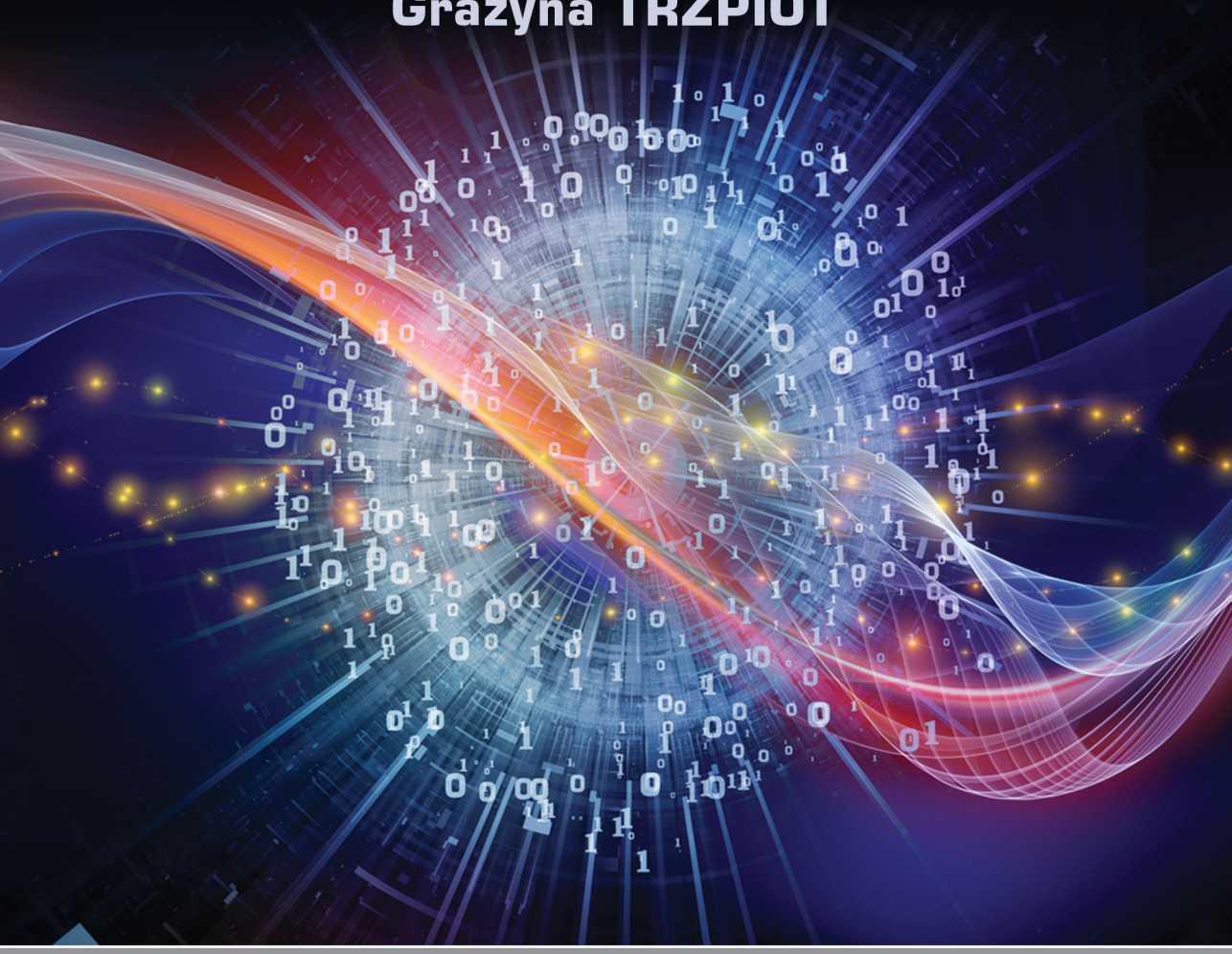
Multidimensional data modelling and risk analysis

Edited by

Grażyna TRZPIOT



Publishing House of the University of Economics
in Katowice

Multidimensional data modelling and risk analysis

Edited by
Grażyna Trzpiot

Katowice 2023

Scientific publication

Editorial Board
Janina Harasim (chair), Monika Ogrodnik (secretary),
Małgorzata Pańkowska, Jacek Pietrucha, Irena Pyka, Anna Skórska,
Maja Szymura-Tyc, Artur Świerczek, Tadeusz Trzaskalik, Ewa Ziemba

Reviewer
Stanisław Heilpern

Publishing editor
Karolina Koluch

Typesetting
Daria Liszowska

Cover design
Emilia Gumulak

Cover picture © agsandrew – Photogenica

ISBN 978-83-7875-868-6

doi.org/10.22367/uekat.9788378758686

Publishing House of the University of Economics in Katowice
ul. 1 Maja 50, 40-287 Katowice, tel.: +48 32 257-76-33
www.wydawnictwo.ue.katowice.pl, e-mail: wydawnictwo@ue.katowice.pl
Facebook: @wydawnictwouekatowice

Contents

Introduction

This scientific monograph presented for readers concerns risk analysis and multivariate data modeling. It contains a wide range of problems that have been addressed, including the understanding of risk in economic theories, the measurement of capital market risk, or the study of the energy market. In addition, demographic issues related to mortality, its analysis and forecasting are addressed, as well as issues related to youth unemployment and analysis of the COVID-19 pandemic.

That monograph which is being prepared is the outcome of the research work of the staff and doctoral students of the Department of Demography and Economic Statistics in recent years. Last year, a nationwide conference SIDVRA 2022 took place, which additionally celebrated the tenth anniversary of the establishment of our Department and was at the same time a presentation of preliminary research results.

The guest of honour at this conference was Professor Józef Stawicki, Ph.D., who presented a lecture on "Perception of Risk in Economic Theory". The transcript of this lecture is chapter one of the presented monograph. So the opening chapter has more didactic character. It is an extremely interesting overview of theories, views on risk appearing in various aspects or periods related to economic science. There are also considerations of the concept of probability, decision-making issues, insurance or investment activities in the wider sense.The next two chapters focus the authors' attention on capital market risk measurement analysis.

In chapter two, Grażyna Trzpiot replaces the classical linear regression model with a Gini regression model. Specifically, she introduces the Gini regression coefficient instead of the classic beta coefficient, a measure of systematic risk. The Gini regression coefficient is robust to outlier observations and does not require quite limiting practical assumptions, including the assumption of normality of distributions. It also presented a multivariate version of it and introduced the extended Gini regression coefficient. In addition, it allows for the characterization of the researcher's approach to risk in the market. The empirical study of market risks was used both versions of the proposed approach: a Gini regression model and a Gini regression model with EGRC (extended Gini regression coefficient) to reflect the investigator's perception of risk aversion in the market. The issues raised in this chapter are quite important from the point of view of equity investment, or risk analysis.

Dominik Krężołek, in chapter three, applied non-parametric econometric models in risk analysis. Non-parametric econometric modelling is a statistical method used to estimate the same relationships that parametric models but making any assumptions about the functional form of the relationship. Non-parametric models are more flexible than classical models, provide more accurate estimates of the relationship between variables, do not require knowledge of distributions and do not require additional assumptions. In the theoretical part, kernel estimators were introduced, basic definitions and their properties were given. Then the basic risk measures VaR and ES are discussed. The empirical part is devoted to the kernel estimation of the mentioned risk measures and comparing the accuracy of the estimation with classical estimators assuming student's distribution and GED.

Chapter four was prepared by Alicja Ganczarek-Gamrot, who performs research on the energy market. She analysing multivariate time series, with the problem of non-uniform frequency of observations. The data from multiple sources is registered at intervals of varying length. She deals with the comparison of distance measures of time series. Such distance measures are used to group multivariate time series. Especially when dealing with the problem of non-homogeneous frequency of observations, non-stationarity of time series, or the presence of time-varying correlations between them. Classical distance measures such as Euclidean, Frechet, or DTW do not always pass the test in these cases. In addition to these three, the author considers three versions of the CORT measure, two measures based on the correlation coefficient, measures using ACF and PACF functions, and three using periodograms. In order to compare the aforementioned distance measures, the author used energy prices on electricity published on the Noord Pool platform. She considered two time series groupings, and used the Silhouette index to assess the quality of the grouping.

In an era of prolonging human life and risking longevity, a recent and essential topic is being addressed by the Justyna Majewska in chapter five. She deals with the study of mortality trends in Europe by applying multivariate functional analysis for this purpose. The data was taken from the Human Mortality Database and concerned 20 European countries and covers the years 1960-2019. Mortality pattern curves were created for each country and year pair, which were smoothed using glued functions (B-splines). The author presented the evolution of countries in terms of mortality: over the years from 1960 to 2019. She conducted an analysis of these changes taking into account infant mortality and accidental and premature mortality. She compared the development path of each country. She used functional principal component analysis to describe a group of countries. The topic covered in this chapter is important, not only from the point

of view of demography, but also from the point of view of the economy or health care. The mortality analysis makes it possible to examine the current demographic situation, as well as those in the future. Among other things, these projections are derived from an analysis of observed trends. In addition, mortality rates make it possible to predict the future labor market situation.

In the sixth chapter, Agnieszka Orwat-Acedańska addresses the issue of mortality forecasting in powiats in Poland. The purpose of the chapter was to assess the quality of mortality rates and life expectancy. This assessment was examined using the standard MAE measure. Six prediction models were considered: naive, standardized mortality rate, standardized mortality rate at the provincial level (used by the Central Statistical Office), rate ratio model, mortality surface and Brass's relational model. Three cases were considered: all counties combined, age grouping separately for men and women. For each model, the average MAE for the mortality rate forecast and life expectancy was determined. The issue of mortality forecasting is very important from the point of view of the pension system and health care. It is very good that this problem has been addressed in this monograph. It is also important to analyze the accuracy of the forecasts of the various models considered in this chapter.

The problem of unemployment occurring among young people who have completed their education is addressed in chapter seven by Magdalena Kawecka. This is an extremely important topic both for those affected by unemployment and for the further development of the country. It requires an effective and prompt solution. In the introduction, the author skillfully presented the importance of the problem of unemployment occurring among young people, and illuminated the situation of these people. The author constructed a generalized linear model. She used this model to isolate a group of variables affecting the decline in the unemployment rate and to examine the situation in each voivodeship.

The purpose of the study, prepared by Zuzanna Krysiak and Grażyna Trzpiot, described in chapter eight was to analyze time series describing the COVID-19 pandemic phenomenon. The analysis itself, conducted on time series for 6 countries: Poland, Italy, Mexico, Chile, India and Israel is good. In this analyses some specific model was estimated based on ARIMA and SARIMA class of models, allows further analysis of the problem, forecasting observations on the incidence of the disease, studying the relationship with vaccination or deaths.

The authors of the research are employees and doctoral students of the Department of Demography and Economic Statistics. Selected issues of multivariate modeling of demographic and economic data are covered in this monograph. We use available datasets published by Eurostat, stock exchanges and commodi-

ty exchanges, as well as the source of the data used in the empirical analyses is the CSO database and the Human Mortality Database. The analyzed sets have different structures, different dimensions, and are linked to different temporal and spatial measurements. The methodological layer of the research takes up the latest results and developments in the area of inference and analysis of multidimensional data sets, the utilitarian one covers detailed applications, and in addition, a risk analysis is carried out.

The authors of the monograph sincerely thank the Reviewer for his insightful review, of such numerous and thematically diverse parts of this monograph, which made a contribution to the quality of the final version of this book.

Grażyna Trzpiot

Chapter I

Perception of risk in economic theory

Józef Stawicki

The thoughts I want to share are didactic rather than exploratory and scientific by nature. Education in mathematics allowed me to practice the study of economics formally. Learning about the history of economics, schools of economics, directions was an unattractive activity and rather difficult. It can be said that the 'spirit' of mathematics has stuck with me all the time, and hence the classical approach of the Lausanne school was very familiar to me. The analysis of probabilistic models in economics started at the doctoral level (Markov chains) remained the basis for thinking about economics in the spirit of Prof. Zbigniew Czerwinski's reflection – "If you don't want to be a philosopher, then don't wonder where the residual component ε in an econometric model came from" [Czerwiński, 1992, p. 204]. This spirit is still active; however, there comes a time when we all become philosophers; so do I. I did not learn probability from Bruno de Finetti's textbook, hence the phrase 'probability does not exist' was strange for me for a very long time. In his lecture I (Tuesday, March 13, 1979), he wrote: "In my own view − as many will already know, if they are familiar with this subject − probability has only a subjective meaning. That is, I think it is senseless to ask what probability an event has per se, abstractly. By an 'event', on the other hand, I mean a single well-defined fact" [Finetti, 2008, p. 3].

Growing up and growing into 'classical' or axiomatic probability theory has its consequences. And dealing with logics during my studies made me look for axiomatic systems in economics. Over time, I got used to 'natural economics' and began to be surprised by questions from my colleagues in the Mathematics Department like "then what measure do you apply to research...?". I have been teaching subjects with the keyword 'risk' for many years. It must be said that such a concept in the economics of socialism practically did not exist and hardly entered into the canon of teaching during the transformation period. I learned about risk in economics and management from the first MPaR[1] conferences as an active or passive participant. These conferences, organized by the strong Ka-

[1] A series of scientific conferences organized by the Department of Operational Research at the University of Economics in Katowice (formerly the Academy of Economics in Katowice).

towice based center, to which I had been associated since the 1970s, had a great impact on my understanding of risk. Over the past years I got acquainted with many monographs, articles; I reviewed a sizable number of promotions works "with risk inside", an even larger number of articles; I wrote something by myself, created something... It will sound trivial, but only now I know how few I understand of what we define as risk, associating it with economics. If I share my reflections with you today it is a bit on the grounds of "I understood something and I want to talk about it" − this is following Prof. Marian Grabowski and his thoughts in his book *Istotne i nieistotne w nauce* [Grabowski, 1998, p. 28]. I do not want to talk about the history of the concept of risk, although it is necessary to refer to it. In this regard, Peter Bernstein's book [Bernstein, 2017] and many works by Terie Aven [Aven, 2012; Aven and Krohn, 2014; Aven and Reniers, 2013] (very little cited in economic works) are excellent. It is impossible to see risk without knowing the history of the concept of probability. That's why it's worth referring to Blais Pascal (think about the so-called Pascal's bet) or Nicolas and Daniel Bernoulli (the St. Petersburg paradox).

In my classes I like to make provocative introductions to encourage students to think. I start a lecture on random variables by having the bursar at our university throw a dice to know how many thousands of zlotys to pay Stawicki. The students, who are already after a series of lectures on statistics, smile, but already when I ask them about the meaning of studying the distribution of wages as a random variable, they start to get lost in thinking what a random variable is.

One time I tried to repeat the experiment described in John Allen Paulos' book *Innumeracy. Matematyczna ignorancja i jej konsekwencje* presented there as a stock market scam [Paulos, 2012]. Unfortunately, with a small number of students, the experiment failed. Only the discussion was interesting. The experiment consists in the fact that I sent a message to half of the students: "The quotation of company X will rise tomorrow", and to the other half: "The quotation of company X will fall tomorrow". After the quotation was realized, to the right group (dividing it into halves) I sent information about the rise or fall of the quotation of company X. After the third message sent in this way and the realization of the phenomenon, I asked only those who had received three correct 'forecasts' − their number being 1/8 of the group − whether they were willing to pay for another forecasting message. The discussion revolved around the questions: Where is the randomness? What is the risk in the experiment and for whom? What measure of probability to use? What if, like Paulos, the experiment was repeated many times for a group of thousands?

During a course on ‘project management’ when discussing the topic of ‘project risk’ I presented a classic table with linguistic categories of probability (large, medium, small). These are interval probabilities, so to speak, successfully used by Prof. Grażyna Trzpiot in other topics [Trzpiot, 1999].

Table 1. Assessment of probability

Probability of factor	Possible rate of occurrence
Very large (occurrence almost certain)	> = 1 for 2 cases 1 for 3 cases
Large (large probability of occurrence)	1 for 8 cases 1 for 20 cases
Medium (appearing occasionally)	1 for 80 cases 1 for 400 cases 1 for 2000 cases
Low (relatively low possibility of occurrence)	1 for 15 000 cases 1 for 150 000 cases
Marginal (occurrence is almost unlikely)	< = 1 for 1 500 000 cases

Questions were sprinkled in a natural way. If one case out of two occurred, can you count probabilities and classify them? And if one case in a hundred and fifty cases occurred, what are these situations when a project is supposed to be an innovative endeavour?

After the first lecture on the theory of risk and insurance, I propose to students a game that I will toss a coin to everyone leaving the room. If a head falls out, I pay 10 PLN, if a tail falls out I pay 0 PLN. I ask the question if anyone joins the game expecting one of three questions:

1. Is the coin I will toss symmetrical?
2. What will be the number of tosses for each student, i.e., is it possible to repeat participation in the game?
   and the basic question:
3. How much do you have to pay to join the game?

No question is asked, this third one too, and everyone wants to play. Maybe the scope of knowledge imparted in the first lecture needs to be changed. My dream is to teach this subject in the laboratory by ‘experimenting’ on a group of students with the help of appropriate tools (interactive computer programs with appropriate immediate analysis).

The basic problem in classical risk analysis is the concept of probability and the concept of decision which is usually made on the basis of a utility function (also a random or chaotic utility function or on the basis of an appropriate preference relation). Risk does not exist if the decision-making process is not considered. The existence of even a certain distribution of states of

the external world does not create a situation of risk until decision-making is considered. The decision-making situation generates a state of risk, which varies depending on the decision-making process being considered. Considering the probability of certain weather conditions is just a game. It becomes a risk situation different for a farmer, different for a tourist, different for a driver.

The problem posed by Bruno de Finetti [1975] (that the probability does not exist) is therefore not trivial. If we use the model of the toss of a symmetrical coin then we have to assume the probabilities of tossing a heads or tails are the same equal to ½. And if we take a particular coin then we should statistically check under which model it belongs. We can do this several times. If we observe an economic phenomenon of the rise or fall of a quotation, we can assume after verification (even in a very sophisticated way) what model we will use. However, this model will refer to the past and therefore to the history, and we bear the risk of choosing a model. Without the assumption that the model is valid for the future, there is no point in using it. This brings us back to the conflict between the historical school of economics and abstract-deductive economics. Without a paradigm about the ergodicity of economic processes (as realizations of stochastic processes) or a paradigm about the stability of dynamic systems, practicing economics will not be possible. But the model is, after all, a part of economic theory. The second fundamental question is the problem of individual and mass phenomena. The individual ones are very numerous (if not all) in economics. One would refer to the considerations of Ludwig von Mises [2011] and the Austrian school of economics. The recall of the probability of classes according to von Mises should be understood as the recurrence of the sampling − a natural recurrence, such as the toss of a coin. If we even use the correct model and the sampling is singular, then our knowledge is useless. Von Mises warns gamblers against making the mistaken assumption that knowledge of probability will help them win. A gambler's hope of winning is not based on knowledge of the distribution, but on the desire to win and the belief that one is lucky. However, if the Chevalier de Mere cared about knowing what was more likely: tossing out the six at least once in six tosses of the dice or the twelve when tossing twenty-four times with two dices, it is because he was a great gambler and played several times. And yet the trivial problem of the gambler Chevalier de Mere became the basis for the development (perhaps the birth) of the calculus of probability and beyond. Blaise Pascal, Pierre de Fermat are the names from this period that cannot be forgotten. But already the St. Petersburg paradox makes us link probability with utility. This raises the question of so-called subjective probability. This probability is only in our brains. How to understand that assigned probability in a decision problem called Pascal's bet is an open problem. Marek Wójtowicz

[2016] of the University of Silesia presented some of them in his monograph. The problem of what probability is became even more troublesome when the many paradoxes associated with it were recognized. Krystyna Simons [2019] in her paper talked about some of them. I also refer you to a short paper on subjective probability by Prof. Mirosław Szreder [2004] in Statistical Review. He calls it the personalistic conception of probability. So, if risk does not exist without probability, this concept has many definitions and is interpreted in many ways. The problem of defining probability arose at the very beginning of scientific thought on decision theory. Reference must be made to Ramsey, who died very young in 1930 at the age of 26. He was working on a book entirely devoted to the concepts of truth, belief and probability. Thus, the article *Truth and Probability*, written in 1926 [Ramsey, 1926], is not a definitive statement of the concepts discussed in it. Ramsey would write later that he was not fully satisfied with his explanation of the concept of probability, mainly because he considered it too psychologically based. However, he believed that he had laid the foundation for a new way of approaching the concept of probability. Few modern researchers refer to Ramsey. And after all, it is his theory and Bruno de Finetti's understanding of probability that are the basis of today's thinking. As I understand it, the most complete overview of this topic can be found in the works of Prof. Terje Aven. Based on an extensive literature, not only in economics, he lists nine insights into risk [Aven, 2012]:

1) Risk = Expected value (loss) (R = E),
2) Risk = Probability of an event (undesirable) (R = P),
3) Risk = Objective Uncertainty (R = OU),
4) Risk = Uncertainty (R = U),
5) Risk = Potential/possibility of loss (R = PO),
6) Risk = Probability and scenarios of consequences (severity of consequences) (R = P&C),
7) Risk = Event or consequence (R = C),
8) Risk = Consequences/damages/predictability of those + Uncertainty (R = C&U),
9) Risk determines the impact of uncertainty on objectives (R = ISO).

Some of these insights are very similar; for many people the measure of risk imposes itself, for others it is difficult to imagine a reasonable measure. Going back to the basic idea of risk as a consequence of decision-making (failure to make a decision is also a decision), it is necessary to consider the way the decision problem is presented and the methodology of support in making that decision. A near-perfect overview on support methods was provided by Prof. Tadeusz Trzaskalik [2014]. The question arises how the type of decision problem, the way it describes the situation, and the chosen method fits into specific economic

and management theories. Economics is the science of household management. So, when considering economic problems, we touch management in an important way. In human nature is the desire to make decisions under favourable or unfavourable circumstances. Not making decision is also a decision. The problem of preference as one of the initial concepts (understood as a primary concept) is considered in economics in the broadest sense. It takes its origin in consumer and production decisions. Nowadays, the theory of preferences is most developed in market research. The second important trend of preference analysis is the application of this category in modelling and/or supporting decisions on so-called projects. While in the case of consumer theory on the basis of classical economics, the analysis leads to the aggregation of consumer attitudes, in project analysis decisions are singular in nature and should be analysed that way. Just as in the case of risk there is a fundamental difference between declared attitude and realized attitude (a meaningful example is the research on risk attitudes and taking a form of payoff for the effort of participating in the study in the form of a random payoff or a certain [Tyszka, 2010]) so there is a fundamental difference between declared preference and realized preference. A whole new trend of research very mathematized launched with a discussion by Paul A. Samuelson in a short paper, *A Note on the Pure Theory of Consumer's Behaviour* [Samuelson, 1938] has developed incredibly in the numerous literature under the name of 'revealed preferences' that is, revealed preference theory, in which the risk plays an important role and constitutes a coherent whole [Chambers and Echenique, 2016].

When analysing the history of decision-making problems presented or described in economics, it is worth to characterize them by some dimensions. In decision-making modelling, many elements related to the description of preference relations should be taken into account. The determinants presented below are operational and contribute significantly to the methods of analysing the decision problem:

1. The set of decision variants a_i where $i \in I$. The indices set of decision variants can be either a finite, countable or uncountable set. In classical decision--making problems, it is assumed to be either a finite set or an infinite set (mathematical programming methods); in the case of consumer theory, an incalculable set is assumed (baskets of goods being vectors belong to the n-dimensional real space $\bar{x} \in R^n$).
2. The decision-making process being modelled is a single decision, such as the selection of an investment project. The selection of another good from a specific group of consumer goods or services is a decision that is repeated many times; the repetitiveness depends on the type of good or service.

3. The decision-making process may involve a sequence of many decisions over time that are not independent (e.g., choosing an educational path in successive educational degrees – bachelor's, master's, doctorate). Of course, the risk exists, for example, of receiving an interesting job and salary. This issue is known as the intertemporal choices. An example would also be the decision of income distribution during the working and retirement periods (under the risk of whether I will be paid a fair pension).
4. An important role is played by the measurement of the objective achieved in the decision selection process (it is a precise description of the decision options and their order and the corresponding criterion function), its unidimensionality or multidimensionality and the scale of measurement (nominal, ordinal, interval, ratio).
5. The description of decision options can be either stochastic or fuzzy.
6. The order relation can be a classical binary relation, it can be a fuzzy relation, or it can be a stochastic relation.

In addition, it is now necessary to place the decision-making process in a specific world environment (the so-called external world). However, the basic question is what paradigm (axioms) about a person, his behaviour, perception of the world stands for a given model concept including the suggested algorithms. An interesting example is the bipolar method developed by Prof. Ewa Konarzewska-Gubala [1991] and beautifully developed by Dorota Górecka [2009]. Another example is the fuzzy method using Herbert Simon's theory of bounded rationality. Does the psychological paradigm give direction to the development of a particular economic trend? A great example of this is Austrian economics and the concept of the 'acting man' described by L. von Mises [2011].

There are few papers that I know that link the problem of risk and economic theory. From Polish authors, I would cite Prof. Mirosław Bochenek [2012] and Karol Klimczak [2008]. In selected economic theories or selected branches of economics, risk is obviously present and plays an important role. However, there is no category of risk incorporated into a coherent and complete theory of economics (I omit the question of a coherent and complete theory of economics as a whole but rather its parts that can perhaps be axiomatized). At this point it would be appropriate to present an overview of economic theories, those in history that formed the foundations of scientific thought and those of today. I will only refer or rather refer to the interesting overview in the book by Prof. Adam Glapiński *Meandry historii ekonomii. Między matematyką a poezją* [Glapiński, 2006] or *Paradoksy ekonomii. Rozmowy z polskimi ekonomistami* [Konat and Smuga, eds., 2016] to show the thinking of Polish economists about the history of economics and philosophizing about economics. It is still worth mentioning

here the name of an outstanding contemporary Polish scholar − Prof. Lukasz Hardt associated mainly with the philosophy of economics. In the aforementioned book in 2006, Prof. Glapiński wrote: "(...) the main and quantitatively speaking, the dominant part of the contemporary mainstream of theoretical economic thought is a priori deductive economics, establishing a pattern of economic science based on the methodological status characterizing not empirical sciences but mathematics and logics" [Glapiński, 2006, p. 11]. You are also familiar with the notable title of Roy Weintraub's book *How economics became a mathematical science* [Weintraub, 2002]. Tracking contemporary directions such as behavioral economics and experimental economics, one can conclude that Glapiński's view is not quite so true. Perhaps it is necessary to go back to the division proposed by John Neville Keynes, the father of John Maynard. Leaving aside the aspect of definitional chaos in economics, it is worth recalling that John Neville Keynes, in his work first edition in 1890 [Keynes, 1999], distinguished between descriptive economics, normative economics and applied economics. He characterized this division using the example of paying taxes. Descriptive economics examines, why entrepreneurs pay taxes in certain situations and what determines the amount of taxes. Normative economics analyses whether taxes should be paid, and if so, what amount of taxes is fair. In contrast, applied economics studies whether intervention in the process of paying taxes is desirable and, if it is, bring them closer to a fair value. Other scholars have also presented a similar division. Is there a place for risk in this classification. Neville Keynes says nothing about risk. In contemporary terms, when talking about risk, one must invoke Knight's name. But after all, John Maynard Keynes, at the same time as Frank H. Knight, presented a similar concept of risk as a measurable uncertainty. In his paper *A Treatise on Probability* [Knight, 1921] which is rarely cited by Keynesians, he wrote that risk is an immediate sacrifice necessary to be made in the hope of achieving a certain value. Risk has a measurable characteristic and can be insured which cannot be said of uncertainty. It should be added that John Maynard was an enthusiastic investor in the stock market and the art market. He did quite well and died as a wealthy man. It should be said that both Keynes and Knight gained from the achievements contained in the work of Allan H. Willett [1901]. It is interesting that Keynes and Willett are not cited by the authors of the respective entries in the Palgrave Economic Dictionary. An interesting observation was made by the authors analysing risk in terms of economics in *Handbook of Risk Theory* [Roeser et al., eds., 2012]. Perhaps Willett's paper is the first one where economics and insurance became formal theories to be solved in common problems. Until now, insurance has remained on the side-lines and has tended not to relate to economic theory. Interesting in this regard is the paper of a Polish scholar, Danielewicz and Dickstein [1910].

The fact that risk plays a central role in economic theory is accepted by all economists. There are obvious similarities between problems of economic risk and those of other risks, such as health and environmental risks. Sven Ove Hansson [2012, pp. 27-55] − the author of the chapter on the philosophy of risk − looks at economic risk from two points of view: aggregation and positive risk-taking.

The problem of aggregation concerns how to compare risks accruing to different individuals. Standard risk analysis is based on the principles of classical utilitarianism. All risks are aggregated into one and the same balance sheet, regardless of to whom they accrue. Thus, all risks are assumed to be fully comparable and summable. In the analysis of risk benefits and risk hazards, the benefits are added in the same way, and finally the sum of the benefits is compared with the sum of the risks to determine whether the total effect is positive or negative. In such a model, as in classical utilitarianism, individuals play no role merely as carriers of utility and non-utility, whose values are independent of by whom they are carried. The obvious alternative to this utilitarian approach is to treat each individual as a separate moral unit. Then the risks and benefits relating to one and the same individual can be weighted against each other, while the risks and benefits for different individuals are added or somehow 'aggregated' because they are considered incomparable. Such 'individualistic' risk weighting is very different from summation, which is standard in risk analysis. Individualistic risk weighting, however, is prevalent in medicine. It is used, for example, in the ethical evaluation of clinical trials in medicine. The two traditions of risk assessment differ in the same way as the 'old' and 'new' schools of welfare economics. In Arthur Pigou's so-called 'old' welfare economics, values relating to risk are added up to one grand total [Hansson, 2012, p. 49]. This approach is also used in the risk analysis of mainstream economics. This is justified in the field of classical insurance. The new school of welfare economics, which has dominated the mainstream since the 1930s, abandons the aggregation of individual values. Instead, it treats the welfare of different individuals as incomparable. This became the standard approach after Lionel Robbins showed how economic analysis can dispense with interpersonal comparability (Pareto optimality is the main tool needed to achieve this). The individualistic approach is widely used in special insurance or the new trend of risk insurance penetrating the characteristics of insured objects and policyholders (automobile insurance). Undoubtedly, an important role in insurance analysis is played by asymmetry of information. This will be mentioned later.

The risk-benefit analysis contains the implicit message that a rational person should accept exposing himself to risk if it brings greater benefits to others. Modern welfare economics (new school) values self-interested behavior to a much greater extent. The issue of positive risk-taking seems to be more or less specific to economic risk. Risk is by definition undesirable, and we expect a rational person to avoid it as much as possible. However, in economics, risk-taking is often considered desirable. Risk-taking by the capitalist is considered essential to the efficiency of the capitalist system, and justifies the owner's privilege to exercise ultimate control over enterprises and reap the benefits. As Adam Smith already stated in *The Wealth of Nations*: "something must be intended for the profits of the labor contractor who risks his stakes in this adventure" [Smith, 1976, p. 66].

That's probably why Prof. Krzysztof Jajuga's lecture on June 21, 2022 [Jajuga, 2022, p. 48] in Dąbrowa Górnicza at the awarding of a honorary doctorate from WSB University was entitled *Ryzyko – piękność czy bestia* [*Risk – Beauty or Beast*].

The risk, discussed by Smith, was a very large one, namely the risk of bankruptcy. According to Smith, bankruptcy is "perhaps the greatest and most humiliating misfortune that can befall an innocent man" [Smith, 1976, p. 2:741]. It was a risk that the capitalist was expected to take and for which he would be compensated. Its severity was crucial to Smith's argument, as can be seen from his negative attitude toward solutions that reduce the level from the risk of bankruptcy to the risk of losing invested capital. The most important of such solutions was the joint stock (limited liability) company.

However, since the time of Smith, capitalism has undergone a fundamental transformation. There have been two major reductions in capitalist thinking in risk-taking. The first occurred in the second half of the 19th century, when limited liability corporations became the dominant legal form of private enterprise in the industrialized parts of the world. Because of the massive spread of limited liability, personal risk-taking in most large industrial and financial ventures was reduced from bankruptcy to the loss of the original investment, which is exactly what Adam Smith warned against.

The second reduction in economic risk-taking occurred about 100 years later. From the late 20th century onward, private investment in companies was increasingly undertaken through institutions and funds that diversified their securities in sophisticated ways to reduce risk-taking. Portfolio theory and modern financial markets have made risk spreading much more efficient than previously possible. Today, an owner who employs reasonable risk diversification bears risks like those of the general economy.

Risk-sharing ownership has fundamentally changed the economic system, but its philosophical implications have not been much discussed. For example, it is not unreasonable to ask what effect this change has on the legitimacy of the owner's prerogative, which was previously based at least in part on the role of the risk-taking owner.

Understanding risk as an element of economic theory, incorporating it into that theory that many later referred to, was undoubtedly George A. Akerlof's paper *The Market for "lemons": Quality, Uncertainty and the Market Mechanism* [Akerlof, 1970]. Considering the market of cars (category: new and used, and category: good and bad ('lemon')) he concludes about the role of asymmetry of information and lack of information as the basis of the risk considered in the market of goods and the mechanisms governing this market. The random utility function plays here an important role. This thought was the basis for the extension of Bertrand's model to the duopoly market by the team of Prof. Bogumił Kamiński (known to me from the doctoral thesis of Mateusz Zawisza – defended in 2022). The introduction of customer decision rules based on a random utility function leads to Markov models of doupolist market behaviour. I confess that in applying these models (e.g., for research on brand choice) I was not fully aware of the immersion of the methodology used in the economic theory pioneered by Akerlof and others [Jensen and Meckling, 1976] in relation to companies − managers and shareholders in equity companies).

To conclude, it is worthwhile to think in terms of classical or neoclassical economics, which assumes a rational consumer and a rational producer, recall, so to speak, the opposition theory, which cannot be described by mathematics due to the 'total' negation of mathematics in this thinking. It is about the Austrian school of economics. The main thoughts are contained in the work of Ludwig von Mises *Human Action* [von Mises, 2011]. I will refer to this problem based on the published master's thesis of Mr. David Megger (written under my supervision) *Sprawiedliwość ekonomii dobrobytu. Libertarianizm i szkoła austriacka* [Megger, 2021]. Murray N. Rothbard accepted the existence of only demonstrated preferences (a different understanding from revealed preferences). The Austrian School only accepts the existence of time preferences; it even denies the existence of a consistent transitive preference relationship.

The most important concept is praxeology, that is, the action of a human being who wants to change a certain situation. Here arises the answer to the previously introduced probability of classes and probability of individual events. These categories are praxeological and not ontological (de Finetti's thought returns). Thus, in human action we are dealing with subjectivist probability, which does not lend itself to measurement. Uncertainty becomes the other side of the

axiom of action (the whole Austrian theory is built on the axiom of action or praxeology). We are dealing with the subjectivist belief of achieving a certain goal. Rather, it gives rise to the belief of choosing a path to achieve a goal not necessarily a set goal in the sense of maximum utility however understood. To achieve a goal is, as in Herbert Simon's, to get something better than one had so far. Perhaps this way of looking at risk is useful in the analysis of projects so fashionable in modern times (UMK's chief economist expressed in a private conversation that all the university's activities are in the nature of projects and so one must learn project management). Whether a project is to go to the store to buy milk and bread − I do not know, but this approach allows you to take one economics without dividing into micro-economics and macro-economics, as representatives of the Austrian school do. If I mention the Austrian school, it is not to discourage mathematics, statistics, observation of what happened. Rather, to intensify this effort without forgetting and the most important role of man, his choices of values and including goods in terms of economics.

References

Akerlof G.A. (1970), *The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*, "Quarterly Journal of Economics", Vol. 84(3), pp. 488-500.

Aven T. (2012), *The Risk Concept − Historical and Recent Development Trends*, "Reliability Engineering and System Safety", Vol. 99, pp. 33-44.

Aven T., Krohn B.S. (2014), *A New Perspective on How to Understand, Assess and Manage Risk and the Unforeseen*, "Reliability Engineering and System Safety", Vol. 121, pp. 1-10.

Aven T., Reniers G. (2013), *How to Define and Interpret a Probability in a Risk and Safety Setting*, "Safety Science", Vol. 51, pp. 223-231.

Bernstein P.L. (2017), *Przeciw bogom. Niezwykłe dzieje ryzyka*, Kurhaus Publishing, Warszawa.

Bochenek M. (2012), *Ryzyko i niepewność w naukach ekonomicznych – rozważania semantyczne*, „Ekonomia (Economics)", Vol. 4(21), pp. 46-63.

Chambers C.P., Echenique F. (2016), *Revealed Preference Theory*, Cambridge University Press.

Czerwiński Z. (1992), *Dylematy ekonomiczne*, PWE, Warszawa.

Danielewicz B., Dickstein S. (1910), *Zarys arytmetyki politycznej*, Wydawnictwo Szkoły Handlowej im. Leopolda Kronenberga, Warszawa.

Finetti B. de (2008), *Philosophical Lectures on Probability*, Springer Science+Business Media B.V.

Finetti B. de (1975), *Theory of Probability. A Critical Introductory Treatment*, John Wiley and Sons, New York.

Glapiński A. (2006), *Meandry historii ekonomii. Między matematyką a poezją*, Szkoła Główna Handlowa w Warszawie, Warszawa.

Górecka D. (2009), *Wielokryterialne wspomaganie wyboru projektów europejskich*, TNOiK, Toruń.

Grabowski M. (1998), *Istotne i nieistotne w nauce*, Wydawnictwo Rolewski, Toruń.

Hansson S.O. (2012), *A Panorama of the Philosophy of Risk* [in:] S. Roeser, R. Hillerbrand, P. Sandin, M. Peterson (eds.), *Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, Springer, Dordrecht.

Jajuga K. (2022), *Krzysztof Jajuga Doktor Honoris Causa Akademii WSB*, Wydawnictwo Naukowe WSB, Dąbrowa Górnicza.

Jensen M.C., Meckling W.H. (1976), *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, "Journal of Financial Economics", October, Vol. 3, No. 4, pp. 305-360.

Keynes J.M. (1921), *A Treatise on Probability*, Macmillan and Co, London.

Keynes J.N. (1999), *The Scope and Method of Political Economy*, Batoche Books, Kitchener.

Klimczak K.M. (2008), *Ryzyko teorii ekonomii*, „Central European Management Journal", Vol. 16(6), pp. 64-68.

Knight F.H. (1921), *Risk, Uncertainty and Profit*, Houghton Mifflin Company, The University Press Cambridge, New York.

Konarzewska-Gubała E. (1991), *Wspomaganie decyzji wielokryterialnych systemem „BIPOLAR"*, Seria: „Monografie i Opracowania", nr 76, Wydawnictwo Uczelniane Akademii Ekonomicznej we Wrocławiu, Wrocław.

Konat G., Smuga T., red. (2016), *Paradoksy ekonomii. Rozmowy z polskimi ekonomistami*, Wydawnictwo Naukowe PWN, Warszawa.

Megger D. (2021), *Sprawiedliwość ekonomii dobrobytu. Libertarianizm i szkoła austriacka*, UMK, Toruń.

Mises L. von (2011), *Ludzkie działanie. Traktat o ekonomii* [*Human Action. A Treatise on Economics*], Instytut Ludwiga von Misesa, Warszawa.

Paulos J.A. (2012), *Innumeracy. Matematyczna ignorancja i jej konsekwencje w dobie nowoczesnej technologii*, CeDeWu, Warszawa.

Ramsey F.P. (1926), *Truth and Probability, & "Further Considerations", 1928 and "Probability and Partial Belief", 1929* [in:] F.P. Ramsey (1931), *The Foundations of Mathematics and Other Logical Essays*, Google Books, Routledge and Kegan Paul Ltd.

Roeser S., Hillerbrand R., Sandin P., Peterson M., eds. (2012), *Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, Springer, Dordrecht.

Samuelson P.A. (1938), *A Note on the Pure Theory of Consumer's Behaviour*, "Economica", Vol. 5, pp. 61-71.

Simons K. (2019), *Paradoksy prawdopodobieństwa*, Wydawnictwo Naukowe PWN, Warszawa.

Smith A. ([1776] 1976), *An Inquiry into Nature and Causes of the Wealth of Nations* [in:] R.H. Campbell, A.S. Skinner, W.B. Todd (eds.), *The Glasgow Edition of the Works and Correspondence of Adam Smith*, Vol. 2, Clarendon, Oxford.

Szreder M. (2004), *Od klasycznej do częstościowej i personalistycznej interpretacji prawdopodobieństwa*, „Wiadomości Statystyczne", nr 519, s. 1-10.

Trzaskalik T., red. (2014), *Wielokryterialne wspomaganie decyzji. Metody i zastosowania*, PWE, Warszawa.

Trzaskalik T. (2014), *Wielokryterialne wspomaganie decyzji. Przegląd metod i zastosowań*, Zeszyty Naukowe Politechniki Śląskiej, Seria: Organizacja i Zarządzanie, z. 74, pp. 239-263.

Trzpiot G. (1999), *Wielowartościowe zmienne losowe w badaniach ekonomicznych*, Akademia Ekonomiczna, Katowice.

Tyszka T. (2010), *Decyzje. Perspektywa psychologiczna i ekonomiczna*, Wydawnictwo Naukowe Scholar, Warszawa.

Weintraub E.R. (2002), *How Economics Became a Mathematical Science*, Duke University Press, London.

Willett A.H. (1901), *The Economic Theory of Risk and Insurance (Reprint)*, Richard D. Irwin, Inc., Homewood, Illinois, Copyright, 1951.

Wójtowicz M. (2016), *Zakład Pascala – argumentacja i działanie*, Wydawnictwo Uniwersytetu Śląskiego, Katowice.

Chapter II

Extended Gini regression coefficient as robust estimator of systematic risk in capital market

Grażyna Trzpiot

1. Introduction

One of the basic market models is the Sharp model, which is used to position equity investments. The use of linear regression is a classic approach used in modelling, replacing this approach with a Gini regression model is subject of this chapter. Outliers and extreme values, which we observe in the distribution of rates of return on listed assets, were the motivation to use the Gini regression model with EGRC (extended Gini regression coefficient) to reflect the investigator's perception of risk aversion in the market.

The main aim of this work is to look close for the equivalent parameters to the covariance and correlations that are required for the decomposition of a sum of random variables to systematic and specific investment risk. It opens for estimating a robust regression based on those measures and also opens a discussion how the coronavirus has exposed three flawed assumptions of modern portfolio theory. Indication of the advantages of such an approach and verification of suitability for market data is the main goal. The application part is modelling data from the Warsaw Stock Exchange[1].

Gini's mean difference (GMD) was first introduced by Corrado Gini in 1912 as an alternative measure of variability. GMD and the parameters which are derived from it (such as the Gini coefficient, also referred to as the concentration ratio) have been in use in the area of income distribution for almost a century. In other areas it seems to make sporadic appearances and to be 'rediscovered' again and again under different names. It turns out that GMD has at least more than one different alternative representations. Each representation can be given its own interpretation and naturally leads to a different analytical tool such as L_1 metric, order statistics theory, extreme value theory, concentration

[1] Key words: Gini regression model, systematic risk, portfolio analysis.

curves, and more. Some of the representations hold only for nonnegative variables while others need adjustments for handling discrete distributions. On top of that, the GMD was developed in different areas and in different languages. Corrado Gini himself mentioned this difficulty [Gini, 1921].

