Correlation analysis of multivariate time series

Submission date: 06.07.2020

Aymen Tiss, Hamza Mnari and Nikolaos Chrysafidis

Project report Database project
Faculty IV
Electrical Engineering and Computer Science
of the Technical University Berlin

Field of expertise:
Database systems and information management

Summer Semester 2020
Berlin

Contents

List of Figures

Abstract

In the last months humanity faced an unexpected pandemic that changed the daily lives of people all around the world. Lockdown measures were issued, travel restrictions were applied to every country and economy faced a new threat, having recorded a new market crash, thus raising concerns to many experts who expected a new economic recession. It is clear that the pandemic paved the way for a new era because it revolutionized the way home-office work is being done and taught some important lessons to the masses, which refer to public hygiene and disease countermeasures. Unfortunately, the crisis lead to many fears and worries about the future, which are being projected on the topics of unemployment and distress about the unknown coming times, since it is unclear when the whole situation will be diffused. By taking these important matters into account, we tried to analyse different fields that were influenced by the crisis by applying correlation analysis techniques. The goal was to find variables correlated with the covid-19 indicators.

Chapter 1

Introduction

The Correlation analysis is a statistical method used to evaluate the strength of relationship between two quantitative variables. "A high correlation means that two or more variables have a strong relationship with each other, while a weak correlation means that the variables are hardly related."[Franzese and Iuliano, 2018] In particular, our variables here are Multivariate time series. Time series data consists of sequences of values or events changing with time recorded at regular time intervals, e.g., hourly, daily, weekly. In case of multivariate time series , this data has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables.

Under this critical circumstances, we tried to determine and understand the relevant related factors so that we can finally reduce their risk and control the pandemic. In this project, we have studied the correlation of several variables of time series using actual data from different Fields (industrial, social, geographical, political, etc).

In section 2.1.1 we described the applied methodology and the tasks step by step based on an overview scheme. In section 2.1.2 we described our program requirements and explained why we decided to improve our implementation process with specific libraries and frameworks. Additionally, in the section 2.2, we explained our implementation process based on some examples. We took samples that relate to society, economy, online gaming community and temperature. In the last section we summarized the results that we obtained and based on the statistical results, we extracted realistic conclusions, which are the influencing and influenced factors by this pandemic. As an outlook, we proposed some methods to improve the system's performance in the future and how could we go further with this approach to discover more complex factors.

Chapter 2

Body

2.1 Methodology

2.1.1 Tasks Definition

In order to achieve our goal we had to follow specific predefined tasks that were self explanatory and could be applied to every use case. The first important step into doing our analysis is to discuss and choose the fields we want to further investigate and to make sure that our analysis is being done on a non-redundant scale. Non-redundant fields mean that our objective was to find use cases of many factors (Economy, Technology, Society) rather than laying our focus on only on one of them. After our fields definition our next step was to research the internet in order to find representative datasets that are accurate and in a desirable format. Our main priority was to find datasets in form of Time series. "A Time series is a set of observations, each one being recorded at a specific time. (Annual GDP of a country, Sales figure, etc)." [Das, 1994]. After the desired databases were imported everything was ready for the implementation. In (2.4) an overview of our group's task are presented as a guideline to follow in each field we discover. A short definition of each task is described as follows:

- **Filter**: Datasets can be vast and big, containing millions of rows and numerous columns. Our goal for each Dataset was to find ways to reduce its dimensions without removing crucial data. Missing values and other dataset structure problems should be solved in this stage.
- **Define Dimension**: In order to proceed with our analysis, defining the dimensions of the dataset plays an important role. We have to make sure that the dimensions are in the format we want (Example: Daily number of new Covid-19 Cases).
- **Convert To Time Series**: Index of each dataset should be in date format. In this specific project, converting to Time series is a major requirement to bring two tables together.

- **Merge Tables**: After reassuring that the dimensions of two Time series are equally defined, tables are brought together. One table contains information about the pandemic and the other contains data about the field being investigated.
- **Analysis, Visualize**: After verifying that the merged table is error-free (missing values created after merging, values dislocated, etc.) the statistical analysis can be initiated. In this stage the project's objective is being measured (Correlation Analysis).
- **Predict/Forecasting**: Final stage of the procedure. After having analysed the merged table, forecasting methods are applied in the direction of predicting into the unknown future.

