Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

Group Project – phase 1

business goal, vision of the system, functional and non-functional description, risk analysis, schedule, division of work

Agata Makarewicz, Jacek Wiśniewski

version 1.2

25.10.2021

Table of Contents

1 Abstract

For over 40 years, scientific works have presented various divisions of European countries into economic and cultural groups, based on different criteria such as GDP per capita, the level of industrialization or HDI. Depending on the considered indicators and the date of the analysis, usually, 2 to 5 groups are defined. For instance, in the article written by C. Gräbner et al. (2019), the central, peripheral and Eastern European countries as well as financial centres were distinguished. The main goal of the project will be to apply several standard clustering methods such as k-means, hierarchical clustering and the fuzzy c-means method to time series of economic growth to group countries and verify the previously proposed divisions. The algorithms will be evaluated using the existing cluster analysis assessment indexes, e.g. inertia, silhouette score, GAP statistic and PBM index. The thesis will be based on publicly available data, including the Penn World Table. The selection of variables itself is one of the tasks. The analysis will cover complete time series and selected segments (e.g. before and after 2008 - the year of the last financial crisis). Another issue to examine will be the aspect of similarity of time series in the context of the assessment of synchronization or non-synchronization of business cycles of selected groups of countries before and after the crisis. The implemented models will be part of the web application in which the user will be able to compare the results of the methods used, select variables and parameters for the models, as well as the development indicators presented in the charts. Visualizations of the clusters obtained with different clustering methods will also be available.

1.1 History of changes

Date	Author	Description	Version
21.10.2021	Jacek Wiśniewski	First version	1.0
24.10.2021	Agata Makarewicz, Jacek Wiśniewski	All chapters completed except for 3.2 (Functional requirements)	1.1
25.10.2021	Agata Makarewicz	Final version for the first checkpoint	1.2

2 Vocabulary

Homepage - a webpage presented after turning on the application. It will have all of the functionalities like filtering data and generating the report.

“Read about the project" page – a webpage that will present all of the information about the project, authors and contact email addresses.

Report – content from homepage consisting of charts and results of clustering algorithms with comments.

Clustering - is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters).

Business cycle - intervals of expansion followed by a recession in economic activity. Fluctuations are usually characterized by general upswings and downturns in a span of macroeconomic variables.

Segmentation – i.e. **time-series segmentation** is a method of time-series analysis in which an input time-series is divided into a sequence of discrete segments in order to reveal the underlying properties of its source.

Model – machine learning algorithm used for clustering.

Risk analysis - the science of risks and their probability and evaluation. Probabilistic risk assessment is one analysis strategy usually employed in science and engineering.

Schedule - a basic time-management tool, consists of a list of times at which possible tasks, events, or actions are intended to take place, or of a sequence of events in the chronological order in which such things are intended to take place.

Vocabulary - a set of familiar words within a person's language. A vocabulary, usually developed with age, serves as a useful and fundamental tool for communication and acquiring knowledge

3 Specification

3.1 Executive summary

The aim of the diploma thesis is to apply various clustering algorithms to the time series of economic growth of European countries. The analyzed data will be pre-processed using methods such as segmentation, normalization or anomaly removal. The results of the work will be presented in the form of an application with a graphical user interface written in Django, which will allow developers, economists and data scientists to compare the indicators for different countries.

