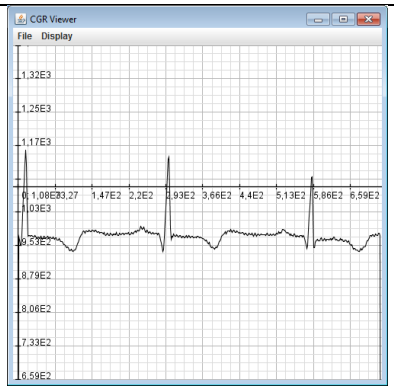
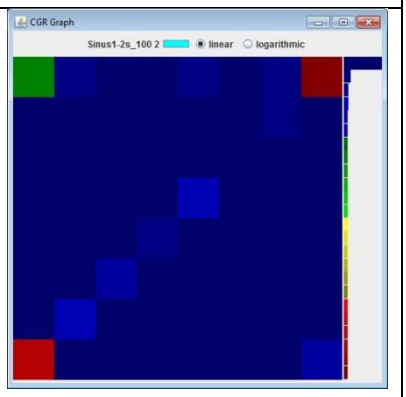
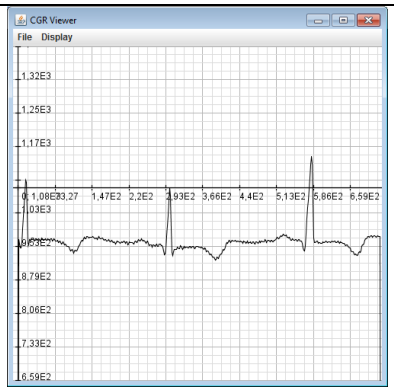
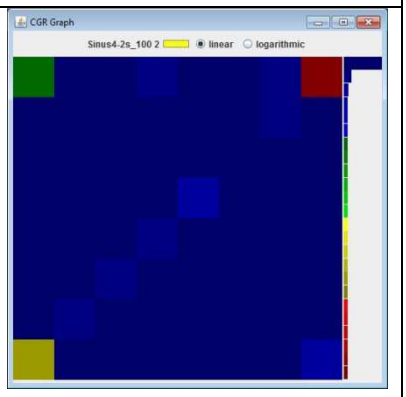
1	320		
1	270		
1	290		
1	280		

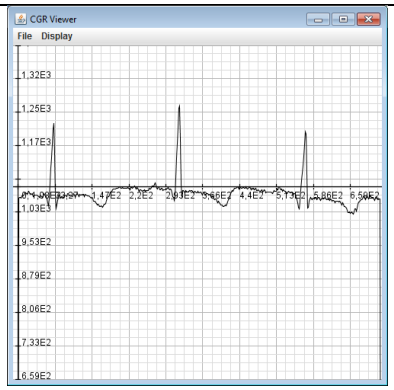
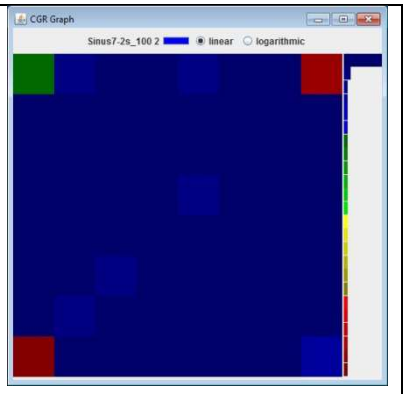
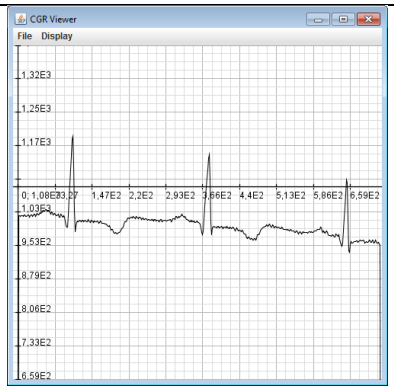
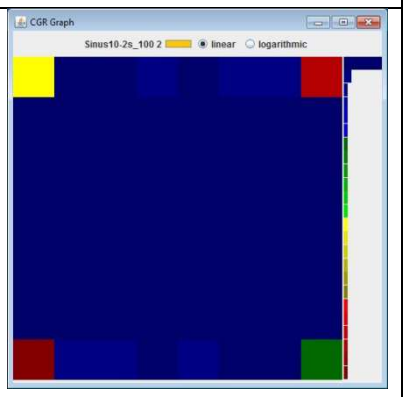
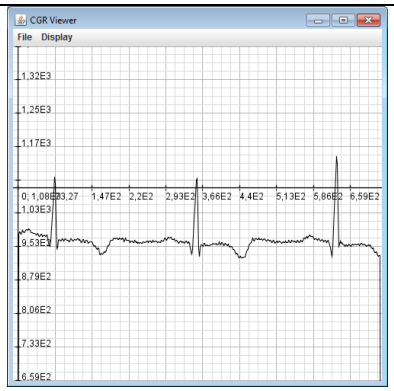
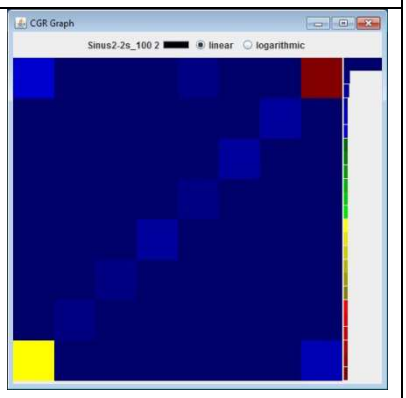
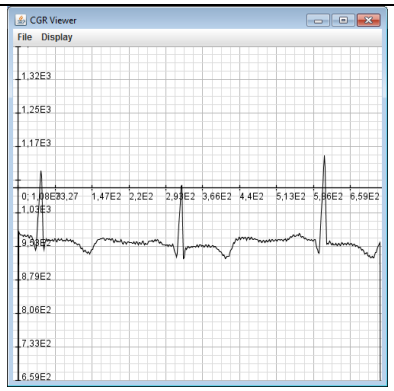
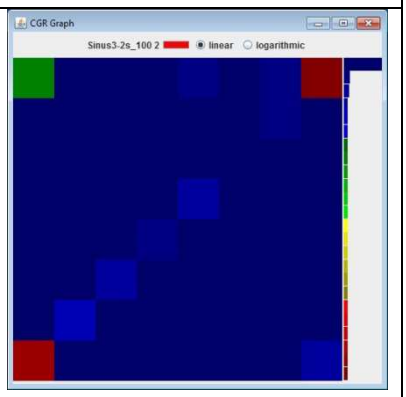
1	290		

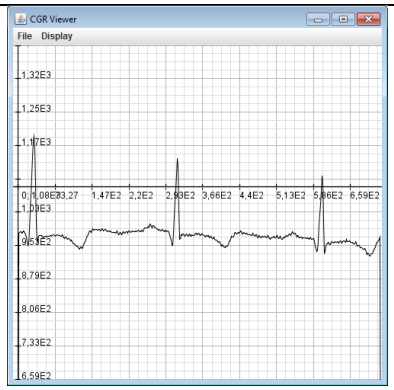
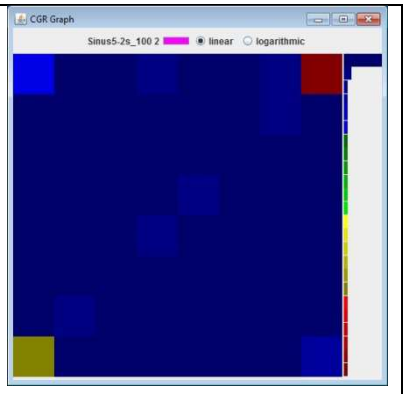
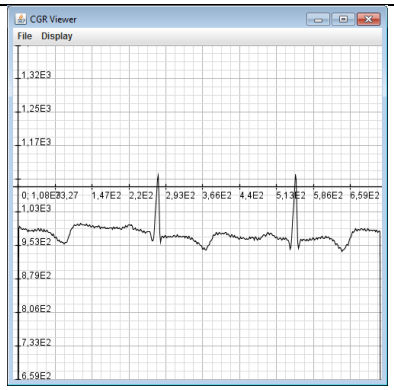
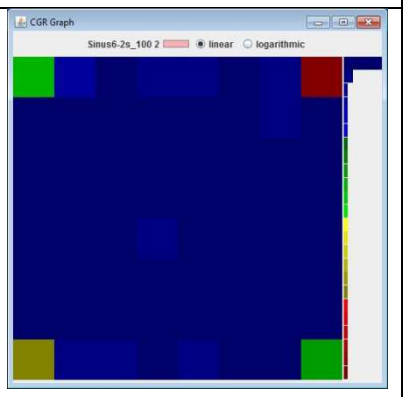
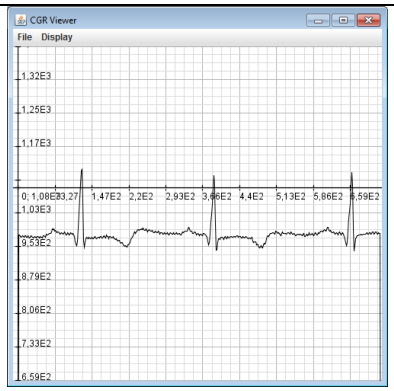
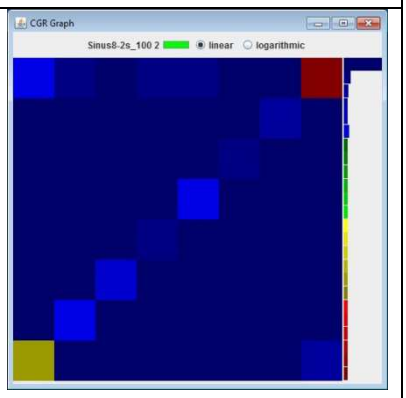
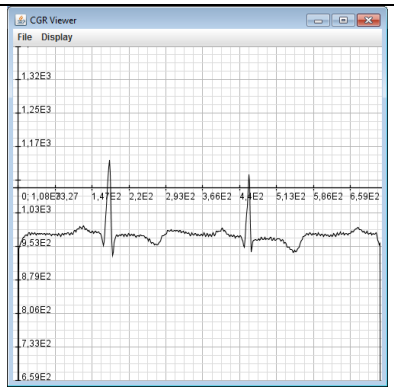
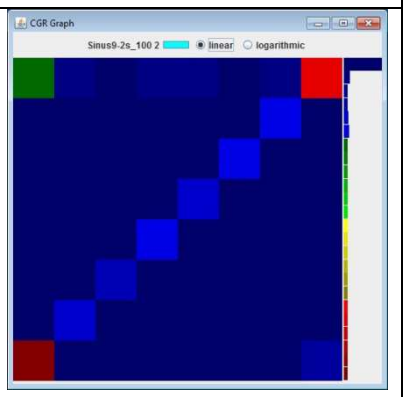
4.6.2 Testreihe 2

