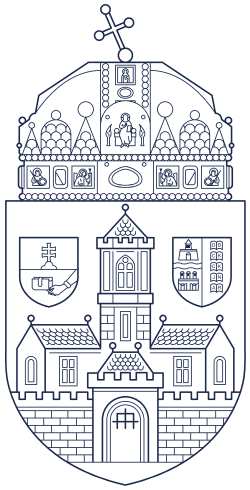
# Óbuda University

## PhD Thesis



Application Possibilities of Robust and Non-parametric Statistical Tools in Respect of Economic Data

by

Ferenc Tolner

Supervisors:

Dr. György Eigner

Dr. Balázs Barta

Applied Informatics and Applied Mathematics Doctoral School

Budapest, 2024

# Contents

# List of Figures

# List of abbreviations

| Abbreviation | Meaning |
|---|---|
| MFV | Most Frequent Value |
| NACE | Nomenclature statistique des activités économiques dans la Communauté européenne |
| NUTS | Nomenclature of territorial units for statistics |
| H2020 | Horizon 2020 |
| LDA | Latent Dirichlet Allocation |
| GDP | Gross Domestic Product |
| NDI | Net Domestic Income |
| PPS | Purchasing Power Standard |
| MAD | Median Absolute Deviation |
| GVA | Gross Value Added |
| NLP | Natural Language Processing |
| SME | Small and Medium-sized Enterprise |
| SMOTE | Synthetic Minority Oversampling Technique |
| MICE | Multiple Imputation by Chained Equations |
| MLE | Maximum Likelihood Estimation |
| IQR | Interquartile Range |

**Acknowledgements**

The outlined work is not yet the closing word, but an important milestone of the research I have been carrying out throughout my Ph.D. program. It could not have been reached without the support of many devoted people, to whom I would like to express my deepest gratitude. First and foremost to my supervisor, Dr. György Eigner from Óbuda University, who closely followed my pursuit and also witnessed my struggling. His prompt and encouraging guidances provided me the faith not to give up and search for new possibilities and options. I would also express my gratitude to Dr. Sándor Fegyverneki from the Department of Applied Mathematics, from University of Miskolc, for his invaluable help, for the time he devoted and for his professional guidance in the field of statistics. His small remarks and excellent mindset revealed how complicated it can be to interpret seemingly easy statistical problems and can conventional techniques lead researchers astray, furthermore drew my attention towards robust statistics. Both of their guidance has significantly contributed to increase the value of my work that might one day be utilized in further research activities.

Further thanks deserves Dr. Balázs Barta, Attila Joós and colleagues from the Pannon Business Network Association for the practical problem definition, access to necessary data and the provided framework throughout the years. Last but not least I would like to express my appreciation for the endurance and loving words of my wife, Mrs. Zsuzsanna Tolnerné Lőrinc who was always there for me and shared my burden in all walks of life.

# 1

# Introduction

Students, researchers and practitioners from various walks of life often bump into numeric or non-numeric data of different origins that have to be analysed in order to validate some kind of assumption or research hypothesis. The need for a comprehensive understanding of accessible data is paramount, since modern Information Communication Technologies and data driven processes literally offer them and therefore generate the driving force for their utilization as well. Conventional statistical techniques offer a repository of a wide range of tools and procedures for answering questions on whether populations are significantly dissimilar, whether trends are observable or there are connections among different observations etc.

On the other hand, for being able to apply the methods of statistics a price has to be paid that is often too high for answering practical, real-life questions. This is especially valid for the usage of parametric statistical tools, where the existence of random samples providing independent observations is assumed. Furthermore, the scale of the variables is typically assumed to be on an interval or ratio measurement scale (i.e.: numeric variables where relative distances between elements are the same or even an absolute zero point exists), sample sizes are assumed to be satisfactorily large, approximately normally distributed and populations are approximately of equal variances [R1].

Econometrics and in general data analytical tasks regarding economics and finance are especially exposed to such circumstances where typical parametric statistical models like least squares estimators, maximum likelihood estimators etc. can underperform than would usually be optimal. This is often due to the underlying processes (e.g.: financial and social phenomena) that are currently not necessarily completely understood and modelled. This can materialize among others in non-normal-, skewed- or heavy-tailed data distributions, high level of (multi)collinearity or sampling biases. Additionally, the practical data at hand can pose challenges in the form of outliers that are often of multivariate kind and data sampling is often unrepeatable. Albeit these problems are general in a mathematical point of view, I will consider them in respect of my economic related investigations since corresponding literature is still not uniform in certain points and offers possibility for robust and non-parametric statistical approaches to generate new scientific added value.

Many classical statistical approaches are well-known for being non-robust and non-resistant. Thus, their results are highly dependent on the stochastic assumptions and properties of the sample elements at hand. Even small deviations from the model assumptions can violate the trustworthiness of such models that is hard to recognize afterwards. Robust statistics is concerned with the creation of statistical procedures that can still serve with satisfactorily reliable results while maintaining its acceptable statistical efficiency under such conditions. It can be viewed as a compromise between parametric and non-parametric statistical approaches. This scientific field gained momentum some 60 years ago

and nowadays counts as one of the highly relevant fields of modern statistics with several practical application possibilities [R2].

## 1.1 Motivation

Selection of the topic that formed the backbone of the current thesis was highly motivated and influenced by the everyday impressions and challenges that I faced during data analytical tasks at the Pannon Business Network Association at Szombathely, Hungary. Our initial goal was to investigate the resilient behaviour of Hungarian Small and Medium-sized Enterprises (SMEs) against various kinds of financial obstacles they bump into throughout their lifetimes based on accessible, annually reported financial attributes and other relevant metadata. This work gained further motivation by the Covid-19 pandemic that suddenly hit on several levels ranging from the individual through microeconomic, macroeconomic up to social level. It was unknown at the beginning of my research work in 2020 what economic aftermath the pandemic would have, but it could be suspected that not every company would be affected at the same level and reactions to the crisis –with an unknown extent– would be different as well.

Although no data was accessed after the pandemic, it became clear during my PhD program that literature statements on the ever-increasing exposure to economic turbulences due to globalization effects and growing interconnectedness were becoming perceivable on the individual's level as well. Rapid financial crisis within several industrial sectors due to the pandemic followed by supply chain problems, raw material shortages, fierce market reactions, political inferences into economic processes on national, EU and broader international level finally an outburst of a new war on the edge of Europe. A short summary of events to support how extremely relevant it is to explore the different aspects of entrepreneurial shock reactions and how it can be characterized, measured and predicted based on accessible data in an objective way since it is evident that collapse of individual economic stakeholders can have further negative impacts on other supply chain members, competitors, partners and even on regional or national level.

Corresponding literature often lacked to possess similar annual balance sheet and income statement information, onto I planned to build up my investigations. Consequently, the accessed materials often misaligned my research setting. On the other hand, additional relevant fields utilizing similar data showed contradictory, case- and data sensitive results that drew my professional attention towards robust and non-parametric statistical application possibilities. For this purpose, detailed analysis of the so-called *Most Frequent Value* (MFV) method was carried out, which was developed by Steiner et al. in relation to earth scientific investigations but no application to economic related fields were discovered up to the beginning of my research work. Therefore, throughout my investigations I devoted special interest to robust and non-parametric statistical techniques – in particular to the MFV concept – and their utilization together with other holistic approaches in order to be able to view our disposable economic related information in a novel way and possibly supplement related fields with my gained results in respect of methodology and practical application as well.

## 1.2 Professional and Scientific Objectives

Throughout my research- and professional work I often experienced that even in the era of big data where data is said to be abundant, how complicated it can be to get access to reasonable, relevant data of satisfactory quantity and quality in order to answer relatively easy to grasp research questions. Therefore, data still represents value and even seemingly

outlying data points shall be handled with care. Otherwise, their careless exclusion – besides altering the apparent type of the distribution of our data or artificially modifying its variance etc. – might lead to the abandonment of potentially important information as well or even lead to biased and false conclusions.

In case of many practically relevant questions, this might be unacceptable, since data sampling is mostly unrepeatable or high costs would be associated to it. Either way, researchers and practitioners have to be equipped with the methodological toolkit to work around with "small data" or data that is contaminated with outlying observations and/or typically does not have a Gaussian distribution.

Since data of necessary quality and quantity is often hard to obtain, researchers have to settle for what they can access and use or develop techniques that promise satisfactorily stable results at a still acceptable statistical efficiency. Fortunately, for natural sciences and engineering related fields in many situations background processes can be recognized and by measurements observed that offers the possibility for reproducible measurements. In case of economics related fields however, – that from application point of view lies in the focus of my interest – this is barely valid. This fact, namely being one of the observational sciences, brings further biases and uncertainties to investigations above naturally existing statistical ones and often makes gained results reasonable only *ex post* and hard to validate based on objective measurement data.

From this end during my theoretical investigations and practical work the following research goals were outlined that I considered as important to achieve in my present thesis or future work:

- The applicability of robust statistical approaches has to be demonstrated in case of economic related data analysis and present gained results in alignment of data-based previous investigations.
- Awareness should be raised for case-sensitive problems, where the non-uniform standpoint of literature can possibly be traced back to the application of data sensitive methods and/or where accessible data typically cannot fulfill assumptions of conventional statistical methods.
- Application opportunities of the MFV concept shall be exhausted in case of economic related investigations as much as possible.
- Practically applicable data analysis "steps" should be investigated from a methodological point of view regarding robustness and resistivity, and recommendations shall be provided for the application of the MFV concept.
- Multidimensional outliers should be revealed in our data as far as possible in an automatic and objective way. If outlying observations are detected, from the practical application point of view an interpretable answer should be outlined based on the applied methodology.
- Robust outlier detection techniques shall be elaborated and developed for economics related questions regarding widely used linear regression and k-means clustering in order to make empirical investigations more accurate and trustworthy.
- The utilization of the MFV concept regarding economics related investigations should be promoted, since for problems similar to earth sciences – for which problems the MFV concept was originally designed – a higher statistical efficiency can be achieved than could be expected generally from conventional parametric techniques. Therefore, better information extraction can be hoped from valuable data that might be collected at high costs.

- If applicable, for certain investigations, applicability of external data sources shall be performed and highlighted in order to gain further support for economics related, data-based findings.
- Investigation of resilient/non-resilient shock reaction behaviour of SMEs shall be performed with appropriate statistical techniques in alignment with the obtainable financial- and other metadata in order to gain generally valid, reproducible and as far as possible data insensitive findings.
- Results of state-of-the-art deep learning methods shall be compared with conventional classification (e.g.: logistic regression, decision tree etc.) techniques for the grouping of resilient/non-resilient behaviour of SMEs.
- Main characteristics and driving factors for resilient behaviour of SMEs shall be identified via appropriate techniques in alignment with obtainable data in order to fabricate a ”predictive model for economic resilience”.

## 1.3 Chronological Order of the Research

The content of the thesis was built up in a logically separated manner instead of following the chronological order of the performed research activities. Nevertheless, the timeline of the research is also provided below in order to elucidate some design decisions regarding the segmentation and structure of the thesis. The subsequent enumeration describes the main cornerstones of my work:

1. Literature research regarding economic resilience and identification of misalignment of own empirical intentions with accessed approaches [T1].
2. Investigation of robust statistical approaches with special emphasis on the MFV concept for being able to investigate economic data in a robust and outlier resistant manner. At this stage, commonly used data analytical steps (correlation analysis, linear regression based on the least squares) were considered that were applied for the examination of the well-known absolute economic $\beta$-convergence problem of the EU regions [T2, T3, T4].
3. The above application area was further assessed by a non-parametric statistical procedure in order to supplement the related field by a further, less data- and case-sensitive mathematical method on economic convergence of West- and East European countries and regions [T5].
4. Investigation of robust outlier detection possibilities regarding linear regression [T3, T4] and seeking for further application possibilities of the MFV concept in outlier detection in unsupervised learning problems [T6].
5. Data acquisition on annual financial data from Hungarian SMEs from the processing industry, data processing and non-parametric hypothesis testing based on a designed resilience index [T7].
6. Analysis of R&D related project data of the Horizon 2020 program as an extension possibility of resilience investigations to a regional scale and identification of outliers on population level [T8, T9].
7. Classification attempt of company-year observations of Hungarian SMEs based on their resilient behaviour via decision trees and logistic regression [T10, T11].
8. Further classification trials of the observations via deep learning techniques.

## 1.4 Outline of the Thesis

The rest of the dissertation is structured in the following way: In Chapter 2. a brief overview of robust statistics is provided together with the introduction of the MFV method. Chapter 3. details robust correlation calculation and robust linear regression based on the MFV concept and its application for the well-known absolute $\beta$-convergence. Chapter 4. connects to this by illustrating robust outlier detection, likewise based on the MFV Concept. In Chapter 5. I present non-parametric hypothesis testing among SMEs regarding their economic shock reactions based on annual financial information. Finally, the main findings are summarised in Chapter 6. and conclusions are outlined in Chapter 7.

# 2

# Methodological Background

Empirical investigations presented in the dissertation highly build upon the advantageous properties of robust statistics and in particular the MFV method compared to conventional statistical approaches in case of data equipped with such properties that I was working with. This chapter details and summarises the main ideas and terms that are essential for the understanding of each thesis point of my work.

## 2.1 Robust Statistics

The literature of robust statistics dates back to the 1960s, when the establishment of the field laid the fundaments for a rich section of modern statistics. The theory of robust statistics is concerned with small deviations measured from model assumptions of conventional statistical methods and seeks solutions for providing procedures that are more reliable in the presence of real-life, contaminated data in the neighbourhood of an assumed model but can still maintain a reasonable efficiency. The theory helps to quantify the extent of deviations from model assumptions, where even small ones can lead to the decrease of statistical efficiency to a great extent. Hence, it is important to ensure whether underlying data is in alignment with model assumptions. Especially in practical use cases, it is favourable to select such models that are mostly affected by the majority of the observations and are less sensitive to single ”abnormalities”. Having this said, it is also desired to utilize practices that help data analysts to perform investigations in a preferably automatic manner by identifying anomalies compared to the ”bulk” of the data systematically and promote usage of simpler descriptive models without the need of introduction of more sophisticated methods to understand data with more variation or the need of data exclusion [R2].

There are however more expressive and more plain techniques to treat gross observations like Tukey’s fence that are based on some kind of rejection of measurements above certain, data specific limits. Winsorisation – substitution of excessive data points with data specific limit values – is also commonly used. Nevertheless, besides artificially altering data variances and introducing extra bias, such outlier spotting tools can become troublesome for multivariate problems or among others for multiple regression tasks. Besides constructing less sensitive models to large deviations, efficiency of given procedures – that can be understood as the information extraction rate at a given sample size – is of great importance. Since statistical efficiency is bound to assumptions about the data it has to be examined what happens to the efficiency of an estimator if the data distribution does not meet assumptions of the model under which the estimator would be optimal. Distribution free procedures offer an alternative for strong dependence on the data. However, these often prove to be less efficient and groundless to be applied by switching from a certain model to an unspecified distribution. Robust statistics is technically a compromise between classical

and distribution-free methods that broadens the assumptions of parametric models while keeping efficiency high [R3].

Modern robust statistics might as well be defined as a combination of computer science and mathematics for data analytical tasks that become more important with the increase of data quantity, dimensionality and the development in computing power. The key idea is to have procedures suited for routine analyses that do not delete outliers, but decrease their influence by a proper weighting. Thereby, protection can be provided up to a certain extent against their negative impact and data structure can be mapped more effectively. Nevertheless, besides outliers, non-independent observations, heavy tailed distributions, autocorrelated or heteroscedastic errors may also exert bias on statistical inferences, which are often in deep relation with econometric related issues. In order to apply robust procedures, the applicability of some kind of model has to be assumed according to which the sample was supposed to be generated. This might be naive in some cases and can question the robustness of our estimate fundamentally. Similarly, outliers can only be identified compared to an assumed model that is not provided in advance, therefore different robust approaches shall not be applied blindly but alongside (also) with conventional models and by combining field relevant experience. Nevertheless, despite having demonstrated the powerful applicability, due to strong traditions of given fields (e.g.: superiority of ordinary least squares in regression problems), the abundant presence of skewed asymmetric distributions to which the main idea does not fit well and the necessity for time-efficient algorithms robust approaches have not become a standard way for data analysis [R3].

In the following, I will introduce the most important notions of robust statistics based on [R4, R5, R6]:

**Robustness:** A statistical method is said to be robust if it is only slightly depending on its assumptions.

**Resistivity:** A statistical method is said to be outlier resistant if the presence of outliers and other anomalies only slightly modify the results.

When considering a parametric model, we often assume to have a sample of identical and independently distributed observations with a common $F_{\Theta_0}$ distribution, which belongs to a set of $F_\Theta$ distributions, where $\Theta \in \mathbb{R}^k$. The task is to estimate $\Theta_0$ based on the observations. In other words, we seek a mapping that takes the value of $\Theta_0$ if the background distribution is $F_{\Theta_0}$. Robust estimates treat the effect of unwanted deviations and anomalies by not considering parametric models, but rather their arbitrary small ambience. The distance of the estimates can be measured for instance by the Prohorov-distance. For this purpose, let $(X, \mathbf{A})$ be a measurable space, where $X$ is a separable, compact metric space and $\mathbf{A}$ is the $\sigma$-algebra induced by the topology. Let $\rho$ be a metric on $(X, \mathbf{A})$, then the vicinity of an observation can be given as $A^\epsilon = \{x \in X, A \in \mathbf{A} | \rho(A, x) < \epsilon\}$.

**Definition:** The *Prohorov-distance* of two $P, Q \in (X, \mathbf{A})$ probabilistic measures is

$$\pi(P, Q) = \inf\{\epsilon > 0 | \forall A \in \mathbf{A} : P(A) \leq Q(A^\epsilon) + \epsilon \text{ and } Q(A) \leq P(A^\epsilon) + \epsilon\}$$

**Definition:** Let $\mathbf{F} \in (X, \mathbf{A})$ a measurable set of probability distributions and $\mathbf{F}_n \subset \mathbf{F}$ for $n \in \mathbb{N}$, which defines a series of distributions determined by the $(X_1, X_2, ..., X_n) \in X^n$ observations. The $\{\theta_n\}$ is called an *estimation series* if $\{\theta_n\} : \mathbf{F}_n \to \mathbb{R}^k$ for $\forall n \in \mathbb{N}$, which is a mapping that assigns a parameter to a $n$-dimensional observation.

In the following when considering estimates we will understand a series of estimates, and we will restrict ourselves to estimates, the distributions of which – with respect to the Prohorov-distance – is a continuous functional of the true distribution.

Let $\xi_1, \xi_2, ..., \xi_n, ..$ a countably infinite set of random variables with a common $F$ distribution, and let $F_n$ be the probability measure defined by the first $n$ random variables. Then the $\theta_n : \mathbf{F}_n \to \mathbb{R}^k$ mapping induces a probability measure on $\mathbb{R}^k$, which is the distribution of $\theta_n$ with respect to $\mathbf{F}$. Let this probability measure be denoted by $L_F(\theta_n)$.

With an increasing number of observations, the true distribution is supposed to be nearer and nearer to the parametric model in order to keep the distribution of the estimate close to the distribution of the given model. For a proper behaviour of the estimation series, we restrict the continuity to be uniform with $n$, which leads us to the qualitative definition of robustness.

**Definition:** The $\{\theta_n | n \in \mathbb{N}\}$ estimation series *qualitative robust with respect to the $F$ probability measure* if

$$\forall \epsilon > 0 : \exists \delta > 0 : \forall n \in \mathbb{N}, \forall G \in \mathbf{F} :$$

$$\pi(F, G) < \delta \Rightarrow \pi(L_F(\theta_n), L_G(\theta_n)) < \epsilon.$$

**Definition:** The $\{\theta_n\}$ estimation series is qualitative robust on the $\mathbf{G}$ set of distributions if it is robust for $\forall G \in \mathbb{G}$. If $\mathbf{F} = \mathbf{G}$ then $\{\theta_n\}$ is *qualitative robust*

With an increasing proportion of faulty observations, there is a limit expected above which the estimate will no longer serve with adequate results and can lead to a complete inefficiency. The breakdown point is supposed to measure the extent of the robustness of an estimate in a sense how far the real distribution can be from the parametric model by providing the percentage of the observations that can be an anomaly without having the model being "destroyed".

**Definition:** The $\epsilon^*$ *breakdown point* of the $\{\theta_n\}$ estimation series with respect to $F$:

$$\delta^* = \sup\{\delta \leq 1 : \exists K_\delta \text{ compact subset of the parameter-space for which}$$

$$\pi(F, G) < \delta \Rightarrow \lim_{n \to \infty} G(\{\theta_n \in K_\delta\}) = 1\}.$$

In order to characterize the sensitivity of the estimate against deviations, we introduce the influence curve, that measures the asymptotic bias exerted onto the estimate caused by point-like contaminations present in the sample.

**Definition:** For a $\theta$ functional in case of a $F$ distribution function the *influence function* or *IC-function*:

$$IC(x, F, \theta) = \lim_{t \to \infty} \frac{\theta((1-t) \cdot F + t \cdot \Delta_x) - \theta(F)}{t},$$

for those $x \in X$ points where the above limit exists and $\Delta_x$ denotes the point-like probability measure.

With the help of the IC-function, we can gain quantitative information on the bias in case of an assumed $F$ distribution and given algorithm of the estimate caused by gross errors. Furthermore, for contamination-free samples, it can be seen to which extent each sample element can contribute to the calculation of our estimate. This follows from the definition, since with a good approximation a single $x$ data point arising with $\Delta$ probability will contribute to our estimate with $\Delta\theta \approx IC(x, F, \theta) \cdot \Delta$, where in case of finite samples $\Delta = \frac{1}{n}$ can be used. The analytic form of the IC-function in most of the cases is rather complicated, therefore in practice the shapes of the curves are investigated and *IC-curve* nomenclature is used instead.

Utilizing the IC-function, the asymptotic normality of the estimates can be given:

$$L_F(\sqrt{n}[\theta_n - \theta(F)]) \xrightarrow[n \to \infty]{} \mathcal{N}(0, V(F, \theta)),$$

where the convergence is stochastic and $V(F, \theta) = \int IF(x, F, \theta)^2 dF(x)$ is the asymptotic variance.

For the characterisation of robustness from a quantitative point of view the upper limit of the influence function is used that properly represents the local behaviour of the estimate. This supremum represents the maximal effect of outliers, therefore it can be denoted as the sensitivity of the estimate.

**Definition:** Let us suppose that the $IC(x,F,\theta)$ exists, then the *gross-error sensitivity* of the estimate is:

$$\gamma^*(F,\theta) = \sup_x |IF(x,F,\theta)|,$$

where the supremum is interpreted for all $x$, where $IC(x,F,\theta)$ exists.

**Definition:** A $\theta$ estimation is said to be *B-robust* in $F$ if $\gamma^*(F,\theta) < \infty$

In case of a real sample, the maximum-likelihood principle serves with an extended theoretical background for a proper parameter estimation when there is an assumption on the background distribution. The principle illustratively states that given a probability distribution that sample is selected with the highest probability for which:

$$\prod_{i=1}^{n} f(x_i,\theta) = \max. \Rightarrow \sum_{i=1}^{n} -\ln f(x_i,\theta) = \min.$$

According to [R7] if we substitute an arbitrarily differentiable function into the place of $-\ln f$ we can generalize the maximum-likelihood principle even for those cases when there is no prior knowledge on the background distribution.

**Definition:** Let us suppose a $(X, \mathbf{A}, P_\theta), \theta \in \Theta \subset \mathbb{R}^k$ statistical space dominated by a $\mu$ measure, thus with existing distribution functions and a $\xi_1, \xi_2, ...\xi_n, ...$ ideally and independently distributed sample with a common $f(x,\theta)$ distribution function. The $\rho : X \bigtimes \Theta \to \mathbb{R}$ estimates are called *M-estimates* if they fulfill the

$$\sum_{i=1}^{n} \rho(\xi_i,\theta) = \min.$$

criterion for the $\theta$ parameter for the given sample.

The $\rho(\xi_i,\theta)$ M-estimates are often calculated based on their $\psi(\xi_i,\theta)_m = \dfrac{\partial\rho(\xi_i,\theta)}{\partial\theta_m}$, partial derivatives $(m = 1,2,..,k)$ if they exist, which transforms the above implicit expression to the following equation system:

$$\sum_{i=1}^{n} \psi(\xi_i,\theta)_m = 0 \qquad (m = 1,2,..,k).$$

In the following, we restrict ourselves to $X = \mathbb{R}$, $\Theta = \mathbb{R}$. Furthermore, let us denote the location parameter by $T$ and be $F_T(x) = F(x-T)$, $T_0 = 0$. In this case we have $\psi$ functions for which $\psi(x,T) = \psi(x-T)$ holds. The next theorem highlights the main features of the M-estimates:

**Theorem (Hampel):** Let $\psi$ be a increasing function that takes up positive and negative values as well, furthermore let $T(F)$ be such for which

$$\int \psi(x - T(F))dF(x) = 0$$

holds. Then the estimate for the $T$ location parameter is B-robust and qualitative robust at $F_0$ if and only if $\psi$ is bounded and $T(F_0)$ is unique. The breakdown point:

$$\epsilon^* = \frac{\eta}{1+\eta},$$

where

$$\eta = \min\left\{-\frac{\psi(-\infty)}{\psi(\infty)}; -\frac{\psi(\infty)}{\psi(-\infty)}\right\}$$

If $\psi$ is unbounded, then $T$ is neither B-robust nor qualitative robust and $\epsilon^* = 0$. The best possible $\epsilon^* = \dfrac{1}{2}$ breakdown point can be achieved if $\psi(-\infty) = -\psi(+\infty)$.

We call the joint estimate of the location- and scale parameter of an M-estimate all of those $(T_n, S_n)$ pairs that are determined by the following pair of equations:

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{S_n}\right) = 0$$

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - T_n}{S_n}\right) = 0$$

From this follows that $T_n = T(F_n)$ and $S_n = S(F_n)$ can be expressed by $T$ and $S$ functionals that are defined by:

$$\int_{-\infty}^{\infty} \psi\left(\frac{x_i - T_n}{S_n}\right) F(dx) = 0$$

$$\int_{-\infty}^{\infty} \chi\left(\frac{x_i - T_n}{S_n}\right) F(dx) = 0$$

For the calculations $\psi$ and $\chi$ shall be chosen in an adequate way. It is a basic requirement for instance for $\psi$ that in case of a symmetric $f(x)$ distribution function it provides the $T$ symmetry point.

If the above two equations hold, it can be shown that the following two equations have to be fulfilled simultaneously:

$$IC(x,F,T)\cdot\int_{-\infty}^{\infty}\psi'\left(\frac{y-T}{S}\right)F(dy)+IC(x,F,S)\cdot\int_{-\infty}^{\infty}\psi'\left(\frac{y-T}{S}\right)\cdot\frac{y-T}{S}F(dy)=S\cdot\psi\left(\frac{y-T}{S}\right)$$

$$IC(x,F,T)\cdot\int_{-\infty}^{\infty}\chi'\left(\frac{y-T}{S}\right)F(dy)+IC(x,F,S)\cdot\int_{-\infty}^{\infty}\chi'\left(\frac{y-T}{S}\right)\cdot\frac{y-T}{S}F(dy)=S\cdot\chi\left(\frac{y-T}{S}\right)$$

This equation system theoretically serves with the $IC(x,F,T)$ and $IC(x,F,S)$ function pairs when the $\psi$ and $\chi$ function pairs – that determines the M-estimate – and the distribution function of the data are known. Technically however, the exact numerical calculations are problematic in general cases. Therefore, often symmetric distributions are considered ($\psi$ is odd and $\chi$ is even) that makes each of the integrals equal to zero and the IC-functions can be expressed in a much simpler way:

$$IC(x,F,T) = \frac{S(F)}{\int_{-\infty}^{\infty}\psi'\left(\frac{y}{S(F)}\right)F(dy)}\cdot\psi\left(\frac{x}{S(F)}\right)$$

$$IC(x,F,S) = \frac{S(F)}{\int_{-\infty}^{\infty}\chi'\left(\frac{y}{S(F)}\right)\frac{y}{S(F)}F(dy)}\cdot\chi\left(\frac{x}{S(F)}\right)$$

This also serves with an illustrative and from a practical point of view significant interpretation possibility of the $\psi$ and $\chi$ functions. The denominators of the fractions result to be definite numbers, therefore the $\psi$ and $\chi$ functions are proportional to the IC-functions. The knowledge of the $\psi$ and $\chi$ functions is of great practical importance as well, since they are needed for the determination of the statistical efficiency of the parameter estimates that can be used for comparing algorithms regarding their information extraction rate at a fixed sample size.

## 2.2 The Most Frequent Value

The derivation of the Most Frequent Value (MFV) and elaboration of the connected theory and applications were done by a Hungarian research group involving among others Ferenc Steiner, László Csernyák or Béla Hajagos in the 1970s and 1980s. The main application areas published are related to earth sciences [R8, R9, R10, R11], but even astrology related fields can be found [R12, R13]. The concept originates from a very practical and demonstrative problem: namely, the definition of a "most characteristic attribute" of a given data set that can be even of small sample size but of practical importance (e.g.: has engineering origin).

The sample median may serve as a robust choice, however it may be too radical since it does not consider distances of far-lying observations from the data concentration and by definition lies symmetric within the distribution even in asymmetric cases. On the other hand, mean values might be too sensitive to such gross observations that are frequently hard to be judged as erroneous and may prove invaluable just to trim away from the rest of the sample. A weighted average can be an intermediary alternative that considers observations closer to the data concentration with a higher weight by viewing them more "characteristic" or important for the bulk of the sample. By denoting the location parameter of the distribution by $T$ the weighted average representing it given the $\{x_i\}, (i = 1...n)$ sample:

$$T = \frac{\sum_{i=1}^{n} x_i \cdot w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} x_i \cdot w(x_i - T)}{\sum_{i=1}^{n} w(x_i - T)}, \tag{2.1}$$

where the $w_i$ weight functions are symmetric to the $T$ location parameter. Unfortunately, the above approach leads to an iterative procedure, since the location parameter should be known in advance in order to be able to select the weights that have a maximum around it. By selecting the weights according to a Cauchy distribution, a practical sample dependent location parameter-like quantity $M_n$ arises as the empirical estimate of the true location parameter that is called the Most Frequent Value:

$$M_n = \frac{\sum_{i=1}^{n} \frac{\epsilon^2}{\epsilon^2 + (x_i - M_n)^2} \cdot x_i}{\sum_{i=1}^{n} \frac{\epsilon^2}{\epsilon^2 + (x_i - M_n)^2}}. \tag{2.2}$$

The selection of the $\epsilon$ scale parameter characterizes the steepness of the weight functions. If $\epsilon$ is large, the corresponding weight functions do not downweight far-lying observations too much, while selecting it small results in a strong downweighting and effectively trimming of the remote data measured from the bulk of the data and leads to a decreased sample size. In this sense, the $\epsilon$ parameter can be linked to the dispersion of the data around the empirical estimation of the true location parameter and therefore called as *dihesion*. The same formulation can be kept for continuous distributions characterized by a $f(x)$ density function after simplification by $\epsilon^2$:

$$M = \frac{\int_{-\infty}^{\infty} \frac{x}{(M - x)^2 + \epsilon^2} \cdot f(x)dx}{\int_{-\infty}^{\infty} \frac{1}{(M - x)^2 + \epsilon^2} \cdot f(x)dx} \tag{2.3}$$

The so far unknown dihesion value can be used as a measure of data dispersion and can be estimated based on the Schwarz inequality by seeking the Cauchy distribution used as

the weight function that is the most similar to the given data distribution. According to the Schwarz inequality:

$$\left| \int_a^b g(x) \cdot h(x) dx \right| \leq \sqrt{\int_a^b g^2(x) dx \cdot \int_a^b h^2(x) dx}, \tag{2.4}$$

where equality holds if $g(x) = \text{const} \cdot h(x)$. Consequently, the

$$\frac{\left| \int_a^b g(x) \cdot h(x) dx \right|}{\sqrt{\int_a^b g^2(x) dx \cdot \int_a^b h^2(x) dx}} \tag{2.5}$$

expression can be used to characterize the similarity in question. By substituting the proposed weight function and $f(x)$ distribution function for achieving the maximum similarity the following criterion has to be fulfilled:

$$\int_{-\infty}^{\infty} \frac{\epsilon^{3/2}}{\epsilon^2 + (x - M)^2} f(x) dx = \max. \tag{2.6}$$

In this context the dihesion provides an illustrative meaning as the half of the interquartile range – also the probable error in this case – of the Cauchy distribution that is the most similar to the data distribution at hand. For the practical calculation of the dihesion based on Eq. 2.6. Steiner et al. suggested the following iterative algorithm:

$$\epsilon_{l+1}^2(x_i) = \frac{3 \cdot \sum_{i=1}^{n} \frac{(x_i - M_n)^2}{[\epsilon_l^2 + (x_i - M_n)^2]^2}}{\sum_{i=1}^{n} \frac{1}{[\epsilon_l^2 + (x_i - M_n)^2]^2}} \tag{2.7}$$

Since a wide range of distribution types are conceivable in order to maintain a high statistical efficiency for the usage of the MFV at given cases a tuning constant is recommended to be used and finally the following formula is advised to calculate the MFV:

$$M_{n+1}(k, x_i) = \frac{\sum_{i=1}^{n} \frac{(k \cdot \epsilon_l)^2}{(k \cdot \epsilon_l)^2 + (x_i - M_n)^2} \cdot x_i}{\sum_{i=1}^{n} \frac{(k \cdot \epsilon_l)^2}{(k \cdot \epsilon_l)^2 + (x_i - M_n)^2}}, \tag{2.8}$$

which means the weighting functions to have the form of:

$$w(k, x_i) = \frac{(k \cdot \epsilon_l(x_i))^2}{(k \cdot \epsilon_l(x_i))^2 + (x_i - M_n(k, x))^2}. \tag{2.9}$$

In practical calculations the tuning parameter is advised to be $k = 2$ if there is no *a priori* knowledge of the distribution, however for long-tailed distributions $k = 3$ is recommended [R6, R14].

Equation 2.7. and 2.8. leads to a "ping-pong" iteration that converges typically within $10 - 15$ iterations and serves with a robust location and scale parameter for the unknown data distribution. For initialization, the median is advised for the MFV and the median absolute deviation (MAD) for the dihesion in order to reach a local minimum.

The same formulas can be achieved by investigating the minimization problem of the I-divergence:

$$I(f||g) = \int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)} dx = \min., \tag{2.10}$$

where $f(x)$ is the a priori unknown distribution function and $g(x)$ is the substitution function for the former one. Supposing that the substituting distribution is in form of $g(x,T)$, – that is a symmetric, two-times differentiable function by the location parameter $T$ and the order of integration and differentiation is exchangeable – the following conditions need to be fulfilled to minimize the distance of the two functions defined by the above Kullback-Leibler distance:

$$\int_{-\infty}^{\infty} \frac{\partial g(x,T)}{\partial T} \cdot \frac{f(x)}{g(x,T)} dx = 0 \tag{2.11}$$

$$\int_{-\infty}^{\infty} \left[ \frac{\partial g(x,T)}{\partial T} \cdot \frac{1}{g(x,T)} \right]^2 f(x) dx - \int_{-\infty}^{\infty} \frac{\partial^2 g(x,T)}{\partial T^2} \cdot \frac{f(x)}{g(x,T)} dx > 0. \tag{2.12}$$

Assuming that the second part of Eq. 2.12. is $0$ the minimum value of $I(f||g)$ is granted since $f(x)$ is a distribution function. Hence, the above formulas determine a location and scale parameter that in case of a Cauchy distribution as a substituting function leads to the same iterative formula as given in Eq. 2.7. and Eq. 2.8. The corresponding $\psi(x)$ function of the MFV:

$$\psi(x) = \frac{x}{1+x^2} \tag{2.13}$$

and of the dihesion:

$$\chi(x) = \frac{3x^2 - 1}{(1+x^2)^2} \tag{2.14}$$

Numerical calculations have shown however that in a few cases convergence cannot be achieved, therefore Fegyverneki in [R15] advised a modification where the minimization of the I-divergence did not assume the second part of Eq. 2.8. to be zero. The obtained equation system for the modified $\psi(x)$ function for location and scale parameter with a Cauchy substituting function:

$$\psi(x) = \frac{x}{1+x^2} \tag{2.15}$$

$$\chi(x) = \frac{x^2 - 1}{1+x^2} \tag{2.16}$$

Only the function defining the dihesion changed, but thereby only one solution exists, and it is a maximum likelihood estimate which ensures various favourable attributes. A numeric algorithm is provided for the calculation of the estimates in [R16], which are B-, V- and qualitative robust. Let $c_n$ and $s_n$ denote the estimates for the MFV and dihesion. According to the results of [R15] the joint distribution of $(T_n, S_n)$ is asymptotically normal and the corresponding breakdown points:

$$\delta^*(T_n) = 0.5 \tag{2.17}$$

$$\delta^*(S_n) = \frac{-\chi(0)}{\chi(-\infty) - \chi(0)} = \frac{1}{3}. \tag{2.18}$$

The code 2.1. shows the implementation for the modified MFV algorithm in a python environment based on [R16]. The algorithm uses the median and MAD for the initialization of the location and scale parameters respectively of the Cauchy distribution most similar to the distribution of the given sample. The algorithm shows convergence typically in $5-7$ iterations (if $|\epsilon_{k+1} - \epsilon_k| \leq 10^{-5}$ is requested), however the sample size and convergence threshold can highly influence the computational time needed. As Fig. 2.2. illustrates on a

logarithmic scale of the sample size and convergence threshold there is a steep increase in time needed for larger datasets and more precise convergence criteria. This latter can be relieved, since in practical calculations $10^{-5} - 10^{-3}$ thresholds lead to sufficiently precise results. Regarding samples size, the MFV algorithm is not necessary for estimation of location and scale when disposable data are abundant. In such cases median and MAD values are reliably calculated as parameter estimates, therefore the MFV algorithm is primarily advised to be used in case of small sample sizes when data are scarce, invaluable and costly to obtain. Nevertheless, it has to be mentioned that in higher dimensions due to the "curse of dimensionality" much larger sample sizes might be required for adequate estimates.

```
import numpy as np
from scipy import stats
import math

def modified_MFV(y, threshold = 10**(-5)):
'''
Calculation of MFV and dihesion.
:y: List of numeric data
:threshold: Parameter for finetuning convergence limit.
:return: MFV and dihesion of the input data
'''
M_old = np.median(y)
eps_old = stats.median_abs_deviation(y)
n = len(y)
try:
    u_old = (y-M_old)/eps_old
    e0_old = 1/n*sum([1/(1+math.pow(x,2)) for x in u_old])
    e1_old = 1/n*sum([x/(1+math.pow(x,2)) for x in u_old])
    e2_old = 1/n*sum([math.pow(x,2)/(1+math.pow(x,2)) for x in u_old])

    i = 1
    diff = 1
    while abs(diff) > threshold:
        M_new = M_old + eps_old * e1_old/e0_old
        eps_new = eps_old * math.sqrt(1/e0_old-1)

        u_new = (y-M_new)/eps_new
        e0_new = 1/n*sum([1/(1+math.pow(x,2)) for x in u_new])
        e1_new = 1/n*sum([x/(1+math.pow(x,2)) for x in u_new])
        e2_new = 1/n*sum([math.pow(x,2)/(1+math.pow(x,2)) for x in u_new])

        diff = max( abs(M_new-M_old), abs(eps_new-eps_old) )

        M_old, eps_old = M_new, eps_new
        e0_old, e1_old, e2_old = e0_new, e1_new, e2_new

        i = i+1
except:
    raise ValueError
return M_new, eps_new
```

**Figure 2.1:** Python code for the calculation of MFV- and dihesion values of a simple data set.

Figure 2.3. represents the effect of outliers on selected location and scale parameters. 30 data points have been generated from a $\mathcal{N}(0,1)$ distribution and outliers have been added to it based on a $\mathcal{N}(5,1)$ distribution in a $5\%$ and $15\%$ amount. Even a small portion of outliers increases the standard deviation dramatically and drags the mean value significantly towards the newly added instances. This latter might be a bigger problem from a practical point of view, since outlier resistivity is often key for stability of predictions, nonetheless bigger standard deviations may contribute to wider confidence intervals and higher uncertainty of the estimates.

From the presented demonstrative example it is obvious that the MFV and dihesion has similar properties as the median and MAD, however from computational point of

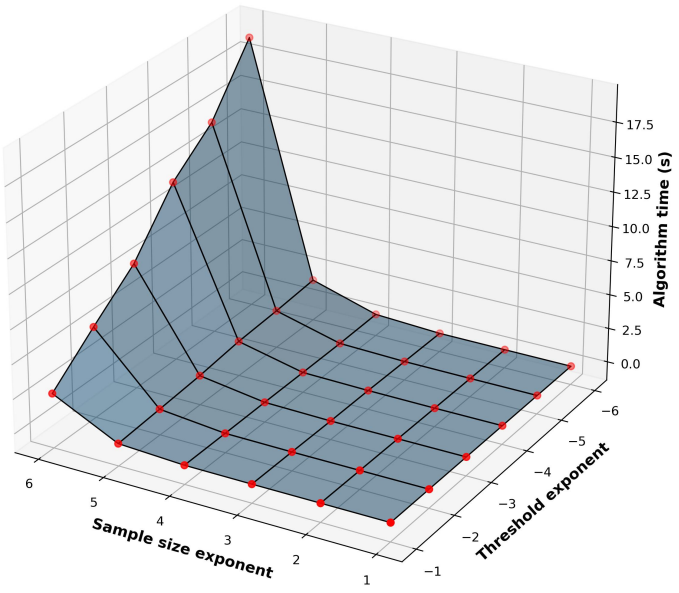



**Figure 2.2:** Computing requirements of the MFV algorithm as a function of sample size and threshold of convergence.

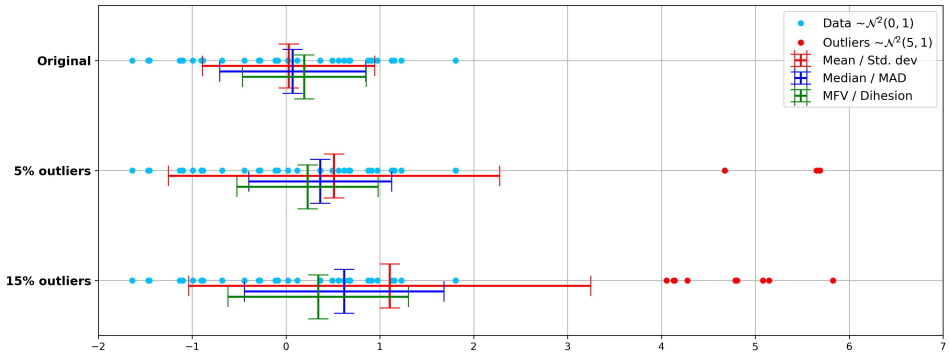



**Figure 2.3:** Effect of outliers in case of 1D data on location and scale parameters.

view it is more expensive. The reason behind the usefulness of the concept is the higher statistical efficiency for a wider range of distributions compared to the median and MAD estimates [R14]. Moreover, there is a possibility for generalization for higher dimensions and applicability for other optimization procedures like regression problems. It also has to be highlighted that besides mathematical considerations the MFV, as a weighted average of data points, where the weights are assigned according to closeness to the data concentration, has deeper physical meaning as well. Since repeated measurements on a single value more often serve with similar values close to the real value, it can be straightforwardly assumed that points at the data concentration are from more precise observations and shall be considered more seriously for the investigation of the ”bulk” of the data. Therefore, the MFV concept has more practical sense at interpretation stages [R6]. While the median can only be one of the data points, the MFV can be any arbitrary value in between that is not necessarily included in the database. Due to the basic complexity of robust statistical approaches in general, it is nice that the MFV concept can be made illustrative for practical purposes and can be applied routine-wise in statistical analyses.

# 3

# MFV-based Linear Regression for Regional Economic Convergence

## Thesis Group 1: Application of robust and non-parametric statistical tools regarding economic convergence

***Thesis 1***
**I have identified via robust and non-parametric statistical approaches that convergence among EU countries and regions based on the absolute $\beta$-convergence regarding GDP per capita is of lesser degree than could be estimated based on the conventional Ordinary Least Squares (OLS) linear regression. I have also ascertained that economic convergence regarding GDP per capita also exists on the whole data-population level between EU member states and regions connected to the EU before and after 2004.**

Publications relevant to the theses: [T2, T3, T4, T5, T9].

***Thesis 1.1***
**With robust correlation calculations based on the Most Frequent Value concept, I have shown that there is less correlation among initial values and growth rates of GDP and NDI financial indicators of EU member states and regions, which suggest a slower pace of convergence than is suggested by OLS linear regression based absolute $\beta$-convergence theorem.**

Publications relevant to the theses: [T2].

***Thesis 1.2***
**With robust linear regression, based on the Most Frequent Value concept, I have shown that convergence of economies of EU member states and regions is of slower pace than provided by OLS linear regression based absolute $\beta$-convergence theorem.**

Publications relevant to the theses: [T3, T4].

***Thesis 1.3***
**With non-parametric hypothesis testing I have shown the existence of convergence among subgroups of economies of EU member states and regions connected to the EU before and after 2004 regarding GDP per capita on the whole data-population level. Based on the non-parametric approach, I have suggested a method for the estimation of rate of convergence and for the time needed for member regions to catch up.**

Publications relevant to the theses: [T5].

***Thesis 1.3.1***
**Having created the regional level social networks of participating organizations of the Horizon 2020 framework, I have shown that regions of EU member states connected after 2004 had generated more connections at the same level of average annual GDP or R&D subsidy given by the EU as member states and regions connected before 2004. This observation further supports the existence of regional convergence.**

Publications relevant to the theses: [T9].

In the following, I will demonstrate the effective applicability of the MFV procedure on data of economic origin regarding robust treatment of a linear regression problem. From the application point of view my aim with the outlined case study is to join to other scientific results that agree with the existence of economic convergence among regions of the European Union. Moreover, I would also like to fine-tune the existing results with new methodological approaches in the field with a holistic approach and by utilizing different, to each other complementary datasets.

## 3.1 Overview of Related Literature

Economic convergence is a macroeconomic concept that is in direct relation with economic growth. According to economic growth theories, economies of countries and regions tend to develop towards an equilibrium growth path. Deviations from the equilibrium can lead to unemployment or prolonged inflation. This convergence can be characterised among others with the absolute $\beta$-convergence theory, which postulates an inverse relationship between GDP (or similar income measure) growth per capita and its initial economic level [R17].

Nevertheless, literature elaborates on numerous approaches of economic growth and convergence. The model developed by Sir Roy Harrod and Evsey Domar already in the 1930s consideres the outstanding role of capital investment and savings with regard to economic growth. It concludes that higher saving rates can lead to more insvestment and thereby faster growth when fixed capital-output ratio is assumed. Based on their description, if poorer regions increase their investment and capital accumulation faster growth will be the consequence and eventually convergence to more wealthier territories. This model already draws attention to the importance of capital formation in regional growth. Nonetheless, it is only applicable to regions of low-income where primaprily capital constrains are in the way of convergence. Furthermore, it does not consider human capital and innovation, which are important pillars of sustained growth [R18].

The neoclassical growth model is a simple and intuitive approarch that is widely applied for empricial convergence calculations in studies since it allows easy comparison across

regions. The concept dates back to the 1950s and suggests that economic growth is a consequence of capital- and labor accumulation and exogenous technological progress. According to this approach, poorer regions grow faster (absolute convergence) in case of decreasing capital accumulation while keeping other factors constant. Because of its assumption of exogenous technology progress though, the origin of technological advances remain unexplained and it cannot consider human capital and innovation as endogenous growth drivers either [R19, R20, R21].

Addressing these limitations, the endogenous grwoth models developed in the 1980s aimed to incorporate factors that can drive internal growth like human capital, innovation or knowledge spillover. By acknowledging such predictors of economic growth an alternative to the assumption of deminishing returns of the neoclassical model could be provided. This approach suggests that regional divergence is also depending on human capital and R&D activities, furthermore regions investing in these areas can produce faster growth and may not necessarily converge with others. Although the model explains divergence by emphasizing the role of education and R&D investments, it can overestimate their role in regions where basic capital constrains are more important. Additionally, the required data to perform the relatively complex calculations – especially in regions with less data abundance – makes its applicability limited [R22, R23].

Conditional convergence models try to further elaborate on the similarities of regions when calculating growth trajectories. Here, not just previously listed factors then complete structural characteristics (e.g.: saving rates, technology, innovation activities, population growth, etc.) of regions shall match as a neccessary condition for convergence and the assumed convergence can only take place among comparable groups. By taking into consideration a broad palette of possible structural differences among regions, the conditional convergence model offers a framework for the explanation of the phenomenon why certain areas converge while others do not when conditions like infrastructure, education or policy do not match. Even though this model is more realistic than the absolute convergence, it requires extensive data on structural factors that might influence regional steady states of convergence making it hard to apply to heterogeneous regions with high variability of such conditions [R19, R21].

The problem of diverging regions to different levels led to the so called club convergence investigations that were often supported by practical observations and empirical findings. The theory suggests that regions, even with similar initial conditions can tend to form diverse ”clubs” that converge to a common income level. These models offer a flexible framework that can incorporate various initial conditions and provide an opportunity to policymakers to have an insight into specific convergence clubs and consequently help them tailor their policies groupwise. On the other hand, such models require the identification of groups and corresponding threshold analysis in connection to them. Moreover, it can also oversimplify convergence by assuming that group members will remain in their clubs and no between-group changes will take place later [R24, R25, R26].

The dynamic approach of convergence is covered by stochastic convergence models in more detail where time-series methods are utilized to decide whether income disparities among regions are temporary or permanent. The concept suggest that income levels should converge in the long-run in case disparities can be treated as stationary. These models can provide insights into temporal stability of convergence and can enable the testing whether regions return to common income paths after economic shocks or not. Albeit stochastic convergence lets regional-level resilience in income levels to be investigated, long-term data availability is requested. Furthermore, it still struggles to explain which structural factors might be relevant for differences when observing economic divergence [R26, R27].

Several scientific investigations regarding economic convergence are built upon the usage of the absolute $\beta$-convergence theory, that assumes negative correlation between

initial income levels and average growth rates of the selected financial indicators. This theory leads to the conclusion that regions with weaker initial conditions shall grow faster due to the free movement of capital to locations where it can be invested with lower costs. According to this concept, weaker economies with higher growth rates in the long-run shall produce a decreasing gap among the subgroups of richer- and poorer regions [R19, R20]. Although besides the reviewed convergence theories other economic approaches exist, due to the abundant presence of absolute $\beta$-convergence based investigations of economic growth and -convergence in the screened literature and out of the methodological purposes of the current chapter, I will connect to authors building upon this convergence concept in order to be able to produce comparable results in an interpretable and reproducible way.

The fundamental equation for economic absolute $\beta$-convergence is:

$$\frac{1}{T} \cdot ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \cdot ln(y_{i,0}) + \epsilon_i, \tag{3.1}$$

where $y_{i,T}$ and $y_{i,0}$ are the per capita economic measures for the $i$-th sample in the end and in the beginning of the investigated time interval correspondingly with length $T$ given in years, $\epsilon_i$ the error term and $\alpha, \beta$ are the intercept and slope parameters for the line in the linear regression problem [R17].

Nevertheless, in spite of empirical evidences of linear regression based results with negative slopes on the existence of convergence among regions and countries, more authors pointed out that the theory shall be used if certain boundary conditions are fulfilled and warn against drawing quick conclusions from the application of the theory. Such conditions that have to be mentioned are:

- The equation of absolute $\beta$-convergence can only be applied in the vicinity of an "equilibrium growth", therefore the theory assumes that countries' individual growth paths are close to their own equilibrium growth paths.
- The model calculates with per capita income indicators, therefore it has to be assumed that the growth of population and workforce are in the same ratio within the investigated time interval.
- The theory can only be applied when the investigated economies are in a growing phase. There is no evidence that it is still applicable for shrinking economies.
- Only in case of homogeneous economies with identical structural parameters (e.g.: technological progress, labour development ratio, savings rates, depreciation of physical capital etc.), where just the initial state differs may it be valid that these economies are converging towards the same equilibrium over time and consequently "the poor catches up with the rich" measured in per capita indicators.

Since countries and regions cannot be regarded as closed economies and there are various disturbances, the model can unfortunately be applied on heterogeneous economies only, where it has technically no predictive power. It can even occur that stronger economies have higher growth rates and differences increase in time. All that can be certainly stated that regions with higher growth rates are further away from their individual equilibrium states [R28, R17].

Although there are many factors influencing economic growth (e.g.: workforce migration, technology transfer among regions, wars and economic crises etc.) there have been numerous analyses performed that proved that in many situations initially poorer regions tend to show higher economic growth rates measured in per capita income indicators concerned [R19, R29, R20].

On the other hand, there is a large segment of publications that criticize results gained based on the absolute $\beta$-convergence concept and there are mathematical and practical

facts as well that prove the insufficiency of the theorem. Although, alternatives like $\sigma$-convergence are also applied simultaneously, techniques that utilize parametric approaches are often questioned in literature on their creditworthiness [R28, R30, R31, R32, R33].

Having this said, corresponding literature is apparently divided regarding the applicability of the convergence theory and finds its assumptions too restrictive with regard to real-life conditions. Divergent conclusions of cohesion tendencies are also characteristic for corresponding literature focusing on the regions and countries of the European Community. On the other hand, in case of many regional and country-level financial data analyses a negative linear relationship does exist between initial levels and growth rates of the same per capita financial indicators, that support the validity of the concept.

Nonetheless, differing conclusions can be found in literature regarding economic growth that concerns economic convergence within the European Union (EU). Several authors observed convergence of a certain extent [R30, R34, R31], while others found clear evidence for divergence [R35, R25, R36, R33], while still others interpreted convergence differently and detected signs of conditional- or even "club convergence" (tendency towards multiple clusters) of the regions that might speed up or lose momentum in certain time intervals [R24, R37, R38, R39]. However, the direct comparison of the various studies is circumstantial due to the different time periods, range of incorporated regions, applied methods and involved financial indicators investigated [R30].

It also has to be noted that there might be several aspects of economic convergence defined and accordingly various approaches have been utilized in literature. Besides parametric statistical methods, non-parametric alternatives –e.g.: stochastic dominance techniques – have been used as well that considers the whole distribution of data and enables an insight into the time dependent intra-distribution mobility of the regions [R31, R36].

From the methodological perspective, the widespread and conventional usage of parametric techniques may contribute to the high variability of conclusions listed in literature. Since these approaches utilize location- and scale parameters of annual distributions that are outlier-sensitive – namely mean and standard deviation values – and are therefore non-robust, the non-normal characteristic of underlying data distributions can significantly distort them and bias conclusions of the economic theory that is based on them. Violations to general assumptions of conventional statistical procedures inevitably lead to model misspecifications and erroneous results that are generally to be avoided.

Additionally, application of parametric statistical methods in the present case faces problems with the sampling procedure as well. In case of regional investigations there are no random samples, since technically the whole data-population can be acquired and the sample members cannot be regarded as independent either, since there are deep interconnections among them if nothing else on a spatial basis (e.g.: relations among neighbouring regions) [R40, R41, R32, R25].

Regression problems – to which the empirical investigation of Eq. 3.1. also belongs – constitute one of the core fundamental elements of statistical learning procedures. Nevertheless, real-life data contaminated by a-typical elements and of skewed non-normal distributions can pose challenges in proper model building and parameter estimation. In case of multidimensional investigations, complications caused by such anomalies are even harder to detect and handle, notwithstanding various techniques are known in order to keep unwanted deviations under control [R42, R43].

Robust statistical procedures aim to address negative influences caused by outliers and data non-normality by using proper weighting of a-typical observations that are deviating greatly from the "bulk" of the data. In practice, this leads to a trade-off between the maximization of statistical efficiency and the level of breakdown point [R44, R45, R43, R46].

Although, parametric models often rely on strict assumptions hard to hold in practice,

they can serve with estimates more accurate and easier to interpret when specific conditions are met. In contrast, non-parametric models can serve with less sharp estimates or results may be harder to interpret, but the range of applicability is wider. As a middle road, robust statistical approaches assume small deviations from expected distributions and models. They aim to operate in the "neighbourhood" of the theoretical assumptions (e.g.: Gaussian error distribution) and can be regarded as an extension of parametric statistical procedures [R47]. Robust procedures utilize the fact, that parametric models are also just an approximation of reality and offer approximate parametric models while maintaining a compromise between strict parametric models and potentially hard to interpret non-parametric models [R47]. Robust regression techniques are aiming to reduce negative influences caused by outliers by assigning appropriate weights to a-typical observations that deviate greatly from the bulk of the data.