3.2 Functional requirements

Use case

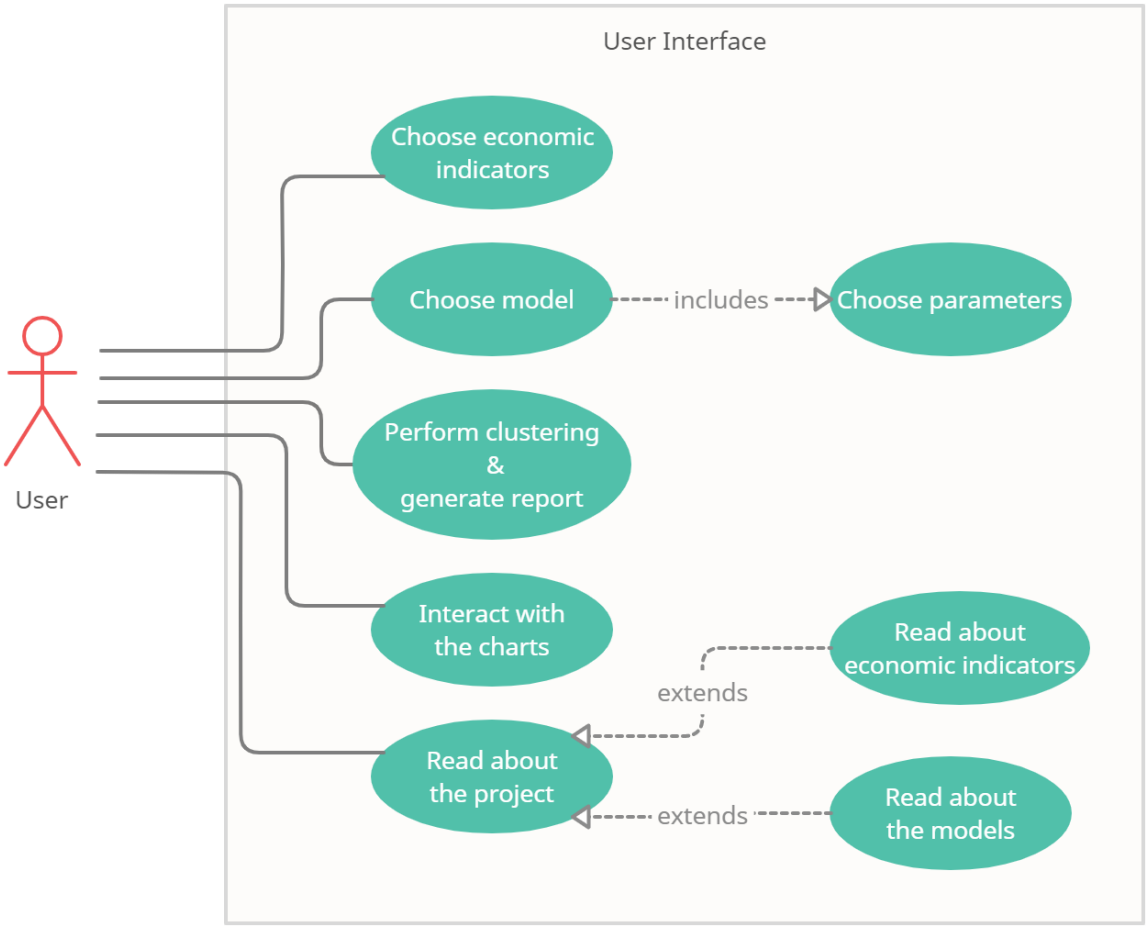

Figure 1. Use case showing actor's interaction with the User Interface

ID	Actor	Name	Description	System response
usr	User	Choose economic indicators	Use dropdown list / checkboxes to select variables for modelling	Homepage with filters (and optionally, previous charts if clustering was already performed once)
		Choose model	Use radio buttons to select clustering algorithm from 3 available ones	Homepage with filters (and optionally, previous charts if clustering was already performed once)
		Choose parameters	Use dropdown list / radio buttons / checkboxes to set parameters of chosen model	Homepage with filters (and optionally, previous charts if clustering was already performed once)
		Perform clustering & generate report	Click the button to perform clustering using chosen model and data	Homepage with a report containing the results of clustering and generated charts
		Interact with charts	Perform actions: zooming, hovering, choosing a variable and more to display preferred information	Homepage with previously generated charts modified accordingly to user's actions
		Read about the project	Click the button to see project information	Webpage with project information
		Show list of available economic indicators	Click the button / use search box to display information about variables	Webpage with a table containing variable name and description
		Show list of available clustering algorithms	Click the button / use search box to display information about algorithms	Webpage with a table containing algorithm name and parameters with description

Table 1. Description of a use case for actors and User Interface

User Stories

The functionality of the application is to present clustering results and visualize them. The user can be anyone interested in time series analysis, clustering (generally data science), economics, or European geography. The most probable users are data scientists and economists, therefore user stories below are presented in two groups because some of them assume the user already has some background knowledge about machine learning (data scientists) or business cycles theory (economists).

