Outlier Identification with MFV-robustified Linear Regression in case of Economic Convergence of EU NUTS Regions

Ferenc Tolner[1,2], Balázs Barta[1], and György Eigner[3,4]

[1]Pannon Business Network Association, Szombathely, Hungary
ferenc.tolner@am-lab.hu, balazs.barta@pbn.hu

[2]Applied Informatics and Applied Mathematics Doctoral School, Óbuda University, Budapest, Hungary,

[3]University Research and Innovation Center, Physiological Controls Research Center, Óbuda University, Budapest, Hungary, eigner.gyorgy@uni-obuda.hu

[4]John von Neumann Faculty of Informatics, Biomatics and Applied Artificial Intelligence Institute, Óbuda University, Budapest, Hungary,
eigner.gyorgy@uni-obuda.hu

Abstract: Despite being questioned, absolute economic β-converge is a widely applied method for investigating cohesion tendencies among EU regions. Corresponding data further supports the application of the theory. Relying on our previous results the present study utilizes MFV-robustified linear regression in order to identify over- and underperforming regions. For this purpose regional GDP and NDI data of EU NUTS2 and NUTS3 level regions are used from the time period of 2000-2020. Since underlying data distributions have typically long tails, are highly skewed and contaminated by several outliers robust statistical approaches are advised. The outlined procedure suggests that economic convergence tendency among EU regions is less expressed than conventional β-convergence would claim. Furthermore, by substituting mean values by "Most Frequent Values" introduced by Stener et. al. in corresponding calculations regions can be found to be greatly deviating predicted by conventional convergence theories that would otherwise be masked due to data characteristics.

Keywords: Most Frequent Value; MFV-robustified linear regression; outlier detection; non-normal distribution; economic absolute β-convergence

1 Introduction

Regression problems 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, however various techniques are known in order to keep unwanted deviations under control [1, 2].

Violations to general assumptions of conventional statistical procedures inevitably lead to model misspecifications and erroneous results that are to be avoided at all costs. 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[1] and the level of break-down point[2] [2–5].

Parametric models often rely on strict assumptions hard to hold in practice but can serve with more accurate and easier to interpret estimations when they are met. In contrast non-parametric models can serve with less sharp estimates or results may be harder to interpret. 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 [6].

It has to be emphasized though that a-typical observations can be identified only compared to an existing model, which of course are generated by incorporating those elements that presumably cause the distorting effects. 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 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 [2].

In the present study our intention is to utilize the *Most Frequent Value* (MFV) procedure[3] for the well-known economic growth model problem described by absolute β-convergence. This theorem implies that poorer regions shall grow faster in the long-run compared to stronger ones due to free capital movement to locations with cheaper investment options. According to this concept weaker economies with higher growth rates in their per capita financial indicators shall lead to decreasing differences among richer and poorer regions [7, 8].

Literature lists various studies building upon economic absolute β-convergence,

1 Amount of information that can be extracted from data. Is to be calculated as the ratio of the minimal asymptotic variance (originated from the Cramer-Rao bound) and of the estimated asymptotic variance of the given statistical estimate.

2 Largest amount of a-typical observations the applied procedure can still handle without failing to serve with a reasonable estimate.

3 Robust statistical technique introduced by Hungarian researchers under the coordination of Ferenc Steiner and found applications mostly in earth science related fields.

nevertheless several authors question its applicability and finds its assumptions too restrictive with regard to real-life conditions [9–13]. On the other hand – as our data will demonstrate as well – a negative linear relationship does exists between initial levels and growth rates of the same per capita financial indicators, which supports the validity of the concept. Furthermore, corresponding literature seems quite divided regarding conclusions on cohesion tendencies measured within country and NUTS[4] regional levels of the European Community. Numerous studies observe certain level of convergence [10, 12, 14], while there are many, who point out diverging tendencies [11, 15–17], find converging clubs of regions or merely time dependent cohesion tendencies [18–21].

The rest of the paper is structured as follows: The investigated dataset is introduced in Sec. 2. with its general attributes then the MFV procedure is introduced for linear regression as a two dimensional robust parameter estimation alternative in Sec. 3. and subsequently its resistance against outliers is demonstrated. Thereupon, the application of the depicted algorithm is outlined in case of economic absolute β-convergence of EU NUTS regions in Sec. 4. with special emphasis on differences between data characterized even by their mean- or by their MFV values and on the identification of emerged outlier points compared to the resulted model. The provided approach highly reduces the distorting effects of outliers and non-Gaussian characteristic of the data at hand and thereby enables a robust and resistant model for comparison with similar works of different time periods, range of regions, applied statistical approaches or involved financial indicators.

2 Investigated Data Set

For our research 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. The exact time intervals and limitations regarding missing data are listed in Table 1. The data were accessed at *Eurostat* that is the responsible institution within the European Community for the dissemination and harmonization of statistical information [22]. As can be seen in Table 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 outlined in Sec. 4. and the amount of missing data was also marginal within the listed time periods therefore could not have significant influence on our findings either.