In case of a proper weighting, data does not have to be trimmed and information to be discarded, since the effect of atypical data on the estimates is about to be decreased. Nevertheless, as a trade-off, mostly iterative procedures can provide the estimates sought. Therefore, robust methods often require more computation time than conventional statistical procedures. The selection of numerical algorithms, the way of implementation and proper initialization can also play an essential role for achieving converging results, while maintaining satisfactory statistical efficiency (of course within certain limits) [R43].

## 3.2 Robust Linear Regression

Considering the original iterative equation system of Eq. 2.7. and Eq. 2.8. for the MFV- and dihesion values of a data distribution provided in Chapter 2. we have the following formulas for the converged values:

$$M(k,x) = \frac{\sum_{i=1}^{n} \frac{(k\cdot\epsilon(x))^2}{(k\cdot\epsilon(x))^2 + (x_i - M(k,x))^2} \cdot x_i}{\sum_{i=1}^{n} \frac{(k\cdot\epsilon(x))^2}{(k\cdot\epsilon(x))^2 + (x_i - M(k,x))^2}} \tag{3.2}$$

$$\epsilon(x)^2 = \frac{3\cdot\sum_{i=1}^{n} \frac{(x_i - M(k,x))^2}{(\epsilon(x)^2 + (x_i - M(k,x))^2)^2}}{\sum_{i=1}^{n} \frac{1}{(\epsilon(x)^2 + (x_i - M(k,x))^2)^2}}. \tag{3.3}$$

Since the $M(k,x)$ in Eq. 3.2. is technically a weighted average, – where the weights are given in a form of a Cauchy-distribution – the far-lying points measured from the location of data-concentration are taken into consideration with a less weight that results in an increased outlier resistance. The $k$ parameter is a constant that shall be chosen according to the underlying data distribution to be investigated, – which in practice is only rarely known *a priori*. Thus, based on results of Monte Carlo simulations the selection of $k = 2$ is recommended for an optimal statistical efficiency – while $\epsilon(x)$ denotes the scale parameter [R48, R6].

According to the results of further extensive Monte Carlo simulations, the MFV method can be applied on real-life data contaminated with outliers in a wide range of distributions "far" from being Gaussian at a much higher statistical efficiency than conventional statistical procedures that are mainly based on the minimization of the $L2$-norm [R14, R49].

After reordering, Eq. 3.2. and multiplying both sides with the denominator of the left side (which is definitely greater than zero) and multiplying both sides with $2/(k\cdot\epsilon(x))^2$ as well the following equivalent formula can be obtained that is identically zero:

$$g(\epsilon(x), M(k,x)) = 2 \cdot \sum_{i=1}^{n} \frac{M(k,x) - x_i}{(M(k,x) - x_i)^2 + (k \cdot \epsilon(x))^2} = 0 \tag{3.4}$$

From Eq. 3.4. it can be seen that an equivalent task is finding the extremum value given in Eq. 3.5. with respect to $M(k,x)$, where due to convexity considerations, a minimum place is sought:

$$G(\epsilon(x), M(k,x)) = \sum_{i=1}^{n} ln\left[(M(k,x) - x_i)^2 + (k \cdot \epsilon(x))^2\right] \tag{3.5}$$

Therefore, finding the roots of $g(\epsilon(x), M(k,x))$ – that is equivalent to estimate the MFV value – is possible by solving the $G'_M(\epsilon(x), M(k,x)) = 0$ instead, which technically transforms the task to a minimization problem with respect to the $M(k,x)$ value and enables the generalization of the definition of the *Most Frequent Value* method for higher dimensional cases.

Consequently, in the present case instead of minimizing the $\sum_{i=1}^{n}(E(x) - x_i)^2 = min.$ expression with respect to $E(x)$ as done for least squares procedures (where $E(x)$ denotes the expected value of the distribution) the $\sum_{i=1}^{n} ln\left[(M(k,x) - x_i)^2 + (k \cdot \epsilon(x))^2\right] = min.$ expression with respect to $M(k,x)$ has to be fulfilled in order to gain the location parameter of interest. Albeit, the theory promises to serve with highly outlier resistant and on wide distribution-range robust location- and scale parameters, the calculations have to be done in an iterative way.

For a two-dimensional case of $(x_i, y_i), i \in [1,n]$ observations, for the better understanding let us stick to conventional notation and refer to the $M(k,x)$ location parameter estimate in a form of $T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}}) = a \cdot x_i + b$, where $\underline{\mathbf{p}} = [a,b]^T$ is a parameter vector containing the slope and intercept values of the regression line and let simply $\epsilon$ denote the dihesion. Hence, Eq. 3.5. can be written up in the following form [R46]:

$$\sum_{i=1}^{n} ln\left[(T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}}) - y_i)^2 + (k\epsilon)^2\right] = min. \tag{3.6}$$

To fulfill the minimum criterion with respect to $T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}})$ the equation system of the partial derivatives set to zero ($r = 1, 2$) has to be solved:

$$\frac{d}{dp_r} \sum_{i=1}^{n} ln\left[(T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}}) - y_i)^2 + (k\epsilon)^2\right] = 0 \tag{3.7}$$

Having performed the designated operations:

$$\sum_{i=1}^{n} \frac{2 \cdot (T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}}) - y_i)}{(T_i(\underline{\mathbf{p}}, \underline{\mathbf{x}}) - y_i)^2 + (k\epsilon)^2} \cdot \frac{dT_i(\underline{\mathbf{p}}, \underline{\mathbf{x}})}{dp_r} = 0 \tag{3.8}$$

After substituting the expression for the MFV value and performing the necessary calculations, the equation system given by Eq. 3.9. and 3.10. can be obtained. These formulas still contain the $\epsilon$ dihesion parameter, therefore they have to be extended by Eq. 3.3. in order to find the corresponding regression line parameters.

$$\sum_{i=1}^{n} \frac{1}{(k\epsilon)^2 + (ax_i + b - y_i)^2} \cdot [(ax_i + b - y_i) \cdot x_i] = 0 \tag{3.9}$$

$$\sum_{i=1}^{n} \frac{1}{(k\epsilon)^2 + (ax_i + b - y_i)^2} \cdot [ax_i + b - y_i] = 0 \tag{3.10}$$

For the initialization we use the solution of the OLS regression as suggested by Steiner et al. Furthermore, from the linear equation system of the OLS regression it can also be seen that the equation system presented by Eq. 3.9. and Eq. 3.10. for the MFV-robustified linear regression is illustratively a weighted form of the latter and in this aspect could be declared as a type of partial least squares regression models [R43]:

The initialization vector for the "MFV iteration" from the OLS regression solution in matrix notation can be given as:

$$\begin{bmatrix} a_0 \\ \\ b_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \tag{3.11}$$

Since Eq. 3.9. and 3.10. cannot be solved directly even for given $\epsilon$ values, a nonlinear equation system has to be handled in every iteration step of the "MFV iteration". Fortunately the partial derivatives of the Jacobian matrix can be expressed analytically and for the generalized Newton's algorithm the following 2D equation system can be written up for the $(m+1)$-th iteration step:

$$\begin{bmatrix} a^{(m+1)} \\ \\ b^{(m+1)} \end{bmatrix} = \begin{bmatrix} a^{(m)} \\ \\ b^{(m)} \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} x_i^2 A_i^{(m)} & \sum_{i=1}^{n} x_i A_i^{(m)} \\ \sum_{i=1}^{n} x_i A_i^{(m)} & \sum_{i=1}^{n} A_i^{(m)} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} x_i B_i^{(m)} \\ \sum_{i=1}^{n} B_i^{(m)} \end{bmatrix} \tag{3.12}$$

Where the values denoted by capital letters can be provided in the $(m)$-th step of the iteration as:

$$A_i^{(m)} = \frac{(k\epsilon^{(m)})^2 - (y_i - a^{(m)}x_i - b^{(m)})^2}{[(k\epsilon^{(m)})^2 + (y_i - a^{(m)}x_i - b^{(m)})^2]^2} \tag{3.13}$$

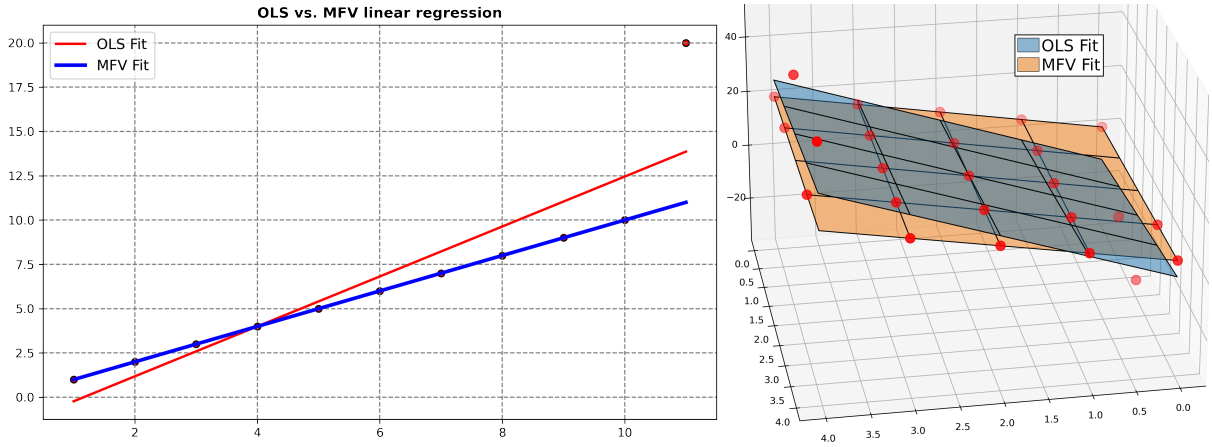
$$B_i^{(m)} = -\frac{(y_i - a^{(m)}x_i - b^{(m)})}{(k\epsilon^{(m)})^2 + (y_i - a^{(m)}x_i - b^{(m)})^2} \tag{3.14}$$

The algorithm for obtaining the parameters for the MFV-robustified linear regression can be summarised as follows:

1. Initialize $a_0$ and $b_0$ parameters from OLS linear regression.
2. Initialize dihesion with the maximal residuals from the OLS-line as $\epsilon_0 = max(r_i^+) - max(r_i^-)$
3. "Inner iteration": Solve nonlinear equation system given in Eq. 3.12. by generalized Newton's or Broyden's method. (Stop condition: $max(a^{(k+1)} - a^{(k)}, b^{(k+1)} - b^{(k)}) \leq 10^{-5}$).
4. "MFV-iteration": Update dihesion parameter using the calculated $a^{(k)}$ and $b^{(k)}$ regression parameters in accordance with Eq. 3.3. (Stop condition: $\epsilon^{(k+1)} - \epsilon^{(k)} \leq 10^{-5}$).

Code samples for the implementation of the MFV-based linear regression can be seen in Appendix A. As an illustration for the difference between OLS and MFV-robustified linear regression, example datasets have been generated in 2D and 3D cases (see Fig. 3.1.). In the 2D case, 10 data points were perfectly aligned on a straight line, while the 11-th data point has been given as an outlier. As Fig. 3.1a. shows the OLS line fits the points to minimize the overall sum of squares of the deviations, while the MFV-robustified line remains on the "bulk" of the data. The same advantageous behaviour is true for the 3D case with

vertical outliers. It has to be noted though that bad leverage points (outlier observations in ”horizontal directions”) can have many severe effects on the regression surface, therefore resistivity is much less against such observations.




**(a)** **(b)**

**Figure 3.1:** Demonstration of the difference between OLS-based and MFV-robustified linear regression in the presence of vertical outliers. The latter fits the ”bulk” of the data more efficient.

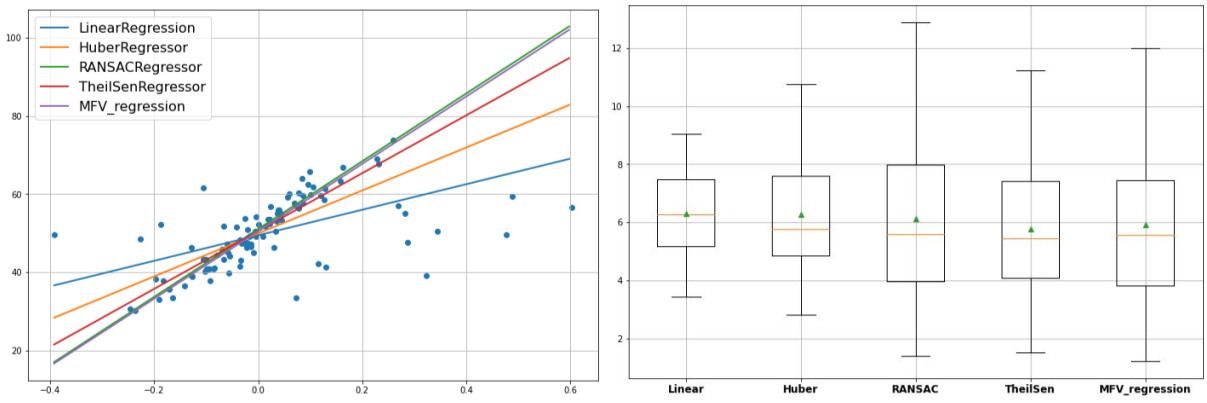
Although it is important to have information on the break-down point of the outlined algorithm, deeper theoretical considerations would be necessary. By further increasing the number of points to $100$ and experimenting with the amount of points modified to outliers, a practical value for the break-down point could be estimated, which in this example was 23%. Unfortunately, this statement is valid only for vertical outliers, since based on our experience bad leverage points can have a more drastic influence on the MFV-robustified regression line similar to other robust statistical techniques [R43, R50].

A further illustration of the MFV-based linear regression can be gained by the application of the method alongside with other robust linear regression techniques. Fig. 3.2a. shows a sample dataset of 100 data points equipped with some statistical noise that is aligned to a line. The data sample is then contaminated with 15 outliers that are horizontally placed in a distance of 2–4 times of the standard deviation of the independent variable. Via three times repeated 10-fold cross validation, the contaminated data set was split, and corresponding regression algorithms were applied on them according to Fig. 3.2a., where the figure illustrates the best fits per method with regard to the mean absolute error (MAE). Fig. 3.2b. represents the gained MAE distributions summarised on box-plots for the selected robust regression methods of the algorithms which were implemented in the *sklearn* python package[1] together with the MFV-based 2D linear regression algorithm.

Depending on randomization and initialization, the mean values and corresponding standard deviations (in brackets) of the resulted box-plots of Fig. 3.2b. were for the OLS-based linear regression $6.310(1.421)$, for Huber regression $6.263(1.980)$, for RANSAC[2] $6.124(3.188)$, for TheilSen $5.763(2.627)$ and $5.918(3.003)$ for the MFV-based regression. The mean of the mean absolute errors for MFV-based regression is resulted to be similar, albeit a bit higher, as for the TheilSen regression and the best fit line provided approximately the same result as the best fit line for the RANSAC regression. From further empirical analysis by increasing the number of bad leverage points, the MFV-based regression had a breakdown point at 20 added contamination point. Thus, it provided with a better fit – in this specific case

[1]For the python implementation of other robust liner regression algorithms, the code snippets provided in https://machinelearningmastery.com/robust-regression-for-machine-learning-in-python/ (accessed: 2023.09.11.) were utilized.

[2]Random Sample Consensus




(a) (b)

**Figure 3.2:** Illustration of fitting behaviour of MFV-based linear regression alongside with other robust regression methods in case of a generated data sample by the investigation of mean absolute errors gained with 3-times repeated 10-fold cross validation.

– to the bulk of the data than Huber- or TheilSen regressions, but regarding breakdown properties it could only over-perform the Huber regression.

## 3.3 Data Analysis

For application purposes, as a primary data source the Gross Domestic Product (GDP) and Net Disposable Income (NDI) data of the EU countries and regions have been analysed within the time period of 2000-2020 regarding economic convergence. The exact time intervals and limitations regarding missing data are listed in Table 3.1. The data were accessed at *Eurostat* that is the responsible institution within the European Community for the dissemination and harmonization of statistical information [R51]. As can be seen in Table 3.1. there were no accessible information regarding GDP for France before 2015 and no NDI data for Malta at all that can result in some distortion for any kind of further statistical investigation. However, the slightly differing time intervals posed no difficulties to the applied analysis and the amount of missing data was also marginal within the listed time periods, therefore could not have significant influence on our findings either.

| Economic Indicator | Dimension | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|
| GDP | EUR per inhabitant | 2009-2020 | 2000-2019 France: 2015-2019 | 2000-2018 France: 2015-2018 |
| | PPS per inhabitant | 2009-2020 | 2008-2019 France: 2015-2019 | 2000-2018 France: 2015-2018 |
| NDI | EUR per inhabitant | 2000-2018 Without Malta | 2000-2018 Without Malta | No data |
| | PPS per inhabitant | 2000-2018 Without Malta | 2000-2018 Without Malta | No data |

**Table 3.1:** GDP and NDI per capita income measures for different regional levels accessed at *Eurostat*.

Corresponding literature often builds both upon GDP and NDI measures when characterizing regional convergence. While GDP can be regarded as a measure of market value

of all the goods and services produced, NDI represents the income of the population after taxes. Thus, GDP can be taken as a metric of economic progress and can be used to compare development of economies, whereas NDI as a metric of the standard of living [R52]. The outlined financial indicators were obtained in *EUR per inhabitant* and *PPS per inhabitant* dimensions that are conventional for convergence investigations. The latter dimension expresses the per inhabitant financial indicator with respect to the average within the European Union that is set to be 100. Representing data in PPS can reduce differences caused by various price levels and enable a better comparison among EU member countries [R51].

Although GDP is most probably the most commonly used growth measure in related literature, its limitations has to be addressed regarding indicating societal well-being or even long-term sustainability. As its primary focus lies on market transactions it does not consider other essential aspects of quality of life like healthcare, education or life satisfaction in general. Therefore, GDP can show growth tendencies while social welfare decline is present. This can make GDP an incomplete proxy for human progress [R53].

GDP also does not account for environmental degradation and resource depletion, does not consider long-term economic potential than rather fouses on short-term economic activity. It can depict an over-optimistic illusion of growth that might lead to unsustainable environmental practices. On the other hand, it also ignores non-market contributions (e.g.: domestic work, unpaid caregiving etc.) that veils significant economic activities and leads to an underestimation of actual economic contributions [R54].

Increasing GDP does not necessarily mean that amenities are shared among citizens in a fair manner. High GDP per capita can also lead to elevated levels of inequality that might cause some layers of society benefit unequally and more modest compared to those of distinguished positions [R55].

Due to the above limitations and potential drawbacks of GDP usage, other indicators are advised in literature that might provide a more comprehensive picture of economic growth and convergence. A non-exhaustive list of such indicators are:

- Total Factor Productivity (TFP), that can measure all inputs in the production process and can can capture contribution of technological innovation and skills. Nevertheless, it might be complex to measure accurately and external factors (e.g.: regulatory changes) can make it difficult to interpret consistently across countries. Furthermore, it also does not consider environmental related resource depletion and citizen well-being [R56].

- Gross National Income (GNI) reflects extra contributions of net income from abroad, therby highlights the role of foreign investments. As a drawback however, it can be influenced by multinational corporations' tax strategies and still neglects environmental and social welfare related issues [R57].

- Human Development Index (HDI) combines life expectancy, education and per cpita income to provide a comprehensive view of growth beyond economic output alone. Albeit serving with a more holistic picture with drivers of long-term development, HDI is still a simplification of complex societal factors by arbitrarily weighting them. This can fade differences among regions with highly differing education and healthcare profiles [R54, R58].

- Genuine Progress Indicator (GPI) takes positive and negative aspects of economic growth (income distribution, environmental cost, value of non-market activities etc.) into consideration in order to examine whether they have benefited citizens or not. Nonetheless, because of its complex calculation and subjective interpretation the comparability across regions and over time can be problematic [R54].

- Green GDP considers environmental costs by substracting them from GDP, thereby strives to serve with a measure of economic growth that can be connected to objectives of sustainability. Unfortunatelly, data shortages and opaque definitions of the exact measurement of environmental impacts over countries and regions can make its usage hardly comparable in calculations [R59].
- Gini coefficient captures the income inequality within regions that can help to understand distributional aspects of economic growth. However, growth progresses can have varying effects on income distribution causing an inverted U-shaped relationship that might necessitate more in-depth analysis of time dynamics of regional economic development [R54].
- Consumption per capita can be used to analyse effects of economic growth on individual well-being by calculating average consumption of goods and services. It can help to better understand how growth can be mapped to the improvements of quality of life, but it lets uneven consumption among the poor and rich be veiled and does not account for sustainability considerations either [R60].

Besides acknowledging the above mentioned deficiencies of GDP utilization, with my data collection and analysis in the following I will connect to those authors who present results applying these data. Since most of the reviewed literature build upon these indicators – even the latest and most fresh ones – from methodological point of view in my investigations I stay aligned with the mainstream in order to keep interpretability and results comparable.

Having finished the data collection process the number of available data points were 27 on country level, 213 on NUTS2 level and 1066 on NUTS3 level that might slightly vary according to the limitations listed in Table 3.1. The financial indicators of interest followed highly skewed annual distributions to the left. Representing annual distributions on box plots according to Tukey's Fences that – as a non-parametric outlier detection method – marks points as outliers when they are beyond the 1st and 3rd quartiles more than 1.5·IQR[3]. It can be seen that the 1D distributions may contain several outlier suspicious items (see Fig. 3.3).

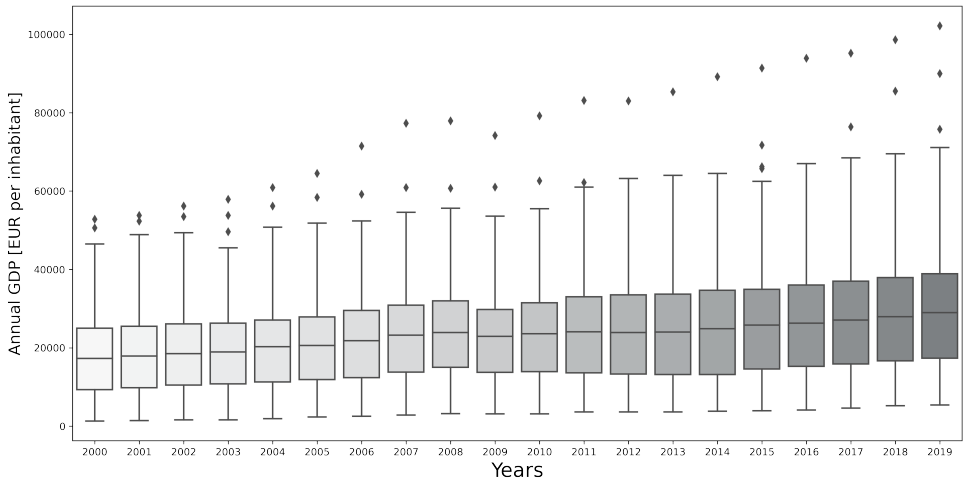



**Figure 3.3:** Annual distribution of GDP per capita values measured in EUR in case of NUTS2 regions.

Although it has to be noted, that in case of data from economic origin it cannot be unequivocally stated whether a data point is outlier since they cannot be attributed to any

[3]Interquartile Range

kind of measurement error or being a member of other data-populations. Furthermore, our data cannot be treated as a random sample of a larger data-population either. Furthermore, since we possess the whole data-population the usage of statistical error estimations (e.g.: confidence intervals) should be treated with reservations. Therefore, application of conventional statistical procedures are arguable and robust- or non-parametric methods are to be used that can increase the amount of statistical information to be extracted out of the underlying sample and reducing the risk of biasing the resulting estimates [R6, R14].

| | **Country** | | | | **NUTS2** | | | | **NUTS3** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **GDP** | | **NDI** | | **GDP** | | **NDI** | | **GDP** | |
| | **EUR** | **PPS** | **EUR** | **PPS** | **EUR** | **PPS** | **EUR** | **PPS** | **EUR** | **PPS** |
| 2000 | - | - | 0.9123 (0.0298) | 0.9514 (0.2505) | 0.9580 (0.0000) | - | 0.9157 (0.0000) | 0.9676 (0.0000) | 0.9283 (0.0000) | 0.9171 (0.0000) |
| 2001 | - | - | 0.9122 (0.0297) | 0.9521 (0.2600) | 0.9609 (0.0000) | - | 0.9208 (0.0000) | 0.9631 (0.0000) | 0.9232 (0.0000) | 0.9071 (0.0000) |
| 2002 | - | - | 0.9068 (0.0223) | 0.9483 (0.2113) | 0.9635 (0.0000) | - | 0.9139 (0.0000) | 0.9568 (0.0000) | 0.9281 (0.0000) | 0.9105 (0.0000) |
| 2003 | - | - | 0.9039 (0.0192) | 0.9533 (0.2761) | 0.9641 (0.0000) | - | 0.9063 (0.0000) | 0.9599 (0.0000) | 0.9309 (0.0000) | 0.9085 (0.0000) |
| 2004 | - | - | 0.9035 (0.0188) | 0.9531 (0.2745) | 0.9641 (0.0000) | - | 0.9050 (0.0000) | 0.9610 (0.0000) | 0.9330 (0.0000) | 0.9049 (0.0000) |
| 2005 | - | - | 0.9088 (0.0247) | 0.9551 (0.3039) | 0.9638 (0.0000) | - | 0.9108 (0.0000) | 0.9600 (0.0000) | 0.9283 (0.0000) | 0.9005 (0.0000) |
| 2006 | - | - | 0.9084 (0.0242) | 0.9541 (0.2891) | 0.9644 (0.0000) | - | 0.9087 (0.0000) | 0.9592 (0.0000) | 0.9290 (0.0000) | 0.9016 (0.0000) |
| 2007 | - | - | 0.9098 (0.0261) | 0.9498 (0.2290) | 0.9637 (0.0000) | - | 0.9128 (0.0000) | 0.9632 (0.0000) | 0.9261 (0.0000) | 0.8971 (0.0000) |
| 2008 | - | - | 0.9194 (0.0436) | 0.9519 (0.2571) | 0.9661 (0.0001) | 0.9666 (0.0001) | 0.9199 (0.0000) | 0.9646 (0.0000) | 0.9272 (0.0000) | 0.8962 (0.0000) |
| 2009 | 0.8731 (0.0034) | 0.8294 (0.0005) | 0.9148 (0.0341) | 0.9537 (0.2827) | 0.9630 (0.0000) | 0.9646 (0.0000) | 0.9076 (0.0000) | 0.9629 (0.0000) | 0.9259 (0.0000) | 0.8937 (0.0000) |
| 2010 | 0.8643 (0.0022) | 0.8229 (0.0004) | 0.9137 (0.0321) | 0.9628 (0.4491) | 0.9628 (0.0000) | 0.9635 (0.0000) | 0.9139 (0.0000) | 0.9734 (0.0002) | 0.9192 (0.0000) | 0.8798 (0.0000) |
| 2011 | 0.8486 (0.0011) | 0.7928 (0.0001) | 0.9241 (0.0562) | 0.9657 (0.5158) | 0.9593 (0.0000) | 0.9609 (0.0000) | 0.9195 (0.0000) | 0.9774 (0.0007) | 0.9097 (0.0000) | 0.8641 (0.0000) |
| 2012 | 0.8489 (0.0011) | 0.7958 (0.0001) | 0.9255 (0.0605) | 0.9632 (0.4593) | 0.9567 (0.0000) | 0.9585 (0.0000) | 0.9260 (0.0000) | 0.9760 (0.0004) | 0.9078 (0.0000) | 0.8598 (0.0000) |
| 2013 | 0.8401 (0.0007) | 0.7925 (0.0001) | 0.9240 (0.0560) | 0.9599 (0.3902) | 0.9530 (0.0000) | 0.9550 (0.0000) | 0.9250 (0.0000) | 0.9739 (0.0002) | 0.9024 (0.0000) | 0.8529 (0.0000) |
| 2014 | 0.8308 (0.0005) | 0.7729 (0.0000) | 0.9180 (0.0405) | 0.9548 (0.2995) | 0.9495 (0.0000) | 0.9514 (0.0000) | 0.9235 (0.0000) | 0.9777 (0.0008) | 0.9020 (0.0000) | 0.8530 (0.0000) |
| 2015 | 0.8453 (0.0009) | 0.7939 (0.0001) | 0.9196 (0.0441) | 0.9600 (0.3910) | 0.9512 (0.0000) | 0.9389 (0.0000) | 0.9244 (0.0000) | 0.9783 (0.0010) | 0.8963 (0.0000) | 0.8460 (0.0000) |
| 2016 | 0.8434 (0.0009) | 0.7842 (0.0001) | 0.9206 (0.0464) | 0.9585 (0.3636) | 0.9521 (0.0000) | 0.9435 (0.0000) | 0.9246 (0.0000) | 0.9793 (0.0014) | 0.8709 (0.0000) | 0.8131 (0.0000) |
| 2017 | 0.8498 (0.0011) | 0.7916 (0.0001) | 0.9186 (0.0416) | 0.9498 (0.2293) | 0.9498 (0.0000) | 0.9419 (0.0000) | 0.9258 (0.0000) | 0.9794 (0.0015) | 0.8779 (0.0000) | 0.8227 (0.0000) |
| 2018 | 0.8427 (0.0008) | 0.7879 (0.0001) | 0.9198 (0.0444) | 0.9477 (0.2043) | 0.9445 (0.0000) | 0.9354 (0.0000) | 0.9285 (0.0000) | 0.9820 (0.0039) | 0.8705 (0.0000) | 0.8148 (0.0000) |
| 2019 | 0.8334 (0.0005) | 0.7784 (0.0001) | - | - | 0.9389 (0.0000) | 0.9303 (0.0000) | - | - | - | - |
| 2020 | 0.8162 (0.0003) | 0.7604 (0.0000) | - | - | - | - | - | - | - | - |

**Table 3.2:** Results of Shapiro-Wilk tests with corresponding p-values in brackets.

The normality assumption of the data was tested by Shapiro-Wilk tests. In most of the cases, the test rejected with high significance that the investigated data were normally distributed. Only in case of NDI indicators on country level could be seen that the Shapiro-Wilk test could not reject the normality assumption in every case. NDI per capita measured in EUR in some years could be regarded as normal on 99% confidence level but could not be regarded as normal on 95%, while measured in PPS the test did not reject normality even on 95% confidence level (see Table 3.2).

## 3.4 Application of the MFV-based Robust Linear Regression

In the following, I approximate the growth rates given in the fundamental equation of absolute $\beta$-convergence with its first order Taylor expansion that is identical to the average of annual growth rate. With this step, my aim is to add more information to the time evolution of growth rates into subsequent calculations. The transformation of the left side of Eq. 3.1. can be done as:

$$\begin{aligned} ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = ln\left(\prod_{t=0}^{T}\frac{y_{i,t+1}}{y_{i,t}}\right) = \sum_{t=0}^{T} ln\left(\frac{y_{i,t+1}}{y_{i,t}}\right) \approx \\ \sum_{t=0}^{T}\left(\frac{y_{i,t+1}}{y_{i,t}} - 1\right) = \sum_{t=0}^{T}\left(\frac{y_{i,t+1} - y_{i,t}}{y_{i,t}}\right), \end{aligned} \tag{3.15}$$

where we took advantage of $ln(1+x) \approx x$ if $x$ is close to zero. This latter holds in the present case, since the annual growth rates of the EU countries and NUTS regions are generally less than 10%. By dividing the resulted formula with the length of the time interval, the average annual growth rate is obtained for each economy.

Putting together, Eq. 3.1. and Eq. 3.15. a linear regression problem has to be solved with more information on the annual changes of each economy on the left. However, it is well known that averages are highly sensitive to the presence of outliers and statistical procedures using averages as empirical means are prone to serve with fallacious results when the underlying data distribution is not normal or outliers may ”contaminate” our data [R49].

As a first step for checking relationships among the generated variables according to Eq. 3.1. and 3.15. correlation coefficients were calculated. This was done based on Pearson's formula that is well-known to be prone to outliers and non-normality:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.16}$$

This formula indicates the strength of linear relationship between two random variables. If they are normally distributed, it even indicates independence if the calculated value is zero. However, the presence of outliers can cause a dramatic bias in this formula and longer-tailed distributions than the Gaussian tend to increase correlation values even in case of trivially non-existing relationships [R14].

Besides Pearson's formula that serves with a quantitative result, Spearman's rho and Kendall's tau provides well-known indicators on the strength of (even nonlinear) relationships between random variables. These measures provide rather with a qualitative than quantitative result, where ranks or concordance values of data are calculated. According to test cases, however, the significance of these coefficients can be questioned occasionally, and the calculated results should seriously be taken into consideration only when they are close to $\pm 1$ [R61].

A practical modification of Pearson's formula can be achieved by the combination of Eq. 3.16. with the MFV methodology. According to the results of Steiner et al. provided in [R14] a practically proven ”robustification” of the conventional correlation coefficient can be achieved by substituting the MFV values instead of arithmetic means and applying the

weighting function as given below:

$$r_{xy} = \frac{\sum_{i=1}^{n} \Big[w(x)(x_i - M(k,x))\Big] \cdot \Big[w(y)(y_i - M(k,y))\Big]}{\sqrt{\sum_{i=1}^{n} w^2(x) \cdot (x_i - M(k,x))^2} \cdot \sqrt{\sum_{i=1}^{n} w^2(y) \cdot (y_i - M(k,y))^2}} \tag{3.17}$$

This "MFV-robustified" correlation coefficient is not just more robust against deviations from the normal distribution of the variables but even suppresses the distorting effects of outliers, therefore it is more resistant as well. The formula of Eq. 3.17. serves together with Eq. 3.2. and Eq. 3.3. a well-defined practical procedure that is relatively easy to implement for calculating robustified correlations.

In order to highlight the distorting effect of outliers within the data besides investigating the annual growth rates of corresponding income measures their MFV values were also calculated and were considered on the left side of Eq. 3.1. These MFVs calculated provide higher outlier resistance for each instance of the data set. Consequently, robust and resistant location parameters are obtained that represent the time evolution of annual development more and are more representative for the "bulk" of the data in the presence of outliers and distributions with long-tails or of non-Gaussian error distributions. For a visual comparison of the different measures of growth rates, Fig. 3.4. represents the calculated measures on the same graph for the analysis of economic absolute $\beta$-convergence in case of GDP per capita on the NUTS2 region level.

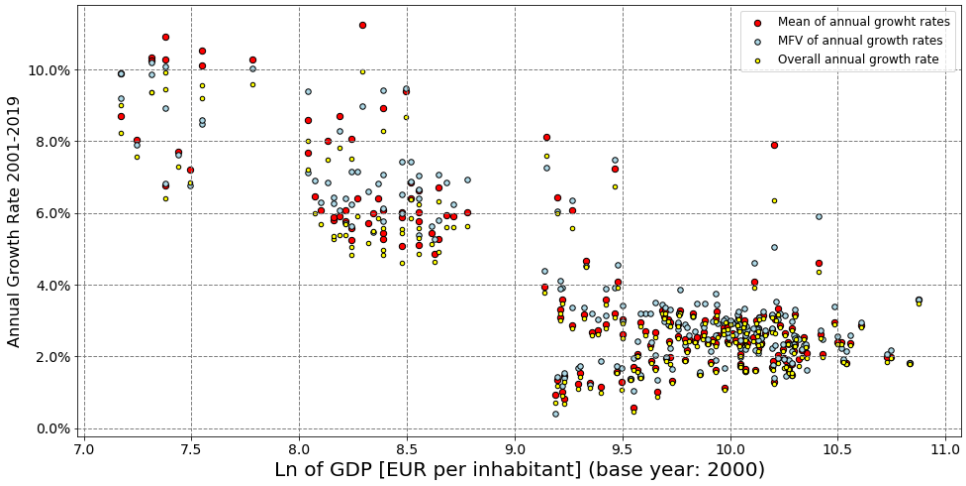



**Figure 3.4:** Different measures of growth rates within the investigated time period for NUTS2 regions regarding GDP [EUR per inhabitant].

In order to get a comprehensive picture on the differences between the original relationship described by Eq. 3.15. and the altered versions with mean- and MFV values on the left, average values of relative changes have been calculated among "overall annual growth rates" and means- or MFVs of annual growth rates for every instance in each dataset. According to Table 3.3. the average of relative changes in case of substituting mean values on the left ($\%_{Mean}$) remain below 10% in all the cases, while substituting MFVs ($\%_{MFV}$) result average relative changes even higher than 55% but at least 16% for all datasets at hand. This latter draws attention to great deviations within the data that are generally masked by "overall" or "average" growth measures that might not represent the typical annual growth, which corresponds to the "bulk" of each data distribution.

Furthermore, by generating *Cullen and Frey* graphs (see Fig. 3.5.) in R utilizing the *fitdistrplus* package for some of the data series with 1000 bootstrap samples, it was revealed

| | Country | | | | NUTS2 | | | | NUTS3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GDP | | NDI | | GDP | | NDI | | GDP | |
| | EUR | PPS | EUR | PPS | EUR | PPS | EUR | PPS | EUR | PPS |
| $\%_{Mean}$ | 8,68 | 5,31 | 5,60 | 4,34 | 5,60 | 6,04 | 6,25 | 4,97 | 6,86 | 6,44 |
| $\%_{MFV}$ | 55,33 | 49,44 | 17,28 | 16,70 | 16,79 | 44,59 | 16,60 | 19,62 | 17,92 | 21,11 |

**Table 3.3:** Averages of relative changes in percentage among overall annual growth rates and means- ($\%_{Mean}$) and MFVs ($\%_{MFV}$) of annual growth rates for each dataset.

that the distribution of annual financial indicators and growth rates can be best described by Gamma distributions. Since skewness and kurtosis are not robust measures and as higher moments have bigger variances, the skewness-kurtosis plot can therefore be taken only as an indicative visualisation for narrowing the set of possible probability distributions that might best describe the data at hand [R62].

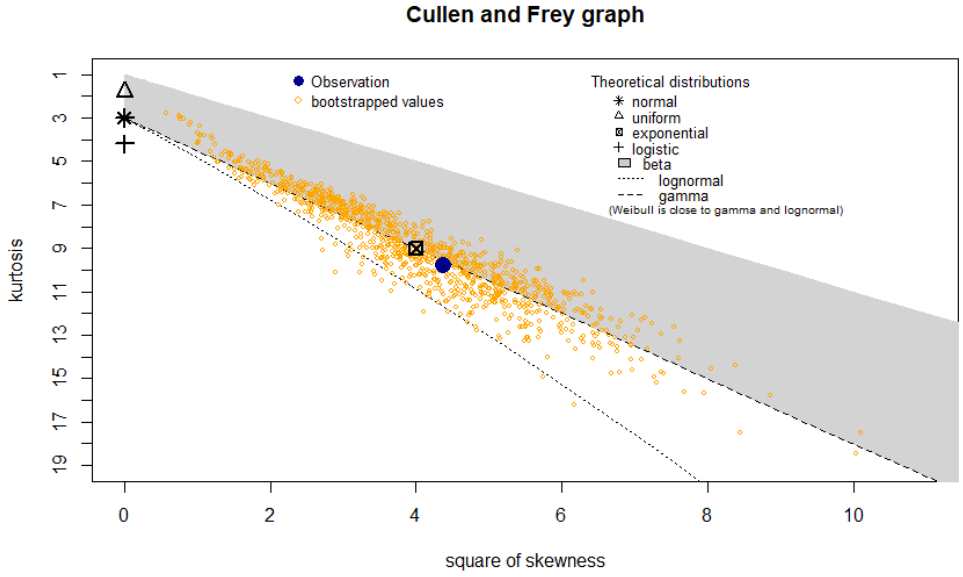



**Figure 3.5:** Cullen and Frey graph for GDP per inhabitant data of NUTS2 regions for the year 2004.

The visual comparison of average and MFV values makes it clear that throughout the almost 20 years of observation, even dramatic differences could take place. On the left of Fig. 3.4. MFV values tend to be smaller than the averages, while on the right they tend to be higher. This leads to weaker linear relationship in case of treating MFV values. On the other hand, Fig. 3.6. shows the calculated MFV values for the same data with an indication on the date of connection of the regions to the European Union.

By investigating Fig. 3.4. and Fig. 3.6. it is obvious that economic convergence according to the absolute $\beta$-convergence theorem – thus the decrease in the gap regarding the given economic measure – was not that outstanding as could have been calculated by utilizing outlier-sensitive average values.

Closely inspecting Fig. 3.4. even a data point (region of *Voreio Aigaio* in Greece) can be observed that had **negative** MFV of annual growth rate in the investigated time period, while in average there were no negative values and the dramatic problem might have been veiled within the framework of a conventional investigation.

The calculated Pearson's correlation coefficients and MFV-robustified correlation coefficients are summarised in Table 3.4. for average annual GDP and NDI growth rates and in Table 3.5. for the MFVs of annual GDP and NDI growth rates. Country, NUTS2 and NUTS3 levels were all considered. In each cell of the tables, the upper value corresponds to the

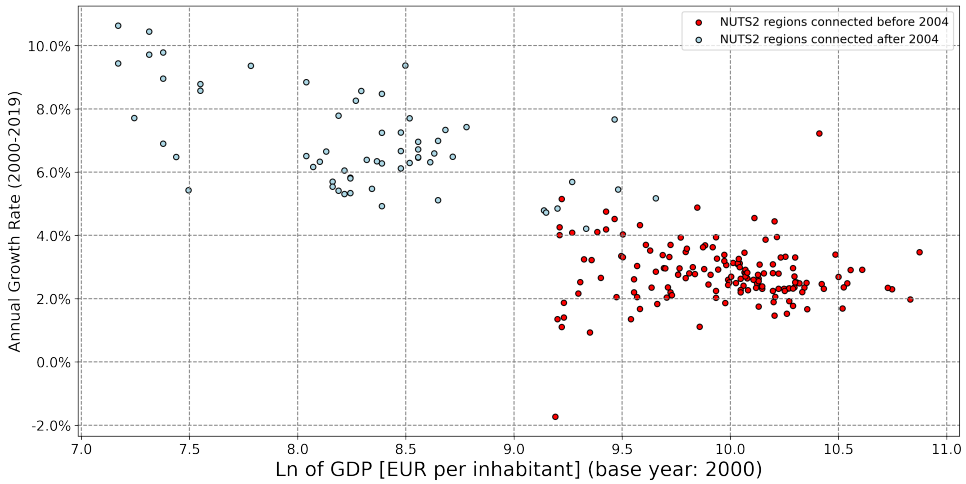



**Figure 3.6:** MFV values of annual growth rates of GDP per inhabitant values for NUTS2 regions with indication to the date of connection to the European Union.

Pearson's correlation coefficient, while the lower value corresponds to the MFV-robustified correlation coefficient.

| Economic Indicator | Dimension | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|
| GDP | EUR per inhabitant | -0.47<br>-0.52 | -0.82<br>-0.35 | -0.82<br>-0.34 |
| | PPS per inhabitant | -0.43<br>-0.54 | -0.41<br>-0.31 | -0.69<br>-0.44 |
| NDI | EUR per inhabitant | -0.92<br>-0.77 | -0.89<br>-0.33 | N.A. |
| | PPS per inhabitant | -0.94<br>-0.91 | -0.86<br>-0.49 | N.A. |

**Table 3.4:** Calculated Pearson- and MFV-robustified Pearson correlation coefficients for averages of annual GDP growth rates.

| Economic Indicator | Dimension | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|
| GDP | EUR per inhabitant | -0.69<br>-0.68 | -0.82<br>-0.52 | -0.77<br>-0.37 |
| | PPS per inhabitant | -0.60<br>-0.69 | -0.51<br>-0.38 | -0.63<br>-0.42 |
| NDI | EUR per inhabitant | -0.91<br>-0.86 | -0.84<br>-0.42 | N.A. |
| | PPS per inhabitant | -0.87<br>-0.84 | -0.79<br>-0.44 | N.A. |

**Table 3.5:** Calculated Pearson- and MFV-robustified Pearson correlation coefficients for MFVs of annual GDP growth rates.

According to literature suggestions, robust statistical methods shall be applied alongside with conventional statistics and in case of big differences in the obtained values for the same issue the presence of outliers and non-normality is to be assumed. In such scenarios, the result gained via conventional statistics should be questioned, and further investigations are advised [R63].

The cells marked with blue in Table 3.4. and 3.5. indicates a dramatic drop in the MFV-robustified correlation compared to the conventional Pearson's correlation. This fact might indicate – as was elaborated previously – that the presence of outliers and non-normality of the background distribution had a huge distorting effect on correlation. Out of this

reason, researches and practitioners can overrate the strength of linear relationship in these cases, which in economical aspect can lead to exaggerated assessment of economic growth of NUTS2 and NUTS3 regions towards an equilibrium growth path.

In other words, the decrease of the gap measured in GDP and NDI between 2000 and 2018 is not that significant on NUTS2 and NUTS3 level, as could be concluded based on conventional correlation analysis. At the same time, on country level the more aggregated and averaged measures are able to mask the unevenness of economic growth and a clear trend towards equilibrium with lowering differences among strong and weak economies can be assumed.

For performing MFV-robustified linear regression for the investigation of regional economic convergence within the concept of absolute $\beta$-convergence, first the generalized Newton's method was selected for solving the nonlinear equation system (see Eq. 3.12.). However, it is assumed that Broyden's method can help to further reduce computation complexity and therefore the necessary computation time as well [R64]. In order to achieve a sensible and reliable comparison of the selected two algorithms, $1000$ runs have been carried out in case of each data set. The run-times of the algorithms showed a highly skewed or even bimodal distribution (see Fig. 3.7.), therefore besides average values the medians and the MFVs have been calculated as well. These values have been provided in the same order in each of the cells of Table 3.6. and the smaller values in each case have been highlighted by green.

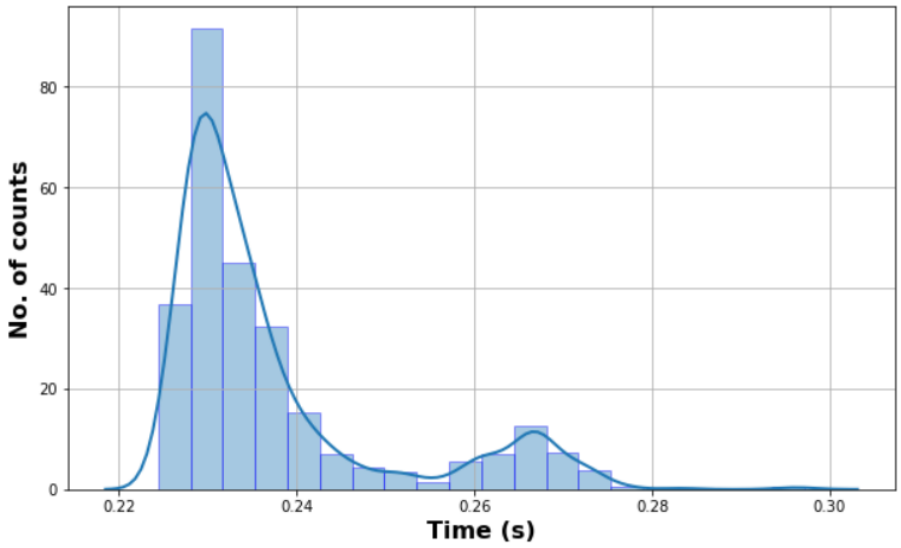



**Figure 3.7:** The distribution of measured run-times for the MFV-robustified linear regression with Broyden nonlinear solver in case of NUTS3 regions for $1000$ executions.

Although the elements of the Jacobian matrix in Eq. 3.12. could be generated based on analytical formulas, the Broyden's algorithm turned out to be the better choice when the sample size increased. Nevertheless, Table 3.6. clearly shows that the selection of the method for solving the nonlinear system is problem dependent and has no influence on the number of iterations necessary for the MFV algorithm to achieve the given limit for the convergence. The typical curves for visualising the convergence of the estimated parameters are illustrated on Fig. 3.8.

Moreover, the rate of convergence for the MFV-robustified linear regression resulted to be the same for two digits in case of both nonlinear solvers and fluctuated around 1.13. The estimation of the rate of convergence can be done based on Eq.3.18. for sufficiently large $m$ values, where $\underline{\mathbf{p}}^{(m)}$ denotes the parameter vector in the $m$-th iteration step and $\underline{\mathbf{p}}^{*}$ denotes the converged value of the parameter-vector [R65]. In the applied settings

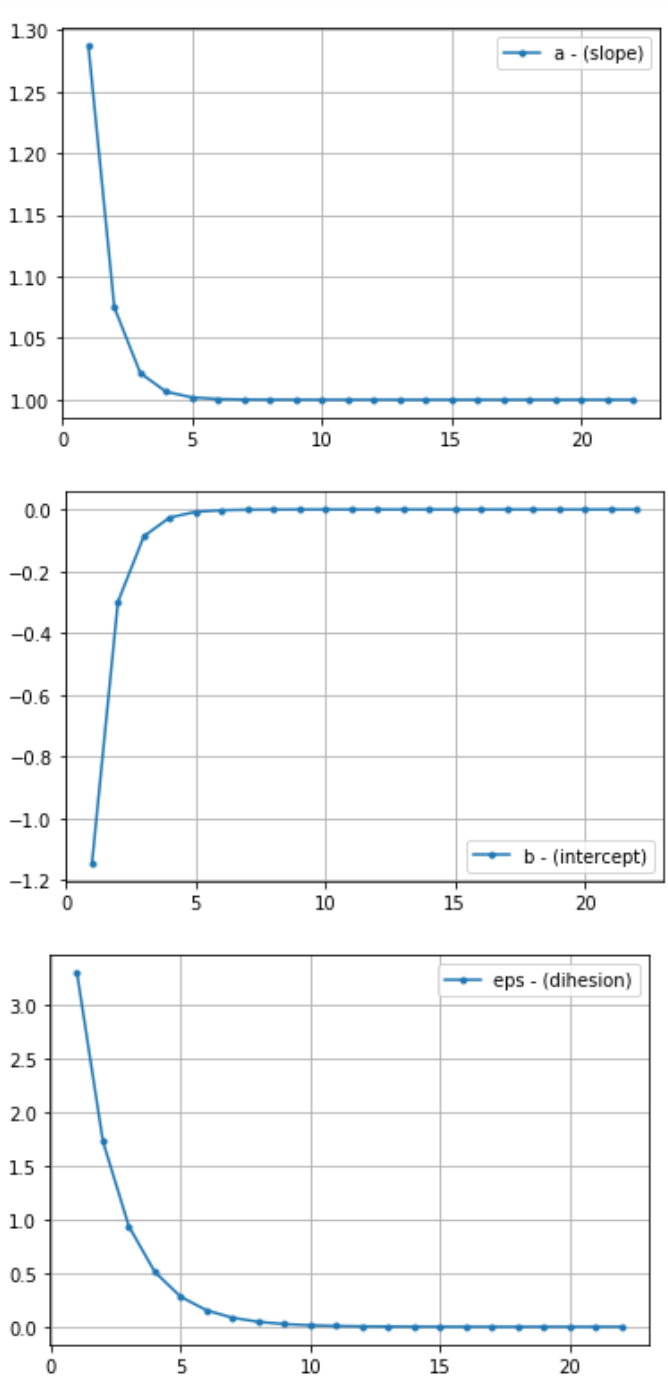



**Figure 3.8:** Typical curves for illustrating the convergence of slope-, intercept- and dihesion parameters of the MFV-based linear regression.

$\|\underline{\mathbf{p}}^{(m+1)} - \underline{\mathbf{p}}^*\| < 10^{-5}$ condition had to be fulfilled.

$$\lim_{m\to\infty} \frac{ln\|\underline{\mathbf{p}}^{(m+1)} - \underline{\mathbf{p}}^*\|}{ln\|\underline{\mathbf{p}}^{(m)} - \underline{\mathbf{p}}^*\|} \tag{3.18}$$

For the calculations a personal computer equipped with an Intel(R) Core(TM) i7-9750H CPU at 2.60GHz with 6 cores and was equipped with 15.9 GB physical memory in total was used.

Robust statistical procedures can often highlight ”abnormalities” that are unfortunately characteristic for the data at hand. Nevertheless, it shall be avoided to rely only on

| Dataset | Sample size | No. of MFV-iterations | $t_{Newton}$(s) | $t_{Broyden}$(s) |
|---|---|---|---|---|
| Demo task | 11 | 6 | 0.0028 | 0.0030 |
| | | | 0.0030 | 0.0030 |
| | | | 0.0030 | 0.0030 |
| Country | 27 | 39 | 0.0217 | 0.0232 |
| | | | 0.0219 | 0.0214 |
| | | | 0.0216 | 0.0220 |
| NUTS2 | 213 | 24 | 0.1071 | 0.1068 |
| | | | 0.1057 | 0.1047 |
| | | | 0.1059 | 0.1054 |
| NUTS3 | 1066 | 26 | 0.5354 | 0.5146 |
| | | | 0.5266 | 0.5107 |
| | | | 0.5260 | 0.5127 |

**Table 3.6:** Average-, median- and MFV values of run-times for Newton's and Broyden's solver for the four different investigated examples.

the results provided by such algorithms, since the statistical efficiency (the information extraction rate) of the applied methods can be highly different. On the other hand, the theory of the MFV procedures are not yet elaborated to specify significance levels or confidence internals for the estimates. Thus, the values provided for the slope parameters in Table 3.7. are also given without these measures in an indicative way, with the main purpose to highlight the differences gained by OLS and MFV-robustified linear regression models [R14, R49].

| Dataset | OLS | MFV | Relative change |
|---|---|---|---|
| Country | -1.582 | -1.791 | 13.2% |
| NUTS2 | -2.226 | -1.911 | -14.2% |
| NUTS3 | -2.008 | -1.791 | -11.6% |

**Table 3.7:** Slope parameters for OLS and MFV-robustified lines for different NUTS region data together with the relative change compared to the OLS values.

The corresponding slope parameters in Table 3.7. shows that a negative trend is present in case of all investigated NUTS level. This indicates the validity of the economic $\beta$-convergence theorem, at least in the sense that the negative linear relationship exists for the accessed EU NUTS regional GDP per capita data (the greater absolute value of the slope parameter indicates higher economic convergence). For country-level data the MFV-robustified fit even shows stronger linear relationship, however it can be suspected that these data contain rather averaged values that might further incorporate distortions from locally deviating regions or purely just from the typically non-normal distribution of the lower level data points. For the NUTS2 and NUTS3 level data however, the MFV-robustified fit already suggests a weaker linear relationship than the OLS fit, which suggest a less clear evidence for the stronger development tendency of weaker regions throughout the past almost $20$ years. (In comparison, [R29] presented -1.83 in the time period 2007-2014, while [R17] -2.7 between 2001-2012 for country level. Direct comparison with authors due to application to different regional sets and investigated time periods is problematic.)

## 3.5 Further Findings Regarding Economic Convergence

As the investigation of Fig. 3.6. regarding absolute economic $\beta$-convergence indicates EU member regions connected before 2004 and afterwards constitute two separate groups. Therefore, viewing the former economic indicators, economic convergence on data-population

level necessitates the convergence of the two subpopulations (group of countries, regions) to each other as well. As Fig. 3.9. and 3.10. show the weaker regions originate mainly from member states connected to the European Union before 2004, that had different development curves due to non-capitalist institutional arrangement for several decades before the change of regime. The investigation of economic convergence in the aspect of such decomposition is of high interest in corresponding literature, since the great Eastern enlargement of the EU has increased its population approximately by 20% but contributed only with around 4% GDP increase [R34].

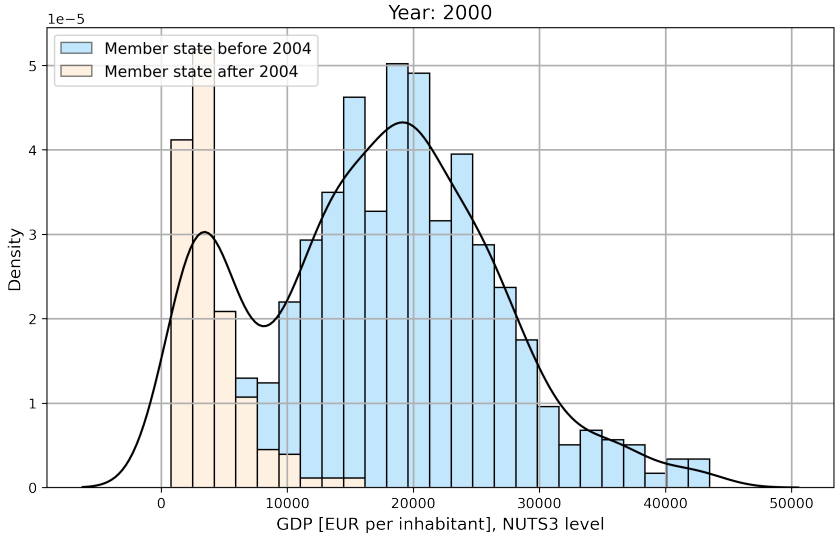



**Figure 3.9:** Histogram representation for the intra-distribution share of NUTS3 regions for year 2000 regarding GDP [EUR per capita] that become a part of the EU before 2004 and afterwards, with corresponding kernel density estimations.