2. Beta as a measures of systematic risk

Systematic risk is the volatility that affects the entire stock market across many industries, stocks, and asset classes. Systematic risk affects the overall market and is therefore difficult to predict and hedge against. Unlike with unsystematic risk, diversification cannot help to smooth systematic risk, because it affects a wide range of assets and securities. Investors can still try to minimize the level of exposure to systematic risk by looking at stock's beta, or its correlation of price movements to the broader market as a whole. Here, we take a closer look at how beta relates to systematic risk.

Beta is a measure of a stock's volatility in relation to the market. It essentially measures the relative risk exposure of holding a particular stock or sector in relation to the market. If you want to know the systematic risk of your portfolio, you can calculate its beta [Sharpe, 1977]. Beta effectively describes the activity of a security's returns as it responds to swings in the market. A security's beta is computed by dividing the product of the covariance of the security's returns (R_k) and the market's returns (R_M) by the variance of the market's returns over a specified period, using this formula:

$$\beta_{MNK} = \frac{cov(R_k, R_M)}{cov(R_M, R_M)} \tag{1}$$

Covariance explains how changes in a stock's returns are related to changes in the market's returns. Variance explains how far the market's data points spread out from their average value. We can calculate beta by running a linear regression on a stock's returns compared to the market using the capital asset pricing model (CAPM) [Sharpe, 1964]. But OLS requires:

- linear relationship between conditional expectation of the dependent variable and explanatory variables,
- errors are identically distributed and uncorrelated with the independent variables.

Often monotonic transformations are applied to linearize the model, can lead to changes of the sign of the estimated coefficients. And OLS is sensitive to outliers. The variance is the most popular measure of variability. There are two

properties which are implicit when dealing with the variance − the symmetry and the decomposition:

1. Symmetric relationship: There are two kinds of symmetric relationships that are imposed in the conventional statistical analysis in general. The first one is the symmetry of the variability measure with respect to the underlying distribution and the second one is the symmetry in the relationship between variables.
2. Decompositions: There are two types of decompositions. One is the decomposition of a variability measure of a linear combination of random variables into the contributions of the individual variables and the contributions of the relationships between them. The other decomposition is the one that decomposes the variability of a population that is composed of several subpopulations into the contributions of the subpopulations and some extra terms.

The Gini approach deviates from this conventional approach in both cases symmetric relationship and the decomposition of the GMD (Gini's Mean Difference) includes the structure of the decomposition of the variance as a special case. The usefulness of the GMD and its contribution to our statistical analysis is especially important whenever the concepts that are used are not symmetric by definition. Among those concepts are regression in statistics and elasticity in economics. The Gini describes the variability by two attributes: the variate and its rank [Schröder and Yitzhaki, 2016].

3. Gini regression

Idea: replace the (co-)variance in an ordinary least squares (OLS) regression with the Gini notion of (co-)variance, i.e., the Gini's Mean Difference (GMD) as the measure of dispersion. Gini Mean Difference is defined as: $G_{YX} = E|Y - X|$ with Gini covariance: ${}_{G}cov(Y; X) = cov(Y; F(X))$, where $F(X)$ is the cumulative population distribution function. GMD was first introduced by Corrado Gini in 1912 as an alternative measure of variability.

Gini regressor is defined, in general case as:

$$\beta^{G} = \frac{cov(Y, F(X)))}{cov(X, F(X)))} \tag{2}$$

can be interpreted as an IV regression, with $F(X)$ as an instrument for X (using the instrumental variable approach).

There are several regression methods that compete with the ordinary least squares (OLS), among them:

- Mean Absolute Deviation (MAD) regression,
- Least Absolute Deviation (LAD) regression,

- Quantile regression (QR − the absolute deviation from a quantile of the residuals),
- Maximum Likelihood (ML) regression.

MAD, LAD and QR are based on variability measures like the GMD, to the L_1 metric (city bloc) and as such, we should expect them to have properties that are identical [Choi, 2009].

Gini regression do not depend on symmetric correlation and variability measure, linearity of the model and coefficients do not change after monotonic transformations of the explanatory or independent variables.

GMD (Gini Mean Difference) here definition has two asymmetric correlation coefficients, one can be used for the regression, the other can be used to test the linearity assumption [Yitzhaki and Schechtman, 2013; Yitzhaki, 2015]. Two regression methods can be interpreted as based on Gini's Mean Difference (GMD). First relies on a weighted average of slopes defined between adjacent observations (a semi-parametric approach). Second is based on minimization of the GMD of the residuals.

The semi-parametric approach is based on estimating a regression coefficient that is a weighted average of slopes defined between adjacent observations (or all pairs of observations) of the regression curve. It resembles the OLS in the sense that the estimators can be explicitly presented, and all the expressions used have parallels in OLS regression. The derivation of the estimators and their properties are discussed in detail in Schechtman, Yitzhaki and Artsev [2005]. The semi-parametric regression does not require specification of the model. Unlike the minimization approach, there is no problem of non-uniqueness of the estimated regression coefficient. The point estimators of the semi-parametric approach can be calculated easily using the instrumental variable approach therefore standard regression software can be used.

The minimization approach is based on minimization of the GMD of the residuals. This approach requires the assumption of a linear model. It is similar to Least Absolute Deviation (LAD) regressions [Trzpiot, 2008; Trzpiot, 2019]. Instead of minimizing the sum of absolute deviations of the residuals, the GMD of the residuals which is the mean of the absolute differences between all pairs of residuals is minimized. Similar to the case in LAD, the estimators can be derived numerically but there are no explicit expressions for them.

3.1. Gini's multiple regressions

Let $(Y, X_1, \ldots, X_K)$ be a $(K + 1)$-variate random variable with expected values $(\mu_Y, \mu_1, \ldots, \mu_K)$, respectively and a finite variance-covariance matrix Σ.

Assume that we have a general regression curve defined by:

$$g(x_1, \ldots, x_K) = E\{Y|X_1 = x_1, \ldots, X_K = x_K\} \tag{3}$$

The resulting vector of regression coefficients of step 2, β_N, is given by:

$$\beta_N = [E(V^T X)]^{-1} E(V^T Y) \tag{4}$$

where: $\beta_N = \{\beta_{N1}, \ldots, \beta_{Nk}\}$ is a $(K \times 1)$ column vector of the (conditional) regression coefficients, V is an $(n \times K)$ matrix of the cumulative distributions of $X_1, \ldots, X_K$, Y is an $(n \times 1)$ vector of the dependent variable and X is an $(n \times K)$ matrix of the deviations of the explanatory variables from their expected values. The elements of $E(V^T Y)$ and $E(V^T X)$ are $\text{cov}(Y, F_k(X_k))$ and $\text{cov}(X_j, F_k(X_k))$, respectively. It is assumed that the rank of $V^T X$ equals K, the number of explanatory variables. Next the constant term can be estimated by minimizing a function of the residuals. The exact function used determines whether the regression passes through the mean, the median, or any other quantile. The multiple regression procedure, although it is not based on an optimization procedure, generates equivalents to the OLS's normal equations. By defining the error term and substituting for the multiple regression coefficients, it can be shown that:

$$\text{COV}(\varepsilon, F_k(X)) = 0 \text{ for } k = 1, \ldots, K \tag{5}$$

3.2. Properties Gini regressions

The Gini semi-parametric approach has the advantage of relying on a few assumptions, no linearity hypothesis is needed [Olkin and Yitzhaki, 1992]. The estimator β_N is less sensitive to extreme values since it is built on the matrices $V'X$.

Among those concepts is R^2 of the regression, which can be considered as a measure to assess the share of the (square of the) GMD which is explained by the model [Trzpiot, 2021]:

$$_G R^2 = 1 - [\text{cov}(e, r(e))/\text{cov}(y, r(y))]^2 \tag{6}$$

$r(x)$ denotes ranks in the sample, where $e = y - x\beta_N$.

3.3. Extended Gini index and extended Gini regression coefficients

The extended Gini variability index is a member of a family of indices defined by:

$$E\_GINI(X, v) = -(v + 1)\mathrm{COV}(X, [1 - F(X)]^{v}) \tag{7}$$

where:
$v > -1$, $v \neq 0$

The role of v in the extended Gini variability index is to reflect the investigator's attitude toward variability. The higher v is, the more stress is put on the lower portion of the distribution of the independent variable. In the extreme case ($v \to \infty$), the investigator cares only about the lowest part of the cumulative distribution, as if he is guided by the max-min criterion. By using the Gini extended index definition, we can write the coefficient in the new extended version:

$$\beta_{E\_GINI}(v) = \frac{-(v + 1)cov(Y, 1 - F_X(X)]^{v})}{-(v + 1)cov(X, 1 - F_X(X)]^{v})} \tag{8}$$

By changing v and reestimating the model, the investigator can learn about the curvature of the regression curve. The higher v is, the higher is the weight that is given to the slopes of the regression curve at the lower end of the range of the independent variable. If $\beta_{E\_GINI}(v)$ turns out to be a declining (increasing) function of v, then the regression curve is convex (concave) [Schechtman, Yitzhaki and Artzev, 2005].

4. Assessment of risk in empirical portfolio

We have selected companies from the stock market index WIG-INFO. The 15 selected companies were evaluated daily closing prices from 17.02.2020 to 15.02.2022. As of 20.03.2022, the index comprises 23 companies. The highest share in the sectoral portfolio is held by Asseco Poland SA (39.55%), the second highest share is held by LiveChat Software SA (13.24%) and the third highest share is held by Comarch SA (8.64%). The composition of the index WIG-INFO is recorded in Table 1.

Table 1. Companies selected for analysis − share in the WIG-Informatics index

Company name	Index share	Company name	Index share	Company name	Index share
1. AILLERON	0.72%	6. COMP	1.92%	11. LIVECHAT	13.24%
2. ASSECOPOL	39.5%	7. DATAWALK	6.06%	12. NTTSYSTEM	0.20%
3. ATENDE	0.96%	8. ELZAB	0.10%	13. SIGNITY	1.08%
4. BRAND24	0.17%	9. IFIRMA	0.74%	14. TALEX	0.12%
5. COMARCH	8.64%	10. K2HOLDING	0.21%	15. WASKO	0.26%

The classic graphical representation of results for empirical data is the rate of return – risk, usually expressed by the value of the standard deviation S (Fig. 1). Using the chart below as a measure of risk, the following can be used respectively semivarince or mean absolute semideviation of the rate of return.

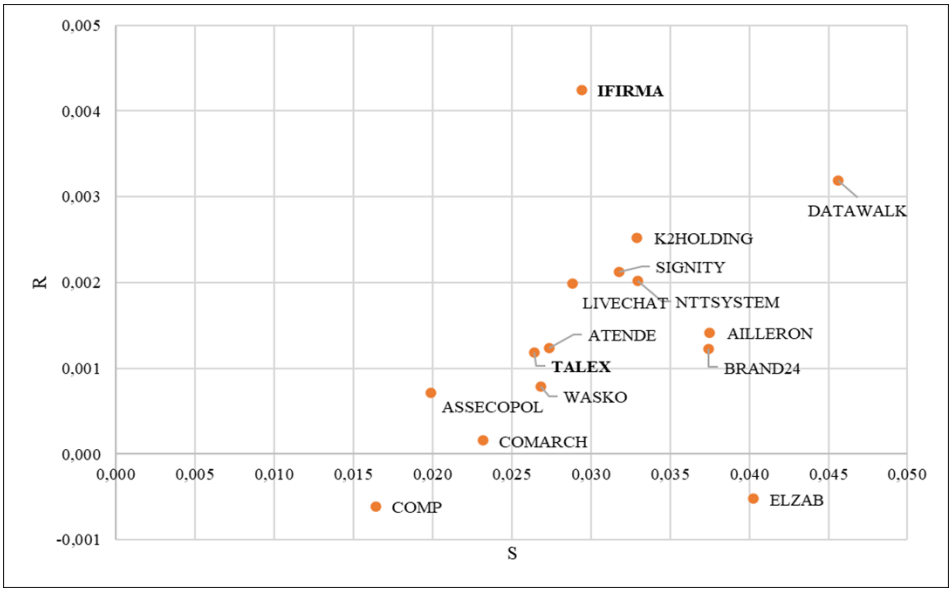


Fig. 1. Risk-return map for 15 selected companies

The analysis begins with the determination of descriptive parameters and selected risk measures (Table 2). We can notice than for the observation of the distribution of losses (negative risk concept) the best interpretation we can obtained using semivariance and semi-slope.

Table 2. Values of parameters and selected risk measures of the companies surveyed

Assets	R	V	S	SV	SS	d	Sd
AILLERON	0.0014	0.0014	0.0375	0.0006	0.0246	0.0251	0.0126
ASSECOPOL	0.0007	0.0004	**0.0199**	0.0002	**0.0134**	0.0144	0.0072
ATENDE	0.0012	0.0007	0.0273	0.0003	0.0183	0.0185	0.0092
BRAND24	0.0012	0.0014	0.0374	0.0005	0.0234	0.0267	0.0134
COMARCH	0.0002	0.0005	**0.0232**	0.0002	0.0157	0.0164	0.0082
COMP	**−0.0006**	0.0003	**0.0164**	0.0001	**0.0114**	**0.0108**	**0.0054**
DATAWALK	0.0032	0.0021	**0.0456**	0.0008	0.0284	**0.0309**	**0.0155**
ELZAB	**−0.0005**	0.0016	**0.0402**	0.0008	0.0277	0.0251	0.0125
IFIRMA	0.0042	0.0009	0.0294	0.0004	0.0198	0.0211	0.0105
K2HOLDING	0.0025	0.0011	0.0329	0.0004	0.0212	0.0232	0.0116
LIVECHAT	0.0020	0.0008	0.0288	0.0003	0.0185	0.0205	0.0102
NTTSYSTEM	0.0020	0.0011	0.0330	0.0005	0.0213	0.0226	0.0113
SIGNITY	0.0021	0.0010	0.0318	0.0004	0.0203	0.0208	0.0104
TALEX	0.0012	0.0007	0.0264	0.0003	0.0178	0.0178	0.0089
WASKO	0.0008	0.0007	0.0269	0.0003	0.0185	0.0166	0.0083

The classical approach to portfolio modelling requires an assumption about the type of return distribution: it should be a normal distribution. We performed two statistical tests: Kolmogorov-Smirnov test and Shapiro-Wilk test.

Table 3. Compliance tests for rate of return distributions the companies surveyed

Assets	Kolmogorov-Smirnov test		Shapiro-Wilk test	
	test value	*p*-value	test value	*p*-value
WIG-INFO	0.066	<0.001	0.952	<0.001
AILLERON	0.104	<0.001	0.913	<0.001
ASSECOPOL	0.069	<0.001	0.967	<0.001
ATENDE	0.094	<0.001	0.929	<0.001
BRAND24	0.095	<0.001	0.931	<0.001
COMARCH	0.087	<0.001	0.958	<0.001
COMP	0.127	<0.001	0.923	<0.001
DATAWALK	0.123	<0.001	0.908	<0.001
ELZAB	0.128	<0.001	0.886	<0.001
IFIRMA	0.078	<0.001	0.956	<0.001
K2HOLDING	0.086	<0.001	0.946	<0.001
LIVECHAT	0.095	<0.001	0.942	<0.001
NTTSYSTEM	0.092	<0.001	0.924	<0.001
SIGNITY	0.118	<0.001	0.880	<0.001
TALEX	0.126	<0.001	0.941	<0.001
WASKO	0.130	<0.001	0.865	<0.001

H_0: Rates of return have a normal distribution.

H_1: Rates of return do not have a normal distribution.

Table 3 summarizes the tests carried out for the conformity of the distribution to the normal distribution. Two tests were performed: Kolmogorov-Smirnov and Shapiro-Wilk, for which the *i*-th hypothesis H_0 is H_0: The returns of the *i*-th company have a normal distribution, while hypothesis H_1 is a simple negation of hypothesis H_0. For both the Kolmogorov-Smirnov and Shapiro-Wilk tests for each company and industry index, the *p*-value is less than the accepted significance level of 0.05. This means that H_0 should be rejected. With a probability of 0.95, it can be argued that the returns of the analyzed companies do not have a normal distribution. Conclusion: rates of return do not have a normal distribution.

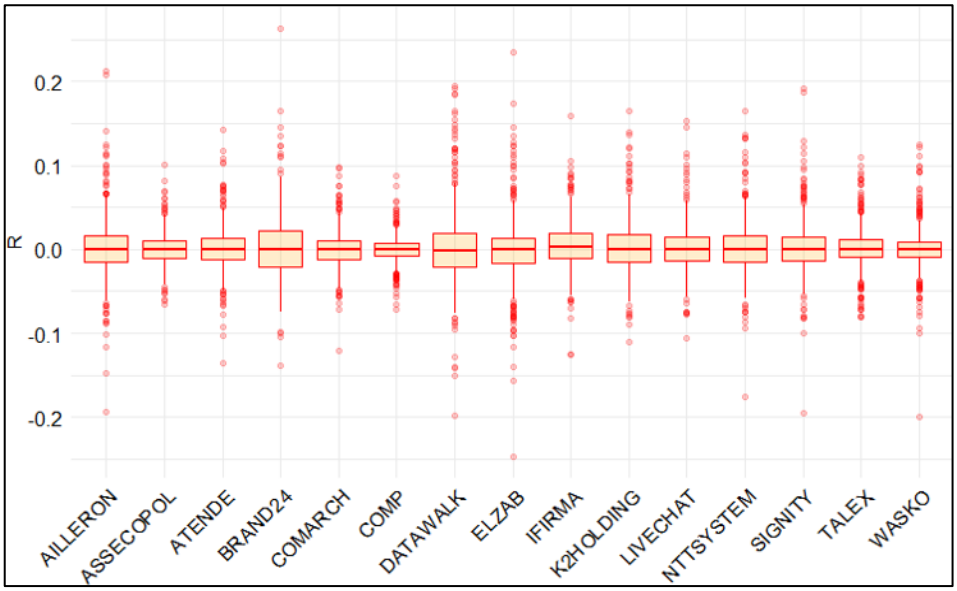


Fig. 2. Examination of rate of return distributions

Before examining whether the distributions of stock returns of the companies selected for analysis are characterised by a normal distribution, box-plots were drawn for the distributions of returns for each company (Fig. 2). Analysing this graph, it can be concluded that outliers were observed for all companies.

 # Portfolio I

The aim of this stage of analysis was to determine a multicomponent portfolio. It was specified that this portfolio was to consist of 5-7 shares of companies out of the 15 taken for analysis and no short selling was allowed. An attempt was first made to construct a portfolio with minimum risk (S_P) at a specified required rate of return $R_p \geq R_0$. As R_0 was fix the average rate of return (with fix equal wages).

Table 4. Shares of companies in the portfolio I

Asset	w_i
ASSECOPOL	**0.146**
ATENDE	**0.060**
BRAND24	**0.039**
COMARCH	**0.051**
COMP	**0.149**
DATAWALK	**0.007**
IFIRMA	**0.150**
K2HOLDING	**0.052**
LIVECHAT	**0.059**
NTTSYSTEM	**0.044**
SIGNITY	**0.050**
TALEX	**0.112**
WASKO	**0.081**

The results of the calculations are included in Table 4. Due to the fact that the obtained optimal portfolio consists of almost all companies, an attempt was made to impose additional limiting conditions, but it was not possible to obtain a smaller portfolio. We receive portfolio with positive rate of return R_p = 0.0014 and S_p = 0,0107.

 # Portfolio II

Therefore, a second smaller portfolio was constructed, this time maximizing the expected return on the portfolio at a given level of risk $S_p \leq S_0$. As S_0 was fix the average risk level of rate of return (with fix equal wages). In this model, by applying an additional constraint imposed on the size of the individual holdings ($w_i \leq 0.15$). The results of the calculations are included in Table 5. This portfolio has an expected return of 0.26% with a risk of 1.73%.

Table 5. Shares of companies in the portfolio II

Asset	w_i
AILLERON	**0.10**
DATAWALK	**0.15**
IFIRMA	**0.15**
K2HOLDING	**0.15**
LIVECHAT	**0.15**
NTTSYSTEM	**0.15**
SIGNITY	**0.15**

We can assess the efficiency of the portfolios we build against the market. For this assessment we choose two known measures, Sharp ratio and Treynor ratio defined as:

Sharpe ratio $$WS_M = \frac{R_M - R_F}{\sigma_M} \quad (9)$$

Treynor ratio $$WT_M = \frac{R_M - R_F}{\beta_M} \quad (10)$$

where:
R_F – risk-free rate of return,
R_M – expected rate of return of the market portfolio,
σ_M – standard deviation of the rates of return of the market portfolio,
β_M – beta coefficient of the market portfolio.

Figure 3 shows the determined magnitudes for the Sharpe ratio both for the two portfolios compared and for the market as a whole. Portfolio II has a (slightly) higher value of the measure and is therefore more efficient than Portfolio I. It should be noted, however, that the Sharpe ratio for the market is lower than the value of the measure for both portfolios, indicating that it is worth investing in both portfolios.

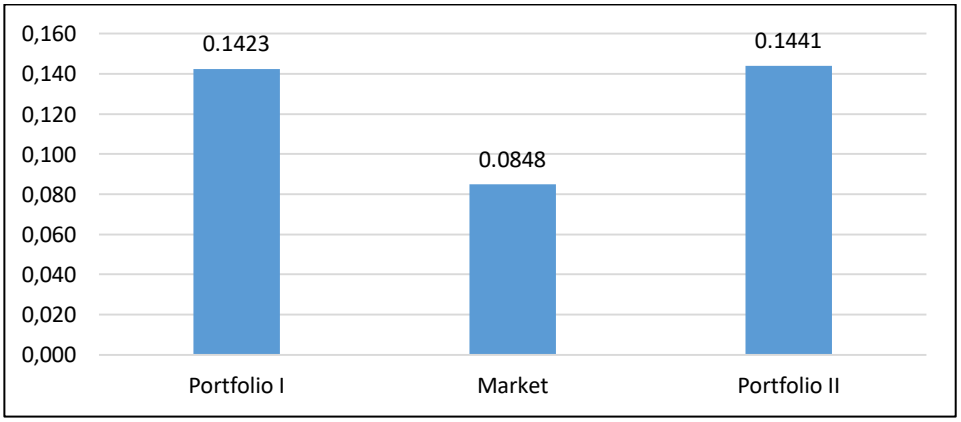


Fig. 3. Results of Sharp ratio

An analogous situation is taking place for the Treynor measure (Fig. 4). Portfolio II is more efficient than Portfolio I, but by comparing the values of the portfolios' measures with the measure determined for the market, it can be concluded that investing in both portfolios is a rational choice.

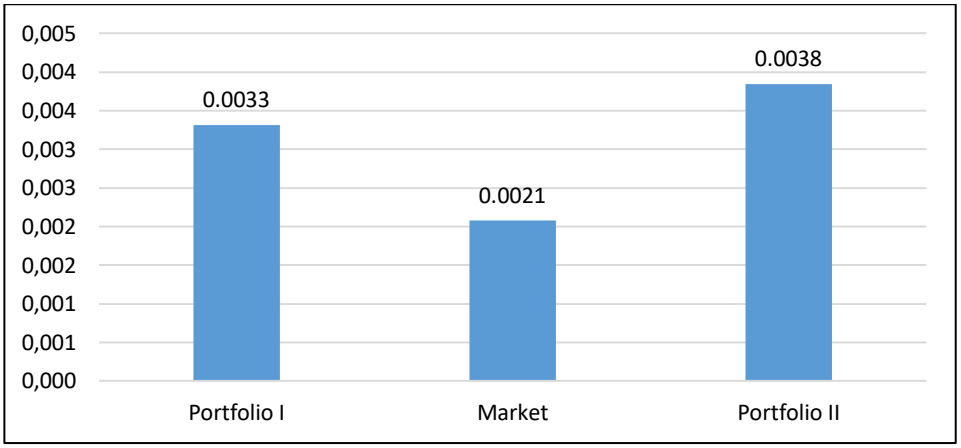


Fig. 4. Results of Treynor ratio

For searching of a portfolio with more profitable parameters for the investor the density function of the distributions of returns for each company were drawn (Fig. 5). An even more detailed characterisation of the distribution of returns can be made on the basis of the analysis of this graph − the company DATAWALK stands out with the most flattened distribution. The highest concentration of observations around the average value can be seen for COMP. The most flattened distribution DATAWALK stands out. Return rate distributions of the companies IFIRMA and LIVECHAT are characterised by right-handed asymmetry. It should also be noted that the returns of IFIRMA and LIVECHAT are characterised by right asymmetry, which means that most of the returns were lower than the average return. The highest concentration of observations around the average value can be seen for COMP.

In addition, the third and fourth central moments standardized (Table 6) were determined in order to be able to more precisely determine the type and strength of asymmetry and concentration of return distributions. Only in the case of WASKO was left-hand asymmetry (weak) observed, while for the remaining companies, right-hand asymmetry is noted − the returns of BRAND24 are characterized by particularly strong right-hand asymmetry. This shows that for 14 of the 15 companies analyzed, the majority of returns are lower than the expected return. Analyzing the indications of the fourth standardized moment, it can be concluded that the rates of return of all companies are characterized by a slender distribution, which means that a significant proportion of observations are similar to each other, and observations that differ significantly from each other are few in number.

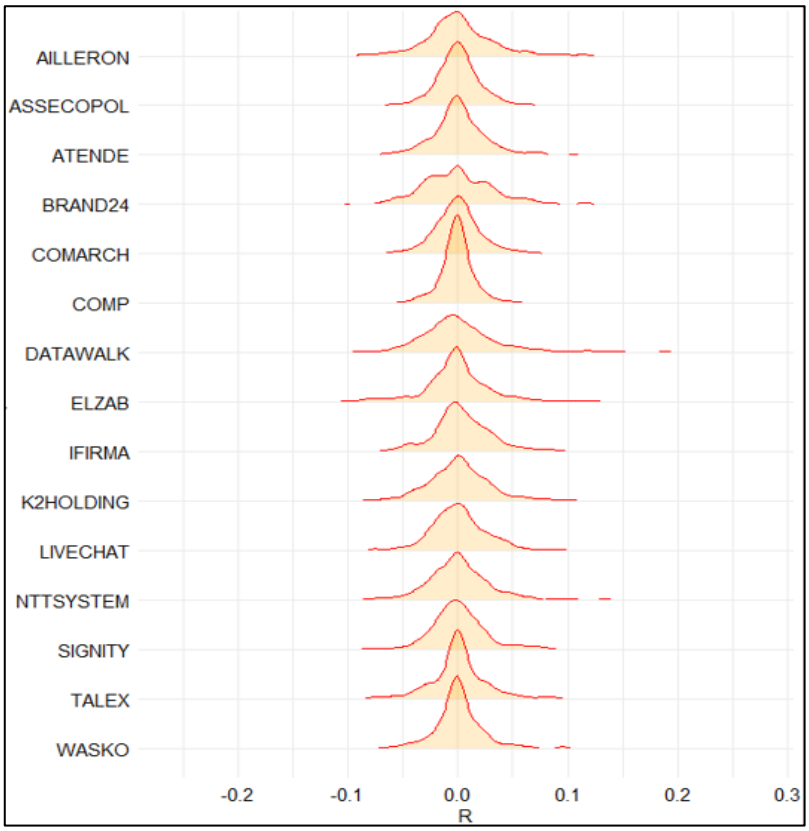


Fig. 5. Density functions of the return distributions

Table 6. Measures of asymmetry and concentration of the companies surveyed

Assets	Measure of asymmetry	Measure of concentration
	standardized third central moment	standardized fourth central moment
AILLERON	0.652	9.013
ASSECOPOL	0.477	5.455
ATENDE	0.434	7.506
BRAND24	1.194	9.111
COMARCH	0.275	6.088
COMP	0.275	7.299
DATAWALK	0.856	6.891
ELZAB	0.204	9.952
IFIRMA	0.314	6.173
K2HOLDING	0.822	6.104
LIVECHAT	0.854	6.848
NTTSYSTEM	0.658	7.702
SIGNITY	0.828	11.463
TALEX	0.397	5.305
WASKO	−0.136	12.574

 # Portfolio III

In the classical Markowitz model, the distributions of returns are assumed to be variables with normal distributions, i.e. symmetric distributions, and two basic characteristics − the expected value and the variance − are sufficient to select the optimal stock portfolio. However, the assumptions made in the Markowitz model regarding the normality of returns have no practical justification, so the model does not correspond to reality. Many modifications and extensions of the Markowitz model are proposed in the literature. One line of research is to modify the model by extending the criteria to include an asymmetry factor (skewness). Investors, preferring a higher probability of large returns and smaller possible losses, prefer positive skewness of the distribution of random returns. Positive skewness of the distribution refers to the right tail of the density function and is objectively desirable as it entails a lower probability of negative returns [Kopańska-Bródka, 2014]. The most commonly used measure of skewness in the selection of the optimal portfolio is the third central moment (normal or standardized) − in this study, the standardized third central moment was used and the values of the previously determined asymmetry coefficients are presented (Table 7).

Table 7. Shares of companies in the optimal portfolio − with the asymmetry factor

Asset	w_i
BRAND24	0.792
DATAWALK	0.045
LIVECHAT	0.104
SIGNITY	0.059

Firstly, the portfolio maximizing the expected rate of return was determined (as in portfolio I) with an additional constraining condition taking into account the asymmetry coefficient.

Secondly, maximizing the expected return on the portfolio at a given level of risk $S_p \leq S_0$ (as in portfolio II).

At the last step, an attempt was made to construct a portfolio that maximizes the portfolio asymmetry ratio $A_p \geq A_0$. As A_0 was fix the average level of asymmetry measures of rate of return (with fix equal wages). The portfolio constructed in this way is characterized by very high right-handed asymmetry.

The portfolio was composed of 4 companies (DATAWALK, IFIRMA, K2HOLDING and SIGNITY) and the portfolio's skewness coefficient was 0.6802. It should be noted that both the expected return and the standard deviation of the portfolio are higher than was the case for the classic Markowitz portfolio, so the portfolio including the skewness aspect is not dominant. We receive portfolio with positive rate of return R_p = 0,0032 and S_p = 0.0216 and A_p = 0.6802.

5. Robust estimation of systematic risk

Classic modelling has highlighted an important aspect that improves portfolio results-taking into account the fact of observing outliers in rate of return distribution. So in the last part of the analysis, the aim of the study is to compare systematic risk estimates using the approaches previously discussed but notice according to the market: R_k – return of k-th assets and R_M – market return:

1. Classical Sharpe coefficient:

$$B_{OLS} = \frac{cov(R_k, R_M)}{cov(R_M, R_M)} \tag{11}$$

2. Gini Regression coefficient:

$$\beta_{GINI} = \frac{cov(R_k, F_M)}{cov(R_M, F_M)} \tag{12}$$

3. Extended Gini Regression coefficient:

$$\beta_{E\_GINI}(v) = \frac{-(v+1)cov(R_k, [1-F_M]^v)}{-(v+1)cov(R_M, [1-F_M]^v)} \tag{13}$$

for two v value $v = 1{,}01$ and $v = 1{,}05$ (according to appropriate investor's utility function), for evaluation all results a multivariate Gini regression was used. The calculations were carried out for the whole set of companies selected for the analysis, and the results are presented in the following table.

Table 8. Systematic risks measures of the companies surveyed

Assets	B_{OLS}	R^2	β_{GINI}	$\beta_{E\_GINI}(1.01)$	$\beta_{E\_GINI}(1.05)$
AILLERON	0.690	0.061	−2.823	−2.825	−2.834
ASSECOPOL	1.184	0.636	0.158	0.158	0.157
ATENDE	0.379	0.034	0.297	0.296	0.294
BRAND24	0.484	0.034	**4.936**	**4.937**	**4.945**
COMARCH	0.777	0.202	0.542	0.542	0.541
COMP	0.243	0.039	−1.022	−1.022	−1.025
DATAWALK	1.588	0.218	**3.765**	**3.766**	**3.768**
ELZAB	0.760	0.064	2.751	2.752	2.757
IFIRMA	0.501	0.052	−1.178	−1.179	−1.181
K2HOLDING	0.415	0.029	**3.070**	**3.071**	**3.073**
LIVECHAT	1.148	0.285	**3.903**	**3.904**	**3.906**
NTTSYSTEM	0.275	0.012	0.983	0.982	0.982
SIGNITY	0.585	0.061	1.443	1.442	1.441
TALEX	0.139	0.005	−1.590	−1.591	−1.593
WASKO	0.239	0.014	−3.016	−3.017	−3.022

The results obtained are consistent with those obtained previously for Portfolio III. They indicate the companies that will allow the best result for the portfolio. A robust modelling approach is an equivalent tool in portfolio modelling. We can use it instead of optimization models.

The classical beta coefficient (β_{OLS}) determines the degree of sensitivity of a stock to changes in the return of a stock index when the reference point is the average change in the change in the return of the stock index.

The directional Gini regression coefficient (β_{GINI}) determines the degree of sensitivity of a given stock to changes in the stock index return rate when the reference point is the median sweep of the empirical distribution of the stock index return rate.

The generalized directional coefficient of the Gini regression ($\beta_{E\_GINI(v)}$) determines the degree of sensitivity of a given stock to changes in the stock market index return rate, taking into account additionally the investor's risk aversion.

6. Concluding remarks

Systematic risk is the covariance between the marginal utility of capital in the analyzed portfolio and the stock return. As shown, these concepts can be measured using the assumption of linearity − we then determine the moments of the random variable − as well as departing from these assumptions. The choice of method for determining the regression is also a choice of the marginal utility of capital function as well as risk aversion.

The used Gini regression methodology in the classical and extended version was compared with classical approaches in portfolio modelling. For the analysis dataset with outlier observations, the measurement of systematic risk is more efficient using this robust approach.

References

Choi S.W. (2009), *The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment*, "Quarterly Journal of Political Science", Vol. 4, pp. 153-165.

Gini C. (1921), *Measurement of Inequality of Incomes*, "The Economic Journal", Vol. 31, Iss. 121, March, pp. 124-125.

Gini C. (1912), *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*, C. Cuppini, Bologna.

Kopańska-Bródka D. (2014), *Optymalny portfel inwestycyjny z kryterium maksymalnej skośności*, „Studia Ekonomiczne", nr 208, s. 46-58, Uniwersytet Ekonomiczny w Katowicach.

Olkin I., Yitzhaki S. (1992), *Gini Regression Analysis*, "International Statistical Review", Vol. 602, pp. 185-196.

Schechtman E., Yitzhaki S., Artsev Y. (2005), *Who Does Not Respond in the Household Expenditure Survey: An Exercise in Extended Gini Regressions*, mimeo, http:// ssrn.com.

Schröder C., Yitzhaki S. (2016), *Reasonable Sample Sizes for Convergence to Normality*, Communications in Statistics − Simulation and Computation, 0918, pp. 1-14.

Sharpe W.F. (1964), *Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk*, "Journal of Finance", Vol. 19 (September), pp. 425-442.

Sharpe W.F. (1977), *The Capital Asset Pricing Model: A 'Multi-Beta' Interpretation* [in:] H. Levy, M. Sarnat (eds.), *Financial Decision Making Under Uncertainty*, Harcourt Brace Jovanovich, Academic Press, New York.

Trzpiot G. (2008), *O wybranej metodzie estymacji beta* [in:] P. Chrzan, E. Dziwok (red.), *Metody matematyczne, ekonometryczne i komputerowe w finansach i ubezpieczeniach*, AE, Katowice, pp. 345-354.

Trzpiot G. (2019), *Application Quantile-Based Risk Measures in Sector Portfolio Analysis −Warsaw Stock Exchange Approach* [in:] W. Tarczynski, K. Nermend (eds.), *Effective Investments on Capital Markets*, Springer Proceedings in Business and Economics, Springer, pp. 405-422.

Trzpiot G. (2021), *Gini Regression in the Capital Investment Risk Assessment − Sensitivity Risk Measures in Portfolio Analysis* [in:] K. Jajuga, K. Najman, M. Walesiak (eds.), *Data Analysis and Classification. Methods and Applications*, "Studies in Classification, Data Analysis, and Knowledge Organization", s. 101-115.

Yitzhaki S. (2015), *Gini's Mean Difference Orders a Response to Leamer's Critique*, "Metron", Vol. 73, pp. 31-43.

Yitzhaki S., Schechtman E. (2013), *The Gini Methodology. A Primer on a Statistical Methodology*, Springer Series in Statistics, Vol. 272.

Chapter III

Non-parametric econometric models in risk analysis

Dominik Krężołek

1. Introduction to econometric modelling

Econometric modelling is a statistical method used to estimate and analyse the relationships between economic variables. It is an important tool for empirical research in economics, finance, and business, and can be used to forecast future economic trends, evaluate policy interventions, and understand the behaviour of economic agents. Econometric models typically involve estimating the relationship between a dependent variable and one or more independent variables, using data collected over time or across individuals. These models can take a variety of forms, depending on the nature of the data and the research question being addressed [Madalla, 2006].

One common type of econometric model is a regression model, which estimates the relationship between a dependent variable and one or more independent variables using a linear or nonlinear equation. These models can be used to estimate the impact of changes in one variable on another, and to test hypotheses about the nature of the relationship between variables. Another type of econometric model is a time-series model, which estimates the behaviour of a single variable over time, taking into account the influence of past values and other relevant factors. These models can be used to forecast future trends in the variable of interest and to analyse the impact of shocks or policy interventions. Econometric modelling can be used to estimate a wide range of economic relationships, from the demand for a particular good or service, to the impact of monetary policy on inflation or economic growth. These models are often used in conjunction with statistical software packages, which provide tools for data analysis, model estimation, and hypothesis testing [Davidson and MacKinnon, 2004].

There are several types of econometric modelling, each of which is suited to different research questions and data structures. Among others we can mention here cross-sectional models, time-series models, panel data models or structural models, etc. However, in a general sense, econometric modelling includes parametric, non-parametric and semi-parametric models [Henderson and Parmeter, 2015].

Parametric econometric model is a statistical tool used to estimate the relationships between economic variables by assuming a particular functional form for the relationship, typically based on prior knowledge or theoretical considerations. In parametric models, the parameters of the functional form are estimated using data, and the model is evaluated based on its ability to fit the data and make accurate predictions. Parametric models are commonly used in economics and finance to estimate demand and supply functions, evaluate the effectiveness of policy interventions, and forecast future trends. Common examples of parametric models include linear regression models, logistic regression models, and autoregressive models [Hayashi, 2000].

One advantage of parametric models is that they are relatively easy to interpret and explain, as the functional form of the model is explicitly defined. This allows for straightforward inference about the nature of the relationship between variables and the effects of policy interventions. However, parametric models also have several limitations. One limitation is that they may not be flexible enough to capture complex and nonlinear relationships between variables, especially when the underlying functional form is unknown or difficult to specify. Another limitation is that parametric models require strong assumptions about the distribution and structure of the data, which may not always hold in practice.

On the other hand, we have non-parametric modelling. Non-parametric econometric modelling is a statistical method used to estimate the same relationships that parametric models but making any assumptions about the functional form of the relationship. In non-parametric models, the relationship between variables is estimated using the data itself, rather than assuming a particular functional form. Non-parametric models are very flexible and can capture complex and nonlinear relationships between variables, without requiring strong assumptions about the functional form of the model. This makes them particularly useful in situations where the underlying relationship between variables is unknown or difficult to specify. But they also have some limitations. For example, they may be more computationally intensive than parametric models, especially when dealing with large datasets or high-dimensional data structures. Another limitation is that they may be less interpretable than parametric models, as the functional form of the model is not explicitly defined. A compromise between parametric and non-parametric approaches are semi-parametric models [Greene, 2017].

Semi-parametric econometric models include a statistical method that combine the flexibility of non-parametric models with the efficiency and interpretability of parametric models. In semi-parametric modelling, some aspects of the model are specified parametrically, while others are left unspecified or modelled

non-parametrically. In a typical semi-parametric econometric model, a parametric model is used to model the mean or trend of the data, while a non-parametric model is used to model the variance or residuals of the data. This approach allows for more accurate estimation of complex and nonlinear relationships between variables, without requiring strong assumptions about the functional form of the model [Ichimura, 1993; Horowitz, 1998].

Semi-parametric models can be used to estimate various types of econometric models, such as regression models, time-series models, and panel data models. The specific choice of parametric and non-parametric components depends on the research question, the data structure, and the available resources. Semi-parametric models are widely used in various fields, including economics, finance, and environmental studies. They are particularly useful in situations where the underlying relationship between variables is complex, and the traditional parametric models may not be appropriate. Semi-parametric models are also used in the analysis of censored and truncated data, where non-parametric methods may be required to handle the non-normality and non-linearity of the data [Xia, Tong and Li, 2012].

The choice of econometric model depends on the research question, the data structure, and the available resources. Each type of model has its strengths and limitations, and careful consideration is required when choosing an appropriate model and estimation technique.

2. Non-parametric estimation of density function

As mentioned in the previous subsection, in non-parametric econometric models any assumptions about the functional form of the relationship between variables are not required. This relationship is estimated using the data itself, rather than assuming a particular functional form. Non-parametric models are commonly used in economics and finance to estimate demand and supply functions, model consumer behaviour, and forecast future trends. Common examples of non-parametric models include among others: kernel regression models, spline regression models, and local polynomial regression models.

2.1. Kernel density estimator

The most popular representation of a density function is histogram. A histogram is a graphical representation of a non-parametric density function that estimates the probability distribution of a random variable by dividing the data into

equal-width bins or intervals and counting the number of observations that fall into each bin. The resulting histogram shows the frequency or count of observations in each bin and provides a visual representation of the shape of the probability distribution. The histogram can be a useful tool for exploratory data analysis and for comparing the shapes of different probability distributions. It is particularly useful when the underlying distribution is not known or when it is difficult to specify a parametric model that fits the data [Scott, 2015].

One advantage of histograms is that they are relatively easy to interpret and can provide insights into the shape, skewness, and kurtosis of the probability distribution. They can also be used to identify outliers or gaps in the data and to detect potential issues such as data truncation or censoring. However, histograms also have some limitations. The choice of bin width can affect the shape and appearance of the histogram, and different bin widths can lead to different interpretations of the data. Additionally, histograms may not be as accurate or precise as other non-parametric density estimation methods, such as kernel density estimation, especially for small or irregularly shaped datasets [Sheather, 2004].

Let $X_1, X_2, \dots, X_n$ be a simple sample of random variable X and let $f(x)$ be the unknown density function. Therefore, the kernel estimator of the density function $f(x)$ is of the form [Rosenblatt, 1956; Parzen, 1962]:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \tag{1}$$

where h is a bandwidth (width of the bin) and $K(\psi)$ is a kernel function.

The kernel function $K(\psi)$ used in non-parametric methods has several important properties that affect the accuracy and performance of the estimator. Some common properties of the kernel function $K(\psi)$ include:

- the kernel function must be non-negative, so that the density estimator is also non-negative,
- the kernel function should be symmetric around its center, so that it can capture both positive and negative effects,
- the kernel function must integrate to one over its support, so that the density estimator integrates to one over the entire domain,
- the kernel function is often chosen to be unimodal, so that it can provide a smooth estimate of the density function,
- the kernel function should be smooth, so that the density estimator is also smooth and can capture local patterns in the data,
- the kernel function is dependent on the bandwidth parameter, which controls the width of the smoothing window around each data point; the choice of bandwidth can affect the bias-variance trade-off of the density estimator, as well as its accuracy and performance.

In general, we can assume that for the kernel function $K(\psi)$ we have:

$$\int_{-\infty}^{+\infty} K(\psi)d\psi = 1 \tag{2}$$

$$\int_{-\infty}^{+\infty} \psi\, K(\psi)d\psi = 0 \tag{3}$$

$$\int_{-\infty}^{+\infty} \psi^2\, K(\psi)d\psi = \kappa_2 < \infty \tag{4}$$

where κ_2 is the central moment of the second order [Pagan and Ullah, 1999].