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

[4] Nomenclature of Territorial Units for Statistics (NUTS): Geocoding standard among the EU member states for referencing the subdivisions of countries for statistical purposes.

Table 1
GDP and NDI per capita income measures for different regional levels accessed at *Eurostat*.

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

progress and can be used to compare development of economies, whereas NDI as a metric of the standard of living [23]. The outlined financial indicators were obtained in *EUR per inhabitatnt* 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 [22].

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 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 point as outliers when they are beyond the 1st and 3rd quartiles more than 1.5·IQR[5] it can be seen that the 1D distributions may contain several outlier suspicious items (see Fig. 1).

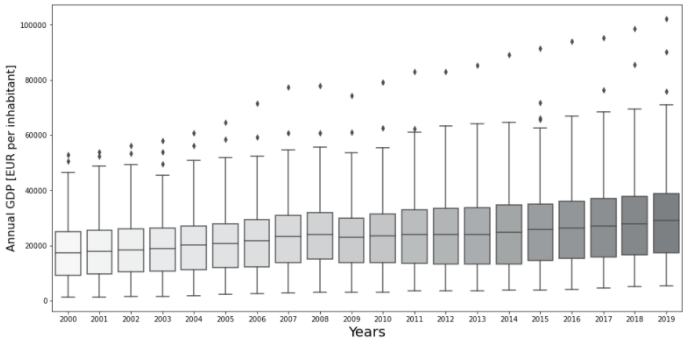


Figure 1
Annual distribution of GDP per capita values measured in EUR in case of NUTS2 regions.

5 Interquartile Range

Table 2
Results of Shapiro-Wilk tests with corresponding p-values in brackets.

	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)	-	-	-	-	-	-	-	-

It has to be noted though, 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 kind of measurement error or being a member of other populations. Furthermore, our data cannot be treated as a random sample of a larger population, since we posses the whole population thus 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 [24, 25].

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 2).

3 Methodological Approach

For handling non-normal data contaminated by outliers the *Most Frequent Value* (MFV) approach will be utilized. The concept of the MFV approach has been developed by Steiner et. al. relies on the basics of robust statistics, where one of the main goals is to reduce the negative effects caused by objects that are far-lying from the "bulk" of the data. Within the MFV theorem this is achieved by a weighting procedure, where the weights correspond to a Cauchy-distribution that takes far-lying points with less degree into consideration. The thereby resulted location parameter of a data distribution $\{x_i\}$ is called its Most Frequent Value – or shortly MFV – and it is represented by $M_{k,x}$ in Eq. 1. (not identical with the mode of the distribution). This value has to be computed via an iterative procedure where besides the location parameter the corresponding scale parameter is calculated as well. This latter given in Eq. 2. is called the *dihesion* (ε) that characterises the dispersion of the data in a robust and outlier resistant way.

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

$$\varepsilon^2 = \frac{3 \cdot \sum_{i=1}^{n} \frac{(x_i - M_{k,x})^2}{(\varepsilon^2 + (x_i - M_{k,x})^2)^2}}{\sum_{i=1}^{n} \frac{1}{(\varepsilon^2 + (x_i - M_{k,x})^2)^2}} \tag{2}$$

The above formulas can be derived from the minimization of the Kullback-Leibler information divergence as well, where the parameter k is a constant for which recommendations are provided by the authors depending on the distribution of data at hand. In practice however the exact mathematical distribution of the data is not known in advance, therefore the selection of $k = 2$ is recommended in order to maintain overall acceptable statistical efficiency [24, 26].

As described above, the MFV of a data distribution can be interpreted for practitioners as a weighted average that has to be calculated in an iterative way. Nevertheless, the theory offers possibility for further generalizations and statistical applications in higher dimensions. For this purpose the so called P-norm has

been introduced that relying on the application of the MFV theorem enables a robust alternative for optimization problems instead of the minimization of the L^2-norm. For data of non-Gaussian error distribution or contaminated with outliers this approach promises much higher statistical efficiency for a wide range of distributions [4, 24, 25, 27, 28].

For an "MFV-robustified" linear regression problem the minimization of the expression given in Eg. 3. has to be done that can be shown to be an equivalent form of Eq. 1. accompanied by the equation for the dihesion given in Eq. 2.

$$G(\varepsilon, M_{k,x}) = \sum_{i=1}^{n} ln\left[(M_{k,x} - x_i)^2 + (k\varepsilon)^2\right] \tag{3}$$

Eq. 3. essentially represents that instead of the fulfillment of the $\sum_{i=1}^{n}(E_x - x_i)^2 = min.$ expression with respect to the expected value (E_x) as done in case of the L^2-norm the $\sum_{i=1}^{n} ln\left[(M_{k,x} - x_i)^2 + (k\varepsilon)^2\right] = min.$ expression should be considered with respect to the Most Frequent Value where $M_{k,x}$ can denote a higher dimensional robust estimate. For the particular case of linear regression this estimate is sought in the form of: $M_{k,\mathbf{x}} = a \cdot \mathbf{x} + b$.