2.1.2 State of the art

Importing and manipulating data may be a time consuming process. Algorithms should be implemented in order to parse the lines of a .csv file and properly transform them into a dataset. Additionally, filtering and other data manipulation methods should be manually implemented. In the past years Data Science, defined by Vasant Dhar as "the study of the generalizable extraction of knowledge from data"[Dhar, 2013], became a hot topic in a global scale and many libraries came into the surface, simplifying these prolonged procedures needed to be done in order to analyze massive amounts of Data. Two big examples are R and Pandas libraries, both of them having a big community and being our candidates for our project implementation. By implementing with these two libraries we were able to have many tools at our disposal, which enabled us to progress our development without having to face the implications of the aforementioned processes.

2.1.3 Evaluation Plan

Unfortunately, in the context of statistical analysis there is no direct method that evaluates if a procedure was better than the other. The goal of the project was to find factors that were influenced or influenced the pandemic. By having a statistical output such as a numerical value or a plot there is no guarantee that this specific output is a result of a correct defined process. As a result, we tried to do crossover examination on our outputs with other outputs after having reconfigured some parts of our models. One example is comparing the outputs of a model with daily values with the outputs of the same model in hourly values. In big datasets which require immense computation, performance evaluation methods can be used in order to decide which implementation option is the better one.

2.2 Implementation

2.2.1 Tasks implementation

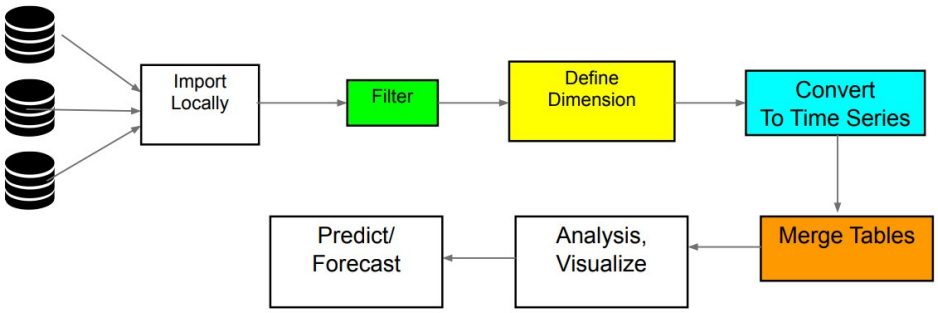

Figure 2.1: Tasks Overview

In 2.1.1 the tasks shown in 2.4 were defined. The goal of the group was to coordinate its processes based on these specific tasks and applying them to every area of our research. In the following examples topics from different areas will be covered and analysed, each one providing an insight to the current pandemic.

2.2.2 Gdelt

Media has proved itself as a very useful way to express the societies opinions, fears, expectations and interests. Therefore, analyzing the media data during a pandemic will reflect in a way the effect of the virus on the societies . Based on the data gathered by the GDELT Project, that retrieved a huge amount of themes "from realtime translation of the world's news in 65 languages, to measurement of more than 2,300 emotions and themes from every article, to a massive inventory of the media of the non-Western world" [1] and the COVID data given by the World Bank, we tried to understand how societies interacted with the situation. Despite the fact that the data gathered covers a huge number of countries, if not all of them, we have noticed that the reliability of this data may vary from one country to another. In other words, for some countries, the amount of data was not consistent enough to work with on the project. Examples are lack of data for several consecutive days or low number of appearances of themes.

That is why we choose to select a set of countries to work on, which were Germany, Italy and the United States of America. First of all we had to extract the data concerning each of these countries and write it in separate csv files to improve the manipulation of the data with the pandas library, which we found less effective when it comes to large amount of data.