The choice of kernel function depends on the research question, the data structure, and the desired properties of the density estimator. The most popular kernel functions are [Fan and Yao, 2005]:

- standard normal (Gaussian): $K(\psi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\psi^2}{2}\right)$,
- Epanechnikov: $K(\psi) = \frac{3}{4\sqrt{5}}\left(1 - \frac{\psi^2}{5}\right)$ for $|\psi| < \sqrt{5}$,
- triangular: $K(\psi) = 1 - |\psi|$ for $|\psi| < 1$,
- uniform: $K(\psi) = \frac{1}{2}$ for $|\psi| < 1$,
- biweight: $K(\psi) = \frac{15}{16}(1 - \psi^2)^2$ for $|\psi| < 1$,
- triweight: $K(\psi) = \frac{35}{32}(1 - \psi^2)^3$ for $|\psi| < 1$.

The kernel density estimator has several advantages over parametric methods of density estimation. It does not require making any assumptions about the functional form of the distribution and can be used to estimate the density function for distributions that are asymmetric, heavy-tailed, or have other complex shapes. Additionally, the kernel density estimator can be used for both univariate and multivariate density estimation. However, the kernel density estimator also has some limitations. It can be sensitive to the choice of kernel function and bandwidth parameter and may require cross-validation or other methods to choose appropriate values. Additionally, the kernel density estimator may be computationally intensive for large datasets.

2.2. Bandwidth selection

Bandwidth selection is a critical step in non-parametric density estimation using kernel methods, such as kernel density estimation. The choice of bandwidth parameter affects the smoothness of the density estimator, the bias-variance trade-off, and the accuracy of the estimation. There are several methods for bandwidth selection, some of which are [Wand and Jones, 1995]:

1. Cross-validation method – is a commonly used method for bandwidth selection in kernel density estimation. The data is split into training and validation

sets, and the estimator is trained on the training set using a range of bandwidth values. The performance of the estimator is then evaluated on the validation set using a suitable criterion, such as mean square error (MSE) or integrated mean square error (IMSE), and the bandwidth that minimizes the criterion is chosen as the optimal bandwidth.

2. Rule-of-thumb method – is a simple heuristic method for bandwidth selection in kernel density estimation. It is based on the standard deviation of the data and assumes a Gaussian kernel. The optimal bandwidth is given by $h = 1.06\sigma n^{-\frac{1}{5}}$, where σ is the standard deviation of the data and n is the sample size.
3. Plug-in method – is a data-driven method for bandwidth selection that is based on estimating the optimal bandwidth from the data. It involves replacing the unknown density function with an estimator and then using the estimator to estimate the optimal bandwidth. One common estimator used in the plug-in method is the Gaussian kernel estimator.
4. Maximum likelihood – is a method for bandwidth selection that involves maximizing the likelihood function of the data with respect to the bandwidth parameter. This method requires specifying a parametric form for the density function and assumes a Gaussian kernel.
5. Bayesian methods – these are probabilistic methods for bandwidth selection that involve specifying a prior distribution over the bandwidth parameter and updating the prior using the data to obtain the posterior distribution. The optimal bandwidth is then obtained by maximizing the posterior distribution.

Overall, the choice of bandwidth selection method depends on the research question, the data structure, and the desired properties of the estimator.

2.3. Accuracy of the kernel estimator

For assessing the properties of many kernel methods, different criteria are used. The first one is the pointwise mean squared error (MSE), which can be written as [Pagan and Ullah, 1999]:

$$MSE\hat{f}(x) = E\left[\hat{f}(x) - f(x)\right]^2 = D^2\left(\hat{f}(x)\right) + \left[bias\hat{f}(x)\right]^2 \qquad (5)$$

where $bias\hat{f}(x) \approx \frac{h^2}{2} f''(x)\kappa_2$ and $D^2\left(\hat{f}(x)\right) \approx \frac{f(x)}{nh}\int_{-\infty}^{+\infty} K^2(\psi)d\psi$.

Is worth of emphasizing that both the bias and variance depend on the bandwidth (bias falls as h decreases, variance rises as h decreases). The bias also increases with $f''(x)$, hence is highest in the peaks of distributions. But, as long as the conditions for consistency are met, namely $h \to 0$ as $n \to \infty$ ($bias \to 0$)

and $nh \to \infty$ as $n \to \infty$ ($D^2\left(\hat{f}(x)\right) \to 0$), then the bias related to $f''(x)$ will diminish as the available data increases and will vanish in the limit.

The second accuracy measure is the integrated mean square error (IMSE) which can be defined as [Li and Racine, 2007]:

$$IMSE\hat{f}(x) = \int_{-\infty}^{+\infty} MSE\hat{f}(x)\,dx =$$

$$= \int_{-\infty}^{+\infty} D^2\left(\hat{f}(x)\right)dx + \int_{-\infty}^{+\infty}\left[bias\hat{f}(x)\right]^2 dx = \frac{\int_{-\infty}^{+\infty} K^2(\psi)d\psi}{nh} +$$

$$+ \frac{h^4}{4}\kappa_2^2 \int_{-\infty}^{+\infty}(f''(x))^2\,dx = \frac{\Phi_0}{nh} + \frac{h^4}{4}\kappa_2^2\Phi_1 \qquad (6)$$

where $\Phi_0 = \int_{-\infty}^{+\infty} K^2(z)dz$ and $\Phi_1 = \int_{-\infty}^{+\infty}[f''(x)]^2 dx$.

Minimizing MSE and IMSE with respect to h we can obtain optimal bandwidths:

$$h_{opt} = \left[\frac{\int_{-\infty}^{+\infty} K^2(z)dz}{\left(\int_{-\infty}^{+\infty} z^2K(z)dz\right)^2 \int_{-\infty}^{+\infty}[f''(x)]^2 dx}\right]^{\frac{1}{5}} n^{-\frac{1}{5}} = bn^{-\frac{1}{5}} \qquad (7)$$

Note that the constant b depends on $f''(x)$ and $K(\cdot)$.

Both MSE and IMSE can be used as criteria for bandwidth selection, and the optimal bandwidth is the one that minimizes the criterion. Cross-validation is a commonly used method for minimizing the MSE criterion, while the plug-in method is a commonly used method for minimizing the IMSE criterion.

3. Risk and measures of risk

Risk is the potential for a negative outcome or loss that arises from an uncertain event or situation. It is the chance that an actual outcome will differ from the expected outcome, and can arise from a variety of factors, including natural disasters, market volatility, technological failures, human error, and geopolitical events. Risk is often quantified in terms of probability and impact, with the probability representing the likelihood of an event occurring and the impact representing the potential consequences of the event. The severity of the risk is a function of both the probability and the impact of the potential loss. In finance and investments, risk is an essential concept, as investors face a variety of risks when making investment decisions. These risks include market risk, credit risk, liquidity risk, and operational risk, among others. Understanding and managing these risks is critical for achieving investment objectives and avoiding potential losses [Jajuga, 2018].

Taking into account the information we have, decision-making problems can be divided into two types of decisions:

1. Deterministic decisions which involve situations where the outcome is certain and there is no uncertainty. The decision maker has complete knowledge of the consequences of each decision and can make the best decision based on their preferences and objectives.
2. Probabilistic decisions which involve situations where the outcome is uncertain and there is risk involved. The decision maker has incomplete knowledge of the consequences of each decision and must consider the likelihood of different outcomes before making a decision.

Within probabilistic decisions, there are two types of decisions: decisions under risk and decisions under uncertainty. Decisions under risk involve situations where the probabilities of different outcomes are known or can be estimated with some degree of confidence. The decision maker can use probability theory to calculate the expected value of each decision and can choose the decision with the highest expected value. Decisions under uncertainty involve situations where the probabilities of different outcomes are unknown or cannot be estimated with confidence. The decision maker must rely on subjective judgments, expert opinions, or scenario analysis to evaluate the potential outcomes of each decision [Gilboa, 2009].

3.1. Extreme risk measures

Extreme risk, also known as tail risk, refers to the possibility of an unlikely but highly impactful event occurring, with potential consequences that are significantly greater than those of typical events. Extreme risk is often associated with events that occur in the tails of a probability distribution, where the likelihood of occurrence is very low, but the potential impact is very high. Examples of extreme risk events include natural disasters such as earthquakes, hurricanes, and tsunamis, as well as economic and financial crises such as market crashes, sovereign defaults, and system-wide banking failures. Extreme risk events can have severe consequences for individuals, businesses, and society as a whole, and can lead to significant economic, financial, and social disruption [McNeil, Frey and Embrechts, 2015].

Measuring and managing extreme risk is an important challenge for investors, financial institutions, and policymakers. Measures of extreme risk include Value at Risk (VaR), Expected Shortfall (ES), among others. These measures provide a framework for estimating the probability of extreme events and their

potential impact and can help investors and policymakers make informed decisions about risk management and risk mitigation strategies [Dowd, 2005].

One of the most popular extreme risk measures is Value at Risk (VaR). VaR is a measure of the maximum potential loss that can be incurred under normal market conditions, at a given confidence level $1 - \alpha$. VaR estimates the potential loss over a specific time horizon, and provides a threshold below which the probability of losses is unlikely to exceed. VaR can be estimated using various parametric and non-parametric methods, including historical simulation, Monte Carlo simulation, and kernel density estimation [Wang and Wang, 2013]. Mathematically VaR can be expressed as:

$$VaR_\alpha(X) = \inf\{x|F_X(x) \geq \alpha\} = F_X^{-1}(\alpha) \tag{8}$$

VaR can be estimated using various parametric and non-parametric methods, including historical simulation, Monte Carlo simulation, and kernel density estimation.

Value at Risk is a widely used risk measure in finance, but it is not an ideal risk measure in all circumstances. While VaR provides a useful estimate of the potential loss that an investor or portfolio may incur at a given confidence level over a specific time horizon, it has some limitations and weaknesses that may make it less suitable for certain applications. One limitation of VaR is that it only considers the potential losses beyond the VaR threshold and ignores the severity of the losses beyond that threshold. This means that VaR may not provide a complete picture of the potential losses that an investor or portfolio may face and may underestimate the risk of tail events or extreme losses. Another limitation of VaR is that it assumes that the underlying distribution of returns is known or can be estimated accurately. In practice, however, the true distribution of returns may be unknown or may change over time, which can lead to inaccurate estimates of VaR and may lead to underestimation of risk. Furthermore, VaR does not consider the timing of the potential losses, which can be an important factor in risk management and portfolio optimization. In some applications, such as liability management, the timing of potential losses may be more important than the size of the losses themselves. VaR does not account for the potential benefits or diversification effects of different investments or risk management strategies. Finally, VaR is not coherent according to the axioms of Artzner et al. [1999]:

- monotonicity – if the outcomes of one random variable stochastically dominate those of another, the risk measure of the former should be no less than the risk measure of the latter,
- subadditivity – the risk measure of the sum of two random variables should be no greater than the sum of their individual risk measures,

- translation invariance – adding a constant to the outcomes of a random variable should not affect its risk measure,
- positive homogeneity – multiplying the outcomes of a random variable by a constant should multiply its risk measure by the same constant.

These axioms ensure that the risk measure is consistent and meaningful, and that it reflects the true nature of risk. A risk measure that violates one or more of these axioms may lead to inconsistent or counterintuitive results and may not be suitable for making informed decisions about risk management and risk mitigation strategies. VaR does not satisfy the axiom of subadditivity. Subadditivity means that the total risk of a portfolio is always less than or equal to the sum of the risks of its individual components. VaR does not satisfy this axiom because the VaR of a portfolio is not necessarily equal to the sum of the VaRs of its individual components. In fact, the VaR of a portfolio can be greater than the sum of the VaRs of its individual components, due to the potential for diversification effects. Diversification can reduce the overall risk of a portfolio and can lead to a lower VaR than would be expected from the sum of the VaRs of its individual components [Poon and Granger, 2003].

This limitation of VaR has led to the development of other risk measures, such as Expected Shortfall (ES) that does satisfy the axiom of subadditivity and provide a more comprehensive measure of portfolio risk. ES takes into account the severity of losses beyond the VaR threshold and provide a more accurate estimate of the potential losses that a portfolio may incur. ES is more suitable for evaluating the risk and return trade-offs of different investments and risk management strategies [Acerbi and Tasche, 2002]. For a continuous loss distribution with density function $f_X(x)$ and a given confidence level α Expected Shortfall can be expressed as:

$$ES_\alpha(X) = \frac{1}{1-\alpha}\int_{-\infty}^{VaR_\alpha(X)} x f_X(x)\,dx \tag{9}$$

ES measures the expected loss of an asset or portfolio beyond the VaR threshold. It represents the average of the losses that exceed the VaR threshold and provides a more accurate estimate of the potential losses that a portfolio may incur in extreme market conditions. ES can be calculated using a variety of methods, including historical simulation, Monte Carlo simulation, and analytical methods. CVaR, on the other hand, is a risk measure that represents the average of the losses beyond a certain confidence level, typically the VaR threshold. CVaR is similar to ES in that it incorporates information about the severity of losses beyond the VaR threshold but differs in the way it is calculated and interpreted [Denault, 2001].

CVaR is a coherent risk measure that satisfies important axioms of risk measurement, including subadditivity, positive homogeneity, and translation invariance. While ES and CVaR are similar in concept, they have different strengths and weaknesses. ES is more intuitive and easier to calculate but may be less coherent than CVaR in some cases. CVaR, on the other hand, is more coherent and provides a more comprehensive measure of portfolio risk but may be more difficult to calculate and interpret [Rockafellar and Uryasev, 2002].

3.2. Kernel estimates of VaR and ES

As mentioned in subsection 2, kernel density estimation is a non-parametric method for estimating the probability density function of a random variable. Kernel methods can also be used to estimate VaR and ES, by estimating the distribution of portfolio returns and using this distribution to estimate the probabilities of extreme losses.

Let $X_1, X_2, \ldots, X_n$ be a simple sample of random variable X and let $F(x)$ be the unknown cumulative density (cdf) function. Therefore, the non-parametric estimator of the cdf is of the form [Lee and González-Rivera, 2008]:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{10}$$

where $I(\cdot)$ denotes an indicator function taking the value 1 when the expression in parentheses is true. The empirical distribution $\hat{F}_n(x)$ is an unbiased estimate of $F(x)$, but has a larger variance than alternative nonparametric methods. Having $\hat{F}_n(x)$ we can easily reformulate (8) and (9) as the empirical estimates of risk measures:

- empirical VaR:

$$\widehat{VaR}_\alpha(X)_n = \inf\{x | \hat{F}_n(x) \geq \alpha\} = \hat{F}_n^{-1}(\alpha) \tag{11}$$

- empirical ES:

$$\widehat{ES}_\alpha(X)_n = \frac{\sum_{i=1}^{n} X_i I(X_i \geq \widehat{VaR}_\alpha(X)_n)}{\sum_{i=1}^{n} I(X_i \geq \widehat{VaR}_\alpha(X)_n)} \tag{12}$$

To estimate VaR and ES using kernel estimation, we have to estimate first the kernel estimator of (10):

$$\hat{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(u)\, du = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{13}$$

Finally, the kernel estimates of VaR and ES are of the form:

- kernel estimator of VaR:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \alpha \tag{14}$$

– kernel estimator of ES:

$$\widehat{ES}_n(x) = \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} X_i \left[1 - K\left(\frac{\widehat{VaR}_\alpha(X)_n - X_i}{h}\right)\right] \tag{15}$$

where $\hat{S}_n(x) = 1 - \hat{F}_n(x)$ is the kernel estimator of survival function $S_n(x)$.

Kernel estimates of VaR and ES offer several advantages over other methods, including flexibility, non-parametric nature, and computational efficiency. However, they also have some limitations and should be used with caution, particularly with regard to the choice of kernel function and bandwidth [Yang and Härdle, 2007].

4. Empirical research

In subsection 4, we present the results of extreme risk measurement based on the kernel estimator of the density function. In the analysis, we use quotations of two indices: WIG20 and S&P500 over the period from January 2018 to December 2021, by calculating daily log-returns using as $R_t = \ln \frac{p_t}{p_{t-1}} \cdot 100$. We compare empirical estimates of VaR and ES with standard normal kernel estimators of density function (based on three bandwidth selection methods: rule-of--thumb, plug-in, and cross-validation) with two theoretical distributions: Student t and Generalized Error Distribution (GED). Extreme risk measures are calculated for two quantiles: 0.005 and 0.001.

The purpose of the study is to reveal the validity of using non-parametric methods for estimating extreme risk compared to parametric approaches. The hypothesis being tested says that estimates of VaR and ES based on the kernel estimator of the density function are more accurate than estimates using parametric distributions. As a measure of accuracy, we use the root mean square error (RMSE).

Empirical time series of WIG20 and S&P500 log-returns are presented in Fig. 1-2 whereas descriptive statistics in Table 1.

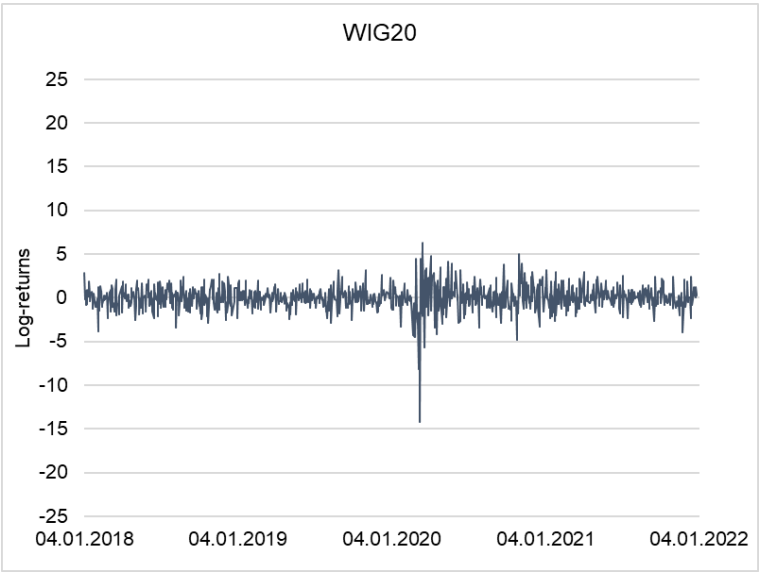


Fig. 1. Log-returns of WIG20

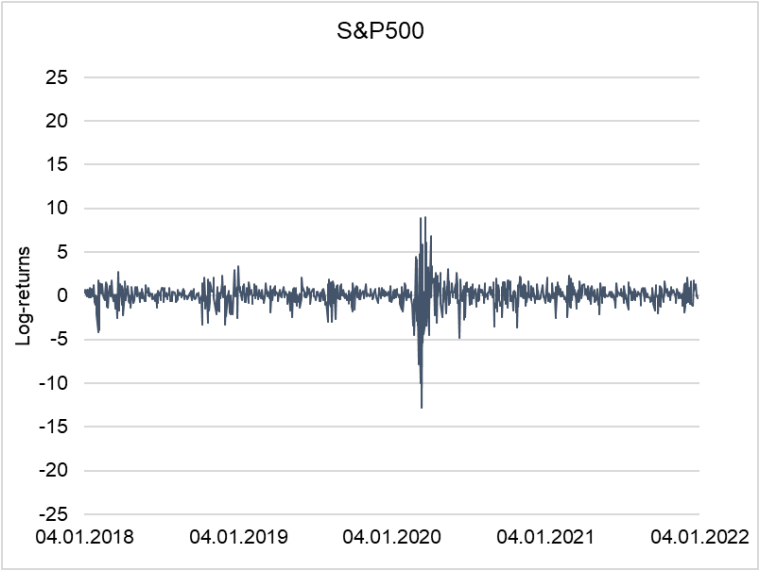


Fig. 2. Log-returns of S&P500

Table 1. Descriptive statistics for log-returns of WIG20 and S&P500

Descriptive statistics	WIG20	S&P500
Mean	−0.00833	0.05678
Standard error	0.04465	0.04199
Median	−0.01927	0.09922
Standard deviation	1.40996	1.32572
Kurtosis	12.64007	19.00007
Skewness	−1.15041	−1.07600
Range	20.58084	21.73353
Minimum	−14.24558	−12.76521
Maksimum	6.33526	8.96832
Anderson-Darling (*p*-value)	<0.001*	<0.001*

* Statistical significance at 0.01.

Source: Own calculations.

Analysing both Fig. 1-2 and the data in Table 1, we observe that the average return of the WIG20 during the examined period was negative, while positive for the S&P500. Empirical distributions are leptokurtic, skewed to the left and non-normally distributed. In addition, on both charts we can see the clustering of variance and high volatility in February 2020, which is due to the WHO announcement of the COVID-19 pandemic. Empirical and theoretical distributions of returns are presented in Fig. 3-4.

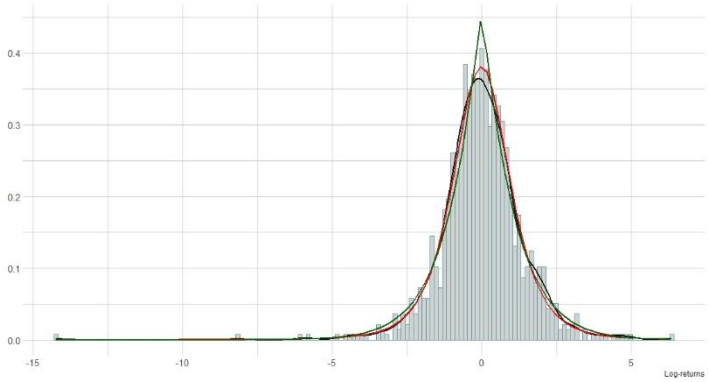


Fig. 3. Histogram with kernel density function (black), Student t density (red) and GED density (dark green) – WIG20

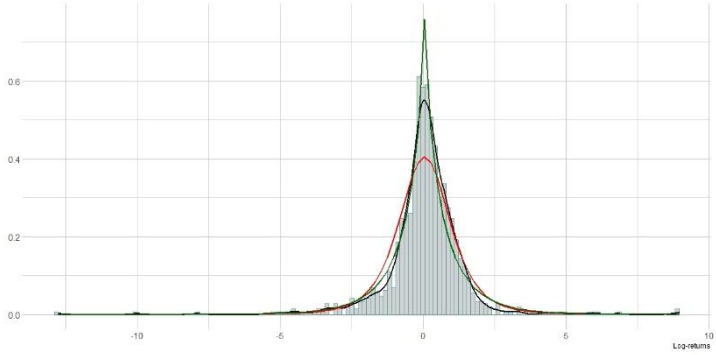


Fig. 4. Histogram with kernel density function (black), Student t density (red) and GED density (dark green) – S&P500

In the next step of the analysis, empirical density functions for both indices were estimated assuming a Gaussian kernel function and three bandwidth selection methods. The results are shown in Table 2.

Table 2. Optimal bandwidth (Gaussian kernel function) for three estimation methods

Estimation method	WIG20		S&P500	
	Optimal bandwidth	Optimal no. of intervals	Optimal bandwidth	Optimal no. of intervals
Rule-of-thumb	0.28979	70	0.20342	107
Plug-in	0.26142	79	0.13882	157
Cross-validation	0.28486	72	0.11756	185

Source: Own calculations.

As we can see, the optimal bandwidth varies depending on the estimation method adopted. For both indices, the widest bandwidth was obtained for the rule-of-thumb method. The optimal number of intervals is also the smallest for this method.

The final step of the study is the estimation of VaR and ES risk measures. Three theoretical distributions were used: the kernel estimator of the density function, Student's t distribution and the GED distribution. For the kernel estimator, the estimation was carried out using three methods: rule-of-thumb, plug-in and cross-validation. VaR and ES was calculated for two quantiles: 0.001 and 0.005. The results are shown in Tables 3-4, whereas the root mean square error (RMSE) values are presented in Tables 5-6.

Table 3. VaR and ES estimates for WIG20

Risk measure	Quantile	WIG20					
		Empirical	Student t	GED	Gaussian kernel estimates		
					rule-of-thumb	plug-in	cross-validation
VaR	0.001	−8.21683	−4.36542	−5.01049	**−8.13466**	−8.94813	−8.85865
	0.005	−4.52038	−3.64014	−3.18393	−4.33956	**−4.38296**	−4.25147
ES	0.001	−11.21910	−4.73912	−9.80704	−10.99472	−11.33555	**−11.10884**
	0.005	−7.79906	−4.06919	−6.34229	−7.64308	−8.02523	**−7.94498**

Source: Own calculations.

Table 4. VaR and ES estimates for S&P500

Risk measure	Quantile	S&P500					
		Empirical	Student t	GED	Gaussian kernel estimates		
					rule-of-thumb	plug-in	cross-validation
VaR	0.001	−10.00557	−4.04001	−3.59397	**−9.90551**	−10.89607	−10.78711
	0.005	−4.78230	−3.35805	−1.69321	−4.59101	**−4.63692**	−4.49781
ES	0.001	−11.37985	−4.52061	−10.04144	−11.15225	−11.49797	**−11.26801**
	0.005	−8.19865	−3.89070	−9.42921	**−8.11667**	−8.52250	−8.43727

Source: Own calculations.

Table 5. RMSE for WIG20

Risk measure	Quantile	WIG20				
		Student t	GED	Gaussian kernel estimates		
				rule-of-thumb	plug-in	cross-validation
VaR	0.001	3.85141	3.20634	**0.08217**	0.73130	0.64182
	0.005	0.88024	1.33646	0.18082	**0.13742**	0.26891
ES	0.001	6.47998	1.41206	0.22438	0.11645	**0.11026**
	0.005	3.72987	1.45677	0.15598	0.22617	**0.14592**

Source: Own calculations.

Table 6. RMSE for S&P500

Risk measure	Quantile	S&P500				
		Student t	GED	Gaussian kernel estimates		
				rule-of-thumb	plug-in	cross-validation
VaR	0.001	5.96556	6.41160	**0.10006**	0.89050	0.78154
	0.005	1.42425	3.08909	0.19129	**0.14538**	0.28449
ES	0.001	6.85924	1.33841	0.22760	0.11812	**0.11184**
	0.005	4.30795	1.23056	**0.08199**	0.32385	0.23862

Source: Own calculations.

The results of risk measurement show that the most accurate estimates of VaR and ES were obtained for the kernel estimator method (estimates with the smallest values of RMSE are marked in bold). For the WIG20, the smallest RMSE values were obtained for the plug-in method (VaR for quantiles 0.001 and 0.005) and for the cross-validation method (ES for quantiles 0.001 and 0.005). In contrast, for the S&P500, the result is similar to VaR for WIG20, while the cross-validation method is reported for the ES measure for the 0.001 quantile, and the rule-of-thumb method for the 0.005 quantile. Comparing the empirical estimates of the risk measures, it was observed that the U.S. index quotations were riskier during the examined period.

5. Conclusions

Non-parametric econometric models are becoming increasingly important in the contemporary world due to several factors. One of them is increased complexity of data. As the amount and complexity of data continue to grow, non-parametric models offer a flexible and robust approach for analysing data that may not conform to traditional parametric models. Non-parametric models can also provide more accurate estimates of relationships between variables by allowing for more flexible relationships and accounting for outliers and other irregularities in the data. As was mentioned at the beginning, these types of models can be used to estimate relationships between variables that may not have a known functional form or may be difficult to model parametrically. Non-parametric models do not require assumptions about the distribution of the data or the functional form of the relationship between variables, making them more robust to deviations from these assumptions. Moreover, non-parametric models have a wide range of applications in various fields such as finance, economics, environmental sciences, social sciences or medical research, and are particularly useful when data do not follow a specific parametric model or when there is a need for more flexible and robust models that can handle complex relationships between variables.

In this chapter we presented the possibility of using nonparametric econometric models in risk analysis. We described the approach based on the kernel estimator of the density function. It has several advantages. The kernel estimator is a useful tool for risk analysis due to its flexibility, robustness, and ability to accurately estimate the tails of the distribution. It is particularly useful when the underlying distribution is complex or unknown and can provide valuable insights into the risk profile of a portfolio or investment.

In the empirical part of the chapter, we presented the results of extreme risk analysis for two indices: WIG20 and S&P500 for the period from January 2018 to December 2021. We calculated daily log-returns and estimated two risk measures: VaR and ES for quantiles of 0.001 and 0.005. We compared the results for estimates of risk measures based on parametric distributions (Student t and GED) with estimates obtained using a kernel estimator of the density function (where we used a Gaussian kernel and three different bandwidth estimation methods). The results showed that the average return of the WIG20 during the examined period was negative, while positive for the S&P500. Empirical distributions were leptokurtic, skewed to the left and non-normally distributed. In addition, the time series of returns for both indices exhibited a high level of volatility. In turn, risk analysis showed that VaR and ES estimates based on the kernel estimator of the density function exhibited lower RMSE values to parametric distributions. In addition, S&P500 returns exhibited higher risk than returns of WIG20. This can be due to several reasons. The S&P500 is composed of 500 large-cap U.S. companies across multiple industries, while the WIG20 is composed of 20 large-cap Polish companies. This means that investments in the S&P500 are more diversified across sectors and companies, but also more susceptible to the overall health of the U.S. economy. On the other hand, investments in the WIG20 are more concentrated in a few sectors and companies, but less susceptible to U.S. economic conditions. The other reason is currency risk. Investing in the S&P500 involves exposure to fluctuations in the U.S. dollar, which can add an additional layer of risk for investors in other currencies. The WIG20, on the other hand, is denominated in Polish zloty, which can reduce currency risk for Polish investors. Moreover, the U.S. stock market is generally more volatile than the Polish stock market due to the larger size and greater number of companies in the S&P500, as well as the overall size and complexity of the U.S. economy. This can lead to higher levels of volatility in returns for investments in the S&P500. An important factor is related to the regulatory risk. The U.S. and Polish markets have different regulatory environments and legal systems, which can create additional risks for investors in the S&P500. For example, changes in U.S. tax laws or regulations can have a significant impact on the value of U.S. stocks.

In summary, we can confirm the validity of using non-parametric econometric models in risk analysis (including extreme risk). We also confirmed the research hypothesis which assumed that estimates of VaR and ES based on the kernel estimator of the density function are more accurate than estimates using parametric distributions.

References

Acerbi C., Tasche D. (2002), *Expected Shortfall: A Natural Coherent Alternative to Value at Risk*, "Economic Notes", Vol. 31(2), pp. 379-388.

Artzner P., Delbaen F., Eber J.-M., Heath D. (1999), *Coherent Measures of Risk*, "Mathematical Finance", Vol. 9(3), pp. 203-228.

Davidson R., MacKinnon J.G. (2004), *Econometric Theory and Methods*, Oxford University Press, Oxford.

Denault M. (2001), *Coherent Risk Measures and Their Applications in Financial Risk Management*, "Risk Analysis", Vol. 21(3), pp. 433-447.

Dowd K. (2005), *Measuring Market Risk (2nd ed.)*, John Wiley & Sons, Chichester.

Fan J., Yao Q. (2005), *Nonlinear Time Series*, Springer Series in Statistics, Springer, New York.

Gilboa I. (2009), *Theory of Decision under Uncertainty*, Cambridge University Press, New York.

Greene W.H. (2017), *Econometric Analysis*, Pearson Education Limited.

Hayashi F. (2000), *Econometrics*, Princeton University Press.

Henderson D.J., Parmeter C.F. (2015), *Applied Nonparametric Econometrics*, Cambridge University Press, Cambridge.

Horowitz J.L. (1998), *Semiparametric Methods in Econometrics*, Springer-Verlag, New York.

Ichimura H. (1993), *Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-index Models*, "Journal of Econometrics", Vol. 58(1-2), pp. 71-120.

Jajuga K. (2018), *Zarządzanie ryzykiem*, Wydawnictwo Naukowe PWN, Warszawa.

Lee T.H., González-Rivera G. (2008), *Nonparametric Estimation of Value-at-Risk Based on Extreme Value Theory*, "Journal of Econometrics", Vol. 147(1), pp. 23-35.

Li Q., Racine J.S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton.

Maddala G.S. (2006), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.

McNeil A.J., Frey R., Embrechts P. (2015), *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition*, Princeton University Press, Princeton.

Pagan A., Ullah A. (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Parzen E. (1962), *On Estimation of a Probability Density Function and Mode*, "The Annals of Mathematical Statistics", No. 33, pp. 1065-1076.

Poon S.-H., Granger C.W.J. (2003), *Forecasting Volatility in Financial Markets: A Review*, "Journal of Economic Literature", Vol. 41, No. 2, pp. 478-539.

Rockafellar R.T., Uryasev S. (2002), *Conditional Value-at-Risk for General Loss Distributions*, "Journal of Banking and Finance", Vol. 26(7), pp. 1443-1471.

Rosenblatt M. (1956), *Remarks on Some Nonparametric Estimates of a Density Function*, "Annals of Mathematical Statistics", No. 27, pp. 832-377.

Scott D.W. (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.

Sheather S.J. (2004), *Density Estimation*, "Statistical Science", Vol. 19(4), pp. 588-597.

Wand M.P., Jones M.C. (1995), *Kernel Smoothing*, (Vol. 60), CRC Press.

Wang S., Wang X. (2013), *The Statistical Properties of Value at Risk*, "Journal of Financial Econometrics", Vol. 11(2), pp. 449-478.

Xia Y., Tong H., Li W.K. (2012), *A Review on Semiparametric Regression*, "Annual Review of Statistics and Its Application", pp. 311-352.

Yang F., Härdle W.K. (2007), *Nonparametric Risk Management with General Risk Function*, "Journal of Econometrics", Vol. 141(2), pp. 492-516.

Chapter IV

Comparative analysis of selected distance measures dedicated to time series

Alicja Ganczarek-Gamrot

1. Introduction

When analysing multivariate time series, we are often faced with the problem of non-uniform frequency of observations. The data from multiple sources is registered at intervals of varying length. We can solve this problem by aggregating data, losing information about the variability within in shorter periods. Taking into account additionally the non-stationary character of time series as well as time-varying correlations between them, methods allowing for the analysis of phenomena observed at different time intervals become interesting.

The aim of this chapter is to identify, among the distance measures dedicated to time series, those that can be used to group multidimensional time series. Cluster analysis was carried out using the average linkage agglomeration method. The Silhouette index was used to assess the quality of the clustering.

The electricity price [EUR/MWh] published on the Noord Pool platform in the period 10.02.-04.10.2021 was used for the comparative analysis.

2. Distance measures

For the analysis, both classical distance measures as well as those allowing for changes of studied values over time and the difference in length of the series were used. Among the measures considered were those based on observed values as well as on time series representations. Let: $\boldsymbol{X_T} = [X_1, X_2, \dots, X_T]^T$, $\boldsymbol{Y_T} = [Y_1, Y_2, \dots, Y_T]^T$ respectively the note to time series. Formulas of the selected distances measures are presented below.

The classical Euclidean distance is given by the formula:

$$\boldsymbol{d_E}(\boldsymbol{X_T}, \boldsymbol{Y_T}) = \sqrt{\sum_{t=1}^{T}(X_t - Y_t)^2} \quad (1)$$

It is based on time-consistent pairs of observations.

The Frechet [1906] distance:

$$\boldsymbol{d_F}\left(\boldsymbol{X_{T_1}}, \boldsymbol{Y_{T_2}}\right) = \min_{r \in M}\left(\max_{i=1,\ldots,m}\left|X_{a_i} - Y_{b_i}\right|\right) \tag{2}$$

where M is the set of all possible sequences of pairs m preserving the order of observations:

$$r = \left(\left(X_{a_1}, Y_{b_1}\right), \ldots, \left(X_{a_m}, Y_{b_m}\right)\right)$$

in which all combinations of pairs of points in the analysed time series are considered. On the one hand, it takes into account all time shifts relative to each other, but, on the other hand, it increases computation time and can distort the true similarity by taking into account outdated values.

A distance that reduces computational complexity and decreases the chances of comparing outdated values is DTW (Dynamic Time Warping) [Sankoff and Kruskal, 1983; Berndt and Clifford, 1994]:

$$d_{DTW}\left(X_{T_1}, Y_{T_2}\right) = \min_{r \in M}\left(\sum_{i=1}^{m}\left|X_{a_i} - Y_{b_i}\right|\right) \tag{3}$$

All three measures above do not take into account the chronological order of events that is so important in a time series. This problem is to some extent solved by a distance based on the similarity index [Douzal-Chouakria and Nagabhushan, 2007]:

$$d_{CORT}(X_T, Y_T) = f_k(CORT(X_T, Y_T))\, d_l(X_T, Y_T) \tag{4}$$

where:

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}}$$

$$f_k(u) = \frac{2}{1 + exp(ku)}, \quad k \geq 0$$

where l = a, b, c (a − Euclidean, b − Frechet, c − DTW distance).

The CORT coefficient takes values in the range <−1;1>. It is a measure of the similarity of two time series in the neighbourhood of one analysed period. Negative values suggest an opposite direction of change of the two time series, which we perceive as the values of the analysed series moving away from each other. When the CORT coefficient is positive we observe that respective realizations of time series tent to change in the same direction. We can thus treat them as closer to each other. In order to use the CORT coefficient to assess distance, it has been proposed [Douzal-Chouakria and Nagabhushan, 2007] to use a decreasing logistic function $f_k(u)$. This alleviate the problem of negative values and,

depending on the parameter *k*, increases the contribution of information about the degree of interdependence between the time series in assessing the distance between them (Fig. 1).

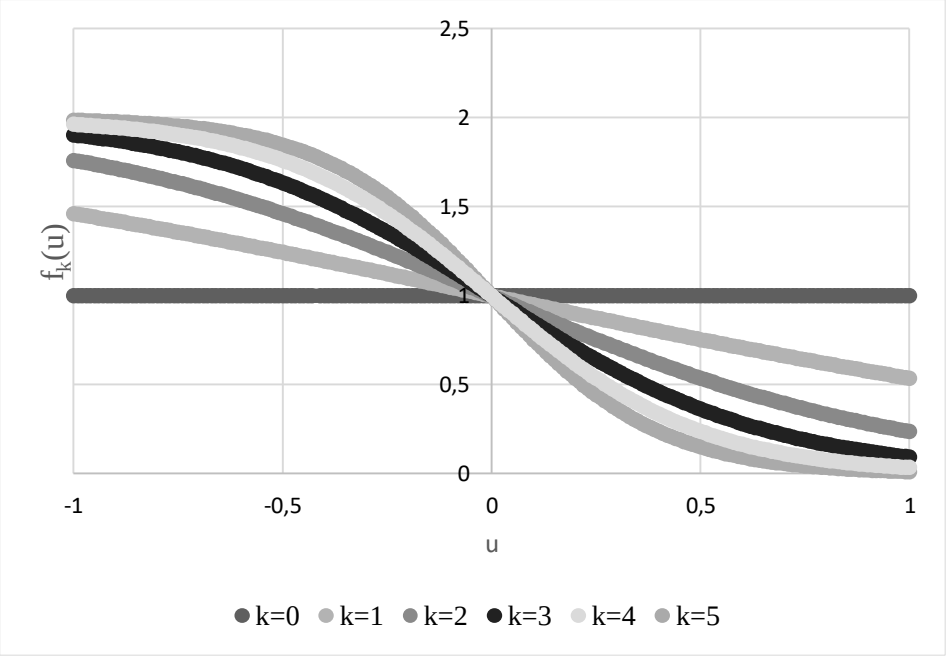


Fig. 1. The value of exponential adaptive tuning function $f_k(u)$ for various *k* values

Source: Own studies based on: Douzal-Chouakria and Nagabhushan [2007].

A similar measure to the classical Euclidean distance is the distance measured by the correlation coefficient [Golay et al., 2005]:

$$d_{COR.1}(X_T, Y_T) = \sqrt{2 - 2COR(X_T, Y_T)} \tag{5}$$

$$d_{COR.2}(X_T, Y_T) = \sqrt{\left(\frac{1-COR(X_T,Y_T)}{1+COR(X_T,Y_T)}\right)^{\beta}} \tag{6}$$

where:

$$\beta \geq 0$$

$$COR(X_T, Y_T) = \frac{\sum_{t=1}^{T}(X_t-\bar{X})(Y_t-\bar{Y})}{\sqrt{\sum_{t=1}^{T}(X_t-\bar{X})^2}\sqrt{\sum_{t=1}^{T}(Y_t-\bar{Y})^2}}$$

It does not take into account the dynamics of the phenomena, but the interaction between them at the same time. The parameter β acts here as a dependency weight. The higher the value of the parameter, the greater the distance between negatively correlated series and the smaller the distance between positively correlated series (Fig. 2).

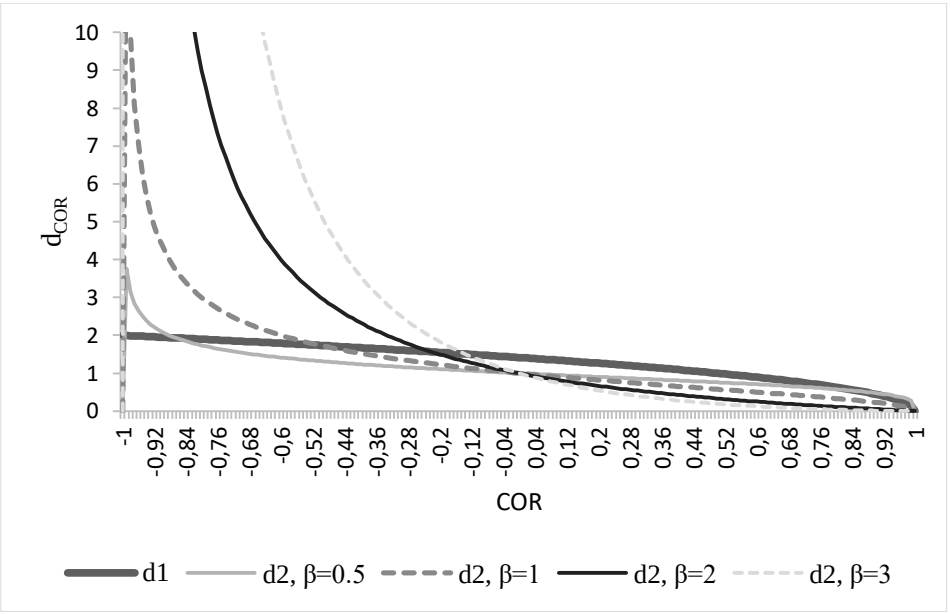


Fig. 2. The value of d_{COR} dependent on β value

Source: Own studies based on: Montero and Vilar [2014].

Measures based on direct observations of series reflect real phenomena but, on the other hand, often enforce equal lengths and frequencies of observed series at analogous points in time with the exception of Frechet distance and DTW, but these have their drawbacks too.

Subsequent measures based on time series representations do not require the researcher to have the same length of observed time series or the same frequency. Of course, they should refer to the same research period. One could say that they are Euclidean distances determined on unambiguous time series representations. In the case of the ACF and PACF functions, we focus on the characteristics of the stochastic processes represented by the time series under study. That is, by estimating the distance in this way, we can afford to compare any length of time series for a predetermined order L [Galeano and Peña, 2000]:

$$d_{ACF}\left(X_{T_1}, Y_{T_2}\right) = \sqrt{\left(ACF_X(L) - ACF_Y(L)\right)' \Omega \left(ACF_X(L) - ACF_Y(L)\right)} \quad (7)$$

$$d_{PACF}\left(X_{T_1}, Y_{T_2}\right) = \sqrt{\left(PACF_X(L) - PACF_Y(L)\right)' \Omega \left(PACF_X(L) - PACF_Y(L)\right)} \quad (8)$$

where Ω is a matrix of weights.

If Ω = I, the measure (7), (8) becomes the Euclidean distance between the estimated autocorrelation or partial autocorrelation functions. Usually Ω involves geometric weights decaying with the autocorrelation lag.