Tabelle 10 zeigt zehn zweisekündige Intervalle verschiedener Sinusrhythmen eines gesunden Herzens. Die Herzrhythmen liegen zwischen 74 und 80 Schläge pro Minute. In den CGR-Bitmaps sind auf allen vier Eckpunkten Treffer zu erkennen. Auf der Diagonale von links unten nach rechts oben sowie auf den Vertikalen zwischen den oberen und unteren Eckpunkten sind vermehrt Treffer zu erkennen.

Tabelle 10: Zwei Sekunden Intervalle verschiedener Sinusrhythmen eines gesunden Herzens als Linienplots und den dazugehörigen CGR-Bitmaps in linearer Farbdarstellung

Herzrhythmus - Schläge pro Minute	Linienplot	CGR-Bitmap
77		
74		

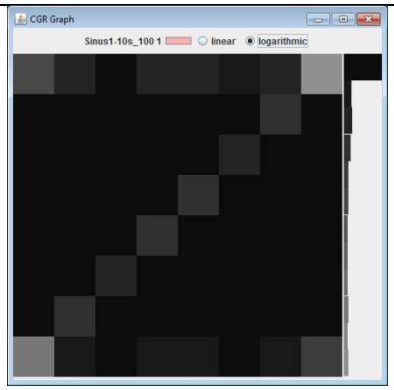
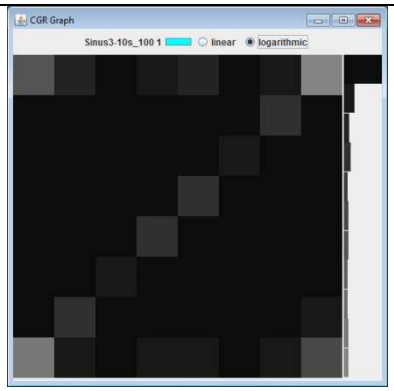
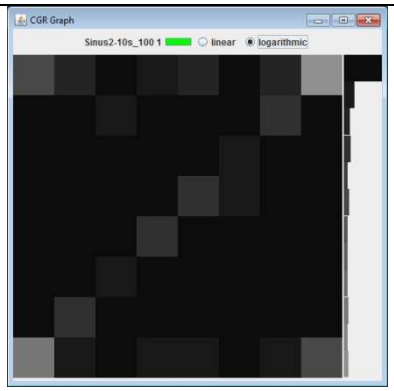
74	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus7-2s_100 2 linear logarithmic
78	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus10-2s_100 2 linear logarithmic
74	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus2-2s_100 2 linear logarithmic
80	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus3-2s_100 2 linear logarithmic

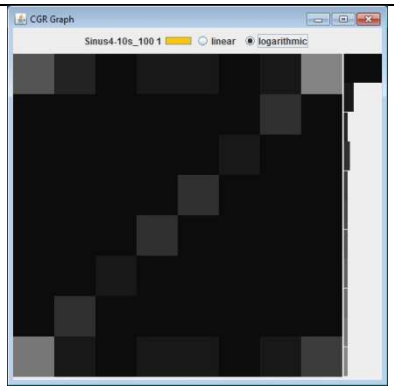
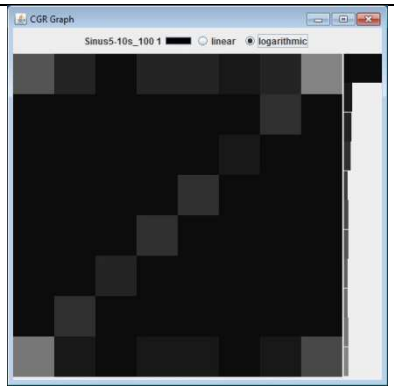
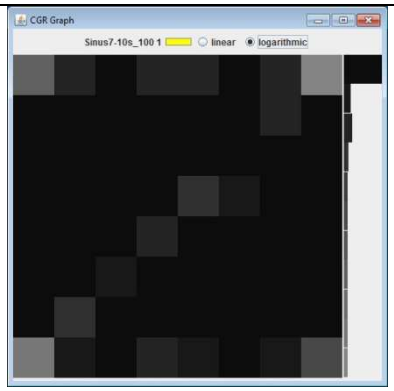
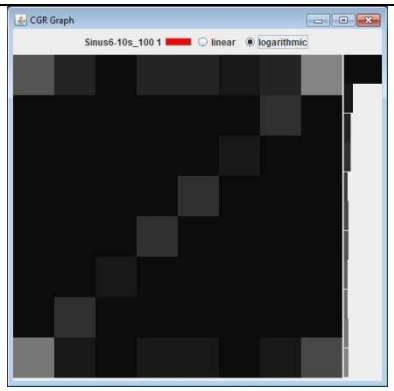
74	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus5-2s_100 2 linear logarithmic
77	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus6-2s_100 2 linear logarithmic
80	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus8-2s_100 2 linear logarithmic
78	 CGR Viewer File Display 1,32E3 1,25E3 1,17E3 1,03E3 9,53E2 8,79E2 8,06E2 7,33E2 6,59E2	 CGR Graph Sinus9-2s_100 2 linear logarithmic

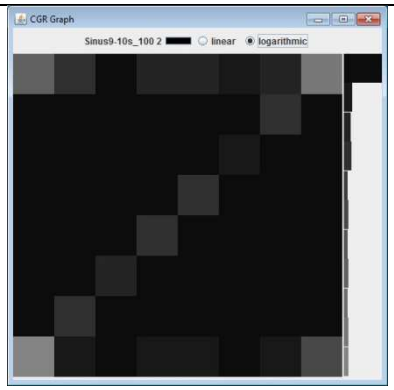
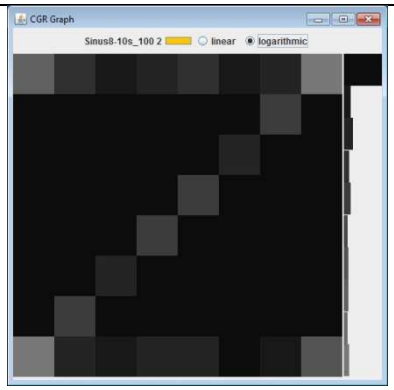
4.6.3 Testreihe 3

Tabelle 11 zeigt zehn zehnsekündige Intervalle verschiedener Sinusrhythmen eines gesunden Herzens, die als CGR-Bitmaps dargestellt sind. Die Herzrhythmen liegen zwischen 72 und 84 Schläge pro Minute. Für eine bessere Unterscheidung der Kästchen mit Treffern wurde die logarithmische Darstellung gewählt. Es ist zu erkennen, dass sich eine Art "Z-Formation" als Trefferbild herauskristallisiert.