[1] https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

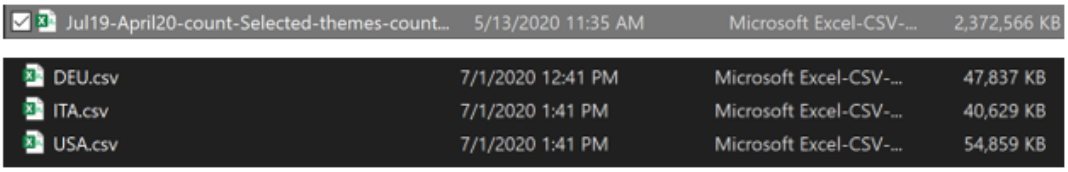

Figure 2.2: Datasets for each country

As we can see in the Figure above, it has reduced the amount of data from a scale of GBs to a scale of MBs. The results of the data extraction became now the input of the next step that consists mainly of filtering these themes and keep only those who have satisfy a couple of thresholds: their correlation with the daily new cases of the Covid-19 virus is greater than 0.69 or less than -0.69 and the theme appeared in the medias more than 1000 times between 1.1.2020 and 30.4.2020 and have complete data for that period of time. For that we kept track of relevant columns (date, themes, daily new cases, daily number of appearances in the media) and for each theme we checked if the thresholds are met and wrote csv files containing the ones who do for each country. This is an example of results found which represents the highly correlated data from Italy :

Column1	theme	correlation	total appearences
0	CRISISLEX_T04_INFRASTRUCTURE	0.72517794	21526
1	CRISISLEX_T11_UPDATESSYMPATHY	0.80728663	140709
2	WB_2165_HEALTH_EMERGENCIES	0.79050486	317578
3	WB_1072_FISCAL_POLICY_AND_GROWTH	0.712177713	2099
4	WB_1070_ECONOMIC_GROWTH_POLICY	0.710960763	2272
5	WB_621_HEALTH_NUTRITION_AND_POPULATION	0.701468748	497017
6	ARMEDCONFLICT	0.742593893	105677
7	WB_697_SOCIAL_PROTECTION_AND_LABOR	0.706585207	101544
8	EPU_CATS_NATIONAL_SECURITY	0.75739943	50144
9	WB_635_PUBLIC_HEALTH	0.769252812	345420
10	CRISISLEX_CRISISLEXREC	0.720597426	690733
11	CRISISLEX_T02_INJURED	0.88260378	140425
12	CRISISLEX_T01_CAUTION_ADVICE	0.705268607	71323
13	EPU_POLICY_SPENDING	0.850270559	20827
14	WB_1350_PHARMACEUTICALS	0.726453601	79081
15	KILL	0.841639451	329716
16	WB_1331_HEALTH_TECHNOLOGIES	0.746011407	82094
17	CRISISLEX_T08_MISSINGFOUNDTRAPPEDPEOPLE	0.848345808	69943
18	UNGP_HEALTHCARE	0.796466511	267904
19	WB_1618_FOOD_DISTRIBUTION	0.785354565	59383
20	CRISISLEX_T03_DEAD	0.869744715	195337
21	WB_1609_FOOD_AND_IN_KIND_TRANSFERS	0.770248484	66488
22	WB_1620_ELDERLY	0.781224615	58752
23	WB_1466_SOCIAL_ASSISTANCE	0.754251261	69148
24	WB_1448_DEMOGRAPHIC_CHANGE	0.692602971	12846
25	WB_1305_HEALTH_SERVICES_DELIVERY	0.737751774	44286
26	ECON_BUDGET_DEFICIT	0.701523422	11282
27	UNEMPLOYMENT	0.738350387	20346
28	SOC_TRAFFICACCIDENT	-0.719199187	3266
29	MANMADE_DISASTER_TRAFFIC_ACCIDENT	-0.733104918	1863

Figure 2.3: Themes and their correlations in Italy

At this point we have noticed that several words appears multiple times in

those csv and for example Public Health and Health Emergencies should be aggregated with other themes containing the word Health and then analyzed together. In the health case it is easily remarkable but for other cases it is not and for example Economy and Economical should be aggregated together but they need to be steemed before . Therefore we have applied a frequent items analysis algorithms on the frequent themes found previously which resulted into some wordclouds and barplots (the ones below refers to Italy).