- Economist

  1. As an economist, I want to choose economic indicators, so that I can select the most important/accurate ones for the identification of the business cycle.

  On the homepage will be placed a list with checkboxes corresponding to economic indicators available to select for the model.

  2. As an economist, I want to interact with charts, so that I can get more detailed information about a particular group of countries/country.

  Generated charts (report) will be interactive, enabling the user to perform zooming, hovering, and more to display preferred information.

  3. As an economist, I want to read about models used in the project, so that I can gain intuition about how they work.

  On the 'read about the project' page will be placed a table with 3 implemented clustering algorithms descriptions.

- Data scientist

  1. As a data scientist, I want to choose a model, so that I can compare the results of different clustering algorithms.

  On the homepage will be placed a dropdown list with 3 implemented clustering algorithms to choose which one should be used for modelling.

  2. As a data scientist, I want to choose the model's parameters, so that I can verify different hypotheses (groupings).

  After choosing a model there will be an option to set its parameters or leave default ones.

  3. As a data scientist, I want to read about models used in the project, so that I can develop my data science skills.

  On the 'read about the project' will be placed a table with 3 implemented clustering algorithms descriptions.

  4. As a data scientist, I want to read about economic indicators, so that I can understand what they mean.

  On the 'read about the project' page will be placed a table with economic indicators description.

3.3 Non-functional requirements

Requirements area	Requirement No.	Description
Utility (*Usability*)	1	The application does not require creating an account.
	2	The application should have an intuitive user interface with all tools presented on the homepage.
	3	All functionalities of the application available to the user must fit on a single screen with a resolution of 1920x1080 and a font no smaller than 12pt.
Reliability (*Reliability*)	4	The application should have built-in redundancy into the system in order to eliminate single points of failure.
	5	The application is to be available 24x7 at least 95% of the time between 7:00 a.m. and 7:00 p.m.
Performance (*Performance*)	6	The application should generate clustering results and charts in no more than 10 seconds.
	7	User's interaction with generated charts will not cause any significant delay.
Maintenance (*Supportability*)	7	The contact email addresses are to be provided on the "read about the project" webpage of the application in order to report failure.
	8	Final documentation will have the user's guide attached.

Table 2. List of non-functional requirements

4 Project schedule

The project is planned to be implemented in accordance with the following schedule:

Nr	Task	Start Date	End Date	Duration
1	**Documentation**	**02.06.2021**	**07.11.2021**	**159**
1.1	Proposal for thesis topic	02.06.2021	19.07.2021	48
1.2	Literature overview	02.06.2021	13.10.2021	134
1.3	Data overview	20.08.2021	27.10.2021	69
1.4	First part of documentation	12.10.2021	**25.10.2021**	14
1.5	Second part of documentation	26.10.2021	**07.11.2021**	13
2	**Exploratory Data Analysis & data preprocessing**	**01.09.2021**	**10.11.2021**	**71**
2.1	Exploratory Data Analysis	01.09.2021	28.10.2021	58
2.2	Variables selection	06.10.2021	28.10.2021	23
2.3	Missing data imputation	29.10.2021	02.11.2021	5
2.4	Anomaly removal/Outlier detection	03.11.2021	05.11.2021	3
2.5	Preprocessing data of different granularity	06.11.2021	07.11.2021	2
2.6	Normalization/standarization	06.11.2021	07.11.2021	2
2.7	Segmentation	08.11.2021	10.11.2021	3
3	**Modeling**	**11.11.2021**	**22.11.2021**	**12**
3.1	K-Means clustering	11.11.2021	13.11.2021	3
3.2	Hierarchical clustering	11.11.2021	13.11.2021	3
3.3	Fuzzy C-Means clustering	14.11.2021	16.11.2021	3
3.4	Evaluation	17.11.2021	22.11.2021	6
3.5	Testing	17.11.2021	22.11.2021	6
4	**Application**	**20.11.2021**	**23.12.2021**	**34**
4.1	Framework	20.11.2021	22.11.2021	3
4.2	Backend	23.11.2021	**06.12.2021**	14
4.3	Visualisation module	07.12.2021	15.12.2021	9
4.4	Frontend	16.12.2021	23.12.2021	8
5	**Testing**	**23.11.2021**	**10.01.2022**	**49**
5.1	Unit tests	23.11.2021	**06.12.2021**	14
5.2	Acceptance tests	27.12.2021	**10.01.2022**	15
5.3	Guide	27.12.2021	**10.01.2022**	15
5.4	Final documentation	27.12.2021	**10.01.2022**	15

Figure 2. Project schedule.