Periodograms are another representation of the series. They enable an unambiguous mapping of the series to the frequency interval [0, 0.5] of the magnitude of the individual amplitudes. This representation also allows the comparison of series of different lengths and the representation is not limited by any additional parameter. In the context of stochastic process representation, however, the periodogram is treated as a biased estimator of the spectral density. In this paper, distance based periodograms (9), normalised periodograms (10) and normalised and logarithmised periodograms (11) are considered [Caiado, Crato and Peña, 2006]:

$$d_P\left(X_{T_1}, Y_{T_2}\right) = \frac{1}{n}\sqrt{\sum_{l=1}^{n}(I_X(\omega_l) - I_Y(\omega_l))^2} \tag{9}$$

$$d_{NP}\left(X_{T_1}, Y_{T_2}\right) = \frac{1}{n}\sqrt{\sum_{l=1}^{n}(NI_X(\omega_l) - NI_Y(\omega_l))^2} \tag{10}$$

$$d_{LNP}\left(X_{T_1}, Y_{T_2}\right) = \frac{1}{n}\sqrt{\sum_{l=1}^{n}(logNI_X(\omega_l) - logNI_Y(\omega_l))^2} \tag{11}$$

where $\omega_k = \frac{2\pi l}{T}$, $l=1,\ldots,n$,with $n=\frac{T-1}{2}$.

3. Analysis

Twenty two time series of daily electricity prices [EUR/MWh] quoted on the Nord Pool platform over the period 10.02-04.10.2021 were used to compare the distance measures discussed earlier. Among them, Poland (PL), Sweden (SE1, SE2, SE3, SE4), Finland (FI), Denmark (DK1, DK2), Norway (N1, N2, N3, N4, N5, N6), Estonia (EE), Lithuania (LV), Latvia (LT), Austria (AT), Belgium (BE), Germany and Luxemburg (DE-LU), France (FR), the Netherlands (NL) were included.

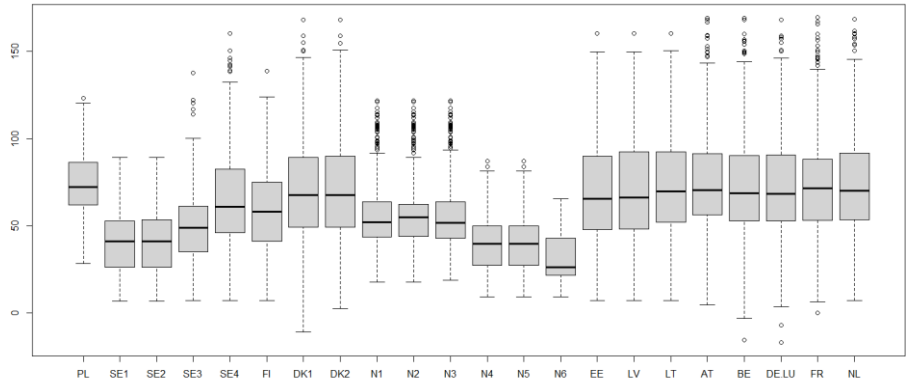


Fig. 3. Prices distributions

In the analysed series, differences in the distributions in both level, variation and skewness can be observed (Fig. 3).

Based on the distances discussed earlier (from formula 1 to 11), the analysed time series representing standardised electricity prices in 22 countries were grouped into two (Table 1) and four (Table 2) clusters. Every column in the table represents the result of clustering using respective distance formulas. Silhouette index values are shown in last row.

Table 1. Division into two groups

Distance measures / Market	d_E (1)	d_F (2)	d_{DTW} (3)	d_{CORT} E (4a)	d_{CORT} F (4b)	d_{CORT} DTW (4c)	$d_{COR.1}$ (5)	$d_{COR.2}$ (6)	d_{ACF} (7)	d_{PACF} (8)	d_P (9)	d_{NP} (10)	d_{LNP} (11)
PL	1	1	1	1	1	1	1	1	1	1	1	1	1
SE1	2	2	2	1	1	2	1	1	2	2	1	2	1
SE2	2	2	2	1	1	2	1	1	2	2	1	2	1
SE3	1	2	1	2	2	1	1	1	2	2	1	1	1
SE4	1	2	1	2	2	1	2	2	2	2	1	2	1
FI	1	2	1	2	2	1	1	1	2	2	1	1	1
DK1	1	2	1	2	2	1	2	2	2	2	1	1	1
DK2	1	2	1	2	2	1	2	2	2	2	1	1	2
N1	1	2	1	1	1	1	1	1	1	1	1	1	1
N2	1	2	1	1	1	1	1	1	2	1	1	1	1
N3	1	2	1	1	1	1	1	1	1	1	1	1	1
N4	2	2	2	1	1	2	1	1	2	2	1	2	1
N5	2	2	2	1	1	2	1	1	2	2	1	2	1
N6	2	2	2	1	1	2	1	1	1	2	1	1	1
EE	1	2	1	2	2	1	1	1	1	2	1	1	1
LV	1	2	1	2	2	1	1	1	1	2	1	1	1
LT	1	2	1	2	2	1	1	1	1	2	1	1	1
AT	1	2	1	2	2	1	2	2	2	2	1	1	1
BE	1	2	1	2	2	1	2	2	2	2	2	1	1
DE-LU	1	2	1	2	2	1	2	2	2	2	1	1	1
FR	1	2	1	2	2	1	2	2	2	2	2	1	1
NL	1	2	1	2	2	1	2	2	2	2	1	1	1
Silhouette	**0.49**	0.27	**0.56**	**0.58**	**0.55**	**0.56**	0.23	**0.52**	0.44	0.19	0.35	0.31	0.1

The best result was obtained for the division into four country groups on the basis of DTW distance supported by the CORT similarity coefficient (Table 2, Fig. 6-7). It corresponds to a Silhouette index value equal 0.7. The second best result was achieved for the Euclidean distance and the third one for the correlation coefficient with β = 2. For the division into two groups of countries, the results are considerably weaker. In this grouping, the Euclidean distance supported by the CORT similarity coefficient gave the best grouping quality (Table 1, Fig. 4-5) with a Silhouette index value of 0.58. Countries were usually classified into analogous groups.

Table 2. Division into four groups

Distance measures / Market	d_E (1)	d_F (2)	d_{DTW} (3)	d_{CORT} E (4a)	d_{CORT} F (4b)	d_{CORT} DTW (4c)	$d_{COR.1}$ (5)	$d_{COR.2}$ (6)	d_{ACF} (7)	d_{PACF} (8)	d_P (9)	d_{NP} (10)	d_{LNP} (11)
PL	1	1	1	1	1	1	1	1	1	1	1	1	1
SE1	2	2	2	2	2	2	2	2	2	2	1	2	1
SE2	2	2	2	2	2	2	2	2	2	2	1	2	1
SE3	3	3	3	3	3	3	3	3	2	2	1	1	2
SE4	3	3	3	4	4	3	4	4	2	2	1	2	1
FI	3	3	3	3	3	3	3	3	2	2	2	1	1
DK1	4	4	4	4	4	4	4	4	3	2	1	1	1
DK2	3	3	4	4	4	3	4	4	3	2	1	1	3
N1	1	1	3	2	2	1	1	1	1	1	2	3	1
N2	1	1	3	2	2	1	1	1	4	1	1	3	1
N3	1	1	2	2	2	1	1	1	1	1	2	3	1
N4	2	2	2	2	2	2	2	1	2	2	1	2	1
N5	2	2	2	2	2	2	2	1	2	2	1	2	1
N6	2	2	2	2	2	2	1	1	1	2	2	4	1
EE	3	3	3	3	3	3	3	3	1	3	2	4	4
LV	3	3	3	3	3	3	3	3	1	3	2	4	4
LT	3	3	3	3	3	3	3	3	1	3	2	4	4
AT	4	3	4	4	4	4	4	4	2	2	1	1	1
BE	4	4	4	4	4	4	4	4	2	2	3	1	1
DE-LU	4	4	4	4	4	4	4	4	3	2	1	1	1
FR	4	4	4	4	4	4	4	4	2	2	4	1	1
NL	4	4	4	4	4	4	4	4	3	4	1	1	1
Silhouette	**0.64**	0.17	0.39	**0.51**	0.47	**0.7**	0.35	**0.6**	0.32	0.2	0.29	0.29	0.09

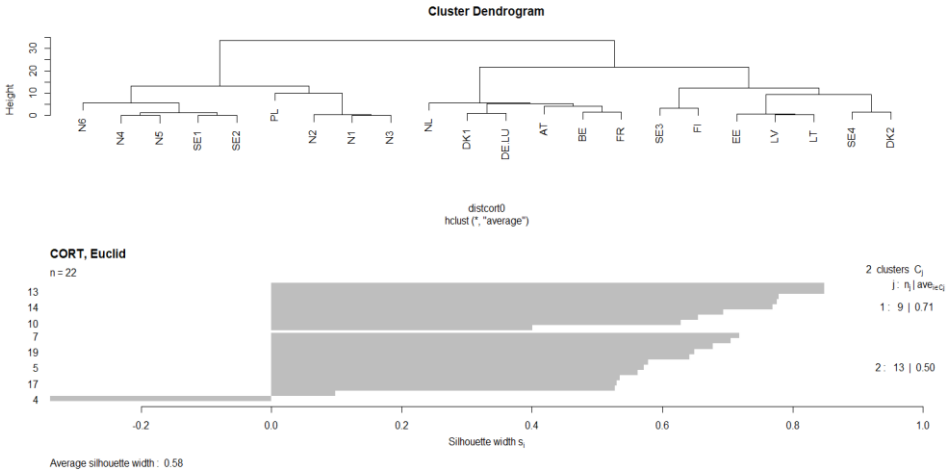


Fig. 4. The best result of separate time series of electric energy prices on two group

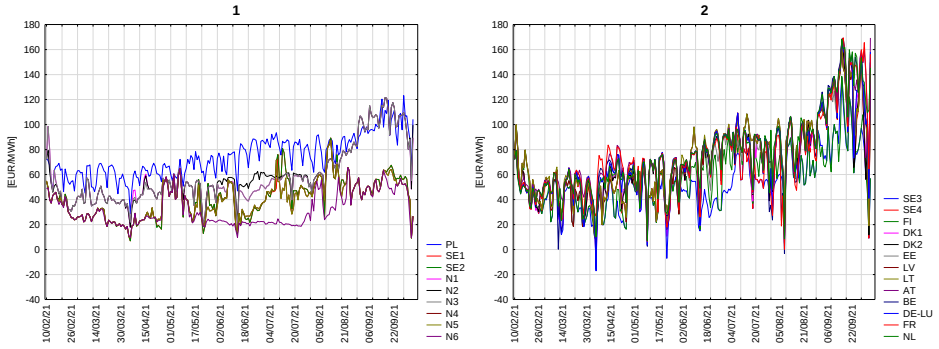


Fig. 5. Two groups of time series of electric energy prices obtained by Euclidean distance with CORT

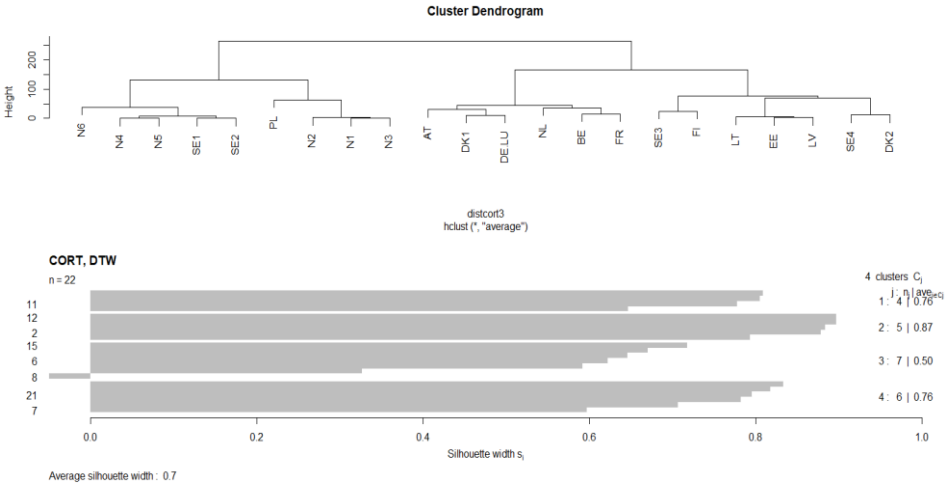


Fig. 6. The best result of separate time series of electric energy prices on four group

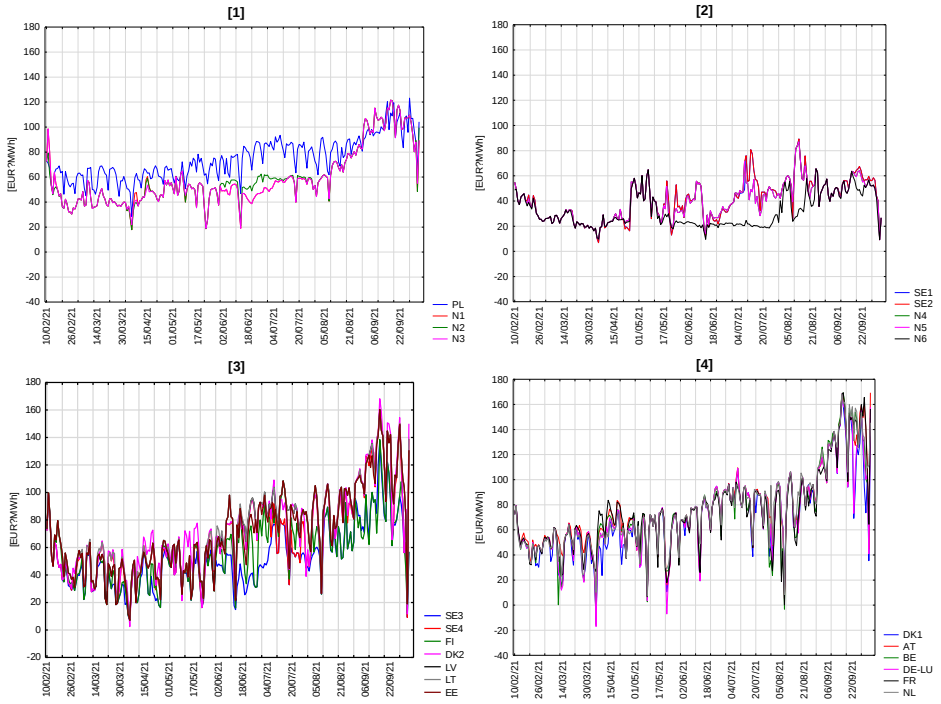


Fig. 7. Four groups of time series of electric energy prices obtained by DTW with CORT

4. Conclusion

The best result was obtained for DTW distance measure with the CORT similarity index and four groups. Classification results are improved by adding CORT similarity index to the distance measure.

Substantially weaker results were obtained for the ACF, PACF functions and the periodogram representation. This indicates a considerable loss of information regarding the variation of the phenomena under study over time, which unfortunately does not support the usefulness of these measures in assessing the similarity of time series measured with different observation frequencies.

References

Berndt D.J., Clifford J. (1994), *Using Dynamic Time Warping to Find Patterns in Time Series*, KDD Workshop, pp. 359-370.

Caiado J., Crato N., Peña D. (2006), *A Periodogram-Based Metric for Time Series Classification*, "Computational Statistics & Data Analysis", Vol. 50(10), pp. 2668-2684.

Douzal-Chouakria A., Nagabhushan P.N. (2007), *Adaptive Dissimilarity Index for Measuring Time Series Proximity*, "Advances in Data Analysis and Classification", Vol. 1(1), pp. 5-21.

Frechet M.M. (1906), *Sur Wuelques Points du Calcul Fonctionnel*, "Rendiconti del Circolo Matematico di Palermo (1884-1940)", Vol. 22(1), pp. 1-72.

Galeano P., Peña D. (2000), *Multivariate Analysis in Vector Time Series*, "Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo", Vol. 4(4), pp. 383-403.

Golay X., Kollias S., Stoll G., Meier D., Valavanis A., Boesiger P. (2005), *A New Correlation Based Fuzzy Logic Clustering Algorithm for fMRI*, "Magnetic Resonance in Medicine", Vol. 40(2), pp. 249-260.

Montero P., Vilar J.A. (2014), *TSclust: An R Package for Time Series Clustering*, "Journal of Statistical Software", November, Vol. 62, Iss. 1, pp. 1-43.

Sankoff D., Kruskal J.B. (1983), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley Publishing Company.

Chapter V

A multivariate functional analysis of mortality trends in Europe

Justyna Majewska

1. Introduction

Mortality data provide information on the current demographic situation of the population and explain its demographic future. In addition to their role in demographic accounting, mortality data serve as important indicators of progress or of socio-economic and health problems. They provide an overview of progress in one of the areas of greatest human concern – life expectancy and the prevention of premature death.

Mortality data show consistent patterns of risk in specific populations and trends in specific causes of death over time. They are also sensitive indicators of differences within the population and can help to identify target groups for special programmes in health care and development. The analysis of trends is still important in the forecasting of mortality. The trends observed in the past will determine the method and the historical period to be used.

A variety of mortality measures are commonly used to monitor trends and explore patterns within and between populations. There is a lack of a single perspective for understanding and interpreting mortality trends. Primary measures are changes over time and absolute and relative differences between countries and groups (geographical, gender, ethnic, socio-economic).

The analysis of the evolution of mortality in one or more populations involves a choice between a wide range of mortality indicators and a focus either on global mortality or on a specific component. The literature in this area is vast. In most cases, analysis of mortality trends is carried out with a focus on summary measures (e.g. life expectancy at birth, life expectancy gap) [Vaupel, Zhang and van Raalte, 2011; van Raalte, Sasson and Martikainen, 2018; Amin and Steinmetz, 2019]. In other cases, researchers focus on specific components of mortality without considering the global pattern [Medford et al., 2019; Kannisto, 2001; Zanotto, Canudas-Romo and Mazzuco, 2020].

However, there is little work that looks at mortality trends using a multidimensional approach. Research tends to focus on cluster analysis. Meslé, Vallin and Andreyev [2002] have already tested a clustering solution for several European countries based on their age-specific probability of dying and have found significant differences in the life expectancies and age structures of eastern and western countries. Debón et al. [2017] grouped EU countries using fuzzy c-means cluster analysis of mortality surfaces, with similar results. Léger and Mazzuco [2021] used functional data analysis to identify the role played by all mortality components, and analyzed whether there were different patterns of mortality decline among low-mortality countries.

Multivariate trend analysis results are used in mortality projections. In some European countries, information on trends in other countries is directly included in the projection (e.g. Poland assumes a 'catch-up' with developed countries in 21-22 years). In the case of stochastic multi-population mortality models, it is important to indicate which countries have similar patterns and should be included in the cohort mortality projection model [e.g. Li and Lee, 2005; Hyndman, Booth and Yasmeen, 2013].

In order to provide a comparative framework, in this paper we treat mortality rates as functional data and examine the evolution of age-specific mortality in European countries since 1960. We use functional cluster and principal component analysis due to the representation of mortality rates as functions. Changes in mortality profiles provide information on whether countries are evolving in the same way (i.e. following the same cluster sequence) or whether there are different patterns, as they are based on the membership of the population in a cluster at a given point in time. We focus on the results for the countries of Central and Eastern Europe, looking at the lag in the mortality patterns of the countries with the lowest mortality rates. Functional principal component analysis is also applied to each group's characteristics, providing a continuous framework for interpretation and comparison.

The structure of the study is as follows. In the second section we emphasize the importance of an understanding of mortality patterns through the lens of demographic transitions. Functional methods are described in the third section. The results of the empirical analysis are presented in the fourth section.

2. Understanding mortality patterns through the perspective of the demographic transitions

Demographic transition theory describes typical mortality trends by age, leading to increased life expectancy over time [United Nations, 2011]. Historically, populations have tended to shift from high fertility and high mortality to low fertility and low mortality over time. This process is referred to as demographic transition.

Pre-transition societies are characterized by high fertility − on average more than 5 or 6 children per woman − and during the transition period, fertility declines towards the replacement level, i.e. 2.1 children per woman or even lower. Pre-transition societies have high mortality rates in all age groups, but as the transition progresses, mortality rates decline, first among children and gradually among adults.

Figures 1-2, which show estimated levels and trends in life expectancy at birth[1] over the past 60 years for selected groups of countries or regions, partially illustrate the decline in mortality that has occurred because of the demographic transition.

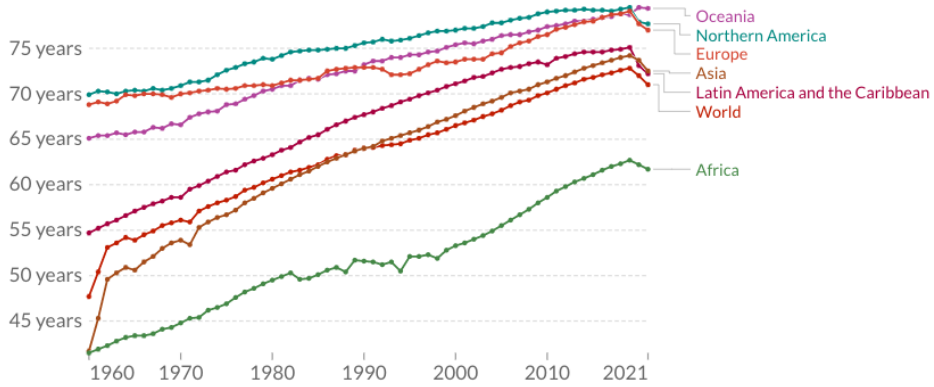


Fig. 1. Period life expectancy at birth from 1960 to 2021 – geographic regions

Source: United Nations [2022].

The trends in life expectancy of the world's population are well known. Asia is the region with the largest increase in life expectancy over the period studied. Africa is the region with the lowest life expectancy at birth. More interesting, however, is the case of Europe. There are significant differences between

[1] The average number of years that a newborn could expect to live, if he or she were to pass through life exposed to the sex- and age-specific death rates prevailing at the time of his or her birth, for a specific year, in a given country, territory, or geographic area (www 1).

Western and Eastern and Central European countries. In the former, life expectancy at birth was over 66 years in the early 1960s, and growth in survival has been somewhat slower than in many of the other groups of countries shown in Fig. 1, with life expectancy exceeding 80 years before 2010.

Trends in life expectancy in Eastern Europe differ markedly from the average for the more developed regions (Fig. 2). Life expectancy at birth in Eastern Europe increased from around 66 to 70 years between the early 1960s and the late 1970s, but then stagnated and even declined slightly at various times over the following decades.

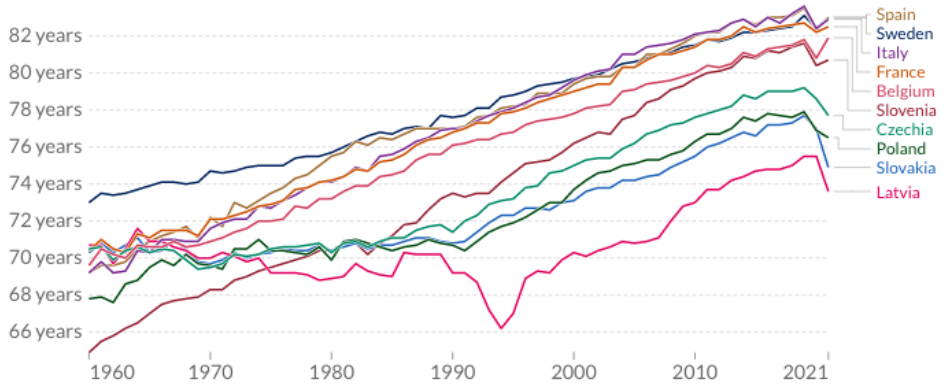


Fig. 2. Period life expectancy at birth, Eastern Europe from 1960 to 2021 – selected European countries

Source: United Nations [2022].

Improvements in life expectancy are almost always the result of advances in both child and adult survival, but the relative contribution of the different age groups changes with the stage of the demographic transition (Fig. 3). In populations with low life expectancy at birth in the early stages of their demographic transition, the proportion of progress in longevity due to improvements in child survival tends to exceed that due to improvements in adult survival. As life expectancy at birth increases, the marginal improvement contributed by progress in survival in each age group shifts to older ages. Overall, all regions have achieved improvements in survival at all ages, but in some regions or groups of countries certain age groups have played a special role. In regions where life expectancy at birth was low in the 1960s (e.g. Africa, Oceania, Asia), most of the gains in survival rates in recent years (but before 2020) have been achieved by reducing mortality below the age of 15. The pattern of improvement was quite different in regions that were more advanced in their demographic transition, represented by the more developed regions (excluding Eastern Europe). In these regions, sur-

vival beyond the age of 60 accounted for a much larger share of the total gains in life expectancy over the past half century. Despite having reached the age of 60 in the 1960s, the pattern of age contributing to longevity improvement in Eastern Europe differs markedly from other countries in more developed regions. Only a small part of the overall increase in life expectancy can be attributed to the decline in mortality above the age of 60, as indicated by the United Nations [2011]. As mortality among adults aged 30-59 increased in Eastern Europe during this period, these age groups contributed to the loss in life expectancy, while improvements in survival among children under five accounted for almost all the increase in life expectancy at birth.

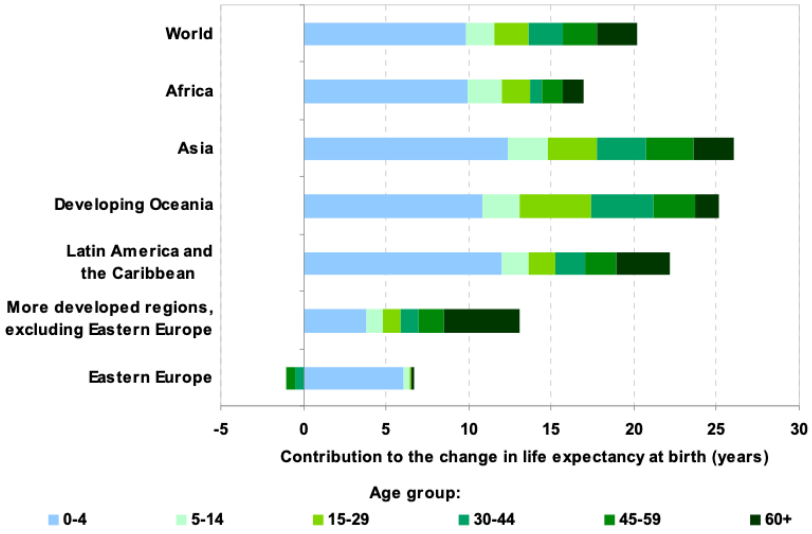


Fig. 3. Contribution of age-specific mortality decline to the change in life expectancy at birth between 1950-1955 and 2005-2010 for the world and selected regions

Source: United Nations [2011].

Although there may be deviations from the usual pattern of mortality decline [Caselli, Mesle and Vallin, 2002], the typical course of the demographic transition is characterized by declines in mortality across all age groups. In the early stages of the transition, the decline in child mortality dominates improvements in survival, and in the advanced stages of the transition, the decline in older adult mortality becomes increasingly dominant.

Declining fertility and mortality rates are leading to a third key feature of the demographic transition: population ageing. At the onset of this transition, large birth cohorts combined with a low probability of survival to old age means that a larger proportion of the population is concentrated in younger age groups

than in older ones. As birth cohorts shrink relative to the size of their parents' generation, and as longevity increases and survival to old age becomes more common, the proportion of children in the population decreases and the proportion of older age groups increases. Declining mortality rates among young people, combined with an ageing population, result in an increasing concentration of deaths at older ages as the population undergoes a demographic transition. Again, Eastern European countries deviate from the typical pattern of demographic transition, with a lower proportion of deaths among children under five (1%).

In order to group similar mortality trends, it is essential to know the changes in overall and age-specific mortality patterns that characterize the demographic transition. This is usually combined with the changing pattern of distribution of deaths by broad cause groups, a feature of the epidemiological transition model, to provide a complete picture.

3. Functional methods

3.1. Functional data

Let $y(t_1), \dots, y(t_N)$ denote age-specific mortality data (mortality rates or life-table death counts) at ages $t_1, \dots, t_N$, which can be single years of age or 5-years-old age groups. A functional approach assumes that the discrete observations come from a continuous underlying function $x(t)$ defined on $t \in [0, T]$.

For observations at the same instants on a common interval, functional data consists of a set of n curves denoted as $x_i(t_j)$, with $t_j \in [0, T], j = 1, \dots, N$, $i = 1, \dots, n$:

$$y_{ij} = x_i(t_j) + \varepsilon_{ij} \tag{1}$$

where the error term ε_{ij} contributes to the roughness of the raw data. The curves are assumed to be independent realizations drawn from the same continuous stochastic process $X(t)$ belonging to $L_2[0, T]$ space.

The first step in analyzing functional data is to reproduce the functional form from discrete data. The basis function system is used, which is a set of known functions that are independent on each other and that can arbitrarily approximate any function. Formally, let's consider p known basis functions $\psi(t) = (\psi_1(t), \dots, \psi_p(t))$. The basis function procedures represent the function $X(t)$ by a linear expansion:

$$X(t) = \sum_{j=1}^{p} \gamma_j \psi_j(t) \tag{2}$$

where $\gamma = (\gamma_1, \dots, \gamma_p)'$ are the basis function coefficients to be estimated by the ordinary least squares method minimizing the sum of squared residuals.

B-spline functions are the most common choice for non-periodic functional data. In practice, the interval over which the function is to be approximated is divided into L subintervals separated by values τ_l, with $l = 1, \dots, L-1$, that are called knots. Over each subinterval, a spline is a polynomial of specified order *m*, and adjacent polynomials join up smoothly at the knots.

In order to capture the structural component of the data and reduce the noise in the data, the underlying functions x_{ij} must be smooth. There are many ways to control the irregularity of the curve and obtain a better approximation. Regression splines use the number of knots as a control parameter. The more knots used, the smoother the curve will be. The knots are chosen to be equally spaced or placed at the quantiles of the distribution in many applications.

It is convenient to use smoothing splines. These introduce a roughness penalty term into the objective function. The integrated squared second derivative is a natural measure of a function's roughness. This becomes the penalized least squares estimation criterion:

$$PSSE_\lambda(x_i(t)|y) = \sum_{j=1}^{N}[y_{ij} - x_i(t)]^2 + \lambda \int [D''(x_i(t))]^2 dt \qquad (3)$$

where $x_i(t) = \sum_{j=1}^{p} \gamma_{ij}\psi_j(t)$ is the basis expansion of each curve, and y_{ij} with $j = 1, \dots, N$ are discrete observations for the *i*-th curve.

The smoothing parameter λ in Eq. (3) controls the trade-off between the closeness of fit to the average of the data and the variability of the curve and is commonly chosen subjectively or selected through the generalized cross-validation criterion.

Other smoothing techniques have also been developed in mortality analysis to improve the accuracy of projections. Hyndman and Ullah [2007] used penalised regression splines with a partial monotonic constraint to smooth log mortality rates. P-splines smoothing, which combines (fixed-knot) B-splines with a roughness penalty, was used by Camarda [2012].

3.2. Functional cluster analysis

Clustering functional data is generally a difficult task due to the nature of the functional data itself (belonging to an infinite dimensional space). Some common problems include not defining probability density, defining distances between curves and estimating from noisy data. To overcome these problems, several methods have been developed, which may be classified into three main approaches [Jacques and Preda, 2014]: two-step clustering, nonparametric clustering (also called distance-based clustering) and model-based clustering. The last one approach is described below for the purposes of analysis.

A model-based approach constructs homogeneous clusters by means of a density mixture model and allows the prediction of membership of each observation to one of the clusters. Conditional to the membership of a cluster, the observations are supposed to come from a common distribution with cluster-specific parameters. In the finite dimensional setting, the main tool to estimate the model is the multivariate probability density. In the case of functional data, the probability density is not defined, so a density probability on the parameters describing the curves need to be assumed. The first model-based clustering method for functional data was developed by James and Sugar [2003].

Formally, let $Z = (Z_1, \dots, Z_K) \in \{0,1\}^K$ be an unobserved random variable indicating the group membership of $x(t)$: Z_K is equal to 1 if Z belongs to the k-th group, and 0 otherwise. The clustering task aims to predict the value $z_i = (z_{i1}, \dots, z_{iK})$ of Z for each observed curve $x_i(t)$. Each curve x_i can be summarized by its basis expansion coefficient vector γ_i, as defined in Eq. (2), whose distribution is assumed to be a mixture of Gaussians with density:

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; \mu_k, \Sigma_k) \tag{4}$$

where ϕ is the Gaussian density function and $\pi_k = P(Z_K = 1)$ the prior probability of group k.

Other distributions can be used, but in finite mixture models Gaussian densities are by far the most used, as they can reasonably approximate a wide class of probability distributions. This model is referred to as the functional latent mixture model FLM by Bouveyron and Jacques [2011] because it can be reparametrized to represent the curves through their group-specific eigenspace projection. The spectral decomposition of the matrix Σ_k allows the modelling and interpretation of the variance of the data of the k-th group through the parameters $a_{k1}, \dots, a_{kd_k}$ and the variance of the noise through parameters b_k, where d_k can be considered as the intrinsic dimension of the latent subspace of the k-th group, and Q_k is the matrix containing the basis expansion coefficients of the eigenfunctions ($FLM_{[a_{ki} b_k Q_k d_k]}$). In contrast to the two-stage methods, where the estimation of these parameters is done before clustering, the two tasks are performed simultaneously in this approach.

3.3. Functional principal component analysis

Using principal component analysis to investigate mortality is not new and has been used with parameter estimation proposals in mortality prediction [Lee and Carter, 1992; Booth, Maindonald and Smith, 2002; Renshaw and Haberman, 2006; Hyndman and Ullah, 2007]. Functional Principal Component Analysis (FPCA) extends traditional multivariate PCA to functional data.

FPCA is a way of looking at the covariance structure that can be much more informative and can be a complement to a direct examination of the variance-covariance function. The values of the variables in PCA are replaced by function values $x_i(t)$ in FPCA and the discrete index by the continuous index t. Given n functional observations $x_i(t)$ with $1 \leq i \leq n$ and $\bar{x}(t)$ as the estimate of the mean function, the estimated covariance function, analogous with the covariance matrix in the multivariate case, is defined as:

$$S(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t)) \quad (5)$$

The spectral decomposition performs the task of finding the most important modes of variation in the covariance or correlation matrix of the curves. It provides a countable set of positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ associated with a basis expansion of orthonormal basis functions $\phi_l(t)$ with $l = 1, \ldots$ such that:

$$S(s,t) = \sum_{l=1}^{\infty} \lambda_l \phi_l(s)\phi_l(t) \quad (6)$$

The basic functions $\phi_l(t)$ are the eigenfunctions or harmonics and define the most important modes of variation in the curves and are orthogonal of each other. The eigenvalues measure the variability in the directions corresponding to the eigenfunctions.

The projection of $x_i(t)$ in the direction of the eigenfunctions $\phi_l(t)$ provides us with the functional principal components, a set of zero-mean linearly uncorrelated random variables, defined on the same interval of the functional data, with variance λ_l. As $x_i(t)$ and $\phi_l(t)$ are functions, summations of variables in the multivariate context are replaced by integrations over t to define an inner product. The principal component scores of the i-th curve are defined as:

$$c_{i,l} = \int x_i(t)\phi_l(t)dt \quad (7)$$

The decomposition of Karhunen-Loève allows the expression of the curve through its functional principal component expansion:

$$x_i(t) = \sum_{l=1}^{\infty} c_{i,l}\phi_l(t) \quad (8)$$

The FPCA gives with a group of basic functions $\phi_1(t), \ldots, \phi_l(t)$ and returns functional data as a linear combination of the new basis functions, where the coefficient of the $\phi_l(t)$ is the estimated score of the l-th principal component of the corresponding curve. The decomposition of Karhunen-Loève facilitates the dimension reduction in that if the first q terms (for a large enough q) provide a good approximation to the infinite sum, the information contained in the curve $x_i(t)$ is essentially synthesized by the q-dimensional vector $c = (c_{i1}, \ldots, c_{iq})$ and one can work with this approximation.

The eigenfunctions allow the identification of the main directions of variability in the complete mortality profile with respect to the mean curve, and the corresponding scores for each curve can be used to characterize the countries in the clusters in a reduced dimensional space.

4. Multivariate analysis of mortality trends

4.1. Data

The data source is the Human Mortality Database [2023]. This database provides high quality and quantity data on mortality profiles for many European and some non-European countries over a wide range of years. For the purpose of this paper, we selected 20 European countries, excluding those whose time series are considered too short. The data range from 1960 to 2019. For most countries, this was the most recent year available.

We examined life table death counts d_x, where the life table radix (i.e., the population experiencing 100,000 births per year) was fixed at 100 000 at age 0 for each year. This means that for each combination of country and year we have constructed a curve of the mortality pattern for the ages from zero to 110 years.

The distribution of mortality by age was used, following the work of Léger and Mazzuco [2021], because one of the most prominent changes in mortality patterns in developed countries over the last few decades is the shift in the age at death mode for adulthood [see, e.g., De Beer and Janssen, 2016] and the compression of mortality above this mode [Thatcher et al., 2010]. Due to the different mortality trends in the past, analyses were performed separately for men and women.

4.2. Smoothing the mortality rates

Due to including Eastern European populations in the analysis and having noisy data in the early 1960s (which is usually associated with poor data quality), the data needs to be smoothed. To obtain a smoothed representation of the data, we used a basis expansion of B-splines. The roughness penalty method described in section 3.1 has been used because it allows us to control the smoothness of the data. Following the work of Léger and Mazzuco [2021] we employed the same set of knots for every curve so that the estimation of the splines coefficients was performed on the same age intervals. To maintain the data structure a sequence of 31 unequally distributed knots – one every three

months over the age interval [0, 2] and one every 5 years over the age interval [2, 110] – was applied. The smoothing parameters were selected through the generalized cross-validation criterion, i.e., a mean-squared error based measure, twice discounted by a term taking into account the number of parameters and the magnitude of the smoothing parameter. When conducting clustering analysis, it is good to have the same set of knots for all the curves. Non-smoothing mortality curves for male and female separately are shown in the Fig. 4 and Fig. 5. An example of smoothing curves for a specific year is presented in the following Fig. 6 and Fig. 7.

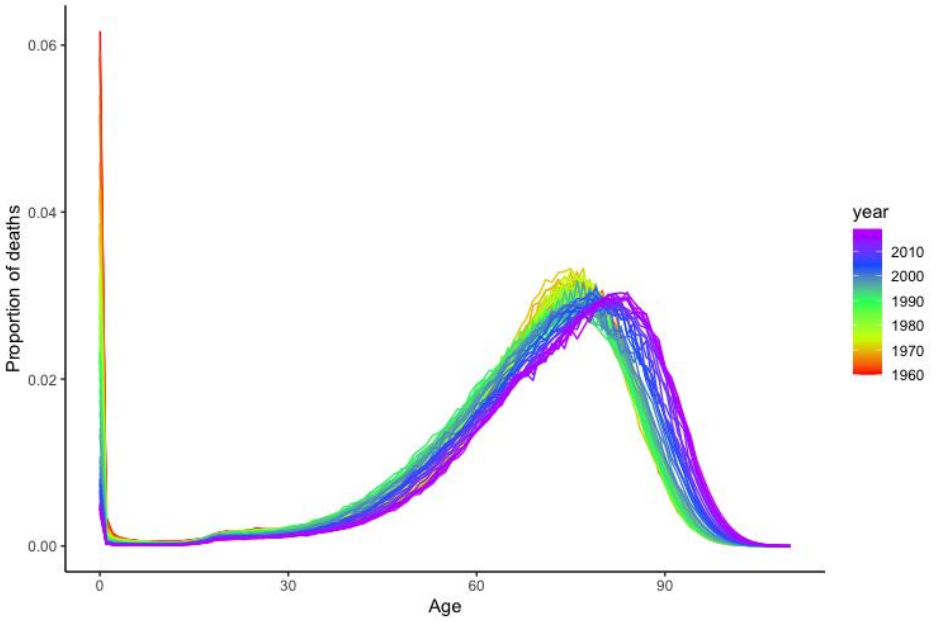


Fig. 4. Life table death counts from 1960 to 2019 in a single year group for males in Poland (curves are ordered chronologically, the oldest years are shown in red and the most recent in purple)

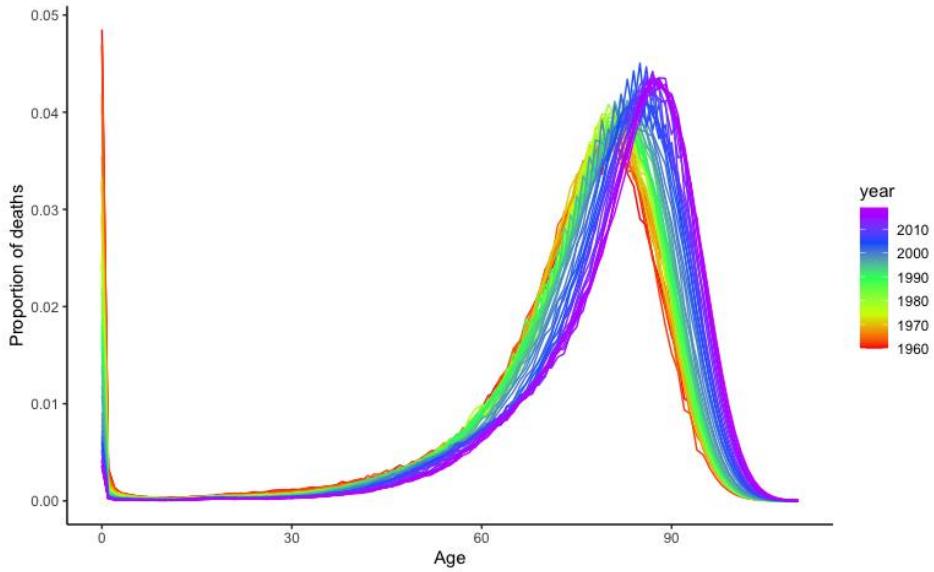


Fig. 5. Life table death counts from 1960 to 2019 in a single year group for females in Poland (curves are ordered chronologically, the oldest years are shown in red and the most recent in purple)

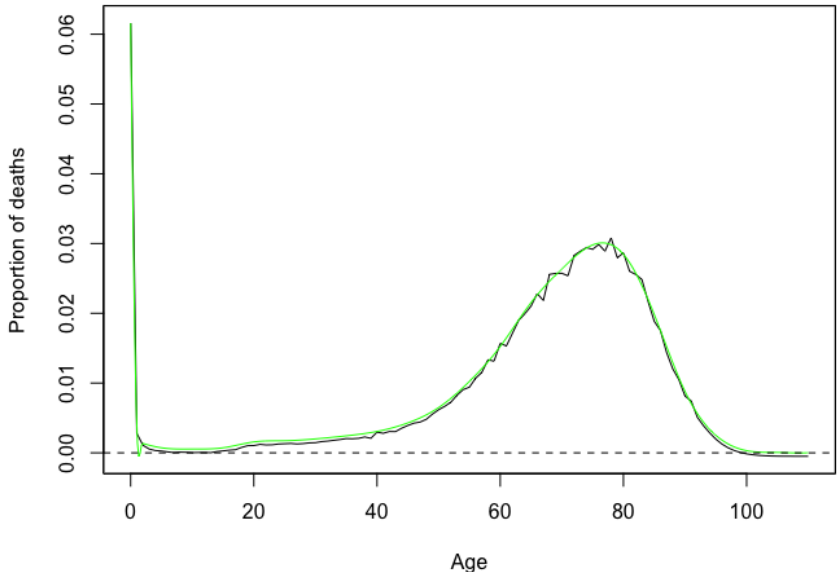


Fig. 6. The smoothing of the curve for Poland (male) in 1960 is shown with a sequence of 31 knots. Each curve is smoothed with a specific parameter λ

Source: Package fda in R [Ramsay, Hooker and Graves, 2010].