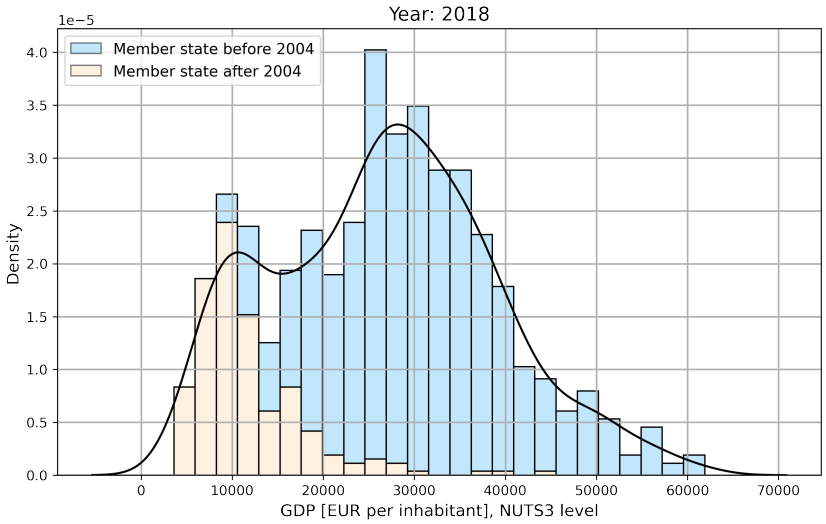



**Figure 3.10:** Histogram representation for the intra-distribution share of NUTS3 regions for year 2018 regarding GDP [EUR per capita] that become a part of the EU before 2004 and afterwards, with corresponding kernel density estimations.

The subsequent snapshots of the decomposed density curves to discrete histograms indicates that the bimodal attribute (at least in former years) is definitely due to the assumed differences between these groups (see Fig. 3.11. and 3.12.). In spite of the visually perceivable decrease of bimodality and the tendency of two subpopulations (group of

countries, regions) towards a uniform, common distribution, the differences in the location parameters of the two sub-data-distribution might not illustrate an obvious trend. In some cases, an even increasing tendency is suggested that can be a direct consequence of the increase of disparities and economic divergence (see Fig. 3.13.)!

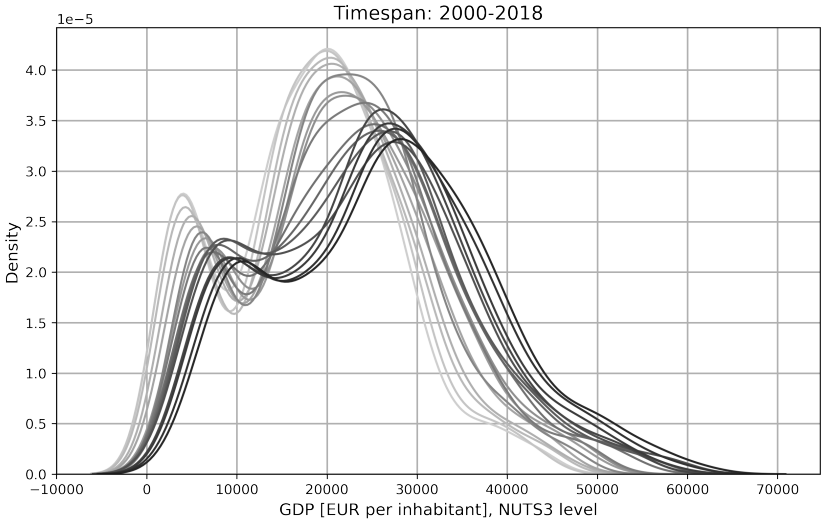



**Figure 3.11:** Kernel density estimation plots for GDP per capita values measured in EUR on NUTS3 levels. The dimmer a curve is, the former year corresponds to it.

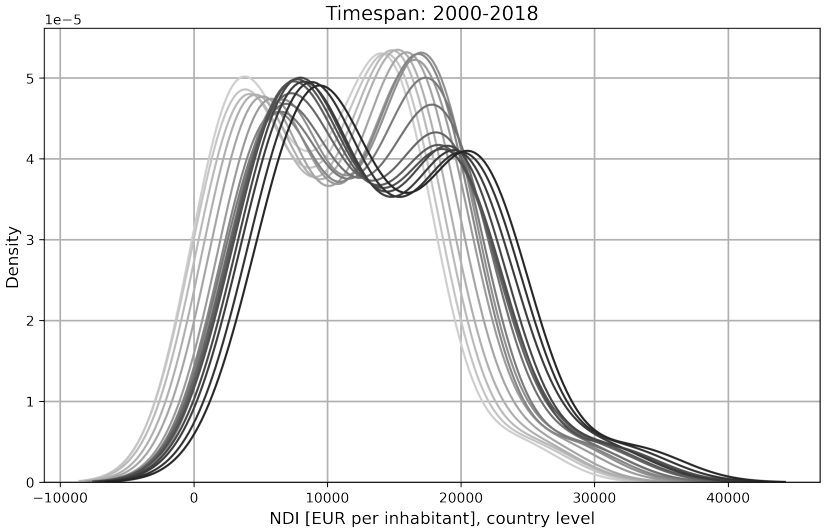



**Figure 3.12:** Kernel density estimation plots for NDI per capita values measured in EUR on country levels. The dimmer a curve is, the former year corresponds to it.

This paradox situation posed by the whole distribution and agglomerating location parameters has also been observed in corresponding literature, to which phenomenon the authors attribute the varying or even contradictory findings regarding economic convergence within the EU member states and regions [R31, R36]. Therefore, in the following I will apply non-parametric statistical procedures that consider the whole of the distribution functions, which – unlike location parameters – might be influenced by non-Gaussian distributions and far-lying data points to a lesser degree.

For the verification of the outlined hypothesis of the convergence of the two group of countries and regions, Mann-Whitney U-tests and Wilcoxon signed rank tests were utilized. The Mann-Whitney U-test (or Wilcoxon rank-sum test) is a non-parametric alternative of

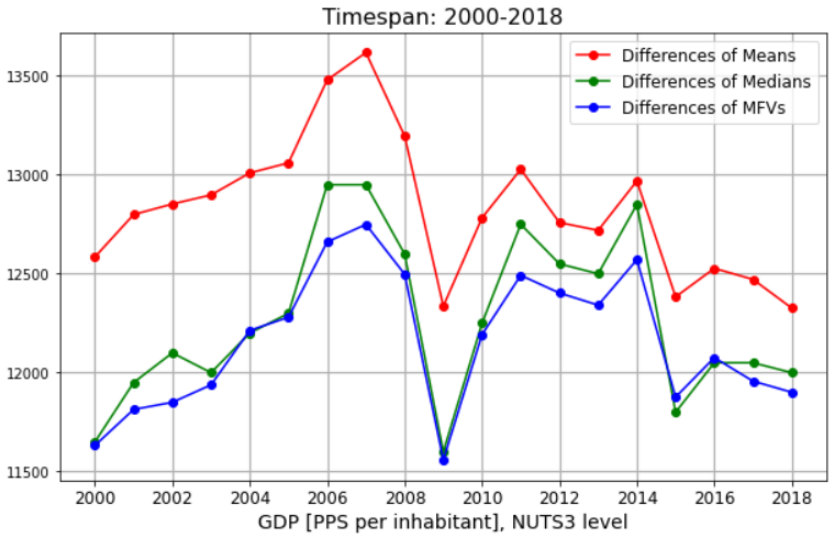



**Figure 3.13:** Annual differences of mean-, median- and MFV location parameters of the member NUTS3 regions (GDP per capita measured in PPS) joined the EU before 2004 and afterwards.

the two sample t-test and tests the $H_0$ hypothesis whether two samples are from the same data-population via investigating the equality of medians of the two distributions (the two distributions to be compared has similar type in each year). In contrast to its parametric alternative the Mann-Whitney test is more robust and less outlier-sensitive, therefore a higher statistical efficiency can be expected by applying it to our data. For smaller sample sizes, statistical tables are available for calculating critical values, while for larger element numbers the normal approximation can be used. In the latter we interpret the transformed statistic value as a $N(0,1)$ distributed variable and the $H_0$ hypothesis for the identification of the two distributions in case of a two-tailed test at $95\%(p = 0.05)$ significance level can be rejected if $|Z| \leq Z_{crit} = 1.96$ [R1]. However, utilizing the extra information that the member states joined after 2004 have smaller values of financial indicators than those joined before, a one-tailed test with $|Z| \leq Z_{crit} = 1.645$ can be used that has more statistical power.

Let us denote the number of elements in the two subpopulations (group of countries and regions) by $n_1$ and $n_2$, the sum of ranks by $R_1$ and $R_2$ accordingly, and let us assume that $R_1 < R_2$. Then the U-statistic can be computed as $U = R_1 - n_1(n_1 + 1)/2$. The total number of elements is $n_1 + n_2$, while the sum of ranks can be given as $n_1 n_2$. The normal approximation of the U-statistic can be calculated according to $Z_u = (U - m_u)/\sigma_u$, where $m_u = n_1 n_2/2$ and $\sigma_u = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$ are the corresponding mean and standard deviation for the U-values (for $n_1$, where $n_1 \geq 20$).

Since the independence of our real-life regional data cannot be guaranteed (that is an assumption of Mann-Whitney U-test) due to geographical interconnections, two-sample Wilcoxon signed rank tests have been performed on paired samples as well for each year in order to support the results gained by the Mann-Whitney U-test. In this case, the calculated $T$ statistic is the minimum of the positive and negative rank sums of the differences of the paired data and for $n \geq 50$ the normal approximation can be used according to $Z_t = (T - m_t)/\sigma_t$, where $m_t = n(n + 1)/4$ and $\sigma_u = \sqrt{n(n + 1)(2n + 1)/24}$ are the corresponding mean and standard deviation for the T-values and $n$ is the number of selected pairs.

Having calculated the corresponding $Z_u$ and $Z_t$ statistic values for each year, the tendency of the statistics can be utilized to search for a possible intersection point with the critical U-value and predictions can be made for the economic convergence tendencies between the groups of countries and regions concerned on data-population level. This technique is analogous to the usage of sequential hypothesis tests used in mass production sites to test

the validity of a certain hypothesis. By observing the distance of serially calculated statistic values from the critical value and/or its tendency towards it, larger deviations from the admissible can be prevented, and the measured parameter can be kept within desired limits with higher certainty [R66].

According to the visual sanity check of the time evolution of bimodality, in case of the obtained data listed in Table 3.1. the estimated normal approximation of the U- and T-statistic values followed a similar trend. Although, in case of country level the usage of the normal approximation is not necessarily as precise and founded as in case of the NUTS2- and NUTS3 level data, for comparative purposes regarding convergence tendencies the corresponding Z-statistics have been calculated for these cases as well.

Since the number of regions for member states connected to the EU before 2004 is greater than those that connected after 2004 and the regions cannot be paired based on any obvious attribute, the pairing for the Wilcoxon signed rank test could not be performed straightforwardly. Therefore, in each year a random selection of the pairs from the two data-subpopulations has been done 1000 times in each year and the Most Frequent Value of the resulting distributions of the $Z_t$ values have been selected for further trend analyses [R6].

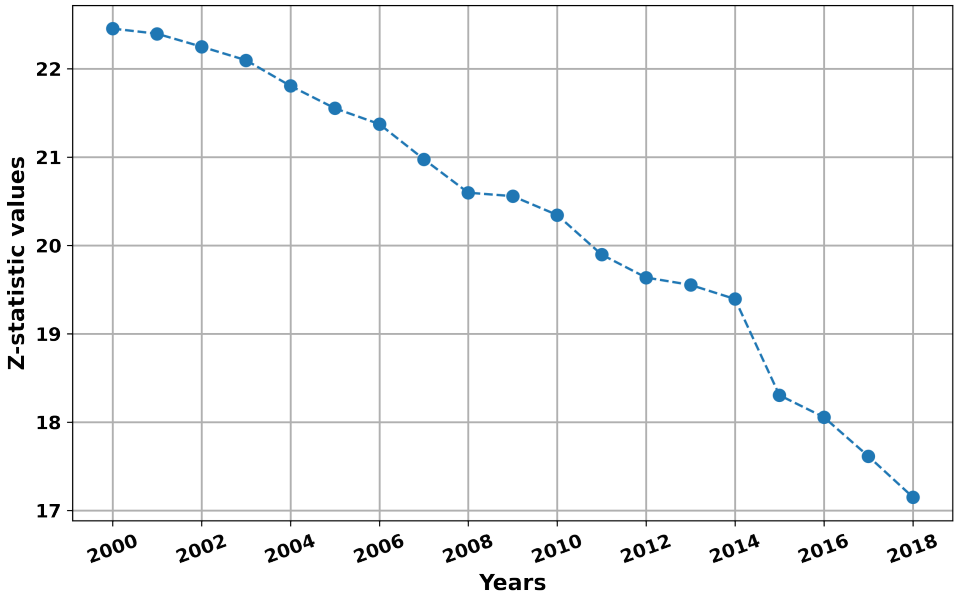



**Figure 3.14:** Calculated $Z_u$-statistic values from Mann-Whitney-U test for the available time period in case of GDP per capita values in PPS for NUTS3 regions.

The estimated Z-values followed a stagnating trend in time for NDI values for country level. However, the Z-values in case of NDI for the NUTS2 level and GDP values for all regional level showed a decreasing tendency (see Fig. 3.14.). The resulted statistics turned out to be highly significant for all the inspected financial variables in every given year, indicating the statistically distinguishable distributions of the EU member states joined before 2004 and afterwards.

The time dependency of the Z-statistic values clearly showed a nonlinear trend that could be seen by fitting linear and higher order polynomial functions in the least squares manner onto them. Having investigated the $R^2$ correlation index, the second order polynomial function turned out to serve with a better fit than the linear one. Therefore, – in order not to overfit our data – parabolic functions have been used to characterize the trend of the Z-values and consequently the rate of economic convergence (where it was applicable, see Table 3.8. and 3.9.) can be approximately viewed as quadratic.

After having fitted $ax^2 + bx + c$ type polynomials the uncertainties of the parameters

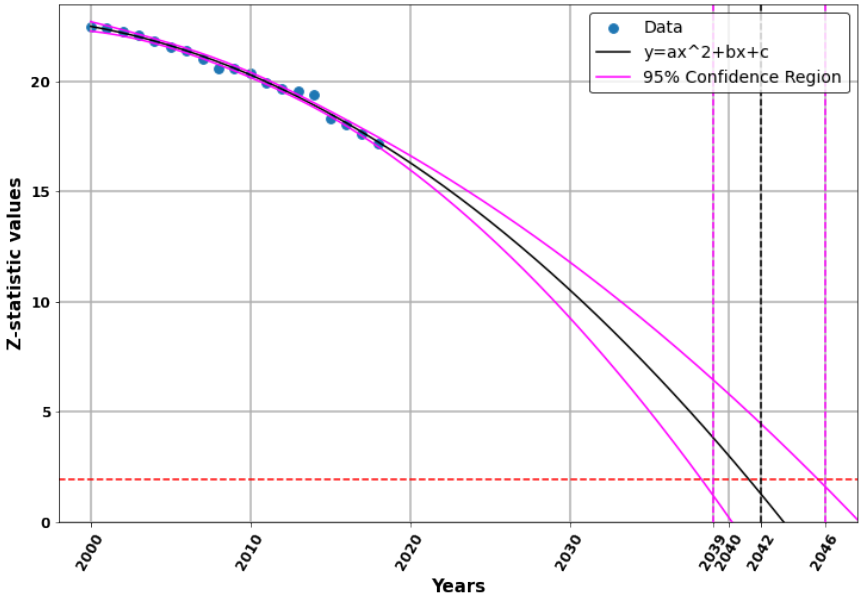



**Figure 3.15:** Estimated parabolic regression of $Z_u$-statistic values in case of GDP per capita values in PPS for NUTS3 level and the estimated intersection point with the critical $Z_u$-value.

and the intersection points with the $Z_{crit} = 1.645$ values have been calculated together with the intersection points of the 95% confidence intervals where it was possible[4]. The corresponding parameter values and the intersection points are provided in Table 3.8. and Table 3.9. The intersection points, as estimated years for the two data-subpopulations to become statistically indistinguishable forecasted by the applied non-parametric tests, were defined as the first year afterwards the corresponding fitted curve descended under the $Z_{crit}$ horizontal line (see Fig. 3.15.).

As Table 3.8. and Table 3.9. indicates there is a certain level of convergence in case most of the data. Years of convergence are marked by green, where a clear descending tendency of Z-values is observable with relatively narrow confidence intervals. Where the trend predicts a slow convergence and intersection points vaguely far in the future the years are marked by orange, while cases of no convergence are marked by red. In case of GDP [EUR per inhabitant] on NUTS2 level, GDP [PPS per inhabitant] on NUTS3 level and NDI [PPS per inhabitant] on NUTS2 level showed a rather definite convergence tendency and in these cases the $R^2$ correlation index for the parabolic regression curve proved to be outstandingly high as well.

By the above analysis of the time evolution of non-parametric statistics calculated between groups of member states and regions joined the European Union before 2004 and afterwards it is shown that on subpopulation level of the data a converge tendency is perceivable. Furthermore, the rate of convergence and a rough estimate for the time of ”merging” of the groups of countries and regions (or at least the methodology of estimating it) are also outlined. According to the obtained results on economic convergence properties of EU NUTS regions, a less expressed convergence tendency is present than would be suggested by other literature sources [R17, R29].

The careful investigation of the annual distribution of GDP and NDI values on different

[4]Since regional data cannot be considered as random sampling (re-sampling of the data cannot be performed or extended by other region members), values gained from confidence interval intersection points shall be treated as an indicative measure for prediction accuracy.

| Economic Indicator | Dimension | Parameter | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|---|
| GDP | EUR per inhabitant | a | -0.0023 ± 0.0014 | -0.0039 ± 0.0006 | -0.0028 ± 0.0009 |
| | | b | 13.39 ± 5.46 | 15.61 ± 2.53 | 11.35 ± 3.47 |
| | | c | -13444.47 ± 5496.67 | -15584.64 ± 2544.05 | -11283.43 ± 3487.49 |
| | | $R^2$ | 0.9143 | 0.9805 | 0.9702 |
| | | Year | 2037 (2032, 2053) | 2048 (2044, 2054) | 2081 (2069, 2111) |
| | PPS per inhabitant | a | -0.0018 ± 0.0014 | -0.0083 ± 0.0050 | -0.0089 ± 0.0015 |
| | | b | 7.13 ± 5.57 | 33.08 ± 20.16 | 35.55 ± 6.09 |
| | | c | -7105.37 ± 5608.25 | -33012.37 ± 20296.62 | -35399.56 ± 6119.65 |
| | | $R^2$ | 0.9694 | 0.9742 | 0.9899 |
| | | Year | 2035 (2030, -) | 2030 (2027, 2042) | 2042 (2040, 2047) |
| NDI | EUR per inhabitant | a | 0.0005 ± 0.0002 | -0.0016 ± 0.0004 | No data |
| | | b | -1.97 ± 0.79 | 6.49 ± 1.42 | |
| | | c | 1990.99 ± 789.32 | -6493.07 ± 1421.69 | |
| | | $R^2$ | 0.6482 | 0.9236 | |
| | | Year | NO CONVERGENCE | 2080 (2070, 2100) | |
| | PPS per inhabitant | a | 0.0007 ± 0.0008 | -0.0081 ± 0.0008 | No data |
| | | b | -2.79 ± 3.06 | 32.44 ± 3.08 | |
| | | c | 2821.51 ± 3069.25 | -32436.45 ± 3091.18 | |
| | | $R^2$ | 0.5201 | 0.9888 | |
| | | Year | NO CONVERGENCE | 2036 (2034, 2038) | |

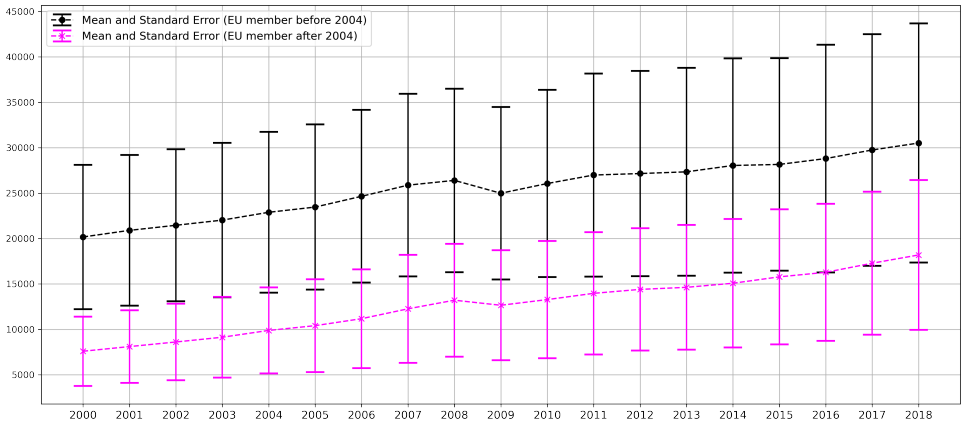
**Table 3.8:** Estimated regression parameters with corresponding uncertainties, correlation indexes and estimated intersection points of the regression line and confidence interval curves with the critical Z-value (in brackets) where applicable for Mann-Whitney U-test.

| Economic Indicator | Dimension | Parameter | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|---|
| GDP | EUR per inhabitant | a | -0.0005 ± 0.0002 | -0.0011 ± 0.0002 | -0.0006 ± 0.0002 |
| | | b | 2.08 ± 0.96 | 4.58 ± 0.63 | 2.33 ± 0.74 |
| | | c | -2084.39 ± 968.02 | -4568.78 ± 630.18 | -2307.31 ± 740.19 |
| | | $R^2$ | 0.9528 | 0.9814 | 0.9468 |
| | | Year | 2060 (2050, 2116) | 2067 (2062, 2075) | 2139 (2114, 2209) |
| | PPS per inhabitant | a | -0.0020 ± 0.0003 | -0.0034 ± 0.0012 | -0.0021 ± 0.0005 |
| | | b | 8.11 ± 1.30 | 13.72 ± 5.02 | 8.51 ± 1.95 |
| | | c | -8147.54 ± 1313.59 | -13681.67 ± 5050.06 | -8416.61 ± 1959.41 |
| | | $R^2$ | 0.9845 | 0.9913 | 0.9933 |
| | | Year | 2037 (2034, 2040) | 2033 (2031, 2039) | 2059 (2053, 2068) |
| NDI | EUR per inhabitant | a | $(-2.71 \pm 0.08)\cdot10^{-11}$ | $-0.0002 \pm 3.24\cdot10^{-5}$ | No data |
| | | b | $(1.09 \pm 0.08)\cdot10^{-7}$ | 0.62 ± 0.13 | |
| | | c | $3.06 \pm 3.23\cdot10^{-6}$ | -616.98 ± 130.77 | |
| | | $R^2$ | 0.9999 | 0.8874 | |
| | | Year | NO CONVERGENCE | 2186 (2159, 2237) | |
| | PPS per inhabitant | a | $(2.6 \pm 3.2)\cdot10^{-4}$ | -0.0029 ± 0.0002 | No data |
| | | b | -1.05 ± 1.29 | 11.15 ± 0.83 | |
| | | c | 1060.61 ± 1299.93 | -11156.51 ± 835.39 | |
| | | $R^2$ | 0.2274 | 0.9867 | |
| | | Year | NO CONVERGENCE | 2047 (2044, 2049) | |

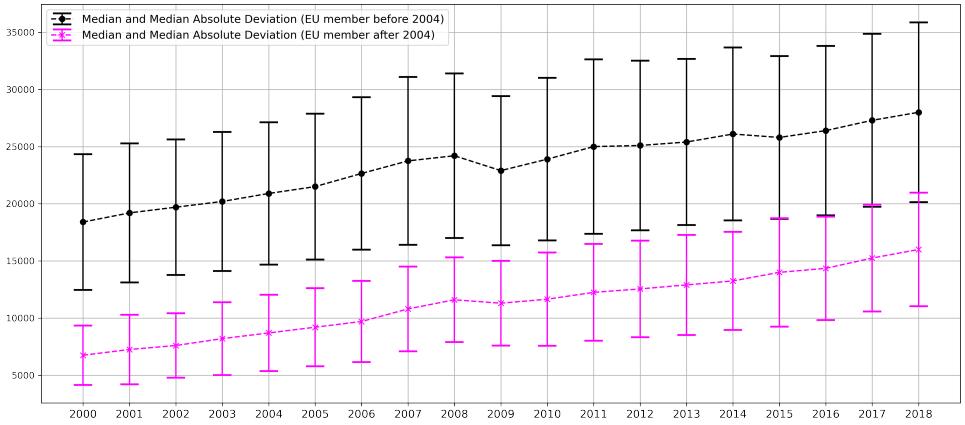
**Table 3.9:** Estimated regression parameters with corresponding uncertainties, correlation indexes and estimated intersection points of the regression line and confidence interval curves with the critical Z-value (in brackets) where applicable for Wilcoxon signed rank test.

regional levels also shed light on potentially growing inequalities among the observed elements. Comparing the x-axes of Fig. 3.9. and Fig. 3.10. it can be seen that over the investigated time period the GDP distirbutions on NUTS3 level have spread out. This applies for both groups implicating growing inequalities within each cluster even if converging tendencies on data-population level are present. The increasing trend in the time-series of each location- and scale parameter of GDP [PPS per inhabitant] are visualised on Fig. 3.16. and their overall changes between the start and end date with regard to the investigated financial measures in percentages are listed in Table 3.10. and Table 3.11.

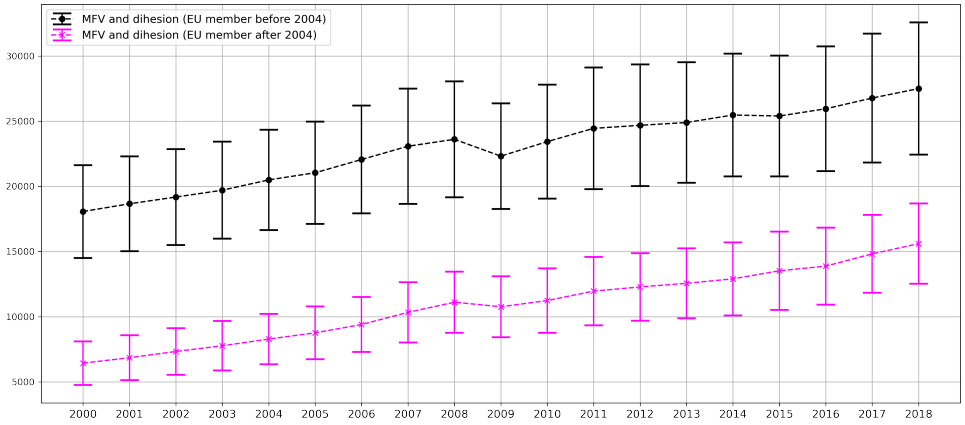
Increasing regional income inequality as a by-product of economic growth has been already treated it the 1960s within the framework of the so-called Williamson hypothesis. It suggests that economic development increases regional inequality first which then tends to decrease with time forming an inverted U-shaped curve. According to the hypothesis economic expansion usually concentrates only in few urbanized areas which then show




**(a)** Time evolution of mean and standard deviation values.




**(b)** Time evolution of median and median absolute deviation values.




**(c)** Time evolution of MFV and dihesion values.

**Figure 3.16:** Increase in location- and scale-parameters of annual GDP [PPS per inhabitant] distributions on NUTS3 level for regions connected before 2004 to the European Union and afterwards.

| Economic Indicator | Dimension | Parameter | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|---|
| GDP | EUR per inhabitant | $\Delta(mean)\%$ | 28.1% | 53.7% | 51.3% |
| | | $\Delta(std)\%$ | 66.7% | 70.4% | 59.4% |
| | | $\Delta(median)\%$ | 25.6% | 48.9% | 52.5% |
| | | $\Delta(MAD)\%$ | 113.0% | 33.9% | 32.0% |
| | | $\Delta(IQR)\%$ | 78.2% | 40.1% | 31.3% |
| | | $\Delta(MFV)\%$ | 24.5% | 48.4% | 52.1% |
| | | $\Delta(dihesion)\%$ | 88.6% | 51.5% | 42.2% |
| | PPS per inhabitant | $\Delta(mean)\%$ | 0.6% | -8.1% | 51.3% |
| | | $\Delta(std)\%$ | 29.4% | 9.8% | 65.6% |
| | | $\Delta(median)\%$ | 0.4% | -10.9% | 52.2% |
| | | $\Delta(MAD)\%$ | 65.0% | -10.0% | 32.5% |
| | | $\Delta(IQR)\%$ | 76.9% | -12.2% | 36.7% |
| | | $\Delta(MFV)\%$ | -1.9% | -10.5% | 52.2% |
| | | $\Delta(dihesion)\%$ | 58.9% | 4.6% | 42.4% |
| NDI | EUR per inhabitant | $\Delta(mean)\%$ | 50.9% | 44.8% | |
| | | $\Delta(std)\%$ | 61.5% | 52.2% | |
| | | $\Delta(median)\%$ | 54.4% | 50.0% | |
| | | $\Delta(MAD)\%$ | 0.0% | 20.0% | No data |
| | | $\Delta(IQR)\%$ | 55.0% | 36.0% | |
| | | $\Delta(MFV)\%$ | 55.5% | 50.0% | |
| | | $\Delta(dihesion)\%$ | 27.7% | 26.9% | |
| | PPS per inhabitant | $\Delta(mean)\%$ | 43.1% | 41.5% | |
| | | $\Delta(std)\%$ | 17.5% | 39.6% | |
| | | $\Delta(median)\%$ | 47.0% | 41.8% | |
| | | $\Delta(MAD)\%$ | -21.0% | 15.8% | No data |
| | | $\Delta(IQR)\%$ | -32.7% | 12.2% | |
| | | $\Delta(MFV)\%$ | 53.3% | 43.5% | |
| | | $\Delta(dihesion)\%$ | -26.2% | 27.1% | |

**Table 3.10:** Relative changes of location- and scale parameters of GDP and NDI values on each regional level between the beginning and ending dates of the investigated time period for countries and regions connected to the European Union before 2004.

elevated growth rates. This can cause consequently regional disparities. Then further market divergence can be a result of concentrated investment, urbanization and migration (e.g.: brain-drain) towards economically thriving regions that aggrevate disparities in income and infrastructure compared to rural areas. Nevertheless, due to market saturation, tecnology diffusion, policy interventions or other spillover effects economic growth eventually spreads over undeveloped areas thereby decreasing regional differences. This leads to a catch up effect with core regions and convergence in economic sense. This latter stage of the Williamson model is in alignment with the concept of absolute $\beta$-convergence. Furthermore, it suggests that with appropriate policy instruments (e.g.: timely investment in human capital, technology and infrastructure etc.) less developed regions can be helped to catch up with richer ones faster in this way closing income gaps and contributing to cross-regional well-being and citizen satisfaction [R67, R68, R69].

The indcated relative changes in location- and scale parameters of the annual distributions on different regional levels given in Table 3.10. and Table 3.11. further imply that a larger positive change in a location value is typically accompanied by a larger increase in the corresponding scale parameter. This is especially dominant for the NUTS3 case, where more data lets deeper understanding of the ongoing processes and does not let agveraging veil them, that is in particular expected for country-level investigations. According to calculated robust location- and scale parameters, besides data-population-level convergence of regions connected to the European Union before and after 2004, growing within-cluster inequalities are more explicit in the east-block where overall convergence speed was higher. This is in line with the suggestions of the Williamson hypothesis. Nonetheless, further research is advised, where not just the spereading tendencies described by the increasing relative changes in scale parameters between the start- and end dates of annual distributions of regional-level financial measures are considered, then in-depth time-series analysis and

| Economic Indicator | Dimension | Parameter | Country | NUTS2 | NUTS3 |
|---|---|---|---|---|---|
| GDP | EUR per inhabitant | $\Delta(mean)\%$ | 48.0% | 216.0% | 189.3% |
| | | $\Delta(std)\%$ | 0.9% | 159.3% | 131.5% |
| | | $\Delta(median)\%$ | 60.3% | 211.9% | 183.1% |
| | | $\Delta(MAD)\%$ | 54.1% | 220.0% | 75.9% |
| | | $\Delta(IQR)\%$ | 13.7% | 250.0% | 115.9% |
| | | $\Delta(MFV)\%$ | 67.7% | 203.4% | 177.9% |
| | | $\Delta(dihesion)\%$ | 54.6% | 164.2% | 108.1% |
| | PPS per inhabitant | $\Delta(mean)\%$ | 14.5% | 15.8% | 139.7% |
| | | $\Delta(std)\%$ | -29.6% | 4.1% | 116.1% |
| | | $\Delta(median)\%$ | 16.9% | 22.2% | 137.0% |
| | | $\Delta(MAD)\%$ | -8.3% | -7.1% | 91.4% |
| | | $\Delta(IQR)\%$ | -44.4% | -7.4% | 107.2% |
| | | $\Delta(MFV)\%$ | 23.6% | 21.2% | 142.4% |
| | | $\Delta(dihesion)\%$ | -2.8% | -5.7% | 84.4% |
| NDI | EUR per inhabitant | $\Delta(mean)\%$ | 164.8% | 173.6% | No data |
| | | $\Delta(std)\%$ | 18.1% | 72.8% | |
| | | $\Delta(median)\%$ | 249.0% | 152.6% | |
| | | $\Delta(MAD)\%$ | 72.2% | 100.0% | |
| | | $\Delta(IQR)\%$ | 107.8% | 113.2% | |
| | | $\Delta(MFV)\%$ | 242.2% | 146.4% | |
| | | $\Delta(dihesion)\%$ | 107.6% | 84.1% | |
| | PPS per inhabitant | $\Delta(mean)\%$ | 127.7% | 130.5% | No data |
| | | $\Delta(std)\%$ | -9.3% | 53.1% | |
| | | $\Delta(median)\%$ | 122.1% | 123.0% | |
| | | $\Delta(MAD)\%$ | -27.6% | 30.4% | |
| | | $\Delta(IQR)\%$ | 18.2% | 8.7% | |
| | | $\Delta(MFV)\%$ | 133.7% | 125.2% | |
| | | $\Delta(dihesion)\%$ | -10.1% | 29.4% | |

**Table 3.11:** Relative changes of location- and scale parameters of GDP and NDI values on each regional level between the beginning and ending dates of the investigated time period for countries and regions connected to the European Union after 2004.

country-level breakdown of regions [R68]. This investigation however lies beyond of the methodological focus of the present thesis.

As an extension of the previous findings regarding economic convergence of European countries and regions, a further aspect had been considered with respect of regional R&D activities. Albeit economic convergence is hard to connect with R&D investments due to the relatively small order of magnitudes compared to other regional spendings, it is well-known that knowledge generation and innovation is the foundation of long-term development and thereby economic growth. As such, it is expected that a convergent geographical entity shall have an increasing collaboration and less developed regions shall be embedded even more into a common scientific community.

The increase of EU-level competitiveness in technological fields builds on the concept that knowledge generation is rather a collaborative process than done by individuals. Furthermore, profitable utilization of ”assembled” knowledge in the form of innovation is a driving force of economic growth. Knowledge is being transferred through such networks, and uneven connection distributions might put obstacles in reaching this goal [R70, R71]. Therefore, understanding of such structures on country- and regional level can contribute to better funding allocation and identification of right local strategic goals for the long run [R72, R73, R74, R75].

In the following, I will demonstrate that on a regional level, the cohesion among regions regarding scientific partnership establishment had an increasing tendency throughout the Horizon 2020 program period. This finding cannot bring in direct relation with the existence and pace of regional economic convergence, however the increasing embeddedness is at least a requirement that might establish further cohesive tendencies on economic levels as well. Therefore, the following results further support the previous findings in the sense

that they logically build a basis for a common economic growth and do not contradict the regional economic tendency investigation results.

Previous researches utilized among others scientific articles and reference lists [R70, R74], patent data [R76], national level tender information [R77], regional databases [R78, R79, R80] or the CORDIS[5] database [R81, R82, R83] in order to build up social networks of H2020 program attendees. Nevertheless, their main focus was the investigation of individual stakeholder connections and country level aggregated metrics [R84, R85]. Therefore, my aimed contribution in the following is to further extend corresponding literature with NUTS level regional investigations and elaborate on the time evolution of regional level aggregated social network metrics.

The utilization of social networks relies on graph theory, that can be considered as a topological abstraction of various real-life problems. In many cases however, self-organizing networks can be only rarely characterized by random graphs, scale-free properties prevail instead. One of the main differences of the two types is that the latter contains nodes with finite probability that possess outstanding number of connections and thereby can ”rule” over the network as being a high preferential point or strongly connected point. These networks cannot be described by single scales (e.g.: average number of connections) and their degree distribution usually follow some kind of power law [R86].

In case of scale-free graphs, the edge number is theoretically unbounded from above and the formation of new edges is not by chance. New edges tend to connect to nodes with more pre-existing edges, there is a preference towards nodes that are already ”preferred” within the network. Therefore, the scale-free attribute can be originated from preferential growth, new connections are more likely to be formed with more experienced, more valued or more creditable partners, which process also seems ”natural” in various walks of life. As a strict consequence, in such models the ”rich becomes richer” and there exists (an imbalanced) competition for new connections [R87]. Since social networks are typically formed on a natural basis, it might also be interesting to see for the regional level collaboration whether the previous statements hold.

Regarding H2020 funded project information, the CORDIS database was used. This was further extended by regional data from the EUROSTAT[6] database. These two sources were combined in order to be able to look into regional level projections of the H2020 resource utilization.

By this means $30$ countries ($27$ EU member states together with economically and regarding scientific activity tightly related Norway, Switzerland and United Kingdom), $284$ NUTS2- and $1215$ NUTS3 regions were taken into consideration. GDP data (from within 2011-2020), population data (from within 2014-2022) and density of population data (from within 2015-2019) as general metrics for characterising regional economies.

Information on $35367$ different projects and $40858$ different stakeholder organizations were accessed with $48$ different ”fields of science” activity area. The obtained variables were:

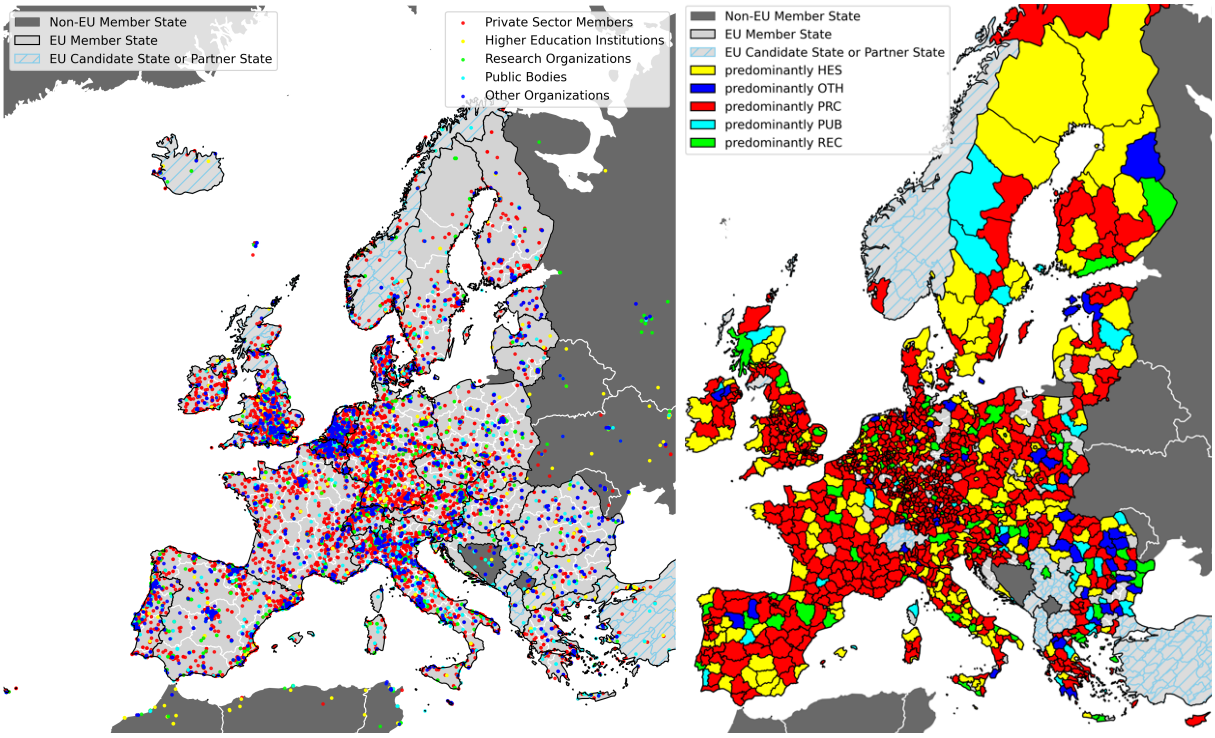
- project title
- ”teaser” of the project (can be considered as an abstract)
- project objective
- project start date
- project end date

[5]Community Research and Development Information Service, `https://data.europa.eu/data/datasets/cordish2020projects?locale=hu`, accessed: 2023.02.01.

[6]European Statistical Office, `https://ec.europa.eu/eurostat/web/main/data/database`, accessed: 2023.03.30.

- status (ongoing / terminated / closed)
- project activity area (”field of science”)
- names of involved organizations’
- geolocation of involved organizations
- EU contribution (*ecContribution* in EUR)
- organization activity type (Private Sector Members - PRC, Higher Education Institutions - HES, Research Organizations - REC, Public Bodies - PUB, Other Organizations - OTH, see Fig. 3.17a. and 3.17b.)
- info on whether organizations are SMEs[7] or not
- organization type (participant / coordinator / third party)
- total project budget (in EUR)
- budget share per project member

Since stakeholder geolocations were provided no geocoding of address information was needed to position project members on the map. This offered the possibility to assign each geolocation to their containing NUTS region. To this end and further geographical investigations, the *geopandas 0.8.0* python package [R88] was used in a Google Colab environment.




**(a)** **(b)**

**Figure 3.17:** Geographical distribution of partaking organisations within H2020 research framework program together with their organisation types on individual and aggregated NUTS3 level.

Having built up our database with unique organisation- and project IDs, firstly the assignment of the NUTS regions to each of the stakeholders have been done. As a next step,

[7] Small- and Medium sized Enterprise

the connections formed among stakeholders have been counted. This was later aggregated on regional level and investigated on an annual basis throughout the whole time period of the accessible data. Two-dimensional visualisation of the connection numbers as a function of the selected regional attributes has been performed in order to reveal possible anomalies and discrepancies among member state regions connected to the EU before 2004 and afterwards (see Fig. 3.18.). The gained data distributions were highly skewed, therefore log-log plots have been used.

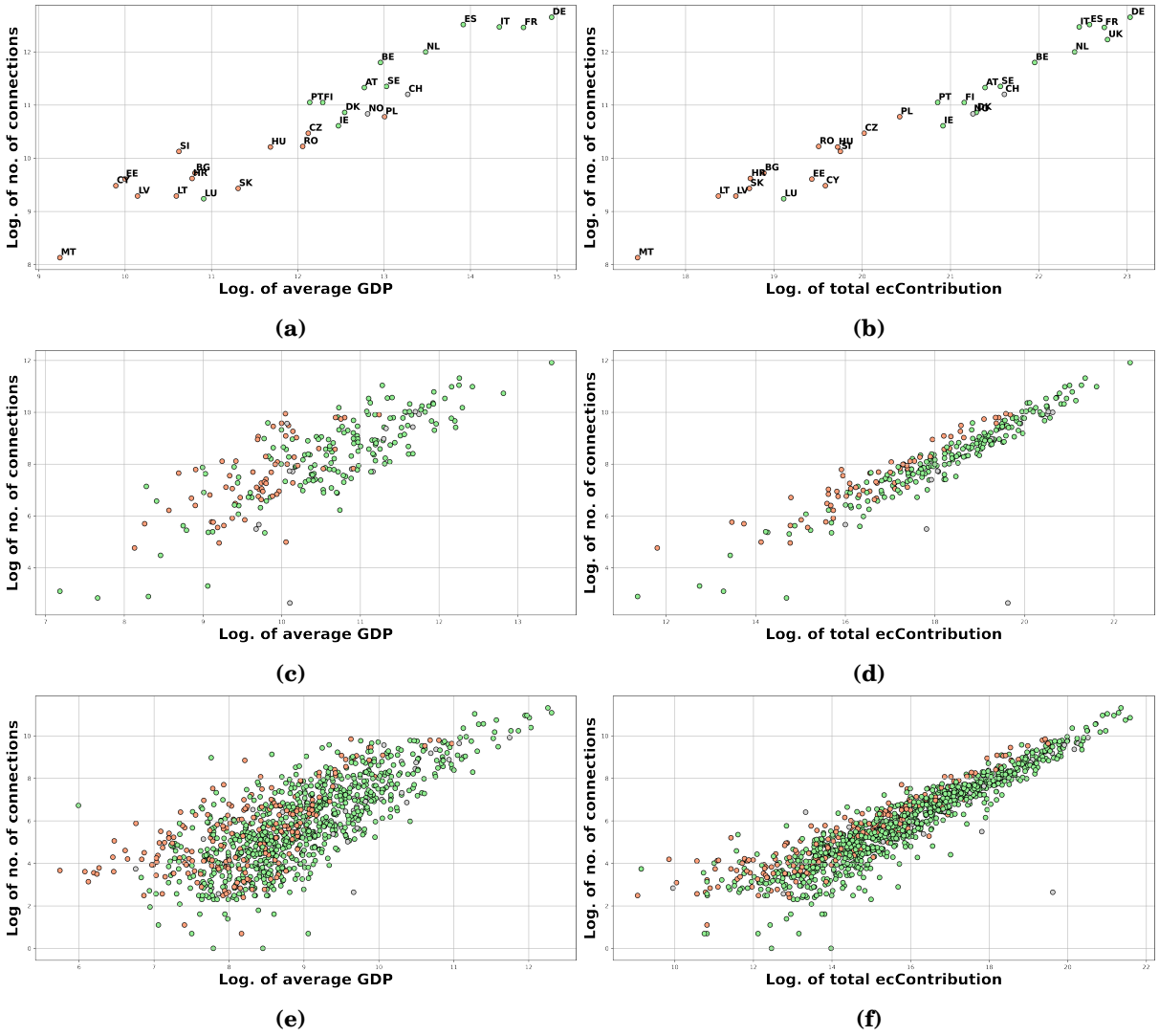



**Figure 3.18:** Log-log plot of selected regional attributes and total formed connection numbers generated within the investigated funding period for country-, NUTS2- and NUTS3 levels in each row. Green dots correspond to regions connected to EU before 2004, while the orange ones mark those that connected after 2004.

From the labelled project-connections, connection matrices have been created, upon which regional level social network analysis can rely. The investigation of the resulted social networks was carried out by the generation of graph centrality measures (see Table 3.12.) and general graph attributes (see e.g.: [R89], [R90]) like:

- Graph degree centrality distribution that represents the fraction of nodes connected to each node. The annual average ($\overline{Deg\_c}$) and median ($\widetilde{Deg\_c}$) values were computed.
- Graph eigenvector centrality distribution that represents the influence of each node to the network. The annual average ($\overline{Eig\_c}$) and median ($\widetilde{Eig\_c}$) values were computed.

- Graph closeness centrality distribution that represents centrality tendencies within the network. The annual average ($\overline{Clo\_c}$) and median ($\widetilde{Clo\_c}$) values were computed.
- Betweenness centrality distribution that represents centrality tendencies via viewing information flow bottlenecks. The annual average ($\overline{Betw\_c}$) and median ($\widetilde{Betw\_c}$) values were computed.
- Density, that gives how many edges are present compared to a fully connected graph.
- Diameter, that is based on the maximal distance among nodes measured in edge number.
- Transitivity, that represents the fraction of closed clubs in networks by calculating the number of triangles.
- Average shortest path length ($\overline{SPL}$) that characterizes the compactness of a network.
- Number of nodes and edges that characterizes the size of the network.

| | $\overline{\text{Deg\_c}}$ | $\widetilde{\text{Deg\_c}}$ | $\overline{\text{Eig\_c}}$ | $\widetilde{\text{Eig\_c}}$ | $\overline{\text{Clo\_c}}$ | $\widetilde{\text{Clo\_c}}$ | $\overline{\text{Betw\_c}}$ | $\widetilde{\text{Betw\_c}}$ | density | diam. | trans. | $\overline{\text{SPL}}$ | nodes | edges |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Country level** | | | | | | | | | | | | | |
| 2014 | 12.76 | 8.32 | 0.22 | 0.18 | 0.59 | 0.58 | 0.10 | 0.01 | 0.30 | 3 | 0.43 | 1.76 | 12 | 20 |
| 2015 | 25.02 | 26.75 | 0.27 | 0.32 | 0.73 | 0.72 | 0.10 | 0.01 | 0.58 | 2 | 0.77 | 1.42 | 10 | 26 |
| 2016 | 25.86 | 19.51 | 0.24 | 0.21 | 0.70 | 0.65 | 0.08 | 0.03 | 0.52 | 2 | 0.69 | 1.48 | 12 | 34 |
| 2017 | 26.13 | 25.09 | 0.24 | 0.25 | 0.68 | 0.67 | 0.10 | 0.05 | 0.50 | 3 | 0.68 | 1.52 | 12 | 33 |
| 2018 | 29.61 | 28.24 | 0.29 | 0.29 | 0.80 | 0.76 | 0.07 | 0.06 | 0.71 | 2 | 0.75 | 1.29 | 10 | 32 |
| 2019 | 24.63 | 23.50 | 0.24 | 0.26 | 0.68 | 0.65 | 0.08 | 0.04 | 0.48 | 3 | 0.65 | 1.53 | 12 | 32 |
| 2020 | 25.66 | 21.71 | 0.26 | 0.24 | 0.72 | 0.67 | 0.07 | 0.02 | 0.56 | 3 | 0.69 | 1.45 | 11 | 31 |
| 2021 | 23.70 | 20.08 | 0.24 | 0.24 | 0.69 | 0.65 | 0.07 | 0.05 | 0.50 | 3 | 0.64 | 1.52 | 12 | 33 |
| 2022 | 21.67 | 14.84 | 0.21 | 0.17 | 0.61 | 0.61 | 0.08 | 0.00 | 0.36 | 3 | 0.60 | 1.71 | 15 | 38 |
| | **NUTS2 level** | | | | | | | | | | | | | |
| 2014 | 14.78 | 5.64 | 0.11 | 0.07 | 0.50 | 0.50 | 0.03 | 0.00 | 0.12 | 4 | 0.27 | 2.03 | 39 | 91 |
| 2015 | 11.69 | 3.46 | 0.11 | 0.06 | 0.53 | 0.51 | 0.03 | 0.00 | 0.09 | 2 | 0.15 | 1.91 | 38 | 63 |
| 2016 | 11.89 | 4.65 | 0.12 | 0.08 | 0.54 | 0.52 | 0.03 | 0.00 | 0.11 | 2 | 0.16 | 1.89 | 35 | 63 |
| 2017 | 12.51 | 3.85 | 0.10 | 0.06 | 0.52 | 0.51 | 0.02 | 0.00 | 0.07 | 2 | 0.13 | 1.93 | 51 | 93 |
| 2018 | 12.42 | 3.64 | 0.10 | 0.06 | 0.53 | 0.51 | 0.02 | 0.00 | 0.08 | 2 | 0.17 | 1.92 | 44 | 79 |
| 2019 | 10.87 | 3.76 | 0.12 | 0.07 | 0.53 | 0.51 | 0.03 | 0.00 | 0.09 | 2 | 0.14 | 1.91 | 36 | 57 |
| 2020 | 12.08 | 3.50 | 0.12 | 0.06 | 0.54 | 0.51 | 0.03 | 0.00 | 0.11 | 2 | 0.22 | 1.89 | 34 | 61 |
| 2021 | 12.11 | 2.91 | 0.10 | 0.04 | 0.50 | 0.48 | 0.03 | 0.00 | 0.09 | 4 | 0.21 | 2.04 | 43 | 80 |
| 2022 | 12.83 | 5.45 | 0.10 | 0.06 | 0.44 | 0.44 | 0.03 | 0.00 | 0.08 | 5 | 0.19 | 2.37 | 51 | 96 |
| | **NUTS3 level** | | | | | | | | | | | | | |
| 2014 | 14.27 | 4.77 | 0.03 | 0.01 | 1.00 | 1.00 | - | - | 0.11 | 3 | 0.17 | 1.97 | 36 | 71 |
| 2015 | 8.89 | 1.91 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.22 | 4 | 0.41 | 1.96 | 28 | 83 |
| 2016 | 8.56 | 1.76 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.14 | 3 | 0.28 | 1.94 | 34 | 80 |
| 2017 | 9.39 | 2.31 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.15 | 4 | 0.36 | 2.12 | 35 | 90 |
| 2018 | 11.87 | 2.52 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.11 | 3 | 0.27 | 2.11 | 49 | 129 |
| 2019 | 8.69 | 1.82 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.12 | 3 | 0.28 | 2.03 | 42 | 101 |
| 2020 | 11.41 | 2.04 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.12 | 3 | 0.28 | 2.02 | 47 | 130 |
| 2021 | 10.56 | 2.37 | 0.01 | 0.00 | 1.00 | 1.00 | - | - | 0.11 | 4 | 0.24 | 2.16 | 47 | 114 |
| 2022 | 16.51 | 9.81 | 0.04 | 0.02 | 1.00 | 1.00 | - | - | 0.06 | 6 | 0.22 | 3.07 | 58 | 99 |

**Table 3.12:** Annual graph centrality measures and basic descriptive attributes on different regional levels.

Whereas between two regions several connections are possible and within-region connections are also omnipresent, in the following graphs are filtered to the 5 most important connections per region in order to gain a comprehensible network and the extraction of as much information as possible but still keep a satisfactory transparency. Thereby, asymmetric connection matrices were considered[8] for building up undirected graphs. The number of connections among the same regions were understood as a measure of "connection strength" and were used as weights for calculating graph attributes listed above. For the calculations

[8]An "important connection" of a region to another one does not necessitate that this connection is as important for the second one conversely.

the *networkx 2.8.4.* python package had been utilized within an *Anaconda* framework designed to support data scientific calculations [R91, R92].

The gathered total connection numbers throughout the total funding period were viewed together with regional mean GDP data and the total EU contribution to the regions via each project. Each log-log plot of Fig. 3.18. shows correlation among the variables concerned, however in different extent. On country level the member states connected after 2004 can easily be separated from the countries connected before, only Luxembourg is embedded in the cluster of East-European countries. This shows that West European countries regardless of size constitute a larger scientific centre of gravity and could acquire more social capital from within the H2020 projects in proportion to their population and GDP and could receive more EU funding as well.

As a non-trivial observation, it can be remarked that on NUTS2 and NUTS3 level, the East-European countries could gain more connections at the same GDP levels (see Fig. 3.18c. and 3.18e.) which also hold for the same *ecContribution* levels (see Fig. 3.18d. and 3.18f.). This is joyful for the countries and regions on the uptake, since connections are expected to form an inevitable basis for future collaborations and anticipates higher dynamics of increase in scientific project involvement and funding received.

Fig. 3.19. illustrates the resulted H2020 connection network on NUTS2 level. The red lines with different line widths indicate the strength of the ”among-region” connections, while the green colouring of the regions the strength of the ”within-region” connections on a square-root scale (for a better visualisation of highly skewed distributions).