Tabelle 11: Zehn Sekunden Intervalle verschiedener Sinusrhythmen eines gesunden Herzens als CGR-Bitmaps in logarithmischer Farbdarstellung

Herzrhythmus - Schläge pro Minute	CGR-Bitmap
72	
72	
78	

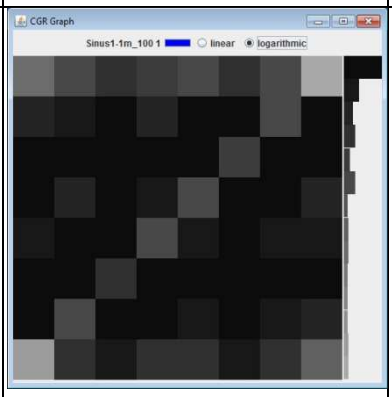
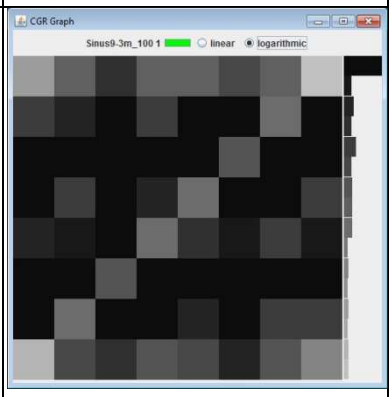
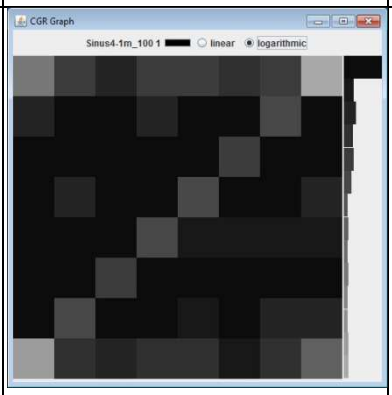
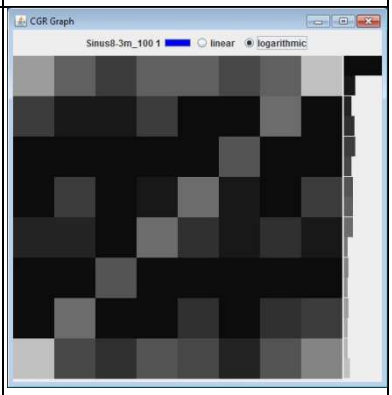
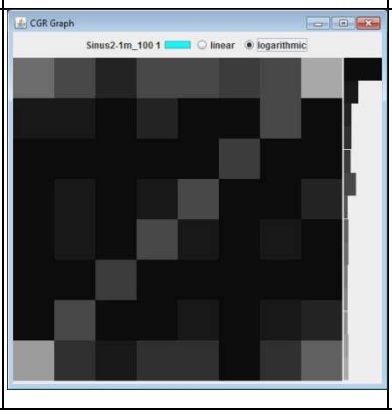
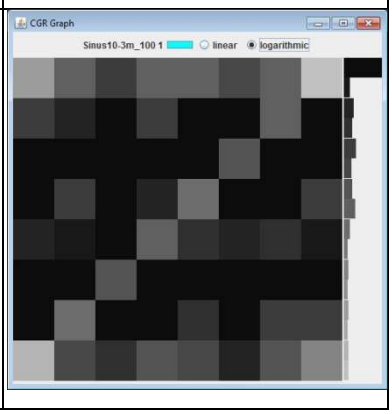
72	 CGR Graph Sinus4-10s_100 1 linear logarithmic
72	 CGR Graph Sinus5-10s_100 1 linear logarithmic
84	 CGR Graph Sinus7-10s_100 1 linear logarithmic
72	 CGR Graph Sinus6-10s_100 1 linear logarithmic

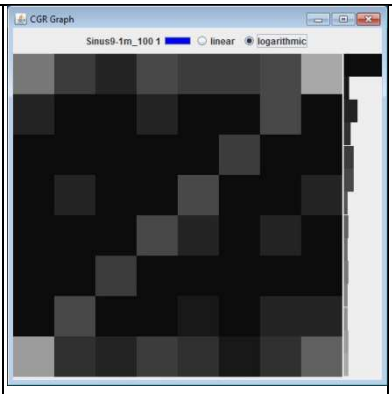
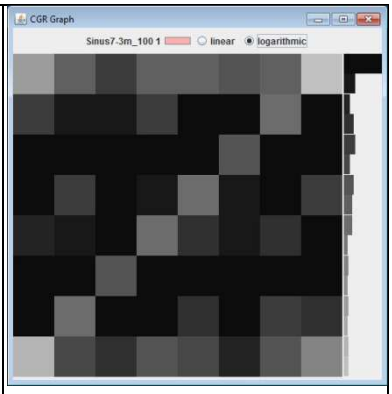
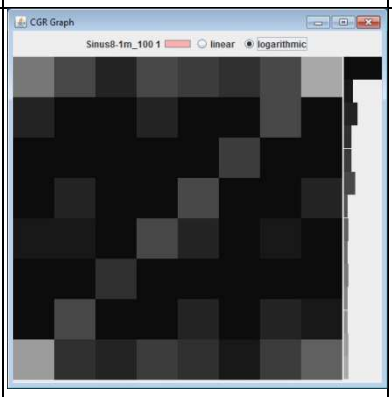
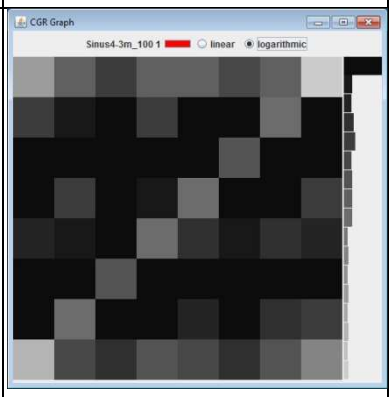
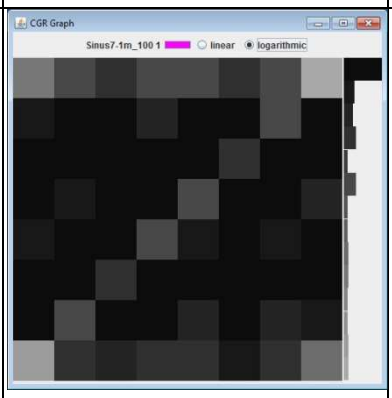
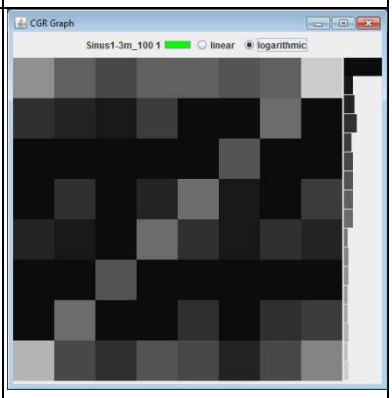
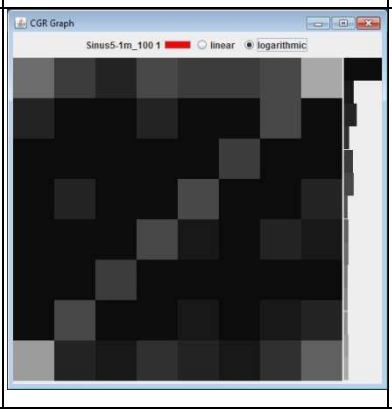
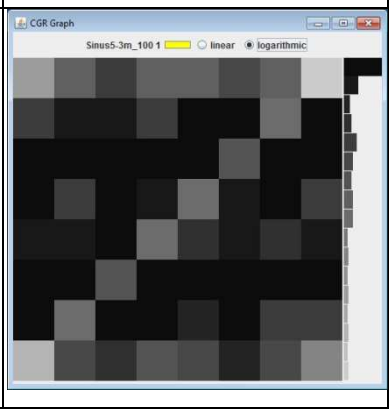
78	
78	
78	

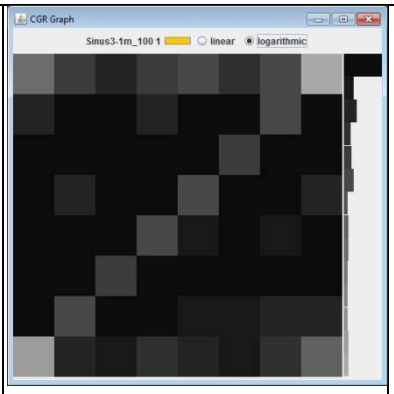
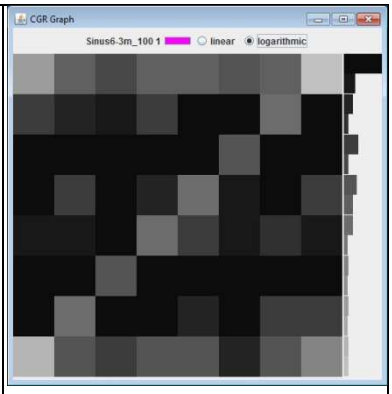
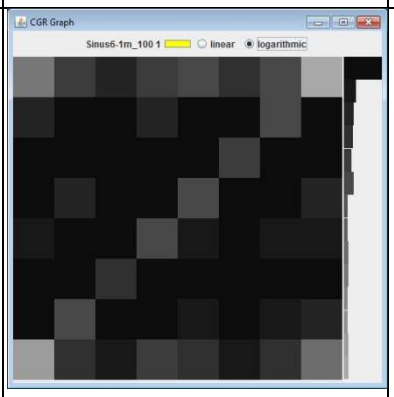
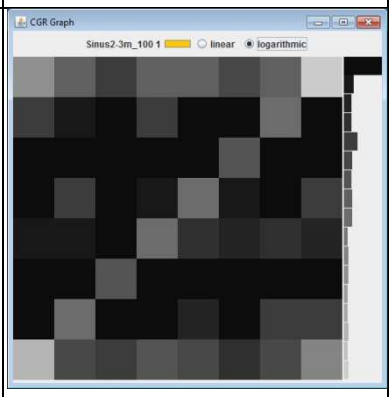
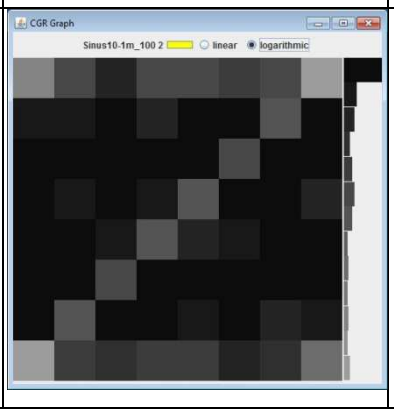
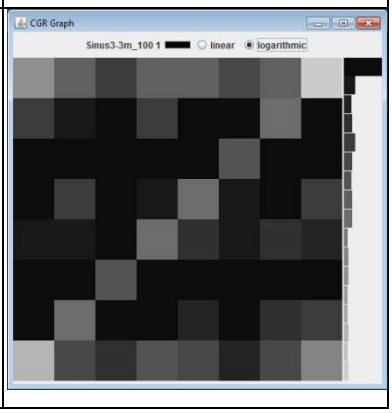
4.6.4 Testreihe 4