After performing the minimization procedure Eq. 4. and Eq. 5. can be obtained for the "MFV-robustified" linear regression. It can be seen that the equation system is a weighted form of the regression problem based on the ordinary least squares (OLS) method [2]. Nevertheless, it has to be extended with Eq. 2. for the calculation of the ε dihesion parameter and therefore has to be solved in an iterative way[6].

$$\sum_{i=1}^{n} \frac{1}{(k\varepsilon)^2 + (ax_i + b - y_i)^2} \cdot [(ax_i + b - y_i) \cdot x_i] = 0 \tag{4}$$

$$\sum_{i=1}^{n} \frac{1}{(k\varepsilon)^2 + (ax_i + b - y_i)^2} \cdot [ax_i + b - y_i] = 0 \tag{5}$$

Unfortunately Eq. 4. and Eq. 5. describes a nonlinear system, therefore cannot be solved directly in each "MFV-iteration" steps. This nonlinear system can be solved for instance by the generalized Newton's method where the partial derivatives of the Jacobian matrix can be given in an analytical form. The iterative formula to be handled in the $(m+1)$-th step ca be given as:

[6] In the following each step of the iteration for solving the equation system will be denoted as "MFV-iteration".

$$\begin{bmatrix} a^{(m+1)} \\ b^{(m+1)} \end{bmatrix} = \begin{bmatrix} a^{(m)} \\ b^{(m)} \end{bmatrix} - \begin{bmatrix} \sum_{i=1}^{n} x_i^2 A_i^{(m)} & \sum_{i=1}^{n} x_i A_i^{(m)} \\ \sum_{i=1}^{n} x_i A_i^{(m)} & \sum_{i=1}^{n} A_i^{(m)} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} x_i B_i^{(m)} \\ \sum_{i=1}^{n} B_i^{(m)} \end{bmatrix} \tag{6}$$

Where $A_i^{(m)}$ and $B_i^{(m)}$ in the (m)-th step are:

$$A_i^{(m)} = \frac{(k\varepsilon^{(m)})^2 - (y_i - a^{(m)}x_i - b^{(m)})^2}{[(k\varepsilon^{(m)})^2 + (y_i - a^{(m)}x_i - b^{(m)})^2]^2} \tag{7}$$

$$B_i^{(m)} = -\frac{(y_i - a^{(m)}x_i - b^{(m)})}{(k\varepsilon^{(m)})^2 + (y_i - a^{(m)}x_i - b^{(m)})^2} \tag{8}$$

In order to obtain an MFV-robustified estimate for the linear regression task a nested iteration has to be performed. In each step of the MFV-iteration the ε dihesion has to be calculated according to Eq. 2. then the resulted nonlinear system has to be solved. Since the initialization of nonlinear problems is crucial, Steiner et. al. suggested the OLS estimates (a_0, b_0) for the first iteration that in case of a 2D problem can be gained from [24]:

$$\begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} \tag{9}$$

The implemented algorithm for the MFV-robustified linear regression with corresponding stop conditions in our case were as follows:

1. Initialization a_0 and b_0 parameters from OLS line regression.
2. Initialization of dihesion with the maximal residuals measured from the fitted OLS line in positive and negative directions: $\varepsilon_0 = max(r_i^+) - max(r_i^-)$
3. "Inner iteration": Solve nonlinear equation system given in Eq. 6. by generalized Newton'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. 2. (Stop condition: $\varepsilon^{(k+1)} - \varepsilon^{(k)} \leq 10^{-5}$).

The outlier resistance of the MFV-robustified linear regression model compared to the OLS-fitted line is illustrated on Fig. 2., where a single vertical outlier is placed among the data that otherwise perfectly fits a straight line. The OLS fit is

biased by the outlier, while the MFV-robustified regression line remains on the ”bulk” of the data.

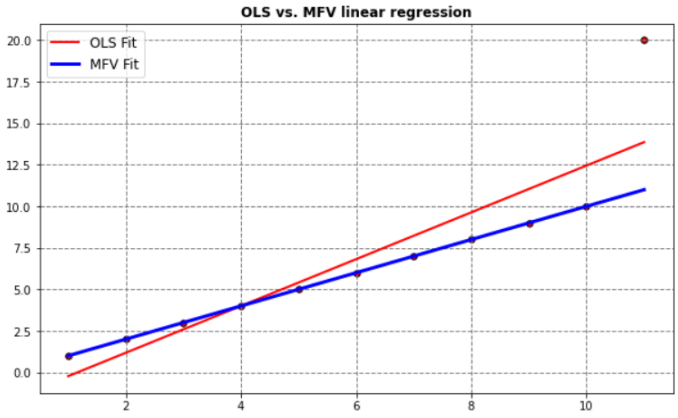


Figure 2
Demonstration of the difference of OLS and MFV-robustified linear regression in the presence of a simple outlier (in the top-right corner). The latter fits the ”bulk” of the data more efficient.