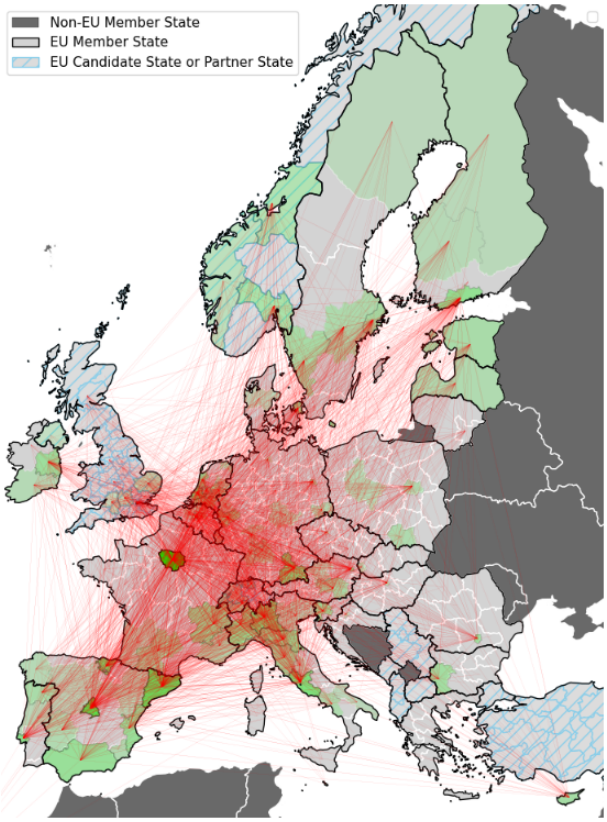



**Figure 3.19:** Connection intensities on NUTS2 level with a square root scale for colouring and line widths.

Figures on NUTS2 and NUTS3 levels indicate that some regions dominate the network as strongly-connected points. This was also expected from a network that is equipped with scale-free attributes. The same procedure was performed on country level as well that showed outstanding connection strength within and among Italy, France, Germany, Spain

and the United Kingdom (IT, FR, DE, ES, UK cluster as central region on country level). As non-trivial relatively strong within-country connection intensity could be observed in case of Bulgaria and among-country connections between Finland, Austria and the Czech-Republic with the (IT, FR, DE, ES, UK) group.

Further detailed investigation of the formed connections aggregated on regional level was done by selecting the 5 most important connections of each region in order to make the resulting network clear-cut. Furthermore, for the same purpose within the constructed weighted network, where the weights were constituted by the number of connections among nodes only connections with weights $> 2$ were taken into account as a practically arising threshold in order to gain reasonably understandable transparent but still connected networks.

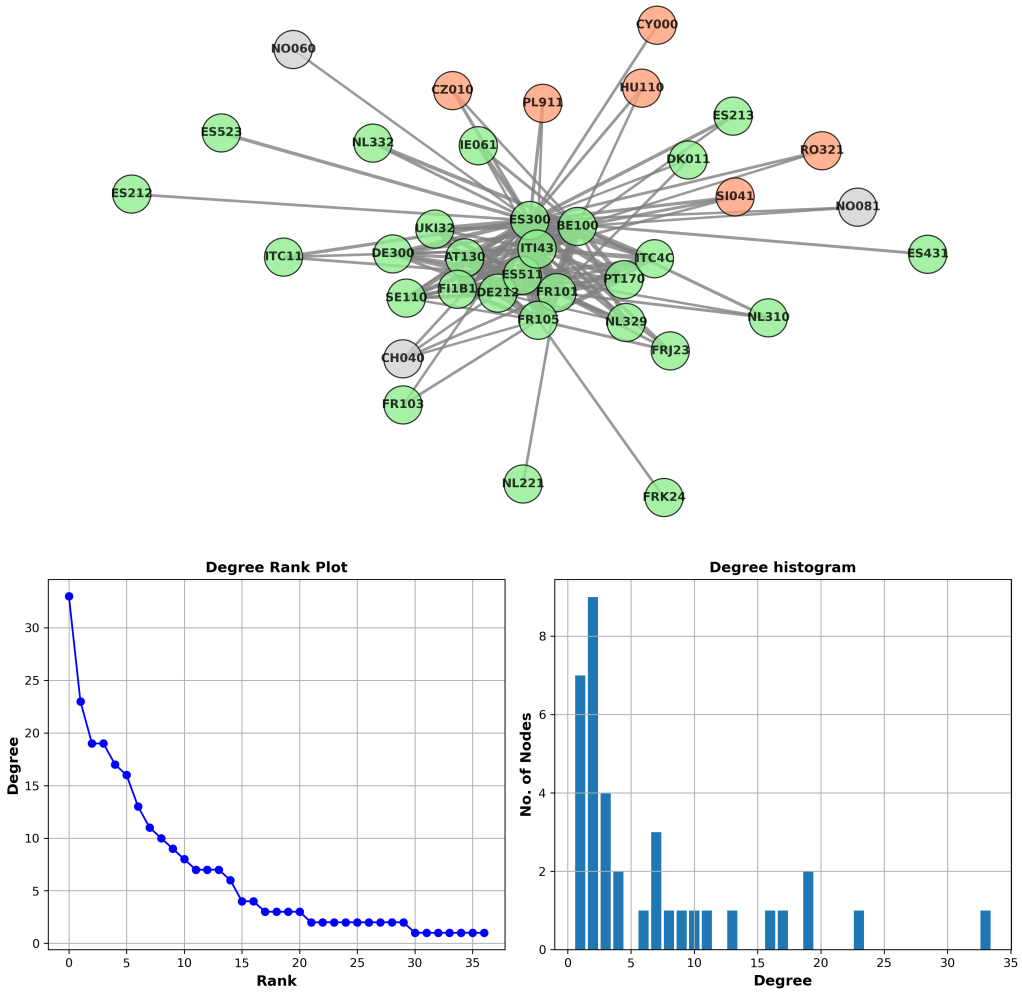



**Figure 3.20:** Social network of NUTS3 regions when most important connections are considered together with corresponding graph degree distributions.

Besides geographical investigations, the connections of regions representing central players can be investigated in detail according to Fig. 3.20. where the web of regions is illustrated with their degree distributions. This latter indicates a power law distribution, however due to low node numbers the MLE estimation of this graph would be quite uncertain, and it is of less importance, since the social network is already filtered to the most relevant connections only. Similarly to the NUTS3 case, for NUTS2 regions a highly interconnected group of nodes is to be seen. These nodes also arise mainly from capital regions of member states connected to the EU before 2004 and other nodes are attached to

this interconnected cluster from capital regions of other countries or from other developed West-European territories (regions of Paris, Madrid, Barcelona, Bilbao, Rome, Munich etc.).

The calculated annual graph attributes and centrality measures listed in Table 3.12. are provided for Country-, NUTS2 and NUTS3 levels, however for NUTS3 level due to the high interconnection of nodes some of the centrality measures are not applicable. On the other hand data from 2022 are not representative, since the H2020 period had completely ended by then, therefore for the investigation of tendencies that year should be excluded and regarded only as informative.

The results show that on the whole the number of nodes and edges for the region-level social network of most relevant connections had an increasing tendency together with the average shortest path length ($\overline{SPL}$), average- and median degree centrality and graph diameter, furthermore a slightly decreasing closeness of centrality by the end of the H2020 funding period. These data indicate that the size of the connection network of most relevant connections had an increasing trend but showed a tendency towards less centralization as well.

# 4

# MFV-based Robust Outlier Detection

## Thesis Group 2: Robust outlier detection based on the MFV concept and its application regarding economic convergence

***Thesis 2***
**I have developed a clustering algorithm based on the MFV concept and performed robust outlier detection in case of linear regression problem with application area of the economic convergence of EU regions. I have developed a methodology to identify regions in a robust way that show faster or slower economic convergence than the bulk of the regions suggested by the absolute economic $\beta$-convergence theorem.**

Publications relevant to the theses: [T3, T4, T6, T8].

***Thesis 2.1***
**I have developed a method to identify outliers based on the MFV concept compared to the MFV-robustified linear regression. Regarding the absolute economic $\beta$-convergence of EU member states and regions, I have identified regions converging faster or slower than the bulk of the data suggested by the fitted robust trend line.**

Publications relevant to the theses: [T3, T4].

***Thesis 2.2***
**Based on the Minimum Covariant Determinant estimator, I have shown by the usage of robust Mahalanobis distances that regions of EU member states connected to the EU before 2004 and after are forming two separate groups regarding H2020 financial support. Furthermore, I have demonstrated the applicability of the method in identifying outlying regions compared to the bulk of the data when regional GDP, population and R&D funding is considered.**

Publications relevant to the theses: [T8].

***Thesis 2.3***
**I have developed a clustering algorithm by using robust location- and scale parameters based on the MFV concept that considers outliers compared to the bulk of the data but attenuates their biasing effect. The developed method can perform similarly or better regarding computation time measured in the number of centroid swaps in case of larger sample sizes simultaneously with more clusters to be identified than the robust k-Medians.**

Publications relevant to the theses: [T6].

The present chapter details findings regarding robust outlier detection combined with linear regression and k-Means clustering methods. For the linear regression case, the economic convergence of EU countries and regions will be used as an application area. The background of economic absolute $\beta$-convergence and MFV-based linear regression is discussed in Chapter 3. Further findings within a holistic perspective will be provided based on H2020 project information where robust Mahalanobis distances are utilized for outlier detection. Clustering related investigations will be demonstrated on sample datasets.

## 4.1 Overview of Related Literature

In several real-life scenarios measurements are run, parameters are monitored, data are collected where measurement errors are omnipresent or barely avoidable. To characterize "outlyingness" often marginal distributions are used separately for each variable observed. This is a viable concept and practically the simplest and most straightforward choice if we consider serial production sites or simple blood sample laboratory reports etc. For multivariate investigations, however, combinations of original variables are considered (e.g.: linear combinations for orthogonal projections etc.). Consequently, for multivariate models, we mix outliers and other anomalies (e.g.: missing values) into all projected and combined data points even before data cleaning or any downweighting procedures. Furthermore, outlyingness exists on the multivariate level as well. For the example of serial production this is of high interest, since parts can be judged from any univariate aspect as non-outlier, however can still cause or lead to errors in the application field that might be non-trivial to trace back. In this multivariate context, outliers might easily remain invisible when they are viewed variable-wise [R3].

It has to be emphasized though that a-typical observations can be identified only compared to an existing model, which is of course generated by incorporating those elements that presumably cause the distorting effects. In case of a regression setting, a single observation can be identified as an outlier with respect to a linear regression model, but for instance with respect to a nonlinear regression model the same observation will not appear as anomalous. On the other hand, in many situations such observations cannot be attributed to measurement errors and may hold invaluable information on the underlying processes that should not be overlooked by eliminating them. Robust procedures do not eliminate these items, since the weighting of instances only reduces the often dramatic impact of a-typical observations on conventional statistical procedures. The cost of this favourable characteristic is the increased computational time caused by iterative algorithms, therefore besides the selection of robust statistical alternatives a careful choice of numerical approaches and initialization are also of great importance [R43].

The aim of outlier identification can be of multiple purposes. Outliers can often be distorting elements, and the behaviour of the majority of data is of interest. Nonetheless, the same outliers can also be of special interest and their detection might reveal processes

that are to be eliminated or shall be explored. On the other hand, albeit outliers might draw attention to malfunctions or processes towards unwanted states, they are not necessary influential to the same extent to the estimator at hand to be computed. The impact on the sought estimator highly depends on the distribution of the sample and the model specification as well, that makes the definition of outliers in the given context precise (e.g.: linear regression). In case of other applications however, (e.g.: clustering) outliers cannot be determined like that. Therefore, they are sometimes characterized as observations that behave significantly differently compared to the majority of the remaining observations. Their isolation can be tricky, since due to masking and swamping effects, outliers might remain hidden. Furthermore, heavy-tailed error distributions, autocorrelated- and heteroscedastic error terms can cause similar biases for statistical inferences as well [R3].

Specifically, in case of clustering we speak of an unsupervised ordering of unlabelled data into separable, relatively homogeneous groups, which process is often used in itself or as an intermediate step of data pre-processing and for the understanding of the structure of the underlying sample [R93]. A major goal of clustering is to split big data sets into smaller, less heterogenous segments based on similarity features that can be further investigated in smaller scale [R94]. Thereby, the hidden information can be broken down into smaller units that are easier to interpret [R95, R96].

Outliers, noises and non-sharp cluster borders may however pose challenges to many clustering algorithms. Therefore, robustification of algorithms when working with real-life data is of particular importance in order to be able to stabilize efficiency and predictive power [R97]. Nonetheless, robustness and stability depend not just on the underlying data then on the clusters themselves. Outliers often form heterogeneous groups with the bulk of the data, therefore clustering can theoretically isolate them [R98, R42, R99]. However, it is not necessarily worth to seek outliers this way, since noisy observations can lead to ”chaining effects” via ”bridging points” that can result in the density-based interconnection of different data groups and low breakdown points – in the present case the percentage amount of outlier points that leads to an unacceptable classification of the data points – of clustering algorithms (e.g.: k-means) [R100].

During clustering, outliers and separate groups can be identified without any prior knowledge about the data. Many times different approaches are combined in literature (e.g.: partition based-, hierarchical-, density based-, grid based- or model based techniques) in order to reach higher efficiency, but there are limitations of validity depending on size- and nature of data, number of dimensions, data distributions etc. [R99]. Thereby, selection of appropriate clustering algorithm shall be done accordingly and by no means automatically [R101].

Authors of related works often build upon robust approaches like k-Medoids / k-Medians that enjoy widespread popularity in the literature [R102, R103, R104, R105, R106, R98, R107, R108, R109]. In general, these show higher accuracy than the well-known k-Means, but their run-time increase fast with sample size that makes them unfavourable in case of bigger problems [R110, R101]. On the other hand, robust statistical methods are gaining more attention, which instead of only focusing on robust location parameters try to rely on the ”bulk” of the data by performing adequate trimming or suppressing of ”far-lying” observations [R97, R100, R111, R112]. A great advantage of latter approaches is that they enable higher-dimensional investigations as well and can be extended to Principal Component Analysis problems or multivariate outlier detection tasks [R113, R114, R115].

## 4.2 Data Analysis

The present chapter, regarding MFV-based robust outlier detection in case of the absolute economic $\beta$-convergence, relies on the same data sets that were introduced in Chapter 3.,

Sec. 3.3., where NUTS regional GDP and NDI information are provided. Additionally, Horizon 2020 project information on R&D funding are also investigated together with regional GDP and population information that are presented in Sec. 3.5. of the same chapter.

In relation to MFV-based robust outlier detection in case of cluster analysis, sample datasets have been selected. 5 different, real-life datasets were accessed from the UCI database (UC Irvine Machine Learning Repository) [R116]. Datasets equipped with known classification and of relatively small cluster sizes have been selected due to interpretation purposes. Different sample sizes and different feature numbers were also considered for the investigation, in parallel to other similar literature sources [R117]. The main characteristics of the data selected are listed in Table 4.1.

| Dataset name | Sample size | No. of features | No. of clusters | Distribution per cluster |
|---|---|---|---|---|
| Long Jump | 58 | 1 | 2 | 33-25 |
| Iris | 150 | 4 | 3 | 50-50-50 |
| Wine | 178 | 13 | 3 | 71-59-48 |
| Ecoli | 336 | 7 | 4 | 143-116-52-25 |
| Breast cancer | 569 | 30 | 2 | 357-212 |

**Table 4.1:** Descriptive information on datasets accessed for robust clustering related investigations.

## 4.3 Robust Outlier Detection

Throughout practical data analyses of real-life and presumably contaminated data, besides specifying the "bulk" of the data, the identification and classification of outliers (e.g.: compared to a robust and resistant regression line) is of great interest. Therefore, a definition for "outlyingness" is required, that in the following is defined based on the MFV concept. In order to generate comparable results of data with Gaussian error distribution, we use the dihesion as a consistent estimator of the standard deviation that can be calculated in an iterative manner according to Eq. 2.7. and Eq. 2.8. In case the amount of data around the $\mu$ mean value is covered by $\pm\epsilon$ distance is known (let us indicate this portion by "$R$") the probability of observing data within this distance can be expressed as:

$$\mathbb{P}(|x-\mu| \leq \epsilon) = \mathbb{P}\Big(\Big|\frac{x-\mu}{\sigma}\Big| \leq \frac{\epsilon}{\sigma}\Big) = R \tag{4.1}$$

Therefore, for normally distributed data we must have:

$$\Phi\Big(\frac{\epsilon}{\sigma}\Big) - \Phi\Big(-\frac{\epsilon}{\sigma}\Big) = R \tag{4.2}$$

$$\Phi\Big(-\frac{\epsilon}{\sigma}\Big) = 1 - \Phi\Big(\frac{\epsilon}{\sigma}\Big) \tag{4.3}$$

Summing Eq.4.2. and 4.3 the relationship between the dihesion and standard deviation results to be:

$$\epsilon = \Phi^{-1}\Big(\frac{R+1}{2}\Big) \cdot \sigma \tag{4.4}$$

Thus, the estimate for the standard deviation can be calculated as:

$$\hat{\sigma} = A \cdot \epsilon = \left(\Phi^{-1}\left(\frac{R+1}{2}\right)\right)^{-1} \cdot \epsilon, \tag{4.5}$$

where "$A$" denotes a constant distribution dependent scale factor. With a consistent estimate for the characterization of far-lying data points, the recommendations of [R118, R119, R120] are followed that considers an observation as an outlier if one of the following selected criteria is met:

$$\left|\frac{x_i-\mu}{\sigma}\right| \geq 3 \Longrightarrow \text{ Very conservative (less than 1\% of the data)}$$

$$\left|\frac{x_i-\mu}{\sigma}\right| \geq 2.5 \Longrightarrow \text{ Moderately conservative (compromise)}$$

$$\left|\frac{x_i-\mu}{\sigma}\right| \geq 2 \Longrightarrow \text{ Poorly conservative (less than 5\% of the data)}$$

These criteria can be applied appropriately to linear regression problems what will be present in Sec. 4.4. where the MFV-robustified linear regression line will be taken as representative of the data condensation. As an arbitrary selection, in the further investigations the outliers will be classified according to the moderately conservative approach, i.e.:

$$\textbf{Weak outlier: } |x_i - M(k,x)| > \epsilon \text{ and } |x_i - M(k,x)| \leq 2.5 \cdot A \cdot \epsilon$$

$$\textbf{Strong outlier: } |x_i - M(k,x)| > 2.5 \cdot A \cdot \epsilon$$

## 4.4 Application to Economic Convergence of EU Regions

According to the iterative parameter estimation procedure outlined in Chapter 3., Sec. 3.2., slope- ($a_{MFV}, \tilde{a}_{MFV}$), intercept- ($b_{MFV}, \tilde{b}_{MFV}$) and dihesion ($\epsilon_{MFV}, \tilde{\epsilon}_{MFV}$) values have been computed for the means of annual growth rates and MFVs of annual growth rates of country level and regional GDP and NDI measures respectively with the MFV-based robust linear regression method. For comparative purposes, the slope- and intercept parameters have also been calculated for linear regression based on the minimization of the $L^2$-norm by using the ordinary least squares method ($a_{OLS}, \tilde{a}_{OLS}$) as well. Besides parameters of linear regression, the number of "MFV-iterations" necessary to reach the specified exit criteria for convergence ($n, \tilde{n}$), the ratio of data lying within a "one-dihesion-distance" measured from the fitted line ($R, \tilde{R}$) and scale factors ($A, \tilde{A}$) in order to be able to use the resulted dihesion values as consistent estimators of the standard deviations have also been given in Table 4.2.

For characterizing the rate of convergence among the investigated spatial entities within the framework of economic absolute $\beta$-convergence, the slope parameters of the fitted lines have to be used. Negative slopes with larger absolute values correspond to faster convergence. As can be seen from the slope parameters listed in Table 4.2., investigations performed on means of annual growth rates resulted similar or even larger convergence tendency on country level while smaller convergence for other regional levels except NUTS2 level for GDP [PPS per inhabitant] data. In case of investigating MFVs of annual growth rates, the MFV slope parameters followed the same tendency compared to the slope parameters fitted by the ordinary least squares method. In case of countries, however, it might be assumed that highly aggregated and averaged data into less than 30 data points can be uncertain or less accurate. For NUTS2 or NUTS3 regions, much more data are at hand that provides more trustworthiness. For the accessed data with more regional instances and consequently higher territorial resolution, the absolute $\beta$-convergence theorem based on MFV-robustified line regression (except NUTS2 level for GDP [PPS per inhabitant]) served with a conclusion that a less exaggerated convergence among the EU regions shall be expected compared to what are provided by those analyses that are based on conventional statistical procedures.

| | Country | | | | NUTS2 | | | | NUTS3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GDP | | NDI | | GDP | | NDI | | GDP | |
| | EUR | PPS | EUR | PPS | EUR | PPS | EUR | PPS | EUR | PPS |
| $a_{OLS}$ | -1.5821 | -1.9679 | -2.8551 | -3.5044 | -2.2468 | -1.5868 | -2.5856 | -3.0220 | -2.1345 | -2.0112 |
| $b_{OLS}$ | 18.5016 | 9.4796 | 29.3284 | 34.9894 | 25.0051 | 7.1442 | 26.5802 | 30.4550 | 23.7478 | 22.4653 |
| $a_{MFV}$ | -1.7910 | -2.5753 | -2.9047 | -3.2907 | -1.9052 | -1.7265 | -2.1025 | -2.5921 | -1.8082 | -1.7766 |
| $b_{MFV}$ | 20.6659 | 12.1594 | 29.9497 | 33.0436 | 21.6766 | 7.8563 | 22.2062 | 26.5490 | 20.5669 | 20.1701 |
| $\epsilon_{MFV}$ | 1.0033 | 0.8793 | 1.1449 | 0.5579 | 0.6817 | 1.0691 | 0.4238 | 0.5315 | 0.7249 | 0.7196 |
| $n$ | 39 | 40 | 20 | 29 | 25 | 18 | 33 | 19 | 27 | 25 |
| $R$ | 0.5556 | 0.4815 | 0.7308 | 0.6154 | 0.5446 | 0.6573 | 0.5000 | 0.6387 | 0.5657 | 0.5750 |
| $A$ | 1.3077 | 1.5489 | 0.9051 | 1.1502 | 1.3397 | 1.0540 | 1.4826 | 1.0955 | 1.2791 | 1.2534 |
| $\tilde{a}_{OLS}$ | -1.7455 | -2.3188 | -2.5445 | -3.5144 | -2.1727 | -1.7539 | -2.4192 | -2.7915 | -2.0862 | -1.9320 |
| $\tilde{b}_{OLS}$ | 20.2831 | 10.9508 | 26.5844 | 35.2308 | 24.4841 | 7.9164 | 25.1065 | 28.4242 | 23.3954 | 21.8306 |
| $\tilde{a}_{MFV}$ | -1.6873 | -2.6208 | -2.5603 | -2.9129 | -2.0484 | -1.8393 | -2.1952 | -2.3504 | -1.9138 | -1.6870 |
| $\tilde{b}_{MFV}$ | 19.7949 | 12.2610 | 26.6369 | 29.5613 | 23.2992 | 8.3528 | 23.1009 | 24.3534 | 21.7506 | 19.4399 |
| $\tilde{\epsilon}_{MFV}$ | 0.4868 | 0.8331 | 0.8321 | 0.2868 | 0.6215 | 0.9147 | 0.4701 | 0.4635 | 0.7956 | 0.7546 |
| $\tilde{n}$ | 80 | 30 | 18 | 32 | 29 | 22 | 27 | 27 | 24 | 26 |
| $\tilde{R}$ | 0.4074 | 0.5926 | 0.7308 | 0.4615 | 0.5399 | 0.6291 | 0.5588 | 0.5588 | 0.5854 | 0.5619 |
| $\tilde{A}$ | 1.8689 | 1.2071 | 0.9051 | 1.6256 | 1.3537 | 1.1176 | 1.2984 | 1.2984 | 1.2258 | 1.2896 |

**Table 4.2:** Estimated model parameters on Country-, NUTS2- and NUTS3 region level for GDP and NDI data with dimensions of [EUR per inhabitant] or [PPS per inhabitant]. Parameters with tilde stand for calculations performed on MFVs of annual growth rates.

According to Table 4.2., the slopes of the fitted MFV-robustified regression lines indicate diverse differences when comparing results of fitted lines on mean- and MFV values of annual growth rates. It cannot be univocally stated that MFVs of annual growth rates would bring results on stronger convergence of the investigated regions, although in most of the cases for NUTS2- and NUTS3 levels slightly higher slopes were calculated in absolute value. However, observing thoroughly the distribution of the data points in each case differences among the weak- and strong outliers turned out to be more relevant.

The identification of outliers has been done compared to the fitted MFV-robustified regression line. Data points lying within one-dihesion-distance constituted the ”bulk” of the data. Instances lying within one-dihesion-distance and the distance specified by the scale parameter times the corresponding dihesion value (see Table 4.2.) were labelled as weakly outlying, while those that can be found further from the fitted MFV regression line than this distance were labelled as strongly outlying points (see Fig. 4.1. and Fig 4.2.) in accordance with Sec. 4.3.

The more robust and outlier resistant a linear regression technique is, the identification of ”interesting” objects becomes possible that would have otherwise been masked by the inflated variance of the data. In our case, for linear regression based on the MFV concept, not just the less expressed speed of economic convergence within the framework of absolute $\beta$-convergence was pointed out, but the over- or under-performing regions within the past two decades – in terms of the convergence theorem – were directly specified and labelled as well. These regions are visualised for NUTS2 and NUTS3 levels for each investigated case on Fig. 4.3., 4.4. and 4.5.

In case of some countries, strikingly different outliers occurred. For GDP in EUR per capita values on NUTS2 level differences for the case of Sweden and Poland seems to be the most prominent, while for NDI likewise on NUTS2 level and in EUR per capita dimension French regions gained highly differing labels that might be surprising. Nevertheless, the provided geographical visualisations can be further utilized as a basis for better comparison and the gained classification of regions in general as input for further field relevant researches that are at present out of the scope of the present thesis from methodological point of view.

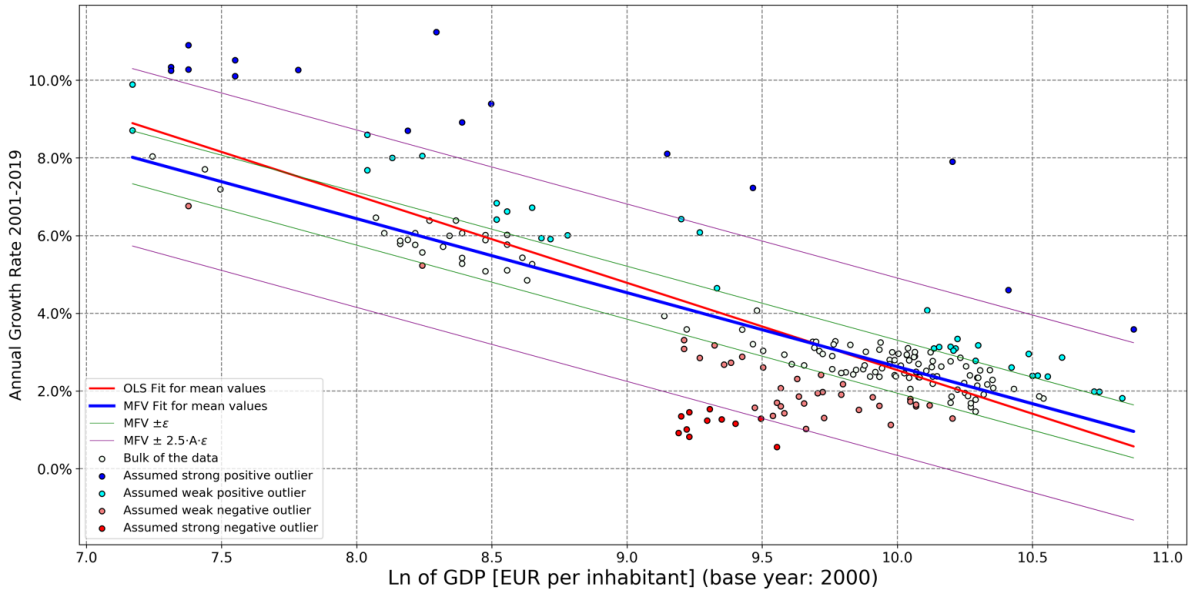



**Figure 4.1:** Fitted MFV regression line with corresponding classification of outliers and OLS line fit for comparative purposes (dataset: NUTS2 level, GDP [EUR per inhabitant].

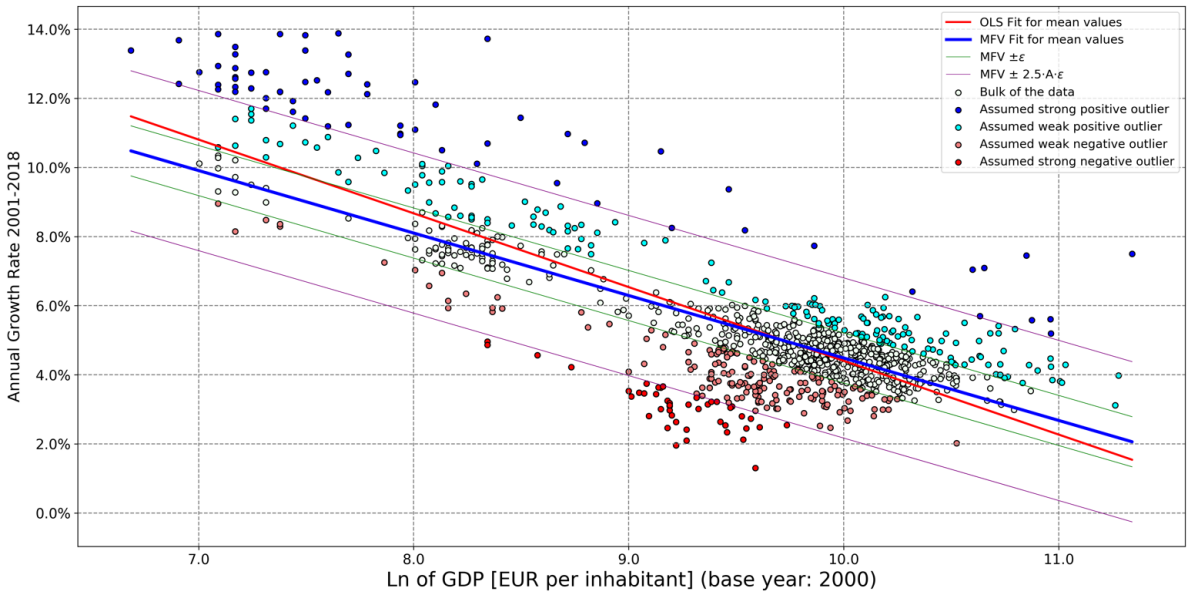
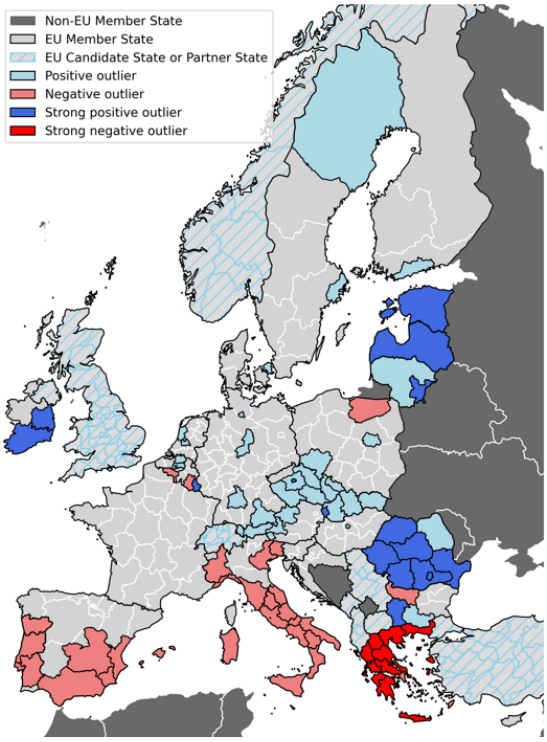
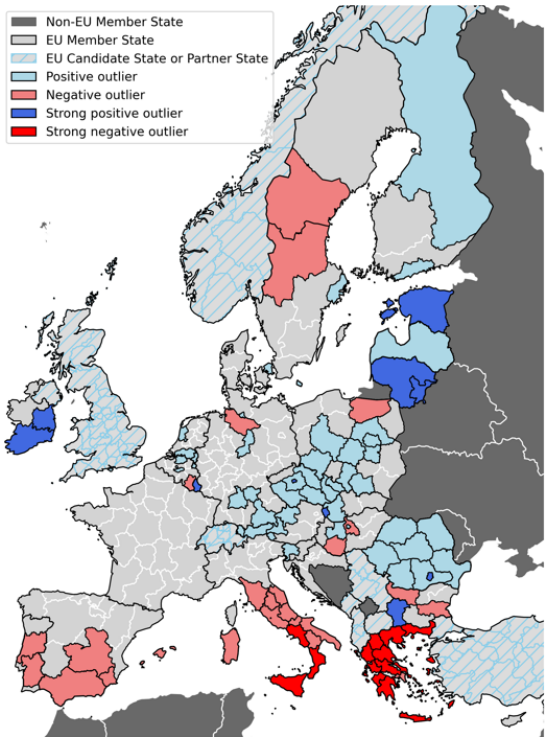



**Figure 4.2:** Fitted MFV regression line with corresponding classification of outliers and OLS line fit for comparative purposes (dataset: NUTS3 level, GDP [EUR per inhabitant].

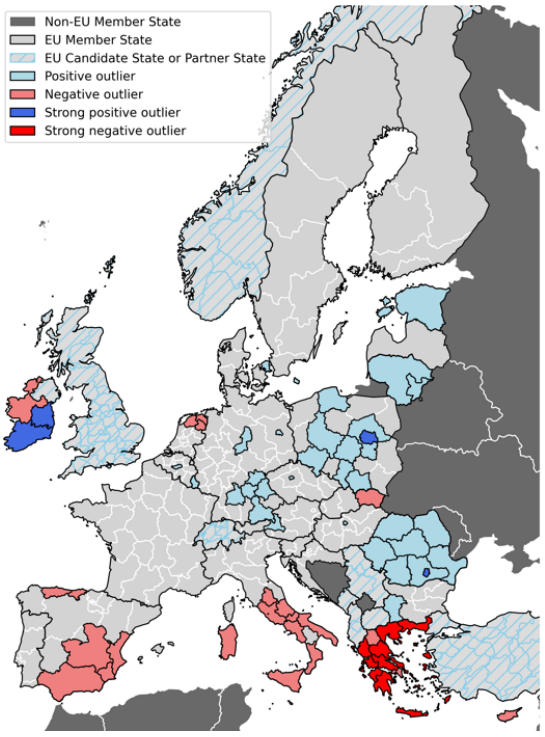
As a holistic approach, similar to the investigations on economic convergence outlined in Chapter 3.3., Sec. 3.5., data on Horizon 2020 R&D project information were also considered, where funding contributions provided by the European Union were combined with regional population-, density of population- and GDP data with the aim of exploring regions with outlying positioning compared with the ”bulk” of the data and observing possible clustering phenomena. Since all the accessed data by nature showed strongly skewed distributions to the left, the natural logarithms were used in order to stabilise standard deviations and obtain uni- and multivariate distributions that are ”closer” to Gaussian. Contributions provided by the European Community (marked as *ecContribution* on Fig. 4.6., 4.7. and 4.8. in accordance with database notations) were aggregated to each region throughout the whole funding period, while the associated regional data were time-averaged since these



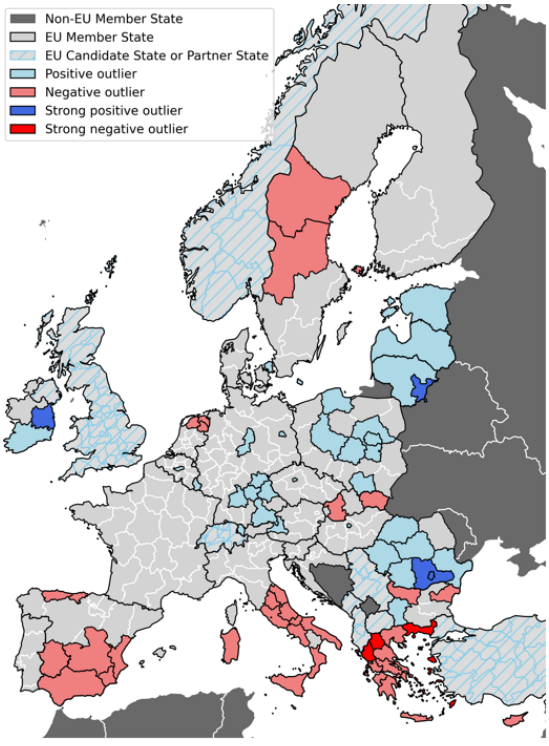

**(a)** GDP [EUR per inhabitant] - Fit on Mean values




**(b)** GDP [EUR per inhabitant] - Fit on MFV values

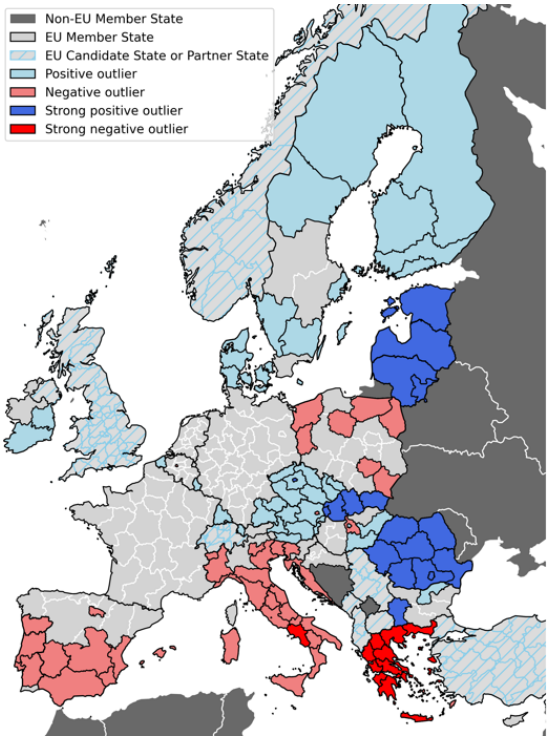



**(c)** GDP [PPS per inhabitant] - Fit on Mean values

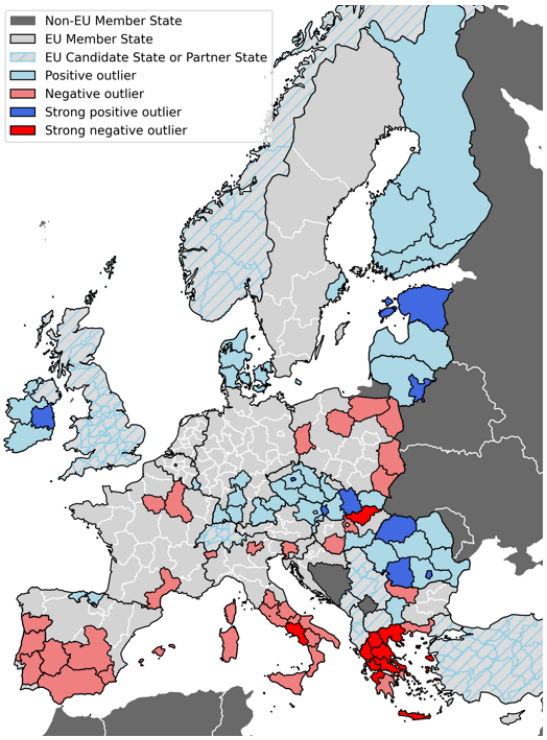



**(d)** GDP [PPS per inhabitant] - Fit on MFV values

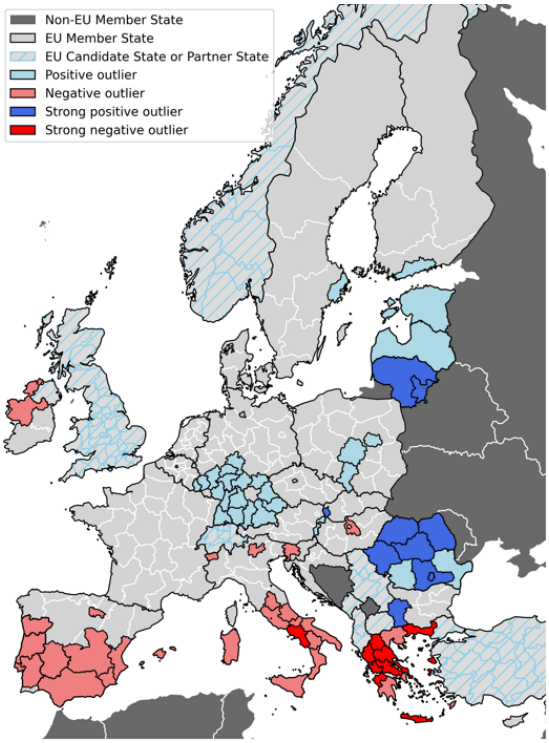
**Figure 4.3:** Estimated outliers for NUTS2 level using GDP per capita values.



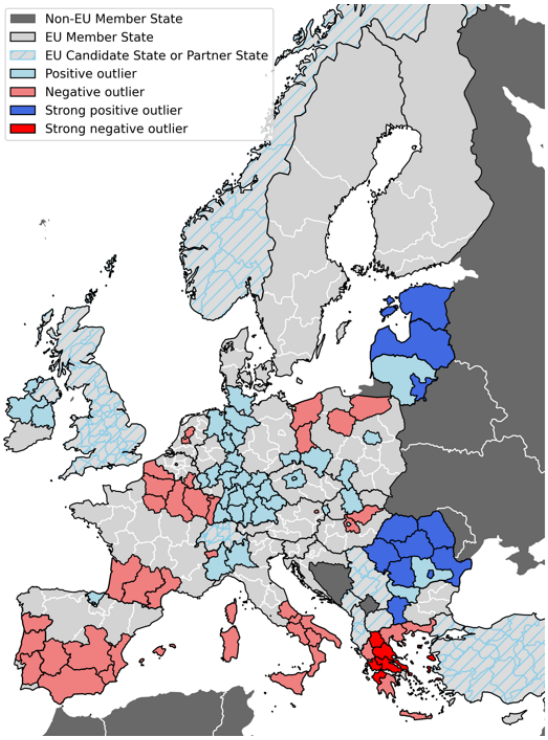

**(a)** NDI [EUR per inhabitant] - Fit on Mean values




**(b)** NDI [EUR per inhabitant] - Fit on MFV values

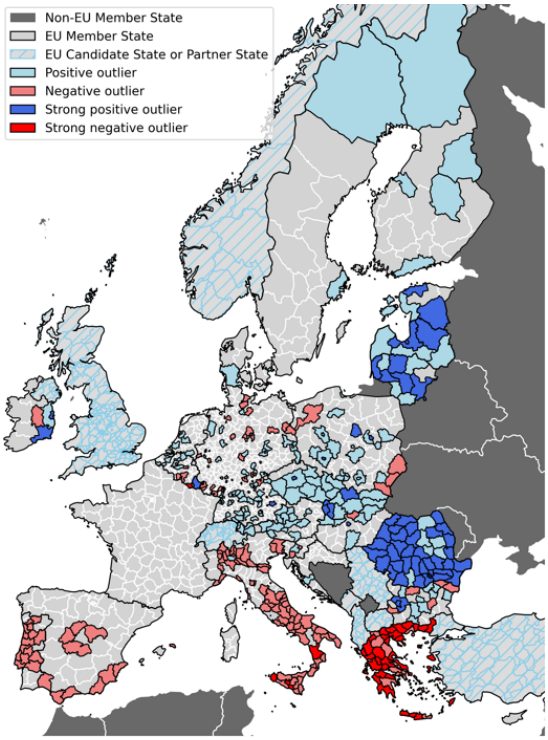



**(c)** NDI [PPS per inhabitant] - Fit on Mean values

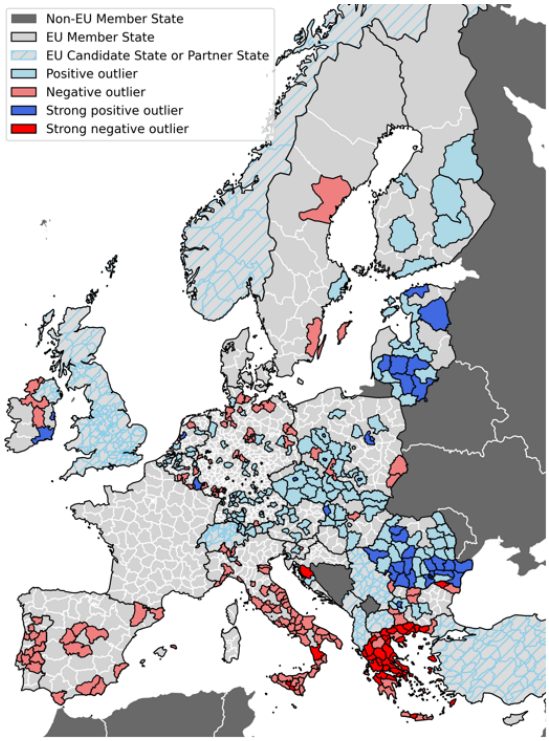



**(d)** NDI [PPS per inhabitant] - Fit on MFV values

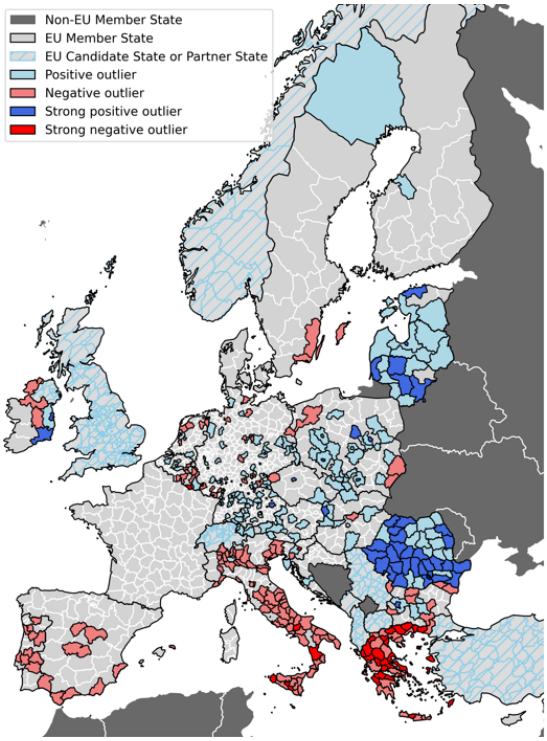
**Figure 4.4:** Estimated outliers for NUTS2 level using NDI per capita values.



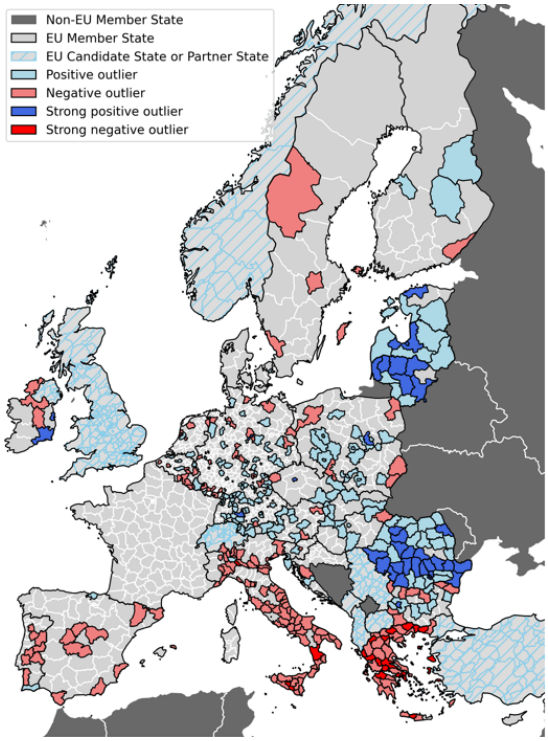

**(a)** GDP [EUR per inhabitant] - Fit on Mean values




**(b)** GDP [EUR per inhabitant] - Fit on MFV values




**(c)** GDP [PPS per inhabitant] - Fit on Mean values




**(d)** GDP [PPS per inhabitant] - Fit on MFV values

**Figure 4.5:** Estimated outliers for NUTS3 level using GDP per capita values.

values did not show significant time evolution.

Robust Mahalanobis (RD) distances – as less outlier sensitive within-distribution distances – can be calculated for arbitrary large dimension numbers, therefore *ecContribution* can be pooled technically with all the gathered regional variables or with even more variables of interest. However, due to the curse of dimensionality, I restricted myself to two dimensions that also enabled easier graphical interpretation. For "far-lying" observations compared to the "bulk" of the data, observations with $RD > 2.5$ values were considered – in accordance with the classification provided in Sec. 4.3. – that is an often used thumb rule suggested in literature and a close approximation of $\sqrt{\chi^2_{2,0.95}}$ value [R113].

The gained results suggest that in general, the H2020 funding of regions tend to form one, unique data-population (see e.g.: Fig. 4.6.). Nevertheless, on country level the method suggests a clear distinction between member states connected before and after 2004 with respect to population (except for Estonia that resulted to be closer to the West-European states in every investigated aspect), therefore in this single case a grouping or clustering could be observed with less H2020 financial funding for the marked East-European members on country level (see Fig. 4.8.).

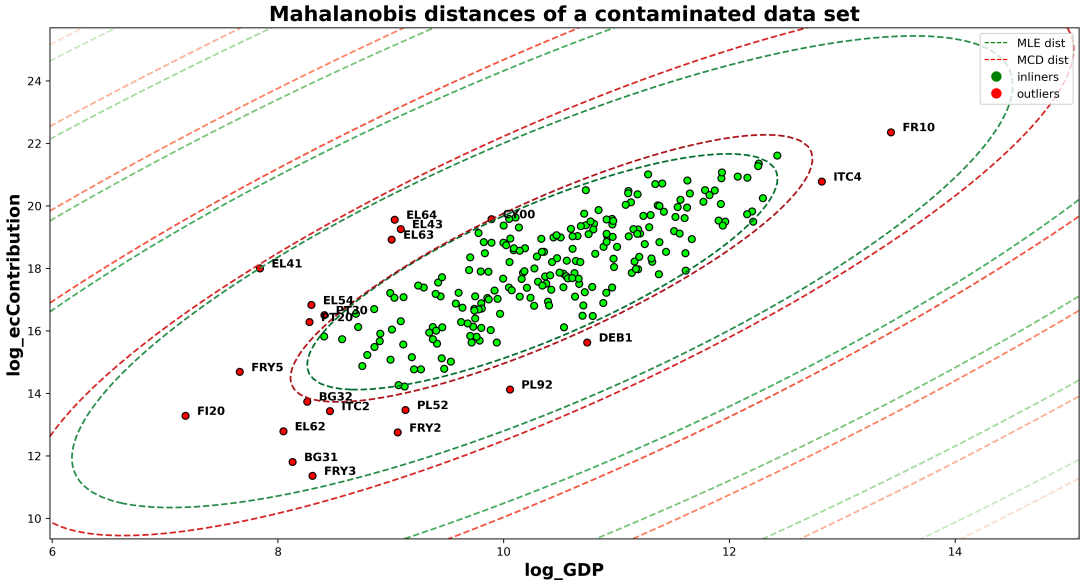



**Figure 4.6:** 2D distribution of the NUTS2 level logarithms of EU contributions and GDP time averages with calculated MLE and MCD contours and outliers with $RD > 2.5$.

For higher level analysis, graphical investigations can be suggested (see Fig. 4.9.) viewed together with the 2D scatter plots coloured according to robust Mahalanobis distances for researchers and policymakers concerned. Thereby, the exact location of each region compared to the bulk of the data can be specified and input for further strategic planing with the aim of improving within-data-distribution positioning (regarding "H2020 community") can be acquired.

The outlined investigation of H2020 spendings of the European Community enables the rethinking of regional positioning that might be relevant for acquiring future research funding. This can contribute to further innovative activities and the boosting of economic growth, that latter in a controlled manner can lead to economic convergence on population level – in mathematical sense – of EU countries and regions. Unfortunately, the analysis revealed that in the past decade R&D funding were uneven in a detectable extent projected onto regional population when speaking of countries connected before and after 2004 to the European Union. On regional level the results are more varied and less obvious. Further in-depth, field-specific economic analysis might be advised in order to explore the relationship

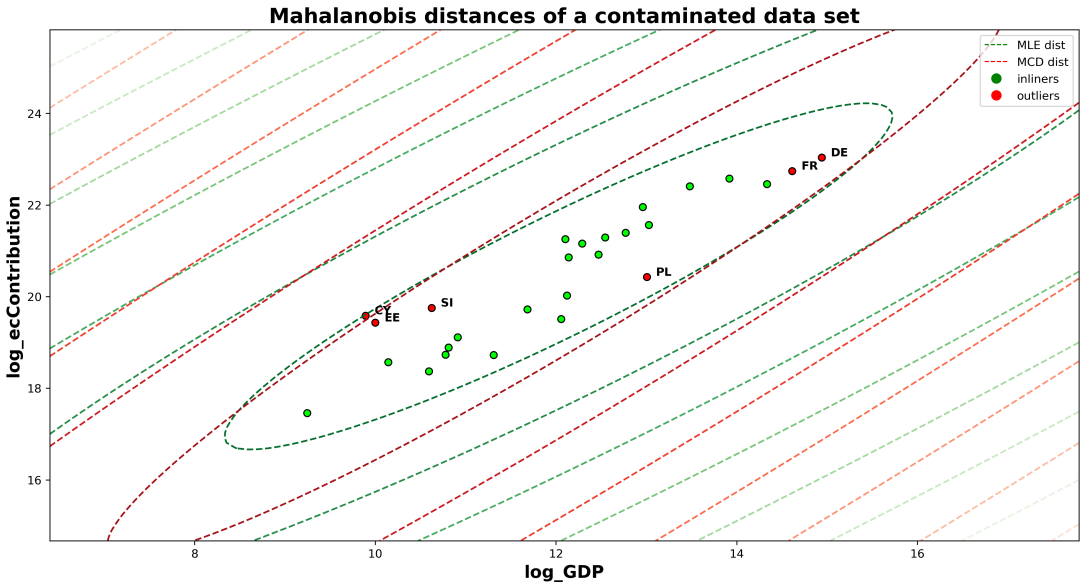



**Figure 4.7:** 2D distribution of the country level logarithms of EU contributions and time-averaged GDP with calculated MLE and MCD contours and outliers with $RD > 2.5$.

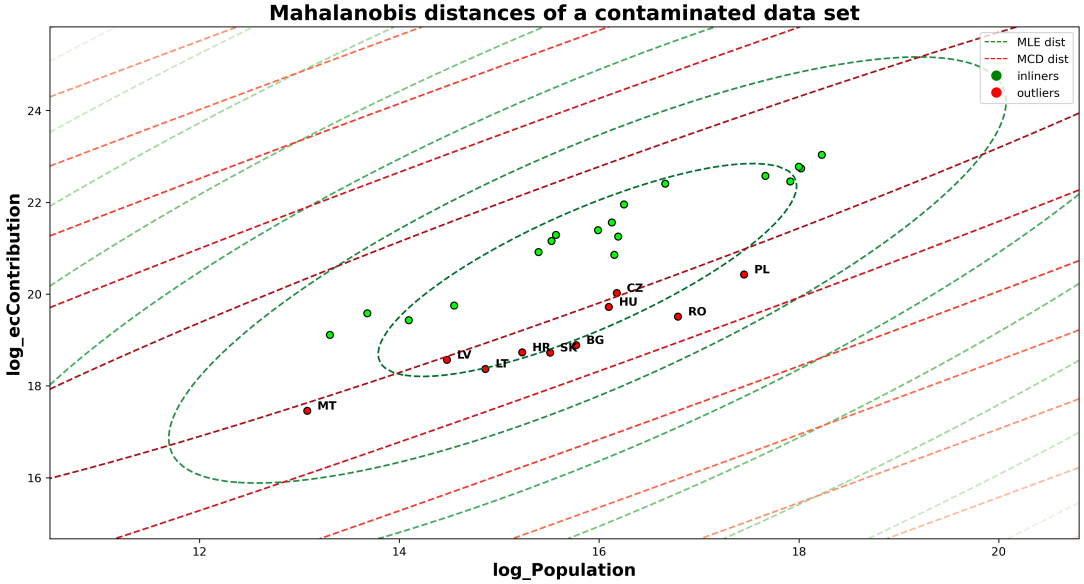
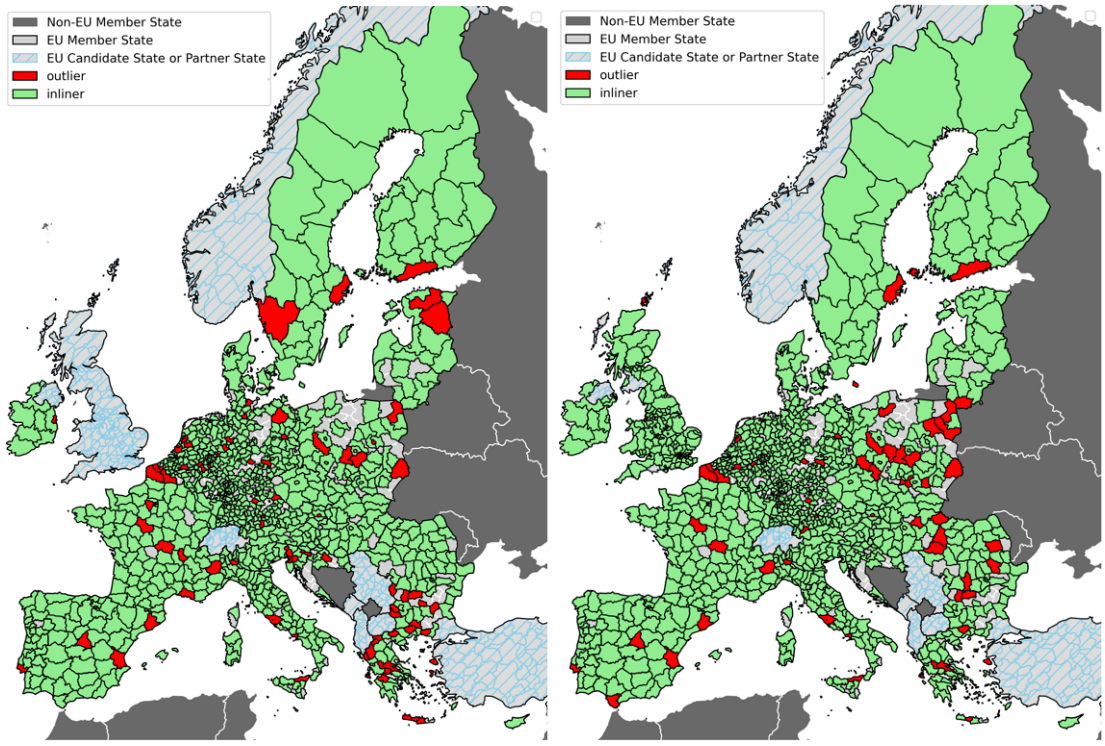



**Figure 4.8:** 2D distribution of the country level logarithms of EU contributions and time-averaged population with calculated MLE and MCD contours and outliers with $RD > 2.5$.

among uneven funding and future economic convergence properties. (Since R&D funding typically affects future economic growth, it is pointless to compare results gained so far with those based on absolute $\beta$-convergence that are primarily based on past economic indicators. Nevertheless, the time horizon and the magnitude of the impact of various R&D investments especially on regional level are hard to estimate.) This unfortunately lies beyond the framework of the present work that had as its main target to explore robust- and non-parametric statistical tools on the application field of economic convergence of EU countries and regions.