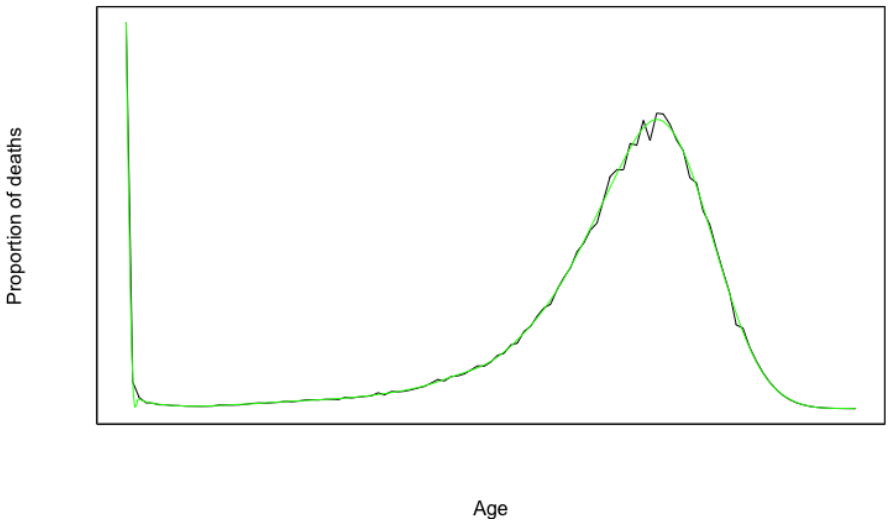


Fig. 7. The smoothing of the curve for Poland (female) in 1960 is shown with a sequence of 31 knots. Each curve is smoothed with a specific parameter λ

Source: Package fda in R [Ramsay, Hooker and Graves, 2010].

4.3. Clusters of mortality trends

Using cluster analysis, countries are grouped by year for each gender based on differences in smoothed curves. The age-specific mortality curves are grouped into clusters in such a way that they are as similar as possible within the same cluster and as dissimilar as possible in different clusters. The number of groups was selected based on the Bayesian information criterion (BIC). BIC for female and male subpopulations for a possible number of groups of 2 to 5 with the complexity of the model (the number of parameters) are presented in Table 1 (male) and Table 2 (female).

Table 1. Model-based clustering, men subpopulation: the BIC values and model complexity for the choice of the number of clusters

Number of clusters *k*	BIC	Complexity
2	−357 039.5	166
3	−534 999.4	296
4	−408 624.7	426
5	−326 152.3	556
6	−163 912.9	687
7	−101 513.9	817
8	−78 983.5	947

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011].

Table 2. Model-based clustering, female subpopulation: the BIC values and model complexity for the choice of the number of clusters

Number of clusters *k*	BIC	Complexity
2	−197 949.7	134
3	−300 08.3	233
4	−119 390.2	395
5	−5670.6	494
6	−45 574.9	562
7	−81 231.9	754
8	−107 305.7	759

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011].

For female the lowest BIC occurred at $k = 5$. For the male subpopulation the choice is not clear. Therefore, solutions for $k = 4$, 5 and 6 have been tested. With the number of clusters 4, insufficient differences in infant mortality are apparent. In contrast, for 6 clusters two of them are very similar in description. Thus, in this study we present solutions for $k = 5$.

The mortality curves and corresponding mean curves within the clusters (Fig. 8 for male and Fig. 9 for female) allow us to distinguish very clearly those with a similar shape but different levels of infant mortality and those with a higher accidental and premature mortality.

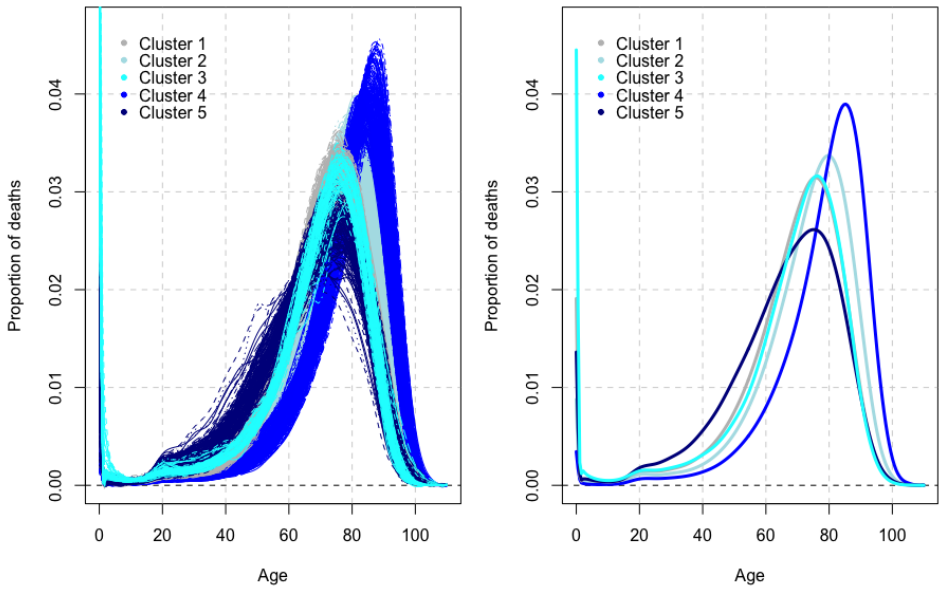


Fig. 8. Results of the model-based clustering on the men's mortality data: mortality curves (left), mean curves (right)

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011] & the way of presentation as suggested by Léger and Mazzuco [2021].

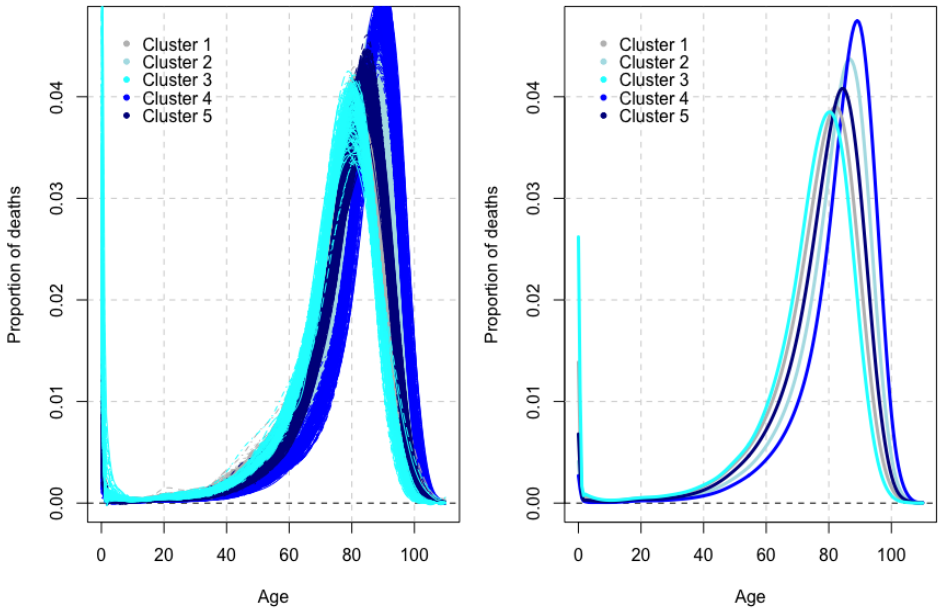


Fig. 9. Results of the model-based clustering on the women's mortality data: mortality curves (left), mean curves (right)

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011] & the way of presentation as suggested by Léger and Mazzuco [2021].

For both sexes, it can be seen the gradual shift toward older ages and the compression above the modal age at death is also clearly visible. The last cluster expresses a high level of premature mortality and a lower number of deaths around the modal age at death compared to other clusters.

Figures 9 and 10 show how mortality curves were classified in the clusters and allow one to follow the evolution of countries (rows) from 1960 to 2018 (columns).

For males, there was a decline in infant mortality in the Nordic, Western and Southern countries. This was followed by a shift in the curves and an increase in the number of deaths around the modal age of death throughout the period. Epidemiological transition is well known in the Nordic countries. The Finnish pattern of mortality (with an extremely high incidence of external causes of death) was already known [Saarela and Finnäs, 2008]. The Netherlands followed a pattern similar to the Nordic countries. Switzerland, France and Austria started slightly behind and moved more quickly to the next group. The southern European countries (Italy, Spain and Portugal) started even further behind, but also underwent a rapid transformation that brought Italy and Spain into the last cluster at the same time as Sweden and Norway. The analyses also identified the

higher infant mortality of Southern countries in the first twenty years. In the latter part of the decade, it appears that the differences have narrowed and that all countries have followed the above-mentioned process of shift and compression. The countries of central Europe reduced their high levels of infant mortality in the first decade of the period, but then had a lag of about 20 years in relation to the countries that preceded them. A long period of increased premature mortality was observed in Hungary and Poland. Poland, Slovakia and Estonia (Poland first) have reached the point where Scandinavia, Western Europe and Southern Europe reached around the turn of the century. This confirms the assumption that Poland lags the countries with the lowest mortality by about 20 years (CSO projections).

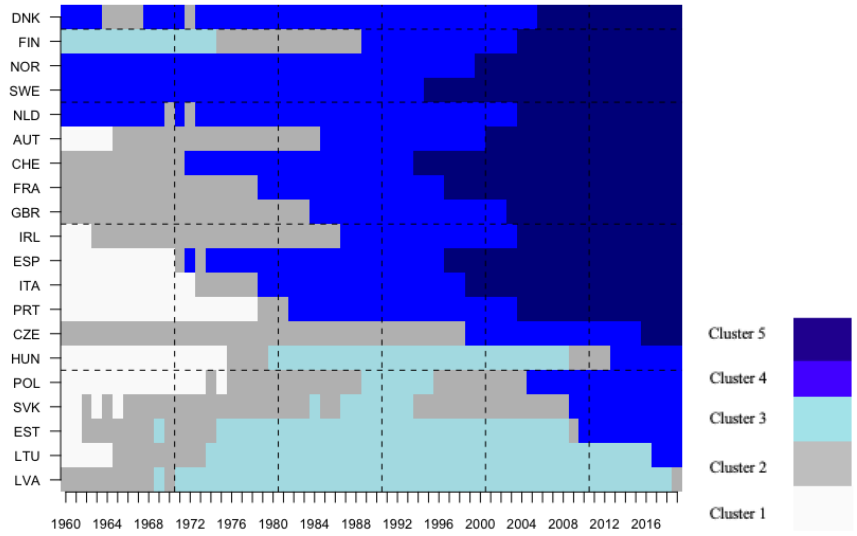


Fig. 10. Results of the model-based clustering on the men's mortality data: the composition of the 5 clusters

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011] & the way of presentation as suggested by Léger and Mazzuco [2021].

For females, over the period, countries experienced a continuous shift and compression of mortality curves towards older ages. Disparities between countries appeared to persist until the end of the period. This was because the transition to the clusters occurred in different years. For example, the shift occurred in the 1970s (cluster 3), during the 1980s (cluster 4) and at the beginning of the 1990s (cluster 5) in Norway, Sweden, Switzerland, France, Spain. Some gender-specific dynamics can also be observed, such as Estonia, which, at the end of the year under analysis, joined the group of countries where deaths are shifted to-

wards older ages. For Central and Eastern Europe, a long period of stagnation can be observed (cluster 2). This is followed by a shift and compression of the curves during the last decade (clusters 3 and 4). The Czech Republic, Poland and the Baltic countries appear to have made some progress (they end up in cluster 4).

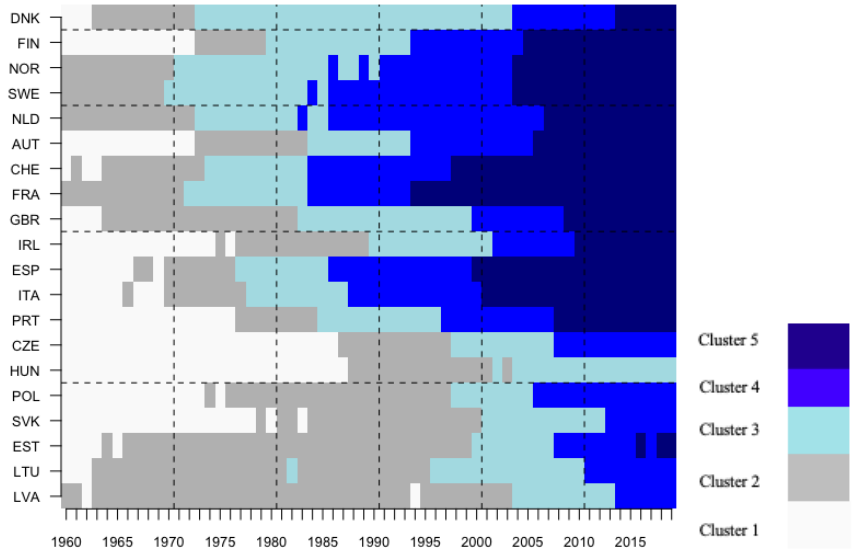


Fig. 11. Results of the model-based clustering on the women's mortality data: the composition of the 5 clusters

Source: Calculations in package funHDDC [Bouveyron and Jacques, 2011] & the way of presentation as suggested by Léger and Mazzuco [2021].

To conclusion, the analyses for the data on both the men and women showed a similar evolution for the Northern, Western and Southern European countries that was characterized by the shift of curves to older ages and by the concentration of adult mortality around the modal age at death. For these regions, we can conclude the existence of a common pattern of evolution. In the case of the men's data, all the countries belonged to the same group at the end of the period, supporting the hypothesis of an increasing homogeneity. The situation was more heterogeneous for the Central and Eastern countries because they did not experience the same evolution and, at the end of the period, they did not arrive at the same cluster.

From the FPCA, it emerged that most of the variability was explained by the first two principal components − 84% (men) and 93% (women) for the first principal component and 11% (men) and 5% (women) for the second principal component. Figures 12 and 13 show a solid curve for each of the first two principal components, which is the overall smoothed mean for the men and the func-

tions obtained by adding (+) to and subtracting (−) from the mean function an appropriate multiple of the eigenfunctions. Thus, the (+)/(−) curves represent the variation around the mean.

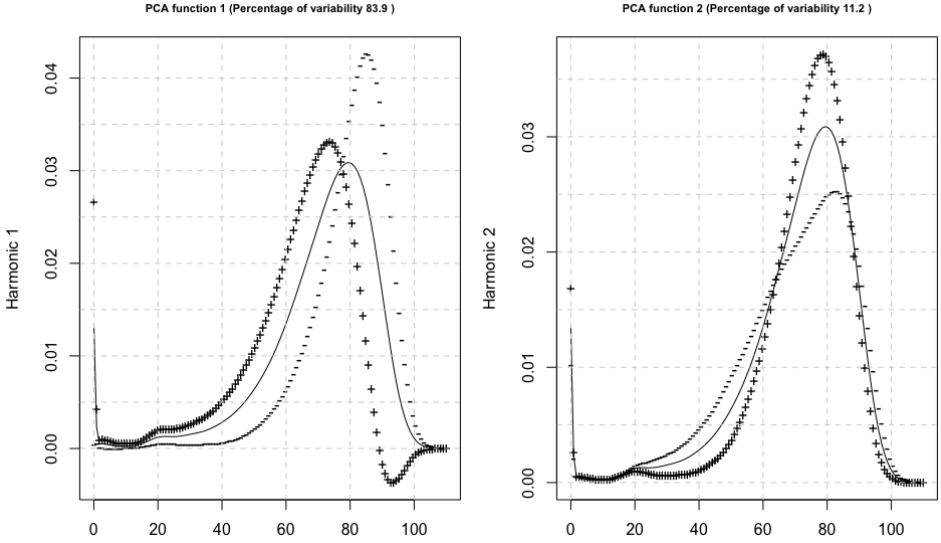


Fig. 12. Results of the FPCA on the men's mortality data: group means and effect of the components

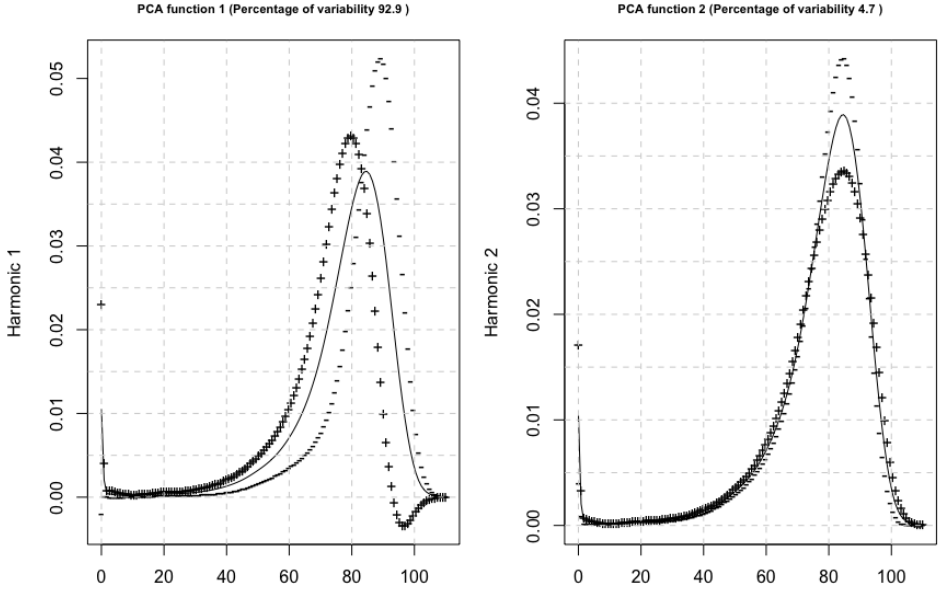


Fig. 13. Results of the FPCA on the women's mortality data: group means and effect of the components

Focusing only on Fig. 12 (left), since adding the first eigenfunction to the mean shifts the (+) curve to the left and subtracting the first eigenfunction from the mean shifts it to the right and compresses the (−) curve, it can be said that the first eigenfunction has the effect of shifting and compressing the overall mean across the age range. The curve of a country year with a large negative value of the first principal component will behave more like the (−) curve, while the curve of a country year with a large positive value of the first principal component will behave more like the (+) curve. Looking at Fig. 12 (right), we can see that the second eigenfunction has the effect of shaping premature mortality (ages 20-65) and adult mortality (ages 65-85), as the addition of the second eigenfunction to the mean decreases premature mortality and increases adult mortality, and the subtraction of the second eigenfunction from the mean increases premature mortality and decreases adult mortality. The curve of a country year behaves similarly to the (−)/(+) curve when the second principal component has a large negative/positive value. The fact that premature mortality and adult mortality are opposite can be seen by the fact that the (−) and (+) curves cross at around 65 years and move in opposite directions with respect to the mean curve. Since we are dealing with a distribution, deaths occurring at younger ages tend to avoid those occurring at older ages. Therefore, we can summarize: The first component is representative of the shift and compression of mortality distributions observed in recent decades, while the second component is related to premature mortality.

5. Conclusions

Analyses of the male data showed a similar evolution for the Northern, Western, Southern European countries, which was characterized by a shift in the curves to older ages and a concentration of adult mortality around the modal age at death. We can therefore conclude the existence of a common pattern of evolution. For male data, all countries belonged to the same group at the end of the period, supporting the hypothesis of increasing homogeneity. The situation was more heterogeneous for the middle Eastern countries, as they did not experience the same evolution and were not in the same cluster at the end of the period.

From the perspective of multi-population mortality modeling, the results of the analyses provide a basis for constructing a reference group of populations with similar mortality trends.

References

Abraham C., Cornillon P.A., Matzner-Løber E., Molinari N. (2003), *Unsupervised Curve Clustering Using b-splines*, "Scandinavian Journal of Statistics", Vol. 30(3), pp. 581-595.

Amin R.W., Steinmetz J. (2019), *Spatial Clusters of Life Expectancy and Association with Cardiovascular Disease Mortality and Cancer Mortality in the Contiguous United States: 1980-2014*, "Geospatial Health", Vol. 14, No. 1, pp. 139-145.

Booth H., Maindonald J., Smith L. (2002), *Applying Lee-Carter under Conditions of Variable Mortality Decline*, "Population Studies", Vol. 56(3), pp. 325-336.

Bouveyron C., Jacques J. (2011), *Model-based Clustering of Time Series in Group--specific Functional Subspaces*, "Advances in Data Analysis and Classification", Vol. 5(4), pp. 281-300.

Camarda C.G. (2012), *Mortalitysmooth: An R Package for Smoothing Poisson Counts with P-splines*, "Journal of Statistical Software", Vol. 50(1), pp. 1-24.

Caselli G., Mesle F., Vallin J. (2002), *Epidemiologic Transition Theory Exceptions*, "Genus", Vol. 58, pp. 9-52.

De Beer J., Janssen F. (2016), *A New Parametric Model to Assess Delay and Compression of Mortality*, "Population Health Metrics", Vol. 14(46).

Debón A., Chaves L., Haberman S., Villa F. (2017), *Characterization of Between-Group Inequality of Longevity in European Union Countries*, "Insurance: Mathematics and Economics", Vol. 75, pp. 151-165.

Ferraty F., Vieu P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer, Berlin.

Human Mortality Database (2023), University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), www.mortality.org.

Hyndman R.J., Ullah M.S. (2007), *Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach*, "Computational Statistics & Data Analysis", Vol. 51(10), pp. 4942-4956.

Hyndman R.J., Booth H., Yasmeen F. (2013), *Coherent Mortality Forecasting: The Product-Ratio Method with Functional Time Series Models*, "Demography", Vol. 50(1), pp. 261-283.

Jacques J., Preda C. (2014), *Functional Data Clustering: A Survey*, "Advances in Data Analysis and Classification", Vol. 8(3), pp. 231-255.

James G.M., Sugar C.A. (2003), *Clustering for Sparsely Sampled Functional Data*, "Journal of the American Statistical Association", Vol. 98(462), pp. 397-408.

Kannisto V. (2001), *Mode and Dispersion of the Length of Life*, "Population: An English Selection", Vol. 13(1), pp. 159-171.

Lee R.D., Carter L.R. (1992), *Modeling and Forecasting U.S. Mortality*, "Journal of the American Statistical Association", Vol. 87(41), pp. 659-671.

Léger A.E., Mazzuco S. (2021), *What Can We Learn from the Functional Clustering of Mortality Data? An Application to the Human Mortality Database*, "European Journal of Population", Vol. 37, pp. 769-798.

Li N., Lee R. (2005), *Coherent Mortality Forecasts for a Group of Population: An Extension of the Lee-Carter Method*, "Demography", Vol. 42, pp. 575-94.

Medford A., Christensen K., Skytthe A., Vaupel J.W. (2019), *A Cohort Comparison of Lifespan after Age 100 in Denmark and Sweden: Are Only the Oldest Getting Older?* "Demography", Vol. 56(2), pp. 665-677.

Meslé F., Vallin J., Andreyev Z. (2002), *Mortality in Europe: The Divergence between East and West*, "Population", Vol. 57(1), pp. 157-197.

Ramsay J.O., Hooker G., Graves S. (2010), *Functional Data Analysis with R and Matlab*, Springer, New York.

Renshaw A.E., Haberman S. (2006), *A Cohort-based Extension to the Lee-Carter Model for Mortality Reduction Factors*, "Insurance: Mathematics and Economics", Vol. 38(3), pp. 556-570.

Riley J.C. (2005), *Estimates of Regional and Global Life Expectancy, 1800-2001*, "Population and Development Review", Vol. 31(3), pp. 537-543.

Saarela J., Finnäs F. (2008), *Cause-specific Mortality at Young Ages: Lessons from Finland*, "Health & Place", Vol. 14(2), pp. 265-274.

Thatcher A.R., Cheung S.L.K., Horiuchi S., Robine J.M. (2010), *The Compression of Deaths above the Mode*, "Demographic Research", Vol. 22(17), pp. 505-538.

United Nations (2011), Department of Economic and Social Affairs, Population Division, Changing Levels and Trends in Mortality: The role of patterns of death by cause (United Nations Publication, ST/ESA/SER.A/318).

United Nations (2022), Department of Economic and Social Affairs, Population Division, World Population Prospects 2022, Online Edition.

Van Raalte A.A., Sasson I., Martikainen P. (2018), *The Case for Monitoring Life-span Inequality*, "Science", Vol. 362(6418), pp. 1002-1004.

Vaupel J.W., Zhang Z., van Raalte A.A. (2011), *Life Expectancy and Disparity: An International Comparison of Life Table Data*, BMJ Open.

Zanotto L., Canudas-Romo V., Mazzuco S. (2020), *A Mixture-function Mortality Model: Illustration of the Evolution of Premature Mortality*, "European Journal of Population", Vol. 37(7), pp. 1-27.

(www 1) https://population.un.org/wpp/Download/Standard/MostUsed/ (access: 15.06.2022).

Chapter VI

Selected relational models of mortality predictions in small regional areas populations in Poland

Agnieszka Orwat-Acedańska

1. Introduction

Mortality forecasts are key to developing strategies for national pension and health care systems. Recent decades have seen rapid development of mathematical methods for modeling and forecasting mortality. Mortality modeling and forecasting focuses mainly on national populations, encompassing a country. Mortality modeling methods for large, single populations are dominated by the seminal Lee-Carter [1992] model and its many extensions, summarized by Booth and Tickle [2011] and Janssen [2018].

However, regional mortality projections are increasingly important for the development and evaluation of regional policies, health care and urban planning. In the case of Poland, they concern provinces (NUTS 2 classification) and small regional areas populations − counties or municipalities (NUTS 3 classification). In the paper, regional small areas (also called small-scale areas) will refer to counties, with typical populations ranging from 55.000 to 110.000 inhabitants. In the world literature, much less attention has been paid to developing and testing methods for mortality forecasting of small-scale areas compared to the population of the country as a whole. Individual districts can be treated as sub-populations. It would seem that multi-population models could be used to forecast mortality in small-scale populations, as they provide the desired coherent forecasts for individual subpopulations. Coherent, that is ensuring non-divergence of mortality trajectories for several subpopulations. The main idea behind coherent forecasting is that mortality forecasts for populations with similar mortality developments will not diverge radically, but that structural differences will remain (for instance, consistently higher mortality for men than for women) [Hyndman, Booth and Yasmeen, 2013]. A wide spectrum of multi-population modeling methods include the functional data approach [Hyndman and Ullah, 2007], nonparametric smoothing methods, extrapolation methods and combinations of many others.

For example, Lee-Carter extension for multiple populations [Li and Lee, 2005], two-population age-period-cohort model of Cairns [Cairns and Blake, 2011], product ratio method of Hyndman [Hyndman, Booth and Yasmeen, 2013], Bayesian approaches [Gongaza and Schmertmann, 2016], multilinear component approach of Bergeron-Boucher [Bergeron-Boucher et al., 2018]. It turns out that multi-population approaches based on the cohort component are not sufficient to predict mortality in small-scale areas such as counties in Poland. This carries the risk of low reliability of estimates of mortality rates. Although most of the aforementioned methods provide the desired coherence of forecasts for subpopulations, multi-population approaches require sufficiently large amounts of data. In the cohort approach of small-scale populations, such as counties in Poland, estimates of mortality rates are based on small numbers of people, which can result in large forecast errors and discrepancies. In addition, given that the projections are estimated for cohorts, defined by gender and age, and across age groups, we face difficulties with time series that are too short to calculate mortality rates. The computational complexity of the aforementioned types of multi-population approaches for small-scale areas also generates large costs of method implementation from the point of view of statistical offices responsible for mortality reporting. The quality of data on county mortality levels can lead to large errors in forecast estimates and discrepancies. Methods that cope with the aforementioned difficulties while ensuring the consistency of forecasts are relational approaches, as classified by Wilson [2018] and Wilson et al. [2022]. These methods combine regional mortality forecasts with national forecasts through simple relationships. The essence of these methods is the scaling of various mortality measures between the subpopulation represented by a given small-scale region and a population covering the entire country or a larger administrative area than the projected one.

The purpose of this study is to evaluate the accuracy of forecasts of mortality rates and life expectancy obtained using selected relational models for all 379 counties of Poland (sub-NUTS-3 regions). We forecast cohort mortality rates by sex and age groups and life expectancy at birth. In the relational models, we use Standardized Mortality Ratio (SMR), as described and applied in Giannakouris [2010] and Rate Ratios (RR), as applied by the Office for National Statistics [ONS, 2016]. SMR is a measure used to compare the level of mortality in regions to the mortality of the country's population.

The layout of the chapter is as follows: The second subsection presents the forms of the basic mortality measures: the Cohort Mortality Ratio and the Standardized Mortality Ratio. Also illustrated is the differentiation of Poland's counties in terms of the level of mortality in individual counties in relation to the mortality of the population of Poland as a whole. The next subsection presents

the characters and ideas of selected relational models. The last subsections deal with empirical analysis. The first part describes the assumptions made in the empirical analysis, while the second part contains a description of the numerically obtained results. The work ends with a summary chapter.

2. Basic measures of mortality and the level of mortality in the counties of Poland compared to the national population

The primary measure for mortality is the cohort mortality rate (by sex and age) $d^i_{c,s}(t)$. We forecast age-sex-specific central rates of mortality for the standard abridged age groups: 0, 1-4, 5-9, 10-14, …, 75-79, 80-84, and 85+. These age intervals are the same for both men and women. The central mortality rate is defined as the ratio of the average number of deaths over 5 years in *i*-th county in cohort *c*, sex *s* for the interval *t* to *t* + 4, and the average population over 5 years in the *i*-th county in cohort *c*, sex *s*:

$$d^i_{c,s}(t) = \left[\frac{1}{5}\sum_{j=0}^{4} Z^i_{c,s}(t+j)\right] / \left[\frac{1}{5}\sum_{j=0}^{4} L^i_{c,s}(t+j)\right] \quad (1)$$

where $Z^i_{c,s}$ denotes number of deaths in *i*-th county, whereas $L^i_{c,s}$ denotes population in *i*-th county in cohort *c* and sex *s*.

Another mortality characteristic is life expectancy, which is a function of mortality rates. Life expectancy of a person aged *x* years is denoted in literature by e_x and expresses the average number of years a person aged *x* in completed years has left to live – given current mortality conditions of population [Holzer, 1989]. While random variable T_x represents the complete future lifetime for a life of exact age, then *complete expectation* of life for a life of age *x* is expected value of the random variable T_x:

$$e_x = E(T_x) = \int_0^\infty t \cdot {}_tp^i_x \, dt \quad (2)$$

where ${}_tp^i_x$ represents probability that a person aged *x* survives at least *t* further years. For practical reasons, exact ages are seldom used, and age is expressed in completed years. Therefore, for a discrete random variable K_x representing the total number of years remaining for a person aged *x* years, complete expectation of life takes the form[1]:

$$e_x = E(K_x) = \frac{1}{2} + \sum_{k=0}^{\infty} k \cdot {}_{k+1}p_x \quad (3)$$

In particular, the term 'life expectancy' often refers to the *life expectancy of newborns*, i.e., at the age of 0 years. We will denote such a measure by the symbol e_0.

[1] This formula is accurate if deaths are distributed uniformly over a given year.

On the other hand, the measure used to compare the level of mortality in the regions to that of the country's population is the Standardized Mortality Ratio (SMR) [Giannakouris, 2010]:

$$SMR^i = \frac{\text{The total death in district } i}{\begin{array}{c}\text{The teoretical number of deaths,}\\ \text{if mortality in district } i \text{ was equal to the country-level mortality}\end{array}} \quad (4)$$

A value of the measure above 1 indicates by how many percent the mortality rate of the i-th district is higher than that of the country as a whole (or a larger administrative unit). Figures 1 and 2 show a map of Poland's counties with SMR values for men and women, respectively, in 2019.

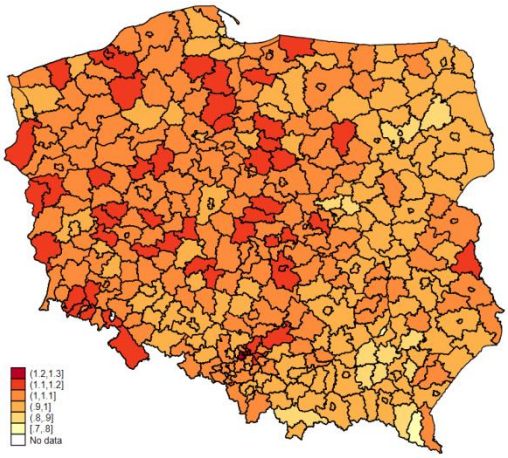


Fig. 1. SMR values in Polish counties in 2019 (men)

Source: Own calculation based on CSO.

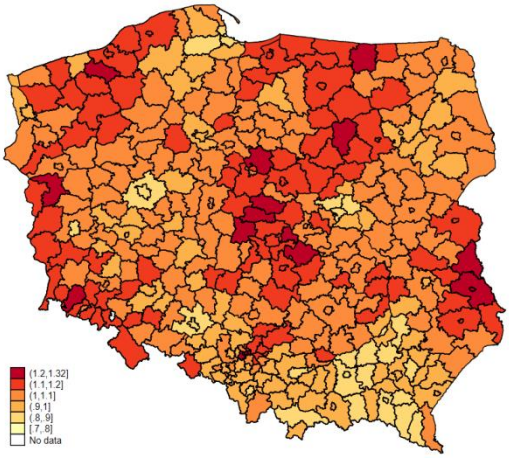


Fig. 2. SMR values in Polish counties in 2019 (women)

Source: Own calculation based on CSO data.

For both men and women, mortality in counties in relation to population mortality on a national scale is a spatially differentiated phenomenon. The spatial distribution of mortality levels measured by SMR for men differs from the distribution of intensity for women. In the case of men, the values of this measure are classified by the axis of Poland running along the north-south direction. The western part is characterized by higher SMR values compared to the eastern part. The distribution of mortality levels measured by SMR for women is more spatially differentiated. The highest SMR values dominate in districts along Poland's western border and districts belonging to the West Pomeranian and Pomeranian Voivodeships, then in the center of the country, in Warmia and in the Lublin Voivodeship. Compared to men, SMS values for women in most districts are higher.

In view of the variation in the intensity of mortality in the counties and the quality of the data, the selection of appropriate forecasting methods and models becomes a key issue, which to a large extent decisions on the quality of demographic forecasts.

3. Assumptions of empirical analysis

The subjects of the study are mortality rates for 379 counties in Poland. Of the counties, 65 are urban counties formed by the largest cities, while the remaining 314 counties include several smaller towns along with surrounding municipalities. The research period covers 2006-2019. The initial year is determined by the availability of detailed data on deaths among the older age groups in the districts. The study period adopted excludes the Covid pandemic period. Forecasting mortality in counties during this period requires imposing an additional approach on relational modeling-namely, an approach that is robust to outliers, influential observations and even extreme values, including robust diagnostics and robust estimation. This topic is a separate subject of the author's research in relational modeling. An important aspect of mortality modeling is the cohort component. Projections are determined for the population of each county characterized by age and gender. Since, for most counties, annual data on the number of deaths (by sex) of people aged from 0 completed years, 1 year, 2 years, etc. up to 85+, contain too few observations, we aggregate these data into 5-year age bins. For many districts, annual data on the number of age-specific deaths contain too few cases to aggregate data into typical five-year intervals. As a result, we consider three subperiods: 2006-2010, 2011-2015, and 2016-2019. We use the first subperiod to calculate the forecasts, while the latter two are treated as ex post testing subperiods.

The predictor variable is mortality rate of the form (1) and life expectancy at birth of the form (3). We calculate these forecasts for i = 379 counties; s = 2 (sex) and for 19 five-year cohorts (0, 1-4, 5-9, etc....., 85+; c = 1, …, 19). We use six relational models analyzed in the paper. In total, 172.824 forecasts are calculated. We evaluate the accuracy of the forecasts based on the Mean Absolute Error of the forecast (MAE). The sources of data acquisition are the Central Statistical Office (CSO) database and the Human Mortality Database (HMD).

4. Selected relational models for forecasting small-scale areas

The main idea of relational models is to scale various mortality measures between a subpopulation represented by a given small-scale region and a population covering an entire country or a larger administrative area than the one for which the forecast is set (Fig. 3).

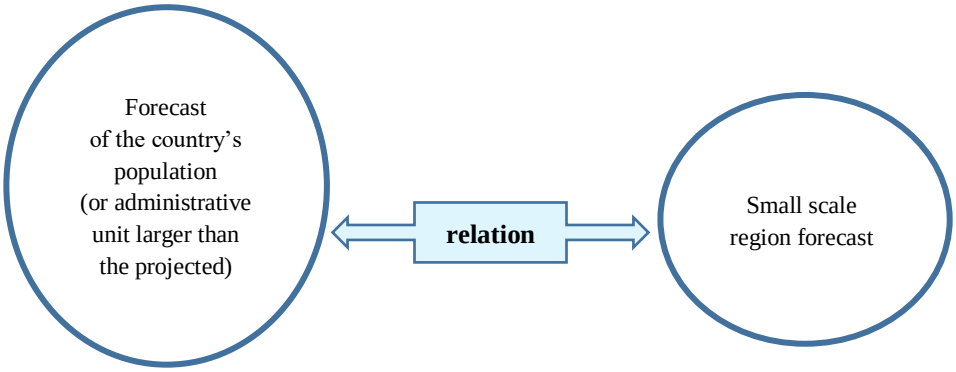


Fig. 3. The idea of relational methods for forecasting small-scale regions

Source: Own elaboration.

Below we describe the idea of selected relational models used in this study to predict mortality in individual counties of Poland. For readability of the description, symbolic names of the models have been adopted.

We will denote the projected death rate of the i-th county by sex in the 5-year age cohort for year T_p by the symbol $d_{c,s}^{i}(T_p)$.

4.1. Naive model (POL)

The POL model assumes that the projected mortality in i-th county is the same as the projected mortality at the national level, denoted by the symbol $d_{c,s}^{POL}(T_p)$. Thus, the forecasts for the i-th county are of the form:

$$d_{c,s}^{i}(T_p) = d_{c,s}^{POL}(T_p) \tag{5}$$

The calculation of the forecast $d_{c,s}^{POL}(T_p)$ is based on a method called *Mortality Surface* (MF). In this approach, the so-called *reference area* is considered. Then, we determine the *reference year* in which the life expectancy at birth in the reference area is as close as possible to the life expectancy in Poland in the base year. For subsequent years, we calculate the changes in mortality for the adopted forecast horizon in the reference country relative to the reference year. Then, projected mortality rates are determined for the Polish population, assuming that they will change at the same rate as in the reference area.

In the practice of determining regional projections by the CSO (provinces), the reference area is a set of several EU countries. The benchmark of life expectancy at birth e_0 is the average of this value determined for the set of these countries. The structure of deaths (also in terms of causality) taking into account the considered time shift (10 years) in such a reference area is most similar to the population of Poland.

4.2. Standardized Mortality Ratio Model (SMR)

In this model, the projected death rate of *i*-th county is the product of the projected death rate for the national population ($d_{c,s}^{POL}(T_p)$) and the standardized death rate for *i*-th county (SMR_s^i). For *i*-th county, the SMR value is calculated separately for sex, but for all age cohorts combined (i.e., we consider a single SMR value for the entire county, according to formula (4)):

$$d_{c,s}^{i}(T_p) = d_{c,s}^{POL}(T_p) \cdot SMR_s^i \tag{6}$$

4.3. Standardized mortality rates at the voivodeship level Model (SMR-REG)

The projected death rate of *i*-th county is calculated as in the model given by formula (6), except that the standardized death rate SMR_s^j is calculated for *j*-th voivodeship (province) to which the county belongs [Wilson, 2018]. Thus, we assume that the level of mortality in counties is the same as in the province to which the county belongs. The model has the form:

$$d_{c,s}^{i}(T_p) = d_{c,s}^{POL}(T_p) \cdot SMR_s^j \tag{7}$$

This method is used by the Statistics Poland.

4.4. Rate Ratio Model (RR)

In this method, sex-age-specific death rate ratios are calculated for each county, and the country-level forecasts are scaled by the $rr^i_{c,s}$ [Wilson, 2018]. The model has the form:

$$d^i_{c,s}(T_p) = d^{POL}_{c,s}(T_p) \cdot rr^i_{c,s} \tag{8}$$

where $rr^i_{c,s}$ is the ratio of the county's mortality rate and the national mortality rate for the base period.

4.5. Mortality Surface Model (MS)

In this approach, the forecasts for counties are calculated in the same way as the country-level forecasts for Poland (model POL). However, in this case, the applicable relationship is that of the *i*-th county and the benchmark country (and not, as in the POL model, the population of Poland and the benchmark country).

In particular, we look for the year in which the life expectancy in *reference area* is the closest to the one observed in a country in the first subperiod and analyze the changes in mortality profiles that took place in *reference area* after five and nine years [Wilson, 2014, 2015, 2018]. We assume that the same changes will occur in the analyzed district.

4.6. Brass Relational Model (BR)

In this method, we consider a linear regression model in which the explanatory variable is the logit of the indicator l^i_x, which is a parameter related to the life tables. This indicator determines how many people will live to age x in the *i*-th county. The explanatory variable of the BR model is the logit of the indicator $l^{POL}_{x,c,s}$, which determines how many people will live to age x on a national scale [Brass, 1971; Slogget, 2015]:

$$Y^i_x = \alpha^i + \beta^i Y^{POL}_{x,\,c,\,s} + \varepsilon^i \tag{9}$$

where $Y^i_x = \text{logit}(l^i_x)$, and $Y^{POL}_{x,\,c,\,s} = \text{logit}(l^{POL}_{x,c,s})$.

We estimate the linear regression models between the logit-transformed surviving populations for districts in the first subperiod and their counterparts for the whole country. Subsequently, the models are used to predict the surviving populations for districts in the next two subperiods, given the country-level forecasts.

5. Assessing forecast quality

The accuracy of the forecasts was evaluated using Mean Absolute Error (MAE). This measure tells how much, on average, during the prediction period, the actual realizations of the forecast variable deviated from the forecasts in absolute terms. We calculate the partial MAEs for mortality rates by the formula:

$$MAE_s^i(T_p) = \frac{1}{19}\sum_{c=1}^{19}\left|d_{c,s}^i(t) - d_{c,s}^i(T_p)\right| \quad (10)$$

where $d_{c,s}^i(t)$ denotes observed mortality rate, namely $d_{c,s}^i(T_p)$ projected death rate in *i*-th county in cohort *c* and sex *s*.

The partial MAEs are calculated separately for each county *i*, sex *s*, and forecasting period T_p, but are subsequently averaged across these dimensions. Analogously, we also calculate MAEs for the life expectancy at birth forecasts.

6. Results

The calculations were made using procedures created by the author in STATA. The evaluation of forecasts obtained with the relational models under consideration takes into account three perspectives. The first, most general perspective concerns the quality of the projections obtained for all 19 age cohorts together. This approach aims to capture the so-called *mortality profile*, which means the set of mortality rates for the all 19 considered age groups. This is a preliminary, general analysis aimed at comparing the quality of the forecasts obtained with the presented models.

We begin by analyzing the accuracy of the mortality profile forecasts. Table 1 contains the MAE for the mortality profiles and the MAE for life expectancy at birth. It reports the values for the full verification sample of 379 counties and two sexes. To improve readability, we use the color scale, where red represents high errors and green indicates more accurate forecasts.