By extending the above simple example with additional points to 100 all together 23 of them could be modified in the same way as it is shown on Fig. 2. without the MFV-fit failing to serve the expected estimate. This can let us have a view on the break-down properties of the algorithm in such extreme cases, although further investigations or even Monte Carlo simulations would be necessary to get founded theoretical background. Moreover, in case of replacing vertical outliers with bad leverage points the estimate can fail at much lower outlier rate, which draws attention to further research options [2, 29].

In the practical data analysis to be outlined in Sec. 4. besides specifying the ”bulk” of the data, the identification and classification of outliers compared to the robust and resistant regression line is of great interest. Therefore, a definition for ”outlyingness” is required. In order to generate comparable results of data with Gaussian error distribution we use the dihesion as a consistent estimator of the standard deviation. In case the amount of data covered by $\pm\varepsilon$ distance is known (let us indicate this portion by ”R”) the probability of observing data within this distance can be expressed as:

$$\mathbb{P}(|x-\mu| \leq \varepsilon) = \mathbb{P}\left(\left|\frac{x-\mu}{\sigma}\right| \leq \frac{\varepsilon}{\sigma}\right) = R \tag{10}$$

Therefore, for normally distributed data we must have:

$$\Phi\left(\frac{\varepsilon}{\sigma}\right) - \Phi\left(-\frac{\varepsilon}{\sigma}\right) = R \tag{11}$$

That eventually gives the relationship between the dihesion and standard deviation of:

$$\varepsilon = \Phi^{-1}\left(\frac{R+1}{2}\right) \cdot \sigma \tag{12}$$

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

$$\hat{\sigma} = A \cdot \varepsilon = \left(\Phi^{-1}\left(\frac{R+1}{2}\right)\right)^{-1} \cdot \varepsilon \tag{13}$$

,

where "A" denotes a constant distribution dependent scale factor. With a consistent estimate for the characterization of far-lying data points compared to the MFV-robustified linear regression line the recommendations of [30–32] are followed that considers a point an outlier if one of the following selected criterion is met:

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

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

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

As an arbitrary selection, in our further investigations the outliers will be classified according to the moderately conservative approach:

$$\textbf{Weak outlier: } |x_i - M_{k,x}| > \varepsilon \text{ and } |x_i - M_{k,x}| \leq 2.5 \cdot A \cdot \varepsilon$$
$$\textbf{Strong outlier: } |x_i - M_{k,x}| > 2.5 \cdot A \cdot \varepsilon$$

4 Results, Discussion

The economic absolute β-convergence states the existence of a negative relationship between growth rate and initial income level in the form of [33]:

$$\frac{1}{T} \cdot ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = \alpha + \beta \cdot ln(y_{i,0}) + \varepsilon_i, \tag{14}$$

where $y_{i,T}$ and $y_{i,0}$ are the per capita economic measures for the i-th sample at the end and beginning of the investigated time period. T given in years is the length

of the time interval, ε_i is the error term for the i-th observation and α, β are the estimated intercept and slope parameters for the regression line respectively [34].

The left side of Eq. 14., the so called "overall annual growth rate" can be approximated with its first order Taylor expansion as[7]:

$$ln\left(\frac{y_{i,T}}{y_{i,0}}\right) = ln\left(\prod_{t=0}^{T}\frac{y_{i,t+1}}{y_{i,t}}\right) = \sum_{t=0}^{T} ln\left(\frac{y_{i,t+1}}{y_{i,t}}\right) \approx \sum_{t=0}^{T}\left(\frac{y_{i,t+1}}{y_{i,t}} - 1\right) = \sum_{t=0}^{T}\left(\frac{y_{i,t+1} - y_{i,t}}{y_{i,t}}\right). \tag{15}$$

By this alteration of Eq. 14. the averages of annual growth rates can be used on the left side of the equation. This approximation enables a more detailed investigation of economic growth because instead of a coarse indicator utilizing only the start and end values a measure containing more statistical information on time evolution can be viewed. Furthermore, besides of the mean values the MFVs were calculated that provides a higher outlier resistance for each instance of the data set. Consequently, robust and resistant location parameters are obtained that represents 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 see Fig. 3.

In order to get a comprehensive picture on the differences between the original relationship described by Eq. 14. 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. the average of relative changes in case of substituting mean values on the left ($\%_{Mean}$) remain below 10% in all of 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.

According to the iterative parameter estimating procedure outlined in Sec. 3. the slope- ($a_{MFV}, \tilde{a}_{MFV}$), intercept- ($b_{MFV}, \tilde{b}_{MFV}$) and dihesion ($\varepsilon_{MFV}, \tilde{\varepsilon}_{MFV}$) values have been computed for the means of annual growth rates and MFVs of annual growth rates respectively. 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"

[7] Since the annual change in the investigated GDP and NDI measures are typically less than 10% for all of the instances the $ln(1+x) \approx x$ approximation is applied, where x is close to zero.