**(a)** Time-averaged GDP **(b)** Time-averaged population

**Figure 4.9:** Cartographic representation of resulted 2D EU-contribution outliers of NUTS3 regions with respect to time-averaged regional attributes where corresponding data were available.

## 4.5 MFV-based Clustering and Outlier Detection

In various walks of life (e.g.: biology, economy etc.) outliers cannot be regarded simply as measurement errors, anomalies or as members of different data distributions, since little is known about the mechanics of the unknown model in the background. It is typically required to be able to unambiguously attach labels to observations and describe general group and cluster attributes (*crisp* clustering). Identification of outlying observations is often an issue when no model assumptions are provided, but similarities among observations can only serve as a guideline. In the case of cluster analysis – as an unsupervised learning process – outliers can pose similar challenges, as detailed previously. In practical terms, clustering algorithms are frequently used for the identification of dissimilar observations, even in a multidimensional sense.

The hereby outlined MFV-robustified clustering approach relies on the well-known Lloyd's algorithm [R121] by calculating a robust location parameter (MFV) as cluster centroid simultaneously with a robust scale parameter (dihesion) of the formed clusters in each iteration step (see Algorithm below). Unlike trimming procedures, all observations are taken into consideration but with different weights corresponding to their location within the data distribution of each group.

For the initialization of the iteration, usually the median is advised to be used as the starting value for the MFV, while the Median Absolute Deviation (MAD) for the dihesion. Since typically no a priori knowledge on the data distribution of each cluster is present, the MFV values offer interpretable alternatives for centroids as weighted averages that are mostly determined by the "bulk" of the data. Therefore, graphically the data condensation points will be identified as cluster centres, which are harder to pull away by far-lying observations or heavy-tailed distributions. An additional advantage of the application of the MFV concept is its generalizability to higher dimensions.

**Algorithm 1** Pseudocode of "k-MFVs" clustering

$k$: number of clusters
$D$: a set of objects of cardinality $n$

1. Initialization of centroids $(c_1, c_2, ..., c_k)$ in $D$

**while** no changes in cluster centroids **do**

2.1. for each data point $x_i$ calculate distance from centroid in each cluster

2.2. assign objects to cluster with nearest centroid (swapping step)

2.3. for each cluster $j = 1, 2, ..., k$ recalculate centroids as MFV value of each actual cluster

**return** Centroids, cluster labels, dihesion values

Robustificating Lloyd's algorithm by using Most Frequent Values as cluster centroids is about to construct an alternative to the numerous unsupervised methodologies that perform the breaking-down of data into smaller parts. With this approach it can be achieved not to discard any data point prior to the clustering. In practice every data point may represent valuable information, but without a proper model (e.g.: economic processes) on the background phenomena or knowledge about the various error types dropping data is less recommended. Therefore, my present focus is on crisp-like clustering approaches with no data exclusion.

Since only the selection of centroids is modified we expect to have somewhat similar clustering results as provided by well-known k-Means and k-Medians algorithms. However, due to the iterative depiction of the MFV values an increased time consumption is expected. Moreover, the MFV-robustified alternative (in the following referred as *k-MFVs*) will also predict spherical-shaped clusters in the multivariate space that instantaneously offers future development directions towards considering elliptical-shaped clusters.

The alteration of centroid calculation shall lead to different classification of the data and new centroid coordinates as well. Thereby, the accuracy and interpretation of the grouping could be different compared to k-Means and k-Medians. In order to look into this, we investigate the k-Means and k-Medians algorithms alongside with the outlined k-MFVs in case of 4 real-life data sets accessible at the UCI database [R116]. The main characteristics of the data are listed in Table 4.1. together with the *Long Jump* data set.

The *Long Jump* data set contains the results of two long jump trials from the 1988 Olympic Games of men decathlon and women heptathlon. Being a set of one-dimensional observations it is adequate for visual comparison of different clustering methods. On Fig. 4.10. besides the original data the results of k-Means, k-Medians, trimmed k-Means at $\alpha = 0.05; 02$ levels and the proposed k-MFVs ($k = 2$) are presented in case of the presence of a single outlier that represents a disqualified jump (therefore with zero value) [R111].

The original data show some overlap among men's and women's outcomes and the group mean of that containing the single outlier is highly biased towards that obviously faulty observation. This overlap of the two groups cannot be differentiated by any of the investigated algorithms. Moreover, the presence of the disqualified observation resulted the k-Means and the trimmed k-Means ($\alpha = 0.05$) to breakdown. At the same time, k-Medians and k-MFVs ($k = 2$) and trimmed k-Means ($\alpha = 0.2$) proved to be resistant enough and the trimmed k-Means served with additional information on the identified outliers.

Since our aim is to cluster the data in the presence of outliers without discarding them the robust Mahalanobis distances are used to specify outliers in the formed groups. The empirical- and robust within-group Mahalanobis distances can be compared to the critical value of $\sqrt{\chi^2_{1;0.975}}$ suggested by Hubert et al. in [R115]. Fig. 4.11. shows the distances for

the two resulted groups for the k-MFVs ($k = 2$). The values above the critical values can be considered as group-wise outliers and their further investigation can be done subsequently. A main advantage is that no data had to be suspended, thus the centroids did not get biased because of that. The applied methodology is easy to interpret and can further be extended for higher dimensional investigations. This possibility holds for the 4 UCI datasets, however the detailed investigation of outliers is beyond our scope.

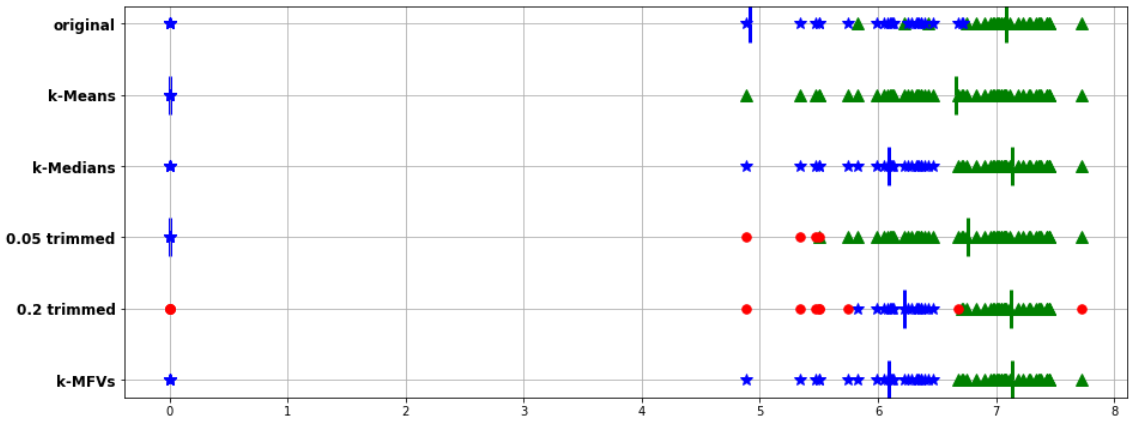



**Figure 4.10:** Comparison of clustering algorithms in case of the *Long Jump* data in the presence of a single outlier.

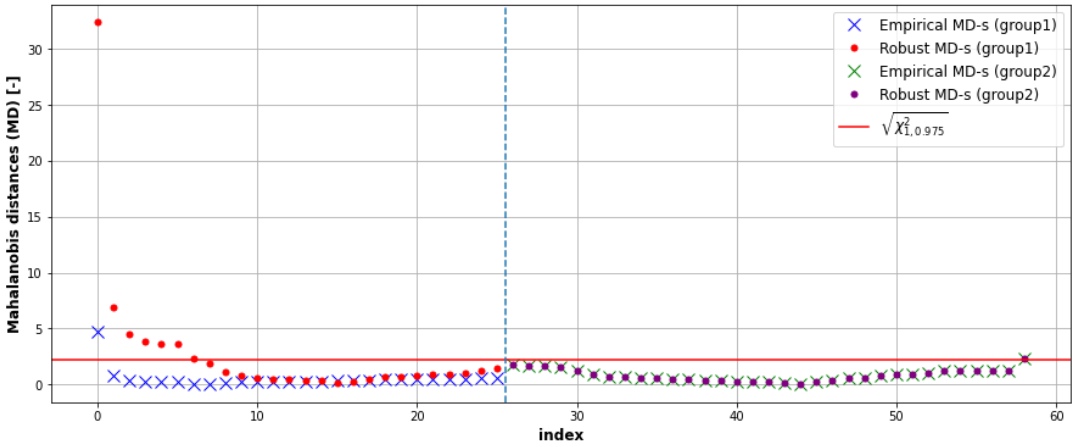



**Figure 4.11:** Empirical- and robustified Mahalanobis distances for the *Long Jump* data extended with a single outlier in case of the k-MFVs ($k = 2$) partitioning into two separate groups as a function of ordered element indexes.

Throughout the experimental investigation all the multidimensional data were used in their ”raw” form without dimensional reduction or standardisation. Since initialization is crucial, k-Means++ and DBSCAN were implemented and tested in order to avoid randomness in the resulted clusters. The k-Means++ proved to be improper for our purposes, because the aim of the robustification with the MFVs is to cluster the ”bulk” of the data and place centroids around high density locations. K-Means++ initialization however, typically led the algorithms to stuck in isolated outlying groups. Therefore, DBSCAN has been selected and by proper parameter sweeping the desired number of initial centroids were defined for each data set in a reproducible way.

The gained results of k-MFVs algorithm in case of the five selected data sets are outlined in Table 4.3. together with the k-Means’ and k-Medians’ for comparative purposes. As important metrics for the judgement of the algorithms the number of ”swap-s” (number of

iteration until no more changes in centroids – and point assignments – is achieved) and computational time for convergence have been recorded besides five clustering validity indices. Table 4.3. contains the results only for the optimal cluster numbers known in advance from the labelled data sets (see Table 4.1.).

| | **Mean** | **Median** | **MFV (k=1)** | **MFV (k=2)** | **MFV (k=3)** | **MFV (k=4)** |
|---|---|---|---|---|---|---|
| | **"Long Jump" dataset** ($n_{clust} = 2$) | | | | | |
| $N_{swap}$ | 4 | 4 | 4 | 4 | 4 | 4 |
| t(s) | 0.0010 | 0.0020 | 0.0479 | 0.0568 | 0.0479 | 0.0409 |
| R | 0.6225 | 0.6225 | 0.6225 | 0.6225 | 0.6225 | 0.6225 |
| $S_{dbw}$ | 0.6511 | 0.6511 | 0.6511 | 0.6511 | 0.6511 | 0.6511 |
| AMI | 0.5087 | 0.5087 | 0.5087 | 0.5087 | 0.5087 | 0.5087 |
| SC | 0.6645 | 0.6645 | 0.6645 | 0.6644 | 0.6644 | 0.6644 |
| DBI | 0.4754 | 0.4754 | 0.4754 | 0.4754 | 0.4754 | 0.4754 |
| Clusters | 33-25 | 33-25 | 33-25 | 33-25 | 33-25 | 33-25 |
| | **"Iris" dataset** ($n_{clust} = 3$) | | | | | |
| $N_{swap}$ | 6 | 6 | 6 | 6 | 6 | 6 |
| t(s) | 0.0030 | 0.0040 | 0.6207 | 0.5513 | 0.6477 | 0.5505 |
| R | 0.7302 | 0.7439 | **0.7437** | 0.7304 | 0.7302 | 0.7302 |
| $S_{dbw}$ | 0.3356 | 0.3373 | 0.3373 | 0.3356 | 0.3356 | 0.3356 |
| AMI | 0.7551 | 0.7631 | **0.7631** | 0.7551 | 0.7551 | 0.7551 |
| SC | 0.5526 | 0.5509 | 0.5509 | 0.5526 | **0.5551** | 0.5526 |
| DBI | 0.6623 | 0.6662 | **0.6662** | 0.6623 | 0.6623 | 0.6623 |
| Clusters | 50-62-38 | 50-61-39 | 50-61-39 | 50-62-38 | 50-62-38 | 50-62-38 |
| | **"Wine" dataset** ($n_{clust} = 3$) | | | | | |
| $N_{swap}$ | 13 | 19 | 12 | 11 | 11 | 12 |
| t(s) | 0.0070 | 0.0119 | 6.0123 | 4.4153 | 4.0440 | 4.5023 |
| R | 0.3518 | **0.3715** | 0.3389 | 0.3415 | 0.3415 | 0.3415 |
| $S_{dbw}$ | 0.4092 | **0.3746** | 0.4134 | 0.4139 | 0.4139 | 0.4139 |
| AMI | **0.4168** | 0.4131 | 0.4068 | 0.4093 | 0.4093 | 0.4093 |
| SC | 0.5596 | **0.5708** | 0.5479 | 0.5447 | 0.5447 | 0.5447 |
| DBI | 0.5496 | **0.5317** | 0.5531 | 0.5541 | 0.5447 | 0.5541 |
| Clusters | 49-27-102 | 62-48-68 | 48-30-100 | 47-31-100 | 47-31-100 | 47-31-100 |
| | **"Ecoli" dataset** ($n_{clust} = 4$) | | | | | |
| $N_{swap}$ | 7 | 7 | 7 | 8 | **6** | **6** |
| t(s) | 0.0049 | 0.0070 | 1.7726 | 1.9345 | 1.4765 | **1.4944** |
| R | 0.6847 | 0.6861 | 0.7541 | 0.6764 | **0.7619** | **0.7619** |
| $S_{dbw}$ | 0.6607 | 0.6607 | 0.6608 | 0.6608 | 0.6607 | 0.6606 |
| AMI | 0.6416 | 0.6483 | 0.6765 | 0.6353 | **0.6836** | **0.6836** |
| SC | 0.4221 | 0.4210 | 0.4210 | 0.4210 | 0.4206 | **0.4226** |
| DBI | **0.9403** | 0.9428 | 0.9423 | 0.9423 | 0.9403 | 0.9403 |
| Clusters | 149-104-75-8 | 149-103-76-8 | 148-104-76-8 | 148-104-76-8 | 149-104-75-8 | 149-104-75-8 |
| | **"Breast Cancer" dataset** ($n_{clust} = 2$) | | | | | |
| $N_{swap}$ | 10 | **6** | **6** | 8 | 8 | 7 |
| t(s) | 0.0076 | 0.0050 | 16.6334 | 20.1386 | 20.2572 | 17.2017 |
| R | 0.4914 | **0.5338** | 0.5286 | **0.5338** | 0.5124 | 0.5019 |
| $S_{dbw}$ | 0.7912 | 0.7857 | **0.7854** | 0.7857 | 0.7881 | 0.7895 |
| AMI | 0.4640 | **0.4973** | 0.4839 | **0.4973** | 0.4805 | 0.4722 |
| SC | **0.6973** | 0.6921 | 0.6911 | 0.6921 | 0.6952 | 0.6965 |
| DBI | **0.5044** | 0.5139 | 0.5154 | 0.5139 | 0.5087 | 0.5064 |
| Clusters | 438-131 | 430-139 | 429-140 | 430-139 | 434-135 | 436-133 |

**Table 4.3:** Validity indexes, main performance metrics and resulted cluster distributions for different location parameter choices and tuning parameter setting (noteworthy results per row indicated by bold).

From Table 4.3. it can be seen that the computational time for the k-MFVs algorithm is much higher, however it cannot be directly compared, since it highly depends on the implementation of the applied built-in functions. Therefore, the relative increase of the time

required has been inspected as a function of dataset size and cluster numbers. According to the results it is not straightforward to expect a time increment with increasing sample size, rather the number of elements in each cluster plays an essential role. For the *Ecoli*, which was the second largest investigated dataset we gained an order of magnitude smaller computational times as in case of the *Breast Cancer*, while the k-Means and k-Medians performed in the same order of magnitude, albeit these required somewhat better run times as well. The latter data had only two relatively large clusters, while the former four clusters out of which three were relatively small in cardinality. Therefore, k-MFVs is expected to serve more cost-efficient result in case of large data sets with more clusters.

The calculation of the MFV values according to Eq. 2.7. and Eq. 2.8. are rather time-consuming. Throughout our investigations the implementation was done in *Python 3.7.13* within an *Anaconda* framework [R92], where the exit criterion from the iterative procedure was established in $\Delta\epsilon_{max} < 10^{-5}$ – similarly, as in case of other presented prior applications – for the dihesion values in two subsequent iterations. This is relatively strict, and its necessity might be data dependent, therefore could be loosened up in order to gain significant time reduction for the k-MFVs algorithm at the same clustering accuracy. For different convergence trajectories of the MFV iterative algorithm in case of the *Long Jump* data set see Fig. 4.12.

The number of centroid swap-s and data reassignments to the groups with cardinality $N_{swap}$ did not show significant variance in case of the algorithms. In case of the *Wine* dataset the k-Medians required more swaps than required by the other algorithms, however in general the k-MFVs resulted in swap numbers between the swap numbers of the k-Means and k-Medoids. The k-Medians and k-MFVs needed approximately the same number of swap-s for higher cluster numbers ($n = 5, 6, 7$), nonetheless k-Medians performed outstandingly in this respect for the *Ecoli* at these cluster number choices. For instance for $n = 7$ (that is far from the optimal clustering setting) the k-Medians required only 5 swap-s, while the k-MFVs 28 and the k-Means 30.

In higher dimensions data are hard to visualise and cluster validity indexes can be used to rely on in order to control the resistance and robustness of the applied procedure in case of specific data. By the investigation of these indexes different methods can be compared and/or optimal cluster numbers can be sought.

Since noises and outliers might influence them even in cases when their presence does not result significantly different groups, corresponding literature draws attention to the possible dependence of such metrics on the selected clustering algorithms [R122]. As a non-sensitive validity index to clustering algorithms the $S_{dbw}$ metric has been selected that has to be minimized in order to gain an optimal grouping [R123, R124].

Whereas the labelling information was also given for all the datasets Silhouette (SC)-, Davis-Bouldin-indicies (DBI) were also calculated besides Adjusted Mutual Information (AMI) [R109, R117] and Rand indices (R) [R110, R117, R125, R126] to compare the resulted groups with the known labels. The SC- [R103], AMI- and R indices had to be maximized while the $S_{dbw}$- and DBI [R106] metrics had to be minimized in the function of cluster numbers (see Table 4.3.). Nevertheless, k-Medians and k-MFVs performed slightly better in all the investigated cases at most of the parameter settings, k-Means were able to serve with better results in case of *Ecoli*- and *Breast Cancer* data sets for DBI or SC, however the differences could only be measured in the third digit.

The calculated validity indexes showed a rather uniform layout for the different cases. This might indicate that the chosen data are not perfectly suitable for spherical partitioning approaches in order to highlight advantageous properties of each investigated algorithm setting. The emerged cluster sample sizes further support this statement. For the previously known optimal cluster numbers the sample size distribution resulted to be approximately the same for the *Long Jump, Iris and Ecoli* data sets. In case of *Wine* data k-Medians led to

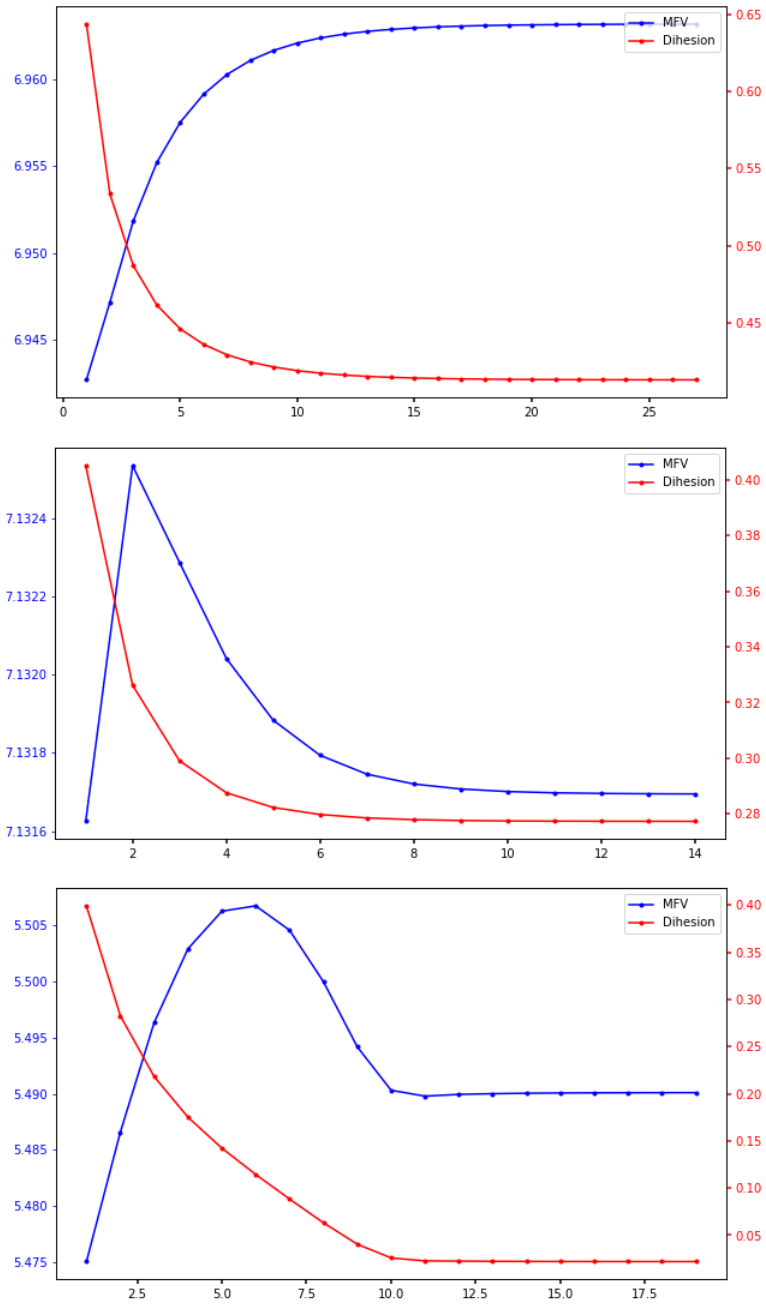



**Figure 4.12:** Typical trajectories towards convergent state of the *Most Frequent Value* and dihesion in arbitrary cases.

more similar cluster distribution to the known one in alignment with the better validity indexes. For the *Breast Cancer* data k-Medians and k-MFVs with $k = 2$ selection provided the same accuracy.

The motivation for the robustified crisp-type algorithm creation was to further enrich the selection of robust clustering methods with an alternative that do not expel any data point by judging it as an outlier. Albeit no significant time improvement – compared to other, from practical point of view proven algorithms – was achieved by the introduction of MFV concept to k-Means like partitioning schemes, the resulted clusters are definitely expected to be slightly different and more outlier resistant. Furthermore, different within-cluster outliers can be identified via e.g.: robust Mahalanobis distances that can give additional information during data acquisition steps. This is of paramount importance with regard

to investigation of data of economic origin by breaking it down into smaller chunks via clustering. In such cases every data point represent valuable information and by neglecting them the variability of the data would be distorted and the derived results biased.

# 5

# Economic Resilience of Small and Medium-sized Enterprises

## Thesis Group 3: Resilience investigations via non-parametric hypothesis testing

***Thesis 3***
**I have designed an indicator number for measuring economic resilience of Small and Medium-sized Enterprises (SMEs) based on fluctuations of their annual sales growth, which was used for the classification of their reaction types to idiosyncratic shocks. With non-parametric hypothesis testing I have demonstrated that showing resilient behaviour is only relevant in the short term regarding individual financial development.**

Publications relevant to the theses: [T1, T7, T10, T11].

***Thesis 3.1***
**Based on the proposed resilience indicator number via matched pair analysis I have shown, that in short- and medium-term the economic attributes of companies reacting resilient to idiosyncratic economic shocks fall behind of those that had not suffered any kind of previous shocks measured in the setback of annual sales growth.**

Publications relevant to the theses: [T7, T11].

***Thesis 3.2***
**Via matched pair analysis, I have shown based on the available balance sheet-, income statement- and other metadata from Hungarian SMEs, that gained experience of short-term successful shock reactions does not influence their survival abilities. The transferability of experiences to the subsequent negative economic events is restricted. However, by enabling longer-term shock reactions the bankruptcy willingness decreases with the length of time enabled to bounce back from the economic shock.**

Publications relevant to the theses: [T7, T11].

***Thesis 3.3***
**I have shown that the proposed sales growth based resilience indicator number and corresponding classification of companies can be extended to a continuous scale of economic shocks and shock reactions, measured in the fluctuations of sales growth. The proposed extension offers a possibility for the comparison of individual observations regarding resilient characteristics. Furthermore, can characterize complete industrial branches regarding resiliency within arbitrary time intervals with respect to their shock reactions.**

Publications relevant to the theses: [T10, T11].

Findings of the present thesis are based on the investigations of real-life annual balance sheet- and income statement data of Hungarian SMEs. The utilized data constitute a property of the *Pannon Business Network Association (PBN), Szombathely*[1]. Any distribution request of the raw data has to be directed towards the managerial board of the organization. The outlined chapter harnessed the accessed financial information in an anonym way and agglomerative, population-level statements were depicted with a pure scientific objective. The following analysis relied on non-parametric statistical tools together with machine learning techniques that did not assume data normality and are less prone to the biasing effect of outliers.

## 5.1 Overview of Related Literature

### 5.1.1 The Concept of Resilience and its Significance

Scientific interest in organizational resilience (or simply resilience) considering the behaviour and reaction of economic stakeholders against turbulent periods and crises has gained substantial momentum in recent years due to the regrettable economic aftermath of the Covid-19 pandemic. This negative economic impact is already affecting everyday life, and it is not yet predictable what further effects are to be expected. Therefore, in order to maintain macroeconomic stability and retain as many workplaces as possible economic resilience is of utmost importance.

According to literature, the term resilience was first introduced in 1973 by Holling et al. in an ecological aspect in order to characterize a system's ability to restore its original state after the occurrence of a negative disturbance [R127]. The pivotal question is how an economic system could react in case of a sudden, unpredictable disturbance. The problem is multifaceted. Albeit being an economic problem human-, technological-, process related-, information related- and other micro- and macroeconomic factors might have a huge and case-dependent influence [R128, R129].

Since the imminent disturbances (Black Swan, tail-event, or X-Event) are unpredictable, in most of the cases there is no means of preparing directly to them. They can originate from numerous sources, might have extreme low probability of occurrence but can also cause significant losses with no theoretical upper bound (e.g.: once-in-a-century flood, Covid-19 pandemic etc.) [R130, R131, R132, R133].

[1]Internet availability of Pannon Business Network Association: `https://www.pbn.hu/`, accessed: 2023.10.05.

In the following I will understand Small and Medium-sized Enterprises (SMEs) under the term ”system” as the subject for further investigations. In this sense, resilience incorporates the system’s flexibility, adaptability and efficacy properties, and it can be characterised by the speed with which the system returns to its original, unperturbed state or by the magnitude of the unexpected environmental perturbation the system can still absorb and survive [R134]. Definitions can be found in the literature about economical resilience – or more directly SME related resilience – already from the year 2000. Nevertheless, the huge variety of definitions and the lack of exact mathematical interpretation due to imprecise knowledge of the system and threatening risk factors makes SME resilience an actual research field [R135, R136].

The term *resilience* originates in the Latin *resilire* that means leap back or rebound. Therefore, the field of resilience tries to cover the ability of the flexible recovery of an entity out of a disturbance may it be of any origin or nature [R137]. A huge amount of definitions and approaches towards economic- and organizational resilience have been created [R127, R138, R139, R140].

Companies were often viewed as standalone systems that have special properties that may contribute to their resilience. However, in real life SMEs are in an interconnected structure, and they are typically forming a part of a supply chain. They tend to outsource activities like bookkeeping, consultations, IT maintenance, tax matters, benchmarks or trainings, which next to the presence of a multi-layered supplier network supports the idea of extending the resilience investigations to a broader environment [R141, R142].

Ruiz-Martin et al. [R143] interpreted resilience as an intermediary step in a four-level resilience maturity model. In their view, depending on the development and abilities of the organisation, the SME can be treated as a fragile, robust, resilient or antifragile system (see Fig. 5.1.). The fragile system collapses right away a disruption occurs. The robust system can tolerate stresses until a certain extent but eventually also collapses. A resilient system not jut tolerates but also survives the turbulent period with manageable losses, while the antifragile system not just survives the turbulent period but also makes advantage out of it (e.g.: by innovations, new contacts, loss of competitors etc.). Therefore, in this context the development goal of the resilient feature of SMEs is clearly selected and attention to a new research area of the ”antifragility of SMEs” is also drawn.

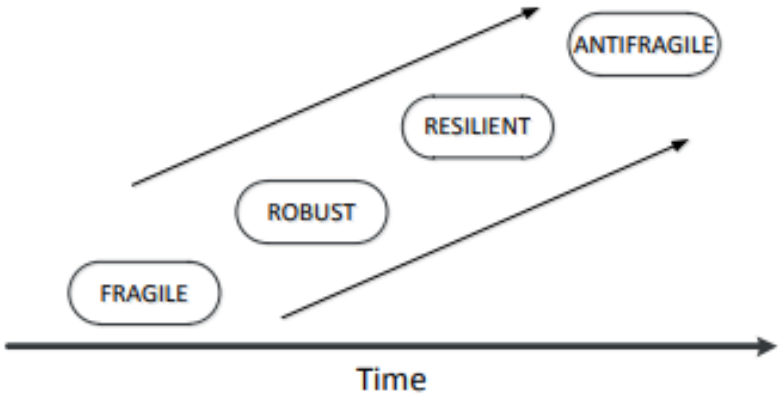



**Figure 5.1:** Stages of four-level resilience maturity model [R143].

In case of a positive reaction to economic distress like antifragility, the goal is not just restore the pre-crisis state but to create a new winning position [R130, R131, R133, R144, R145]. By collecting corresponding best practices or strategies whole economies and thereby regions could be strengthened in an era where economic turbulences are not expected as

only occasionally occurring phenomena any more, but rather a regular part of the everyday life (due to globalization, supply chain embeddings, outsourcing activities, usage of sensitive technologies etc.) [R146, R147].

### 5.1.2 Connection of SME Resilience and Regional Economic Convergence

Economic resilience in terms of withstanding shocks is crucial for SMEs that are essential for decreasing vulnerability in regions. SMEs can contribute to a broader local economic stability via providing jobs, goods and services also amongst unpleasant financial circumstances thereby generating tax and income. Besides job retention, SMEs are important actors of job creation in mordern economies. They continuously maintain tight connections with suppliers, customers and other stakeholders in their reach, which community engagement can collectively contribute to regional economic cohesion [R148, R149, R150]

Increase in number of smaller enterprises tends to support rather than retard economic growth [R151, R152]. Their contribution is more expressed in case of emerging economies or rural areas. In case of economic downturns or crises their locally stabilizing characteristic comes to surface via sustained innovation, maintaining employment levels and consumer spending, which promotes economic recovery and growth. [R153]

In terms of economic convergence, resilient SMEs can play the role of bridging elements that can adapt knowledge and adequate practices from more advanced economies. This helps less developed regions via knowledge transfer to catch up, promote internal growth and reduce income disparities. Economic shocks directly affecting regional resilience and convergence at the same time that might be due to that resilient SMEs often lead to more stable and faster post-crisis recovery [R150, R154].

On the other hand, SMEs often prove to be more innovative than larger firms, which is pivotal for faster productivity gains. This attribute assists competitiveness improvement in lagging regions. Fostering SME resilience with proper policy insturments besides lowering regional disparities can support regional economic convergence as well [R155, R154].

These research insights show that SME resilience is key to long-term economic recovery and narrowing economic disparities, as these firms are many times more agile and responsive to regional economic needs than larger firms. These qualities enable SMEs to support balanced regional growth via smoothing economic fluctuations locally and contribute to reducing the economic gap between urban and rural or lagging regions. Consequenty, understanding the key mechanisms that can lead to favourable shock reactions on corporate level, building acceptable quantitative models based on real-life data or empirical investigations is paramount and promoting the right policies is key for further regional growth and citizen well-being.

### 5.1.3 Literature Aspects of Resilience

Most of the authors in reviewed literature agree that resilience is not just a static but rather a time dependent attribute of an enterprise since resilience is the answer of the system given to a disturbance by withstanding to the negative effects and by quickly returning to stability [R156]. Such measure shall be at least monitored on a yearly basis in order to get a picture of its evolution in time and time-dependent company data shall be investigated, which incorporates recession periods in the enterprise's life.

The time dynamic of the "bouncing-back" phenomenon may depend on several factors (e.g.: company size, type of industry, management decisions, preparatory actions etc.). According to Erol et al. the definition of resilience in general can be phrased as the whole of actions (preventive, defensive and restoring measures) taken before and during the unexpected perturbation [R142]. Thus, in this concept the economic resilience of an enterprise originates from this preparatory stage. This corresponds to the end-result of

a planning activity rather than just a descriptive feature of the actual, static state of the system.

A quite analogous approach is the usage of change management to describe resilience. This involves the actions taken as a response to an event – let it be planned or unexpected – and their efficacy as an important adaptation capability. It also characterises the level of preparedness within the organisation and how fast it can reallocate its technical and organisational capacities in order to withstand the changes. This sheds light on the interesting property of resilience through the general experience of change management – despite being led either by prescribed best practices or standardised procedures – that it is a typical iterative process of trial-and-error phases with its unavoidable pitfalls. Therefore, in this aspect, resilience is rather a result of a development- or learning process than just a characteristic of a system [R157].

Erol et al. in [R134] divides the stages of the reaction to a disturbing phenomenon into eight sections throughout which the companies resistance capability shall be monitored (see Fig. 5.2.). These eight sections already begin with (1) the preparation to the event before (2) the disturbance would occur. Thereafter, the first impacts in the company's operation can be observed and (3) initial measures are taken to soften the negative influences and try to compensate back to the original state. Afterwards, in the presence of more severe circumstances further (4) initial negative effects are observable and later the (5) full impact unfolds itself. In the next stage the company begins to take measures to the (6) restoration of the original state and (7) put the measures of restoration fully into effect. In the last step the (8) long term results and end effects are to be observed that already belong to the new stable state at the same or different level compared to the state prior to the disruptive event.

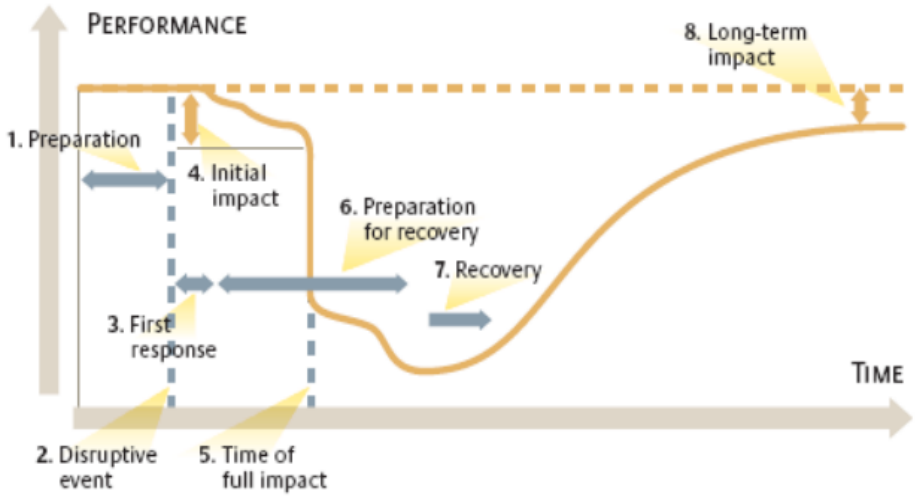



**Figure 5.2:** Phases of a disruptive event viewed in the changes of "hard performance data" as a function of time [R142].

According to Supardi et al.[R158] there are three different approaches of resilience corresponding to the time dependent dynamic of the "bouncing back" phenomenon. In case of perturbation and in business terms resilience shall be interpreted as crisis management. These crisis management approaches are the *proactive-*, the *adaptive-* and *reactive resilience*.

Under proactive resilience the SMEs' preparedness shall be understood, their special business setting and crisis managerial mindset which can be applied under an unexpected situation. This implies business continuity and is bonded to the skill of forecasting and reacting to impacts that might harmfully influence the business processes.

The adaptive resilience is interpreted as a surviving capability under changed circumstances, how tough the business processes of the SMEs' are (e.g.: whether they can maintain their business connections during the depression phase etc.). It also covers a learning willingness and the ability to implement the lessons learned in order to boost development and preparedness. This means a competitive and sustainable adaptation of new knowledge in a turbulent working environment and the capacity to react to changes that – in case of survival – generates motivation and innovation.

The reactive approach of resilience deals with the output side of the change management process, and it is interpreted as the ability of the company to recover after the crisis and consider this period as a learning possibility that helps the observed entity to renew and restructure. Reactive resilience represents the SMEs' ability to come out of the turbulent phase stronger, can restore or even maintain their business processes by limiting and absorbing the negative effects, thus they can still manage to carry out their business priorities.

Vries et al. in [R159] pointed out, that the resilience of SMEs is also highly dependent on the owners' leading attitude and entrepreneurial skills which means further challenges for analytical investigations of resilience.

Based on the work of [R160] economic resilience can be described in three different ways. The *engineering approach* characterizes it with the speed with which the system can return to its pre-crisis stable state. The *ecological approach* considers the absorption capability of the organization before its structure would suffer changes, while the *psychological- and evolutionary approach* understands the adaptation attribute of an economy under the term resilience when perceiving a crisis and what kind of response it can give to it [R129, R161].

### 5.1.4 Prior Approaches Listed in Literature and Practical Difficulties of Resilience Measurement

The time dependency of the recovery process can be quite diverse or even extreme. Some companies may return to a stable position faster than others and this regained stability might be at a different level compared to the pre-disturbance state. Therefore, questions arise, how resilience shall be measured based on time dependent "hard data" that might mean any fiscal data (e.g.: annual revenue, net profit etc.) which were monitored throughout the depression period [R142]. In case of investigating two different company's descriptive fiscal data $(X)$ (see Fig. 5.3.), there can be different levels $(X_A, X_B)$ of the same data and having different time windows of the change in the values of $X$ different relaxation dynamic can be interpreted with different levels of stabilisation compared to the level at time $T_0$. A proper understanding of this dynamic is desirable, which can give an idea of standardisation of time dependent descriptive data or based on their time behaviour the companies in our focus can be segmented into different groups and investigated further separately.

Afgan et al. in [R162] proposed a method based on using "hard data". The authors defined a resilience indicator number as a weighted sum of the time integrals of indicator values that are monitored throughout the period of perturbation (see Fig. 5.4.). As indicator values company profit, company income, final product price and company manpower were considered:

$$R = \sum_{i=0}^{n} w_i \int_{t=T_0}^{t=T_1} \Big[1 - q_i(t)\Big] \, \mathrm{d}t \ , \tag{5.1}$$

where $q_i$ is the $i$-th monitored indicator and $w_i$ is an appropriate weighting factor. At this point however, the selection of indicator values and weighting factors become problematic. In practical cases subjective reasoning must be incorporated and the formula of Eq. 5.1.

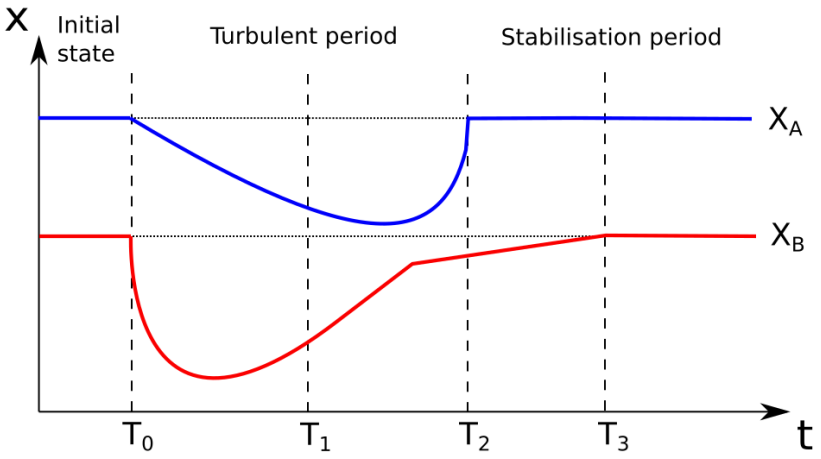



**Figure 5.3:** Different time dependent dynamic of the relaxation of a descriptive ”hard data” throughout an economic depression period [R142].

might not be applicable for every industrial branch (e.g.: service sector with no products produced), therefore generalizability might be restricted.

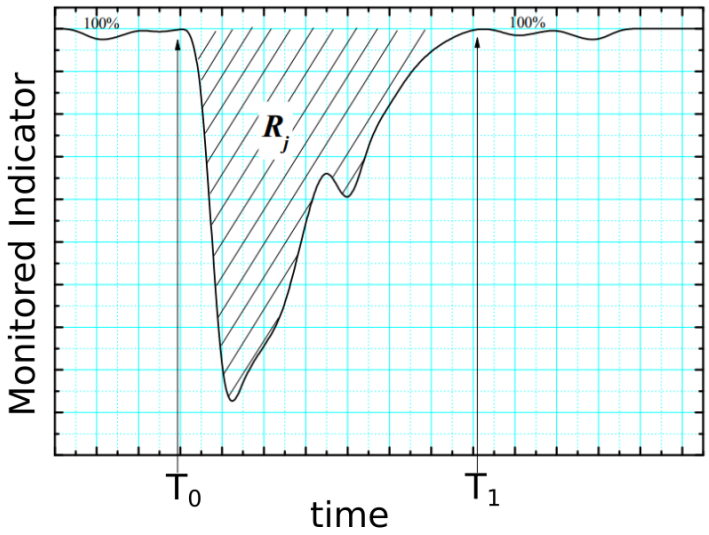



**Figure 5.4:** Sudden change in the *”j”*th monitored indicator throughout a turbulent period. The corresponding resilience metric is depicted from the hatched area [R162].

As a similar concept Coates et al. in [R163] calculated operational resilience from the deviations of production capacity as a monitored indicator factor. The production capacity loss ($PC_{loss}$) was given by the area above curve integral over time. Then this loss was expressed as the percentage of the total production capacity ($PC_{normal}$) – that would have been without disruption – of the investigated time interval. Afterwards the resilience indicator number was calculated as:

$$R = 1 - \frac{PC_{loss}}{PC_{normal}}\,. \tag{5.2}$$

The authors verified their model on simulated data, no real-life verification was presented, however they highlighted that such methods cannot be applied in the lack of a turbulent period and continuously monitored indicators.

Aleksić et al. [R135] investigated ”soft data” collected from senior managers via questionnaires and oral discussions. The importance and vulnerability of certain key areas

were described by linguistic expressions on a Likert scale. They aggregated the individual answers of the respondents – that had unequal importance in the group consensus – with a fuzzy ordered weighted averaging operator (FOWA). The importance of the respondents' answers were set based on their position in the company and compared to reference best practice values from given industries. By setting up a fuzzy decision matrix, the method could serve with simple values for resilience in a range of $[0..1]$ for business processes and for the organisation itself.

Jung et al. in [R164] presented a Cox regression survival analysis based on panel data. Their focus was to verify whether R&D invests and innovative activities have a positive effect on long-term survival. They used time-varying "hard data" and dummy variables depicted from statistical information to presume the presence of innovation or patents etc. (e.g.: R&D activities were assumed in the presence of government or international funds). Based on the presented model correlations between input factors could be investigated together with the survival of companies that had been monitored through 5 years in case of 588 South Korean SMEs.

Sauser et al. in [R165] showed an agent based modelling approach to study the breakdown of a complete system of companies and thereby the resilience of communities and regions. In their simulation SMEs were placed on a grid as abstract points and were equipped with properties like resilience (whether they are able to reopen once), type of their customers (local or global markets) and belonging threshold (how much they depend on the state – open or closed – of the surrounding grid points). Although their model was initialised by random numbers it showed that increasing the severity of the disturbance, after a tipping point the whole system collapses due to the cascade-like failure of the interacting SMEs. Such a model with real-life regional input data can serve with valuable information on regional interconnectedness of SMEs and overall resilience of communities.

According to Somers [R166] there is only sense to interpret the level of resilience after a disturbance occurred (if the company survived it). In his work a latent resilience – that is not presently evident or realised – was estimated based on questionnaire data that provided ordinal-type data on a five point Likert scale regarding factors that were assumed to influence resilience (e.g.: goal-oriented solution seeking, critical situation understanding etc.). Afterwards Pearson's product-moment correlation coefficients, one-way ANOVA and multiple regression methods were used to estimate influences of organisational behaviour on resilience.

Crises typically increase spatial inequalities and can completely reshape total economies. Furthermore, they can trigger resilient behaviour or eventually lead to the death of a business organization, which means that the changing global circumstances often affect economic actors sensitively [R167]. Smaller organizations or countries (e.g.: Singapore) are more exposed to exogenous factors like restricted access to resources and knowledge [R168, R169]. Crises affect companies in different ways both spatially and in time. External- and internal factors may also contribute to deviations in economic performance. Therefore, it would be favourable to characterize resilience of companies in a reproducible way that could serve as a basis of comparison across various industrial branches, regions and countries as well even in different time periods among diverse ambient conditions.

Although literature suggest various metrics for assigning resilient behaviour of organizations and authors highlight highly dissimilar properties that might have an influence on company performance, there is still no consensus on a standardized indicator or mathematical definition that would fulfill every practical expectations [R127]. On the other hand, several literature sources detail concepts of resilience and its measurement but fail to provide financial hard data. Moreover, only few authors had the possibility to access larger amount of longitudinal data that encompasses more than a decade out of various industrial branches [R140].

The approaches via questionnaires and/or identification of human- and company related factors that are case specific, are many times subjective and therefore hard to be measured in a general and reproducible way. This is also due to the problem that financial data are rarely accessible in sufficient depth on the operation of economic organizations. Despite the various studies and considerations most of the authors agree that a consensus should be reached regarding the definition and measurement of resilience, since something that is not measurable cannot be controlled [R128, R130, R146, R170].

### 5.1.5 Small and Medium-sized Enterprises, their Significance, Crisis Management

SMEs build the backbone of the economy of the European Union and this segment constitutes approximately 99% of the total enterprises. More than two-third of the total employment can be attributed to these enterprises which can continuously produce net employment increase. Thus, it is essential to further foster the growth of the SME sector that contributes to the overall competitiveness of the economy [R171]. Likewise, to have a clear understanding of the state of SME resilience and the factors affecting it, it is paramount in order to support government decision makers in reallocating resources destined to the sector in case of an economic depression [R165].

Due to different definitions of the SMEs around the globe there is a different scaling of the companies, which makes the comprehensive definition of SME resilience subtle [R159]. Within the EU, the economic definition for SMEs specifies companies having less than 250 employees and less than 50 million EUR for annual turnover [R172]. For a more detailed specification see Table 5.1.

| SME category | Staff size | Annual turnover | Annual revenue |
|---|---|---|---|
| Medium | $< 250$ | $\leq 50$ m € | $\leq 43$ m € |
| Small | $< 50$ | $\leq 10$ m € | $\leq 10$ m € |
| Micro | $< 10$ | $\leq 2$ m € | $\leq 52$ m € |

**Table 5.1:** The definition of the European Union Commission for SME categories, where annual numbers are given in a million EUR [R173].

The turbulent environment caused by the continuously changing globalised market results in different kinds of challenging situations in the life of an SME. These challenges can be caused by economic recessions, crises (like the Covid-19 pandemic) or among others by competitive changes in the market conditions like the emergence of new competitors, new substituent products or the changing bargaining positions of suppliers and customers [R171].

The challenges that SMEs have to face can be either sudden that might cause big disruptions in certain areas or even slow and insidious, which are hard to recognise in the beginning but can still disrupt key areas and with time can cause cascading troubles. For this reason, it is important to have a comprehensive picture on the linkages among different operation areas and map their vulnerabilities, their exposure to failure in the presence of perturbation, the potentially arising problems and their magnitudes in case of a total failure [R162].

The last global economic crisis took place in 2008-2009, and it has had imbalanced impacts on companies depending on their size with a long-term impact. The smaller SMEs suffered more, they were affected more by market shrinkage and the bank loans were higher for them because of the elevated risk compared to bigger companies. Therefore, they had to make more employees redundant [R164].

Bhamra et al. in [R174] presented results from studies showing that nearly half of the SMEs in the United Kingdom had absolutely no strategy in case of a business discontinuity. While Chen et al. in [R175] concludes, that SMEs do not possess enough resources and

personnel to review and map their improving possibilities, their operations are not transparent enough. For this reason they are unable to get a proper view of their own status of resilience.

According to literature, large companies have more formal crisis management frameworks, while SMEs tend to apply more flexible approaches. They often lack human resources that can be dedicated to crisis response activities or to strategy elaboration accordingly in a properly documented manner. Instead, they tend to focus on more practical and cost-effective approaches, like maintaining closer relationships with key stakeholders and enhance collaboration with partners. Nonetheless, the ability of rapid decision-making and changes in strategies can result pivotal advantages in crucial situations [R176].

It has to be noted though that crises can lead to innovation and resilient behaviour among SMEs as they are forced to re-evaluate their strategies and make creative adjustments. These adaptive strategies and resilience-building via developing the willingness for learning and enhancing general organization-level adaptability can be a key for long-term survival [R177, R178, R150].

Similar opinions have been outlined in [R179] and [R180], where authors emphasize the usage of dynamic capabilities to respond to and recover from disasters. The ability of quick reconfiguration of resources and to innovate make companies more capable to adapt to crises. SMEs in an environment exerted to external shocks (e.g.: disaster, supply-chain disruption etc.) can be treated as more resilient when they invest in resource management, networking to utilize relationships with other businesses for shared resources and learning processes that can also enhance the formation of more robust business models.

According to [R181] SMEs often experience crises as part of a broader cycle of change, where each crisis serves as a potential turning point for innovation and transformation. A successful crisis response should include learning from each disruption and can lead to adaptation in business models, skill- and asset accumulation as well as reconfiguration in everyday operations.

On the other hand however, others argue that planned risk management processes (e.g.: diversifying suppliers, investments in demand forecasting, introduction of flexibility into operations etc.) could significantly contribute to a better anticipation to disruptions and the reduction of overall business vulnerability albeit can significantly foster long-term survival capacities [R180].

Moreover, besides daily operation in a company with limited financial- and labour resources and insufficient strategical thinking, the management is deprived of the possibility of investigating the company's resilience. In order to make the management interested in investing into improving resilience simple and clearly understandable indicators shall be examined, otherwise there will be always some other activity that enjoys more priority [R182].

### 5.1.6 Prospects of Resilience Enhancement of SMEs

From the resilience point of view SMEs shall be treated together with their ambient partner-system (business environment) that is also increasingly exposed to the changing constraints due to globalisation. A greater emphasis shall be put on the training of human workforce and forming of its resilient mindset, because the organisation can become resilient only by the adaptation of proper business processes and by the adequate contribution of the personnel [R135]. Similarly, Patriarca et al. in [R183] suggested SME resilience to be depicted not from the characteristics of the company but rather than from the activities it does.

SMEs are typically centralised organisations that only rarely practice business management processes or does not have a managerial layer at all. They are more influenced by human factors of the decision makers' side compared to the larger companies. Since the majority of SMEs are family businesses, mainly the owners shall be incorporated into

resilience enhancing activities like cooperation with local governments, universities and chambers of commerce. Thereby, more favourable local strategies are to be expected that can respond more adequately to the arising global challenges. [R141]

These companies are great knowledge generators, but poor in knowledge accumulation. With adequate knowledge sharing and flexible treatment of workforce they would be able to give more value to the company, explore customer needs more and retain them [R141].

This knowledge sharing could be utilized among several companies for instance in a form of cluster cooperation. Such a goal oriented and focused business scheme fosters focused business orientation by an interoperable collaborative structure and promotes mutual confidence building. Taking part in such a strategic structure enables the partners to realise mutual advantages that would be otherwise impossible. By working in a cluster scheme the companies can share their knowledge, which can decrease uncertainty and business risks. They can enter new markets with lower invests, increase their negotiation potential, develop new products or compete even with larger firms [R156].

Cooperating in clusters can also contribute to communication with the actors of the supplier chain. This is especially essential with an increasing company size, which – due to the decrease in transparency – might imply further negative impacts on the overall resilient behaviour of the company [R158].

The economic stakeholders are embedded in economic networks, where not just the profit but also the risks are shared. The fall of individual elements might cause domino effects and can jeopardize whole industries or local subsystems [R147, R184]. Therefore, creation of an "early warning system" would be preferred, since a desired reaction to economic shocks is not just company relevant but for regional economies as well. Such early warning systems have already been elaborated regarding bankruptcy prediction in order to warn creditors and decision holders years before a downturn. Authors report different promising classification techniques regarding bankruptcy prediction (kNN, Random Forest, Logistic Regression, SVM, Neural Network etc.) using balance sheet and income statement data [R185, R184, R186, R187].

Traces for a similar, elaborated approach, based on utilization of balance sheet or income statement data in order to create an early warning system for resilient behaviour are extremely scarce in literature and a void can be detected. Although several concept models are published with various metrics that are defined accordingly in order to measure resilience, but mainly questionnaire information is evaluated [R128, R130, R131, R146].

## 5.2 Data Analysis

A unique longitudinal dataset has been analysed with a special emphasis on Hungarian processing industry. This branch of industry contributes typically the most to the Hungarian GDP, consequently it is in focus of high interest. The database contained annual balance sheet and income statement information, data on employee numbers (in form of intervals e.g.: 5-9, 10-25 etc.), industrial branch- and address information supplemented with historical data on bankruptcy, merger and legal status. 26,783 different tax numbers were incorporated within the time period of 2002-2020. According to the suggestions of Virág et al. in [R188] only such companies were involved whose annual net income had at least once exceed 100 million HUF[2] and employed at least once 10 or more people throughout their time series employment data, therefore dominantly SMEs were present in the data pool. Albeit, the gathered data still contained stakeholders with sizes of a wide range this measure was advised in order to filter out untrustworthy annual reports or periodically operating smaller organizations at least to a certain extent.

[2]100 million HUF threshold was taken for the year 2020 and for previous years it had been compensated by the corresponding annual inflation.

