FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

> APPLIED LIFE SCIENCES

MASTERSTUDIUM:
> BIOINFORMATIK

Comparison of categories in industrial process data

MASTERARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES

MASTER OF SCIENCE IN ENGINEERING

DER

FACHHOCHSCHULE FH CAMPUS WIEN

MASTER-STUDIENGANG BIOINFORMATIK

Vorgelegt von:
Johanna Hobiger
Personenkennzeichen: 1810542001

FH-Hauptbetreuer*in:
Ing. Christian J. Binder, BSc

Abstract

Big Data in industriellen Prozessen bringt die Notwendigkeit mit sich, Methoden für effiziente Analysen zu entwickeln. Die Visual Analytics Gruppe des VRVis widmet sich mit ihrer Software Visplore dieser Aufgabe.

Eine noch fehlende Möglichkeit zum qualitativen und quantitativen Vergleich zweier Zeitreihenabschnitte, genannt "Kategorien", wird implementiert. Dazu werden die Programmiersprachen C++ und Python im Rahmen der C-API verwendet, mit dem Ziel, die bereits bestehende Schnittstelle zwischen Visplore und Python ("VisplorePy") zu erweitern.

Die größte Herausforderung besteht darin, zahlreiche bisher entwickelte Zeitreihenanalysemethoden für die bestehende Aufgabenstellung zu begutachten und letztendlich eine eigene Methode zu entwickeln.

Um zwei Kategorien zu vergleichen, werden mit Hilfe von Splines Kurven durch die Datenpunkte gelegt, diese werden integriert und die Größen der Flächen verglichen. Das Größenverhältnis wird den Usern im entwickelten Prototypen in der Kommandozeile des Terminals ausgegeben. Ein mögliches Design des zukünfigen Dashboards in Visplore mit Liniendiagramm und Kurvenvergleich wurde in Jupyter erstellt.

Die Methode wird an unterschiedlichen Datensets getestet (Industrieprozessdaten, Herzfrequenzdaten und Aktienmarktdaten), um Stärken und Schwächen offenzulegen. Diese werden anschließend für weiterführende Implementierungen diskutiert.

Use of *Big Data* in industrial processes implies the necessity to develop methods for efficient analyses. The Visual Analytics Group of VRVis dedicates itself to this task with its software Visplore.

A still missing functionality for the qualitative and quantitative comparison of two time series sections, called "categories", is implemented. For this purpose, the programming languages C++ and Python are used within the C-API with the goal of extending the already implemented API between Visplore and Python ("VisplorePy").

The biggest challenge is to evaluate numerous time series analysis methods developed so far for the given task and finally to design an own method.

To compare two categories, curves are laid through the data points with the help of splines, are integrated and the sizes of the areas are compared. In the developed prototype the size ratio is displayed to the users in the command line of the terminal. A possible design of the future dashboard in Visplore with line chart and curve comparison was created in Jupyter.

The method is tested on different data sets (industrial process data, heart rate data and stock market data) to reveal strengths and weaknesses which are then discussed for further implementations.

Contents

List of Figures

List of Tables

Listings

1 Introduction

1.1 VRVis and process data

The VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH (VRVis) is a research center for visual computing which conducts projects in cooperation with industry and universities. The Visual Analytics group of VRVis focuses on multi- and high-dimensional, time-dependent data. They developed a software - Visplore - which can be used for data exploration. Possible use cases are gaining deeper understanding of data (data quality, outliers, missing values etc.), detecting trends, comparing time periods, pattern search or exporting images. The goal is to make large amount of data accessible and interpretable for users working in various sectors like the energy sector or industrial manufacturing (VRVis 2020).
Process data, which can be analyzed with the help of Visplore, consists of values that are recorded during technical processes. These include, for example, phase voltage (PV), humidity or wind speed. They are typically recorded at regular intervals. Structures such as seasonality, i.e. daily, quarterly or annually recurring patterns, are not uncommon with process data, especially when they are dependent on external influences such as solar radiation or temperature. This kind of data is visualized using time series whereas the time component is plotted on the x-axis, the measured values on the y-axis.