For each of the three variants considered, the highest errors were obtained from the naive POL model, which assumes that mortality in the *i*-th county is the same as mortality at the national level (column two in the table). However, the SMR-based methods are only marginally better. It is worth noting here that the SMR model is slightly better (gives forecasts with smaller errors) compared to the SMR-REG model used by the CSO, with the sample variants analyzed. Recall that the SMR model is a modification of the SMR-REG model, in that instead of replacing the SMR for the county with the SMR for the province, we simply calculate this indicator for the county according to model 2 (formula 6).

This modification results in slightly better results compared to the model used by the CSO. Smaller error values are both for the mortality rate and life expectancy of new-borns projections in each of the sample variants considered, except for the MAE of the mortality rate for women. By far the best model in terms of prediction accuracy turned out to be the MR mortality surface model. Relatively good results were obtained with the Rate Ratio (RR) model. In contrast, the Brass model (BR), despite its greater complexity and implementation costs, did not prove to be the best compared to the other models.

Table 1. Mean absolute errors for the mortality profiles

Model	POL	SMR	SMR-REG	RR	MS	BR
Full sample						
MAE	0.247	0.245	0.246	0.215	0.203	0.218
MAE_e_0	0.872	0.506	0.725	0.497	0.438	0.518
Men						
$MAE\_d^i_{c.s}(T_p)$	0.299	0.296	0.306	0.251	0.245	0.247
MAE_e_0	1.037	0.559	0.871	0.643	0.509	0.674
Women						
$MAE\_d^i_{c.s}(T_p)$	0.194	0.193	0.187	0.179	0.161	0.19
MAE_e_0	0.707	0.454	0.58	0.351	0.367	0.363

The color scales represent values of the accuracy measures: the highest are marked with red and the lowest with green.

Source: Own calculations.

The second perspective for assessing projections considers cohorts of 5-year age ranges but without gender breakdown. Forecast accuracy for each single age group is summarized in Table 2. Comparing the results in Table 2 with the conclusions in Table 1, we note that the POL model, classified as 'worst' compared to the others, now has small prediction errors in the young age groups (up to 19 years). However, for the population in the 50-79 age groups, it generates large errors compared to the other models. The Mortality Surface (MS) model, which has performed well for samples that do not include cohort age ranges, now applied to individual 19 cohorts yields mortality rate predictions with low MAE error only for the 85+ age group. The opposite conclusion is obtained for the SMR model-the largest MAE error of the mortality rate forecast is for the 85+ group. In the other age groups, the forecast errors obtained with this model are the smallest compared to the other models. The structure of error values by age group is more varied in the other models (POL, SMR-REG, RR, BR). The SMR-REG model has the smallest prediction errors in the young age groups (0-49) and the largest for the population of people over 55. The relative forecasting accuracy of POL models and SMR-REG models deteriorates with

age. The RR models and MS models show the opposite effect. In the Brass BR model, the error values in the different age groups are 'sinusoidal' with respect to the age structure.

Table 2. MAE of the mortality rate forecasts for the different age groups

Age group	POL	SMR	SMR-REG	RR	MS	BR
0	0.15	0.147	0.148	0.174	0.171	0.163
1-4	0.013	0.013	0.013	0.017	0.016	0.014
5-9	0.009	0.009	0.009	0.011	0.011	0.009
10-14	0.01	0.01	0.01	0.013	0.013	0.01
15-19	0.019	0.019	0.019	0.023	0.022	0.02
20-24	0.021	0.02	0.02	0.024	0.023	0.02
25-29	0.022	0.02	0.022	0.025	0.024	0.021
30-34	0.028	0.026	0.027	0.03	0.029	0.027
35-39	0.04	0.037	0.038	0.043	0.038	0.039
40-44	0.056	0.055	0.054	0.068	0.056	0.061
45-49	0.079	0.075	0.075	0.099	0.084	0.081
50-54	0.098	0.082	0.09	0.09	0.095	0.087
55-59	0.136	0.107	0.122	0.11	0.108	0.109
60-64	0.166	0.13	0.15	0.154	0.16	0.141
65-69	0.255	0.198	0.236	0.238	0.238	0.215
70-74	0.328	0.275	0.323	0.325	0.286	0.317
75-79	0.496	0.354	0.472	0.394	0.42	0.45
80-84	0.705	0.567	0.699	0.679	0.651	0.788
85-	2.06	2.511	2.153	1.568	1.411	1.577

The color scales represent values of the accuracy measures: the highest are marked with red and the lowest with green.

Source: Own calculations.

Of course, the sheer values of the prediction errors of all models increase as people age. Models increasingly 'differentiate' among senior groups in terms of the level of MAE values. This is illustrated in Fig. 4. So, the apparent contradiction in results between Table 1 and Table 2 results from the fact that the mortality rates for the oldest cohorts are much higher than for the remaining groups of age. Therefore, the MAE values are dominated by forecast errors for these groups. As a result, a method that performs well for most of age intervals but fails for the oldest groups exhibits a poor overall performance in terms of MAE.

The third perspective for assessing projections considers cohorts of 5-year age ranges by gender. Table 3 contains the results of the obtained forecast errors for women, while Table 4 contains these results for men. The results for men and women differ in the error values of the mortality rates, but the performance of the models in each age group is similar.

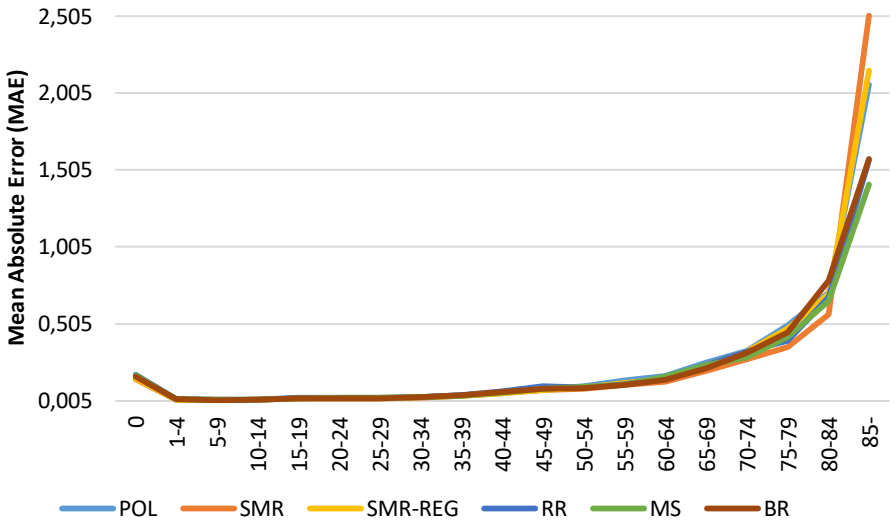


Fig. 4. MAE of the mortality rate forecasts for the different age groups in analyzed models

Source: Own calculations.

Table 3. MAE of the mortality rate forecasts for the different age groups for women

Age group	POL	SMR	SMR-REG	RR	MS	BR
0	0.135	0.132	0.132	0.170	0.165	0.158
1-4	0.012	0.013	0.012	0.016	0.016	0.013
5-9	0.009	0.009	0.009	0.011	0.010	0.009
10-14	0.010	0.010	0.009	0.012	0.012	0.010
15-19	0.014	0.014	0.014	0.016	0.017	0.014
20-24	0.013	0.013	0.013	0.015	0.016	0.013
25-29	0.013	0.013	0.013	0.015	0.016	0.013
30-34	0.015	0.015	0.015	0.019	0.019	0.016
35-39	0.021	0.021	0.021	0.024	0.025	0.020
40-44	0.029	0.027	0.028	0.033	0.036	0.027
45-49	0.047	0.043	0.044	0.051	0.057	0.042
50-54	0.067	0.062	0.062	0.065	0.069	0.061
55-59	0.098	0.092	0.091	0.089	0.078	0.092
60-64	0.118	0.097	0.105	0.117	0.123	0.079
65-69	0.206	0.180	0.195	0.181	0.166	0.164
70-74	0.270	0.225	0.257	0.205	0.202	0.219
75-79	0.362	0.281	0.344	0.283	0.301	0.357
80-84	0.462	0.411	0.437	0.516	0.472	0.725
85+	1.793	2.018	1.750	1.562	1.255	1.581

The color scales represent values of the accuracy measures: the highest are marked with red and the lowest with green.

Source: Own calculations.

Table 4. MAE of the mortality rate forecasts for the different age groups for men

Age group	POL	SMR	SMR-REG	RR	MS	BR
0	0.165	0.161	0.163	0.177	0.177	0.169
1-4	0.014	0.014	0.014	0.017	0.017	0.014
5-9	0.009	0.009	0.009	0.011	0.011	0.009
10-14	0.011	0.011	0.011	0.015	0.014	0.011
15-19	0.024	0.023	0.024	0.030	0.027	0.025
20-24	0.029	0.026	0.028	0.033	0.031	0.028
25-29	0.032	0.028	0.030	0.034	0.033	0.029
30-34	0.040	0.036	0.038	0.041	0.039	0.038
35-39	0.058	0.053	0.056	0.061	0.051	0.058
40-44	0.083	0.083	0.080	0.103	0.076	0.095
45-49	0.112	0.106	0.106	0.147	0.111	0.120
50-54	0.129	0.103	0.117	0.115	0.122	0.112
55-59	0.173	0.121	0.153	0.131	0.137	0.126
60-64	0.214	0.163	0.194	0.191	0.197	0.203
65-69	0.303	0.216	0.277	0.295	0.310	0.266
70-74	0.385	0.324	0.390	0.445	0.369	0.415
75-79	0.630	0.428	0.600	0.505	0.538	0.542
80-84	0.947	0.722	0.960	0.841	0.830	0.851
85+	2.326	3.001	2.553	1.574	1.567	1.574

The color scales represent values of the accuracy measures: the highest are marked with red and the lowest with green.

Source: Own calculations.

7. Summary

The paper evaluates the accuracy of forecasts of mortality rates and life expectancy at birth using six selected relational models. The models were applied to regional mortality forecasting of the population of 379 counties of Poland, which are regions of small scale. Cohort mortality rates by sex and 19st five--year age intervals were analyzed.

In the part of the empirical analysis concerning the evaluation of the quality of the projections obtained for all 19 age intervals combined, the most important conclusions are as follows: The SMR-REG model that the CSO uses gives results that are far from optimal. Compared to the other models, the best results (the smallest MAE errors of forecasts) were obtained using the MS model − both for the population without and with sex division. Slightly inferior were the approaches derived from the RR and BR models. However, given the level of computational cost and structural complexity, the BR model is not recommended as useful for mortality forecasting purposes for small-scale regions. It should also be noted that for projections of life expectancy at birth − the SMR-REG approach is better than the SMR method for the female population only.

Ranking of the methods changes significantly when analyzing the forecasts by single age groups. In general, the relative performance of the various methods is very much age-dependent. Specifically, the SMR-REG method performs well for young cohorts, and the SMR method offer only a slight improvement. Thus, SMR-REG method seems a reasonable choice for infant mortality forecasting. Notably, the methods that are the most accurate for forecasting mortality profiles (RR, MS) are definitely more accurate than the alternatives for the oldest cohort, which dominates our accuracy measures. The advantages of the presented relational models are: the assumption of convergence of forecasts − crucial in the case of small-scale regions (counties), relative simplicity and resistance to too short time series and small amount of data.

References

Bergeron-Boucher M., Simonacci V., Oeppen J., Gallo M. (2018), *Coherent Modeling and Forecasting of Mortality Patterns for Subpopulations Using Multiway Analysis of Compositions: An Application to Canadian Provinces and Territories*, „North American Actuarial Journal", Vol. 22(1), pp. 92-118.

Booth T., Tickle L. (2008), *Mortality Modelling and Forecasting: A Review of Methods*, "Annals of Actuarial Science", Vol. 3(1-2), pp. 3-43.

Brass W. (1971), *On the Scale of Mortality* [in:] W. Brass (ed.), *Biological Aspects of Demography*, Taylor and Francis, London, pp. 69-110.

Cairns A.J.G., Blake D. (2011), *Bayesian Stochastic Mortality Modeling for Two Populations*, "ASTIN Bulletin", Vol. 41(1), pp. 29-59.

Giannakouris K. (2010), *Regional Population Projections EUROPOP2008: Most EU Regions Face Older Population Profile in 2030*, Eurostat statistics in focus 1/2010, European Commission, Luxembourg, http://ec.europa.eu/eurostat/en/web/product sstatistics-in-focus/-/KS-SF-10-001.

Gonzaga M.R., Schmertmann C.P. (2018), *Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas with Defective Vital Records*, „Demography", Vol. 55, No. 4, pp. 1363-1388.

Holzer J. (1989), *Demography*, PWE, Warszawa.

Hyndman R.J., Booth H., Yasmeen F. (2013), *Coherent Mortality Forecasting: The Product Ratio Method with Functional Time Series Models*, "Demography", Vol. 50(1), pp. 261-283.

Hyndman R.J., Ullah M.S. (2007), *Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach*, "Computational Statistics and Data Analysis", Vol. 51(10), pp. 4942-4956.

Janssen F. (2018), *Advances in Mortality Forecasting: Introduction*, "Genus", Vol. 74(21).

Lee R.D., Carter L.R. (1992), *Modeling and Forecasting US Mortality*, "Journal of the American Statistical Association", Vol. 87(419), pp. 659-671.

Li N., Lee R. (2005), *Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method*, "Demography", Vol. 42(3), pp. 575-594.

ONS (Office for National Statistics) (2016), *Methodology Used to Produce the 2014-based Subnational Population Projections for England*, https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/methodologies/methodologyusedtoproducethesubnationalpopulationprojectionsforengland (access: 14.05.2023).

Sloggett A. (2015), *Demographic Models: Model Life Tables* [in:] *Population Analysis for Policy and Programmers*, International Union for the Scientific Study of Population, Paris, http://papp.iussp.org/sessions/papp103_s01/PAPP103_s01_010_010.html.

Wilson T. (2015), *POPACTS*: *Simplified Multi-regional Projection Software for State, Regional and Local Area Population Projections* [in:] T. Wilson, E. Charles-Edwards, M. Bell (eds.), *Demography for Planning and Policy*: *Australian Case Studies*, Springer, Cham, pp. 53-69.

Wilson T. (2014), *Simplifying Local Area Population and Household Projections with POPART* [in:] N. Hoque, L. Potter (eds.), *Emerging Techniques in Applied Demography*, Springer, Dordrecht, pp. 25-38.

Wilson T. (2018), *Evaluation of Simple Methods for Regional Mortality Forecasts*, "Genus", Vol. 74(14).

Wilson T., Grossman I., Alexander M., Rees P., Temple J. (2022), *Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs*, "Population Research and Policy Review", Vol. 41, pp. 865-898.

Chapter VII

Generalized linear model in the study of determinants of youth unemployment rates in Polish provinces

Magdalena Kawecka

1. Introduction

Youth unemployment is a significant problem faced by many countries around the world. It refers to a situation in which young people who are willing and able to work but are unable to find employment. This problem is particularly difficult because it can have long-lasting consequences not only for the individuals concerned, but also for society as a whole. There are many causes of youth unemployment, including a lack of available jobs, a mismatch between the skills young people have and the skills employers are looking for, and economic downturns or crises that make it difficult for employers to hire new employees [Kozlowska, 2022]. To address youth unemployment, governments and organisations often implement a range of policies and programmes to help young people find work. This can include such things as vocational training programmes, apprenticeships and traineeships, as well as initiatives to support entrepreneurship and small business development. Addressing youth unemployment is an important step towards creating a more inclusive and sustainable economy that benefits everyone [Organiściak-Krzykowska and Hrynkiewicz, eds., 2022].

This study will assess unemployment related to the group of young people who remain unemployed after completing their education. The aim of the study is to assess the determinants of youth unemployment rates in the Polish provinces. So far, research in this area has been conducted on a large scale, but in view of the dynamic structure of the labour market, there are still some research gaps in the area of research focusing on young people. The implementation of the research process will make it possible to fill in the gaps concerning the situation of young people on the labour market. At the same time, within the framework of the questions posed, the analysis and conclusions of the research can serve as an indication of the direction of change, which can contribute to an increase in the employment of young people.

The issue of unemployment itself is complex. Taking into account its various aspects, many definitions and analyses of this phenomenon appear in the available literature. Entering the labour market, looking for one's first job, for a person without experience is not and will never be easy. In times of crises, additionally, the situation for those who are starting to take their first steps on the professional path seems to worsen [Piecuch, 2013]. Young people are undeniably one of the most important forces and resources a country can have at its disposal to stimulate socio-economic development. It should be borne in mind that this is a social group which, apart from being numerous, in Poland the 20-34 age group accounts for approximately 18.18% of the total population (the 25-34 group accounts for approximately 13.21%) in 2021. Moreover, we are talking about people who are energetic, courageous and willing to develop their competences, who have a lot of new ideas to offer, which can not only change social economic development (when properly coordinated and involved in the country's economic activity), but also improve processes, thanks to efficient movement in the world of technology [Uścińska and Wiśniewski, eds., 2022]. In view of the above, young people face many challenges, and the only one they face is unemployment, which not only affects material status, participation in social life, but also aspects at the psychological base. Considering material status, therefore, employee remuneration, which affects well-being and therefore quality of life, is extremely important. In Poland, the average gross monthly salary in 2021 was PLN 5682.97, so it increased by approximately 8.7% compared to the previous year [Statistics Poland, 2022d]. Kostrzewski and Worach-Kardas [2013], in their study, attempted to assess the impact of duration of unemployment and isolated socio-demographic characteristics on perceived quality of life, self-assessment of health status, mental health status, and occurrence of chronic diseases among unemployed people aged 45 and over, based on which they found a statistically significant relationship between duration of unemployment and deterioration of mental health status and quality of life. The researchers showed that this phenomenon increases with prolonged duration of unemployment and, in turn, the strength of the health effects of unemployment depends on a number of coexisting factors, i.e.: satisfaction with personal relationships, belief in the ability to be re-employed, physical activity or the presence of another unemployed person in the household. On the other hand, Karwacki and Błędowski [2020], in their study, identify a number of dimensions of young people's experience of unemployment, referring to the quality of available work, the experience of job search, contacts with the labour administration, crises in social relationships and mental health itself. Following the psychological pathway, Drela [2015] notes that it can be concluded that certain elements of the social environment are nec-

essary for mental health, in turn, the absence of these elements both in the workplace and in a situation of unemployment results in a deterioration of mental health (referring, among other things, to the ability to use qualifications, the availability of money or a valued social position). Among such elements, the lack of jobs can also be singled out, as they represent an opportunity to reduce youth unemployment. In Poland, the number of newly created jobs in 2021 was 582.700, up by approximately 23.87% on the previous year. However, liquidated jobs should also be borne in mind: as many as 251.4 thousand jobs were liquidated in 2021, down by about 23.77% on the previous year. In a sense, it is possible that the situation is changing positively, but the ratio of newly created jobs to liquidated jobs, gives us about 331.3 thousand newly created jobs. In summary, on the one hand, one can see the economy developing, while on the other hand, we note that the situation on the labour market is unstable. An important factor affecting the young is the COVID-19 pandemic. Subocz [2022] notes young people, compared to other age groups, are the social group most affected by the crisis caused by the COVID-19 pandemic. The biggest threat in turn is the increase in unemployment and long-term exclusion from the market. This is also corroborated by researcher O'Higgins [2011], who pointed out that previous recessions show that youth unemployment not only increases rapidly and significantly, but above all remains above pre-crisis levels long after the recovery [ILO, 2020; Verick, 2009].

This speaks to the multifaceted nature of the unemployment problem, but since social issues relating to, among other things, mental health were mentioned above, the question arises as to the impact of a country's economic development. The economic literature focuses on factors that upset the balance between the supply and demand sides of the labour market, including the impact of economic prosperity and modernisation processes on the functioning of the labour market. These studies show that developmental processes in the economy are conducive to a reduction in unemployment, while technological progress causes an increase in unemployment, which affects young people in particular [Golnau and Kalinowski, eds., 2007]. Golnau notes that in periods of high growth Gross Domestic Product (GDP), employment of young working-age cohorts grows faster than employment of older people. This is due to the fact that employers hire young people with less experience to meet the growing demand for human resources. On the other hand, during a downturn, the situation is just the opposite: employment of young people falls faster than total employment, as employers are forced to lay off workers, refrain from recruiting younger workers in the first place and dismiss those with the least work experience and the shortest length of service [Grotowska-Leder, 2015]. GDP is a measure of the volume of goods and

services produced in a country over a given period of time, despite its imperfections as discussed by Robert Kennedy, among others, over 50 years ago, GDP still remains the primary measure of the state of the economy. GDP in 4th quarter 2021 seasonally adjusted GDP (at constant prices with a reference year of 2015) grew by 1.7% in real terms compared to the previous quarter and was 7.6% higher than a year ago. Seasonally unadjusted GDP (at constant average prices of the previous year) grew by 7.3% in real terms compared with the fourth quarter of the previous year [Statistics Poland, 2022c]. Measures of economic performance also include output, i.e. the total of products produced during the accounting period, the sum of which in the 4th quarter amounted to about 5.2 million PLN and is higher than a year ago by 14.07% [Statistics Poland, 2022b].

Young people, unable to find a work in the place of residence, often decide to migrate internally or externally (outside the country). Taking into account non-economic motives, factors of political, legal, cultural, as well as historical nature are distinguished [Siek and Bednarczyk, 2009]. However, economic and demographic motives, i.e. poverty, low wages, unemployment, but also socio-cultural motives, i.e. ethnic discrimination and the already mentioned political motives, i.e. violation of human rights, corruption [Dębowska, 2007], are more often mentioned. However, it is among young people that reasons of an economic nature, directly related to, among other things, remuneration, predominate.

On the basis of the above review, it can be concluded that there are a number of factors influencing the unemployment of young people entering the labour market. One such factor is a country's economic development, which influences the shape of the labour market.

2. Data characteristics

The identification and analysis of determinants of youth unemployment rates was based on statistical data provided by the Central Statistical Office (CSO) and Eurostat. The data provide information on Poland's economic development in 2010-2021.

Prior to the analysis and construction of the model, the collected data were assessed for completeness and relevance of the given factors. Guided by substantive reasons, 11 independent factors (explanatory variables) were selected, which correspond to economic development in Poland with provinces as a qualitative variable. In order to focus the study on young people, the unemployment rate was taken into account for the 25-34 age group − corresponding to the group of people who have completed higher education and are entering the labour

market. The choice of age group was guided by the increase in the unemployment rate from 2019 (3.7%). In 2021, the unemployment rate increases by 0.2 percentage points − see Fig. 1.

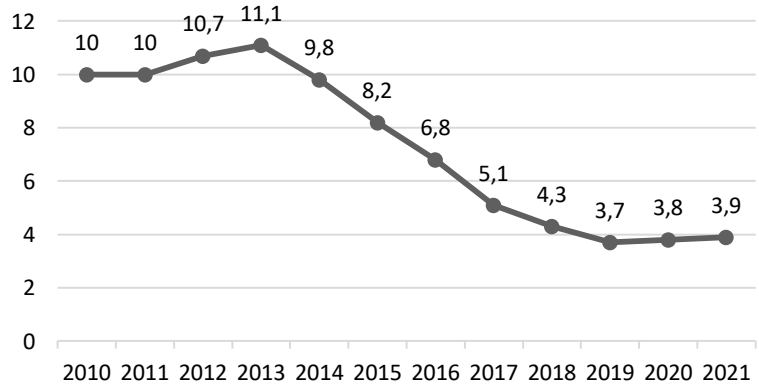


Fig. 1. Unemployment rates by age 25-35 in Poland (2010-2021)

Source: Own calculations based on Eurostat data.

Generalized Linear Models (GLM) were used to assess the impact of independent variables (qualitative and continuous predictors) on the dependent variable, which is the unemployment rate for the 25-35 age group. The variables are presented in detail in Table 1.

Table 1. Variables used in the survey

Variables	Description of the variables	Unit of variables
UR	Unemployment rates for the 25-34 age group	[%]
GDI	Gross disposable income per capita, Poland = 100	[−]
$NNWP$	Number of newly created work places	[thousand places]
GDP	Gross domestic product per capita	[PLN]
NE	New entities of the national economy recorded per 10 thousand population at working age	[−]
OP	Output (Global production)	[million PLN]
$AMGW$	Average monthly gross wages and salary in relation to the average domestic (Poland = 100)	[%]
RGR	Real growth rate of regional gross value added (GVA) at basic prices	Index, 2015=100
NM_I	Net migration internal for permanent residence	[person]
NM_IN	Net migration international for permanent residence	[person]
$IRNE$	Investment rate in national economy	[%]
EVB_SA	Expenditure of voivodships budgets: Social assistance	[PLN]

Source: Own calculations based on Eurostat and Statistics Poland data.

3. Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are a class of models that are extensions of linear models. Both linear and non-linear effects can be analysed for any number and type of predictor variables on the discrete or continuous dependent variable. Systems can include multiple degree of freedom effects for qualitative predictors, single degree of freedom effects for continuous predictors or any combination of effects for continuous and qualitative predictors [Aczel, 2018; Stanisz, 2007; Ptak-Chmielewska, 2013]. Thus, in the GLM, the population expected value depends on a linear predictor (a linear combination of explanatory variables) through a non-linear linking function, while the distribution of the dependent variable is any distribution from the family of exponential distributions [Stanisz, 2007; Ptak-Chmielewska, 2013].

Three fundamental assumptions of the GLM are mentioned above all [Vonesh, 2012; Ptak-Chmielewska, 2013]:

1) the assumption of randomness of the analysed sample,
2) the assumption of statistical independence of the sampled units,
3) the assumption that the observations of the dependent variable are independent of each other and come from the same probability distribution.

In the model notation, the following designations have been adopted: X – matrix of the system of independent (explanatory) variables, x_i – i-th row of the matrix X, Y – dependent variable, β – coefficients relating to the matrix X, $\widehat{\beta}$ – estimators of coefficients β, $\eta = X\beta$ – linear predictor, L – reliability function, LL – the (natural) logarithm of the credibility function, g – link function, $g(Y) = \eta$.

GLM consists of the following three components [por. Nelder and Wedderburn, 1972, p. 32; Ptak-Chmielewska, 2013; Dobson, 1990; Green and Silverman, 1994; McCullagh and Nelder, 1989]:

- dependent variable with a distribution from the family of exponential distributions, which means, among other things, that the variance of the variable depends on the expected value through the variance function:

$$Var(y) = \frac{\phi * V(\mu)}{\varpi} \tag{1}$$

  where μ – expected value, ϕ – dispersion parameter (known or estimated), ϖ – weighting for each observation;
- linear component (linear predictor), i.e. a linear combination of the explanatory variables in the model (as in a linear model), which may include, inter alia, quantitative variables, transformations of these variables, binary variables, polynomials, interactions:

$$\eta = X^T\beta \tag{2}$$

– link function, a monotone differentiable function that determines how the expected value of the dependent variable is related to a linear predictor:

$$g(\mu) = X^T\beta \tag{3}$$

The study used a log-normal model, for which the binding function chosen was Log: $f(\mu) = \log(\mu)$ and identity: $f(z) = z$. Parameterisation is based on sigma-constraints.

Hypothesis verification of the significance of individual model parameters or groups of parameters was carried out using the Wald test and reliability quotient tests [Stanisz, 2007; Ptak-Chmielewska, 2013]:

1. Maximum-likelihood parameter estimation provides parameter estimators and standard error estimates of parameter estimators. We denote the parameter estimator of the *j*-th variable by β_j, but by σ_{β_j} its asymptotic standard error. The Wald statistic is then written with the formula:

$$W = \left(\frac{\beta_j}{\sigma_{\beta_j}}\right) \tag{4}$$

This statistic verifies the null hypothesis: $H_0: \beta_j = 0$, which has a Chi--square distribution with 1 degree of freedom. Rejection H_0 (its materiality) means that the *i*-th variable is a significant predictor.

2. Reliability quotient tests are denoted by L_1 reliability function for a model containing p independent variables $(X_1, X_2, \dots, X_p)$, and by L_2 credibility function for a model that additionally includes k new variables $(X_{p+1}, X_{p+2}, \dots, X_{p+k})$. The form of the statistic is then written with the formula:

$$LR = -2\log\left(\frac{L_1}{L_2}\right) \tag{5}$$

This statistic has an asymptotic Chi-square distribution with $(p + k) - p = k$ degrees of freedom, so the number of variables in the second model minus the number of variables in the first model. The null hypothesis has the form: $H_0: \beta_{p+1} = \beta_{p+} = \cdots = \beta_{p+k} = 0$. Then, when the statistic is significant, we reject the null hypothesis. This means that at least one of the independent variables introduced significantly affects the dependent variable (otherwise the new variables are not significant). A special case is considered when $k = 1$, in which case we test the significance of one variable [Agresti, 1990].

The next step is to determine how well our defined model fits the observed data. To do this, statistics such as:

1. Deviation, described by the formula: $D = 2 * \left(L_p(\beta, y) - L(\beta, y)\right)$, where $L_p(\beta, y)$ denotes the logarithm of the highest reliability for the full model,

and $L(\beta, y)$ is the value of the logarithm of credibility for the model under consideration. The deviation statistic for a normal distribution has the form: $\sum(y_i - \mu)^2$ [Agresti, 1990; Gill, 2000].

2. Generalised Pearson's Chi-square statistic (χ^2) written with the formula: $\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sqrt{Var(\mu)}} \right)^2$, where $Var(\mu)$ is the estimated variance.
3. Akaike information criterion (AIC). The criterion introduces a so-called 'penalisation' of the reliability function, such that simpler models are preferred. When fitting a model with q parameters to the data, the criterion takes the form of: $D = \alpha q \varphi$, where D is the deviation and φ dispersion parameter. We choose the model for which the expression is minimal [Olson, 1974; Akaike, 1973].

4. Result

Based on the prepared data, the model was verified. At the outset, it should be noted that Tables 2 and 3 provide summary analyses of the generalized linear model method used. As mentioned in the method for a normal distribution, two binding functions were used, namely the Log function and the identity function, allowing the estimation of the observed data to be assessed. The first step was to assess the significance of the parameters obtained and to fit the model. The model applied sigma-restricted parametrization to Opole Province (log-normal model with Log function) and Lodz Province (log-normal model with identity function). In accordance with GLM statistics, the values were chosen to facilitate interpretation of the magnitude of the regression coefficient associated with the explanatory variable (predictor) [StatSoft Electronic Statistics Textbook, 2023].

The estimated model, taking into account the normal distribution and the binding function of the Log function, is of the form:

$$\log(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{27} x_{27} \tag{6}$$

In order to find out the regression coefficients, the parameters were evaluated and are presented in Table 2 (see test of all effects and parameter estimates).

Table 2. Results of generalized linear models (Distribution: *normal*; Link function: *log*)

Effect	*UR − Test of all effects*							
						DF	Wald (Stat.)	p-value
Intercept						1	198,2973	0,0000
GDI						1	8,1872	0,0042
NE						1	24,3822	0,0000
OP						1		
RGR						1	230,0419	0,0000
NM_IN						1	52,7265	0,0000
IRNE						1	4,9329	0,0263
EVB_SA						1		
Regions						15	301,3632	0,0000
Effect	***UR − Parameter estimates***							
	Level of (Effect)	Column	Estimate	Standard (Error)	Wald (Stat.)	Lower CL (95,0%)	Upper CL (95,0%)	p-value
Intercept		1	8,1047	0,5755	198,2973	6,9767	9,2328	0,0000
GDI		2	−0,0157	0,0055	8,1872	−0,0265	−0,0050	0,0042
NE		3	−0,0059	0,0012	24,3822	−0,0083	−0,0036	0,0000
OP		4	0,0000	0,0000	12,0494	0,0000	0,0000	0,0005
RGR		5	−0,0379	0,0025	230,0419	−0,0428	−0,0330	0,0000
NM_IN		6	−0,0001	0,0000	52,7265	−0,0002	−0,0001	0,0000
IRNE		7	−0,0103	0,0046	4,9329	−0,0193	−0,0012	0,0263
EVB_SA		8	0,0000	0,0000	2,2213	0,0000	0,0000	***0,1361***
Regions	Lesser Poland	9	−0,1101	0,0524	4,4080	−0,2129	−0,0073	0,0358
Regions	Silesia	10	−1,0674	0,1506	50,2248	−1,3627	−0,7722	0,0000
Regions	Greater Poland	11	−0,4923	0,0884	31,0219	−0,6655	−0,3190	0,0000
Regions	West Pomerania	12	0,3095	0,0768	16,2195	0,1589	0,4601	0,0001
Regions	Lublin Province	13	0,1793	0,0910	3,8782	0,0009	0,3577	0,0489
Regions	Lower Silesia	14	−0,1939	0,0736	6,9428	−0,3381	−0,0497	0,0084
Regions	Kuyavia-Pomerania	15	0,1706	0,0519	10,8093	0,0689	0,2723	0,0010
Regions	Pomerania	16	−0,1350	0,0683	3,9020	−0,2689	−0,0011	0,0482
Regions	Holy Cross	17	0,5708	0,0858	44,2588	0,4027	0,7390	0,0000
Regions	Lubusz Province	18	0,3546	0,0696	25,9967	0,2183	0,4909	0,0000
Regions	Podlasie Province	19	0,6496	0,0886	53,8007	0,4760	0,8232	0,0000
Regions	Masovia	20	−0,8604	0,3056	7,9287	−1,4594	−0,2615	0,0049
Regions	Subcarpathia	21	0,4636	0,0752	38,0344	0,3163	0,6110	0,0000
Regions	Warmia-Masuria	22	0,2168	0,0799	7,3564	0,0601	0,3735	0,0067
Regions	Lodz Province	23	0,0425	0,0444	0,9162	−0,0445	0,1295	***0,3385***
Scale			1,1705	0,0597		1,0591	1,2936	

Effect	*UR − Likelihood Type 1 Test*				*UR − Likelihood Type 3 Test*				
	DF	Log-likelihood	Chi-(Square)	*p*-vaule		DF	Log-likelihood	Chi-(Square)	*p*-value
Intercept	1	−531,9988							
GDI	1	−502,9342	58,1293	0,0000	**GDI**	1	−306,6683	8,0188	0,0046
NE	1	−472,6791	60,5102	0,0000	**NE**	1	−314,3123	23,3068	0,0000
OP	1	−467,7741	9,8099	0,0017	**OP**	1	−308,0132	10,7086	0,0011
RGR	1	−401,5081	132,5321	0,0000	**RGR**	1	−368,7455	132,1732	0,0000
NM_IN	1	−398,4680	6,0801	0,0137	**NM_IN**	1	−326,0211	46,7243	0,0000
IRNE	1	−398,2027	0,5307	**0,4663**	**IRNE**	1	−305,1228	4,9278	**0,0264**
EVB_SA	1	−397,2058	1,9938	**0,1579**	**EVB_SA**	1	−303,9054	2,4930	**0,1144**
Regions	15	−302,6589	189,0937	0,0000	**Regions**	15	−397,2058	189,0937	0,0000

UR − Statistics of goodness of fit			
	DF	Stat.	Stat/DF
Deviance	169	263,0425	1,5565
Scaled Deviance	169	192,0000	1,1361
Pearson Chi2	169	263,0425	1,5565
Scaled P. Chi2	169	192,0000	1,1361
AIC		653,3179	
AICC		660,5035	
BIC		731,4978	
Log-likelihood		−302,6589	

Significance at *p*-value = 0,05.

Parametrization: sigma-restricted (relative to Opole Province).

Source: Own calculations.

The model was estimated based on summary of all effects with the Wald statistic and respective *p*-values for all effects in the model. The variables NNWP, GDP, AMGW and NM_I are statistically insignificant, so it was decided to remove them so that they do not interfere with the model results.

According to the estimation results, the model takes the form of (see Table 2):

$$\begin{aligned}\log(\mu_i) = {} & 8{,}105 - 0{,}016GDI - 0{,}006NE + 0{,}000OP - 0{,}038RGR \\ & + 0{,}000NM_{IN} - 0{,}010IRNE + 0{,}000EVB^{*}_{SA} \\ & - 0{,}110R_{Lesser\ Poland} - 1{,}067R_{Silesia} \\ & - 0{,}492R_{Greater\ Poland} + 0{,}309R_{West\ Pomerania} \\ & + 0{,}179R_{Lublin\ Province} - 0{,}194R_{Lower\ Silesia} \\ & + 0{,}171R_{Kuyavia-Pomerania} - 0{,}135R_{Pomerania} \\ & + 0{,}571R_{Holy\ Cross} + 0{,}355R_{Lubusz\ Province} \\ & + 0{,}650R_{Podlasie\ Province} - 0{,}860R_{Masovia} \\ & + 0{,}464R_{Subcarpathia} + 0{,}217R_{Warmia-Masuria} \\ & + 0{,}042R^{*}_{Lodz\ Province}\end{aligned} \tag{7}$$

* non-significant variables were also included in the model.

A certain tendency affecting the estimated model negatively was observed, namely the removal of the EVB_SA variable results in a worse fit of the model. Moreover, irrespective of the choice of the Opole Provincial as the reference variable, it was the choice of the Opole Provincial that resulted in the best fit of the model.

Interpretation of the results of parameter estimation in the model takes into account exponential transformations (for selected explanatory variables, the scale parameter was assumed constant, equal to 1.0):

- $(\mathrm{e}^{-0{,}015730} - 1) * 100\% = -1{,}56\%$ – an increase in disposable income (gross) per capita by 1 percentage point results in a decrease in the unemployment rate among young people in the 25-34 age group by 1,56%;
- $(\mathrm{e}^{-0{,}010268} - 1) * 100\% = -1{,}02\%$ – an increase in the rate of investment in the national economy by 1 percentage point results in a decrease in the unemployment rate among young people in the 25-34 age group by 1,02%;
- $e^{0{,}463646} = 1{,}60$ – this means that young people living in Subcarpathia are 1.6 times more likely to be unemployed than those living in Opole Province.

Based on the results of the type 1 and type 3 analysis, of the variables included, IRNE (Investment rate in national economy) and EVB_SA (Expenditure of voivodships budgets: Social assistance) were found to be insignificant in the model for the type 1 test. In the case of the type 3 test, only the variable EVB_SA is a non-significant variable (which confirms the evaluation of the model estimation).

Taking into account the significance of the variables on the dependent variable, the next step was to assess the fit of the model (Table 2, see Statistics of goodness of fit). The deviance and statistical value of Pearson's Chi-square statistic for the model is 263.04. It can therefore be concluded that the algorithm has achieved convergence. The overloading is due to not taking into account the degrees of freedom (unloaded estimator σ^2 is the size $\hat{\sigma}^2$, and therefore equals Deviation/DF = 1,556. Therefore, the element $\hat{\sigma} = \sqrt{263{,}042} = 16{,}22$ – value equal to the standard error of the estimation. There was no overdispersion in the model, which is due to the fact that the value of the deviancy when divided by the degrees of freedom is close to the value of 1 ($V = \frac{263{,}042507}{169} = 1{,}56$), the overdispersion phenomenon is not significant and does not exceed a value of 2, so that no correction of parameter estimation errors due to overdispersion needs to be made [Ptak-Chmielewska, 2013, pp. 65-68].

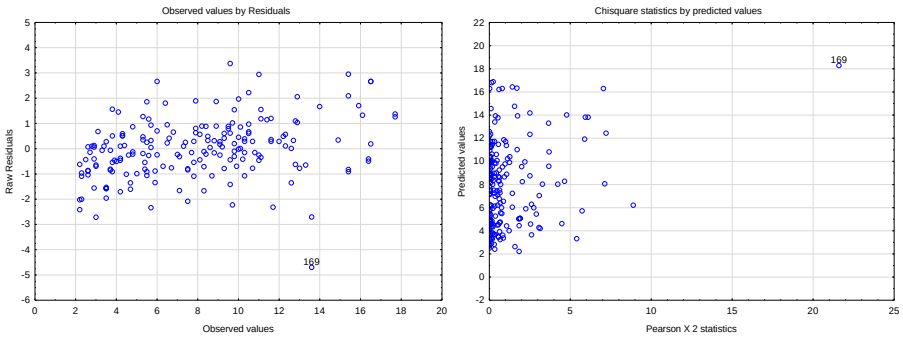


Fig. 2. Graphs of the spread of predicted versus observed values

Source: Own calculations.

As can be seen, case 169 (Subcarpathia for 2010) − a scatter plot of observed values against predicted values. For confirmation, a scatter plot of Chi-square values (indicating the contribution of individual observations to the Chi-square statistic) for each case against predicted values was also made. It should be borne in mind that individual regions in Poland differ, not only in terms of geography (i.e. location), but also in terms of urbanisation. Taking this into account − outliers were not removed.

The model was also verified for the normality of the residuals (Fig. 3). From it, one can infer validity − the residuals have an approximately normal distribution.

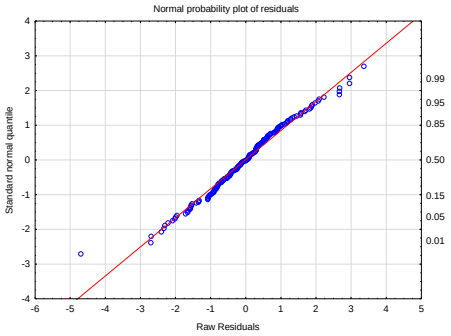


Fig. 3. Normal probability plot of residuals

Source: Own calculations.

In summary, the independent variables influencing the decreasing membership of the unemployed among young people are: GDI (gross disposable income per capita), NE (new entities of the national economy recorded per 10 thousand population at working age), RGR (real growth rate of regional gross value), NM_IN (net migration international for permanent residence) and IRNE (investment rate in national economy). In the case of growth in the other variables, there is an increase in the unemployment rate. On the other hand, for those living in West Pomerania, Lublin Province, Kuyavia-Pomerania, Holy Cross, Lubusz Province, Podlasie Province, Subcarpathia and Warmia-Masuria, the unemployment rate will be higher than for those living in Opolskie Province. For the remaining voivodeships, in relation to the reference variable, the unemployment rate will be lower.

Given these results, it was decided to use a model with a normal distribution with an identity binding function for comparison, the estimated model is of the form:

$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{27} x_{27} \tag{8}$$

Summary results are presented in Table 3. We write the model similarly to the logarithmic function:

$$\begin{aligned} \log(\mu_i) = {} & 61{,}292 - 0{,}244GDI - 0{,}030NE + 0{,}00002OP - 0{,}283RGR \\ & - 0{,}0009NM_{IN} + 0{,}000\mathrm{EVB}_{\mathrm{SA}} - 1{,}39R_{Lesser\ Poland} \\ & - 9{,}616R_{Silesia} - 5{,}052R_{Greater\ Poland} \\ & + 2{,}26R_{West\ Pomerania} + 1{,}811R_{Lublin\ Province} \\ & - 2{,}761R_{Lower\ Silesia} + 1{,}9R_{Kuyavia-Pomerania} \\ & - 1{,}1R_{Pomerania} + 6{,}144R_{Holy\ Cross} \\ & + 4{,}31R_{Lubusz\ Province} + 7{,}03R_{Podlasie\ Province} \\ & - 12{,}95R_{Masovia} + 5{,}8R_{Subcarpathia} + 0{,}9R^{*}_{Opole\ Province} \\ & + 2{,}43R_{Warmia-Masuria} \end{aligned} \tag{9}$$

* non-significant variables were also included in the model.