Since during the investigated time period several changes took place in the Hungarian economic administration these had to be synchronized. The industrial branch coding had evolved and bankruptcies, company mergers and acquisitions etc. caused tax number alterations. During the explanatory data analysis these anomalies were detected and unified. All together 79 different NACE[3] categories were present in the database (see Fig. 5.6. and Table. 5.2.). The most stakeholders were present in the sectors of *Manufacture of fabricated metal products* (25), *Food production* (10) and *Manufacture of machinery and equipment n. e. c.* (28). For the time evolution of the number of employees see Fig. 5.5., where the ranges of employee numbers were unified and aggregated for those years when data were accessible. For further calculations, the employee number were estimated as the average of the interval limits given.

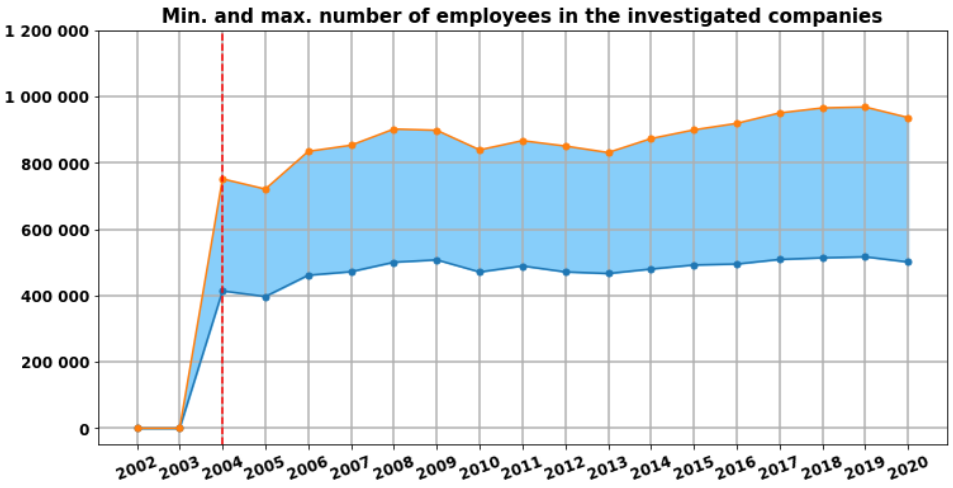



**Figure 5.5:** Number of employees in each year when it was applicable. Since ”from-to” intervals were provided the aggregation was done accordingly.

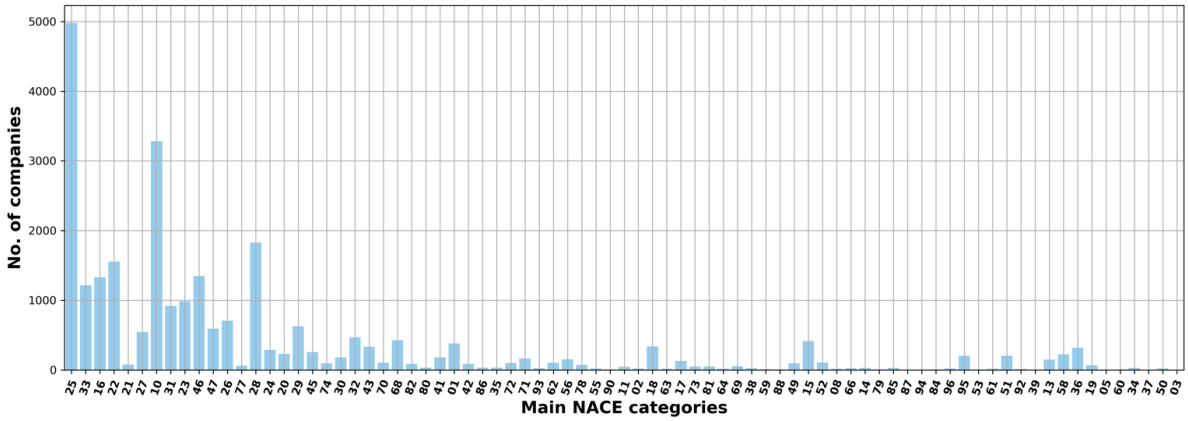



**Figure 5.6:** Number of companies in different industrial branches.

The available address information was geocoded with the open source *geopandas 0.8.0* python module that is the process of generating latitude and longitude coordinates and generation of geographic position from addresses and names of locations [R88]. By visualising the data it can be seen that the whole country is ”covered” by observations, but local condensations around bigger cities are prominent and the capital with its surrounding is overrepresented compared to rural regions. The location data were also used to create new

[3]Standard classification for economic activities used within the European Union.

| NACE | Count | Branch of industry description |
|---|---|---|
| 1 | 380 | Crop and animal production, hunting and related service activities |
| 10 | 3284 | Food production |
| 11 | 47 | Manufacture of beverages |
| 13 | 150 | Manufacture of textiles |
| 14 | 31 | Manufacture of wearing apparel |
| 15 | 417 | Manufacture of leather and related products, footwear |
| 16 | 1330 | Manufacture of wood and of products of wood and cork, except furniture |
| 17 | 132 | Manufacture of paper and paper products |
| 18 | 339 | Printing and other reproduction activities |
| 19 | 69 | Coke production, petroleum refining |
| 2 | 22 | Forestry |
| 20 | 232 | Manufacture of chemicals and chemical products |
| 21 | 78 | Manufacture of pharmaceuticals |
| 22 | 1555 | Manufacture of rubber and plastic products |
| 23 | 987 | Manufacture of non-metallic mineral products |
| 24 | 290 | Manufacture of basic metals |
| 25 | 4981 | Manufacture of fabricated metal products |
| 26 | 709 | Manufacture of computer, electronic and optical products |
| 27 | 548 | Manufacture of electrical equipment |
| 28 | 1827 | Manufacture of machinery and equipment n. e. c. |
| 29 | 629 | Manufacture of road vehicles |
| 3 | 3 | Fishing, fish farming |
| 30 | 179 | Manufacture of other transport equipment |
| 31 | 920 | Furniture production |
| 32 | 468 | Other manufacturing |
| 33 | 1215 | Repair and installation of industrial machinery, equipment and tools |
| 35 | 34 | Electricity, gas, steam and air conditioning supply |
| 36 | 321 | Water production, treatment and supply |
| 37 | 2 | Sewage collection and treatment |
| 38 | 28 | Waste management |
| 39 | 3 | Decontamination and other waste treatment |
| 41 | 182 | Construction of buildings |
| 42 | 90 | Construction of other civil engineering projects |
| 43 | 336 | Special construction |
| 45 | 258 | Trade and repair of motor vehicles and motorcycles |
| 46 | 1347 | Wholesale trade, except of motor vehicles and motorcycles |
| 47 | 591 | Retail trade, except of motor vehicles and motorcycles |
| 49 | 96 | Land transport by pipeline |
| 5 | 1 | Coal mining |
| 50 | 24 | Water transportation |
| 51 | 203 | Air Transport |
| 52 | 106 | Warehousing and support activities for transportation |
| 53 | 6 | Postal and courier activities |
| 55 | 23 | Accommodation service |
| 56 | 154 | Hospitality |

**Table 5.2:** Main NACE categories and their description for the investigated companies with the corresponding number of occurrence among the instances.

variables regarding the level of urbanization proximity. On Fig. 5.7. the companies are illustrated with different colouring according to their closeness to bigger cities or urbanized locations with appropriate classification. In addition to the presence of the capital, the proximity to ”Large Cities” (cities with county rights) and ”Bigger Cities” (district seats) was identified. During further classification of company-year observations regarding levels of resilient behaviours using Logistic Regression and Random Forest models, the generated ”urbanization closeness” served as a supplementary feature variable (see Section 5.5).

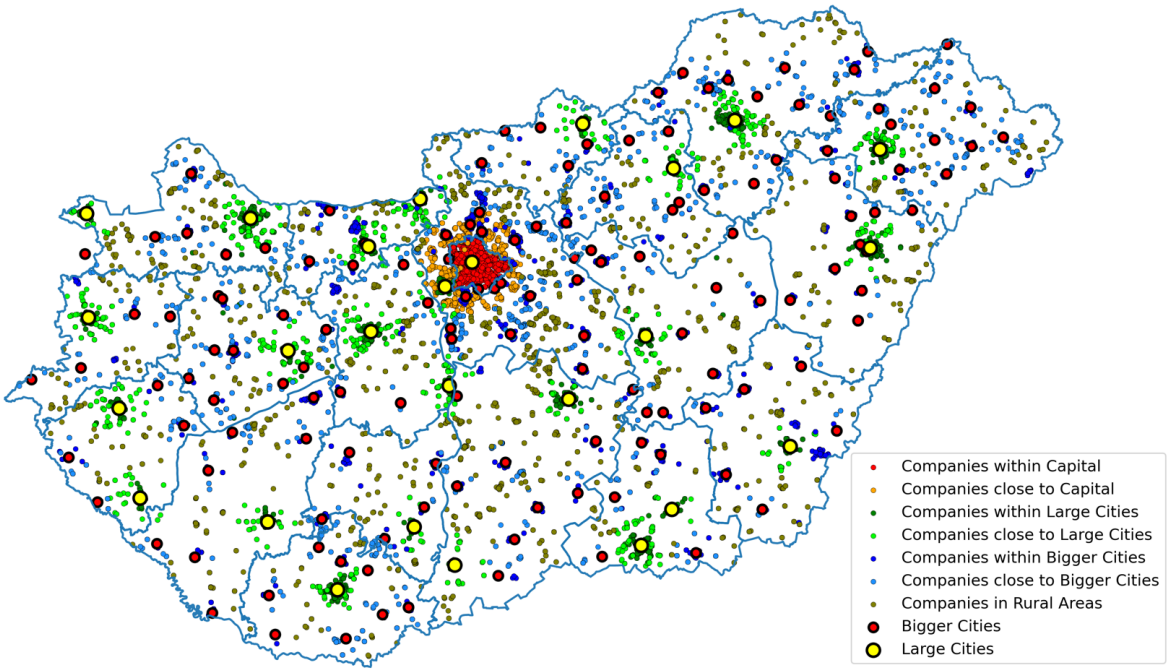



**Figure 5.7:** Classification of Hungarian companies with shock information (setback of minimum -10% sales growth at least once throughout their time series data) according to their geographical closeness to urbanised regions.

## 5.3 Proposed Resilience Indicator Number

Since resilient attribute is in tight connection with economic performance, it shall be investigated based on one or more performance indicators. Literature suggests several financial performance metrics, however it also emphasizes that common financial ratios (e.g.: Return on Equity (ROE), Return on Assets (ROA), Return on Investment (ROI) etc.) can easily be biased [R189]. This is an inevitable consequence of the accounting policy that gives certain freedom in bookkeeping that can cause ambiguity to statistical analyses. Several authors point out that such traditional and widely used metrics can show opposite behaviour when companies are in a growing phase in accordance with management decisions (e.g.: ROE can increase and decrease due to different capital reallocations, investments or change in capital structure of the company etc.) [R190]. Furthermore, in case of statistical investigations regardless of the probably biased nature of such metrics ratios can show unexpected unwanted behaviour, which can easily lead to artificial (otherwise non-existing) correlations among variables and later on can lead to data-dependent, non-robust and contradictory findings [R191, R192, R193, R194, R195].

In order to construct a widely applicable measure for resilience the *sales growth* of each organization has been selected as a financial performance indicator that is often used in corresponding literature [R196, R197, R198]. Sales growth as the relative change in annual turnover can represent the market conditions of a company that can be affected by changes in technology, emerging competitors, changes in customer requirements and habits, macroeconomic crises (see Fig. 5.8.) and so on.

By deciding to utilize only one of the numerous possible financial indicators for characterizing shock response processes definitely has to be noted among the limitations of the investigations to be presented. However, the exclusion of possible biases enables reproducible, objective calculations that can be extended sector independently to enterprises of other countries as well, where highly interpretable results are expected from a practical point of view and also firm connections to related literature can be established. Further

extension of the indicator number to be proposed by involving other financial indicators is about to form a basis of an extended future research with broader company related considerations and the involvement of more in-depth business understanding on SME level. However, due to the primarily methodological focus of the present work this lies outside of the framework of my current investigations.

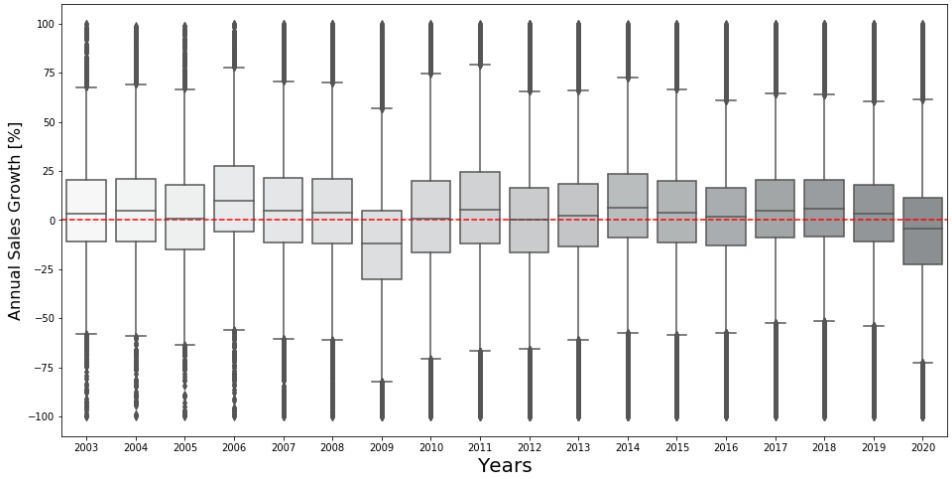



**Figure 5.8:** Application of sales growth in order to indicate macroeconomic crises (years of 2009 and 2020) among Hungarian SMEs of processing industry.

As being said, the sales growth based approach provides an opportunity to create comparable results over different companies and industrial branches regardless of company size and without the presence of unwanted interaction terms in further statistical analyses thereby can provide a basis for standardisation. According to practical investigations however, there are rapid fluctuations in the selected measure (see Fig. 5.9.) that makes long-term investigations arguable, since there are technically no long-term steady states and in many cases only few years separate the consecutive setbacks.

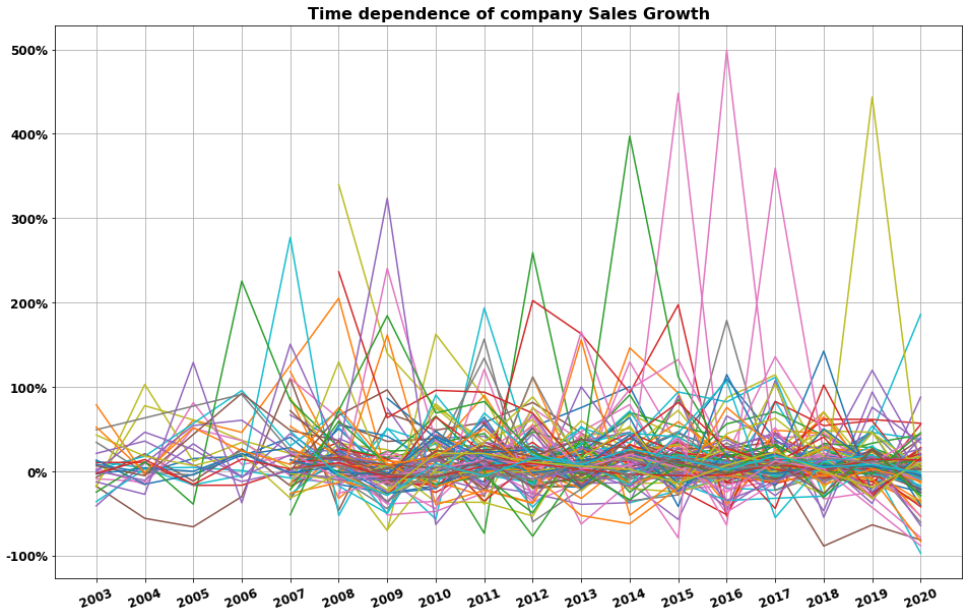



**Figure 5.9:** Annual sales growth values of 100 randomly selected companies (without first 3 years of company data in order to eliminate initial high fluctuations attributable to startup of the enterprises).

Therefore, due to practical reasons we support short-term resilience investigations that

fit real-life conditions more compared to long-term investigations, since the data at hand do not support the analysis of long-term effects [R167, R197, R199]. On the other hand the ”life” and behaviour of an SME is largely influenced by human factors (owner, CEO, managers etc.), which cannot be extended to the long run.

According to literature, in the following resilience will be treated as a behavioural pattern of an economic entity that is a rapid reaction and restructuring after an **unpredictable crisis phenomena** and will characterize how successfully can the organization sort out difficulties from one year to the consecutive one [R139, R167]. Moreover, the subsequent analysis will connect to those studies that regard resilient behaviour as a reaction to a crisis which means that *only those entities can be considered as resilient that at least once in their lifetime successfully survived a crisis* [R200]. In the following an economic downturn will be considered as a necessary prerequisite of resilience and therefore *reactive resilience* of companies is analysed. The presence of a crisis will be measured in the setback of sales growth and the minimum level of -10% setback in sales growth will be defined as a threshold for detecting a crisis in the history of a company [R198, R201].

In order to construct an indicator number for measuring resilient behaviour that is widely applicable and easy to interpret, the four-staged resilient maturity model defined in [R138] is utilized. Nevertheless, the theoretical concept is extended with empirical considerations. In short, we distinguish between *fragile*, *robust*, *resilient* and *antifragile* companies based on the following criteria:

- **Fragile:**

  $\mathrm{R}(\mathrm{year}_{i+1}) < \mathrm{R}(\mathrm{year}_{i})$ or

  $\mathrm{SG}(\mathrm{year}_{i-1,i}) > \mathrm{SG}(\mathrm{year}_{i-1,i+1}) >$ -100%

- **Robust:**

  $\mathrm{R}(\mathrm{year}_{i}) < \mathrm{R}(\mathrm{year}_{i+1}) < \mathrm{R}(\mathrm{year}_{i-1})$ or

  $0\% > \mathrm{SG}(\mathrm{year}_{i-1,i+1}) > \mathrm{SG}(\mathrm{year}_{i-1,i})$

- **Resilient:**

  $\mathrm{R}(\mathrm{year}_{i-1}) < \mathrm{R}(\mathrm{year}_{i+1}) < 2{\cdot}\mathrm{R}(\mathrm{year}_{i-1})$ or

  $100\% > \mathrm{SG}(\mathrm{year}_{i-1,i+1}) > 0\%$

- **Antifragile:**

  $2{\cdot}\mathrm{R}(\mathrm{year}_{i}) < \mathrm{R}(\mathrm{year}_{i+1})$ or

  $\mathrm{SG}(\mathrm{year}_{i-1,i+1}) > 100\%$,

where $\mathrm{R}(\mathrm{year}_{i})$ denotes the annual revenue in year $i$, when the economic shock occurs, while $\mathrm{SG}(\mathrm{year}_{j,k})$ denotes the sales growth from year $j$ to year $k$. The outlined definition implicitly assumes that $\mathrm{R}(\mathrm{year}_{i}) < \mathrm{R}(\mathrm{year}_{i-1})$, namely there is a certain level of setback due to the disturbance. According to the response of the organization measured by its annual revenue or sales growth – that technically represents its input in an economic aspect – the classification regarding the time-dependent, annual ”crisis-reaction” can be done (see Fig 5.10.).

This concept connects to those authors that regard resilient attribute as a short-term, quick reaction and restructuring as a response to a disturbance. Thereby, reaction to idiosyncratic economic shocks are characterized by a time-dependent, annual ”label”. Companies for which it takes several years to reach their pre-crisis level regarding annual revenue are not considered resilient in the given year. Organizations, which can be categorized as resilient or antifragile can be denoted as ”*one-year reactive resilient companies*” (OYRRCs). This concept is in alignment with everyday realistic expectations in the sense that shock

reaction of companies is more interesting in the short-term, furthermore the access to additional data and long-term stability in SMEs' financial operation is of low probability.

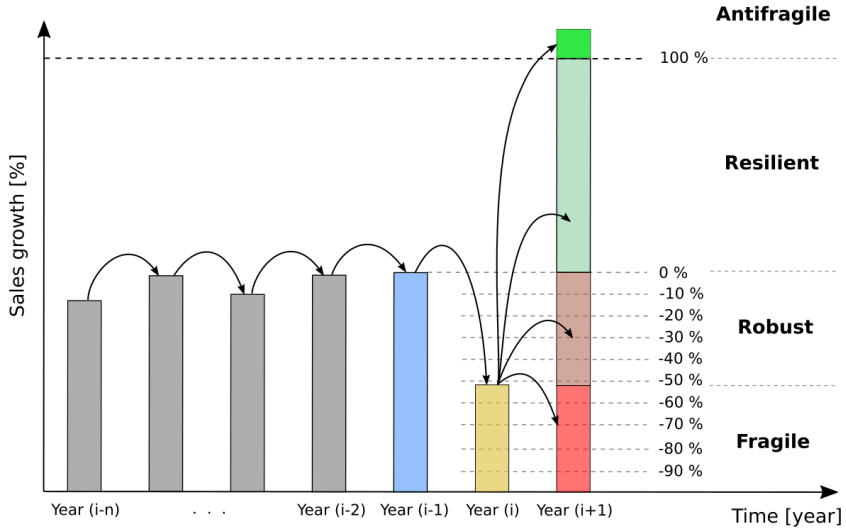



**Figure 5.10:** Classification of companies based on time-dependent, annual "crisis-reaction" with respect to sales growth.

### 5.3.1 Further Extension Possibilities

Although the above definition is straightforward and simple, it still lacks the ability to compare shock reactions of different companies to each other. Just about to survive a huge, for instance -70% setback of annual sales growth in a robust way shall not be considered immediately as a weaker reaction than coming out of a -12% setback in a resilient way. A continuous resilience metric might be advantageous that not just considers the shock reaction itself then the level of survived shock as well.

The available company-year observations with regard to the endured economic shocks and the given reactions to it measured in sales growth can be visualised on a 2D scatter plot according to Fig. 5.11., where the independent variable ($x$) is the sales growth between the year of crisis and the preceding year while the dependent variable ($y$) is the sales growth between the subsequent year and the year of crisis. The base year for calculating the independent- and dependent values for visualising shock – shock-reaction pairs was year$_{i-1}$ i.e. the preceding year to the idiosyncratic shock of a given entity that is technically a sweep parameter of the company-year observations.

The red dots in Fig. 5.11. correspond to observations where at least -10% shock took place, therefore correspond to a shock reaction and are of special interest of the present investigation. To these fragile / robust / resilient / antifragile shock reactions a continuous resilience metric can be assigned that needs to fulfill the following criteria:

1. The $f(x, y)$ resilience metric shall be a continuous function of $x$ = SA(year$_{i-1,i}$) and $y$ = SA(year$_{i-1,i+1}$), where $(x, y \geq -1)$.

2. $f(x, y) \geq -1$, that means that to the least non-resilient behaviour $-1$ is assigned (typically bankruptcy), while no upper limit of resiliency is defined since in theory unbounded post-crisis increase is possible.

3. If $x = 0$ and $y = 0 \Rightarrow f(x, y) = 0$, that is in steady conditions the resilience metric is defined to be zero.

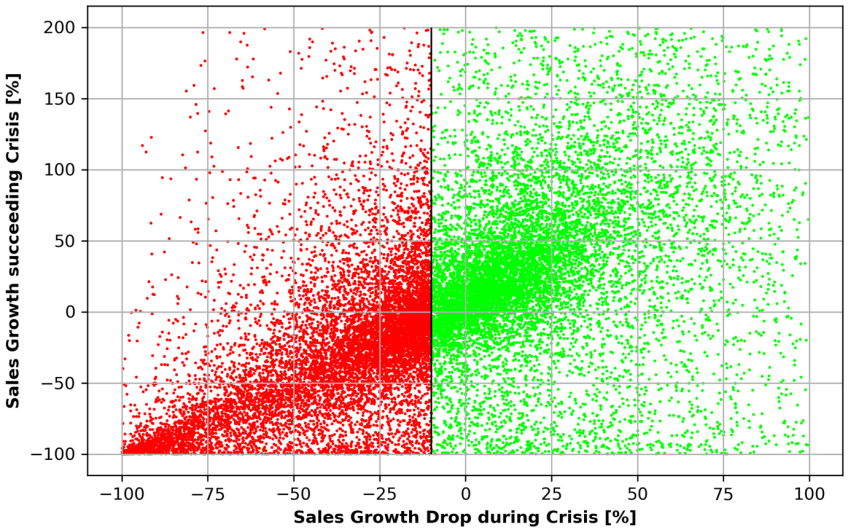



**Figure 5.11:** Relation of two years sales growth to the ”simple” sales growth measured from the same base year.

4. If $x = \text{const.}$ and $y_1 > y_2 \Rightarrow f_1(x, y) > f_2(x, y)$, that is larger post-crisis developments should be rewarded at the same shock levels.

5. If $y = \text{const.}$ and $x_1 > x_2 \Rightarrow f_1(x, y) < f_2(x, y)$, that is surviving more severe shocks should be rewarded at the same post-crisis development levels.

In theory several continuous functions might satisfy the above conditions. In order to fulfill the above requirements a simple logarithmic function can easily be constructed that can suppress the atypical outstanding shock reactions (see Fig. 5.12.):

$$f(x, y) = \frac{ln(y + 2) \cdot ln(-x + 2)}{ln^2(2)} - 1 \tag{5.3}$$

The formula of Eq. 5.3. is applicable for comparing shock-reactions of companies where only annual financial data are available. Fig. 5.12. also illustrates the domains suggested by the four-level resilience model for classification of the company-year observations where an economic shock occurred in the preceding year. Thereby whole industrial branches can be compared to each other with regard to idiosyncratic shock reactions by observing cardinality of observations within specified time limits in each class and the distribution of data can be used as well, where for instance the comparison of robust location and scale parameters of distributions might be applicable.

## 5.4 Matched Pair Analysis for Shocked and Un-shocked Companies

Results detailed in the present section directly build upon the the rich, longitudinal dataset on Hungarian processing industry described in Sec. 5.2. and shock reaction to economic disturbances are analysed according to the resilience indicator number definition provided in Sec. 5.3. As a minimum of setback in sales growth for the identification of an economic shock in the history of a company -10% was defined in accordance with literature suggestion. The geographical distribution of SME headquarters with and without the resulted shock information throughout their lifetimes is depicted on Fig. 5.13.

For further investigations companies with sufficiently long financial history were filtered out and rapid fluctuations belonging to company starting periods (start-up phase) were

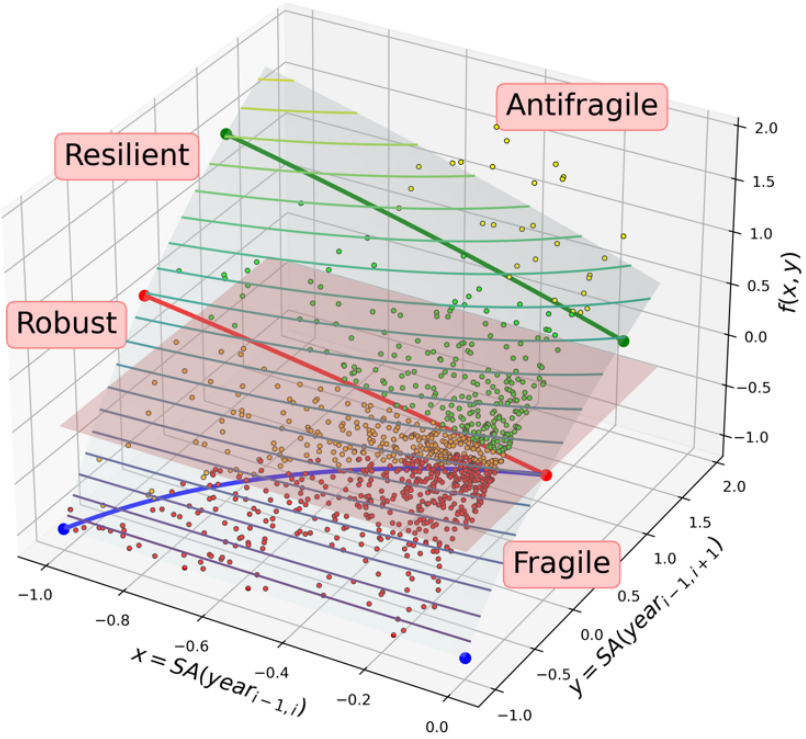



**Figure 5.12:** Proposed logarithmic function given in Eq. 5.3 as a continuous resilience metric option for enabling the comparison of companies' shock reactions.

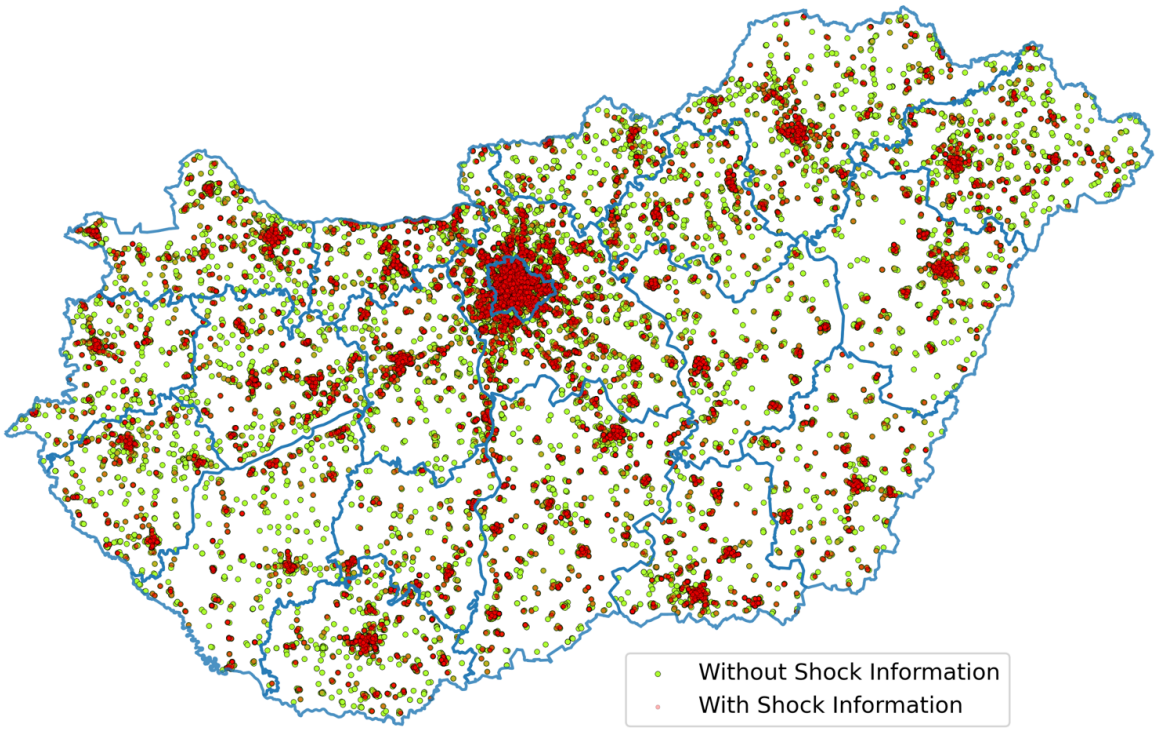



**Figure 5.13:** Headquarters of investigated processing industry stakeholders. The companies that had at least once more than a -10% sales growth drop in their lifetime are marked with red.

also omitted. Therefore, only companies without their first 3 years of financial data and at least with 5 existing financially closed fiscal year were investigated. This led to a dataset of 25,889 different tax numbers and 301,684 country-year observations out of which 40,695

showed a shock reaction to a setback worse than -10% in the measured sales growth.

By investigating changes in sales growth the annual shock phenomena experienced by companies can easily be determined. This enables the identification of annual crises during the operation of the companies regardless their origin or type (e.g.: caused by macroeconomic situation, human factor, market condition etc.), however the real occurrence of economic crisis and the realization in fiscal data are not necessarily coincident. For instance the global economic crises of 2008 made its highest impact regarding economic setback measured in sales growth (see Fig. 5.14. and Fig. 5.15) in 2009, however regarding employee numbers 2010 resulted to be the worst year in general. These visualisations enable us investigations independently of firm size and industrial branch (but with optional incorporation possibility). Moreover, the level of shock, that the organization had to undergo can be further characterized [R198].

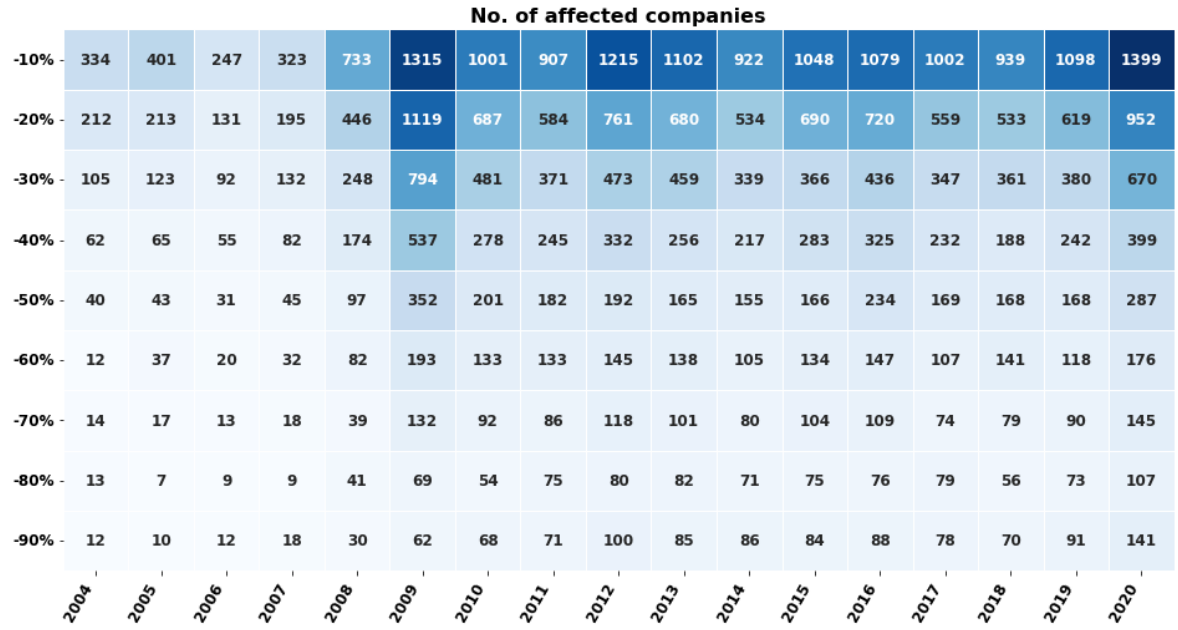



**Figure 5.14:** Number of companies that suffered a certain level of shock in each year measured by the setback of their sales growth.

The number of organizations can be similarly determined in each year that managed to successfully come out of a financial distress measured by the setback in sales growth. In Fig. 5.16. the total number of resilient and antifragile companies are aggregated to different shock levels. These numbers aggregate the company-year observations in each year of the investigated time period when an organization could respond positively to a shock in the consecutive year in accordance with Fig. 5.10.

The presented classification based on the crisis reaction measured by sales growth evolution enables the characterization of the various annual crisis responses of the whole population, namely the Hungarian processing industry. Fig. 5.17. shows the fourfold classification of the industry members on annual basis. Nevertheless, Fig. 5.14., 5.16. and Fig. 5.17. give an insight into crisis related behaviour of the total population concerned that demonstrates less dependence on macroeconomic crises. Individual, company-specific crises can take place independently of global turbulences as well, however they are most probably influenced by them. For this reason, the crisis-influence and shock-response layout (see Fig. 5.18.) as an additional important aggregation of the resulted time-dependent classification data should be generated.

On the resulted shock-response layout the top-left corner compresses those occasions when the investigated actors reacted poorly even to small disturbances, while the bottom-

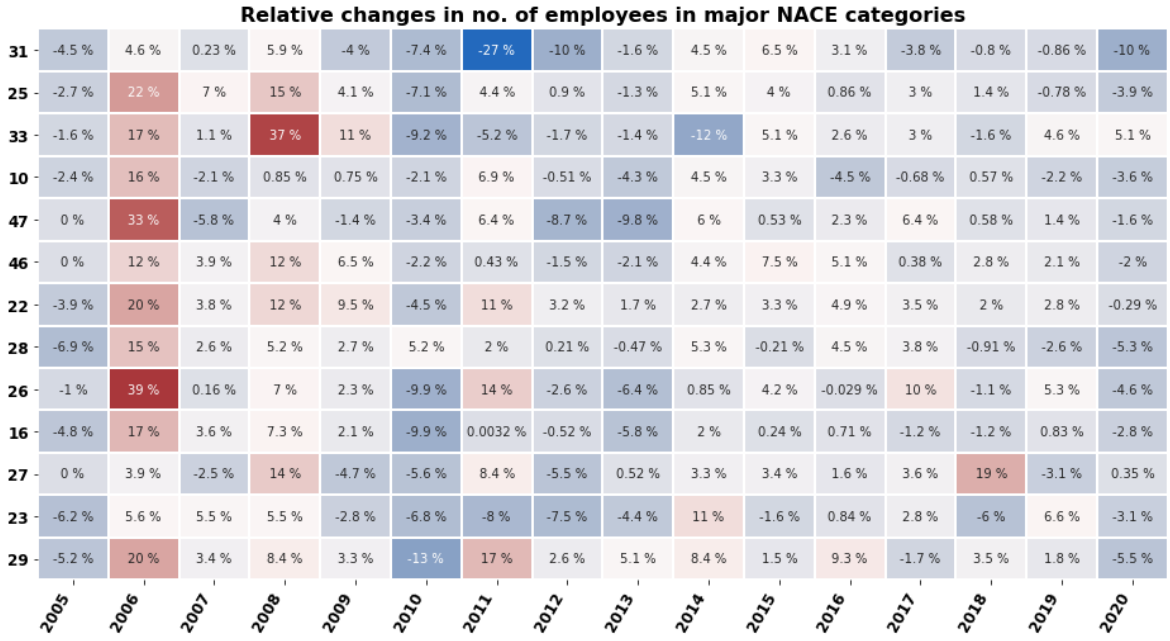



**Figure 5.15:** Relative changes in annually reported average employee numbers of NACE categories with the highest cardinality.

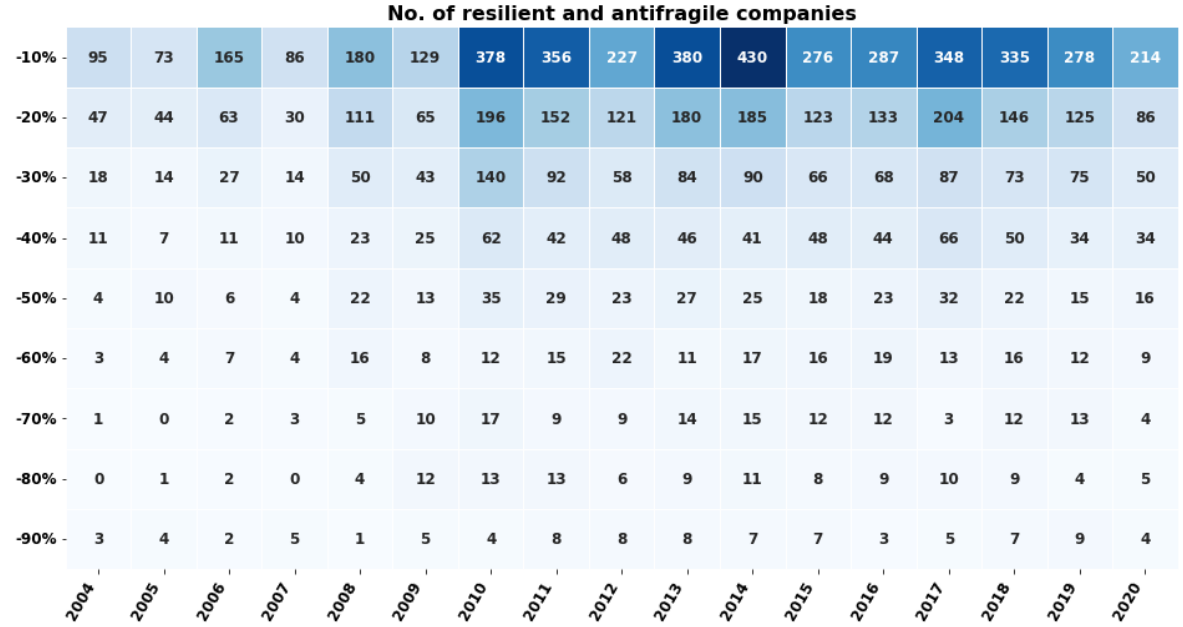



**Figure 5.16:** Number of companies that could manage their annual crisis in a resilient or antifragile way at a certain level of shock measured by the setback of their sales growth.

right corner represents the occasions of powerful- and effective reactions even for extreme negative impacts. The thorough investigation of the happenings aggregated in this section of the heatmap might contribute to the understanding of the general, fundamental factors of organizational resilience, however due to further company specific information it is beyond of the possibilities of the present research.

It might be not superfluous to note that on such aggregated visualisations (see Fig. 5.14., Fig. 5.16., Fig 5.17. and Fig. 5.18.) each processing industry member can be present more than once, since throughout their lifetimes companies might be exposed to economic shocks

several times.

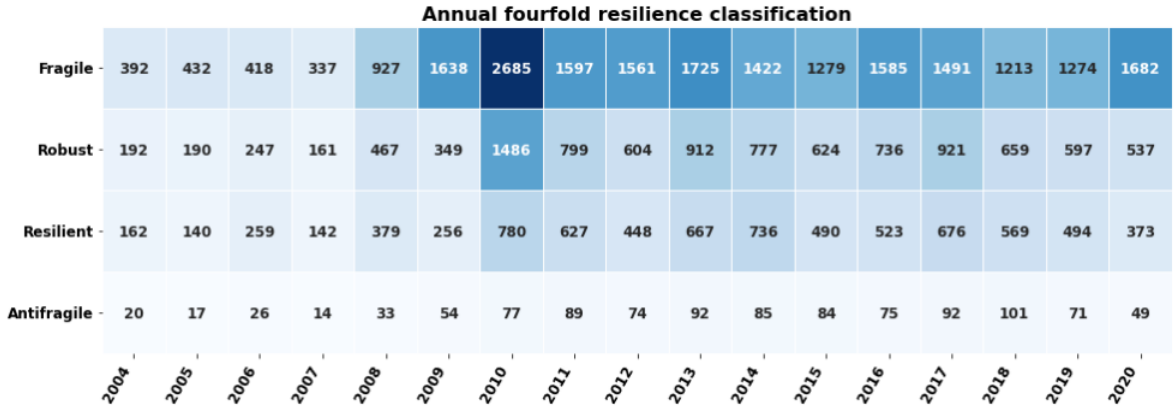



**Figure 5.17:** Annual characterization of the Hungarian processing industry via the improved and data-customized fourfold resilience maturity model.

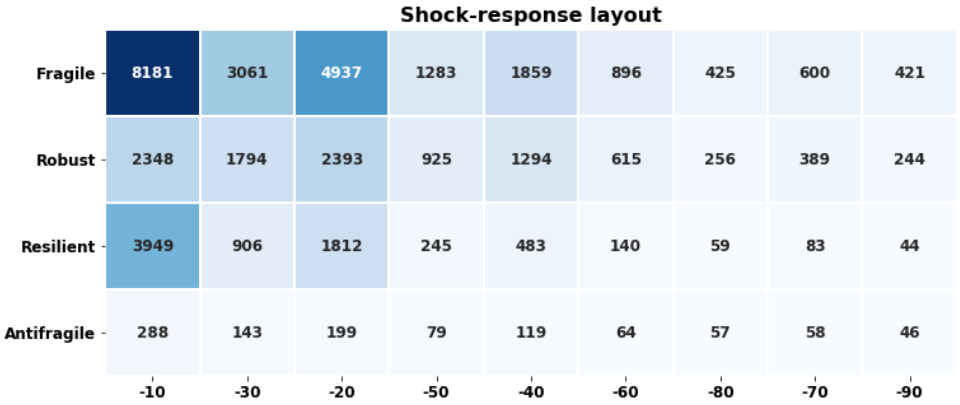



**Figure 5.18:** Crisis-influence and shock-response layout for the investigated time period at given shock levels.

### 5.4.1 Hypothesis testing and pair matching procedure

In order to collect best practices regarding shock preparation and survival the quantitative identification of favourable shock reactions (resilient and antifragile) is necessary. For this purpose the generated ”resilience history log” of the companies, as a classification enables the comparison of resilient and antifragile companies with those that have never had to face any economic disturbance greater than -10% during their lifetimes [R201].

In the present subsection the following seven hypotheses will be examined that – in accordance with overall literature attitude towards organizational resilience – assumes that companies that managed to successfully survive an economic downturn might learn from the gained experience and prosper later on:

**H1: Companies being resilient once lives longer, than companies that have never faced any shock.**

**H2: Companies being resilient once have higher annual net revenue in the long run, than companies that have never faced any shock.**

**H3: Companies being resilient once develop better regarding employee number in the long run, than companies that have never faced any shock.**

**H4: Companies being resilient once could cope with the next shock better, than non-resilient ones.**

**H5: Companies being resilient once have less probability of going bankrupt, than companies that have never faced any shock.**

**H6: Companies being resilient once have higher sales growth in the subsequent years of the shock, than companies that have never faced any shock.**

**H7: Companies being resilient once have less equity-to-asset ratio, than companies that have never faced any shock.**

In order to alleviate problems caused by missing data and the non-normal behaviour of the outlying distributions matched pair analysis was employed in a quasi-experimental setup that treated the economic shock as a treatment on each individual. In the matching procedure a similar process has been followed as it is outlined in [R198]:

- The members of each pair were selected randomly
- Each member within a pair were of the same industrial branch (same main NACE coding)
- Each member within a pair were of same size regarding employment categorization
- Before the year of shock the members had never had any other shocks
- Regarding revenue the difference between members were less than 10% in the year after the shock (when the resilient attribute manifested)

| Hypothesis | N | t-values | W-values | Mean (resilient/control) | Median (resilient/control) | $\chi^2$ (Df=1) | Odds ratio | Decision |
|---|---|---|---|---|---|---|---|---|
| **H1** | 1349 | -0.597 (0.275) | 12559 (0.036) | 5.8/5.9 (years) | 6.0/6.0 (years) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=2) | 1161 | -2.823 (0.002) | 184387 (0.000) | 320.1/395.2 (M HUF) | 129.5/179.3 (M HUF) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=5) | 766 | -5.564 (0.000) | 51920 (0.000) | 331.5/550.9 (M HUF) | 125.1/246.5 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=7) | 530 | -5.892 (0.000) | 24692 (0.000) | 482.1/721.5 (M HUF) | 134.6/314.3 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=10) | 225 | -4.788 (0.000) | 4170 (0.000) | 490.3/1008.5 (M HUF) | 187.5/415.9 (M HUF) | | | Rejected |
| **H3** ($\Delta$t=2) | 1164 | -1.121 (0.131) | 64034 (0.000) | 18.2/20.8 (-) | 7.0/14.5 (-) | N.A. | N.A. | Rejected |
| **H3** ($\Delta$t=5) | 769 | -4.212 (0.000) | 28920 (0.000) | 17.4/25.4 (-) | 7.0/14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=7) | 531 | -2.906 (0.002) | 11839 (0.000) | 20.4/30.8 (-) | 7.0/14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=10) | 228 | -1.841 (0.033) | 2351.5 (0.000) | 27.8/42.2 (-) | 14.5/34.5 (-) | | | Rejected |
| **H4** | 104 | N.A | N.A | N.A | N.A | 0.095 (0.758) | 1.153 (0.759) | Rejected |
| **H5** | 345 | N.A | N.A | N.A | N.A | 0.120 (0.729) | 1.198 (0.729) | Rejected |
| **H6** ($\Delta$t=1) | 1059 | -6.468 (0.000) | 177337 (0.000) | 6.0/20.5 (%) | -1.9/11.6 (%) | N.A. | N.A. | Rejected |
| **H6** ($\Delta$t=2) | 1048 | -2.761 (0.003) | 203523 (0.000) | 12.8/19.3 (%) | 3.3/13.1 (%) | | | Rejected |
| **H6** ($\Delta$t=3) | 1012 | -3.431 (0.000) | 188676 (0.000) | 10.9/18.4 (%) | 2.9/11.3 (%) | | | Rejected |
| **H6** ($\Delta$t=5) | 767 | -2.880 (0.002) | 111607 (0.000) | 10.0/16.8 (%) | 5.1/10.5 (%) | | | Rejected |
| **H6** ($\Delta$t=7) | 530 | 0.471 (0.681) | 60264 (0.002) | 14.2/12.6 (%) | 4.7/8.9 (%) | | | Rejected |
| **H6** ($\Delta$t=10) | 228 | 0.186 (0.574) | 11559 (0.067) | 4.7/4.0 (%) | -0.6/4.8 (%) | | | Failed to reject |
| **H7** ($\Delta$t=2) | 1095 | 4.201 (0.000) | 345400 (0.000) | 0.6/0.5 (-) | 0.6/0.5 (-) | N.A. | N.A. | Rejected |
| **H7** ($\Delta$t=5) | 730 | 3.975 (0.000) | 157209 (0.000) | 0.6/0.6 (-) | 0.6/0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=7) | 506 | 4.415 (0.000) | 78097 (0.000) | 0.6/0.6 (-) | 0.6/0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=10) | 220 | 3.979 (0.000) | 15914 (0.000) | 0.7/0.6 (-) | 0.7/0.6 (-) | | | Rejected |

**Table 5.3:** Corresponding statistical test results applied for testing each hypothesis with different time shifts measured from the date when each company showed resilient attribute ($\Delta$t = 0 corresponds to the first year after the shock phenomenon).

In case of testing H4, fragile and robust companies have been selected into the control group, while in the other cases companies that have never faced a shock greater than -10% throughout their lifetimes. According to Shapiro-Wilk tests, all the resulted sample distributions proved to be non-normal. Therefore, besides the one-tailed 2-sample t-tests, one-tailed Wilcoxon signed-rank tests (W-values) were also calculated in order to check the validity of the assumptions[4]. For H4 and H5 $\chi^2$ statistics were calculated with corresponding odds ratios. In Table 5.3. besides sample sizes (N) appropriate statistic values are given to different time shifts measured from the first year after the economic shock when the surviving items could show their resilient attribute. Furthermore, – where it was applicable – the mean and median values within the treated- and control groups are also listed to further support the decisions in each case.

Although, literature towards being resilient suggested a positive attitude, according to the results of match-pair analysis being resilient in the long-run does not necessarily mean any competitive advantage. On the contrary, resilient companies seem to be less viable and being resilient once does not mean being more resistant against future economic turbulences that might be of various origin either. Therefore, the present results do not contradict to the findings of [R198] where shocked companies turned out to be more prone to bankruptcy and future financial difficulties. Nevertheless, resilient companies as a subset of shocked economic stakeholders possess unique characteristics and significantly do not show higher chance to go bankrupt as non-shocked pairs within the control groups. Their performance seem to lag behind in the post-shock years.

Consequently, results suggest that despite surviving crises successfully, resilient companies loose economic performance and momentum that can be seen in 5-10 years perspective. This is in contradiction to everyday assumptions, since successful crisis management implies a lag compared to control group members. Nevertheless, it has to be emphasized that in the short-term, proper crisis reaction is inevitable for survival and long-term operation. Therefore, it is key to any further development perspectives and for the retrieval of workplaces, which also has numerous beneficial effects on regional level. The main message of the findings, that policymakers should consider resilient organizations as "wounded" in an economic sense and despite their momentary success their operation shall be monitored and in case of national economic interest prompt resources should be allocated to avoid fallback and to clear obstacles in the path of individual and regional economic growth.

## 5.5 Prospects of an "Early Warning System" for Resilience

The ultimate goal of resilience research would be to identify certain patterns of resilient companies as well as their time dependent reaction to crises and based on the gained relationships the creation of an "early-warning system" in order to direct economic actors towards a resilient way of crisis management [R202]. However, the scope of obtainable data on companies is relatively narrow and there are only few publications that report comprehensive analysis on longitudinal data gathered from industrial branches with a wide scope [R140].

In order to obtain such a system, predictive models shall be constructed. According to literature, classification models can be an alternative for this purpose that have already shown promising results in the field of bankruptcy prediction. Albeit, in case of bankruptcy prediction models primarily utilize balance sheet and income statement information likewise the goal is fundamentally different. In case of bankruptcy, the company tends towards its end-state, which – according to various literature sources, listed among others in Subsection. 5.1.6. – can be at a certain extent predicted already in a 3-years distance. Of course,

[4]In case of contradiction, the results of the Wilcoxon signed-rank tests were taken as normative.

models and results show great variability corresponding to the outlying data, therefore analyses should be regarded rather case-specific than generalizable for wider population.

In case of predicting resilient behaviour the problem in mathematical sense is relatively similar, therefore the analogy is quite straightforward. Nevertheless, in case of bankruptcy prediction we speak about a process, for "resilience prediction" a swift reaction to an unpredictable negative deviation in financial variables is sought. Therefore, it is not evident whether the time dependent company data extended with their metadata on industrial branch and geographical location contains the necessary information on predicting the labels of the company-year observations, labelled according to the constructed qualitative definition of resilience (see Sec. 5.3.). Moreover, for bankruptcy prediction more financial variables might tend towards a negative direction, while in our case only the sales growth data serves as a basis for the labelling process of each company-year observation.

To attempt the creation of a resilience prediction model, three cleaned datasets have been prepared from the original data sources. The first one contained those financial variables that were already involved in the balance sheets and income statements in raw form (*Data_-1*). Each company-year observation contained data $4$ years ahead to the idiosyncratic shock of the company (years *"i-4"*, *"i-3"*, *"i-2"*, *"i-1"*), the year of the shock occurred (year *"i"*) and a label according to the data of the consecutive year of the shock (year *"i+1"*, see Fig. 5.10.).

The considered financial variables were: (1) Net sales, (2) Operating income, (3) Profit after tax, (4) Fixed assets, (5) Current assets, (6) Liabilities, (7) Inventories, (8) Liquid assets, (9) Shareholders' equity, (10) Current receivables, (11) Current liabilities, (12) Long-term liabilities and (13) Profit or loss of the year.

The second constructed dataset contained derived financial ratios based on the above listed ones that are generally advised and used in bankruptcy prediction analyses [R188] (*Data_2*). Finally, the third dataset contained transitions of the raw financial variables from one year to the subsequent one in form of relative changes (*Data_3*). The variables of the three datasets were not mixed or pooled to a common dataset in order to avoid expressed multicollinearity. Missing values were handled by the Multiple Imputation using Chained Equations (MICE) algorithm.

Besides the employment data that represented the "size" of the company, other proxy variables were generated and included like the "closeness to urbanized regions" (see Fig. 5.7., where surroundings of urbanised regions were defined with a $10$km radius). The "age at crisis" variable was added in years dimension, the level of shock in percentages and the overall tendency of development in terms of net income of the companies were brought into the analysis in form of:

- *Number of crises before*: Number of crises deeper than -10% in the preceding 3 years to the given year when the actual crisis is investigated.
- *Effect of crises before*: The sum of the negative sales growth drops of the occurrent crises in the preceding 3 years (provided in percentage dimension) to the given year when the actual crisis is investigated.
- *Overall development before*: Average development based on annual sales growth rates taken the year *"i-4"* as a base year. Since annual sales growth values can have large values the one-year-, two-years-, three-years sales growth values measured from the base year of *"i-4"* were averaged as:

$$(SG(year_{i-4,i-3}) + SG(year_{i-4,i-2}) + SG(year_{i-4,i-1}))/3, \tag{5.4}$$

  where two-years- and three-years sales growth values measured from the base year of *"i-4"* can be calculated from the annual values as:

$$SG(year_{i-4,i-2}) = (SG(year_{i-4,i-3}) + 1) \cdot (SG(year_{i-3,i-2}) + 1) - 1 \tag{5.5}$$

$$SG(year_{i-4,i-1}) = (SG(year_{i-4,i-3}) + 1) \cdot (SG(year_{i-3,i-2}) + 1) \cdot (SG(year_{i-2,i-1}) + 1) - 1 \tag{5.6}$$

The aim of the classification purpose is to create a model that is able to predict the shock response of the company in year *"i+1"*, after the shock. Therefore, the three datasets (*Data_1*, *Data_2* and *Data_3*) were built up so that they contained the variables of the three preceding years as distinct feature vectors.