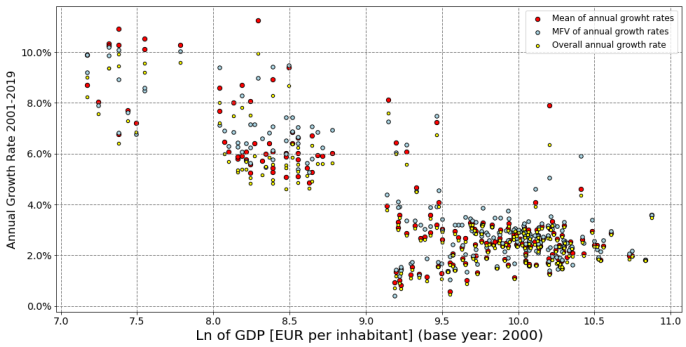


Figure 3
Different measures of growth rates within the investigated time period for NUTS2 regions (dataset: GDP [EUR per inhabitant]).

Table 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.

	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

necessary to reach the specified exit criteria for convergence $(n, \tilde{n})$, the ratio of data lying within 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.

For characterizing the rate of convergence among the investigated spatial entities within the framework of economic absolute β-convergence the slope of the fitted lines have to be used. The bigger negative values correspond to faster convergence. As can be seen from the slope parameters listed in Table 4. 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 the researchers shall assume highly aggregated and averaged data into less than 30 data points that can be uncertain or less accurate. For NUTS2 or NUTS3 regions much more data are at hand that provides more trustworthiness. For these data with more regional instances and consequently higher territorial resolution the MFV-robustified line regression method (except NUTS2 level for GDP [PPS per inhabitant]) served with a conclusion that a less exaggerated convergence among the regions shall be expected compared to the ordinary analyses done regarding absolute β-convergence based on conventional statistical procedures.

Table 4
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.

	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
ε_{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
$\tilde{\varepsilon}_{MFV}$	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

According to Table 4., among the slopes of the fitted MFV-robustified regression lines diverse differences can be observed comparing results for means of annual growth rates and MFVs 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 cases differences among the weakly- and strongly outliers turned out to be more relevant.

The identification of outliers has been done compared to the fitted MFV-robustified regression line. Data points lyin g 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.) were labelled as weakly outlying, while those ones that can be found further from the fitted MFV regression line than this distance were labelled as strongly outlying points (see Fig. 4. and Fig 5.) in accordance with Sec.3.

The more robust and outlier resistant MFV-robustified linear regression enables the identification of ”interesting” objects that would have otherwise been masked by the inflated variance of the data. In our case not just the less expressed speed of economic convergence within the framework of absolute β-convergence was pointed out but the regions over- or under-performing 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 cases on Fig. 6., 7. and 8.

In case of some countries strikingly different outliers occurred. In case of GDP in EUR per capita values on NUTS2 level differences in case of Sweden and Poland seems to be the most prominent, while for NDI likewise on NUTS2 level

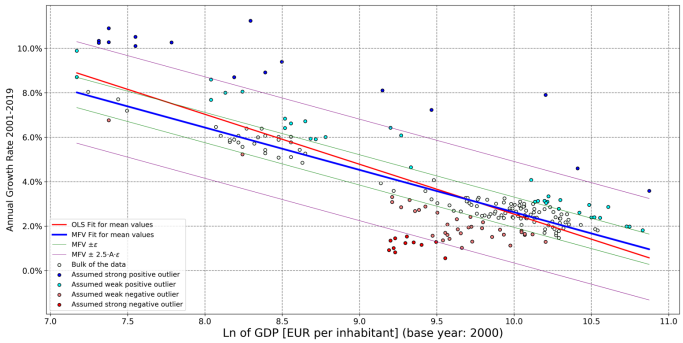


Figure 4
Fitted MFV regression line with corresponding classification of outliers and OLS line fit for comparative purposes (dataset: NUTS2 level, GDP [EUR per inhabitant].

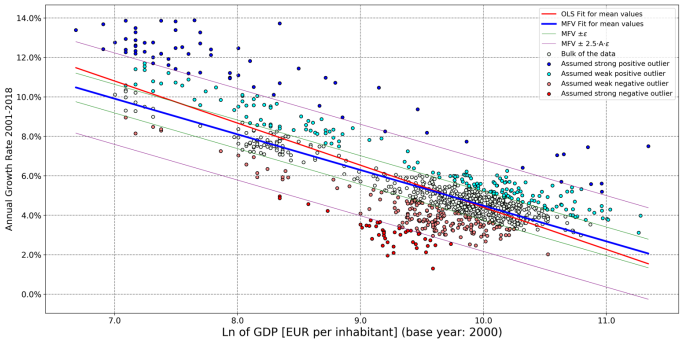
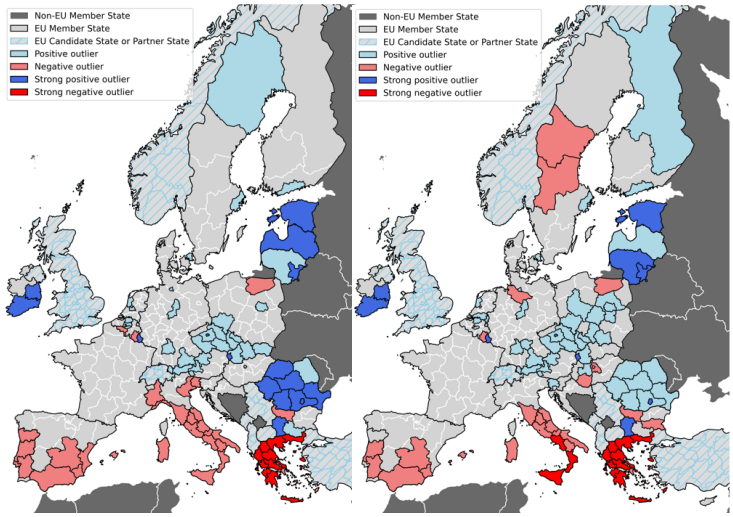