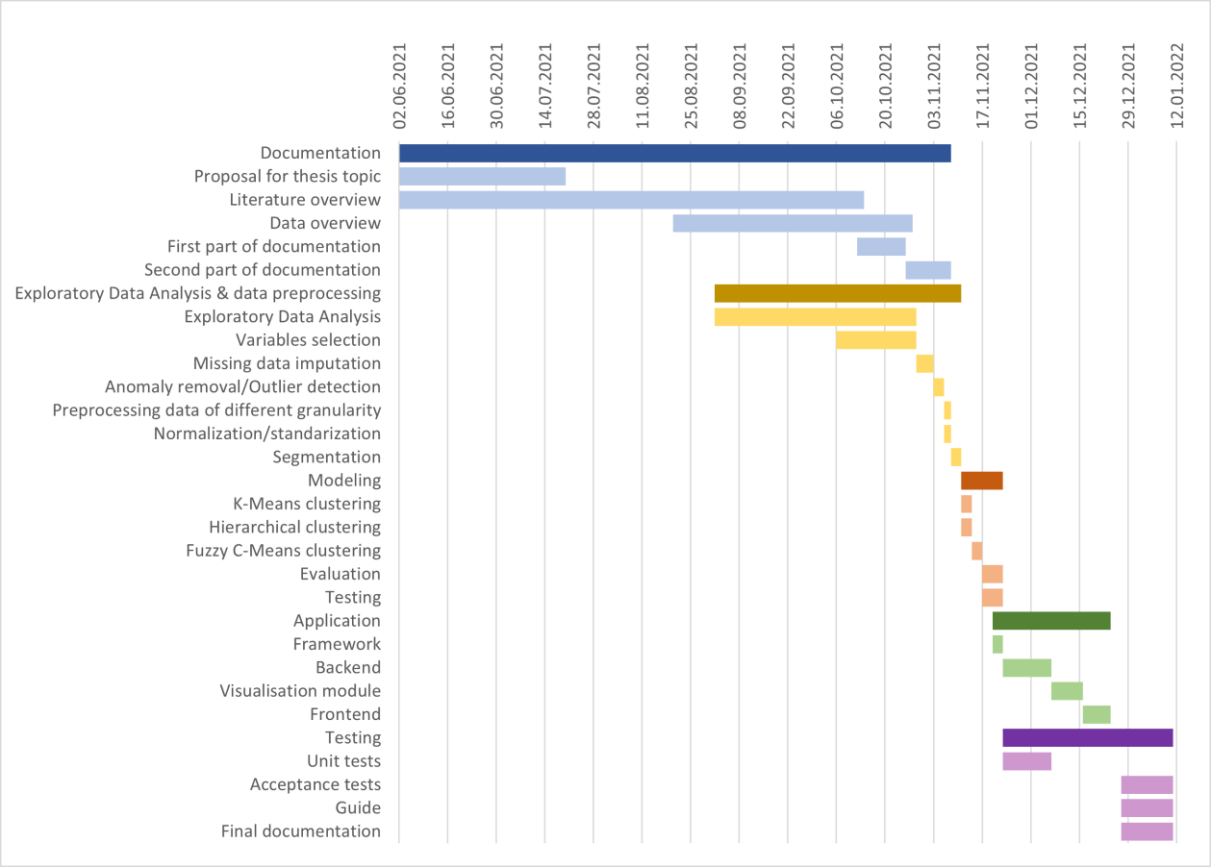

Figure 3. Project schedule - Gantt diagram.

5 Risk Analysis

SWOT	Helpful	Harmful
Internal	1. Schedule created at the beginning of the project 2. Experience in working in the team from previous projects 3. Strong Python skills	1. Not enough time due to other projects 2. Lack of economic background
External	1. Multiple papers covering the subject of this project 2. Multiple economic indexes available, published by different sources	1. Not enough data available to perform efficient clustering 2. Paid access to some of the potentially useful papers

Strengths:

1. Schedule created at the beginning of the project will be helpful in a balanced division of work and effective time management.
   Value: medium

2. Experience in working in a team from previous projects increases the chances of fruitful cooperation and lowers the risk of misunderstandings.
   Value: medium

3. Strong Python skills gained during the last 3 years of studies will fasten completing the technical part of the project.
   Value: high

Weaknesses:

1. Deadlines from other projects may overlap with this project's deadlines and cause a delay in the workflow. In order to prevent such a situation from happening, a well-thought-out plan and effective time management are needed.
   Likelihood: very high

2. Lack of economic background may be harmful during the selection of relevant variables for the model and explaining the clustering results in the report. In such a situation, background reading of economic papers as well as supervising professor’s experience would be useful.
   Likelihood: medium

Opportunities:

1. There have already been published multiple scientific works presenting various divisions of European countries, based on different criteria. Such papers will help gain sufficient economic knowledge for this thesis and provide conclusions which could be used as a benchmark for obtained results.
   Value: medium

2. Thanks to the fact that there are multiple economic indexes available, published by different sources, there is a wide range of variables to analyze and select for modelling phase.
   Value: high

Threats:

1. Some of the European countries have gained full independence (or has been formed) around 1990-1995, and therefore their data was collected for only approximately 20-30 years. This may lead to the problem of too short time series to perform efficient clustering. To overcome this issue, it will be important to gather multiple economic indexes to obtain as many features as possible, and also data with different granularity can be used (collected more often than once a year).
   Likelihood: high

2. Paid access to some of the potentially useful papers may slow down the progress of the first phase of the project. Searching for open-access resources may solve this problem, but it will take more time to find the necessary ones.
   Likelihood: low

6 Bibliography

[1] Aghabozorgi, Saeed, Shirkhorshidi, Ali S., and Wah, Teh Y. Time-series clustering – A decade review. Information Systems 53 16-38, 2015.

[2] Gräbner, C., Heimberger, P., Kapeller, J., and Schütz B. Structural change in times of increasing openness: assessing path dependency in European economic integration. Journal of Evolutionary Economics 30, 1467–1495, 2020.

[3] Bartlett, W. and Prica, I. Interdependence between Core and Peripheries of the European Economy: Secular Stagnation and Growth in the Western Balkans. LSE'Europe in Question' Discussion Paper Series, LEQS Paper No. 104/2016, 2016.

[4] Hamilton, James Douglas Time Series Analysis. Princeton University Press, 1994.

[5] Pal, Avishek, Prakash, PKS. Practical Time Series Analysis. Master Time Series Data Processing Visualization and Modelling Using Python. Packt, 2017