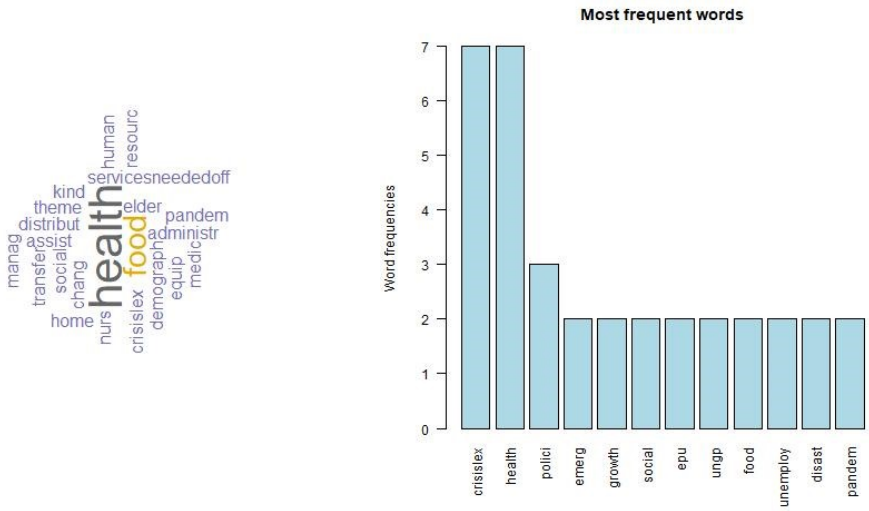

Figure 2.4: Frequent items analysis algorithm outputs

Afterwards we applied the same analysis as before but on the aggregation of themes containing each of the words that appeared more than once in the frequent words set for a country. Of course the number of appearances and the daily new cases are on different scales that is why we normalized the data before plotting it and an example of the results also for Italy is displayed below:

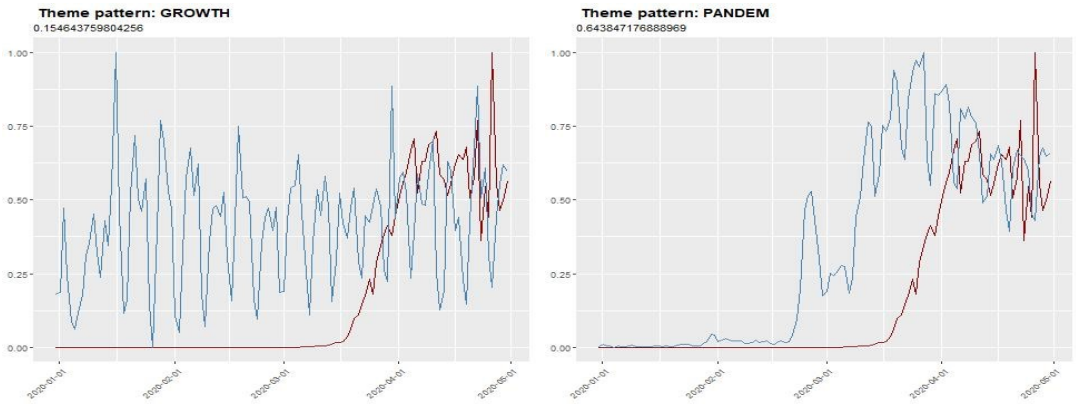

Figure 2.5: Normalized plots post theme aggregation : stemmed words "Pandem"/"Growth" and daily new cases between 01.01.2020 and 30.04.2020 in Italy

We noticed that for example the correlation with the themes containing the word Growth is low despite the fact that for example the correlation with fiscal policy and growth was high. On the other hand , themes related to the word "Pandem" stayed highly correlated and this gives our results more credibility and efficiency.

2.2.3 Economy

The pandemic had a huge impact on the economy section and it is still unsure when the situation will be stabilized. "On average, each additional month of crisis costs 2.5-3% of global GDP."[Fernandes, 2020]. We tried to do a market analysis of many companies based on their daily close market value. The datasets were downloaded from `https://eoddata.com/`. One problem that had to be solved was that the .csv files provided by the site were separated for each company and as a result, when downloading a new dataset, code had to be updated in order to include and import this new dataset. The problem was solved by writing an automation, which checks the folder path (One folder with many .csv files) and doing the analysis based on this folder and not based on the individual files. The only requirement was to insert the files in the folder and the code would adapt to each update.

Since stock markets are affected on a global scale, we decided that the regional dimension of each companies dataset is set on a global scope. The column of interest was the 'Closing Market Value' and after the necessary conversions each company's dataset had the dimension of (Daily 'Closing Market Value'/ Globaly). Covid-19 dataset had to be in a global scale too, so downloading a global dataset with daily recorded cases would suffice. Having a global dimension reduced the need of regional filtering and by having (Daily new cases/ Global) and (Daily closing stock market value/Global) the tables were ready to be merged. After the merged tables were generated we were able to calculate the correlation between cases and stock market value by using the Pearson's method. The results were stored in a new dataset containing each company's name and its corresponding correlation. With this approach we were able to check which companies were heavily affected by the pandemic. In 2.6 and 2.7 we can see to extreme examples of the pandemic influence in the economy sector.