Nowadays, a great deal of data is collected through data mining in production for the purpose of quality assurance or process optimization. A simple example would be the comparison of power consumption of two processes. If there is a difference between two locations for the exact same process, the recorded data can be used to investigate which factors have contributed to varying consumption. It is of utmost interest to assess changes that have occurred on the basis of the time aspect.

1.2 Background to time series comparison methods

Since time series analysis is a wide research field, many methods are already in use, which in many cases depend on a particular use case. To exclude the possibility that a comparison of two time periods, in this context called "categories", is already implemented in some way, it is necessary to gain an insight into different approaches.
The existing methods can roughly be categorized as visualization, search for specific values or trends, finding periodicity or patterns, examination of existing data and doing forecasts.

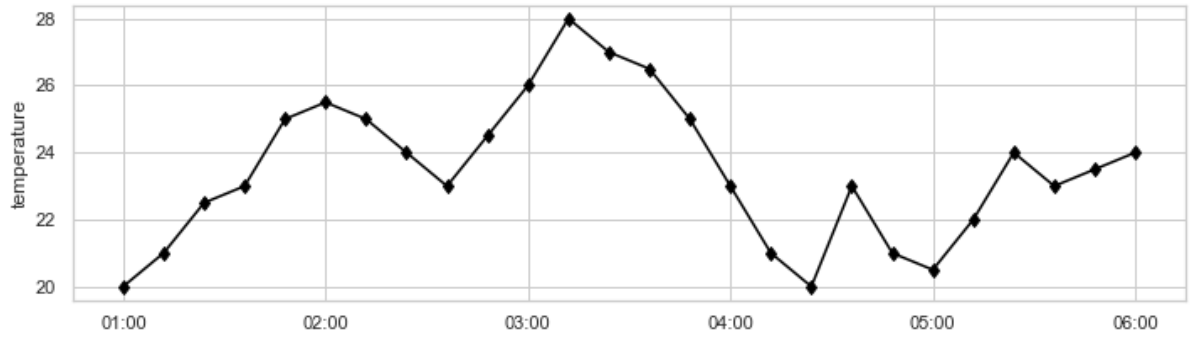
Of special interest are the following points:

- Representation of data or computed differences of data sets as line charts with original values and/or smoothed.
- Testing for differences with various statistical tests.
- Modeling attributes like structure and shape and comparing these models.

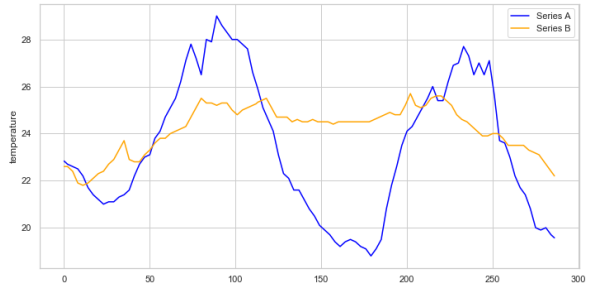
1.2.1 Visualization

Another challenge of large data sets, besides the data analysis, is the visualization of the results. There are countless options to choose from, but one has to consider the properties of the data. First and foremost, the chosen visualization method must be suitable for effectively displaying patterns and creating a mental picture of the data for users (Behrisch et al. 2018).