Tabelle 12 zeigt Intervalllängen von einer und drei Minuten verschiedener Sinusrhythmen eines gesunden Herzens, die als CGR-Bitmaps dargestellt sind. Die Herzrhythmen liegen zwischen 74 und 80 Schläge pro Minute. Es ist zu erkennen, dass sich die "Z-Formation" mit zunehmender Intervalllänge stärker hervorhebt. Auch im Umfeld der Formation sind Treffer zu erkennen.

Tabelle 12: eine Minute und drei Minuten Intervalle verschiedener Sinusrhythmen eines gesunden Herzens als CGR-Bitmaps in logarithmischer Farbdarstellung

Herzrhythmus - Schläge pro Minute	CGR-Bitmap (1 Minute)	Herzrhythmus - Schläge pro Minute	CGR-Bitmap (3 Minuten)
74		76	
74		77	
76		77	

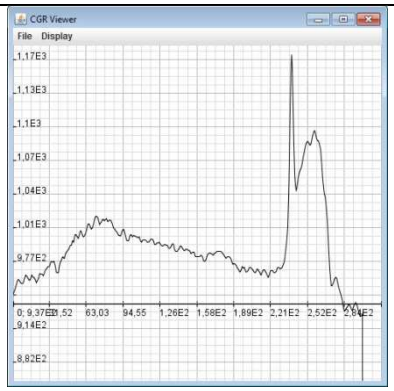
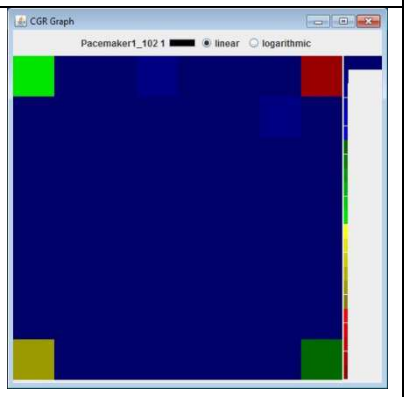
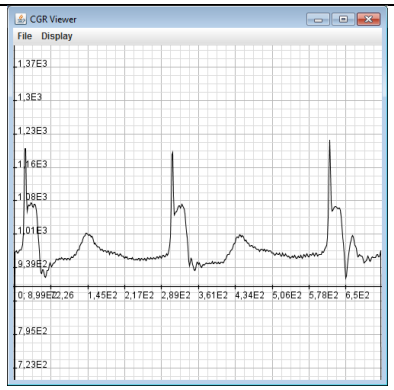
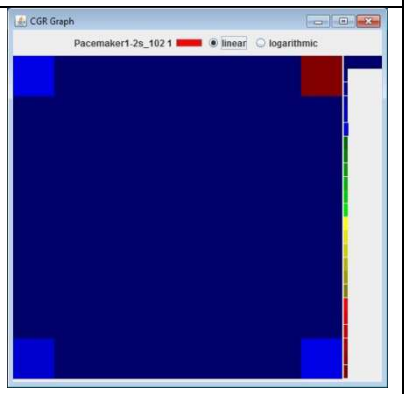
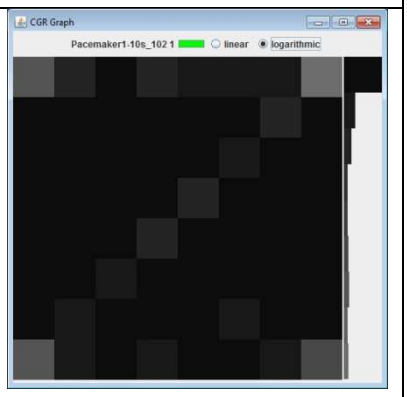
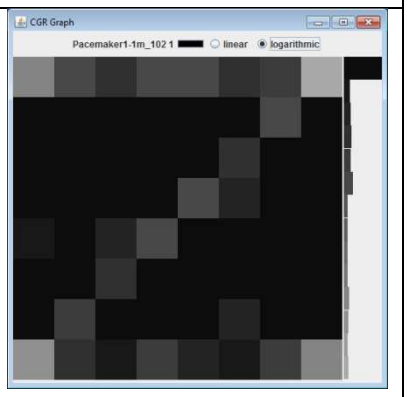
77		80	
80		74	
78		75	
74		75	

75		78	
77		74	
77		75	

4.6.5 Testreihe 5

Tabelle 13 (Herzschrittmacher), 14 (Kammerflattern), 15 (Vorhofflimmern) und 16 (Sinusbradykardie) zeigen verschiedene Arrhythmien mit unterschiedlichen Intervalllängen, die als CGR-Bitmap dargestellt sind . Für einen Herzschlag und einem zwei Sekundenintervall sind die jeweiligen Linienplots mit aufgeführt.

Tabelle 13: CGR-Bitmaps verschiedener Intervalllängen eines Herzrhythmus mit einsetzendem Herzschrittmacher. Bei den Intervalllängen ´ein Herzschlag´ und ´2 Sekunden´ ist der dazugehörige Linienplot zu sehen.

Intervalllänge	Linienplot	CGR-Bitmap
ein Herzschlag		
2 Sekunden		
10 Sekunden		
1 Minute		

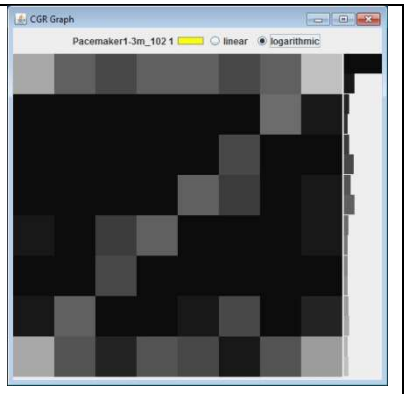
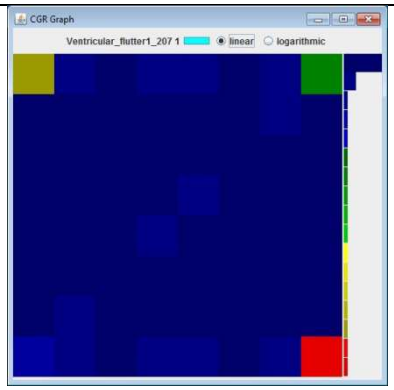
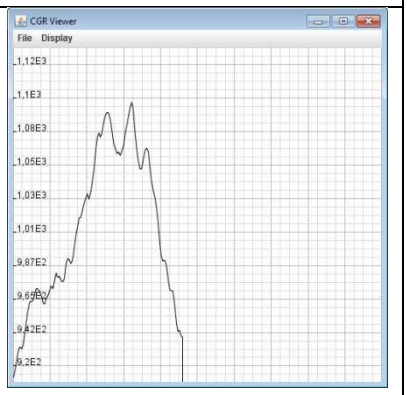
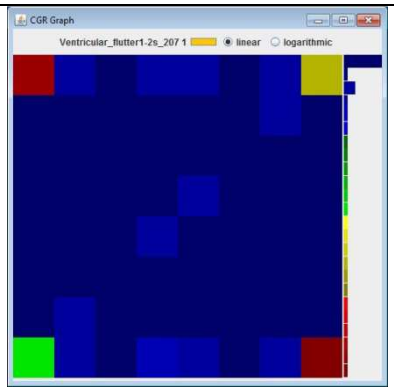
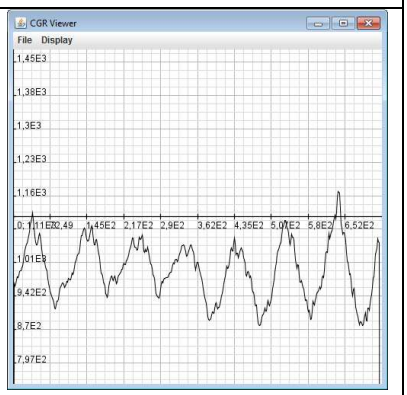
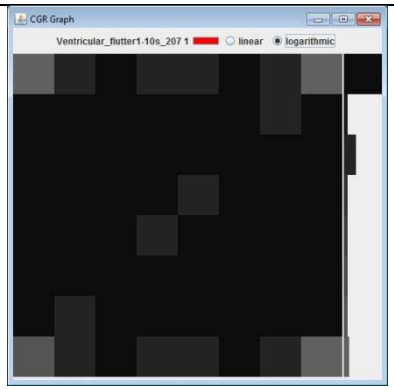
3 Minute		

Tabelle 14: CGR-Bitmaps verschiedener Intervalllängen eines Herzrhythmus mit Kammerflattern. Bei den Intervalllängen ´ein Herzschlag´ und ´2 Sekunden´ ist der dazugehörige Linienplot zu sehen.