2.2.4 Gaming

Lockdown measures had an enormous impact on the daily routine of people. By working from home and meeting their friends in online platforms it is clear that internet usage has been dramatically increased. Reports from OpenVault Broadband Insights Report [2] state that during the coronavirus outbreak data usage has been increased by 47%. We tried to measure the way the gaming

[2] `https://openvault.com/complimentary-report-Q120/`

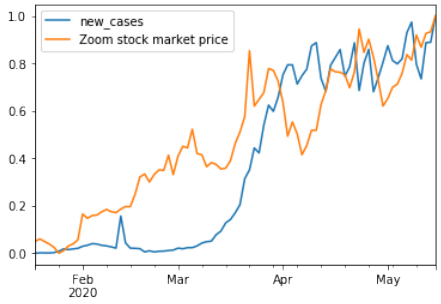

Figure 2.6: Normalized Time-series of Zoom company.

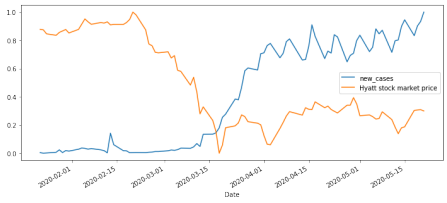

Figure 2.7: Normalized Time-series of Hyatt company.

community reacted to the corona outbreak and decided to analyse twitch.tv, the biggest gaming streaming platform in the world. As an example we took a sample from Counter Strike: Global Offensive, which is in the top 5 streamed games. The dataset was downloaded from twitchtracker.com and it contained daily values of average viewers.

The first problem that had to be solved was that the average daily viewer values were recorded every second day and we had to find a way to approximate the missing values. Two ways to solve this problem were either with Interpolation or with Imputation."The relatively easiest and in many applications often most desired approach to solve the problem is interpolation, where an approximating function is constructed in such a way as to agree perfectly with the usually unknown original function at the given measurement points""[Meijering, 2002]. On the other hand imputation techniques "rely on inter-attribute correlations to estimate values for the missing data"[Moritz and Bartz-Beielstein, 2017]. We assumed that a missing value depends on its closest values, since the instances are daily viewers and there are no peaks to be expected between two known points. Based on this assumption, We applied linear interpolation and fixed the missing values problem. In 2.8 a post interpolation example is been shown. After the missing values were the dataset had the following dimensions: (Daily average viewers/Global).

The coronavirus dataset should have rows of daily new cases in a global scale. The original dataset had the shape of (950670 rows x 18 columns) and some filter methods had to be applied in order to reduce its size and keep the important information, such as removing the unnecessary columns. After applying a groupBy method based on the dates we had the sum of cases in each day and the global dimension requirement was fulfilled. The problem with the dataset was that it stored the number of cumulative cases and that led to the creation of a new column ,storing the difference of the next cases value and the current value of a row. The new column stored the global new cases instead of total new cases. The dimensions were compatible and the tables were ready to be joined. A correlation Pearson value of 0.42 was calculated, which is a moderate value for this use case. In 2.9 the scatterplot output of this

	viewers
2020-03-02	47453
2020-03-03	NaN
2020-03-04	66472
2020-03-05	NaN
2020-03-06	56286

	viewers
2020-03-02	47453.000000
2020-03-03	56962.500000
2020-03-04	66472.000000
2020-03-05	61379.000000
2020-03-06	56286.000000

Figure 2.8: Interpolation Results.

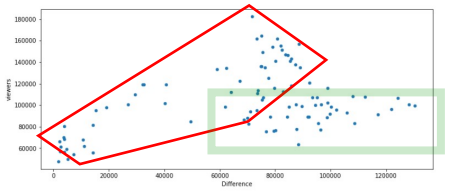

Figure 2.9: Scatter Plot: Daily average viewers/ New cases.

procedure is plotted.