Logistic regression (LR) and Random Forest (RF) models had been used in order to classify labelled company-year observations regarding shock reaction. Because the variables showed non-normal distributions, for each dataset in case of collinearity questions the Mann-Whitney U-statistic had been calculated to select the more important variables to keep. (By this means it could be seen that the binary variable, which labelled the companies corresponding to their **within** city location has a significant effect on the classification, while the ranking of companies based on their vicinity to urbanized regions has less significant effect.)

The most important parameters of each dataset and the results served by the RF- and LR models are listed in Table 5.4. At each classification performance metric the mean values and standard deviations are provided that were gained on the test sets by performing a 10-fold cross validation.

| Classification Results for Resilient + Antifragile Observations | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 40,695 | 60 | 36,728 | 8774 | 31,921 | 0.75 ±0.02 | 0.48 ±0.02 | 0.55 ±0.04 | 0.42 ±0.01 | 0.74 ±0.01 | 0.64 ±0.01 | 0.36 ±0.02 | 0.43 ±0.03 | 0.31 ±0.01 | 0.67 ±0.01 |
| Data_2 | 40,695 | 132 | 115,660 | 8774 | 31,921 | 0.67 ±0.01 | 0.39 ±0.01 | 0.44 ±0.02 | 0.34 ±0.01 | 0.70 ±0.01 | 0.48 ±0.02 | 0.28 ±0.02 | 0.45 ±0.09 | 0.20 ±0.01 | 0.50 ±0.05 |
| Data_3 | 40,695 | 48 | 74,950 | 8774 | 31,921 | 0.74 ±0.01 | 0.46 ±0.02 | 0.49 ±0.03 | 0.43 ±0.02 | 0.75 ±0.01 | 0.57 ±0.01 | 0.34 ±0.01 | 0.50 ±0.06 | 0.26 ±0.01 | 0.58 ±0.04 |
| **Classification Results for Antifragile Observations** | | | | | | | | | | | | | | | |
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 40,695 | 60 | 36,728 | 1053 | 39,642 | 0.88 ±0.02 | 0.30 ±0.05 | 0.41 ±0.09 | 0.25 ±0.05 | 0.95 ±0.01 | 0.78 ±0.04 | 0.14 ±0.03 | 0.36 ±0.05 | 0.08 ±0.02 | 0.88 ±0.02 |
| Data_2 | 40,695 | 132 | 115,660 | 1053 | 39,642 | 0.83 ±0.02 | 0.21 ±0.04 | 0.28 ±0.05 | 0.17 ±0.04 | 0.94 ±0.01 | 0.53 ±0.03 | 0.05 ±0.01 | 0.35 ±0.08 | 0.03 ±0.00 | 0.67 ±0.05 |
| Data_3 | 40,695 | 48 | 74,950 | 1053 | 39,642 | 0.84 ±0.01 | 0.20 ±0.06 | 0.21 ±0.08 | 0.22 ±0.09 | 0.96 ±0.01 | 0.61 ±0.04 | 0.06 ±0.00 | 0.63 ±0.09 | 0.03 ±0.00 | 0.53 ±0.07 |

**Table 5.4:** Dataset info and metrics for classification performance evaluation for resilient + antifragile and for antifragile shock-reactions.

According to the results the RF models performed better than the LR. The RF models could serve with AUC>0.7 scores in some cases that can be a reason for hope considering similar analytical approaches in the field of economic resilience related studies. Nevertheless, beyond analysing observations that proved to be resilient in our approach, the antifragile subset of these observations could be classified with the same methods with higher AUC scores (see Fig. 5.19).

Unfortunately, it has to be noted that considering only the antifragile company-year observations the datasets to be classified became much more imbalanced. This can also be tracked in the deteriorating f1-scores compared to the classification of the resilient and antifragile company-year observations [R203]. This sheds light on the limitations of the present definition and classification attempt or on the possibilities that balance sheet and income data can offer for predicting resilient and/or antifragile shock reactions of economic organizations.

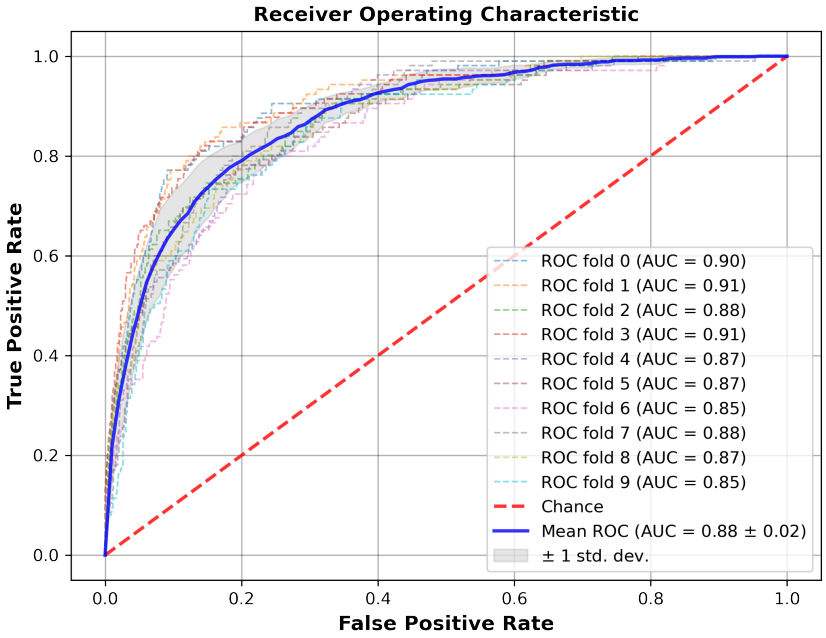



**Figure 5.19:** Receiver Operator Characteristics of 10-fold cross validation with the corresponding Area Under Curve Values for Data_1 classifying *"Antifragile* observations" with Random Forest technique.

As a conclusion it can be summarised that the range of accessible data in the present case with the applied methodology proved to be unsatisfactory for the prediction of resilient shock reactions based on the used Resilience Indicator Number. However, similar to bankruptcy prediction models, several variable combinations can be conceived and other indexes – even based on other economic or literature considerations – can be applied. Furthermore, the heterogeneity of the data could be restricted by applying machine learning based classification techniques to individual industrial branches or based on further considerations.

## 5.6 Improvement Opportunities by Examining Long-term Resilience

Although the view of one-year resilience is inspired by practice and seems to be a logical choice for shock-reaction evaluation it may also be interesting and perhaps not less

practically relevant to analyse longer term shock reactions.

For this purpose the present section aims the extension of the Resilience Indicator Number proposed in Sec.5.3. by considering shock reactions that allow a bouncing back effect 2,3 or 5 years after the shock occurrence instead of restricting itself only for the subsequent calendar year. Thereby an extended reaction to the idiosyncratic shocks can be observed. Longer time periods – albeit could be interesting to view – were omitted due to the length of the available time series and the possibility of further fluctuations within the sales growth within the selected time frame that could question the relevance of the reaction time stretching. The proposed extension of the previously used concept based on annual sales growth fluctuations for investigating prolonged economic shock reactions is sketched on Fig. 5.20.

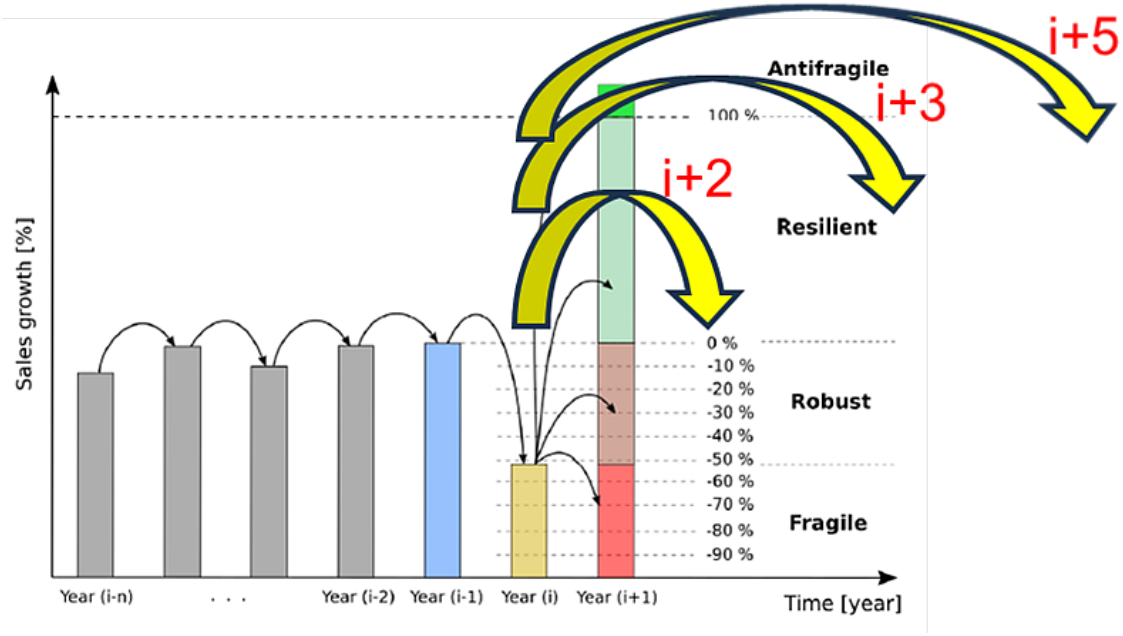



**Figure 5.20:** Extension of the one-year Resilience Indicator Number by allowing longer term (2, 3 and 5 years) shock reactions.

Accordingly, the hypothesis testing outlined in Subsec. 5.4.1. could be repeated straightforwardly.

By expanding and somewhat relaxing the formerly applied resilience indicator number, I have broadened the scope of resilient firm-year observations to include three additional similar data sets. The hypothesis tests have been conducted analogously on these data sets as well as had been formulated in the earlier stages of the research. These tests largely yielded results consistent with previous findings, wherein resilient behaviour was examined based on financial indicators from the first year following the crisis year. Accordingly, the matched-pair analysis indicated that companies that endured a crisis exhibited long-term lag relative to control group companies of similar size and revenue in the crisis year, which did not experience crises in that year. Regarding subsequent crisis management and growth potentials, I similarly observed no positive differences between companies capable of positively responding to shocks and their control pairs.

However, I did observe a deviation in one hypothesis. This concerned the likelihood of bankruptcy for companies that underwent a crisis and responded well. While short-term successful crisis responses did not indicate a lower probability of subsequent bankruptcy compared to control companies, this became increasingly evident over 2-, 3-, and 5-year time horizons for successfully responding companies, according to the calculated Chi-squared statistics and odds ratios (see Table 5.5., 5.6. and 5.7. In other words, companies that

performed well according to the relaxed indicator showed a kind of long-term, persistent resilience, appearing more intent on keeping the companies alive, which aligns better with the everyday notion of crisis resilience and was not observable in case of using the short-term, one-year resilience indicator.

Thus, the logical economic narrative derived from the subjective behaviour of individual companies – which cannot yet claim objective analysis from large-scale data – does not seem entirely unfounded. This suggests that crisis resilience is economically important and advantageous from all perspectives. Nevertheless, my previous analyses did not support this, as I had previously determined that the examined population of domestic small- and medium-sized manufacturing sector enterprises exhibited lag and required external support and nurturing. However, the current extension for prolonged crisis management, beyond the (at least short-term) retention of jobs, highlights the long-term ”survival” of suitable companies, which is also a significant economic factor through the retention of jobs.

| Hypothesis | N | t-values | W-values | Mean (resilient/control) | Median (resilient/control) | $\chi^2$ (Df=1) | Odds ratio | Decision |
|---|---|---|---|---|---|---|---|---|
| **H1** | 2180 | 0.443 (0.671) | 40746 (0.865) | 5.5/5.4 (years) | 5.0/5.0 (years) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=2) | 1847 | -2.374 (0.009) | 541614.5 (0.000) | 346.9/396.5 (M HUF) | 145.6/190.5 (M HUF) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=5) | 1189 | -5.545 (0.000) | 143657 (0.000) | 377.2/568.9 (M HUF) | 147.3/270.8 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=7) | 750 | -6.399 (0.000) | 47620 (0.000) | 414.9/714.5 (M HUF) | 165.1/350.1 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=10) | 151 | -3.344 (0.000) | 1866 (0.000) | 513.2/968.8 (M HUF) | 174.1/489.3 (M HUF) | | | Rejected |
| **H3** ($\Delta$t=2) | 1846 | -0.655 (0.256) | 187304.5 (0.000) | 19.8/20.8 (-) | 14.5 / 14.5 (-) | N.A. | N.A. | Rejected |
| **H3** ($\Delta$t=5) | 1191 | -4.551 (0.000) | 71072.5 (0.000) | 20.3/28.9 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=7) | 753 | -4.380 (0.000) | 26705.5 (0.000) | 22.9/34.5 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=10) | 152 | -2.669 (0.004) | 1373.5 (0.000) | 22.9/39.0 (-) | 14.5 / 34.5 (-) | | | Rejected |
| **H4** | 92 | N.A. | N.A. | N.A. | N.A. | 0.457 (0.499) | 1.332 (0.499) | Rejected |
| **H5** | 49 | N.A. | N.A. | N.A. | N.A. | 7.642 (0.005) | 0.488 (0.005) | Failed to reject |
| **H6** ($\Delta$t=1) | 407 | -4.148 (0.000) | 24655 (0.000) | 6.3 / 22.8 (%) | -3.6 / 11.1 (%) | N.A. | N.A. | Rejected |
| **H6** ($\Delta$t=2) | 468 | -3.988 (0.000) | 33954 (0.000) | 10.6 / 24.5 (%) | 0.9 / 14.8 (%) | | | Rejected |
| **H6** ($\Delta$t=3) | 515 | -2.721 (0.003) | 49121 (0.000) | 13.9 / 22.8 (%) | 4.7 / 13.9 (%) | | | Rejected |
| **H6** ($\Delta$t=5) | 536 | -0.301 (0.382) | 64237 (0.016) | 18.8 / 19.9 (%) | 8.2 / 11.5 (%) | | | Rejected |
| **H6** ($\Delta$t=7) | 410 | -0.292 (0.385) | 37868 (0.038) | 10.3 / 11.2 (%) | 5.3 / 8.8 (%) | | | Rejected |
| **H6** ($\Delta$t=10) | 80 | 0.396 (0.654) | 1373 (0.118) | 5.6 / 3.0 (%) | 0.2 / 4.9 (%) | | | Rejected |
| **H7** ($\Delta$t=2) | 515 | 3.311 (0.000) | 77305 (0.001) | 0.6 / 0.5 (-) | 0.6 / 0.5 (-) | N.A. | N.A. | Rejected |
| **H7** ($\Delta$t=5) | 515 | 1.726 (0.042) | 71500 (0.067) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=7) | 391 | 2.019 (0.022) | 42681 (0.026) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=10) | 76 | -0.663 (0.746) | 1378 (0.670) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |

**Table 5.5:** Corresponding statistical test results applied for testing each hypothesis with different time shifts measured from the date when each company showed resilient attribute ($\Delta$t = 0 corresponds to the second year after the shock phenomenon).

| Hypothesis | N | t-values | W-values | Mean (resilient/control) | Median (resilient/control) | $\chi^2$ (Df=1) | Odds ratio | Decision |
|---|---|---|---|---|---|---|---|---|
| **H1** | 2734 | 1.616 (0.947) | 79986.5 (1.000) | 5.0 /4.9 (years) | 5.0 / 5.0 (years) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=2) | 2158 | -3.474 (0.000) | 681358 (0.000) | 342.3 / 411.1 (M HUF) | 146.9 / 196.8 (M HUF) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=5) | 1356 | -5.276 (0.000) | 207438 (0.000) | 407.7 / 575.3 (M HUF) | 161.2 / 273.0 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=7) | 849 | -5.284 (0.000) | 78118 (0.000) | 492.0 / 749.1 (M HUF) | 195.4 / 351.9 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=10) | 153 | -2.566 (0.005) | 2266 (0.000) | 608.7 / 972.8 (M HUF) | 179.7 / 499.1 (M HUF) | | | Rejected |
| **H3** ($\Delta$t=2) | 2161 | -2.787 (0.003) | 189418 (0.000) | 19.2 / 22.0 (-) | 14.5 / 14.5 (-) | N.A. | N.A. | Rejected |
| **H3** ($\Delta$t=5) | 1357 | -3.250 (0.001) | 93261.5 (0.000) | 22.0 / 28.1 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=7) | 855 | -2.998 (0.001) | 36065 (0.000) | 26.2 / 35.3 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=10) | 153 | -4.875 (0.000) | 694 (0.000) | 21.7 / 40.5 (-) | 14.5 / 34.5 (-) | | | Rejected |
| **H4** | 63 | N.A. | N.A. | N.A. | N.A. | 0.657 (0.418) | 1.511 (0.417) | Rejected |
| **H5** | 657 | N.A. | N.A. | N.A. | N.A. | 40.071 (0.000) | 0.206 (0.000) | Failed to reject |
| **H6** ($\Delta$t=1) | 480 | -4.508 (0.000) | 36988 (0.000) | 7.8 / 21.6 (%) | 1.4 / 13.1 (%) | N.A. | N.A. | Rejected |
| **H6** ($\Delta$t=2) | 539 | -3.290 (0.001) | 48897 (0.000) | 12.7 / 23.8 (%) | 4.6 / 13.9 (%) | | | Rejected |
| **H6** ($\Delta$t=3) | 597 | -2.078 (0.019) | 68558 (0.000) | 15.3 / 22.0 (%) | 6.5 / 13.3 (%) | | | Rejected |
| **H6** ($\Delta$t=5) | 604 | -0.964 (0.168) | 82472 (0.019) | 11.4 / 13.8 (%) | 5.6 / 8.5 (%) | | | Rejected |
| **H6** ($\Delta$t=7) | 480 | 1.232 (0.891) | 51564 (0.021) | 11.5 / 6.6 (%) | 0.0 / 6.5 (%) | | | Rejected |
| **H6** ($\Delta$t=10) | 78 | 0.838 (0.798) | 1464 (0.352) | 11.1 / 4.4 (%) | -1.7 / 5.1 (%) | | | Rejected |
| **H7** ($\Delta$t=2) | 583 | 5.269 (0.000) | 105294 (0.000) | 0.6 / 0.5 (-) | 0.6 / 0.5 (-) | N.A. | N.A. | Rejected |
| **H7** ($\Delta$t=5) | 579 | 2.734 (0.003) | 95422 (0.002) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=7) | 463 | 2.570 (0.005) | 61255 (0.004) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=10) | 78 | -0.111 (0.544) | 1545 (0.491) | 0.6 / 0.6 (-) | 0.7 / 0.7 (-) | | | Rejected |

**Table 5.6:** Corresponding statistical test results applied for testing each hypothesis with different time shifts measured from the date when each company showed resilient attribute ($\Delta$t = 0 corresponds to the third year after the shock phenomenon).

| Hypothesis | N | t-values | W-values | Mean (resilient/control) | Median (resilient/control) | $\chi^2$ (Df=1) | Odds ratio | Decision |
|---|---|---|---|---|---|---|---|---|
| **H1** | 3236 | 2.131 (0.983) | 100211 (1.000) | 4.1 / 4.2 (years) | 4.0 / 4.0 (years) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=2) | 2533 | -1.817 (0.035) | 1243791 (0.000) | 384.0 / 421.1 (M HUF) | 164.5 / 187.4 (M HUF) | N.A. | N.A. | Rejected |
| **H2** ($\Delta$t=5) | 1397 | -3.136 (0.001) | 330595,5 (0.000) | 471.3 / 577.2 (M HUF) | 194.6 / 265.2 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=7) | 407 | -2.784 (0.003) | 23495 (0.000) | 467.1 / 639.4 (M HUF) | 186.1 / 317.4 (M HUF) | | | Rejected |
| **H2** ($\Delta$t=10) | 112 | -3.382 (0.000) | 1424 (0.000) | 474.2 / 955.5 (M HUF) | 259.2 / 514.1 (M HUF) | | | Rejected |
| **H3** ($\Delta$t=2) | 2543 | -0.949 (0.171) | 345006 (0.000) | 21.0 / 22.1 (-) | 14.5 / 14.5 (-) | N.A. | N.A. | Rejected |
| **H3** ($\Delta$t=5) | 1409 | -1.622 (0.052) | 139704 (0.000) | 25.6 / 29.3 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=7) | 409 | -3.300 (0.001) | 11569,5 (0.000) | 21.2 / 31.5 (-) | 14.5 / 14.5 (-) | | | Rejected |
| **H3** ($\Delta$t=10) | 113 | -2.744 (0.003) | 592 (0.000) | 23.2 / 44.4 (-) | 14.5 / 34.5 (-) | | | Rejected |
| **H4** | 47 | N.A. | N.A. | N.A. | N.A. | 0.056 (0.813) | 1.251 (0.813) | Rejected |
| **H5** | 671 | N.A. | N.A. | N.A. | N.A. | 48.332 (0.000) | 0.157 (0.000) | Failed to reject |
| **H6** ($\Delta$t=1) | 561 | -4.777 (0.000) | 54730 (0.000) | 6.7 / 21.4 (%) | 2.4 / 11.2 (%) | N.A. | N.A. | Rejected |
| **H6** ($\Delta$t=2) | 607 | -1.206 (0.114) | 78720 (0.001) | 9.2 / 12.6 (%) | 3.0 / 9.2 (%) | | | Rejected |
| **H6** ($\Delta$t=3) | 637 | 0.143 (0.557) | 93066 (0.033) | 17.2 / 16.7 (%) | 6.0 / 12.0 (%) | | | Rejected |
| **H6** ($\Delta$t=5) | 609 | -0.130 (0.448) | 85469 (0.044) | 12.5 / 12.8 (%) | 3.6 / 6.7 (%) | | | Rejected |
| **H6** ($\Delta$t=7) | 217 | -1.413 (0.079) | 9846 (0.016) | 1.7 / 7.4 (%) | -1.2 / 7.4 (%) | | | Rejected |
| **H6** ($\Delta$t=10) | 58 | 1.013 (0.843) | 758 (0.225) | 20.2 / 9.2 (%) | 4.6 / 10.9 (%) | | | Rejected |
| **H7** ($\Delta$t=2) | 621 | 5.045 (0.000) | 117671 (0.000) | 0.6 / 0.5 (-) | 0.6 / 0.5 (-) | N.A. | N.A. | Rejected |
| **H7** ($\Delta$t=5) | 582 | 3.036 (0.001) | 96028 (0.003) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=7) | 204 | 1.511 (0.066) | 11441 (0.121) | 0.6 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |
| **H7** ($\Delta$t=10) | 56 | 2.013 (0.023) | 1094 (0.008) | 0.7 / 0.6 (-) | 0.6 / 0.6 (-) | | | Rejected |

**Table 5.7:** Corresponding statistical test results applied for testing each hypothesis with different time shifts measured from the date when each company showed resilient attribute ($\Delta$t = 0 corresponds to the fifth year after the shock phenomenon).

It is also interesting to check previous results for the ”resilience-prediction” for the prolonged crisis reaction case. The attributes of the three constructed data sets after data cleaning and variable selection (which include annual corporate financial statements and ratio-based financial indicators derived from them, augmented with geographic, sectoral, and employment indicators) are detailed in Table 5.8., 5.9. and 5.10. together with corresponding performance metrics gained from Random Forest and Logistic Regression based classifications.

Classification was separately conducted for positive crisis responses, where a return to pre-crisis levels was achieved (antifragile + resilient), and for very strong crisis responses, defined by sales growth values at least doubling (100% increase) compared to pre-crisis levels. The results indicate that extending the time horizon of crisis reactions improved the classification results of the applied decision tree and logistic regression models. Alongside AUC values, f1 scores were also emphasized due to varying degrees of imbalanced classification in individual cases. Results were generated using layered 10-fold cross-validation.

Limiting the target group to antifragile crisis responses consistently resulted in $f1 < 0.7$, which I deemed insufficient or weak in quality. However, in cases of 3- and 5-year resilient behaviours, f1 scores greater than 0.7 (or close to this threshold) were also achieved multiple times (indicated by green in the tables). Furthermore, by thoroughly observing the results even a tendency can be seen similarly to the hypothesis test statistics. This trend shows that by enlarging the time period for the bouncing back the classification performance of the models tend to become better. Thus, being resilient in a longer term not just leads to a better survival by not going bankrupt but also to a more distinguished role within the population of company-year observations.

Overall, according to the performance metrics of Table 5.8., 5.9. and 5.10., using decision trees for ”*Data_2*” and ”*Data_3*” and logistic regression for ”*Data_1*” more satisfactory classification results could be achieved, providing a basis of confidence for constructing a ”corporate crisis management predictor” system based on companies’ financial- and metadata.

| Classification Results for Resilient + Antifragile Observations | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | No. of Samples | No. of Variable | No. of Missing Data | No. of True Labels | No. of False Labels | AUC | f1 Score | Recall | Precision | Accuracy | AUC | f1 Score | Recall | Precision | Accuracy |
| Data_1 | 36 883 | 57 | 1 988 | 11 063 | 25 820 | 0.76±0.02 | 0.56±0.02 | 0.63±0.04 | 0.51±0.01 | 0.71±0.01 | 0.81±0.02 | 0.65±0.01 | 0.73±0.03 | 0.58±0.01 | 0.76±0.01 |
| Data_2 | 36 883 | 129 | 78 748 | 11 063 | 25 820 | 0.81±0.01 | 0.63±0.01 | 0.69±0.03 | 0.59±0.01 | 0.76±0.01 | 0.60±0.01 | 0.47±0.02 | 0.73±0.08 | 0.35±0.01 | 0.51±0.03 |
| Data_3 | 36 883 | 45 | 37 034 | 11 063 | 25 820 | 0.82±0.01 | 0.64±0.01 | 0.70±0.03 | 0.60±0.01 | 0.77±0.01 | 0.70±0.02 | 0.54±0.01 | 0.82±0.02 | 0.40±0.01 | 0.58±0.02 |
| **Classification Results for Antifragile Observations** | | | | | | | | | | | | | | | |
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | No. of Samples | No. of Variable | No. of Missing Data | No. of True Labels | No. of False Labels | AUC | f1 Score | Recall | Precision | Accuracy | AUC | f1 Score | Recall | Precision | Accuracy |
| Data_1 | 36 883 | 57 | 1 988 | 1 934 | 34 949 | 0.85±0.02 | 0.36±0.05 | 0.40±0.10 | 0.34±0.03 | 0.93±0.01 | 0.85±0.02 | 0.33±0.02 | 0.66±0.05 | 0.22±0.01 | 0.86±0.01 |
| Data_2 | 36 883 | 129 | 78 748 | 1 934 | 34 949 | 0.88±0.01 | 0.41±0.04 | 0.47±0.09 | 0.36±0.03 | 0.93±0.01 | 0.44±0.03 | 0.07±0.01 | 0.29±0.13 | 0.04±0.01 | 0.61±0.10 |
| Data_3 | 36 883 | 45 | 37 034 | 1 934 | 34 949 | 0.87±0.01 | 0.42±0.03 | 0.51±0.08 | 0.36±0.04 | 0.93±0.01 | 0.79±0.01 | 0.14±0.01 | 0.88±0.03 | 0.08±0.00 | 0.43±0.03 |

**Table 5.8:** Dataset info and metrics for classification performance evaluation for resilient + antifragile and for antifragile shock-reactions considering 2 years for bouncing back from an economic setback.

| Classification Results for Resilient + Antifragile Observations | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 33 898 | 57 | 1 696 | 11 931 | 21 967 | 0.82±0.01 | 0.66±0.01 | 0.70±0.03 | 0.62±0.01 | 0.74±0.01 | 0.85±0.01 | 0.73±0.01 | 0.76±0.02 | 0.71±0.02 | 0.81±0.01 |
| Data_2 | 33 898 | 129 | 75 992 | 11 931 | 21 967 | 0.85±0.01 | 0.69±0.01 | 0.72±0.03 | 0.66±0.02 | 0.77±0.01 | 0.69±0.01 | 0.53±0.02 | 0.62±0.05 | 0.46±0.02 | 0.61±0.02 |
| Data_3 | 33 898 | 45 | 33 912 | 11 931 | 21 967 | 0.86±0.00 | 0.71±0.01 | 0.74±0.03 | 0.69±0.03 | 0.79±0.01 | 0.70±0.02 | 0.59±0.01 | 0.86±0.02 | 0.45±0.01 | 0.58±0.01 |
| **Classification Results for Antifragile Observations** | | | | | | | | | | | | | | | |
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 33 898 | 57 | 1 696 | 2 658 | 31 240 | 0.89±0.01 | 0.49±0.03 | 0.48±0.08 | 0.50±0.04 | 0.92±0.01 | 0.91±0.01 | 0.56±0.03 | 0.73±0.03 | 0.45±0.04 | 0.91±.0.1 |
| Data_2 | 33 898 | 129 | 75 992 | 2 658 | 31 240 | 0.90±0.01 | 0.52±0.02 | 0.53±0.07 | 0.52±0.06 | 0.92±0.01 | 0.52±0.04 | 0.14±0.02 | 0.55±0.12 | 0.08±0.01 | 0.49±0.09 |
| Data_3 | 33 898 | 45 | 33 912 | 2 658 | 31 240 | 0.90±0.01 | 0.52±0.04 | 0.55±0.06 | 0.51±0.08 | 0.92±0.02 | 0.82±0.01 | 0.19±0.01 | 0.95±0.02 | 0.11±0.00 | 0.38±0.03 |

**Table 5.9:** Dataset info and metrics for classification performance evaluation for resilient + antifragile and for antifragile shock-reactions considering 3 years for bouncing back from an economic setback.

| Classification Results for Resilient + Antifragile Observations | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 27 355 | 57 | 1 347 | 11 261 | 16 094 | 0.79±0.01 | 0.67±0.01 | 0.70±0.04 | 0.65±0.02 | 0.72±0.01 | 0.80±0.01 | 0.71±0.01 | 0.78±0.03 | 0.66±0.02 | 0.74±0.01 |
| Data_2 | 27 355 | 129 | 60 203 | 11 261 | 16 094 | 0.85±0.01 | 0.72±0.02 | 0.74±0.04 | 0.71±0.03 | 0.77±0.02 | 0.59±0.04 | 0.53±0.04 | 0.61±0.06 | 0.47±0.03 | 0.56±0.03 |
| Data_3 | 27 355 | 45 | 25 994 | 11 261 | 16 094 | 0.83±0.01 | 0.70±0.01 | 0.71±0.03 | 0.70±0.03 | 0.75±0.01 | 0.68±0.01 | 0.63±0.01 | 0.84±0.02 | 0.50±0.01 | 0.59±0.01 |
| **Classification Results for Antifragile Observations** | | | | | | | | | | | | | | | |
| | Data Info | | | | | Random Forest | | | | | Logistic Regression | | | | |
| | **No. of Samples** | **No. of Variable** | **No. of Missing Data** | **No. of True Labels** | **No. of False Labels** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** | **AUC** | **f1 Score** | **Recall** | **Precision** | **Accuracy** |
| Data_1 | 27 355 | 57 | 1 347 | 3 476 | 23 879 | 0.84±0.01 | 0.48±0.02 | 0.48±0.02 | 0.48±0.03 | 0.87±0.01 | 0.84±0.02 | 0.51±0.03 | 0.77±0.05 | 0.39±0.04 | 0.81±0.03 |
| Data_2 | 27 355 | 129 | 60 203 | 3 476 | 23 879 | 0.89±0.01 | 0.55±0.02 | 0.55±0.07 | 0.56±0.06 | 0.89±0.01 | 0.60±0.01 | 0.25±0.01 | 0.73±0.04 | 0.15±0.00 | 0.43±0.03 |
| Data_3 | 27 355 | 45 | 25 994 | 3 476 | 23 879 | 0.87±0.01 | 0.53±0.01 | 0.52±0.06 | 0.55±0.06 | 0.88±0.01 | 0.76±0.02 | 0.28±0.01 | 0.92±0.02 | 0.17±0.00 | 0.40±0.03 |

**Table 5.10:** Dataset info and metrics for classification performance evaluation for resilient + antifragile and for antifragile shock-reactions considering 5 years for bouncing back from an economic setback.

# 6

# Summary of Scientific Results, Outlook

I have applied the Most Frequent Value (MFV) concept in the field of economy related investigations together with non-parametric statistical investigations in order to refine existing results and generate new added value to the field with methodological development and application.

- It has been demonstrated that the MFV concept is applicable and practical in case of data analytical investigations also with economic background.
- A method has been developed, implemented and demonstrated for the isolation of outliers based on the MFV concept.
- It has been demonstrated, that economic convergence among regions of the European Union is less expressed when robust statistical analysis is applied for the investigation of the $\beta$-convergence theorem. Thereby, the importance of robust approaches have been demonstrated and emphasized in the field of regional economic studies.
- The k-Means algorithm was updated based on the MFV concept and applied on existing, open data sources in order to gain a more robust clustering algorithm version that relies on a more interpretable background than the k-Medians or k-Modes.
- Real-life data has been collected on Hungarian Small and Medium-sized Enterprises in order to investigate their resilient behaviour that is of great interest from national economic aspect.
- A Resilience Indicator Number – that is can be applied in a reproducible manner for similar data sources – has been created based on existing literature concepts and models that can utilize balance sheet and income statement data and offers a possibility to quantitatively analyse reactions to idiosyncratic economic shocks regardless of industrial branch or company size in a comparative way.
- Albeit numerous Resilience Indexes and Indicator Numbers are conceivable, a method has been presented for the quantitative investigation of company resilience that can be further elaborated corresponding to the economic model of selection.
- It has been shown that the resilient behaviour of economic entities has a real significance primarily in the short-term, but in the long-run an economic fallback and lag compared to those organisations is observable who managed not to get into economic setbacks. Since there are no means to prepare or forecast sudden, unpredictable negative events of all kinds this result does not mean that well-preparedness is essential, then it draws attention to the monitoring of key economic stakeholders even in case of

successful crisis management. With proper financial tools and adequate timing besides retrieving workplaces and maintaining organizational operation, long-term development trajectories can be set, and regional development can be promoted. However, it has to be accepted that resilient companies are still ”wounded” economic entities that need time to ”heal” and catch up in analogy with human injuries, for which an outside help might be advantageous.

# 7

# Conclusions

Non-normal data distributions and presence of outliers can still pose challenges to data scientists and statisticians while investigating processes from various walks of life. The interpretation of results gained from different algorithms can often be biased when such data distributions and anomalies are present. These are however abundant in case of real-life data and particularly significant when speaking of data of economic origin, where general, comprehensive models of the background processes are often not known or could not be used – or only partially – since the range and amount of accessible data are many times unsatisfactory. Keeping this in mind, in my dissertation I considered robust and non-parametric statistical techniques in order to handle outliers and non-normal distributions in case of economic data. As a robust technique, the *Most Frequent Value* (MFV) method was selected that was originally developed by Steiner et al. at the University of Miskolc, Hungary and was mainly adopted for statistical investigations of earth sciences related problems.

Throughout my work, I faced economic related problems with ambiguous standpoints within the accessed scientific literature, which I addressed with the robust MFV method and further non-parametric statistical tools. An algorithm for investigating linear regression problems based on the robust MFV procedure has been implemented based on the work of Steiner et al. and extended by considerations for identifying outlying observations. Furthermore, also utilizing the MFV technique a robust clustering method has been developed and implemented as an analogy of the prominent k-Means. The developed and implemented algorithms were able to serve with more reliable parameter estimates in case of investigated model data sets at hand. Nevertheless, they computationally underperform similar robust techniques, and consequently for the time being they might have relevance for smaller data sets or large datasets with high cluster cardinality.

With the applied robust and non-parametric techniques, I have strengthened the view of those in related literature who assume that economic convergence based on the absolute economic $\beta$-convergence among regions of the European Union does exist. Nevertheless, with the applied methodology, I have shown that the rate of convergence is somewhat smaller than forecasted by conventional approaches based on the Ordinary Least Squares. My findings were then further supported by population-level investigations and by the incorporation of social network analysis of R&D related Horizon 2020 project information from an external data source.

I have also addressed the question of resilient behaviour of companies based on large amount of balance sheet and income statement data extended by other meta information on the observations. This field is of special interest due to the growing interconnectedness and vulnerability of economic systems and their increasing exposure to turbulences of financial origin. I have proposed a data-based resiliency definition and via non-parametric statistical

tests I have revealed the short-term positive nature of resilient behaviour against idiosyncratic shocks regarding advantageous development perspectives compared to undisturbed control pairs. Additionally, I have proposed a simple indicator for comparing companies of various industrial sectors and of different shock-history regarding their shock reactions that might be used for further sales growth -based resiliency investigations.

Future research options of the work undertaken are abundant. Primarily, the extension of the MFV-based linear regression and clustering with multidimensional location- and scale parameter would be important. Thereby, parameter estimation considering multivariate distributions would be feasible that can be of high interest regarding theoretical aspect and economic related data investigations as well. On the other hand, despite our trials the application of the MFV-based approaches regarding resiliency related investigations did not prove to be reasonable. It is advised to further elaborate on this highly relevant and for national- and regional economy significant topic with additional non-parametric tests and machine learning based approaches. Nevertheless, my investigations showed that the data at hand were not satisfactory for the development of a predictive model as currently proposed in bankruptcy prediction related literature and therefore only *ex post* analysis of shock reactions were feasible. In the near future however, transaction data might be available that can enable the construction of supply chains and investigations via social network analysis. Shock- and shock reaction propagation might be feasible, may it be a positive or a negative reaction to idiosyncratic or even macroeconomic shock phenomena that – in my view – can offer new perspectives of research.

Appendix A

# Python implementation of MFV-based linear regression

## A.1 Most Frequent Value and Dihesion

```
import numpy as np
from scipy import stats
import math

def modified_MFV(y, threshold = 10**(-5)):
'''
Calculation of MFV and dihesion.
:y: List of numeric data
:threshold: Parameter for finetuning convergence limit.
:return: MFV and dihesion of the input data
'''
M_old = np.median(y)
eps_old = stats.median_abs_deviation(y)
n = len(y)
try:
    u_old = (y-M_old)/eps_old
    e0_old = 1/n*sum([1/(1+math.pow(x,2)) for x in u_old])
    e1_old = 1/n*sum([x/(1+math.pow(x,2)) for x in u_old])
    e2_old = 1/n*sum([math.pow(x,2)/(1+math.pow(x,2)) for x in u_old])

    i = 1
    diff = 1
    while abs(diff) > threshold:
        M_new = M_old + eps_old * e1_old/e0_old
        eps_new = eps_old * math.sqrt(1/e0_old-1)

        u_new = (y-M_new)/eps_new
        e0_new = 1/n*sum([1/(1+math.pow(x,2)) for x in u_new])
        e1_new = 1/n*sum([x/(1+math.pow(x,2)) for x in u_new])
        e2_new = 1/n*sum([math.pow(x,2)/(1+math.pow(x,2)) for x in u_new])

        diff = max( abs(M_new-M_old), abs(eps_new-eps_old) )

        M_old, eps_old = M_new, eps_new
        e0_old, e1_old, e2_old = e0_new, e1_new, e2_new

        i = i+1
except:
    raise ValueError
return M_new, eps_new
```

## A.2 2D case with Newton's method

```
import numpy as np
from scipy import stats
import math


def OLS_init(x,y):
    '''
    Initialization of slope- ('a') and intercept ('b') parameters for
    the linear regression line to be fitted. The function provides a
    simple OLS regression, where the weights of the MFV regression
    method are set to 1.
    :x: List of x-coordinates
    :y: List of y-coordinates
    :return: column vector of slope and intercept parameters
    '''
    x11, x12, x21, x22 = 0, 0, 0, 0
    b1, b2 = 0, 0
    for i in range(len(x)):
        x11 += x[i]**2
        x12 += x[i]
        b1 += x[i]*y[i]
        b2 += y[i]
    x21 = x12
    x22 = len(x)

    X = np.array([[x11, x12], [x21, x22]])
    b = np.array([b1, b2])
    return np.dot(np.linalg.inv(X), b)

def eps_init(x,y):
    '''
    Initialization of dihesion for the MFV-regression. The function uses
    the biggest residuals in their absolute values of the original data
    set measured from the OLS regression line gained from the 'OLS_init' function.
    :x: List of x-coordinates
    :y: List of y-coordinates
    :return: initial value for dihesion (float)
    '''
    initParams = OLS_init(x,y)
    a = initParams[0]
    b = initParams[1]
    resid_max = 0
    resid_min = 0
    for i in range(len(x)):
        resid = y[i] - (a*x[i] + b)
        if resid > resid_max:
            resid_max = resid
        if resid < resid_min:
            resid_min = resid
    return abs(resid_max - resid_min)

def calcJacobi(x, y, a_old, b_old, eps_old):
    '''
    Calculation of Jacobian matrix.
    :x: List of x-coordinates
    :y: List of y-coordinates
    :a_old: slope parameter of regression line
    :b_old: intercept parameter of regression line
    :eps_old: dihesion
    :return: Jacobian matrix of nonlinear system
    '''
    J11, J12, J21, J22 = 0, 0, 0, 0
    f1, f2 = 0, 0
    for i in range(len(x)):
        nominator = eps_old**2 - (y[i]-a_old*x[i]-b_old)**2
        denominator = (eps_old**2 + (y[i]-a_old*x[i]-b_old)**2)**2
        A = nominator/denominator
        J11 += x[i]**2 * A
        J12 += x[i] * A
        J21 += x[i] * A
        J22 += A
    J = np.array([[J11, J12], [J21, J22]])
    return J
```

```
def calcF(x, y, a_old, b_old, eps_old):
    '''
    Calculation of f-fector for nonlinear system.
    :x: List of x-coordinates
    :y: List of y-coordinates
    :a_old: slope parameter of regression line
    :b_old: intercept parameter of regression line
    :eps_old: dihesion
    :return: f-vector
    '''
    f1, f2 = 0, 0
    for i in range(len(x)):
        f1 += -x[i]*(y[i]-a_old*x[i]-b_old)/(eps_old**2+(y[i]-a_old*x[i]-b_old)**2)
        f2 += -(y[i]-a_old*x[i]-b_old)/(eps_old**2+(y[i]-a_old*x[i]-b_old)**2)
    f = np.array([f1, f2])
    return f

def MFV_regression_newton(x,y):
    '''
    Estimation of MFV-robustified regression line parameters by utilizing
    Newton's method as nonlinear solver.
    :x: List of x-coordinates
    :y: List of y-coordinates
    :return: slope, intercept and dihesion parameters for the fitted line.
    '''
    #Initialization
    a_old = OLS_init(x,y)[0]
    b_old = OLS_init(x,y)[1]
    eps_old = eps_init(x,y)

    diff_MFV = 10**6
    noOfMFViterations = 0
    j=0

    ind,a,b,eps=list(),list(),list(),list()
    while diff_MFV > 10**(-5):
        noOfMFViterations += 1
        diff = 10**6
        while diff > 10**(-5):
            #Newton's algorithm -> in each step a linear system has to be solved
            a_vec_old = np.array([a_old, b_old])
            J = calcJacobi(x, y, a_vec_old[0], a_vec_old[1], eps_old)
            f = calcF(x, y, a_vec_old[0], a_vec_old[1], eps_old)
            a_vec_new = a_vec_old - np.dot(np.linalg.inv(J), f)

            #for the stop condition of the nonlinear solver
            diff = max(abs(a_vec_old[0]-a_vec_new[0]),abs(a_vec_old[1]-a_vec_new[1]))
            a_old = a_vec_new[0]
            b_old = a_vec_new[1]

        eps_nominator, eps_denominator = 0, 0
        for i in range(len(x)):
            eps_nominator += (y[i]-a_old*x[i]-b_old)**2/ \
                             (eps_old**2+(y[i]-a_old*x[i]-b_old)**2)**2
            eps_denominator += 1 / (eps_old**2 + (y[i] - a_old*x[i]-b_old)**2)**2

        eps_new = np.sqrt(3*eps_nominator/eps_denominator)
        diff_MFV = abs(eps_old - eps_new)
        eps_old = eps_new
        j=j+1
    return a_old, b_old, eps_old
```

## A.3 2D case with Broyden's method

```
def MFV_regression_broyden(x,y):
'''
Estimation of MFV-robustified regression line parameters by utilizing
Newton's method as nonlinear solver.
:x: List of x-coordinates
:y: List of y-coordinates
:return: slope, intercept and dihesion parameters for the fitted line.
'''
#Initialization
a_old = OLS_init(x,y)[0]
b_old = OLS_init(x,y)[1]
eps_old = eps_init(x,y)
j=0

ind,a,b,eps=list(),list(),list(),list()
diff_MFV = 10**6
while diff_MFV > 10**(-5):
    x_old = np.array([a_old, b_old])
    A_old = calcJacobi(x, y, x_old[0], x_old[1], eps_old)
    f_old = calcF(x, y, x_old[0], x_old[1], eps_old)

    diff = 10**6
    while diff > 10**(-5):
        s = np.linalg.solve(A_old, -f_old)
        x_new = x_old + s
        f_new = calcF(x, y, x_new[0], x_new[1], eps_old)

        #for the stop condition of the nonlinear solver
        diff = max( abs(x_new[0]-x_old[0]), abs(x_new[1]-x_old[1]))

        delta_x = s
        delta_x_norma = np.dot(delta_x, delta_x)
        delta_f = f_new - f_old

        #Creation of Jacobian matrix
        A_new = A_old+(np.outer((delta_f-np.dot(A_old,delta_x)),delta_x))/ \
                delta_x_norma

        x_old = x_new
        f_old = f_new
        A_old = A_new
        a_old = x_new[0]
        b_old = x_new[1]

    eps_nominator, eps_denominator = 0, 0
    for i in range(len(x)):
        eps_nominator += (y[i]-a_old*x[i]-b_old)**2/ \
        (eps_old**2+(y[i]-a_old*x[i]-b_old)**2)**2
        eps_denominator += 1/(eps_old**2+(y[i]-a_old*x[i]-b_old)**2)**2

    eps_new = np.sqrt(3*eps_nominator/eps_denominator)
    diff_MFV = abs(eps_old - eps_new)
    eps_old = eps_new

    if eps_old <0.01:
        return a_old, b_old, eps_old
    j=j+1
return a_old, b_old, eps_old
```

Appendix B

# Python implementation of MFV-based clustering

## B.1 k-MFVs

```
class Kmeans:
    '''Implementing Kmeans algorithm.'''

    def __init__(self, n_clusters, max_iter=100, random_state=123, method = "mean", \
    init = None):
        self.n_clusters = n_clusters # cluster number
        self.max_iter = max_iter # no. of maximum iterations
        self.random_state = random_state
        self.method = method # ("mean", "median", "MFV")
        self.init = init # initialization possibility by user

    def initializ_centroids_random(self, X):
        '''
        Providing coordinates of centroids around data point by random selection.
        '''
        np.random.RandomState(self.random_state) # initialization of random state
        random_idx = np.random.permutation(X.shape[0]) # shuffling of data points
        centroids = X[random_idx[:self.n_clusters]] # selection of centroids
        return centroids

    def plot(self, data, centroids):
        '''
        Visualisation of data points and centroids in 2 dimensions.
        '''
        plt.scatter(data[:, 0], data[:, 1], marker = '.', color = 'gray', \
        label = 'data points')
        plt.scatter(centroids[i-1, 0], centroids[i-1, 1], color = 'black', \
        label = 'previously selected centroids')
        plt.scatter(centroids[i, 0], centroids[i, 1], color = 'red', \
        label = 'next centroid')
        plt.title('Select % d th centroid'%(centroids.shape[0]))

        plt.legend()
        plt.show()

    # initialization algorithm
    def initializ_centroids(self, X):
        '''
        K-means++ initialization of cluster centroids.
        '''
        centroids = np.zeros((self.n_clusters, X.shape[1])) # zero-matrix for centroids
        # Step 1.: random selection of first centroid
        centroids[0] = X[np.random.randint(X.shape[0]),:]  # random initialization
        k = 0
        # Step 2.: selection of second centroid in the farthest possible position
        # compared to the first centroid
        temp = self.compute_distance(X, centroids)
        distancesFromCentroids = temp[:, k] # distance calculation among data and \
        #first centroid
        pointIndex = np.argsort(distancesFromCentroids)[-1] # finding index of \
```

```
        #farthest data point
        centroids[k+1] = X[pointIndex, :] # coordinates of farthest data point

        for k in range(1, self.n_clusters):
            # Step 3.: clustering with resulted centroids
            temp2 = self.compute_distance(X, centroids)
            distancesFromCentroids = temp2[:, :(k+1)] # calculation of distances \
            #among data points and [0..k.]-th centroids
            labels = self.find_closest_cluster(distancesFromCentroids) # point \
            # assignment to existing centroids
            # Step 4.: estimation of new centroids based on max. inter-cluster distances
            max=0
            for j in range(0, k+1):
                # calculation of index corresponding to max. distance in cluster j
                try: # it might happen that there are no points in the cluster
                    maxDist = np.amax(distancesFromCentroids[:, j][labels==j], axis=0)
                    ind = np.where(distancesFromCentroids[:, j] == maxDist)[0][0]
                    if distancesFromCentroids[ind][j] > max:
                        max = distancesFromCentroids[ind][j]
                        pointIndex = ind
                except:
                    print('ERROR! No points belonging to initialized centroid.')
            if (k < self.n_clusters-1):
                centroids[k+1] = X[pointIndex, :] # coordinates of farthest point
                #plot(X,centroids[:(k+2)]) # visualisation of centroids
        return centroids

    def compute_centroids(self, X, centroids, labels):
        '''
        Calculation of new cluster centroids after labelling data points.
        '''
        centroids2 = np.zeros((self.n_clusters, X.shape[1])) # storing of centroids
        if self.method == "mean":
            for k in range(self.n_clusters): # point assignment according to labels
                centroids2[k, :] = np.mean(X[labels == k, :], axis=0)
        if self.method == "median":
            for k in range(self.n_clusters): # point assignment according to labels
                centroids2[k, :] = np.median(X[labels == k, :], axis=0)
        if self.method == "MFV":
            for k in range(self.n_clusters): # point assignment according to labels
                MFV_values = list()
                coordinates = list()
                for variable in range(X.shape[1]): # calculation of MFV values per \
                #dimension (x,y,z etc.)
                    for point in X[labels == k, :]: # calculation of MFV per cluster
                        coordinates.append(point[variable])
                    if len(coordinates) > 1: # treatment of low cluster cardinality
                        mfv = MFV(coordinates, k = 2, isPlot = True)[0]
                    elif len(coordinates) == 1:
                        mfv = coordinates[0]
                    else: #if len(coordinates) == 0:
                        mfv = centroids[k,variable]
                    MFV_values.append(mfv) #(MFV_x, MFV_y, MFV_z etc.)
                    coordinates = [] # new list of coordinates per dimension (x,y,z etc.)
                centroids2[k, :] = np.array(MFV_values)
        return centroids2

    def compute_distance(self, X, centroids):
        '''
        Calculation of Euclidean distances among centroids and data points (L2 norm).
        '''
        distance = np.zeros((X.shape[0], self.n_clusters)) # for storing of distances \
        # among every point and centroid
        for k in range(self.n_clusters):
            row_norm = norm(X - centroids[k, :], axis=1) # calculation of distances \
            # among every point and centroid
            distance[:, k] = np.square(row_norm)
        return distance

    def find_closest_cluster(self, distance):
        '''
        Ordering of data points to the nearest cluster centroids.
        '''
        labels = np.argmin(distance, axis=1) # finding index of smallest element in \
        # the list that is in range of [0..n_cluster] for cluster labelling)
        return labels
```

```
    def compute_sse(self, X, labels, centroids):
        '''
        Calculation of sum of distances within clusters in order to evaluate the \
        "goodness" of clustering.
        '''
        distance = np.zeros(X.shape[0])
        for k in range(self.n_clusters):
            distance[labels == k] = norm(X[labels == k] - centroids[k], axis=1)
        return np.sum(np.square(distance))

    def fit(self, X):
        '''
        Execution of clustering to the given data points. The function calculates \
        centroid coordinates and updates them as a side effect.
        '''
        if self.init is not None:
            self.centroids = self.init
        else:
            self.centroids = self.initializ_centroids(X) # cluster centroids at \
            # initialization
        for i in range(self.max_iter): #max. number of iterations
            old_centroids = self.centroids
            distance = self.compute_distance(X, old_centroids)
            # print('Number of centroid swaps:', i+1)
            self.labels = self.find_closest_cluster(distance) # data point labelling
            # coordinates of new cluster centroids
            self.centroids = self.compute_centroids(X, old_centroids, self.labels)
            if np.all(old_centroids == self.centroids): # stop if no centroids change
                break
        self.error = self.compute_sse(X, self.labels, self.centroids)

    def predict(self, X):
        '''
        Specification of cluster labels of each data point after "max_iter" iterations.
        '''
        distance = self.compute_distance(X, self.centroids)
        return self.find_closest_cluster(distance)
```

```
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
import random
import time
import scipy
import sklearn

def run_kMFV(df, classLabels, n_clusters, method):
    '''
    Example run of MFV based clustering on known dataset for algorithm validation.
    df: input DataFrame
    classLabels: known pre-existing labels for the calculation of validity indexes
    n_clusters: number of clusters we would like to see
    method: 'mean', 'median', 'MFV'
    '''
    df = np.array(df)
    classLabels = np.array(classLabels)

    km = Kmeans(n_clusters = n_clusters,
                    max_iter = 30,
                    random_state = np.random.randint(0, 1000, size = 1),
                    method = method,
                    init = initByDBSCAN(df, n_clusters)) # initialization of instance


    time0 = time.time()
    km.fit( df ) # run of instance
    time1 = time.time()
    print('Time necessary for calculations: {0:.4f}s'.format(time1-time0))
    #km.centroids
    #km.labels
    unique, counts = np.unique(km.labels, return_counts=True)
```

```
print('Cardinality of clusters: ', counts)

print('Cluster validity indexes:')
print('Adjusted Rand Score:', adjusted_rand_score( km.labels, classLabels ) )
print('s_Dbw score:', S_Dbw(df, labels = km.labels, centers_id=None, \
            method='Tong', alg_noise='bind', centr='mean', \
            nearest_centr=True, metric='euclidean') )
print('Adjusted mutual information score: ', \
adjusted_mutual_info_score(km.labels, classLabels) )

print('Silhouette score: ', silhouette_score(df, km.labels, metric='euclidean') )
print('Davis-Bouldin score: ', davies_bouldin_score(df, km.labels) )
print('----------------------------------------------------------')
```

# Bibliography


[R1] Gregory W. Corder and Dale I. Foreman. *Nonparametric Statistics, A Step-by-Step Approach*. Hoboken, New Jersey: Wiley, 2014. ISBN: 978-1-118-84031-3.

[R2] Elvezio Ronchetti. "The Historical Development of Robust Statistics". In: *Proceedings of the 7th International Conference on Teaching Statistics (ICOTS-7)* (Jan. 2006). URL: https://www.researchgate.net/publication/228434224_The_historical_development_of_robust_statistics.

[R3] Stephan Morgenthaler. "A survey of robust statistics". In: *Statistical Methods and Applications* 15 (Jan. 2007), pp. 271–293. DOI: https://doi.org/10.1007/s10260-006-0034-4.

[R4] Pál Kerékfy. "About Robust Estimates". In: *Alkalmazott Matematikai Lapok* 4 (1978), pp. 327–357.

[R5] Csilla Csendes. "Robust Parameter Estimation Procedure of Symmetric Stable Distributions and its Application". Available at http://www.hjphd.iit.uni-miskolc.hu/images/ertekezesek/2014/CsendesCsilla/disszertacio.pdf. PhD thesis. Miskolc, Hungary: University of Miskolc, 2014.

[R6] F. Steiner. "The Bases of Geostatistics (In Hungarian)". In: (1990). Tankönyvkiadó, Budapest, Hungary, 363p., ISBN: 963 18 2819 0.

[R7] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, New York, Second Edition, Sept. 2011. ISBN: 978-1-118-21033-8.

[R8] Peter Szucs and Faruk Civian. "Multi-layer well log interpretation using the simulated annealing method". In: *Journal of Petroleum Science and Engineering* 14.3–4 (1996), pp. 209–220. DOI: https://doi.org/10.1016/0920-4105(95)00048-8.

[R9] Peter Szucs, Faruk Civan, and Margit Virag. "Applicability of the most frequent value method in groundwater modeling". In: *Hydrogeology Journal* 14.1 (Jan. 2006), pp. 31–43. DOI: https://doi.org/10.1007/s10040-004-0426-1.

[R10] Norbert Péter Szabó, Gergely Pál Balogh, and János Stickel. "Most frequent value-based factor analysis of direct-push logging data". In: *Geophysical Prospecting* 66.4 (Mar. 2018), pp. 530–548. DOI: https://doi.org/10.1111/1365-2478.12573.