Intervalllänge	CGR-Bitmap	Linienplot
1 Herzschlag		
2 Sekunden		
10 Sekunden		

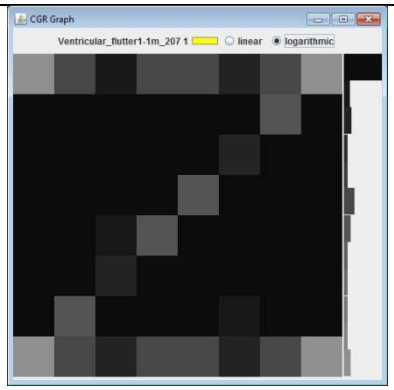
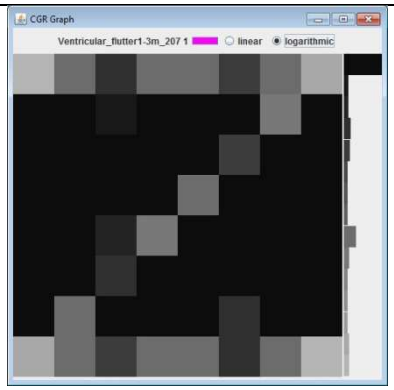
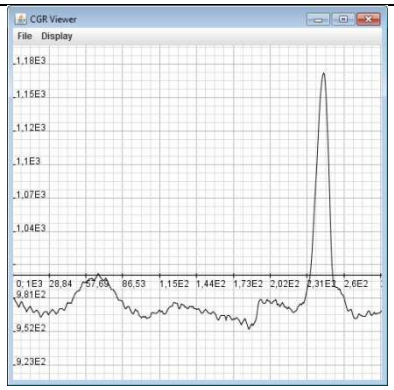
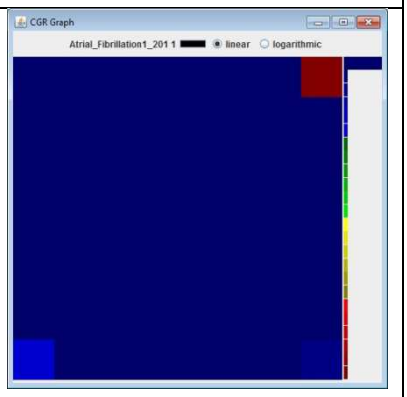
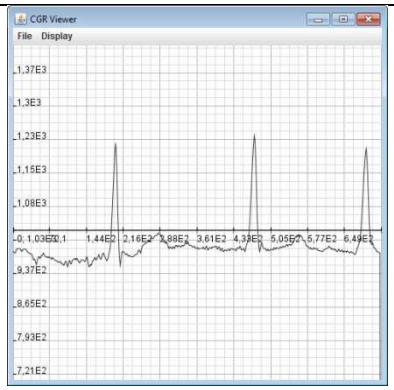
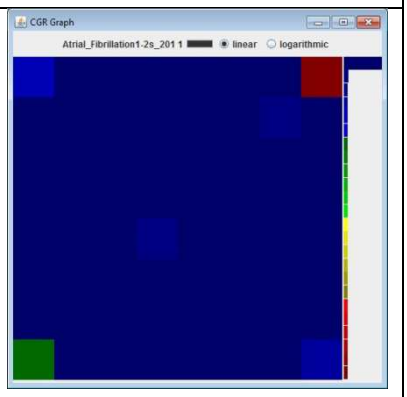
1 Minute		
3 Minuten		

Tabelle 15: CGR-Bitmaps verschiedener Intervalllängen eines Herzrhythmus mit Vorhofflimmern. Bei den Intervalllängen ´ein Herzschlag´ und ´2 Sekunden´ ist der dazugehörige Linienplot zu sehen.

Intervalllänge	Linienplot	CGR-Bitmap
1 Herzschlag		
2 Sekunden		

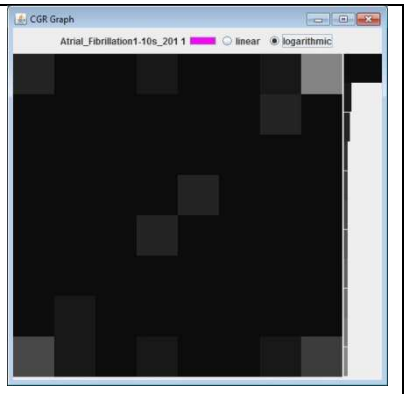
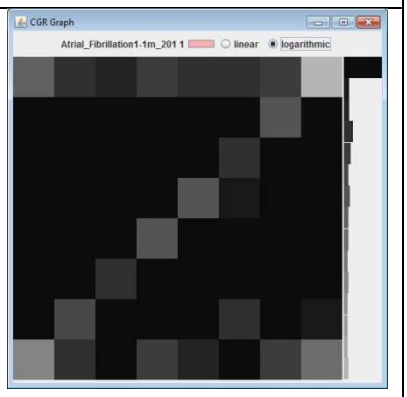
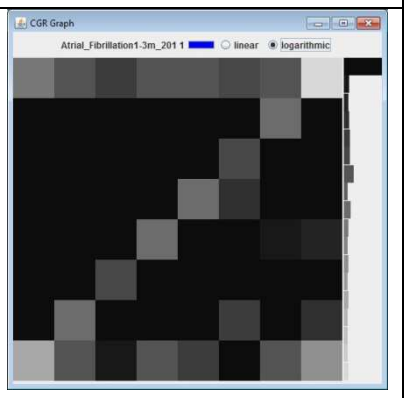
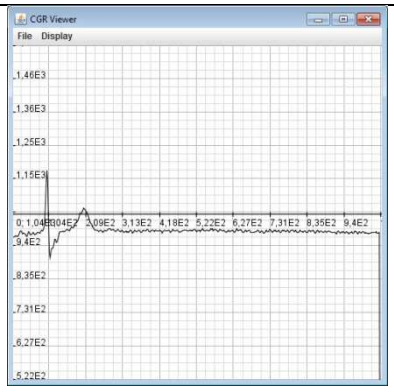
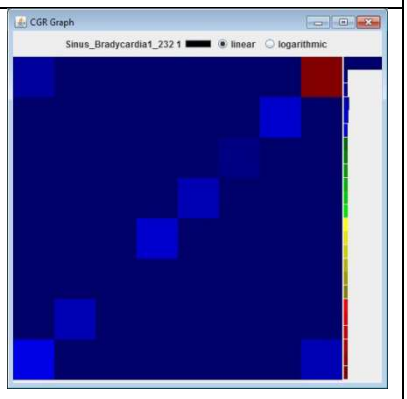
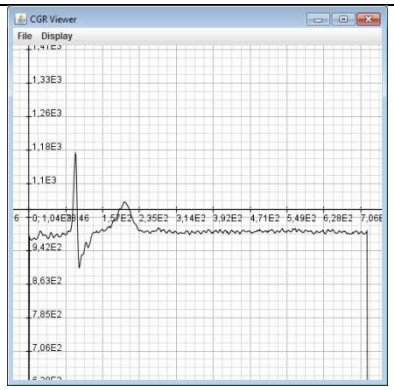
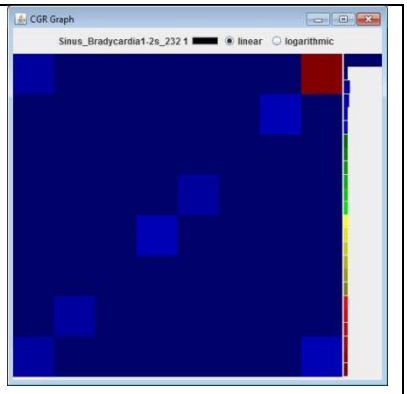
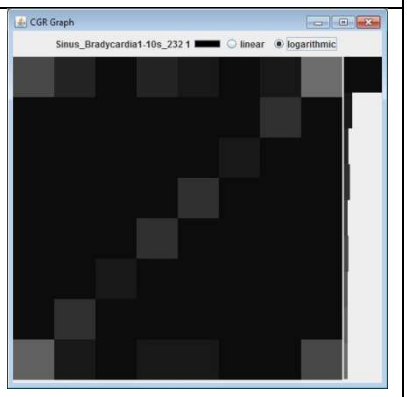
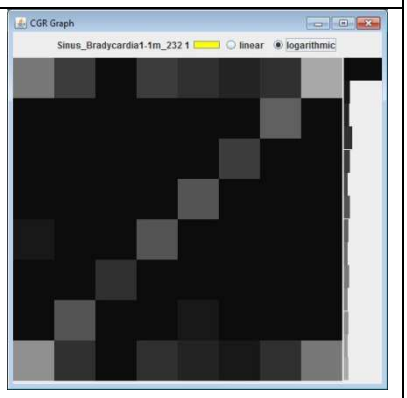
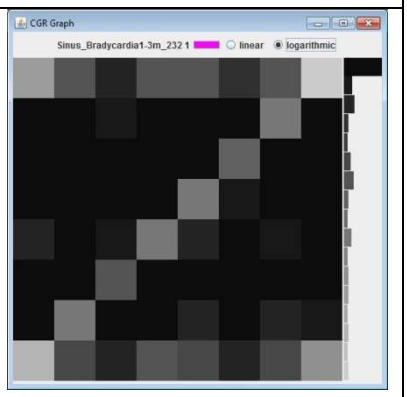
10 Sekunden		
1 Minute		
3 Minuten		

Tabelle 16: CGR-Bitmaps verschiedener Intervalllängen eines Herzrhythmus mit Sinusbradykardien. Bei den Intervalllängen ´ein Herzschlag´ und ´2 Sekunden´ ist der dazugehörige Linienplot zu sehen.