Figure 5
Fitted MFV regression line with corresponding classification of outliers and OLS line fit for comparative purposes (dataset: NUTS3 level, GDP [EUR per inhabitant].

and in EUR per capita dimension French regions gained highly differing labels. The geographical visualisation can be used as a basis for better comparison and the gained classification of regions in general as input for further field relevant researches that analysis is unfortunately far beyond the limits of the current study.

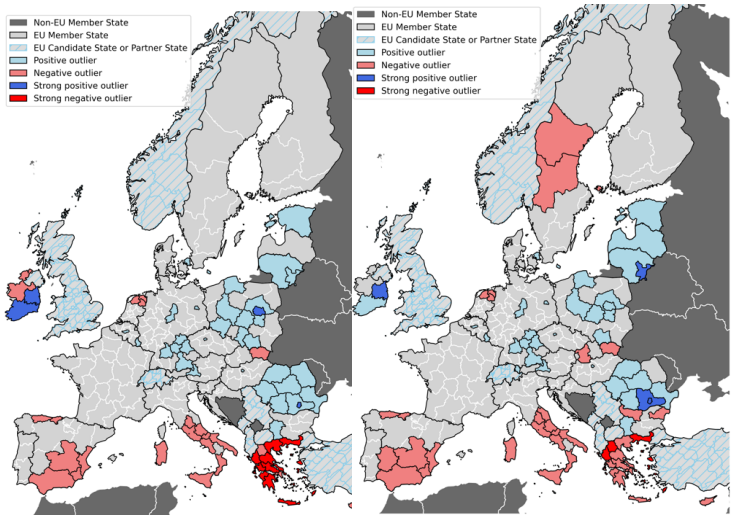
Conclusions, Future Work

The present work investigated the economic absolute β-convergence of EU regions of NUTS2 and NUTS3 level with respect to GDP and NDI data over the past two decades. A robust linear regression technique has been introduced and applied in order to reduce influencing effects of outliers and long-tailedness of skewed distributions that would otherwise give questionable statistical results. The Most Frequent Value procedure developed by Hungarian researchers and until now applied mainly in the field of earth sciences has been utilized for ”fine-tuning” previous findings regarding regional convergence. To the best of the authors knowledge the MFV procedure has not yet been used regarding economical



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

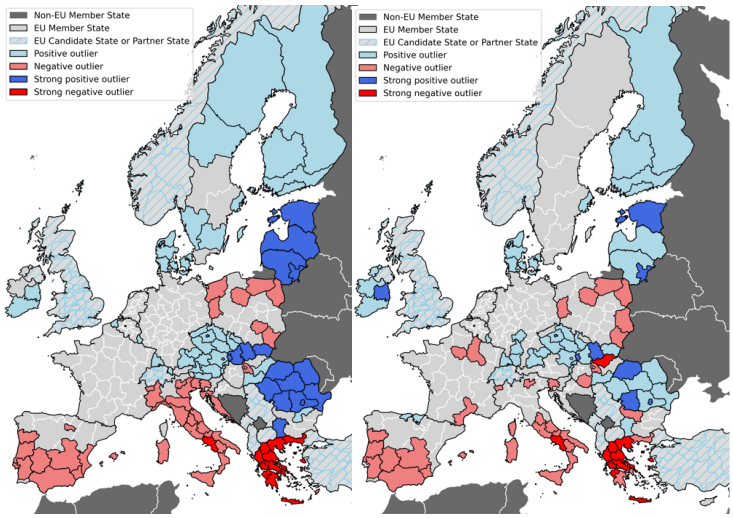
(b)
GDP [EUR per inhabitant] - Fit on MFV values



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

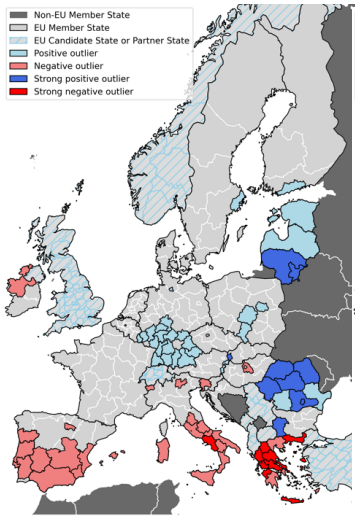
(d)
GDP [PPS per inhabitant] - Fit on MFV values

Figure 6
Estimated outliers for NUTS2 level using GDP per capita values.