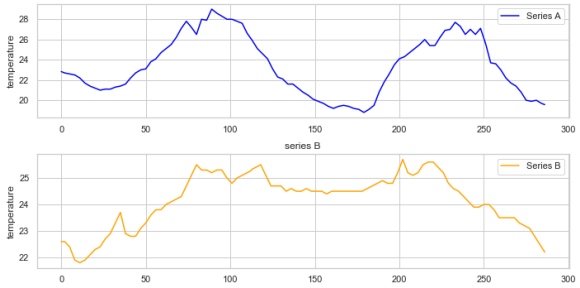
Line charts The best way to display time series are line charts (Behrisch et al. 2018). The values that are plotted along the time component on the x-axis are connected with straight lines, see Figure 1a. With line charts, trends or correlations between two time series can be identified at a glance (Behrisch et al. 2018; Saket, Endert, and Demiralp 2019), but identifying a value of a specific data point is difficult (Saket, Endert, and Demiralp 2019). For this goal, other kinds of visualizations should be chosen.
Another important point is how to visualize a comparison of two or more time series in a meaningful way. Two important terms in this context are "shared space" and "split space", where the latter means that each time series gets its own graph, while "shared space" means that all time series are plotted in one graph, see Figures 1b, 1c.
The advantage of shared space is that a comparison is easier because the eye does not have to travel as far as in split space. The individual series must be clearly distinguishable from one another, the classic approach is to use different colors or line styles. Furthermore, a common baseline is recommended to simplify comparisons (Javed, McDonnel, and Elmqvist 2010). A possible disadvantage resulting from this is cluttering or overlap. If the series are very similar, one should therefore decide for split space.
A problematic aspect of the representation is scaling, i.e. the ratio of height to width of the graph, since it influences the representation of trends (Behrisch et al. 2018). A distorted graph can make trends disappear or give the appearance of existing ones although there are none, see Figure 1d and 1e.



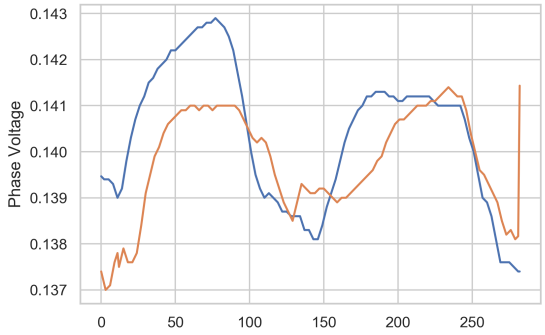
(a) Classic time series line chart



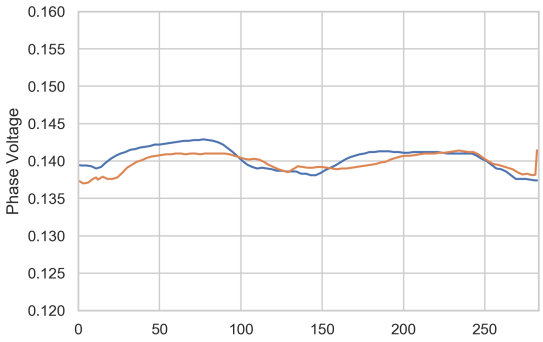
(b) Shared space



(c) Split space



(d) Data with shorter y-scale



(e) Data with larger y-scale

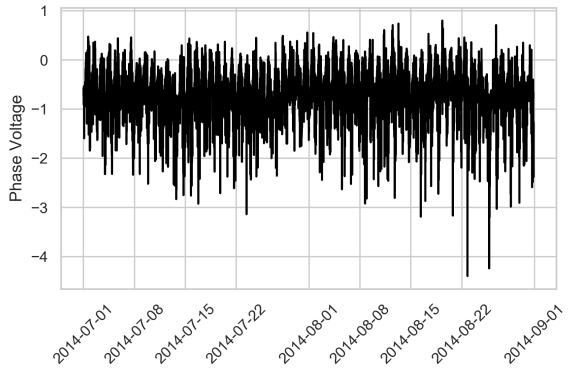
Figure 1: Time series as line charts. 1a shows a line chart example with connected values plotted along the x-axis. 1b and 1c represent two possibilities of plotting two line charts for comparison. 1d and 1e show how scaling of the y-axis influences perception of trends.

Moving averages moving average (MA) is a technique for data smoothing. By calculating the mean of a sliding window of predetermined length over raw data, a new series is created and plotted to ease interpretation, see Figure 2c.