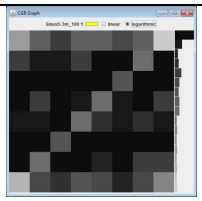
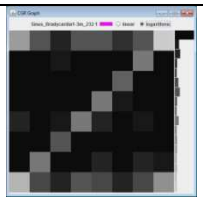
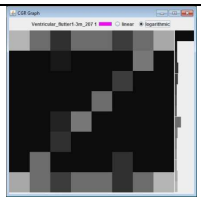
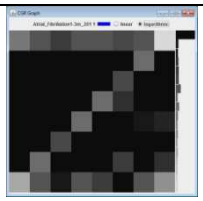
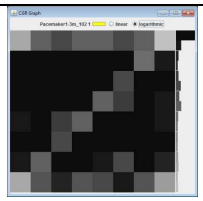
Intervalllänge	Linienplot	CGR-Bitmap
1 Herzschlag		

2 Sekunden		
10 Sekunden		
1 Minute		
3 Minuten		

4.6.6 Testreihe 6

Tabelle 17 zeigt dreiminütige Sequenzen von verschiedener Arrhythmien, die als CGR-Bitmap dargestellt sind. Zum Vergleich wurde ein CGR-Bitmap eines normalen Sinusrhythmus mit hinzugenommen. Deutlich zu erkennen sind bei allen 5 CGR-Bitmaps die "Z-Formation".

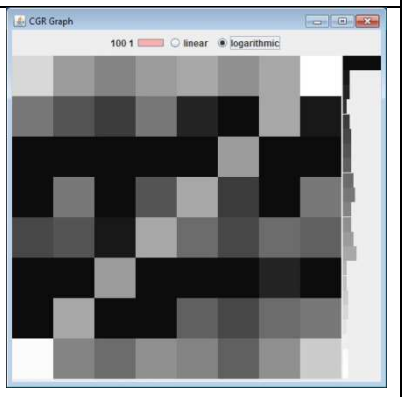
Tabelle 17: CGR-Bitmaps gleicher Länge mit verschiedenen Arrhythmien

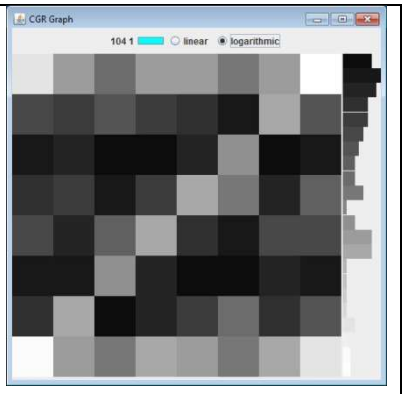
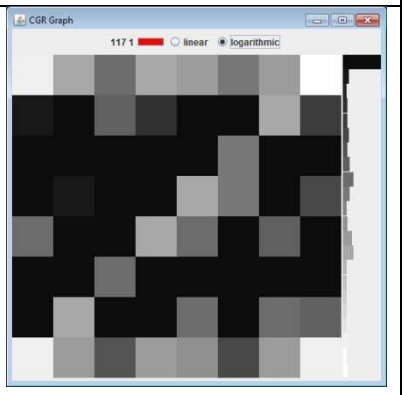
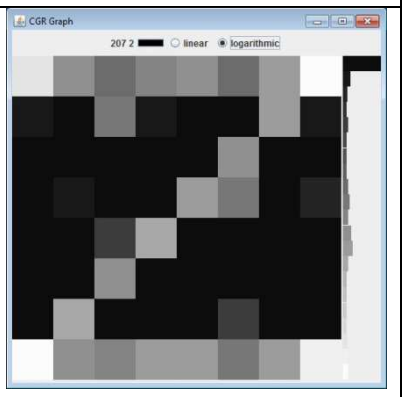
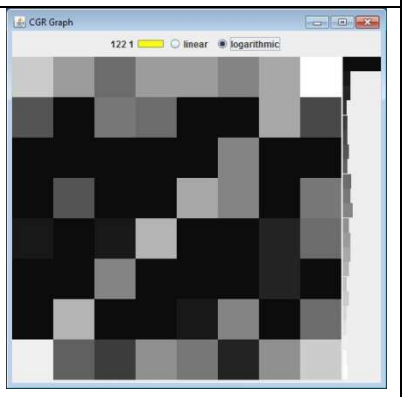
CGR-Bitmap (Normaler Sinusrhythmus)	CGR-Bitmap (Sinusbradykardie)	CGR-Bitmap (Kammerflattern)	CGR-Bitmap (Vorhofflimmern)	CGR-Bitmap (Herzschrittmacher)
				

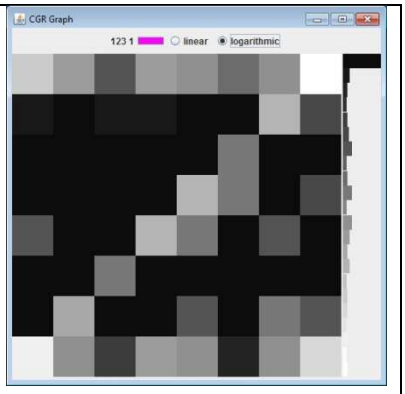
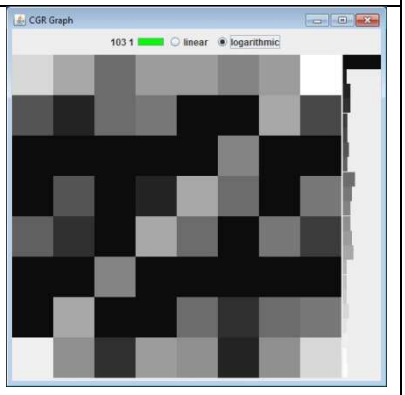
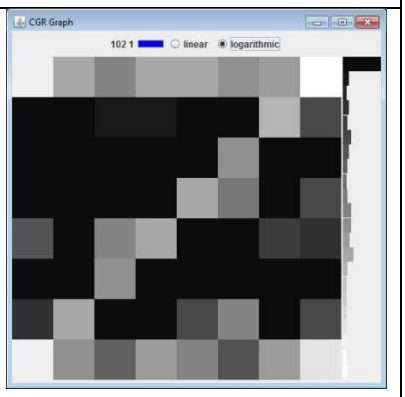
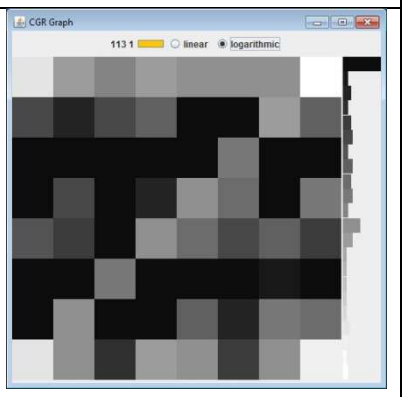
4.6.7 Testreihe 7

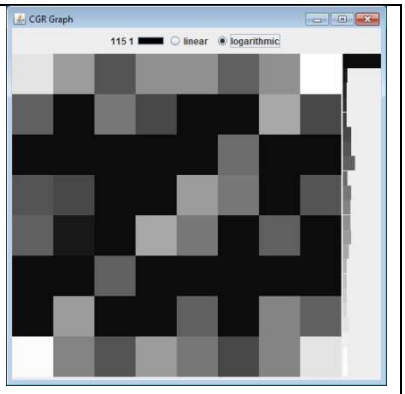
Tabelle 17 zeigt zehn dreißigminütige Sequenzen von verschiedener Arrhythmien, die als CGR-Bitmap dargestellt sind. Innerhalb der EKG-Aufnahmezeit kam es bei den Patienten zu verschiedenen Arrhythmien in unterschiedlicher Ausprägung und Dauer.