2.2.5 Temperature Analysis

One of hottest debates of the current pandemic is in what way a country's climate and temperature may affect the outbreak. It was our group's goal to dive into this matter and produce results that have statistical significance. The complexity of the task was increased, because the goal was to generate a mini report for each country and that required to join big datasets of each country's temperatures and datasets of their pandemic data. In parallel, we should make sure that there are no bugs and missing values, since everything will be automated. Two big datasets were at our disposal, one recording daily temperatures from each major city and one recording the data about the pandemic. The first task was to filter the temperature dataset and bring it in the wanted dimensions. Based on the temperature dataset we created new datasets based on each country's names. As previously mentioned, the daily values stored temperatures of different cities for a specific country and for our purpose we decided to create a Time series based on the average daily values of each city. For example if on 01.01.2018 Berlin has an average temperature of 3°C, Munich 4°C and Hamburg 3°C the temperature stored in this specific day's row for Germany will be the average of the above (3,33 °C)

After the aforementioned configurations were made, datasets (for each country) that stored daily Celsius temperatures were created. A loop through the dataset was done and based on the each country's name, filtering methods were applied on the covid-19 dataset in order to find values that are consonant with the temperature dataset. With this method, we tried to create a "bridge" between two completely different datasets. For each country's temperature table a corresponding covid-19 table was merged and we were able to calculate the Pearson correlation value between the daily new cases and the daily average temperatures. A mini-report was generated, an example of which is shown in 2.10. Everything was stored in a new dataset, as seen in 2.11, which gave as an overview of each country's climate and pandemic correlation.

```
Croatia
Average Temperature: 11.547945205479452
Correlation Between Cases and Temperature
-0.29109506630248083
```

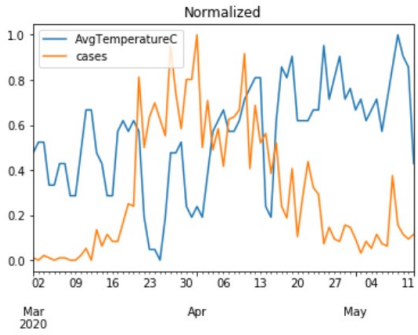

Figure 2.10: Temperature mini report example.

	Country	avg Temperature(C)	temp/cases corr
15	Brazil	21.837838	-0.621895
2	Argentina	18.486486	-0.605970
4	Austria	9.635135	-0.530626
48	Iceland	2.554054	-0.454179
97	South Africa	17.581081	-0.454061
53	Italy	12.067568	-0.385088
99	Spain	13.486486	-0.380621
103	Switzerland	9.337838	-0.374476
107	Tanzania	26.405405	-0.345184
111	Tunisia	16.378378	-0.321802
28	Denmark	6.013514	-0.311632

Figure 2.11: Temperature analysis overview

Chapter 3

Conclusions

From the beginning of the project it was clear that Coronavirus left a huge impact in every part of the world. Our group managed to answer the question of "how big" the impact in different sectors of the society was. By measuring success stories and disaster stories from the marketplace, analysing the way the gaming community reacted to the pandemic, inspecting the most mentioned topics of each country and generating temperature mini reports for each country we were able to enhance our understanding of the crisis by formulating an assumption into statistical outputs. It is still unclear when the situation will be over and there are many other fields ready to be discovered.

At the end of the project's cycle our team was in a position to compare different fields and determine which ones were more influenced by the crisis. Unfortunately we were not able to produce significant results based on the forecasting task. One problem lied within the short time span, which we had at our disposal. Fitting models that are good at predicting requires bigger datasets with many recorded values and unluckily, because the the time passed since the outbreak is still short, we weren't able to generate representative outputs. It is clear that if we had more months we would have bigger datasets and better models and that leaves an open parenthesis for future work.

Bibliography

[Das, 1994] Das, S. (1994). *Time series analysis*, volume 10. Princeton university press, Princeton, NJ.

[Dhar, 2013] Dhar, V. (2013). Data science and prediction. *Commun. ACM*, 56(12):64–73.

[Fernandes, 2020] Fernandes, N. (2020). Economic effects of coronavirus outbreak (covid-19) on the world economy.

[Franzese and Iuliano, 2018] Franzese, M. and Iuliano, A. (2018). *Correlation Analysis*.

[Meijering, 2002] Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3):319–342.

[Moritz and Bartz-Beielstein, 2017] Moritz, S. and Bartz-Beielstein, T. (2017). imputets: time series missing value imputation in r. *R J.*, 9(1):207.