[R11] László Ferenczy and Ferenc Steiner. "Method of the most frequent values in the well-log interpretation (in Hungarian)". In: *Magyar Geofizika* 29.3 (1988), pp. 95–103.

[R12] J. Zhang. “Most Frequent Value Statistics and the Hubble Constant”. In: *The Astronomical Society of the Pacific* 130.990 (Aug. 2018), p. 084502. DOI: https://doi.org/10.1088/1538-3873/aac767.

[R13] J. Zhang. “Most frequent value statistics and distribution of 7Li abundance observations”. In: *Monthly Notices of the Royal Astronomical Society* 468.4 (July 2017), pp. 5014–5019. DOI: https://doi.org/10.1093/mnras/stx627.

[R14] Ferenc Steiner. “(Editor), Optimum Methods in Statistics”. In: (1997). Akadémiai Kiadó, Budapest, Hungary, 370p., ISBN: 963 05 7439 X.

[R15] Sándor Fegyverneki. “Modification of the Most Frequent Value and Dihesion”. In: *Multidisciplinar Sciences* 9.4 (2019), pp. 455–459. DOI: https://doi.org/10.35925/j.multi.2019.4.47.

[R16] Ferenc Nagy. “Parameter Estimation of the Cauchy Distribution in Information Theory Approach”. In: *Journal of Universal Computer Science* 12.9 (2006), pp. 1332–1344. DOI: https://doi.org/10.1016/j.eswa.2011.09.018.

[R17] Kateina Dvoroková. “Sigma Versus Beta-convergence in EU28: Do they lead to different results?” In: *WSEAS Transactions on Business and Economics* 11.1 (Jan. 2014), pp. 314–321. URL: https://www.wseas.org/multimedia/journals/economics/2014/a185707-228.pdf.

[R18] R. F. Harrod. “An Essay in Dynamic Theory”. In: *The Economic Journal* 49.193 (Mar. 1939), pp. 14–33. DOI: https://doi.org/10.2307/2225181.

[R19] Robert J. Barro and Xavier Sala-i-Mart. “Convergence”. In: *Journal of Political Economy,* 100.2 (Apr. 1992), pp. 223–251. DOI: https://doi.org/10.1086/261816.

[R20] Robert M. Solow. “A Contribution to the Theory of Economic Growth”. In: *The Quarterly Journal of Economics* 70.1 (Feb. 1956), pp. 65–94. DOI: https://doi.org/10.2307/1884513.

[R21] Nazrul Islam. “What Have We Learnt from the Convergence Debate?” In: *Journal of Economic Surveys* 17.3 (June 2003), pp. 309–362. DOI: https://doi.org/10.1111/1467-6419.00197.

[R22] Paul M. Romer. “Increasing Returns and Long-Run Growth”. In: *Journal of Political Economy* 94.5 (Oct. 1986), pp. 1002–1037. DOI: https://doi.org/10.1086/261420.

[R23] Robert E. Lucas Jr. “On the mechanics of economic development”. In: *Journal of Monetary Economics* 22.1 (July 1988), pp. 3–42. DOI: https://doi.org/10.1016/0304-3932(88)90168-7.

[R24] Mihály Tamás Borsi and Norbert Metiu. “The evolution of economic convergence in the European Union”. In: *Empirical Economics* 38.2 (Mar. 2014). DOI: https://doi.org/10.1007/s00181-014-0801-2.

[R25] Danny T. Quah. “Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs”. In: *Journal of Economic Growth* 2.1 (1997), pp. 27–59. DOI: https://doi.org/10.1023/A:1009781613339.

[R26] Mihály Tamás Borsi and Norbert Metiu. “The evolution of economic convergence in the European Union”. In: *Empirical Economics, Springer* 48.2 (Mar. 2015), pp. 657–681. DOI: https://doi.org/10.1007/s00181-014-0801-2.

[R27] Andrew B. Bernard and Steven N. Durlauf. “Convergence in international output”. In: *Journal of Applied Econometrics* 10.2 (June 1995), pp. 97–108. DOI: https://doi.org/10.1002/jae.3950100202.

[R28] Konstantin Gluschenko. “Myths about Beta-Convergence”. In: *William Davidson Institute Working Papers Series* 4 (Nov. 2012), pp. 26–44. DOI: https://doi.org/10.2139/ssrn.2188430.

[R29] Wojciech Dyba et al. “Regional Development in Central-Eastern European Countries at the Beginning of the 21st Century: Path Dependence and Effects of EU Cohesion Policy”. In: *Quaestiones Geographicae* 37.2 (Feb. 2018), pp. 77–92. DOI: https://doi.org/10.2478/quageo-2018-0017.

[R30] Mindaugas Butkus et al. “What Is the Evolution of Convergence in the EU? Decomposing EU Disparities up to NUTS 3 Level”. In: *Sustainability* 10.5 (May 2018), pp. 1–37. DOI: https://doi.org/10.3390/su10051552.

[R31] Angelos Liontakis, Christos T. Papadas, and Irene Tzouramani. “Regional Economic Convergence in Greece: A Stochastic Dominance Approach”. In: *50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy"* Conference Paper (Aug. 2010). URL: https://www.econstor.eu/handle/10419/119147.

[R32] Danny T. Quah. “Twin Peaks: Growth and Convergence in Models of Distribution Dynamics”. In: *The Economic Journal* 106.437 (July 1996), pp. 1045–1055. DOI: https://doi.org/10.2307/2235377.

[R33] Danny Quah. “Galtons Fallacy and Tests of the Convergence Hypothesis”. In: *Scandinavian Journal of Economics* 95.4 (Dec. 1993), pp. 427–443. DOI: https://doi.org/10.2307/3440905.

[R34] Roberto Ezcurra and Manuel Rapún. “Regional Dynamics and Convergence Profiles in the Enlarged European Union: A Non-parametric Approach”. In: *Journal of Economic and Human Geography* 98.5 (Dec. 2007), pp. 564–584. DOI: https://doi.org/10.1111/j.1467-9663.2007.00426.x.

[R35] Nikolay Nenovsky and Kiril Tochkov. “The Distribution Dynamics of Income in Central and Eastern Europe relative to the EU: A Nonparametric Analysis”. In: *William Davidson Institute Working Paper* 1063 (Nov. 2013). DOI: https://doi.org/10.2139/ssrn.2370625.

[R36] Gordon Anderson. “Making Inferences about the Polarization, Welfare and Poverty of Nations: A Study of 101 Countries 1970-1995”. In: *Journal of Applied Economics* 19.5 (May 2004), pp. 537–550. DOI: https://doi.org/10.1002/jae.750.

[R37] Monica Rileanu Szeles. “Exploring the Economic Convergence in the EU’s new Member States by using non-parametric models”. In: *Romanian Journal on Economic Forecasting* 14.1 (2011), pp. 20–40. URL: https://ipe.ro/rjef/rjef1_11/rjef1_2011p20-40.pdf.

[R38] Maciej Smetkowski and Piotr Wójcik. “Regional Convergence in Central and Eastern European Countries: A Multidimensional Approach”. In: *European Planning Studies* 20.6 (June 2012), pp. 1–17. DOI: https://doi.org/10.1080/09654313.2012.673560.

[R39] Hiranya K. Nath and Kiril Tochkov. “Relative inflation dynamics in the new EU member countries of Central and Eastern Europe”. In: *Empirical Economics* 45.1 (Aug. 2013), pp. 1–22. DOI: https://doi.org/10.1007/s00181-012-0596-y.

[R40] Sebastien Bourdin. “National and regional trajectories of convergence and economic integration in Central and Eastern Europe”. In: *The Canadian Journal of Regional Science* 38.1/3 (Dec. 2015), pp. 55–63. URL: https://idjs.ca/images/rcsr/archives/V38N1-BOURDIN.pdf.

[R41] Goran Radosavljevi et al. “Income convergence between Southeast Europe and the European Union”. In: *Proceedings of Rijeka School of Economics* 38.2 (Dec. 2020), pp. 499–519. DOI: https://doi.org/10.18045/zbefri.2020.2.499.

[R42] Feng Jiang et al. “Initialization of K-Modes Clustering Using Outlier Detection Techniques”. In: *Information Sciences* 332.1 (Mar. 2016), pp. 167–183. DOI: https://doi.org/10.1016/j.ins.2015.11.005.

[R43] Peter Filzmoser and Klaus Nordhausen. “Robust linear regression for high - dimensional data: An overview”. In: *WIREs Computational Statistics* 13.4 (Aug. 2021), e1524. DOI: https://doi.org/10.1002/wics.1524.

[R44] Ferenc Steiner and Béla Hajagos. “Practical Definition of Robustness”. In: *Geophysical Transactions* 38.4 (1993), pp. 193–210.

[R45] Ferenc Steiner. “Comparison of the L2-, L1- and P-norm based statistical procedures in respect of their asymptotic efficiencies”. In: *Magyar Geofizika* 41.1 (Feb. 2000).

[R46] László Ferenczy. “A short introduction to the most frequent value procedures (in Hungarian)”. In: *Magyar Geofizika* 29.3 (1988), pp. 83–94.

[R47] Marco Avella Medina and Elvezio Ronchetti. “Robust statistics: a selective overview and new directions”. In: *WIREs Computational Statistics* 7 (Dec. 2015), pp. 372–393. DOI: https://doi.org/10.1002/wics.1363.

[R48] Béla Hajagos and Ferenc Steiner. “Investigations Concerning Resistance - Importance of the Choice of the Formula Determining the Scale Parameter”. In: *Geophysical Transactions* 38.4 (1993), pp. 211–230.

[R49] F. Steiner. “(Editor), The Most Frequent Value. Introduction to a Modern Conception of Statistics”. In: (1991). Akadémiai Kiadó, Budapest, Hungary, 315p., ISBN: 963 05 5687 1.

[R50] P. Filzmoser et al. “Cellwise robust M regression”. In: *Computational Statistics and Data Analysis* 147 (Mar. 2020), p. 106944. DOI: https://doi.org/10.1016/j.csda.2020.106944.

[R51] *European Commission, EuroStat for statistical data download*. https://ec.europa.eu/eurostat/web/main/data/database. Accessed: 30/08/2021.

[R52] Michaela Chocholatá and Andrea Furková. “Income disparities and convergence across regions of Central Europe”. In: *Croatian Operational Research Review* 7.2 (Dec. 2016), pp. 303–318. DOI: https://doi.org/10.17535/crorr.2016.0021.

[R53] Joseph Stiglitz, Amartya K. Sen, and Jean-Paul Fitoussi. “The measurement of economic performance and social progress revisited: Reflections and Overview”. In: *The Review of Income and Wealth* 29.3 (Dec. 2009), pp. 297–316. URL: https://sciencespo.hal.science/hal-01069384.

[R54] Ida Kubiszewski et al. “Beyond GDP: Measuring and achieving global genuine progress”. In: *Ecological Economics* 93 (Apr. 2013), pp. 57–68. DOI: https://doi.org/10.1016/j.ecolecon.2013.04.019.

[R55] Robert Costanza et al. “Beyond GDP: The Need for New Measures of Progress”. In: *Pardee Paper No. 4, Boston: Pardee Center for the Study of the Longer-Range Future* (Jan. 2009). URL: https://www.bu.edu/pardee/files/documents/PP-004-GDP.pdf.

[R56] T. K. Rymes. “More on the Measurement of Total Factor Productivity”. In: *The Review of Income and Wealth* 29.3 (Sept. 1983), pp. 297–316. DOI: https://doi.org/10.1111/j.1475-4991.1983.tb00646.x.

[R57] Cristina Terra. “How to Measure International Transactions”. In: *Principles of International Finance and Open Economy Macroeconomics Theories, Applications, and Policies* (2015), pp. 9–30. DOI: https://doi.org/10.1016/B978-0-12-802297-9.00002-6.

[R58] Gustav Ranis, Frances Stewart, and Emma Samman. “Human Development: Beyond the Human Development Index”. In: *Journal of Human Development and Capabilities* 7.3 (Nov. 2006), pp. 323–358. DOI: https://doi.org/10.1080/14649880600815917.

[R59] Pranav Bhaskar and Rashmi Tripathi. “Green GDP integrating economic growth with ecological sustainability”. In: *Biophilia Insights* 1.1 (June 2023), pp. 57–68. DOI: https://doi.org/10.52679/bi.e202311003.

[R60] Robert E. Hall and Charles I. Jones. “Why do Some Countries Produce So Much More Output Per Worker than Others?” In: *The Quarterly Journal of Economics* 114.1 (Feb. 1999), pp. 83–116. DOI: https://doi.org/10.1162/003355399555954.

[R61] Jan Hauke and Tomasz Kossowski. “Comparision of Values of Pearson’s and Spearman’s Correlation Coefficeints on the same Sets of Data”. In: *Quaestiones Geographicae* 30.2 (Feb. 2011), pp. 87–93. DOI: https://doi.org/10.2478/v10117-011-0021-1.

[R62] Marie Laure Delignette-Muller and Christophe Dutang. “fitdistrplus: An R Package for Fitting Distributions”. In: *Journal of Statistical Software* 64.4 (Mar. 2015). DOI: https://doi.org/10.18637/jss.v064.i04.

[R63] Ricardo A. Maronna, R. Douglas Martin, and Víctor J. Yohai. “Robust Statistics. Theory and Methods”. In: (2006). John Wiley & Sons Ltd, England, 403p., ISBN: 0-470-01092-4.

[R64] C. G. Broyden. “A Class of Methods of Solving Nonlinear Simultaneous Equations”. In: *Mathematics of Computation* 19.92 (Oct. 1965), pp. 577–593. DOI: https://doi.org/10.1090/S0025-5718-1965-0198670-6.

[R65] István Faragó and Róbert Horváth. “Numerical Methods (in Hungarian)”. In: (2011). Institute of Mathematics, Budapest University of Technology and Economics, Budapest, Hungary, 399p., ISBN: 978-963-279-456-3.

[R66] István Vincze. “Mathematical Statistics with Industrial Applications (in Hungarian)”. In: (1968). Mszaki Kiadó, Budapest, Hungary, 352p.

[R67] Jeffrey G. Williamson. “Regional Inequality and the Process of National Development: A Description of the Patterns”. In: *Economic Development and Cultural Change* 13.4 (July 1965). DOI: https://doi.org/10.1086/450136.

[R68] Salvador Barrios and Eric Strobl. “The Dynamics of Regional Inequalities”. In: *Regional Science and Urban Economics* 39.5 (2009), pp. 575–591. DOI: https://doi.org/10.1016/j.regsciurbeco.2009.03.008.

[R69] George Petrakos, Andrés Rodríguez-Pose, and Antonis Rovolis. “Growth, integration, and regional disparities in the European Union”. In: *Environment and Planning A* 37.10 (2005), pp. 1837–1855. DOI: https://doi.org/10.1068/a37348.

[R70] Miguel Linhares Pinheiro et al. “The role of social capital towards resource sharing in collaborative R&D projects: Evidences from the 7th Framework Programme”. In: *International Journal of Project Management* 34.8 (July 2016), pp. 1519–1536. DOI: https://doi.org/10.1016/j.ijproman.2016.07.006.

[R71] Marlies Schütz and Rita Strohmaier. “Power relations in European RDI - collaboration networks. Disparities in policy-driven opportunities for knowledge generation in ICT”. In: *Economics of Innovation and New Technology* 31.3 (Sept. 2020), pp. 209–230. DOI: https://doi.org/10.1080/10438599.2020.1799139.

[R72] Oleksandr Husiev, Olatz Ukar Arrien, and Marta Enciso-Santocildes. “What does Horizon 2020 contribute to? Analysing and visualising the community practices of Europes largest research and innovation programme”. In: *Energy Research & Social Science* 95.102879 (Jan. 2023). DOI: https://doi.org/10.1016/j.erss.2022.102879.

[R73] Adrienn Novotni, Zoltán Pásztory, and Zsolt Tóth. “Social Network Analysis in Wood Industry Projects”. In: *Acta Silvatica & Lignaria Hungarica* 18.2 (2022), pp. 89–101. DOI: https://doi.org/10.37045/aslh-2022-0006.

[R74] Sylwia Krzyek-Liburska and Grayna Paliwoda-Pkosz. “Towards Modelling the Success Determinants of International Research Proposals”. In: *30th International Conference on Information Systems Development (ISD2022 CLUJ-Napoca, Romania)* (Sept. 2022). DOI: https://doi.org/10.62036/ISD.2022.54.

[R75] Nikolaos Giarelis and Nikos Karacapilidis. “Understanding Horizon 2020 Data: A Knowledge Graph-Based Approach”. In: *Applied Sciences* 11.23 (Dec. 2021), p. 11425. DOI: https://doi.org/10.3390/app112311425.

[R76] Daniele Archibugi, Rinaldo Evangelista, and Antonio Vezzani. “Regional Technological Capabilities and the Access to H2020 Funds”. In: *Journal of Common Market Studies* 60.4 (Dec. 2021), pp. 926–944. DOI: https://doi.org/10.1111/jcms.13291.

[R77] Simen G. Enger and Fulvio Castellacci. “Who gets Horizon 2020 research grants? Propensity to apply and probability to succeed in a two-step analysis”. In: *Scientometrics* 109.3/12 (2016), p. 1638. DOI: https://doi.org/10.1007/s11192-016-2145-5.

[R78] Antonia Brali. “Social Network Analysis of Country Participation in Horizon 2020 Programme”. In: *Conference: Central European Conference on Information and Intelligent System (28th CECIIS, Varadin, Croatia)* (Sept. 2017). DOI: https://doi.org/10.1016/j.eswa.2011.09.018.

[R79] Pedro Varela-Vázquez, Manuel González-López, and María del Carmen Sánchez-Carreíra. “The uneven regional distributi on of projects funded by the EU Framework Programmes”. In: *Journal of Entrepreneurship, Management and Innovation* 15.3 (Aug. 2019), pp. 45–72. DOI: https://doi.org/10.7341/20191532.

[R80] Iris Wanzenböck and Philipp Piribauer. “R&D networks and regional knowledge production in Europe: Evidence from a space-time model”. In: *Papers in Regional Science* 97.S1 (Mar. 2018), S1–S24. DOI: https://doi.org/10.1111/pirs.12236.

[R81] Francesca Maridina Malloci et al. “A Text Mining Approach to Extract and Rank Innovation Insights from Research Projects”. In: *Web Information Systems Engineering  WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 2024, 2020, Proceedings, Part II* (Oct. 2020), pp. 143–154. DOI: https://doi.org/10.1007/978-3-030-62008-0_10.

[R82] M. J. Barber et al. “The Network of EU-Funded Collaborative R&D Projects”. In: *Physical Review E* 73.036132 (Mar. 2006). DOI: https://doi.org/10.1016/j.eswa.2011.09.018.

[R83] Simen G. Enger. “Closed clubs: Network centrality and participation in Horizon 2020”. In: *Science and Public Policy* (2018), pp. 1–13. DOI: https://doi.org/10.1093/scipol/scy029.

[R84] Teemu Makkonen and Timo Mitze. “Scientific collaboration between old and new member states: Did joining the European Union make a difference?” In: *Scientometrics* 106 (2016), pp. 1193–1215. DOI: https://doi.org/10.1007/s11192-015-1824-y.

[R85] Pierre-Alexandre Balland, Ron Boschma, and Julien Ravet. “Network dynamics in collaborative research in the EU, 20032017”. In: *European Planning Studies* 27.9 (May 2019), pp. 1811–1837. DOI: https://doi.org/10.1080/09654313.2019.1641187.

[R86] László Molnár. “Basic concepts of network analysis - graphs, centrality, adjacency, bridges and the small world”. In: *In: Péter Sárvári (ed.): Systems Theory, Dialóg Campus, Budapest* (2020), pp. 123–140. DOI: https://doi.org/10.36250/00734.07.

[R87] Albert-László Barabási. “The science of networks: From society to the web (in Hungarian)”. In: *Magyar Tudomány* 51.11 (Nov. 2006), pp. 1298–1308.

[R88] K Jordahl. “GeoPandas: Python Tools for Geographic Data”. In: *Web.: https://github. com/geopandas/geopandas* (2014). Available online: https://github.com/geopandas/geopandas (accessed on 15 June 2023).

[R89] Stephanie Mwika Mbiya, Gerhard P. Hancke, and Bruno Silva. “An Efficient Routing Algorithm for Wireless Sensor Networks based on Centrality Measures”. In: *Acta Polytechnica Hungarica* 17.1 (Jan. 2020), pp. 83–99. DOI: https://doi.org/10.12700/APH.17.1.2020.1.5.

[R90] Tamás Réti and István Bárányi. “On the Irregularity Characterization of Mean Graphs”. In: *Acta Polytechnica Hungarica* 18.5 (Jan. 2021), pp. 207–220. DOI: https://doi.org/10.12700/APH.18.5.2021.5.13.

[R91] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring network structure, dynamics, and function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA)* (Aug. 2008). URL: https://aric.hagberg.org/papers/hagberg-2008-exploring.pdf.

[R92] Computer software. Vers. 2-2.4.0. Anaconda. “Anaconda Software Distribution”. In: (Nov. 2016). URL: https://anaconda.com.

[R93] Bharat Pardeshi and Durga Toshniwal. “Improved K-Medoids Clustering Based on Cluster Validity Index and Object Density”. In: *2010 IEEE 2nd International Advance Computing Conference (IACC)* (Mar. 2010). DOI: https://doi.org/10.1109/IADCC.2010.5422924.

[R94] Peter J. Rousseeuw and Mia Hubert. “Robust Statistics for Outlier Detection”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (2011). DOI: https://doi.org/10.1002/widm.2.

[R95] Arnis Kirshners, Arkady Borisov, and Serge Parshutin. “Robust Cluster Analysis in Forecasting Task”. In: *International Conference on Applied Information and Communication Technologies* (Apr. 2012). URL: https://llufb.llu.lv/conference/AICT/2012/KRISHNERS_AICT2012.pdf.

[R96] Sébastien Thomassey and Antonio Fiordaliso. “A hybrid sales forecasting system based on clustering and decision trees”. In: *Decision Support Systems* 42.1 (Oct. 2006), pp. 408–421. DOI: https://doi.org/10.1016/j.dss.2005.01.008.

[R97] Olga Dorabiala, J. Nathan Kutz, and Aleksandr Y. Aravkin. “Robust Trimmed k-means”. In: *ArXiv* abs/2108.07186 (Aug. 2021). DOI: https://doi.org/10.48550/arXiv.2108.07186.

[R98] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. “Unsupervised Clustering Approach for Network Anomaly Detection”. In: *Fourth International Conference on Networked Digital Technologies* (Apr. 2012). DOI: https://doi.org/10.1007/978-3-642-30507-8_7.

[R99] Roshan Chitrakar and Huang Chuanhe. “Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering”. In: *Third Asian Himalayas International Conference on Internet* (Nov. 2012). DOI: https://doi.org/10.1109/AHICI.2012.6408446.

[R100] Luis Angel García-Escudero et al. “A review of robust clustering methods”. In: *Advances in Data Analysis and Classification* 4.2 (Sept. 2010), pp. 89–109. DOI: https://doi.org/10.1007/s11634-010-0064-5.

[R101] Norazam Arbin et al. “Comparative Analysis between K-Means and K-Medoids for Statistical Clustering”. In: *3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)* (Dec. 2015), pp. 117–121. DOI: https://doi.org/10.1109/AIMS.2015.82.

[R102] Kalpit G. Soni and Atul Patel. “Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data”. In: *International Journal of Computational Intelligence Research* 13.5 (2017), pp. 899–906. ISSN: 0973-1873. URL: https://www.ripublication.com/ijcir17/ijcirv13n5_21.pdf.

[R103] Norin Rahayu Shamsuddin and Nor Idayu Mahat. “Comparison Between k-Means and k-Medoids for Mixed Variables Clustering”. In: *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)* (Jan. 2019), pp. 303–308. DOI: https://doi.org/10.1007/978-981-13-7279-7_3.

[R104] T. Velmurugan and A.Dharmarajan. “Clustering Lung Cancer Data by k-Means and k-Medoids Algorithms”. In: *International Conference on Information and Convergence Technology for Smart Society* (Jan. 2015). DOI: https://doi.org/10.20894/IJDMTA.102.003.002.009.

[R105] Tagaram Soni Madhulatha. “Comparision between k-Means and k-Medoids Clustering Algorithms”. In: *International Conference on Advances in Computing and Information Technology Advances in Computing and Information Technology (ACITY 2011), Communications in Computer and Information Science* 198 (Apr. 2011), pp. 472–481. ISSN: 978-3-642-22555-0. DOI: https://doi.org/10.1007/978-3-642-22555-0_48.

[R106] Mediana Aryuni, Evaristus Didik Madyatmadja, and Eka Miranda. “Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering”. In: *International Conference on Information Management and Technology (ICIMTech)* (Sept. 2018). DOI: https://doi.org/10.1109/ICIMTech.2018.8528086.

[R107] Triyanna Widiyaningtyas, Utomo Pujianto, and Martin Indra Wisnu Prabowo. “K-Medoids and K-Means Clustering in High School Teacher Distribution”. In: *International Conference on Electrical, Electronics and Information Engineering (ICEEIE)* (Oct. 2019). DOI: https://doi.org/10.1109/ICEEIE47180.2019.8981466.

[R108] A. Dharmarajan and T. Velmurugan. “Efficiency of k-Means and k-Medoids Clustering Algorithms using Lung Cancer Dataset”. In: *International Journal of Data Mining Techniques and Applications* 5.2 (Dec. 2016), pp. 150–156. ISSN: 2278-2419. DOI: https://doi.org/10.20894/IJDMTA.102.005.002.011.

[R109] Habiba Drias, Nadjib Fodil Cherif, and Amine Kechid. “k-MM: A Hybrid Clustering Algorithm Based on k-Means and k-Medoids”. In: *Advances in Nature and Biologically Inspired Computing* (Nov. 2016), pp. 37–48. DOI: https://doi.org/10.1007/978-3-319-27400-3_4.

[R110] Peter Olukanmi, Fulufhelo Nelwamondo, and Tshilidzi Marwala. “Effect of Data Parameters and Seeding on k-Means and k-Medoids”. In: *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)* (Aug. 2020). DOI: https://doi.org/10.1109/icABCD49160.2020.9183892.

[R111] Luis Ángel García-Escudero and Alfonso Gordaliza. “Robustness Properties of k Means and Trimmed k Means”. In: *Journal of the American Statistical Association* 94.447 (Sept. 1999), pp. 956–969. DOI: https://doi.org/10.1080/01621459.1999.10474200.

[R112] Luis Angel García-Escudero, Alfonso Gordaliza, and Carlos Matrán. “Trimming Tools in Exploratory Data Analysis”. In: *Journal of Computational and Graphical Statistics* 12.2 (June 2003), pp. 434–449. DOI: https://doi.org/10.1198/1061860031806.

[R113] Peter J. Rousseeuw and Mia Hubert. “Anomaly detection by robust statistics”. In: *WIREs Data Mining and Knowledge Discovery* 8.2 (Apr. 2018), e1236. DOI: https://doi.org/10.1002/widm.1236.

[R114] Peter J. Rousseeuw and Bert C. van Zomeren. “Unmasking Multivariate Outliers and Leverage Points”. In: *Journal of the American Statistical Association* 85.411 (Sept. 1990), pp. 633–639. DOI: https://doi.org/10.1080/01621459.1990.10474920.

[R115] Mia Hubert, Peter J. Rousseeuw, and Stefan Van Aelst. “High-Breakdown Robust Multivariate Methods”. In: *Statistical Science* 23.1 (Feb. 2008), pp. 92–119. DOI: https://doi.org/10.1214/088342307000000087.

[R116] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[R117] Joaquín Pérez-Ortega et al. “Improving the Efficiency of the K-medoids Clustering Algorithm by Getting Initial Medoids”. In: *Advances in Intelligent Systems and Computing* (Mar. 2017). DOI: https://doi.org/10.1007/978-3-319-56535-4_13.

[R118] Christophe Leys et al. “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median”. In: *Journal of Experimental Social Psychology* 49.4 (July 2013), pp. 764–766. DOI: https://doi.org/10.1016/j.jesp.2013.03.013.

[R119] Jeff Miller. “Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size”. In: *The Quarterly Journal of Experimental Psychology Section A* 43.4 (Nov. 1991), pp. 907–912. DOI: https://doi.org/10.1080/14640749108400962.

[R120] Peter J. Rousseeuw and Christophe Croux. “Alternatives to the Median Absolute Deviation”. In: *Journal of the American Statistical Association* 88.424 (Dec. 1993), pp. 1273–1283. DOI: https://doi.org/10.2307/2291267.

[R121] David Arthur and Sergei Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. In: *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms* (Jan. 2007). DOI: https://doi.org/10.1145/1283383.1283494.

[R122] Kuo-Lung Wu, Miin-ShenYang, and June-Nan Hsieh. “Robust cluster validity indexes”. In: *Pattern Recognition* 42.11 (Nov. 2009), pp. 2541–2550. DOI: https://doi.org/10.1016/j.patcog.2009.02.010.

[R123] Yanchi Liu et al. “Understanding of Internal Clustering Validation Measures”. In: *2010 IEEE International Conference on Data Mining* (Dec. 2010). DOI: https://doi.org/10.1109/ICDM.2010.35.

[R124] Maria Halkidi and Michalis Vazirgiannis. “Clustering validity assessment: finding the optimal partitioning of a data set”. In: *Proceedings 2001 IEEE International Conference on Data Mining* (Nov. 2001). DOI: https://doi.org/10.1109/ICDM.2001.989517.

[R125] Adimulam Raghuvira Pratap et al. “An Efficient Density based Improved K-Medoids Clustering Algorithm”. In: *International Journal of Advanced Computer Science and Applications* 2.6 (July 2011). DOI: https://doi.org/10.14569/IJACSA.2011.020607.

[R126] Andrea Cerioli, Alessio Farcomeni, and Marco Riani. “Wild adaptive trimming for robust estimation and cluster analysis”. In: *Scandinavian Jorunal of Statistics* 46.1 (July 2018), pp. 235–256. DOI: https://doi.org/10.1111/sjos.12349.

[R127] Ran Bhamra, Samir Dani, and Kevin Burnard. “Resilience: the concept, a literature review and future directions”. In: *International Journal of Production Research* 49.18 (Sept. 2011), pp. 5375–5393. DOI: https://doi.org/10.1080/00207543.2011.563826.

[R128] Adina Aldea et al. “Assessing Resilience in Enterprise Architecture: A Systematic Review”. In: *2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC), Eindhoven, The Netherlands, 5–8 October* (Oct. 2020). DOI: https://doi.org/10.1109/EDOC49727.2020.00011.

[R129] János Varga. “Defining the Economic Role and Benefits of Micro, Small and Medium-sized Enterprises in the 21st Century with a Systematic Review of the Literature”. In: *ACTA Polytechnica Hungarica* 28.11 (Jan. 2021), pp. 209–228. DOI: https://doi.org/10.12700/APH.18.11.2021.11.12.

[R130] Eric Lutters Denzil Kennon Corné S.L. Schutte. “An alternative view to assessing antifragility in an organisation: A case study in a manufacturing SME”. In: *CIRP Annals - Manufacturing Technology* 64.1 (2015), pp. 177–180. DOI: https://doi.org/10.1016/j.cirp.2015.04.024.

[R131] John Johnson and Adrian V. Gheorghe. “Antifragility Analysis and Measurement Framework for Systems of Systems”. In: *International Journal of Disaster Risk Science* 4.4 (2013), pp. 159–168. DOI: https://doi.org/10.1007/s13753-013-0017-7.

[R132] Nassim Nicholas Taleb. “Black Swans and the Domains of Statistics”. In: *The American Statistician* 61.3 (Aug. 2007), pp. 198–200. DOI: https://doi.org/10.1198/000313007X219996.

[R133] Rapha ́el Douady Nassim Nicholas Taleb. “Mathematical Definition, Mapping, and Detection of (Anti)Fragility”. In: *Quantitative Finance* 13.11 (Nov. 2013), pp. 1677–1689. DOI: https://doi.org/10.48550/arXiv.1208.1189.

[R134] Ozgur Erol, Brian J. Sauser, and Mo Mansouri. “A framework for investigation into extended enterprise resilience”. In: *Enterprise Information Systems* 4.2 (May 2010), pp. 111–136. DOI: https://doi.org/10.1080/17517570903474304.

[R135] A. Aleksi et al. “An assessment of organizational resilience potential in SMEs of the process industry, a fuzzy approach”. In: *Journal of Loss Prevention in the Process Industries* 26.6 (Nov. 2013), pp. 1238–1245. DOI: https://doi.org/10.1016/j.jlp.2013.06.004.

[R136] Peter Karácsony. “Analyzing the Relationship between Leadership Style and Corporate Social Responsibility in Hungarian Small and Medium-sized Enterprises”. In: *Acta Polytechnica Hungarica* 17.7 (2020), pp. 183–198. DOI: https://doi.org/10.12700/APH.17.7.2020.7.10.

[R137] James Simmie and Ron Martin. “The economic resilience of regions: towards an evolutionary approach”. In: *Cambridge Journal of Regions, Economy and Society* 39.1 (Jan. 2010), pp. 27–43. DOI: https://doi.org/10.1093/cjres/rsp029.

[R138] Cristina Ruiz-Martin, Adolfo López-Paredes, and Gabriel Wainer. “What we know and do not know about organizational resilience”. In: *International Journal of Production Management and Engineering* 6.1 (Dec. 2017), pp. 11–28. DOI: https://doi.org/10.4995/ijpme.2018.7898.

[R139] Adam Rose and Elisabeth Krausmann. “An economic framework for the development of a resilience index for business recovery”. In: *International Journal of Disaster Risk Reduction* 5 (Sept. 2013), pp. 73–83. DOI: https://doi.org/10.1016/j.ijdrr.2013.08.003.

[R140] Muhammedamin Hussen saad et al. “Conceptualization of SMEs business resilience: A systematic literature review”. In: *Cogent Business & Management* 8.1 (2021), p. 1938347. DOI: https://doi.org/10.1080/23311975.2021.1938347.

[R141] Angappa Gunasekaran, Bharatendra K. Rai, and Michael Griffn. “Resilience and competitiveness of small and medium size enterprises: An empirical research”. In: *International Jorunal of Production Research* 49.18 (Apr. 2011), pp. 5489–5509. DOI: https://doi.org/10.1080/00207543.2011.563831.

[R142] Ozgur Erol et al. “Perspectives on Measuring Enterprise Resilience”. In: *2010 IEEE International Systems Conference, San Diego, CA, USA* (Apr. 2010), pp. 587–592. DOI: https://doi.org/10.1109/SYSTEMS.2010.5482333.

[R143] Cristina Ruiz-Martina, Adolfo López-Paredes, and Gabriel Wainer. “What we know and do not know about organizational resilience”. In: *International Journal of Production Management and Engineering* 6.1 (2018), pp. 11–28. DOI: https://doi.org/10.4995/ijpme.2018.7898.

[R144] Véronique Ambrosini Albert Munoz Jon Billsberry. “Resilience, robustness, and antifragility: Towards an appreciation of distinct organizational responses to adversity”. In: *The International Journal of Management Reviews* 24.2 (Apr. 2022), pp. 181–187. DOI: https://doi.org/10.1111/ijmr.12289.

[R145] Olga Eckardt. “Company Maturity Matrix”. In: *Emerging Markets Journal* 8.1 (2008), pp. 28–30. DOI: https://doi.org/10.5195/emaj.2018.148.

[R146] Tjo-Kin Man. “Measuring and Analysing Resilience of Enterprise Architectures”. In: *31th Twente Student Conference on IT, University of Twente, Enschede, Netherlands,* (July 2019).

[R147] Javaneh Ramezani and Luís M. Camarinha-Matos. “Approaches for resilience and antifragility in collaborative business ecosystems”. In: *Technological Forecasting and Social Change* 151 (2020), p. 119846. DOI: https://doi.org/10.1016/j.techfore.2019.119846.

[R148] Lino Briguglio et al. “Economic Vulnerability and Resilience: Concepts and Measurements”. In: *Oxford Development Studies* 37.3 (Sept. 2009), pp. 229–247. DOI: https://doi.org/10.1080/13600810903089893.

[R149] Adam Rose. “Defining and measuring economic resilience to disasters”. In: *Disaster Prevention and Management* 13.4 (2004), pp. 307–314. DOI: https://doi.org/10.1108/09653560410556528.

[R150] Ambika Zutshi et al. “From Challenges to Creativity: Enhancing SMEs Resilience in the Context of COVID-19”. In: *Sustainability* 13.12 (June 2021), p. 6542. DOI: https://doi.org/10.3390/su13126542.

[R151] David B. Audretsch and A. Roy Thurik. “What’s New about the New Economy? Sources of Growth in the Managed and Entrepreneurial Economies”. In: *Industrial and Corporate Change* 10.1 (Mar. 2001), pp. 267–315. DOI: https://doi.org/10.1093/icc/10.1.267.

[R152] Giulia Faggio, Olmo Silva, and William C. Strange. “Heterogeneous Agglomeration”. In: *The Review of Economics and Statistics* 99.1 (2017), pp. 80–94. DOI: https://doi.org/10.1162/REST_a_00604.

[R153] Maryann P. Feldman et al. “The logic of economic development: A definition and model for investment”. In: *Environment and Planning C: Government and Policy* 34.1 (2016), pp. 5–21. DOI: https://doi.org/10.1177/0263774X15614653.

[R154] Kinga Nagyné Pércsi and Zsolt Fülöp. “Relationships between Sustainable Operations and the Resilience of SMEs”. In: *Sustainability* 16.2 (Jan. 2024), p. 741. DOI: https://doi.org/10.3390/su16020741.

[R155] David Bailey and Lisa De Propris. “Manufacturing reshoring and its limits: the UK automotive case”. In: *Cambridge Journal of Regions, Economy and Society* 7.3 (July 2014), pp. 379–395. DOI: https://doi.org/10.1093/cjres/rsu019.

[R156] Paula Graça and Luís M. Camarinha-Matos. “Performance indicators for collaborative business ecosystems  Literature review and trends”. In: *Technological Forecasting & Social Change* 116 (Mar. 2017), pp. 237–255. DOI: https://doi.org/10.1016/j.techfore.2016.10.012.

[R157] Aylin Ates and Umit Bititci. “Change process: a key enabler for building resilient SMEs”. In: *International Journal of Production Research* 49.18 (Sept. 2011), pp. 5601–5618. DOI: https://doi.org/10.1080/00207543.2011.563825.

[R158] Supardi Supardi and Syamsul Hadi. “New Perspective on the Resilience of SMEs, Proactive, Adaptive, Reactive from Budsiness Turbulence: A Systematic Review”. In: *Journal of Xi’an University of Architecture & Technology* XII.V (May 2020), pp. 4068–4076. ISSN: 1006-7930. DOI: https://doi.org/10.1080/00207543.2011.563825.

[R159] Herb de Vries and Michelle Shields. “Towards a Theory of Entrepreneurial Resilience: A Case Study Analysis of New Zealand SME Owner Operators”. In: *Applied Research Journal* 5.1 (Sept. 2006).

[R160] B. Hennebry. “The economic resilience of Irish counties for subsequent recessions and the impact of population distribution on resilience”. In: *R-Economy* 6.3 (2020), pp. 146–153. DOI: https://doi.org/10.15826/recon.2020.6.3.012.

[R161] György Kadocsa and Anna Francsovics. “Macro and Micro Economic Factors of Small Enterprise Competitiveness”. In: *ACTA Polytechnica Hungarica* 8.1 (Jan. 2011), pp. 23–40. URL: https://acta.uni-obuda.hu/Kadocsa_Francsovics_27.pdf.

[R162] Naim H. Afgan. “Resilience of Company Management System”. In: *PICMET 2010 Technology Management for Global Economic Growth, Phuket, Thailand* (2010), pp. 1–8. ISSN: 2159-5100. URL: https://www.researchgate.net/publication/254451022_Post_Evaluation_Sustainable_Resilience_of_Company_Management_System.

[R163] Graham Coates et al. “Evaluating the operational resilience of small and medium-sized enterprises to flooding using a computational modelling and simulation approach: a case study of the 2007 flood in Tewkesbury”. In: *Philos Trans A Math Phys Eng Sci.* (Apr. 2020). DOI: https://doi.org/10.1098/rsta.2019.0210.

[R164] Hyejin Jung, JungTae Hwang, and Byung-Keun Kim. “Does R&D investment increase SME survival during a recession?” In: *Technological Forecasting & Social Change* 137 (Aug. 2018), pp. 190–198. DOI: https://doi.org/10.1016/j.techfore.2018.07.042.

[R165] Brian Sauser et al. “Resilience of small- and medium-sized enterprises as a correlation to community impact: an agent-based modeling approach”. In: *Natural Hazards* 90.1 (Jan. 2018), pp. 79–99. DOI: https://doi.org/10.1007/s11069-017-3034-9.

[R166] Scott Somers. “Measuring Resilience Potential: An Adaptive Strategy for Organizational Crisis Planning”. In: *Journal of Contingencies and Crisis Management* 17.1 (Mar. 2009), pp. 12–23. DOI: https://doi.org/10.1111/j.1468-5973.2009.00558.x.

[R167] Rudrajeet Pal, HåkanTorstensson, and Heikki Mattila. “Antecedents of Organizational Resilience in Economic Crisesan Empirical Study of Swedish Textile and Clothing SMEs”. In: *Int. J.Production Economics* 147 (2014), pp. 410–428. DOI: https://doi.org/10.1016/j.ijpe.2013.02.031.

[R168] Elias Giannakis and Adriana Bruggeman. “Regional disparities in economic resilience in the European Union across the urbanrural divide”. In: *Regional Studies* 54.9 (Dec. 2019), pp. 1200–1213. DOI: https://doi.org/10.1080/00343404.2019.1698720.

[R169] Lino Briguglio et al. “Economic Vulnerability and Resilience: Concepts and Measurements”. In: *Oxford Development Studies* 37.3 (Sept. 2009), pp. 229–247. DOI: https://doi.org/10.1080/13600810903089893.

[R170] Gerben van der Velde Muhammedamin Hussen saad Geoffrey Hagelaar and S. W. F. Omta. “Conceptualization of SMEs business resilience: A systematic literature review”. In: *Cogent Business and Management* 8.1 (May 2021), p. 1938347. DOI: https://doi.org/10.1080/23311975.2021.1938347.

[R171] Andrea Némethné Gál. “Competitiveness of Small and Medium-sized Enterprises - PhD Thesis (in Hungarian)”. In: (2009). István Széchenyi University, Doctoral School of Regional- and Business Administration, Gyõr, Available online: https://rgdi.sze.hu/files/Ertekezesek,%20tezisek/Magyar%20Tezis%20NGA.pdf (accessed on 1 June 2023).

[R172] Tariq Masood and Paul Sonntag. “Industry 4.0: Adoption challenges and benefits for SMEs”. In: *Computers in Industry* 121.103261 (Oct. 2020). DOI: https://doi.org/10.1016/j.compind.2020.103261.

[R173] Emil Blixt Hansen and Simon Bøgh. “Artificial intelligence and internet of things in small and medium-sized enterprises: A survey”. In: *Journal of Manufacturing Systems* (Aug. 2020). DOI: https://doi.org/10.1016/j.jmsy.2020.08.009.

[R174] Ran Bhamra, Samir Dani, and Kevin Burnard. “Resilience: the concept, a literature review and future directions”. In: *International Jorunal of Production Research* 49.18 (Jan. 2011), pp. 5375–5393. ISSN: 0020-7543. DOI: https://doi.org/10.1080/00207543.2011.563826.

[R175] Danfang Chen et al. “A holistic and rapid sustainability assessment tool for manufacturing SMEs”. In: *CIRP Annals* 63.1 (2014), pp. 437–440. DOI: https://doi.org/10.1016/j.cirp.2014.03.113.

[R176] Brahim Herbane. “Exploring Crisis Management in UK Small- and Medium-Sized Enterprises”. In: *Journal of Contingencies and Crisis Management* 21.2 (June 2013), pp. 82–95. DOI: https://doi.org/10.1111/1468-5973.12006.

[R177] Rachel Doern, Nick Williams, and Tim Vorley. “Special Issue on Entrepreneurship and Crises: Business as Usual? An Introduction and Review of the Literature”. In: *Entrepreneurship & Regional Development* 31.5–6 (2019), pp. 400–412. DOI: https://doi.org/10.1080/08985626.2018.1541590.

[R178] Martina K. Linnenluecke. “Resilience in Business and Management Research: A Review of Influential Publications and a Research Agenda”. In: *International Journal of Management Reviews* 19.4 (Jan. 2017), pp. 4–30. DOI: https://doi.org/10.1111/ijmr.12076.

[R179] Martina Battisti and David Deakins. “The relationship between dynamic capabilities, the firms resource base and performance in a post-disaster environment”. In: *International Small Business Journal* 35.1 (2017), pp. 78–98. DOI: https://doi.org/10.1177/0266242615611471.

[R180] Rudrajeet Pal, Roy Andersson, and Hakan Torstensson. “Organisational resilience through crisis strategic planning: a study of Swedish textile SMEs in financial crises of 2007-2011”. In: *Int. J. Decision Sciences, Risk and Management* 4.3/4 (2012), pp. 314–341. DOI: https://doi.org/10.1504/IJDSRM.2012.053372.

[R181] K.K.N.B. Adikaram and H.A.K.N.S. Surangi. “Crisis management in small and medium scale enterprises: a systematic literature review”. In: *Journal of Business Studies* 7.2 (Dec. 2020), pp. 19–41. DOI: https://doi.org/10.4038/jbs.v7i2.59.

[R182] Amy V. Lee, John Vargo, and Erica Seville. “Developing a Tool to Measure and Compare Organizations Resilience”. In: *Natural Hazards Review* 14.1 (Feb. 2013), pp. 29–41. DOI: https://doi.org/10.1061/(ASCE)NH.1527-6996.0000075.

[R183] Riccardo Patriarca et al. “An Analytic Framework to Assess Organizational Resilience”. In: *Safety and Health at Work* 9.3 (July 2018), pp. 265–276. DOI: https://doi.org/10.1016/j.shaw.2017.10.005.

[R184] Sami Ben Jabeur. “Bankruptcy prediction using Partial Least Squares Logistic Regression”. In: *Jorunal of Reailing and Computer Services* 36.C (Feb. 2017), pp. 197–202. DOI: https://doi.org/10.1016/j.jretconser.2017.02.005.

[R185] Mohamed Oudgou Youssef Zizi and Abdeslam El Moudden. “Determinants and Predictors of SMEs Financial Failure: A Logistic Regression Approach”. In: *Risks* 8.4 (Oct. 2020), p. 107. DOI: https://doi.org/10.3390/risks8040107.

[R186] Tamás Kristóf and Miklós Virág. “A Comprehensive Review of Corporate Bankruptcy Prediction in Hungary”. In: *Journal of Risk and Financial Management* 13.2 (Feb. 2020), pp. 1–20. DOI: https://doi.org/10.3390/jrfm13020035.

[R187] Mario Hernandez Tinoco and Nick Wilson. “Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables”. In: *International Review of Financial Analysis* 30.C (2013), pp. 391–419. DOI: https://doi.org/10.1016/j.irfa.2013.02.013.

[R188] Miklós Virág et al. “Financial analysis, bankruptcy prediction, crisis management (in Hungarian)”. In: *Kossuth kiadó* (2013). ISBN: 978 963 09 7724 1.

[R189] Johannes de Wet and Elda du Toit. “Return on equity: A popular, but flawed measure of corporate financial performance”. In: *South African Journal of Business Management* 38.1 (Mar. 2007), pp. 59–69. DOI: https://doi.org/10.4102/sajbm.v38i1.578.

[R190] Gregory B.Murphy, Jeff W. Trailer, and Robert C. Hill. “Measuring performance in entrepreneurship research”. In: *Journal of Business Research* 36.1 (May 1996), pp. 15–23. DOI: https://doi.org/10.1016/0148-2963(95)00159-X.

[R191] Thomas Brambor, William Roberts Clark, and Matt Golder. “Understanding Interaction Models: Improving Empirical Analyses”. In: *Political Analysis* 14.1 (2006), pp. 63–82. DOI: https://doi.org/10.1093/pan/mpi014.

[R192] Douglas Curran-Everett. “Explorations in statistics: the analysis of ratios and normalized data”. In: *Advances in Physiology Education* 37.3 (2013), pp. 213–219. DOI: https://doi.org/10.1152/advan.00053.2013.

[R193] Ji-Hyun Kim. “Spurious Correlation between Ratios with a Common Divisor”. In: *Statistics & Probability Letters* 44.4 (1999), pp. 383–386. DOI: https://doi.org/10.1016/j.eswa.2011.09.018.

[R194] Paul Barnes. “The Analysis and Use of Financial Ratios: A Review Article”. In: *Journal of Business Finance & Accounting* 14.4 (1987), pp. 3311–3323. DOI: https://doi.org/10.1111/j.1468-5957.1987.tb00106.x.

[R195] Baruch Lev and Shyam Sunder. “Methodological Issues in the Use of Financial Ratios”. In: *Journal of Accounting and Economics* 1.3 (Dec. 1979), pp. 187–210. DOI: https://doi.org/10.1016/0165-4101(79)90007-7.

[R196] Alessandra Tognazzo, Paolo Gubitta, and Saverio Dave Favaron. “Does slack always affect resilience? A study of quasi-medium-sized Italian firms”. In: *Entrepreneurship & Regional Development* 28.9-10 (Nov. 2016), pp. 768–790. DOI: https://doi.org/10.1080/08985626.2016.1250820.

[R197] Natalia Ortiz-de-Mandojana and Pratima Bansal. “The Long-term Benefits of Organizational Resilience Through Sustainable Business Practices”. In: *Strategic Management Journal* 37.8 (May 2015), pp. 1615–1631. DOI: https://doi.org/10.1002/smj.2410.

[R198] Róbert Pálovics, Primoz Dolenc, and Jure Leskovec. “Companies under stress: the impact of shocks on the production network”. In: *EPJ Data Science* 10.57 (Dec. 2021). DOI: https://doi.org/10.1140/epjds/s13688-021-00310-w.

[R199] Gideon M. Markman and Markus Venzin. “Resilience: Lessons from banks that have braved the economic crisisAnd from those that have not”. In: *International Business Review* 23.6 (Dec. 2014), pp. 1096–1107. DOI: https://doi.org/10.1016/j.ibusrev.2014.06.013.

[R200] Devanandham Henry and Jose Emmanuel Ramirez-Marquez. “A Generic Quantitative Approach to Resilience: A Proposal”. In: *INCOSE International Symposium* 20.1 (July 2010), pp. 291–301. DOI: https://doi.org/10.1002/j.2334-5837.2010.tb01071.x.

[R201] Györgyi Nyikos, Bettina Soha, and Attila Béres. “Entrepreneurial resilience and firm performance during the COVID-19 crisis  Evidence from Hungary”. In: *Regional Statistics* 11.3 (2021), pp. 29–59. DOI: https://doi.org/10.15196/RS110307.

[R202] Anthony Soroka et al. “Measuring regional business resilience”. In: *Regional Studies* 54.6 (Sept. 2019), pp. 838–850. DOI: https://doi.org/10.1080/00343404.2019.1652893.

[R203] Aleksandar Peji and Piroska Stani Molcer. “Predictive Machine Learning Approach for Complex Problem Solving Process Data Mining”. In: *ACTA Polytechnica Hungarica* 18.1 (Jan. 2021), pp. 45–63. DOI: https://doi.org/10.12700/APH.18.1.2021.1.4.

# Own Publications Pertaining to Theses


[T1] Ferenc Tolner, György Eigner, and Balázs Barta. “Resilience Interpretations of Small and Medium-sized Enterprises and its Analytical Approaches - Literature Review”. In: *2021 IEEE 19th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, Jan. 2021. DOI: https://doi.org/10.1109/SAMI50585.2021.9378637. URL: https://ieeexplore.ieee.org/document/9378637.

[T2] Ferenc Tolner et al. “Application of MFV-robustified Correlation Coefficient for the Investigation of the Strength of Beta-convergence of EU NUTS regions”. In: *2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, Mar. 2022. DOI: https://doi.org/10.1109/SAMI54271.2022.9780675. URL: https://ieeexplore.ieee.org/document/9780675.

[T3] Ferenc Tolner, Balázs Barta, and György Eigner. “Comparison of Newtons and Broydens Method as Nonlinear Solver in the Implementation of MFV-robustified Linear Regression”. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Oct. 2022. DOI: https://doi.org/10.1109/SMC53654.2022.9945222. URL: https://ieeexplore.ieee.org/document/9945222.

[T4] Ferenc Tolner, Balázs Barta, and György Eigner. “Outlier Identification with MFV-robustified Linear Regression in Case of Economic Convergence of EU NUTS Regions”. In: *ACTA POLYTECHNICA HUNGARICA* 21.8 (2023), pp. 47–66. DOI: https://doi.org/10.12700/APH.21.8.2024.8.3. URL: https://acta.uni-obuda.hu/Tolner_Barta_Eigner_148.pdf.

[T5] Ferenc Tolner, Balázs Barta, and György Eigner. “Economic Cohesion Perspectives of the EU member Regions: A Non-parametric Approach”. In: *2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, May 2022. DOI: https://doi.org/10.1109/SACI55618.2022.9919533. URL: https://ieeexplore.ieee.org/document/9919533.

[T6] Ferenc Tolner et al. “Robust clustering based on the most frequent value method”. In: *Multidisciplinary Sciences* 13.1 (2023), pp. 141–153. DOI: https://doi.org/10.35925/j.multi.2023.1.11. URL: https://ojs.uni-miskolc.hu/index.php/multi/article/view/2202.

[T7] Ferenc Tolner et al. “Long-term Development Perspectives of Resilient Companies”. In: *2023 IEEE 21th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl’any, Slovakia*. IEEE, Jan. 2023. DOI:

https://doi.org/10.1109/SAMI58000.2023.10044504. URL: https://ieeexplore.ieee.org/document/10044504.

[T8] Ferenc Tolner, Balázs Barta, and György Eigner. “Comprehensive Analysis of H2020 Funding Participation Based on LDA Topic Modeling and Robust Outlier Identification”. In: *2023 IEEE 21th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, Sept. 2023. DOI: https://doi.org/10.1109/SISY60376.2023.10417969. URL: https://ieeexplore.ieee.org/document/10417969.

[T9] Ferenc Tolner, Balázs Barta, and György Eigner. “Regional Level Investigation of EU-Funded H2020 Collaboration via Social Network Analysis”. In: *2023 IEEE 23th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, Nov. 2023. DOI: https://doi.org/10.1109/CINTI59972.2023.10381903. URL: https://ieeexplore.ieee.org/document/10381903.

[T10] Ferenc Tolner, Balázs Barta, and György Eigner. “Economic Resilience and Antifragility: Classification of SME’s Shock Reactions based on Balance Sheet and Income Statement Data”. In: *2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania*. IEEE, May 2022. DOI: https://doi.org/10.1109/SACI58269.2023.10158644. URL: https://ieeexplore.ieee.org/document/10158644.

[T11] Ferenc Tolner, Balázs Barta, and György Eigner. “Reaction to Idiosyncratic Economic Shocks - Economic Resilience of Small and Medium-Sized Enterprises”. In: *Sustainability* 16.13 (2024), p. 5470. DOI: https://doi.org/10.3390/su16135470. URL: https://www.mdpi.com/2071-1050/16/13/5470.

# Own Publications Not Pertaining to Theses


[Tx1] Ferenc Tolner et al. “Enhancing Cross-border Co-operation of Business Organizations based on the Investigation of Textual- and Categorical Information”. In: *ACTA POLYTECHNICA HUNGARICA* 19.5 (2022), pp. 235–255. DOI: https://doi.org/10.12700/APH.19.5.2022.5.12. URL: http://acta.uni-obuda.hu/Tolner_Barta_Takacs_Eigner_123.pdf.

[Tx2] Ferenc Tolner et al. “Investigation of High-Growth Firms in the SME sector via the Perspective of Owners and CEOs using Wordclouds”. In: *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, May 2021. DOI: https://doi.org/10.1109/SACI51354.2021.9465560. URL: https://ieeexplore.ieee.org/document/9465560.

[Tx3] Ferenc Tolner et al. “Clustering based on Preferences with K-modes using Categorical Variables”. In: *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, Sept. 2021. DOI: https://doi.org/10.1109/SISY52375.2021.9582485. URL: https://ieeexplore.ieee.org/document/9582485.

[Tx4] Ferenc Tolner et al. “Clustering of Business Organisations based on Textual Data - An LDA Topic Modeling Approach”. In: *2021 IEEE 21th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, Nov. 2021. DOI: https://doi.org/10.1109/CINTI53070.2021.9668337. URL: https://ieeexplore.ieee.org/document/9668337.