As before, the interpretation of the results of the parameter estimation in the model takes into account exponential transformations (for the selected explanatory variables, the scale parameter was assumed constant, equal to 1.0):

- $(e^{-0,2437} - 1) * 100\% = -21,63\%$ – an increase in gross disposable income per capita by 1 percentage point results in a decrease in the unemployment rate among young people in the 25-34 age group by 21,63%;
- $(e^{0,000024} - 1) * 100\% = 0,0024\%$ – an increase in output by PLN 1 million results in an increase in the unemployment rate among young people in the 25-34 age group by 0.0024% (in other words: an increase in output by PLN 1 million results in a 1.00024 times greater chance of being among the unemployed among young people);
- $(e^{-0,000000470761981266007} - 1) * 100\% = -0,0000047\%$ – an increase in expenditure of voivodships budgets: social assistance by 1 PLN results in a decrease of the unemployment rate among young people in the age group 25-34 by 0,0000047%;
- $e^{-12,945786} = 0,0000024$ – this means that people living in Masovia are 0.0000024 times less likely to be among the young unemployed than those living in Lodz Province.

The variables responsible for the reduction in the youth unemployment rate (and therefore the positive effect) include (see Table 3): GDI (gross disposable income per capita), NE (new entities of the national economy recorded per 10 thousand population at working age), RGR (real growth rate of regional gross value), NM_IN (net migration international for permanent residence) − so the same set except for IRNE. The variables OP (output) and EVB_SA refer to growth. The unemployment rate will be lower relative to Lodz Province in: Lesser Poland, Silesia, Greater Poland, Lower Silesia, Pomerania and Masovia.

The information criterion AIC and BIC, which assume that the lower the value the better the model, it can be seen that the difference between the values of the AIC statistics ($[Log]692,73 - [Identity]653,32 = 39,41$), and for BIC ($[Log]731,50 - [Identity]767,65 = -36,15$). From this, we conclude that the difference is too small to conclude that the model with an identity binding function is better than the binding function with a logarithmic function.

The phenomenon of overdispersion, does not occur (it is small, no steps need to be taken), as it has not exceeded the value of 2 and is still close to the value of 1 ($V = \frac{326,362256}{170} = 1,90$).

Table 3. Results of generalized linear models (Distribution: *normal*; Link function: *identity − Analysis sample*)

Effect	*UR − Test of all effects*							
						DF	Wald (Stat.)	*p*-value
Intercept						1	112,4340	0,0000
GDI						1	16,6487	0,0000
NE						1	8,0229	0,0046
OP						1	48,6678	0,0000
RGR						1	266,6253	0,0000
NM_IN						1	27,9015	0,0000
EVB_SA						1		
Regions						15	282,6512	0,0000
Effect	*UR − Parameter estimates*							
	Level of (Effect)	Column	Estimate	Standard (Error)	Wald (Stat.)	Lower CL (95,0%)	Upper CL (95,0%)	*p*-value
Intercept		1	61,2921	5,7804	112,4340	49,9628	72,6214	0,0000
GDI		2	−0,2437	0,0597	16,6487	−0,3608	−0,1266	0,0000
NE		3	−0,0302	0,0107	8,0229	−0,0512	−0,0093	0,0046
OP		4	0,0000	0,0000	48,6678	0,0000	0,0000	0,0000
RGR		5	−0,2835	0,0174	266,6253	−0,3175	−0,2495	0,0000
NM_IN		6	−0,0009	0,0002	27,9015	−0,0012	−0,0006	0,0000
EVB_SA		7	0,0000	0,0000	11,3552	0,0000	0,0000	0,0008
Regions	Lesser Poland	8	−1,3893	0,4146	11,2310	−2,2018	−0,5768	0,0008
Regions	Silesia	9	−9,6155	1,0485	84,1073	−11,6705	−7,5606	0,0000
Regions	Greater Poland	10	−5,0528	0,6125	68,0606	−6,2532	−3,8524	0,0000
Regions	West Pomerania	11	2,2587	0,5935	14,4822	1,0954	3,4220	0,0001
Regions	Lublin Province	12	1,8115	0,6352	8,1341	0,5666	3,0564	0,0043
Regions	Lower Silesia	13	−2,7614	0,5621	24,1342	−3,8631	−1,6597	0,0000
Regions	Kuyavia-Pomerania	14	1,8901	0,4586	16,9870	0,9913	2,7889	0,0000
Regions	Pomerania	15	−1,0845	0,5408	4,0213	−2,1444	−0,0245	0,0449
Regions	Holy Cross	16	6,1449	0,6572	87,4175	4,8568	7,4330	0,0000
Regions	Lubusz Province	17	4,3075	0,5595	59,2721	3,2109	5,4042	0,0000
Regions	Podlasie Province	18	7,0281	0,6818	106,2721	5,6919	8,3643	0,0000
Regions	Masovia	19	−12,9458	2,0343	40,4963	−16,9330	−8,9586	0,0000
Regions	Subcarpathia	20	5,7683	0,6281	84,3289	4,5372	6,9995	0,0000
Regions	Opole Province	21	0,8717	0,6807	1,6400	−0,4624	2,2059	0,2003
Regions	Warmia-Masuria	22	2,4253	0,5847	17,2066	1,2793	3,5712	0,0000
Scale			1,3038	0,0665		1,1797	1,4409	

Effect	*UR − Likelihood Type 1 Test*				*UR − Likelihood Type 3 Test*				
	DF	Log-likelihood	Chi-(Square)	*p*-vaule		DF	Log-likelihood	Chi-(Square)	*p*-value
Intercept	1	−531,9988							
GDI	1	−506,8478	50,3020	0,0000	**GDI**	1	−331,3485	15,9661	0,0001
NE	1	−475,5904	62,5149	0,0000	**NE**	1	−327,2953	7,8598	0,0051
OP	1	−474,2840	2,6128	0,1060	**OP**	1	−345,0539	43,3771	0,0000
RGR	1	−417,3601	113,8477	0,0000	**RGR**	1	−406,9563	167,1817	0,0000
NM_IN	1	−410,5451	13,6301	0,0002	**NM_IN**	1	−336,3911	26,0514	0,0000
IRNE	1	−410,2535	0,5830	0,4451	**EVB_SA**	1	−328,8815	11,0321	0,0009
EVB_SA	15	−323,3654	173,7763	0,0000	**Regions**	15	−410,2535	173,7763	0,0000
Regions	1	−531,9988							

UR − Statistics of goodness of fit			
	Df	Stat.	Stat/Df
Deviance	170	326,3623	1,9198
Scaled Deviance	170	192,0000	1,1294
Pearson Chi2	170	326,3623	1,9198
Scaled P. Chi2	170	192,0000	1,1294
AIC		692,7308	
AICC		699,3022	
BIC		767,6532	
Log-likelihood		−323,3654	

Significance at *p*-value = 0,05.

Parametrization: sigma-restricted (relative to Lodz Province).

Source: Own calculations.

The variables NNWP, GDP, AMGW and NM_I and IRNE are statistically insignificant, so it was decided to remove them so that they do not interfere with the model results. As before, outlier cases were verified (Fig. 4). For the identity function these are: 108 (Pomerania for 2021), 169 (Subcarpathia for 2010) and 187 (Warmia-Masuria for 2016). In both cases, these are regions of Poland characterised by lower urbanisation and therefore higher unemployment.

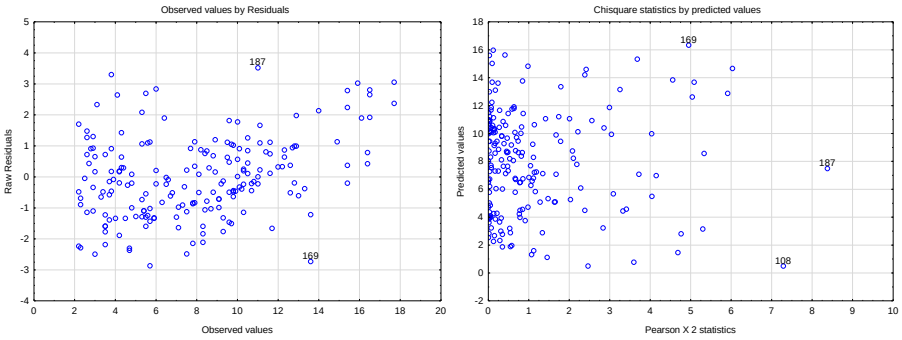


Fig. 4. Graphs of the spread of predicted versus observed values

Source: Own calculations.

The graph (Fig. 5) confirms the presence of normality of the residuals.

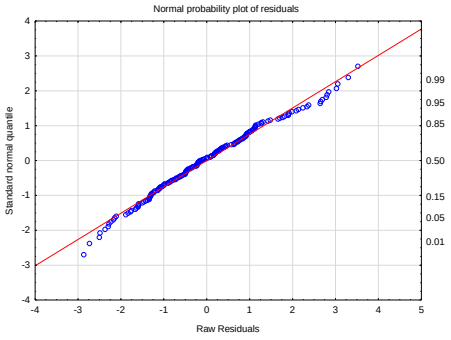


Fig. 5. Normal probability plot of residuals

Source: Own calculations.

5. Conclusions

In summary, the aim of the study was to assess the determinants of youth unemployment rates in the Polish voivodeships. For this purpose, a generalized linear model was used, allowing for the selection of factors responsible for the urbanisation of a given voivodeship. The dependent variable was the unem-

ployment rate among people in the 25-34 age group. In order to precisely verify the determinants, two models were assessed for a normal distribution with a log-bounding function and an identity function.

This allowed the identification of a group of variables that influence the decrease in the unemployment rate of people in the group 25-34, these are the variables GDI (gross disposable income per capita), NE (new entities of the national economy recorded per 10 thousand population at working age), RGR (real growth rate of regional gross value), NM_IN (net migration international for permanent residence) − for both models, for the Log function also IRNE (investment rate in national economy). For the model with the Log function the unemployment rate will be lower relative to Opole Province in: West Pomerania, Lublin Province, Kuyavia-Pomerania, Holy Cross, Lubusz Province, Podlasie Province, Subcarpathia and Warmia-Masuria. For the model with the identity function, the unemployment rate will be lower relative to Lodz Province in: Lesser Poland, Silesia, Greater Poland, Lower Silesia, Pomerania and Masovia. The remaining voivodeships assume that the unemployment rate will be higher relative to the reference variable.

The above results speak of a positive aspect. However, both two models showed that there are variables that affect negatively and therefore increase the chance of being among the unemployed. Both models refer to OP (output). Output, i.e. the value of all output produced in an economy in a year. Production drives consumption and consumption drives production. Due to the automation of labour market processes, as well as the increasing use of technology, some occupations are even considered unnecessary (e.g. potter, shoemaker; more on this is written by Kobosko [2021]; Walczak-Duraj [2022]). The result of such actions may be an increase in unemployment not only among young people, but also among older people on the labour market.

In the case of the choice of the binding function, we could observe that the differences between the two are not very great. Which suggests that the two models are a good fit, and that their use allowed us to identify the explanatory variables influencing the explanatory variable. At the same time, the comparison of the two models allowed verification of the estimated parameters.

The last decades have witnessed radical changes in the labour market − we are not only talking about technological changes, but also about regional, but above all global, competition. For young people, changes in the labour market can often be detrimental or even discriminatory in terms of various factors (for more see Trzpiot and Kawecka [2021a]; Trzpiot and Kawecka [2021b]). Observing the changes that are taking place in the market, as well as the situations of young people, the author intends to continue research in this area.

References

Aczel A.D. (2018), *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa.

Agresti A. (1990), *Categorical Data Analysis*, 2nd ed., John Wiley & Sons, https://mybiostats.files.wordpress.com/2015/03/3rd-ed-alan_agresti_categorical_data_analysis.pdf (access: 9.03.2023).

Akaike H. (1973), *Information Theory and an Extension of the Maximum Likelihood Principle* [in:] B.N. Petrov, F. Csaki (eds.), *International Symposium on Information Theory*, pp. 267-281, https://gwern.net/doc/statistics/decision/1998-akaike.pdf (access: 9.03.2023).

Dębowska O. (2007), *Migracje – wyniki aktualnych badań i analiz*, Lesser Poland Obserwatorium Rynku Pracy i Edukacji, Kraków.

Dobson A.J. (1990), *An Introduction to Generalized Linear Models*, Chapman & Hall, New York.

Drela K. (2015), *Psychologiczno-ekonomiczne problemy bezrobocia*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Zeszyty Naukowe nr 851, Współczesne Problemy Ekonomiczne nr 10, pp. 133-145.

Gill J. (2000), *Generalized Linear Models: A Unified Approach*, Thousand Oaks, Sage Publications.

Golnau W., Kalinowski M., eds. (2007), *Zarządzanie zasobami ludzkimi*, wyd. 3, Wydawnictwo CeDeWu, Warszawa.

Green P.J., Silverman B.W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, New York.

Grotowska-Leder J. (2015), *Wybrane aspekty teoretyczne, metodologiczne i empiryczne analizy bezrobocia ludzi młodych: perspektywa Unii Europejskiej* [in:] K. Górniak, T. Kanasz, B. Pasamonik, J. Zalewska (eds.), *Socjologia czasu, kultury i ubóstwa. Księga jubileuszowa dla profesor Elżbiety Tarkowskiej*, Wydawnictwo APS, pp. 218-231.

ILO (International Labour Organization) (2020), *Preventing Exclusion from the Labour Market: Tackling the COVID-19 Youth Employment Crisis*, Policy brief, https://sdgs.un.org/sites/default/files/documents/26635youth_covid_policy _brief.pdf (access: 27.02.2023).

Karwacki A., Błędowski P. (2020), *Bezrobocie jako współczesna kwestia społeczna – wybrane aspekty socjologiczne i ekonomiczne*, Instytut Filozofii i Socjologii Polskiej Akademii Nauk, „Studia Socjologiczne", nr 1(236), Warszawa, pp. 135-164.

Kobosko M. (2021), *Ginące zawody jako konsekwencja zmian technologicznych na polskim rynku pracy*, „Studia z Polityki Publicznej", nr 8(4(32), pp. 75-95.

Kostrzewski S., Worach-Kardas H. (2013), *Skutki długotrwałego bezrobocia dla zdrowia i jakości życia osób w starszym wieku produkcyjnym*, Oddział Zdrowia Publicznego, Wydział Nauk o Zdrowiu, Uniwersytet Medyczny w Łodzi, Kierownik: prof. dr hab. n.med. Tomasz Kostka, „Nowiny Lekarskie", nr 82(4), pp. 310-317.

Kozłowska J. (2022), *Sytuacja młodych ludzi na rynku pracy*, Zespół Szkół Górniczo-Energetycznych im. S. Staszica w Koninie, Zeszyty Naukowe ZPSB „Firma i Rynek", nr 1(61), pp. 73-84.

McCullagh P., Nelder J.A. (1989), *Generalized Linear Models* (2nd ed.), Chapman & Hall, New York.

Nelder J.A., Wedderburn R.W.M. (1972), *Generalized Linear Models*, "Journal of the Royal Statistical Society: Series A (General)", Vol. 135, Iss. 3, pp. 370-384.

O'Higgins N. (2011), *The Impact of the Economic and Financial Crisis on Youth Employment: Measures for Labour Market Recovery in the European Union*, Canada and the United States, Employment Working Paper No. 70, ILO, Geneva.

Olson C.L. (1974), *Comparative Robustness of Six Tests Multivariate Analysis of Variance*, "Journal of the American Statistical Association", Vol. 69, pp. 894-908.

Organiściak-Krzykowska A., Hrynkiewicz J., eds. (2022), *Depopulacja w ujęciu lokalnym*, Rządowa Rada Ludnościowa, Materiały z III Kongresu Demograficznego. Część 3, Zakład Wydawnictw Statystycznych, Warszawa.

Pasternak-Malicka M. (2013), *Przyczyny i skutki migracji zagranicznych młodych Polaków*, „Zeszyty Naukowe Uniwersytetu Szczecińskiego", nr 7, pp. 177-188.

Piecuch T. (2013), *Przedsiębiorczość. Podstawy teoretyczne*, wyd. 2, C.H. Beck, Warszawa.

Ptak-Chmielewska A. (2013), *Uogólnione modele liniowe*, wyd. 1, Wydawnictwo Oficyna Wydawnicza SGH, Warszawa.

Siek E., Bednarczyk J.L. (2009), *Kryzys ekonomiczny a migracje ludności Polski do wybranych krajów*, „Rocznik Żyrardowski", nr 7, pp. 164-166.

Statistics Poland (2022a), GUS – Bank Danych Lokalnych, *Dane według dziedzin*, https://bdl.stat.gov.pl/ (access: 10.03.2023).

Statistics Poland (2022b), GUS − Główny Urząd Statystyczny, *Informacje statystyczne: Rachunki kwartalne produktu krajowego brutto w latach 2017-2021*, Warszawa, https://stat.gov.pl/obszary-tematyczne/rachunki-narodowe/kwartalne-rachunki-narodowe/rachunki-kwartalne-produktu-krajowego-brutto-w-latach-2017-2021,6,16.html (access: 10.03.2023).

Statistics Poland (2022c), GUS – Główny Urząd Statystyczny, *Rachunki narodowe. Wstępny szacunek produktu krajowego brutto w 4 kwartale 2021 roku*, https://stat.gov.pl/obszary-tematyczne/rachunki-narodowe/kwartalne-rachunki-narodowe/wstepny-szacunek-produktu-krajowego-brutto-w-4-kwartale-2021-roku,3,78.html (access: 10.03.2023).

Statistics Poland (2022d), GUS – Główny Urząd Statystyczny, *Pracujący i wynagrodzenia w gospodarce narodowej w 2021 r. – dane ostateczne*, https://tiny.pl/w94kj (access: 10.11.2023).

StatSoft Electronic Statistics Textbook (2023), Ogólne modele liniowe (GLM), Model z sigma-ograniczeniami a model przeparametryzowany, https://www.statsoft.pl/textbook/stathome_stat.html?https%3A%2F%2Fwww.statsoft.pl%2Ftextbook%2Fstglm.html (access: 22.09.2023).

Stanisz A. (2007), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 2. Modele liniowe i nieliniowe*, Wydanie drugie, zmienione i poprawione, Wydawnictwo StatSoft Polska Sp. z o.o., Kraków.

Subocz E. (2022), *Wpływ pandemii COVID-19 na sytuację osób młodych na europejskim rynku pracy – wybrane aspekty*, Sieć Badawcza Łukasiewicz – Instytut Technologii Eksploatacji, „Edukacja Ustawiczna Dorosłych", nr 2, pp. 43-54.

Trzpiot G., Kawecka M. (2021a), *Evaluation of the Labor Market Status of Young People in Selected Countries of the European Union − The Multiple Regression Approach* [in:] G. Trzpiot (ed.), *Modeling of Complex Data Sets and Risk Analysis*, Publishing House of the University of Economics, Katowice, pp. 46-73.

Trzpiot G.A., Kawecka M. (2021b), *Description of the Labour Market Status of Young People in Selected Countries of the European Union − The Taxonomic Approach*, Ekonometria, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Vol. 25, No. 4, pp. 17-39.

Uścińska G., Wiśniewski Z., eds. (2022), *Zmiany demograficzne a rynek pracy i ubezpieczenia społeczne*, Rządowa Rada Ludnościowa, Materiały z III Kongresu Demograficznego. Część 4, Zakład Wydawnictw Statystycznych, Warszawa.

Verick S. (2009), *Who Is Hit Hardest during a Financial Crisis? The Vulnerability of Young Men and Women to Unemployment in an Economic Downturn*, IZA Discussion Paper No. 4359, Bonn.

Vonesh E.F. (2012), *Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS*, Publisher SAS Institute Inc.

Walczak-Duraj D. (2022), *Zmiany współczesnej pracy, zawodów i profesji*, „Acta Universitatis Lodziensis. Folia Sociologica", Iss. 81, pp. 5-27.

Chapter VIII

Time series analysis of the number of COVID-19 cases during the pandemic in selected countries

Zuzanna Krysiak, Grażyna Trzpiot

1. Introduction

The topic of the chapter is the analysis of time series for observations describing the phenomenon of the SARS CoV-19 virus pandemic. The analysis was conducted for countries located in Europe (Poland, Italy), America (Chile, Mexico) and Asia (India, Israel). The countermeasure to the violent outbreak of the pandemic was the introduction of vaccines against the COVID-19 virus. The number of vaccinated people and the current number of people who have fallen ill are the main observations on which the research is based. These factors showed the nature of time series, as a result of which further attempts were made to analyze.

However, there are many determinants of the course of the disease, which are defined later in this work. It is important to notice the problem of the dependence of the number of people vaccinated in a given society and the course of the disease and the number of deaths and morbidity among the elderly, or the effectiveness of the vaccinations introduced. Each of the analyzed countries was selected due to the different methods of dealing with the pandemic and significantly different ones, e.g. in terms of aging of the population, geographic location or restrictions related to the pandemic introduced by a given country. The effectiveness of vaccines has been questioned both because of the large number of choices given by producers and the development and flare-up of the pandemic over repeated periods of time.

The aim of the chapter was to collect and analyze the variables describing the phenomenon of the COVID-19 virus pandemic depending on the vaccines introduced, and to compare the interdependent observations for selected countries. The analysis of the time series allowed for an in-depth study of the pandemic phenomenon and the creation of models that allow for further analysis or forecasting of observations describing the incidence and the current vaccination doses administered.

2. Identification of the structure of the time series

Time series analysis has two main goals: detecting the nature of the phenomenon represented by the sequence of observations and forecasting (predicting future values of the time series) [Alsan, 2020]. The following analysis focuses on the observation of the disease course of the COVID-19 virus and its variants in given periods of time with a daily frequency and the impact of introducing vaccinations in the following countries: Poland, Italy, India, Israel, Chile, Mexico.

The data presented in the paper relate to a one-dimensional time series, the variables of which Y are ordered by the value of the time variable *t*. The period of time in which the phenomenon was investigated ranges from March 1, 2020 to January 22, 2022. For the analysis of the time series, the observation informing about the current cases of disease was selected due to the correlation with the current vaccination doses administered, shown in the previous cluster analyzes [Stellwagen, 2013]. The new_cases_per_million variable also showed differences in relation to the observations concerning the current tests performed and the current number of deaths, but these observations were strongly dependent on each other, and therefore it was selected for the further part of the analysis.

In the time series analysis, the following components are distinguished: systematic and random[1]. As part of the estimation of information on the stochastic process determining the stationarity of the series, the following are analyzed: the occurrence of the trend and seasonality of the series [Fanelli, Piazza, 2020]. Diagnostics of the order of the processes runs through the evaluation by observation of the order of the ACF function and the PACF function. The same parameter was analyzed for the following countries: Poland, Italy, Chile, Mexico, India and Israel.

The chart below (Fig. 1) shows the time series for daily data in Poland for the current number of sick people. Then the functions of ACF autocorrelation and partial PACF autocorrelation are presented. The ACF function decreases exponentially with the increase of the p parameter, therefore it is known that there is a trend in the process examined below. From the diagram of the partial autocorrelation, the order p of the partial autoregression process was read, which was $p = 7$ with the simultaneous observation of the order $p = 1$. The value of the autocorrelation of the order $p = 1$ is close to 1. The unit root is probable for the current number of cases [Trzpiot, 2017]. This may indicate a trend. The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of monthly seasonal fluctuations.

[1] The estimation methods included in the chapter are based on recursive methods using the Statistica package.

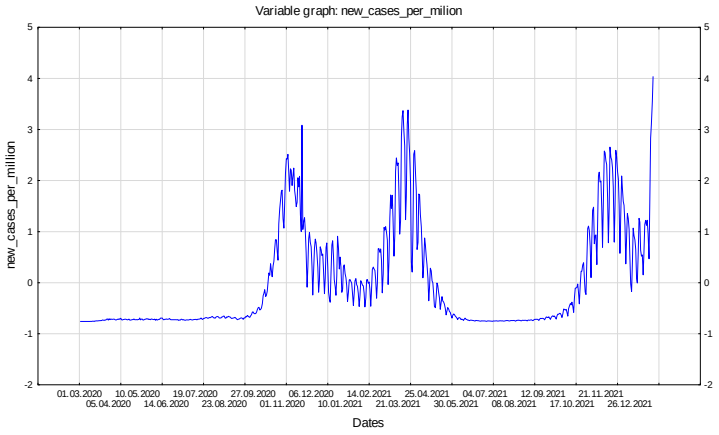


Fig. 1. Variable plot for Poland data

Source: Own elaboration.

After making the differentiation for the first time, the trend was eliminated. However, after two-fold differentiation (Fig. 2) D (−1), D (−7), the seasonality was eliminated from the time series.

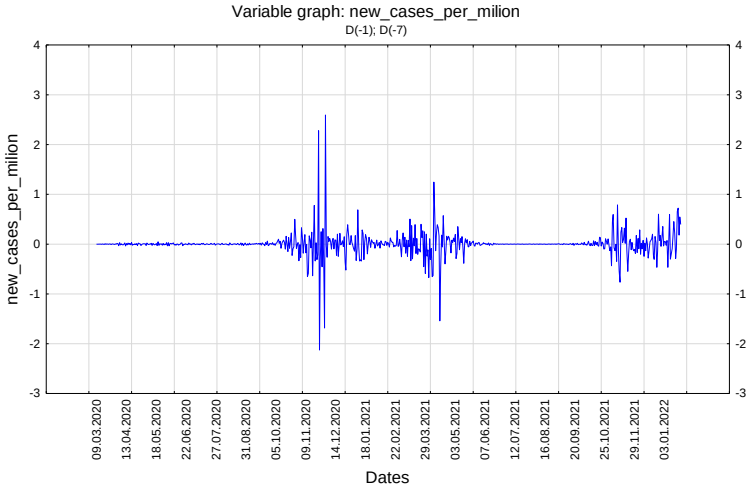


Fig. 2. Diversification for the trend and seasonality for Poland data

Source: Own elaboration.

After differentiation functions ACF and PACF, the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model.

The graph below (Fig. 3), relating to the data observed in Italy, shows the time series for the daily data in Italy for the current number of sick people. The ACF function decreases exponentially with the increase of the *p* parameter, therefore it is known that there is a trend in the process examined. The *p*-order of the partial autoregression process was read, which was $p = 7$ or $p = 9$ with the simultaneous observation of the $p = 1$ order. The value of the autocorrelation of the order $p = 1$ is close to 1. The unit root is probable for the current number of cases. This may indicate a trend.

The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of monthly seasonal fluctuations [Nowak, 2007].

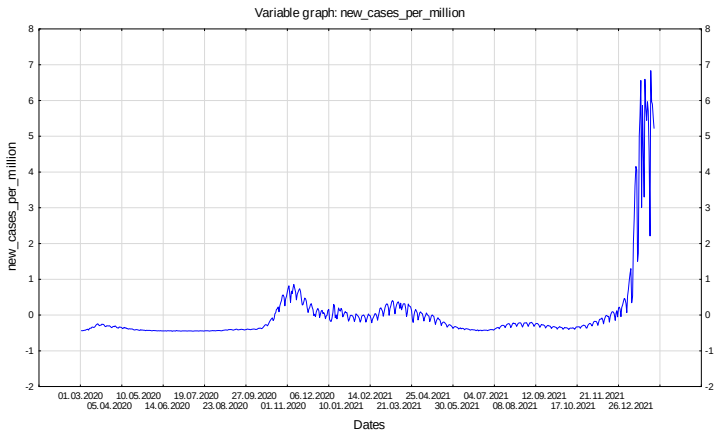


Fig. 3. Variable chart for Italy

Source: Own elaboration.

After making the differentiation for the first time, the trend was eliminated. However, after two-fold differentiation (Fig. 4.) D (−1), D (−7), the seasonality was eliminated from the time series.

Functions after differentiation: ACF and PACF the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model.

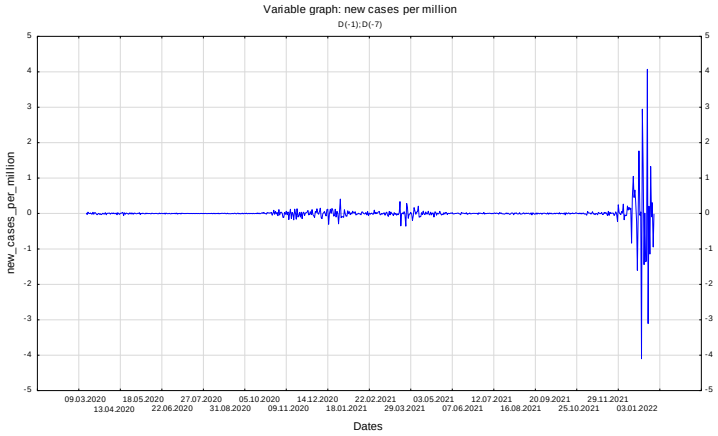


Fig. 4. Differentiation for trend and seasonality for data Italy

Source: Own elaboration.

The graph below (Fig. 5), concerning the data observed in Chile, shows the time series for the daily data in Chile for the current number of sick people. Then, the functions of the ACF autocorrelation and the PACF partial autocorrelation were analyzed. The ACF function decreases exponentially with the increase of the *p* parameter, therefore it is known that there is a trend in the process examined below. From the diagram of the partial autocorrelation the *p*-order of the partial autoregression process was read, which was $p = 7$ or $p = 8$ with the simultaneous observation of the $p = 1$ order. The value of the autocorrelation of the order $p = 1$ is close to 1. The unit root is probable for the current number of cases.

The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of seasonal monthly fluctuations.

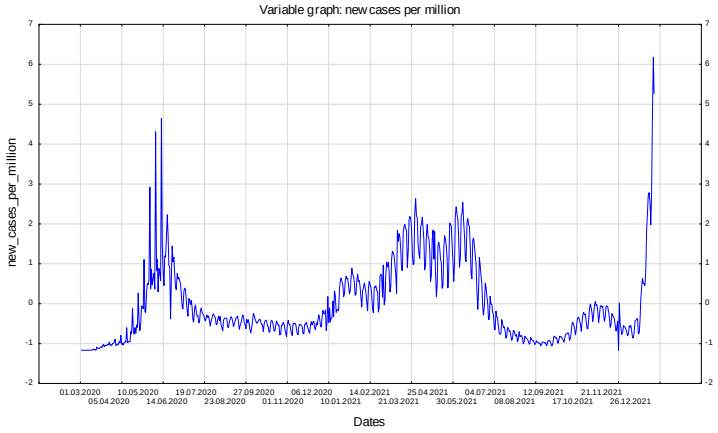


Fig. 5. Variable plot for Chile data

Source: Own elaboration.

After making the differentiation for the first time the trend was eliminated. However, after a two-fold differentiation (Fig. 6) D (−1), D (−7), the seasonality was eliminated from the time series [Luszniewicz and Słaby, 2001].

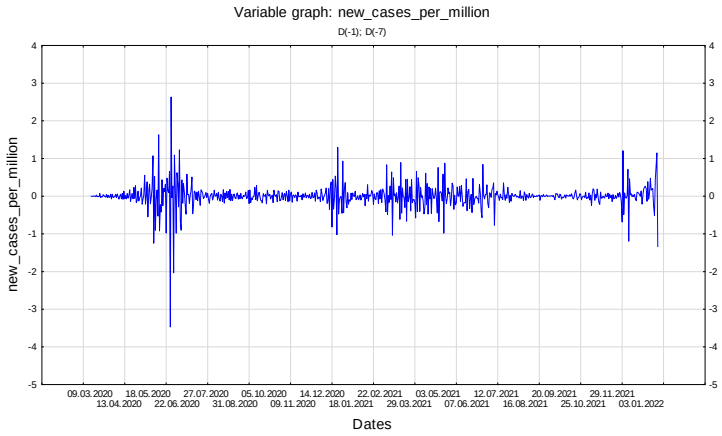


Fig. 6. Differentiation for trend and seasonality for Chile data

Source: Own elaboration.

Functions after differentiation: ACF and PACF the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model.

The graph below (Fig. 7), concerning the data observed in Mexico, shows the time series for the daily data in Mexico for the current number of sick people. The ACF function decreases exponentially with the increase of the *p* parameter, therefore it is known that there is a trend in the process examined below. The order *p* of the partial autoregression process was read from the diagram of the partial autocorrelation which was $p = 7$.

The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of seasonal monthly fluctuations.

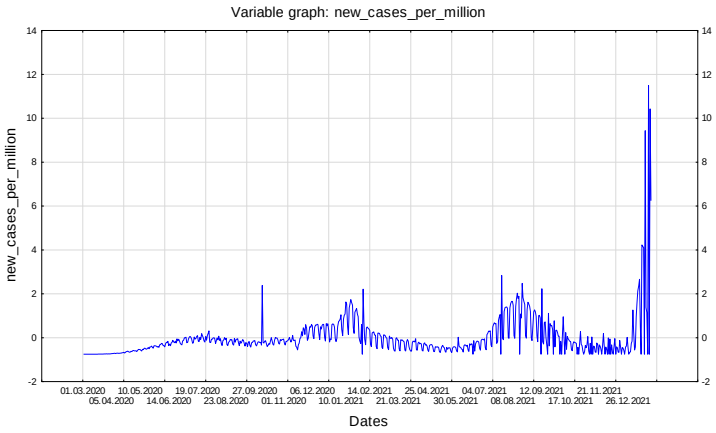


Fig. 7. Variable plot for Mexico data

Source: Own elaboration.

After making the differentiation for the first time the trend was eliminated. However, after a two-fold differentiation (Fig. 8) D (−1), D (−7), the seasonality was eliminated from the time series.

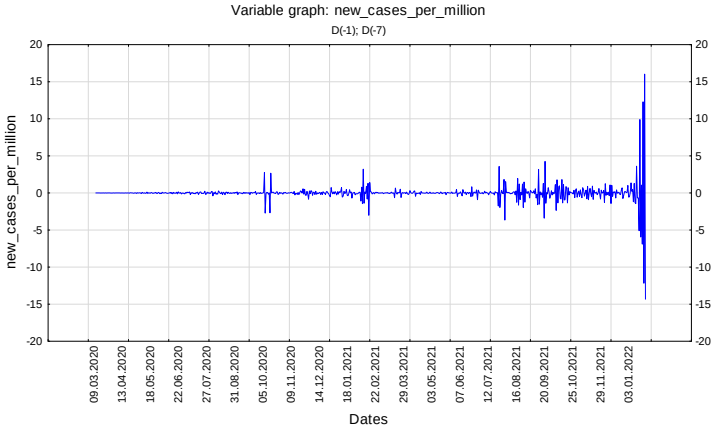


Fig. 8. Differentiation for the trend and seasonality for Mexico data

Source: Own elaboration.

Functions after differentiation: ACF and PACF the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model. The value of the autocorrelation of the order $p = 1$ is close to 1. The unit root is probable for the current number of cases. This may indicate a trend.

The graph below (Fig. 9), concerning the data observed in India, shows the time series for the daily data in India for the current number of sick people. The ACF function decreases exponentially with the increase of the *p* parameter, therefore it is known that there is a trend in the process examined below. The order *p* of the partial autoregression process was read from the diagram of the partial autocorrelation which was $p = 1$. The value of the autocorrelation of the order $p = 1$ is close to 1. The unit root is probable for the current number of cases. This may indicate a trend and in this case long-term changes in an upward direction.

The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of seasonal monthly fluctuations.

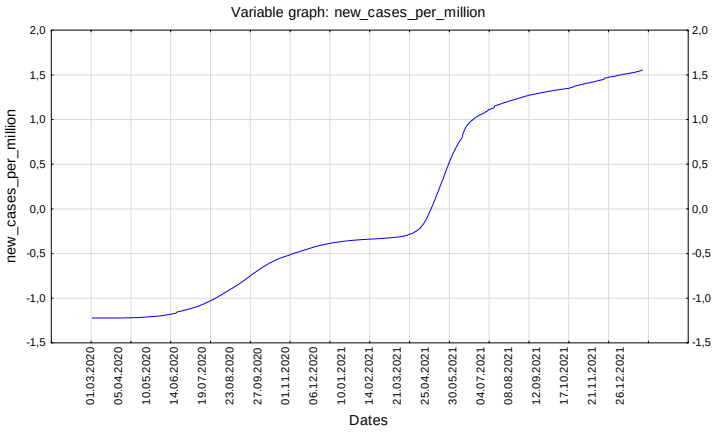


Fig. 9. Variable chart for India

Source: Own elaboration.

After making the differentiation for the first time the trend was eliminated. However, after a two-fold differentiation (Fig. 10) D (−1), D (−7), the seasonality was eliminated from the time series.

Functions after differentiation: ACF and PACF the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model.

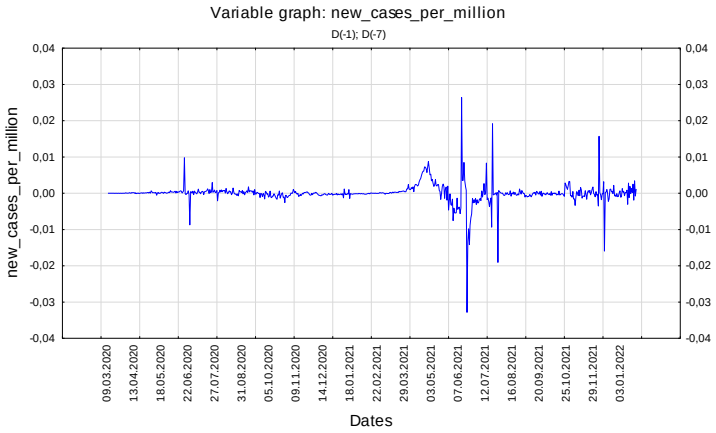


Fig. 10. Diversification for the trend and seasonality for India data

Source: Own elaboration.

The graph below (Fig. 11), concerning the data observed in Israel, shows the time series for daily data in Israel of the current number of sick people. Then, the functions of the ACF autocorrelation and the PACF partial autocorrelation have been designated. The ACF function decreases exponentially with the increase of the *p* parameter, therefore it is known that there is a trend in the process examined below. The order *p* of the partial autoregression process was read from the diagram of the partial autocorrelation which was $p = 7$. The next step in the analysis is the application of the time series differentiation function, where the trend has been eliminated and the ACF function indicated the occurrence of seasonal monthly fluctuations [Luszniewicz and Słaby, 2001].

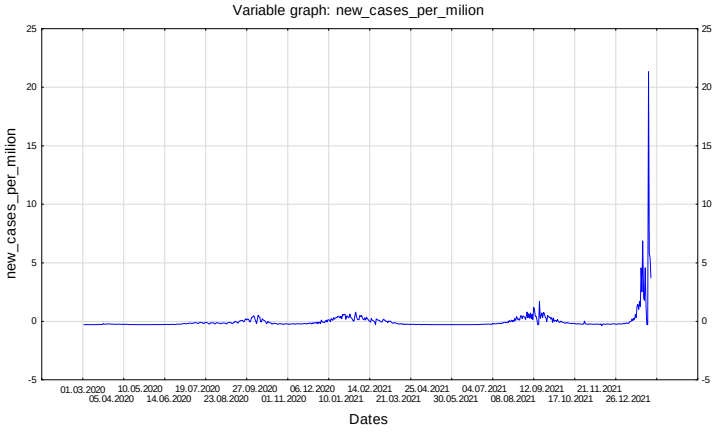


Fig. 11. Variable plot for Israel data

Source: Own elaboration.

After making the differentiation for the first time the trend was eliminated. However, after two-fold differentiation (Fig. 12) D (−1), D (−7), the seasonality was eliminated from the time series.

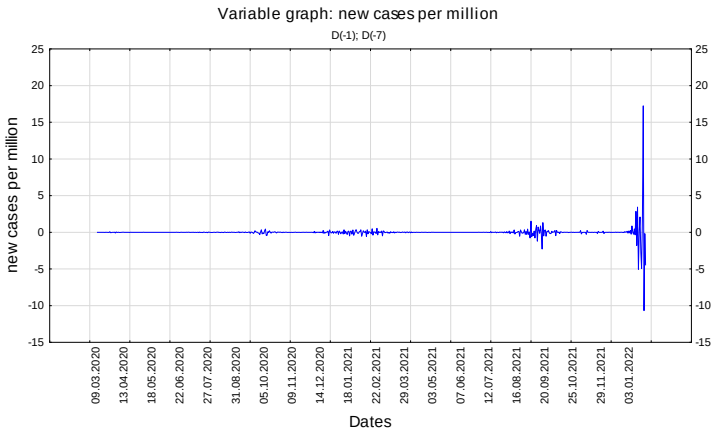


Fig. 12. Differentiation for trend and seasonality for Israel data

Source: Own elaboration.

Functions after differentiation: ACF and PACF the autocorrelation function is still important. The series is not random, there is no white noise yet. It is necessary to estimate parameters and select an appropriate ARMA model.

The analysis of the series structure for Poland and Italy, based on the same parameter of current cases, showed similar results. The time series for Poland may show more seasonality. Structure analysis for Chile and Mexico also showed similar results with the possibility of more seasonality for Chile [Mohammadi, 2021]. There are significant differences in the series structure analysis for India and Israel. For India, the value of the autocorrelation before differentiation was 1, and for Israel, similarly to the previous cases, it was 7. The number of current cases for India is less varied than in the case of data for Israel. This may be due to errors in the database or an external factor influencing the number of cases of disease.

3. ARIMA model and estimation of ARIMA model parameters

In the study, the analysis of time series with the use of ARIMA models was carried out in order to forecast the phenomenon of the dependencies between the COVID-19 data. The Box-Jenkins method was used to create the analysis, which consists in comparing the ACF and PACF functions for a specific stationary series with the theoretical forms of these functions for the AR (p) and MA (q) models.

AR (p) relates to the PACF chart, while MA (q) to the ACF chart. The time series covers data from March 1, 2020 to January 22, 2022. A week in series runs from Monday to Sunday.

3.1. Analysis of ARIMA models with seasonality for selected countries

Data analyzed for the time series of observations from Poland. The graphs show the fitted autocorrelation function (Fig. 13) and the partial autocorrelation function, on the basis of which the analyzes allowing to create the ARIMA model were performed. In order to eliminate the trend, differentiation against the first-order trend and one-time differentiation due to the seasonality of the seventh order were used [Nazarko and Chodakowska, 2022].

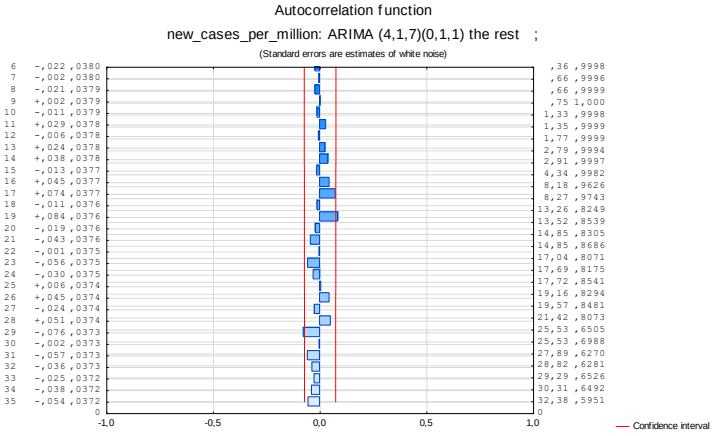


Fig. 13. ACF function for data Poland

Source: Own elaboration.

The ARIMA model parameters selected for analysis are $p = 4$, $q = 7$, as well as $P = 0$ and $Q = 1$. Table 1 presents the estimated parameters. The parameters marked in red are significant because they meet the $p < 5\%$ condition. The resulting AR model was created when autoregressive delay 4 was selected and the MA model was determined by the parameter q from the moving average equal to 7. The autocorrelation function and partial autocorrelation (Fig. 14) of the estimated model is within the confidence interval. Only a single parameter borders on the interval, but due to the very large diversity of data, the resulting model is correctly adjusted in terms of the ACF and PACF functions.