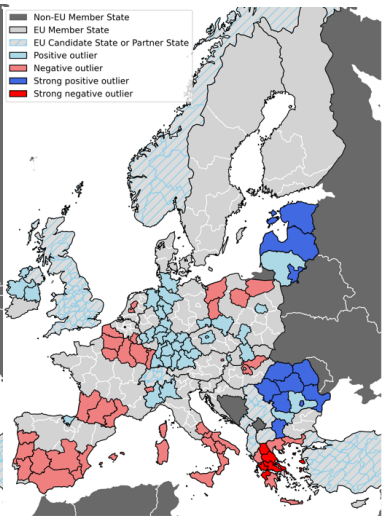


(a)
NDI [EUR per inhabitant] - Fit on Mean values

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

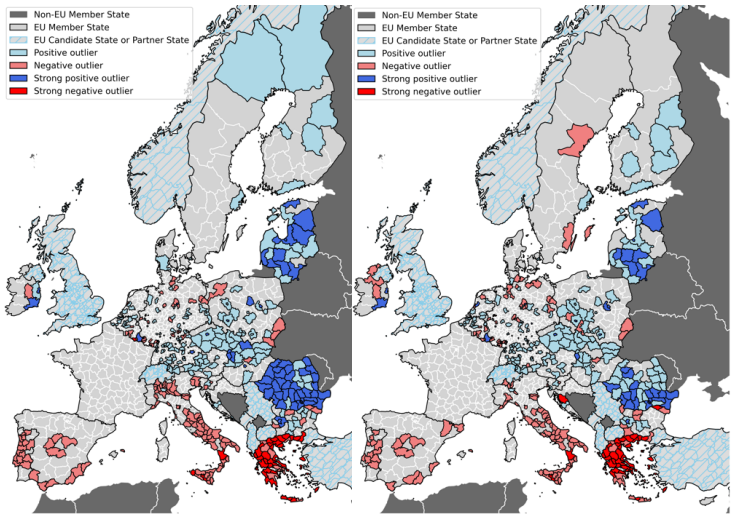


(c)
NDI [PPS per inhabitant] - Fit on Mean values



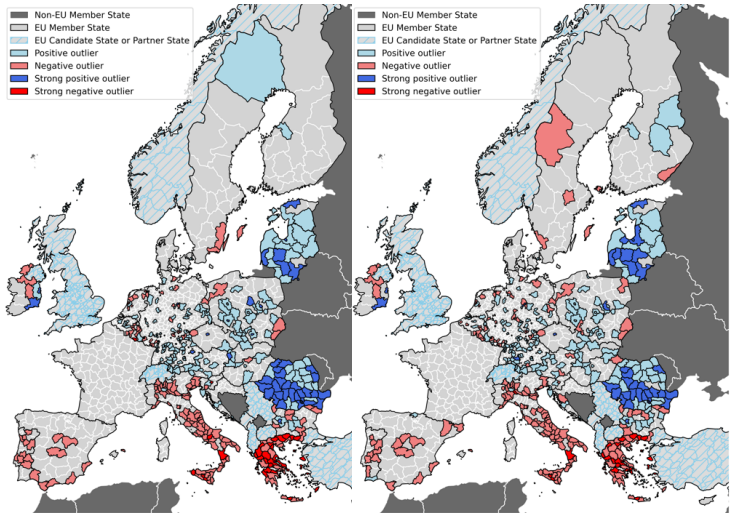
(d)
NDI [PPS per inhabitant] - Fit on MFV values

Figure 7
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 8
Estimated outliers for NUTS3 level using GDP per capita values.

studies until now.

Our results further strengthened the existence of a negative linear relationship between initial per capita financial indicators and growth rates on longer time intervals. Nevertheless, they also suggest that the economic convergence among EU regions is of less speed than the absolute β-convergence assessed by conventional statistical tools would predict, which can basically attributed to distorting effects of outliers and non-normality of the background distributions.

By substituting annual growth rate values of financial indicators with their Most Frequent Values in order to reduce attraction of outliers on the regression line to be fitted and increase the extracted statistical information of the data, regions could be identified that would otherwise be "masked". Such regions have generally over- or under-performed throughout the past two decades compared to the prediction of the convergence theorem. These outlier point together can form the basis of further regional investigations and determination of best practices or remediation plans for corresponding entities.

In a future work we intend to further utilize the higher statistical information extraction capability of the MFV technique and make it use for robustified classification problems in case of annual balance sheet and income statement data. Thereby, our aim is to shed light on contributing factors of organizational resilience by eliminating biasing and distorting effects caused by naturally occurring strong outliers and typically non-normal distributions of financial data.

Acknowledgment

Gy. Eigner was supported by the Eötvös Loránd Research Network Secretariat under grant agreement no. ELKH KÖ-40/2020 ('Development of cyber-medical systems based on AI and hybrid cloud methods'). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 679681). Project no. 2019-1.3.1-KK-2019-00007. has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme. The publication of this article has been supported by the Robotics Special College via the "NTP-SZKOLL-20-0043 Talent management and community building in the Robotics Special College of Óbuda University" project and by the Pannon Business Network Association. The publication was supported by the Applied Informatics and Applied Mathematics Doctoral School of Óbuda University.