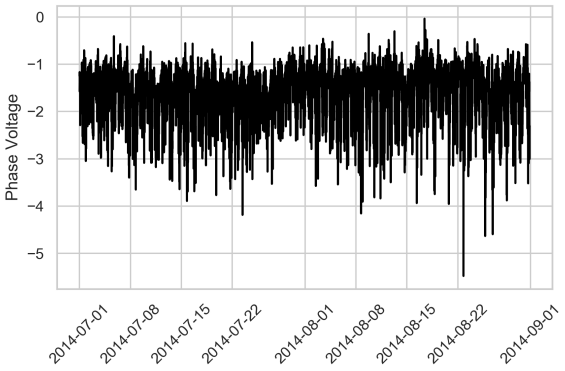
Calculated differences One can also decide for plotting relations of data instead of data itself. There are different ways of computing and presenting those.

- Figure 2d and 2e show the differences and ratios of the values of A and B.

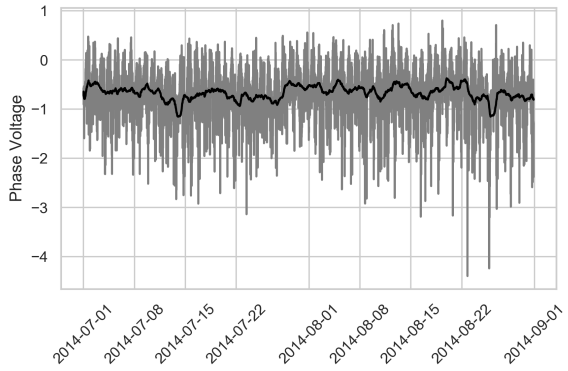
- In Figure 2f the cumulated sum is computed and plotted.



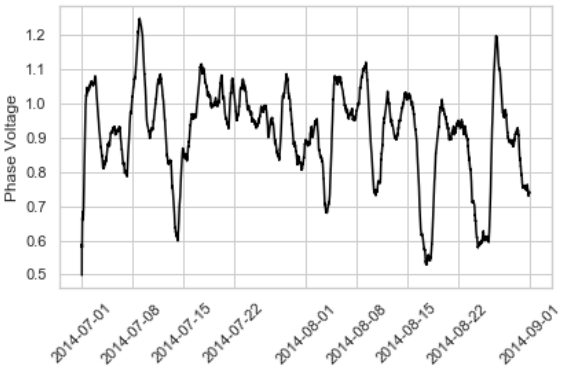
(a) Raw data A



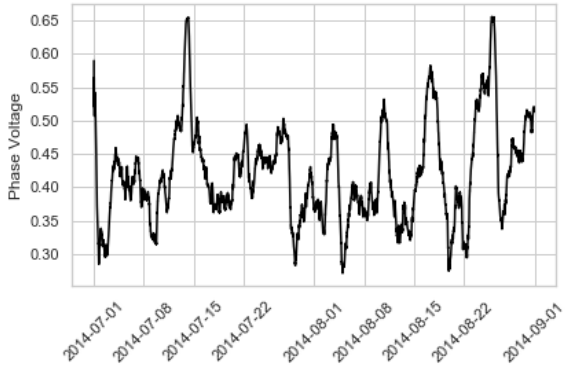
(b) Raw data B



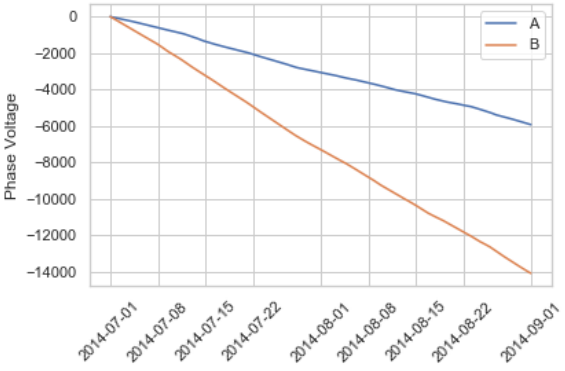
(c) Raw data (gray) and moving average (black)



(d) Difference (A-B)



(e) Ratio (A/B)



(f) Cumulative sum

Figure 2: Visualization of various comparisons. 2a and 2b show the raw data of two series A and B. In Figure 2c a moving average is plotted for comparison to raw data of A. 2d shows the calculated difference of A and B, 2e the ratio of A/B. In 2f the cumulated sum of both series is plotted.