The graph (Fig. 15) shows the mean of the residuals for the estimated model and the normality plot which only after logarithm of these residuals has a normal distribution (Fig. 16).

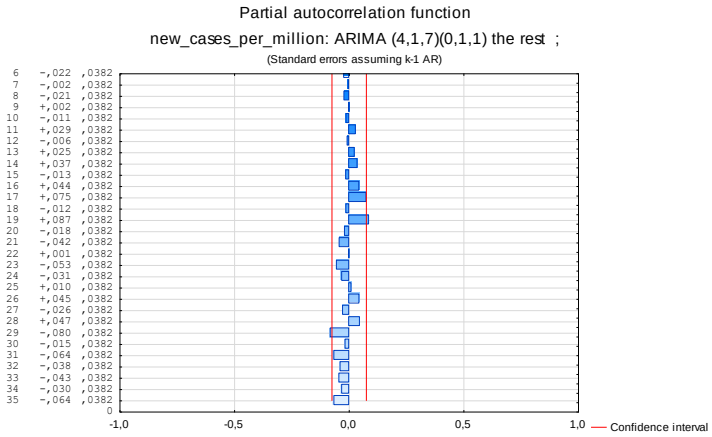


Fig. 14. PACF function for data Poland

Source: Own elaboration.

Table 1. ARIMA Poland model

	Data: new_cases_per_million Transformations: D(1), D(7) Model: (4, 1, 7) (0, 1, 1) Season delay.: 7 Residual MS = 0.04233					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	−0.766909	0.148151	−5.17655	0.000000	−1.05780	−0.476016
p(2)	−0.702138	0.162881	−4.31075	0.000019	−1.02195	−0.382323
p(3)	−0.533456	0.158433	−3.36708	0.000803	−0.84454	−0.222374
p(4)	−0.347464	0.091762	−3.78656	0.000166	−0.52764	−0.167289
q(1)	−0.671533	0.144824	−4.63690	0.000004	−0.95589	−0.387172
q(2)	−0.399408	0.152597	−2.61739	0.009059	−0.69903	−0.099783
q(3)	−0.292844	0.122898	−2.38282	0.017457	−0.53415	−0.051534
q(4)	−0.090927	0.076161	−1.19389	0.232943	−0.24047	0.058614
q(5)	0.180140	0.066963	2.69012	0.007320	0.04866	0.311622
q(6)	−0.039077	0.056741	−0.68869	0.491257	−0.15049	0.072334
q(7)	0.409291	0.088170	4.64207	0.000004	0.23617	0.582413
Qs(1)	0.057157	0.076492	0.74723	0.455187	−0.09303	0.207348

Source: Own elaboration.

Table 1 shows the descriptive statistics for the rest of the current cases. Based on the PACF function, the residual normality plot, and the expected mean value, the distribution is normal. In order to better fit the model or the need for forecasting, another model can be selected. The resulting model is the ARIMA model with seasonality, i.e. the SARIMA model is described in the next section.

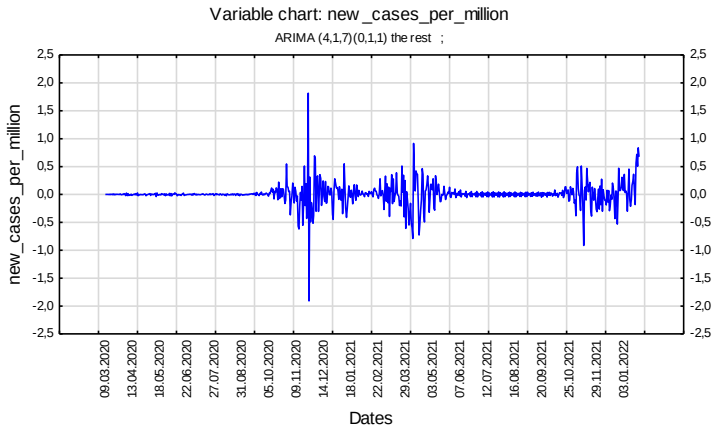


Fig. 15. ARIMA of residuals for data Poland

Source: Own elaboration.

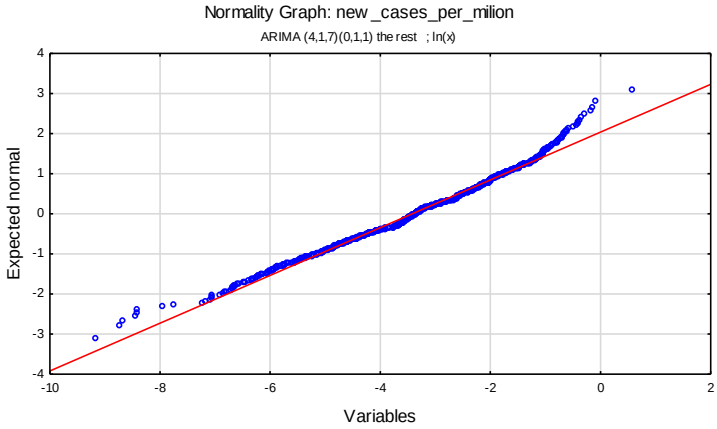


Fig. 16. Residual normality chart for Poland data

Source: Own elaboration.

Data analyzed for the time series of observations from Italy. The graphs show the fitted autocorrelation function (Fig. 17) and the partial autocorrelation function (Fig. 18) on the basis of which the analyzes allowing to create the ARIMA model were performed. In order to eliminate the trend, the seasonality was differentiated against the first-order trend and one-time differentiation was applied due to the seasonality of the seventh order.

Table 2. ARIMA Italy model

	Data: new_cases_per_million Transformations: D(1), D(7) Model: (3,1,6)(1,1,1) Season delay.: 7 Residual MS = 0.05339					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	−0.758765	0.000000	−4.29567	0.000000	−0.758765	−0.758765
p(2)	−0.481537	0.068611	−7.01838	0.000000	−0.616253	−0.346820
p(3)	−0.722761	0.000000	−4.29567	0.000000	−0.722761	−0.722761
q(1)	−0.015380	0.049877	−3.08359	0.757904	−0.113313	0.082553
q(2)	0.150496	0.074367	2.02369	0.043395	0.004477	0.296515
q(3)	−0.339966	0.051735	−6.57125	0.000000	−0.441547	−0.238384
q(4)	0.416150	0.035769	1.16343	0.000000	0.345917	0.486382
q(5)	−0.260897	0.059206	−4.40659	0.000012	−0.377147	−0.144646
q(6)	−0.300761	0.073521	−4.09082	0.000048	−0.445119	−0.156403
Ps(1)	0.323830	0.112848	2.86961	0.004239	0.102254	0.545405
Qs(1)	0.679014	0.098284	6.90869	0.000000	0.486034	0.871994

Source: Own elaboration.

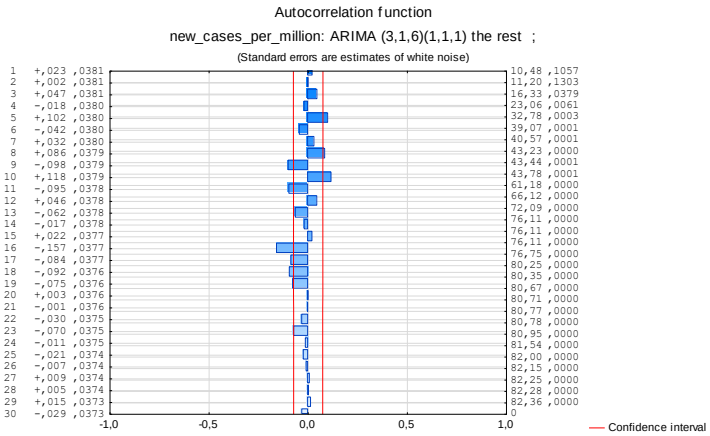


Fig. 17. ACF function for data Italy

Source: Own elaboration.

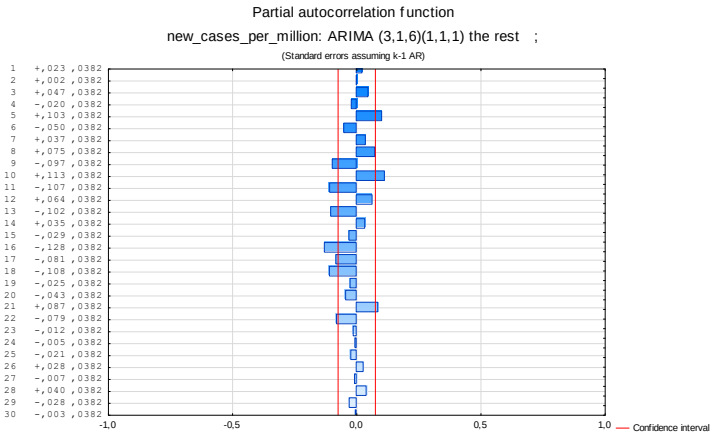


Fig. 18. PACF function for data Italy

Source: Own elaboration.

The ARIMA model parameters selected for analysis are $p = 4$, $q = 6$, as well as $P = 1$ and $Q = 1$, which results in the table above (Table 2). The resulting AR model was created when autoregressive delay 4 was selected and the MA model was determined by the parameter q of the moving average equal to 6. The autocorrelation and partial autocorrelation function of the estimated model is within the confidence interval. Only individual parameters border the range, but due to the very large diversity of data, the resulting model is correctly adjusted in terms of the ACF and PACF functions.

The graph (Fig. 19) shows the mean of the residuals for the estimated model and the normality graph (Fig. 20), which only after logarithm of these residuals has a normal distribution. Based on the PACF function, the residual normality plot, and the expected mean value, the distribution is normal. In order to better fit the model, another model should be selected. This model is not available in Statistica [Stefanowski, 2009]. The resulting model is the ARIMA model with seasonality, i.e. the SARIMA model, which is described in the next section.

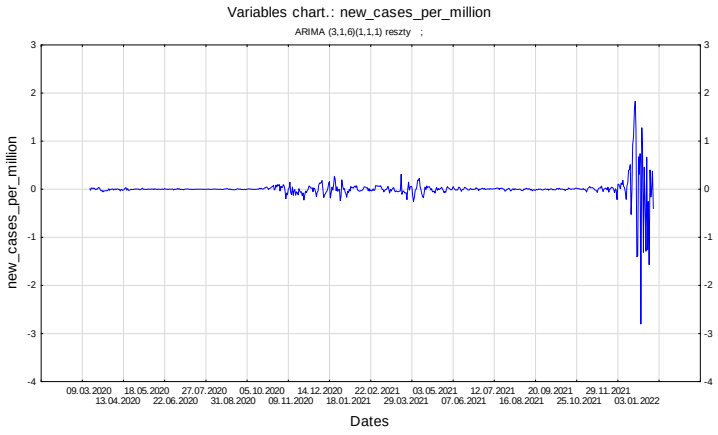


Fig. 19. ARIMA residuals for Italy data

Source: Own elaboration.

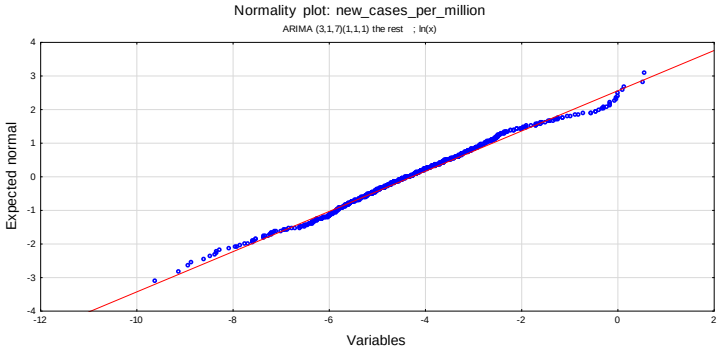


Fig. 20. Residual normality plot for data Italy

Source: Own elaboration.

Data analyzed for the time series of observations from Chile. The graphs show the fitted autocorrelation function (Fig. 21) and the partial autocorrelation function (Fig. 22), on the basis of which the analyzes allowing for the creation of the ARIMA model were performed. In order to eliminate the trend, differentiation against the first-order trend and one-time differentiation due to the seasonality of the seventh order were used.

Table 3. ARIMA Chile model

	Data: new_cases_per_million Transformations: D(1), D(7) Model: (3,1,4)(0,1,0) Residual MS = 0.06514					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	−0.760457	0.032075	−23.7087	0.000000	−0.823435	−0.697478
p(2)	0.728571	0.047440	15.3579	0.000000	0.635424	0.821717
p(3)	0.839891	0.031118	26.9901	0.000000	0.778791	0.900991
q(1)	−0.034000	0.032889	−1.0338	0.301602	−0.098577	0.030576
q(2)	1.374060	0.031157	44.1014	0.000000	1.312884	1.435235
q(3)	0.303391	0.022765	13.3268	0.000000	0.258692	0.348090
q(4)	−0.811314	0.028561	−28.4066	0.000000	−0.867392	−0.755236

Source: Own elaboration.

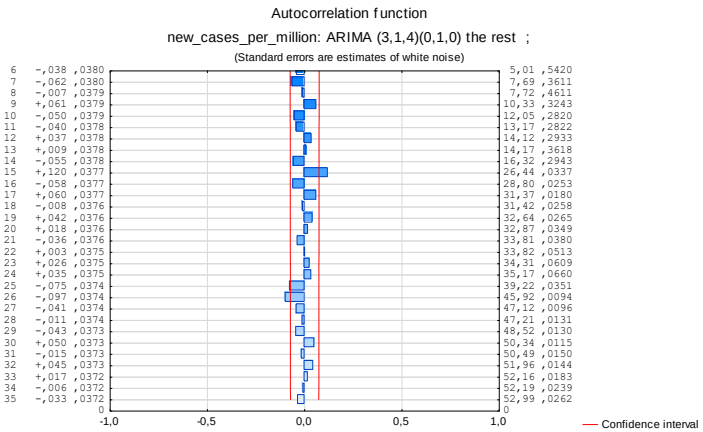


Fig. 21. ACF function for Chile data

Source: Own elaboration.

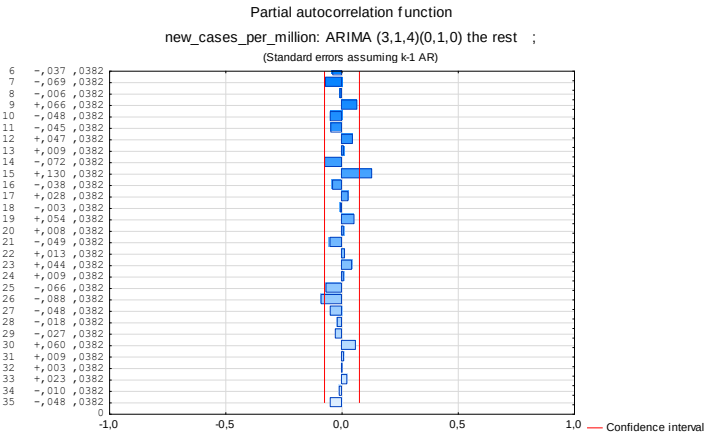


Fig. 22. PACF function for Chile data

Source: Own elaboration.

The ARIMA model parameters selected for analysis are $p = 3$, $q = 4$, as well as $P = 0$ and $Q = 0$. Parameters for the models are presented in the Table 3. The AR model was created when autoregressive delay 3 was selected and the MA model was determined by the parameter q of the moving average equal to 4. The autocorrelation and partial autocorrelation function of the estimated model is within the confidence interval. Only a single parameter borders on the interval, but due to the very large diversity of data, the resulting model is correctly adjusted in terms of the ACF and PACF functions.

The graph (Fig. 23) shows the mean of the residuals for the estimated model and the normality graph (Fig. 24), which has a normal distribution only after logarithm of these residuals. Based on the PACF function, the residual normality plot, and the expected mean value, the distribution is normal. In order to better fit the model, another model should be selected.

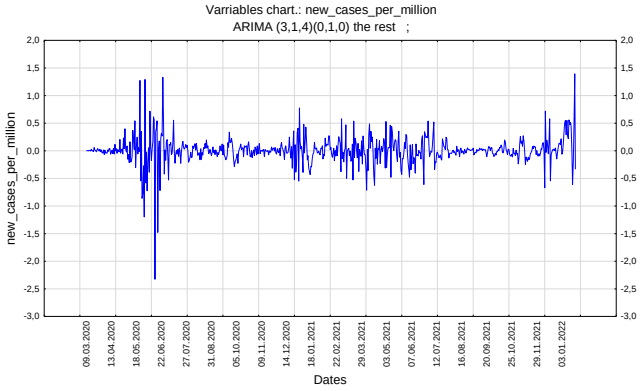


Fig. 23. ARIMA of residuals for Chile data

Source: Own elaboration.

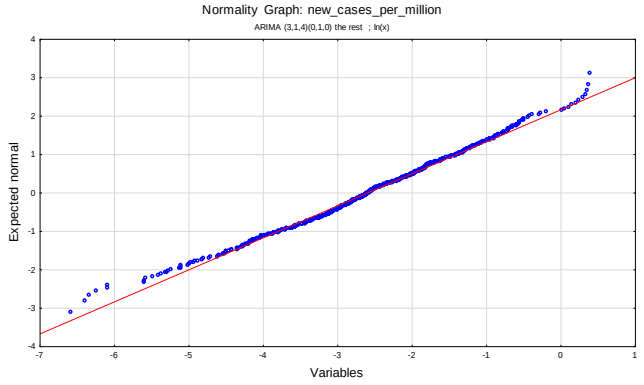


Fig. 24. Residual normality plot for Chile data

Source: Own elaboration.

Data analyzed for the time series of observations from Mexico. The graphs show the fitted autocorrelation function (Fig. 25) and the partial autocorrelation function (Fig. 26), on the basis of which the analyzes allowing to create the ARIMA model were performed. In order to eliminate the trend, the seasonality was differentiated against the first-order trend and one-time differentiation was applied due to the seasonality of the seventh order.

Table 4. ARIMA Mexico model

	Data: new_cases_per_million Transformations: D(2) Model: (2,1,3)(1,0,1) 1) Season delay.: 7 Residual MS = 0.28423					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	−1.06079	0.041277	−25.6992	0.000000	−1.14184	−0.979746
p(2)	−0.40843	0.041685	−9.7979	0.000000	−0.49028	−0.326583
q(1)	−0.99042	0.028513	−34.7358	0.000000	−1.04641	−0.934439
q(2)	0.45386	0.045709	9.9292	0.000000	0.36411	0.543605
q(3)	0.80644	0.028378	28.4181	0.000000	0.75072	0.862153
Ps(1)	0.99964	0.039876	25.0685	0.000000	0.92134	1.077931
Qs(1)	0.19359	0.051546	3.7556	0.000188	0.09238	0.294796

Source: Own elaboration.

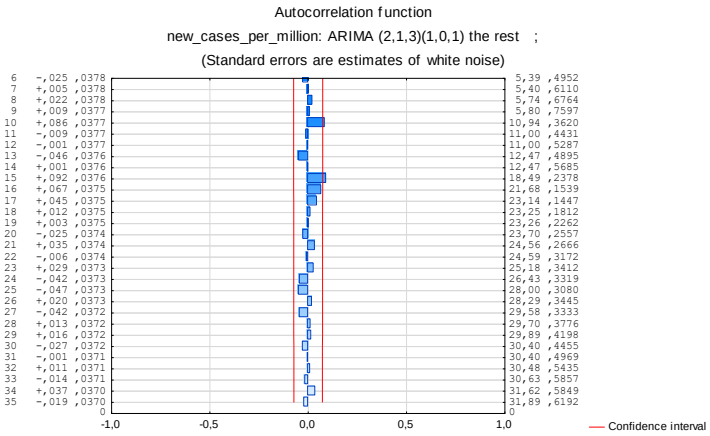


Fig. 25. ACF function for data Mexico

Source: Own elaboration.

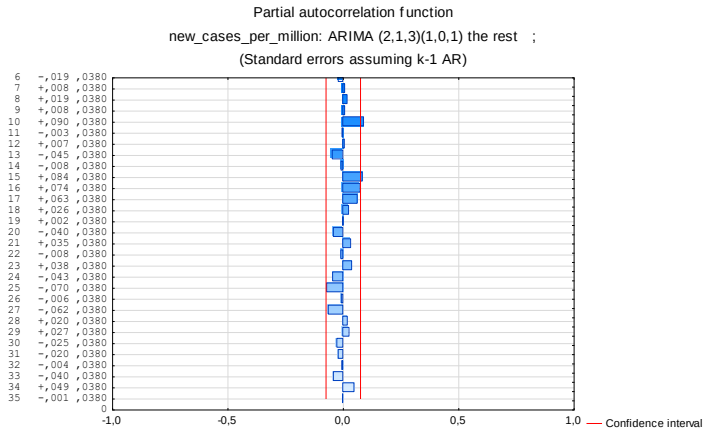


Fig. 26. PACF function for data Mexico

Source: Own elaboration.

Table 4 represents parameters for ARIMA model with seasonality. The ARIMA model parameters selected for analysis are $p = 2$, $q = 3$, as well as $P = 1$ and $Q = 1$. The AR model was created when autoregressive delay 2 was selected and the MA model was determined by the parameter q of the moving average equal to 3. The autocorrelation and partial autocorrelation function of the estimated model is within the confidence interval [Sokołowski, 2003]. The resulting model is correctly fitted in terms of ACF and PACF functions.

The graph (Fig. 27) shows the mean of the residuals for the estimated model and the normality graph (Fig. 28), which only after logarithm of these residuals has a normal distribution. Based on the PACF function, the residual normality plot, and the expected mean value, the distribution is normal. In order to better fit the model, another model should be selected [Liu, 2021]. This model is not available in Statistica.

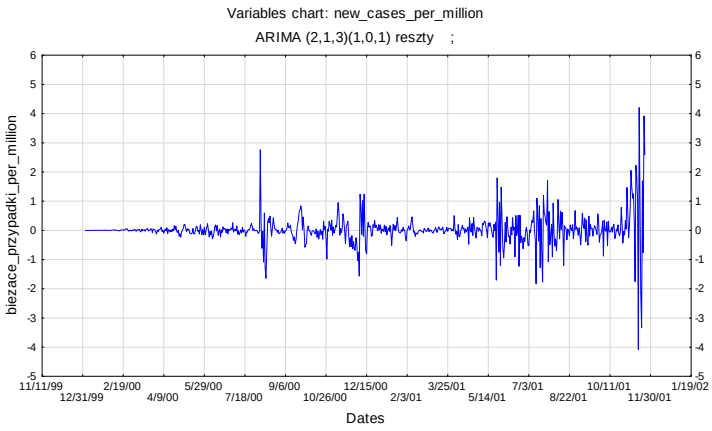


Fig. 27. ARIMA of residuals for Mexico data

Source: Own elaboration.

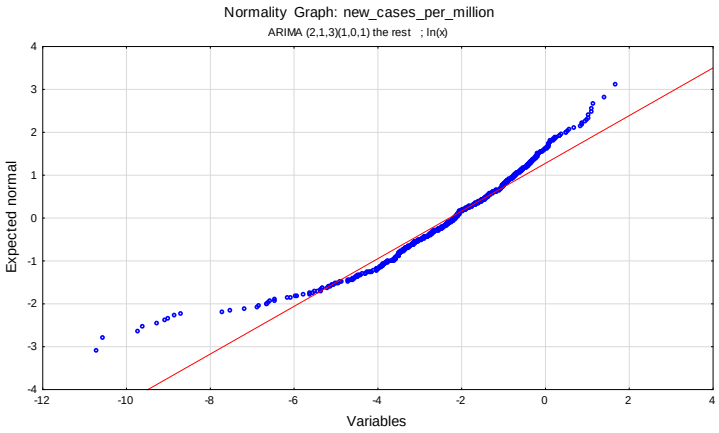


Fig. 28. Residual normality chart for Mexico data

Source: Own elaboration.

Data analyzed for the time series of observations from India. The graphs show the fitted autocorrelation function (Fig. 29) and the partial autocorrelation function (Fig. 30), on the basis of which the analyzes allowing to create the ARIMA model were performed. In order to eliminate the trend, the seasonality was differentiated against the first-order trend and one-time differentiation was applied due to the seasonality of the seventh order.

Table 5. ARIMA India model

	Data: new_cases_per_million Transformations: D(1), D(7) Model: (1,1,1)(0,1,1) Season delay.: 7 Residual MS = 0.00000					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	0.992704	0.004845	204.8756	0.00	0.983190	1.002217
q(1)	0.619534	0.028954	21.3970	0.00	0.562683	0.676384
Qs(1)	0.991891	0.006177	160.5765	0.00	0.979762	1.004019

Source: Own elaboration.

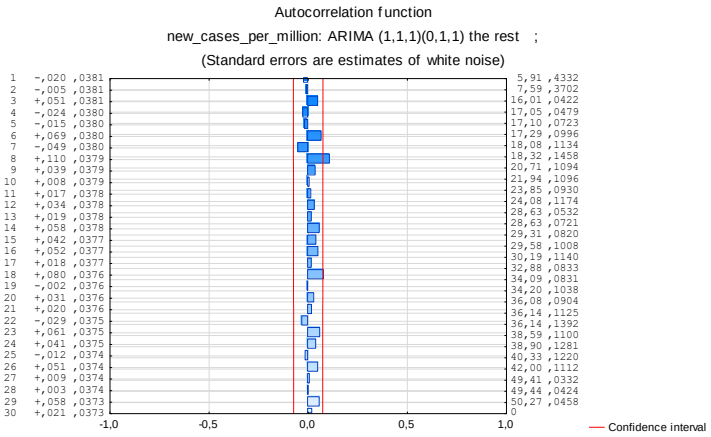


Fig. 29. ACF function for India data

Source: Own elaboration.

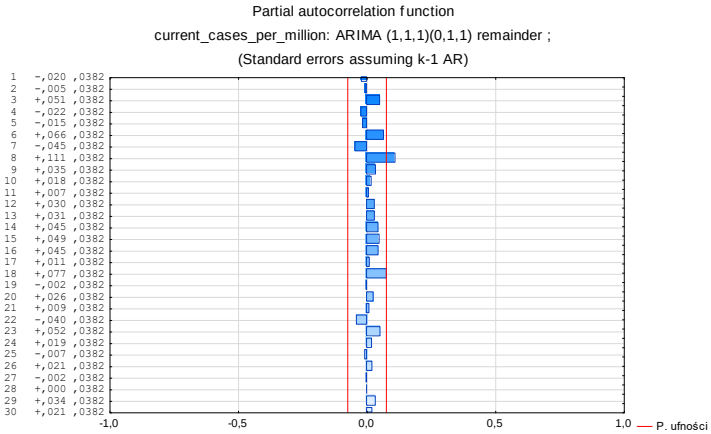


Fig. 30. PACF function for India data

Source: Own elaboration.

The ARIMA model parameters selected for analysis are $p = 1$, $q = 1$, as well as $P = 0$ and $Q = 1$. The results of selecting these parameters are available in the Table 5. The AR model was created when autoregressive delay 1 was selected and the MA model was determined by the parameter q of the moving average equal to 1. The autocorrelation and partial autocorrelation function of the estimated model is within the confidence interval. The resulting model is correctly fitted in terms of ACF and PACF functions [Stellwagen, 2013].

The graph (Fig. 31) shows the mean residuals for the estimated model and the normality graph (Fig. 32), which only after logarithm of these residuals has a normal distribution. Based on the PACF function, the residual normality plot, and the

expected mean value, the distribution is normal. In order to better fit the model, another model should be selected. The resulting model is the ARIMA model with seasonality, i.e. the SARIMA model, the exact models of which is presented in the section 3.2.

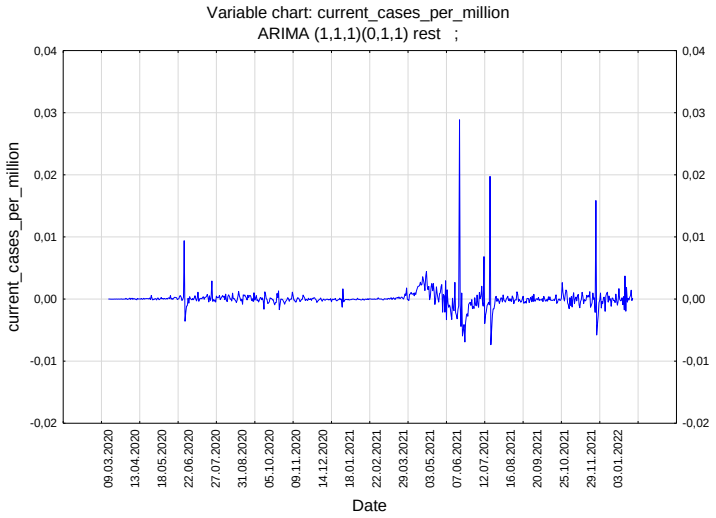


Fig. 31. ARIMA residuals for India data

Source: Own elaboration.

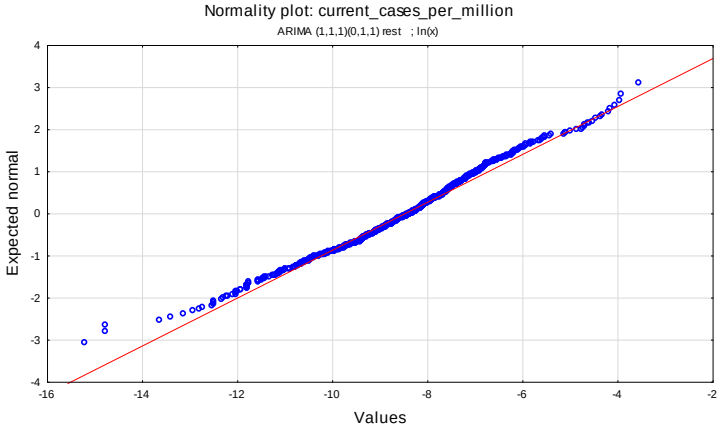


Fig. 32. Residual normality chart for India data

Source: Own elaboration.

Data analyzed for the time series of observations from Israel. The graphs show the fitted autocorrelation function and the partial autocorrelation function, on the basis of which analyzes were made to create the ARIMA model [Płonka, 2014]. In order to eliminate the trend, the seasonality was differentiated against the first-order trend and one-time differentiation was applied due to the seasonality of the seventh order.

Table 6. ARIMA Israel model

	Data: new_cases_per_million Transformations: D(1), D(7) Model: (2,1,2)(0,1,0) 1) Residual MS = 0.39987					
	Parameter	Asympt. Std. error	Asympt. *t*(673)	*p*	Lower limit 95% confidence level	Upper limit 95% confidence level
p(1)	0.841730	0.043578	19.3155	0.000000	0.756167	0.927294
p(2)	−0.321112	0.043606	−7.3639	0.000000	−0.406731	−0.235494
q(1)	1.795076	0.030067	59.7018	0.000000	1.736040	1.854112
q(2)	−0.859717	0.028114	−30.5796	0.000000	−0.914918	−0.804517

Source: Own elaboration.

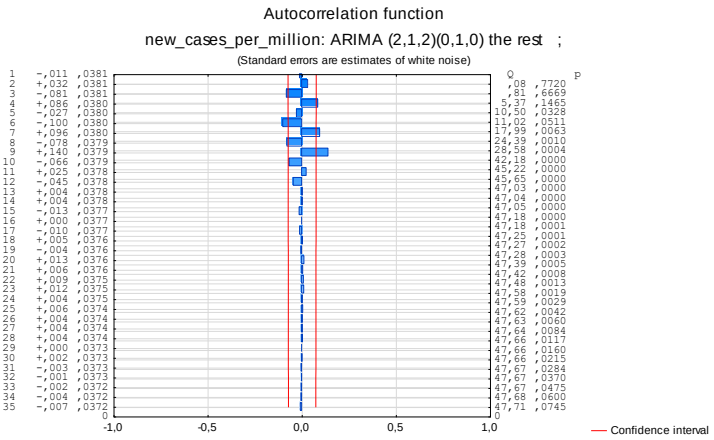


Fig. 33. ACF function for Israel data

Source: Own elaboration.

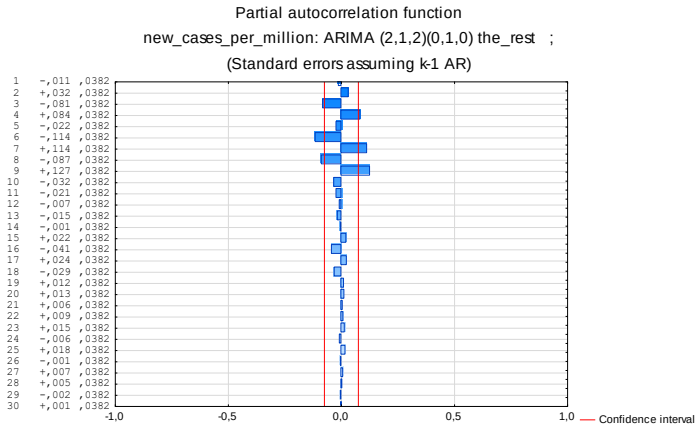


Fig. 34. PACF function for Israel data

Source: Own elaboration.

The ARIMA model parameters selected for analysis are $p = 2$, $q = 2$, as well as $P = 0$ and $Q = 1$, resulting in the Table 6 shown above. The AR model was created when autoregressive delay 2 was selected, and the MA model was determined by the parameter q from the moving average equal to 2. The autocorrelation function (Fig. 33) and the partial autocorrelation (Fig. 34) of the estimated model are within the confidence interval. The resulting model is correctly adjusted in terms of ACF and PACF functions [Stefanowski, 2009].

The diagram (Fig. 35) shows the mean residuals for the estimated model and the normality plot (Fig. 36), which only after logarithm of these residuals has a normal distribution. Based on the PACF function, the residual normality plot, and the expected mean value, the distribution is normal. SARIMA model is described in section 3.2.

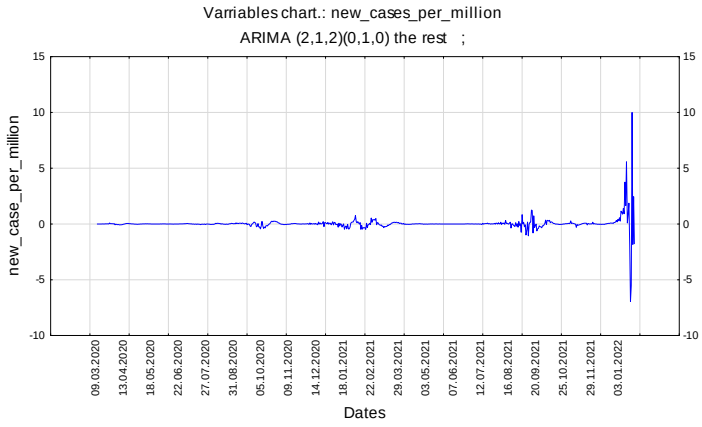


Fig. 35. ARIMA of residuals for Israel data

Source: Own elaboration.

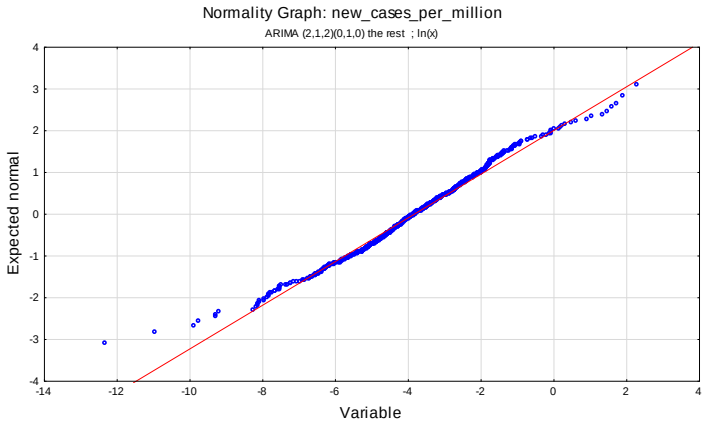


Fig. 36. Residual normality chart for Israel data

Source: Own elaboration.

3.2. SARIMA model

The models presented for the analyzed countries are ARIMA models with seasonality, i.e. SARIMA *(p, d, q) (P, D, Q) m* models with the general formula $\emptyset(B)\Phi(B^s)\Delta_s^d y_t = \theta(B)\Theta(B^s)\varepsilon_t$. Selected model parameters are described in section 3.1 and presented in tables (Tables 1 to 6).

There are three trend elements (the autoregressive part) that require configuration [Nielsen, 2020]. They are the same as in the ARIMA model; specifically:

p: Sequence of trend autoregression.

d: Sequence of trend differences.

q: The trend of the moving average of the order.

There are four seasonal items (part of the moving average) that are not part of ARIMA that need to be configured. They are:

P: Seasonal autoregressive order.

D: Sequence of seasonal differences.

Q: Sequence of the seasonal moving average.

m: Number of time steps for a single seasonal period [Nielsen, 2020].

Poland: Model **SARIMA (4,1,7)(0,1,1)7**

$$(1 + 0{,}76B + 0{,}70B + 0{,}53B + 0{,}35B)(1 - B)(1 - B^7)y_t = \\ = (1 + 0{,}67B + 0{,}39B + 0{,}29B - 0{,}18B - 0{,}40B^2)\varepsilon_t$$

Italy: Model **SARIMA (3,1,6)(1,1,1)7**

$$(1 + 0{,}76B + 0{,}48B + 0{,}72B)(1 - 0{,}32B^2)(1 - B)(1 - B^7)y_t = \\ = (1 - 0{,}15B + 0{,}34B - 0{,}42B + 0{,}26B + 0{,}30B^2)(1 - 0{,}68B^7)\varepsilon_t$$

Chile: Model **SARIMA (3,1,4)(0,1,0)7**

$$(1 + 0{,}76B - 0{,}73B - O, 83B)(1 - B)(1 - B^7)y_t = \\ = (1 - 1{,}37B - 0{,}30B + 0{,}81B^2)\varepsilon_t$$

Mexico: Model **SARIMA (2,1,3)(1,0,1)7**

$$(1 + 1{,}06B + 0{,}41B)(1 - 0{,}99B^2)(1 - B)(1 - B^7)y_t = \\ = (1 + 0{,}99B - 0{,}45B^2)(1 - 0{,}19B^7)\varepsilon_t$$

India: Model **SARIMA (1,1,1)(0,1,1)7**

$$(1 - 0{,}99B)((1 - B)(1 - B^7)y_t = (1 - 0{,}62B^2)(1 - 0{,}99B^7)\varepsilon_t$$

Israel: Model **SARIMA (2,1,2)(0,1,0)7**

$$(1 - 0{,}84B + 0{,}32B)(1 - B)(1 - B^7)y_t = (1 - 1{,}79B + 0{,}86B^2)\varepsilon_t$$

4. Conclusion

The chapter was written to describe the morbidity phenomenon in relation to the administered vaccinations and to analyze the observations describing the COVID-19 virus pandemics in the period from March 1, 2020 to January 22, 2022 and to collect information on the relationships between the introduction into use at the turn of 2020 and 2021 years of vaccines, and the mortality rate from this virus and the incidence of new cases.

As part of the research, it was possible to conclude that there are strong associations between the variables that determine the number of cases and the overall course of the pandemic. The introduction of vaccinations was related to a decrease in the number of cases, mortality, etc. for the analyzed countries. The time series analysis used in the study allowed for the observation of the studied phenomenon and opened the possibility for further work with the created models and using them to forecast the morbidity phenomenon. The large variety of the studied database allows for a detailed analysis of observations and opens up the use of other research methods. Therefore, it is also possible to forecast the observation regarding the relationship between vaccinated persons and the incidence in subsequent studies. The research provides a strong basis for further analyzes of key elements influencing the number of diseases in society.

References

Alsan M. (2020), *Economic Insecurity and the Spread of COVID-19: Evidence from the United States*, "Journal of Public Economics", NBER Working Paper Series, No. 28958, pp. 1-27.

Fanelli D., Piazza F. (2020), *Time Series Analysis and Forecast of COVID-19 Spreading in China, Italy, and France*, "Chaos, Solitons & Fractals", Vol. 134, May, 109761.

Liu R. (2021), *Time Series Analysis of COVID-19 Incidence and Mortality in Ontario*, Canada.

Luszniewicz A., Słaby T. (2001), *Statystyka z pakietem komputerowym Statistica. Teoria i zastosowania*, Wydawnictwo C.H. Beck, Warszawa.

Mohammadi S.R. (2021), *Time Series Analysis of COVID-19 in the United States: The Effects of Seasonality and Mobility Restrictions*, IEEE.

Nazarko J., Chodakowska E. (2022), *Prognozowanie w zarządzaniu*, Oficyna Wydawnicza Politechniki Białostockiej, Białystok.

Nielsen A. (2020), *Szeregi czasowe. Praktyczna analiza i predykcja z wykorzystaniem statystyki i uczenia maszynowego*, Helion, Gliwice.

Nowak E. (2007), *Zarys metod ekonometrii*, Wydawnictwo Naukowe PWN, Warszawa.

Płonka M. (2014), *Co trzeba wiedzieć korzystając z modelu ARIMA?* Predictive Solution, Kraków.

Sokołowski A. (2003), *Prognozowanie finansowych szeregów czasowych*, StatSoft Polska.

Stefanowski J. (2009), *Analiza szeregów czasowych*, Politechnika Poznańska, Poznań.

Stellwagen E. (2013), *ARIMA: The Models of Box and Jenkins*, "The International Journal of Applied Forecasting", Iss. 30, pp. 28-33.

Trzpiot G. (2017), *Statystyka a Data Science*, Wydawnictwo Uniwersytetu Ekonomicznego, Katowice.

About the Authors

Prof. dr hab. Józef Stawicki − Uniwersytet Mikołaja Kopernika w Toruniu, Wydział Nauk Ekonomicznych i Zarządzania, Katedra Ekonometrii i Statystyki
stawicki@uni.torun.pl

Prof. dr hab. Grażyna Trzpiot − Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Katedra Demografii i Statystyki Ekonomicznej
grazyna.trzpiot@ue.katowice.pl

Dr hab. Dominik Krężołek, prof. UE − Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Katedra Demografii i Statystyki Ekonomicznej
dominik.krężołek@ue.katowice.pl

Dr hab. Alicja Ganczarek-Gamrot, prof. UE − Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Katedra Demografii i Statystyki Ekonomicznej
alicja.ganczarek-gamrot@ue.katowice.pl

Dr Justyna Majewska − Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Katedra Demografii i Statystyki Ekonomicznej
justyna.majewska@ue.katowice.pl

Dr Agnieszka Orwat-Acedańska − Uniwersytet Ekonomiczny w Katowicach, Wydział Informatyki i Komunikacji, Katedra Demografii i Statystyki Ekonomicznej
agnieszka.orwat@ue.katowice.pl

Magdalena Kawecka − Uniwersytet Ekonomiczny w Katowicach, Szkoła Doktorska
magdalena.kawecka@edu.uekat.pl

Zuzanna Krysiak − Uniwersytet Ekonomiczny w Katowicach, Szkoła Doktorska
zuzanna.krysiak@edu.uekat.pl

This scientific monograph presented for readers concerns risk analysis and multivariate data modeling. It contains a wide range of problems that have been addressed, including the understanding of risk in economic theories, the measurement of capital market risk, or the study of the energy market. In addition, demographic issues related to mortality, its analysis and forecasting are addressed, as well as issues related to youth unemployment and analysis of the COVID-19 pandemic.

That monograph which is being prepared is the outcome of the research work of the staff and doctoral students of the Department of Demography and Economic Statistics in recent years. Last year, a nationwide conference SIDVRA 2022 took place, which additionally celebrated the 10th anniversary of the establishment of our Department and was at the same time a presentation of preliminary research results.

ISBN 978-83-7875-868-6

University of Economics in Katowice