Tabelle 18: 30 Minuten Intervalle verschiedener Herzrhythmen mit verschiedenen Arrhythmien unterschiedlicher Dauer als CGR-Bitmaps dargestellt

Arrhythmien	CGR-Bitmap
• Normaler Sinusrhythmus • Herzgeräusche • Artefakte • Vorhofextrasystolen	

• Normaler Sinusrhythmus • Kammerflimmern • Herzgeräusche • Herzschrittmacher	
• Normaler Sinusrhythmus • Herzgeräusche • Vorhofextrasystolen	
• Ventrikuläre Bigeminie • Ventrikuläre Couplets • Ventrikuläre Tachykardien • Kammerflattern • Normaler Sinusrhythmus • Herzgeräusche • Arterielle Couplets	
• Normaler Sinusrhythmus • Herzgeräusche	

• Sinusarrhythmien • Vorhofextrasystolen	
• Normaler Sinusrhythmus • Herzgeräusche • Vorzeitige Vorhofkontraktionen	
• Normaler Sinusrhythmus • Herzschrittmacher • Vorhofextrasystolen	
• Basline Wander • Vorzeitige Vorhofkontraktionen • Sinusarrhythmien	

• Normaler Sinusrhythmus • Sinusarrhythmien • Basline Wander • Artefakte	

4.7 Testauswertung

Der Ausgangspunkt der Testreihen 1 bis 4 ist eine ganz normale Erregungsleitung eines gesunden Menschenherzens, wie sie im Abschnitt 4.2 beschrieben ist. Getestet wurden Kurz- und Langezeitserien.

In der Testreihe 1 habe ich verschiedene Intervalllängen eines Herzschlages als CGR-Bitmap dargestellt. Aufgrund der geringen Längenunterschiede (Samples) und der strukturellen Ähnlichkeit war zu erwarten, dass die CGR-Bitmaps sich sehr ähnelten. Bis auf vereinzelte Treffer in der Diagonalen der CGR-Bitmaps hat sich die Annahme bestätigt. Die Testreihen 2 bis 4 zeigen unterschiedliche Sequenzlängen von 2 und 10 Sekunden sowie von einer und 3 Minuten mit verschiedenen Herzrhythmen. Zu beobachten ist, dass sich auch hier keine deutlichen strukturellen Unterschiede in den CGR-Bitmaps zeigen. Aufgrund der zunehmenden Sequenzlängen und der damit verbundenen Auftretenswahrscheinlichkeit der einzelnen SAX-Suffixe bildet sich eine "Z-Formation" im CGR-Bitmap immer deutlicher heraus. Auch der Vergleich einzelner CGR-Bitmaps bei gleichen Herzrhythmen ergaben keine wesentlichen Unterschiede. In der Testreihe 5 und 6 vergleiche ich schwere Herzrhythmusstörungen wie das Einsetzten eines Herzschrittmachers, Kammerflattern, Vorhofflimmern und die Sinusbradykardie. Leichte Unterschiede in den Verteilungshäufigkeiten eines Herzschlages sind zu erkennen. Mit zunehmender Länge ist aber kein Unterschied zu anderen CGR-Bitmaps feststellbar. Wie bereits bei den normalen Sinusrhythmen bilden sich auch hier "Z-Formationen" heraus. Die CGR-Bitmaps der Herzrhythmusstörungen und des normalen Sinusrhythmus sind zu ähnlich, um eine Aussage über eine spezifische Herzerkrankung treffen zu können. Eine Unterscheidung ist damit ausgeschlossen. In der Testreihe 7 vergleiche ich 30 Minuten Intervalle verschiedener Herzrhythmen mit verschiedenen Arrhythmien von unterschiedlicher Dauer.

Unterschiede in der Struktur der CGR-Bitmaps kann ich erkennen. In den Nebenbereichen der "Z-Formation" gibt es verschieden häufige Treffer an unterschiedlichen Stellen. Jedoch können aufgrund der Komplexität der einzelnen Herzerkrankungen und den daraus resultierenden CGR-Bitmaps keine Rückschlüsse auf die Erkrankung selbst oder auf die Herzfrequenz gezogen werden.

Aus meiner Sicht können mehrere Details für die Unwirksamkeit der CGR-Bitmaps bei EKG-Zeitserien verantwortlich sein:

Als Erstes ist die Definition des Sinusrhythmus mit seiner Frequenz zu nennen, die das Erkennen eines Krankheitsbildes unmöglich macht. Nach Definition wird ein Mensch als gesund bewertet, wenn die Herzfrequenz zwischen 60 und 100 Schlägen in der Minute beträgt. Auch das über- oder unterschreiten der normalen Herzfrequenzen müssen nicht automatisch auf eine Erkrankung hinweisen. Beispielsweise sinkt der Herzrhythmus im Schlaf unter 60 Herzschläge pro Minute oder bei vielen sportlichen Aktivitäten steigt die Herzfrequenz oft über 100 Herzschläge an.

Zweitens ist die Veränderung der Wellen und Zacken eines normalen Sinusrhythmus zu nennen. Kein normales Sinusbild gleicht im Detail einem anderen. Kleinste Unterschiede können über ein gesundes Herz oder über krankhafte Veränderungen entscheiden.

Der dritter Punkt ist die Ableitung und dessen Darstellung. Man kann einen Sinusrhythmus darstellen, der von der normalen Wellenform teilweise ganz erheblich abweicht, so dass bei gesunden Menschen mit einem sinusgleichen Sinusrhythmus viele verschiedene Bilder herauskommen und damit keine Aussage über den Gesundheitszustand mittels CGR-Bitmap getroffen werden kann.

Der Vergleich von Herzfrequenzen und verschiedenen einzelnen Wellen des EKG sind voneinander zu unterscheiden. Die Wellenveränderung von unterschiedlichen Einzelbildern (ein Herzschlag - CGR-Bitmap) ist noch miteinander vergleichbar. Nimmt man die Frequenz hinzu, wird die Interpretation der CGR-Bitmaps unüberschaubar und unklar, so dass keine Aussagen mehr getroffen werden können.

Kapitel V: Diskussion und Ausblick

In dieser Arbeit habe ich potentiell lange und kurze Zeitserien mit Hilfe von SAX und CGR untersucht. Dazu habe ich das Programm CGR Viewer geschrieben und als GUI programmiert, um eine mausgestützte und visuelle Möglichkeit zur Bedienung und Ansicht zu haben. Das Laden verschiedener Zeitserien, die Darstellung der Zeitserien als klassische Liniendiagramme und als CGR-Bitmap ist mit dem Programm möglich. Wichtige Parameter für die Darstellung der CGR-Bitmaps, wie zum Beispiel die SAX-Suffixlänge oder die PAA-Länge, sind frei wählbar. Das Hauptaugenmerk der Arbeit war die Implementierung von Interaktionsmöglichkeiten zwischen CGR und Liniendiagrammen, dem sogenannten "Brushing & Linking", die eine Zeitpunkt- oder Feldauswahl ermöglichen und die dazugehörigen Felder in den einzelnen Windows anzeigen. Am Ende der Arbeit habe ich mit dem CGR Viewer auf der Basis von potentiell langen und kurzen EKG-Aufzeichnungen verschiedene Tests durchgeführt. Dabei stellte sich heraus, dass lange EKG-Zeitserien, die als CGR-Bitmaps dargestellt wurden, sich nicht zur Analyse von Herzerkrankungen eignen. Hauptgründe dafür liegen in der Definition des Sinusrhythmus und seiner Frequenz und die Darstellungen der Ableitungen. Vergleicht man jedoch einzelne Herzschläge von gesunden und erkrankten Menschen, sind Unterschiede in der Struktur bzw. in der Häufigkeit der Verteilungen einzelner Kästchen im CGR-Bitmap zu erkennen.

Die Vergangenheit hat gezeigt, dass CGR-Bitmaps zur Darstellung und Untersuchung von DNA-Sequenzen genutzt werden und unbekannte Strukturen und Muster aufzeigen. Diese Muster und Strukturen sind meines Erachtens auch in langen Zeitserien aus der Wirtschaft zu finden, wie der sich sekündlich ändernde DAX-Chart oder im Leistungssport, bei der Frequenzanalyse der Auf- und Abbewegungen von Pedalumdrehungen beim Radfahren. Um eine bessere Interpretation von CGR-Bitmaps zu gewährleisten, wäre aus meiner Sicht eine zuvor durchgeführte Clusteranalyse sehr hilfreich.

Kapitel VI: Anhang

A.1 Gleitender Mittelwert

Der gleitende Mittelwert ist eine einfache Methode zur Glättung von Messdaten. Die Menge der gleitenden Mittelwerte werden iterativ, also "gleitend" über einen Abschnitt einer gegebenen Zeitreihe berechnet. Der verwendete Abschnitt wird überlappend verschoben, d.h. wiederholt wird der erste Wert aus dem betrachteten Abschnitt gestrichen und der erste Wert nach dem Abschnitt hinzugenommen. Die im Abschnitt vorkommenden Werte können gewichtet in den resultierenden Mittelwert eingehen.