References

[1] F. Jiang, G. Liu, J. Du, and Y. Sui. Initialization of k-modes clustering using outlier detection techniques. *Information Sciences*, 332(1):167–183, 2016.

[2] P. Filzmoser and K. Nordhausen. Robust linear regression for high-dimensional data: An overview. *WIREs Computational Statistics*, 13(4):e1524, 2021.

[3] F. Steiner and B. Hajagos. Practical definition of robustness. *Geophysical Transactions*, 38(4):193–210, 1993.

[4] F. Steiner. Comparison of the L2-, L1- and P-norm based statistical procedures in respect of their asymptotic efficiencies. *Magyar Geofizika*, 41(1), Feb. 2000.

[5] L. Ferenczy. A short introduction to the most frequent value procedures (in Hungarian). *Magyar Geofizika*, 29(3):83–94, 1988.

[6] M. A. Medina and E. Ronchetti. Robust statistics: a selective overview and new directions. *WIREs Computational Statistics*, 7:372–393, 2015.

[7] R. J. Barro and X. S. i Mart. Convergence. *Journal of Political Economy,*, 100(2):223–251, 1992.

[8] R. M. Solow. A contribution to the theory of economic growth. *The Quarterly Journal of Economics*, 70(1):65–94, 1956.

[9] K. Gluschenko. Myths about beta-convergence. *William Davidson Institute Working Papers Series*, (4):26–44, 2012.

[10] M. Butkus, D. Cibulskiene, A. Maciulyte-Sniukiene, and K. Matuzeviciute. What is the evolution of convergence in the EU? Decomposing EU disparities up to NUTS 3 level. *Sustainability*, 10(5):1–37, May 2018.

[11] D. Quah. Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics*, 95(4):427–443, 1993.

[12] A. Liontakis, C. T. Papadas, and I. Tzouramani. Regional economic convergence in Greece: A stochastic dominance approach. *50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy"*, Conference Paper, 2010.

[13] D. T. Quah. Twin peaks: Growth and convergence in models of distribution dynamics. *The Economic Journal*, 106(437):1045–1055, 1996.

[14] R. Ezcurra and M. Rapún. Regional dynamics and convergence profiles in the Enlarged European Union: A non-parametric approach.

[15] N. Nenovsky and K. Tochkov. The distribution dynamics of income in Central and Eastern Europe relative to the EU: A nonparametric analysis. *William Davidson Institute Working Paper*, 1063, 2013.

[16] D. T. Quah. Empirics for growth and distribution: Stratification, polarization, and convergence clubs. *Journal of Economic Growth*, 2(1):27–59, 1997.

[17] G. Anderson. Making inferences about the polarization, welfare and poverty of nations: A study of 101 countries 1970-1995. *Journal of Applied Economics*, 19(5):537–550, May 2004.

[18] M. T. Borsi and N. Metiu. The evolution of economic convergence in the European Union. *Empirical Economics*, 38(2), 2014.

[19] M. R. Szeles. Exploring the economic convergence in the EU's new member states by using non-parametric models. *Romanian Journal on Economic Forecasting*, 1:20–40, 2011.

[20] M. Smetkowski and P. Wójcik. Regional convergence in Central and Eastern European Countries: A multidimensional approach. *European Planning Studies*, 20(6):1–17, 2012.

[21] H. K. Nath and K. Tochkov. Relative inflation dynamics in the new EU member countries of Central and Eastern Europe. *Empirical Economics*, 45(1):1–22, 2013.

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

[23] M. Chocholatá and A. Furková. Income disparities and convergence across regions of Central Europe. *Croatian Operational Research Review*, 7(2):303–318, 2016.

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

[25] F. Steiner. (Editor), Optimum Methods in Statistics. 1997. Akadémiai Kiadó, Budapest, Hungary, 370p., ISBN: 963 05 7439 X.

[26] B. Hajagos and F. Steiner. Investigations concerning resistance - importance of the choice of the formula determining the scale parameter. *Geophysical Transactions*, 38(4):211–230, 1993.

[27] F. Steiner. New results on the theory of the most frequent value procedures. *Geophysical Transactions*, 41(1-2):1–21, 1997.

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

[29] P. Filzmoser, S. Höppner, I. Ortner, S. Serneels, and T. Verdonck. Cellwise robust m regression. *Computational Statistics and Data Analysis*, 147:106944, 2020.

[30] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[31] J. Miller. Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology Section A*, 43(4):907–912, 1991.

[32] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[33] W. Dyba, B. Loewen, J. Looga, and P. Zdražil. Regional development in Central-Eastern European countries at the beginning of the 21st century: Path dependence and effects of EU cohesion policy. *Quaestiones Geographicae*, 37(2):77–92, 2018.

[34] K. Dvoroková. Sigma versus beta-convergence in EU28: Do they lead to different results? *WSEAS Transactions on Business and Economics*, 11(1):314–321, 2014.