Der einfache gleitende Mittelwert (Kreiß & Neuhaus, 2006) n-ter Ordnung einer diskreten Zeitreihe $x(t)$ ist die Folge der arithmetischen Mittelwerte der jeweils letzten n aufeinanderfolgender Datenpunkte:

$$m(t) = \frac{1}{n} \sum_{i=0}^{n-1} x(t-i)$$

A.2 Lookup Table

Eine Lookup Table ist eine Tabelle, die Informationen statisch definiert und diese zur Laufzeit eines Programms zur Verfügung stellt. Das Ziel dabei ist, aufwändige Berechnungen während der Laufzeit eines Programms oder den hohen Speicherplatzbedarf zu vermeiden.

A.3 Quartärbaum

Ein Quartärbaum ist in der Graphentheorie eine spezielle Form eines Graphen. Es handelt sich um einen gewurzelten Baum, bei dem jeder Knoten höchstens vier Kinderknoten hat. Ein vollständiger Quartärbaum der Höhe h heißt vollständig, wenn jeder Knoten einer Tiefe kleiner h genau vier Kinderknoten hat.

A.4 Data Mining

Data Mining ist die systematische Anwendung statistischer Methoden auf einen Datenbestand mit dem Ziel, neue Muster zu erkennen. Hierbei geht es auch um die

Verarbeitung sehr großer Datenbestände, wofür effiziente Methoden benötigt werden, deren Zeitkomplexität sie für solche Datenmengen geeignet macht.

A.5 Wavelet

Wavelet werden die einer kontinuierlichen oder diskreten Wavelet-Transformation zugrundeliegenden Funktionen bezeichnet.

A.6 Datenparser

Ein (Daten)-Parser ist ein Computerprogramm, das für die Zerlegung und Umwandlung einer beliebigen Eingabe in ein für die Weiterverarbeitung brauchbares Format zuständig ist. Häufig werden Parser eingesetzt, um im Anschluss an den Analysevorgang die Semantik der Eingabe zu erschließen und daraufhin Aktionen durchzuführen.

A.7 JMotif

JMotif ist eine in Java implementierte Bibliothek, die eine Reihe von Methoden für das Zeitseriendatenhandling, das Data Mining und die Klassifikation zur Verfügung stellt. Insbesondere setzt JMotif auf ein vollständiges Zeitserien-Data-Mining-Workflow welches SAX nutzt.

A.8 Logarithmische Darstellung

Die logarithmische Darstellung basiert auf einer Skale, die nicht den Wert einer physikalischen Größe verwendet, sondern den Logarithmus ihres Zahlenwerts. Bei der logarithmischen Darstellung werden in einem Diagramm die Werte einer oder mehrerer Achsen logarithmiert aufgetragen. Eine solche Darstellung ist vor allem dann hilfreich, wenn der Wertebereich der dargestellten Daten viele Größenordnungen umfasst. Durch die logarithmische Darstellung werden Zusammenhänge im Bereich der kleinen Werte besser überschaubar. (Wikimedia Foundation Inc., 2013)

Abkürzungsverzeichnis

CGR	Chaos Game Representation
DAX	Deutscher Aktienindex
DNA	Desoxyribonukleinsäure
EEG	Elektroenzephalografie
EKG	Elektrokardiogramm
GBO	Grafische Benutzeroberfläche
GUI	Graphicle User Interface
IFS	Iteratives Funktionensystem
MIT	Massachusetts Institute of Technology
PAA	Piecewise Aggregate Approximation
RNA	Ribonukleinsäure
SAX	Symbolic Aggregate Approximation

Literatur- und Quellenverzeichnis

Androulakis, I. P. (2005). *New approaches for representing, analyzing and visualizing complex kinetic.* Barcelona: Proceedings of the 15th European symposium on computer aided process.

Baransley, M. F. (1988). Fractals Everywher. *Morgan Kaufmann Pub* .

Duchene, F., Garbay, C., & Rialle, V. (2004). *Mining heterogeneous multivariate time-series for learning meaningful patterns: application to home health telecare.* Grenoble: Research Report 1070-I, Institue de Informatique et Mathematiques Appliquees de Grenoble.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., et al. (13. 06 2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation Electronic Pages 101(23)* , S. e215-e220.

Google. (2012). *http://code.google.com/p/jmotif/wiki/PAAapproximation.* Abgerufen am 8. 04 2013 von http://www.google.com

Google. (2012). *http://code.google.com/p/jmotif/wiki/SAX*. Abgerufen am 12. 04 2013 von http://www.google.com

Hartung, J., & Elpelt, B. (1984). *Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik.* München: Oldenbourg Verlag.

Jeffrey, H. J. (1990). Chaos game of gene structure. *Nuckeic Acids Research, Oxford University Press* , S. 2163-2170.

Keogh, E. J. (2001). Dimensionality reduction for fast similarity search in large time series databases. In *Knowledge and Information Systems, vol. 3* (S. 263-286).

Keogh, E. J., & Pazzan, M. J. (2000). A simple dimensionality reduction technique for fast similarity search in large time series database. *4th Pacific-Asia Conference.*

Kleindienst, R. (07. 11 2009). *Standardableitungen 12-Kanal-EKG*. Abgerufen am 10. 09 2013 von EKG Online: http://www.ekg-online.de/ableitung/ableitung1.htm

Kreiß, J.-P., & Neuhaus, G. (2006). *Einführung in die Zeitreihenanalyse.* Springer.

Kumar, N., Lolla, N., Keogh, E., Lonardi, S., & Ratanamahatana, C. A. (2005). Time-series bitmaps: a practical visualization tool for working with large time series databases. *SIAM Data Mining Conference,* (S. 531-535).

Kumar, N., Nishanth, V. L., Keogh, E., Lonardi, S., Ratanamahatana, C. A., & Wei, L. (2005). Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases. *SIAM 2005 Data Mining Conference* .

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2006). *Experiencing SAX: a novel symbolic representation of time series.* Springer Science + Business Media, LLC 2007.

McGovern, A., Kruger, A., Rosendahl, D., & Droegemeier, K. (2006). *Dynamic relational models for improved hazardous weather prediction.* Pittsburgh: Proceedings of ICML workshop on open problems in statistical relational learning.

Microsoft Press. (2005). *http://msdn.microsoft.com/de-de/library/ms252076%28v=vs.80%29.aspx*. Abgerufen am 24. 07 2013 von http://msdn.microsoft.com

Moody, G., & Mark, R. (05-06 2001). The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng in Med and Biol 20(3)* , S. 45-50.

NetDoktor.de GmbH. (2013). *Elektrokardiografie (EKG).* Abgerufen am 18. 09 2013 von www.NetDoktor.de: http://www.netdoktor.de/Diagnostik+Behandlungen/Untersuchungen/Elektrokardiografie-EKG-249.html

Schäfer, P. (1. 07 2008). Zeitreihen in Datenbanksystemen. Berlin: Freie Universität Berlin.

Silvent, A. C., Carry, P. Y., & Dojat, M. (2003). *Data information and knowledge for medical scenario construction.* Protaras: Proceedings of the intelligent data analysis in medicine and pharmacology workshop.

Silvent, A., Dojat, M., & Garbay, C. (2004). *Multi-level temporal abstraction for medical scenario.* Int J Adapt Control Signal Process.

Springer Gabler Verlag. (2013). *http://wirtschaftslexikon.gabler.de/Archiv/435/bestandsgesamtheit-v9.html*. Abgerufen am 12. 07 2013 von http://wirtschaftslexikon.gabler.de

Springer Gabler Verlag. (2013). *http://wirtschaftslexikon.gabler.de/Archiv/437/bewegungsgesamtheit-v9.html*. Von http://wirtschaftslexikon.gabler.de abgerufen

Springer Gabler Verlag. (2013). *http://wirtschaftslexikon.gabler.de/Archiv/57589/zeitreihe-v10.html*. Abgerufen am 17. 05 2013 von http://wirtschaftslexikon.gabler.de

Springer-Verlag GmbH. (kein Datum). *Gabler Wirtschaftslexikon*. Abgerufen am 18. 05 2013 von http://wirtschaftslexikon.gabler.de/: http://wirtschaftslexikon.gabler.de/Archiv/2071/normalverteilung-v10.html

Universität Greifswald. (kein Datum). *Institut für Geographie und Geologie*. Abgerufen am 25. 06 2013 von http://www.yepat.uni-greifswald.de:8080/geo/: http://www.yepat.uni-greifswald.de/geo/fileadmin/dateien/Publikationen/GGA/GGA33/Kap5_Faktoranalyse.pdf

Wehner, J. (2011). *Die Erregungsleitung des Herzens*. Abgerufen am 18. 09 2013 von http://www.medizinfo.de/: http://www.medizinfo.de/kardio/herzrhythmus/erregungsleitung.shtml

Wikimedia Foundation Inc. (12. 04 2013). *Logarithmische Darstellung*. Abgerufen am 24. 06 2013 von www.wikipedia.de: http://de.wikipedia.org/wiki/Logarithmische_Darstellung

ZUM Internet e.V. (1995). *http://satgeo.zum.de/reisebuero/methoden/excel_netz_celsius.gif*. Abgerufen am 17. 05 2013 von www.zum.de: www.zum.de

Tabellenverzeichnis

Abbildungsverzeichnis

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben.

Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet.

Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens.

Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht.

Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Datum: Unterschrift: