MASTERARBEIT | MASTER'S THESIS

Titel | Title

Wisdom of crowds and forecasting time series subject to structural breaks

verfasst von | submitted by

Georg Weichselbaumer BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 915
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Betriebswirtschaft
Betreut von | Supervisor:	Assoz. Prof. Steffen Keck PhD

Abstract (deutsch)

Die Vorhersage von Zeitreihen, die plötzlichen Veränderungen unterliegen, auch als strukturelle Brüche bezeichnet, stellt für bestehende Modelle eine Herausforderung dar. Insbesondere statistische Ansätze haben oft Schwierigkeiten, sich diesen abrupten Veränderungen anzupassen. Daher haben Vorhersagen, die auf menschlicher Intuition und Expertise beruhen, als ergänzende Methode zu statistischen Modellen zunehmend an Aufmerksamkeit gewonnen. Zusätzlich dienen sie weiterhin als beliebter Benchmark für neue Prognosemethoden.

Trotz zahlreicher Studien zu der Vorhersage von Zeitreihen, haben sich nur wenige auf die kollektive Leistung von Teilnehmern für Zeitreihen im Allgemeinen und insbesondere für solche mit plötzlichen Veränderungen konzentriert.

Diese Arbeit analysiert daher die Anwendbarkeit des wisdom of crowd concepts („Weisheit der Masse“, eine Methode zur Aggregation mehrerer individueller Schätzungen) bezüglich der Vorhersage von Zeitreihen, die strukturellen Brüchen ausgesetzt sind.

Die kollektive Leistung wurde dabei anhand des Medians, des Mittleren Absoluten Fehlers (MAE) und des Prozentsatzes der Teilnehmer gemessen, deren MAE höher war als der der Gruppe.

Die Ergebnisse der vorliegenden Arbeit stellen dar, dass die Vorhersagen der Gruppe jene der individuellen Teilnehmer für die 12 Zeitreihen in Genauigkeit übertrafen, insbesondere jene mit zwei strukturellen Brüchen. Die mittleren Fehler der Gruppe waren jedoch bei aufwärts gerichteten Trends und aufwärts gerichteten strukturellen Brüchen höher im Vergleich zu abwärts gerichteten Verläufen. Die einzige Situation, in der individuelle Vorhersagen genauer waren als die Gruppenvorhersage, war bei abwärts gerichteten Trends zu Beginn der jeweiligen Zeitreihen. Trotz der besseren Vorhersagen der Gruppe im Vergleich zu den individuellen Vorhersagen hatte die Gruppe Schwierigkeiten, die korrekte Richtung und das Ausmaß der Strukturbrüche vorherzusagen.

Kleinere Gruppen mit zufällig ausgewählten Teilnehmern (3,5,9 oder 15 Teilnehmer) erreichten zum Teil eine gleichwertige oder sogar höhere Genauigkeit im Vergleich zur gesamten Gruppe von 71 Teilnehmern, vor allem bei der Vorhersage von strukturellen Brüchen. Die Ergebnisse leisten

insbesondere einen Beitrag zur bestehenden Literatur über crowd wisdom und Prognosemethoden, indem sie die Wirksamkeit der Aggregation von individuellen Schätzungen bei der Vorhersage von Zeitreihen mit strukturellen Brüchen demonstrieren.

Abstract

Forecasting time series subject to sudden shifts, also known as structural breaks, is a challenging task for existing models, especially for statistical approaches, which often struggle to adapt to these abrupt shifts. Consequently, judgmental forecasting, which relies on human intuition and expertise has gained increasing attention as a complementary method to statistical models and still poses as a popular benchmark for new forecasting methods. Despite numerous literature studying the performance of individuals on forecasting time series, few focused on the collective performance of individuals for time series in general and in particular for those which undergo sudden shifts.

Therefore, this thesis analyzes the applicability of the wisdom of crowd concept (a method of aggregating multiple individual forecasts) in terms of predicting the direction of time series subject to structural breaks and compares it to individual forecasts. The crowd performance was evaluated by using the median, the mean absolute error (MAE) and determining the percentage of individuals displaying a higher MAE than the crowd.

Results revealed that the crowd forecasts outperformed individual predictions for the 12 time series, particularly for those containing two structural breaks. However, the crowd's performance was weaker for upward trends and upward structural breaks compared to downward patterns. The only scenario in which individual forecasts were more accurate than the crowd predictions was for downward initial trends. Despite the outperformance of the induvial forecasts**,** the crowd in general struggled to predict the right direction as well as the magnitude of structural breaks.

Smaller crowds with randomly selected participants (3, 5, 9 or 15 participants) occasionally achieved equivalent or greater accuracy compared to the whole crowd of 71 participants, especially on predicting structural break points. The findings contribute to the literature on crowd wisdom and forecasting by demonstrating the effectiveness of crowd aggregation for time series forecasting subject to structural breaks in comparison to individual judgmental forecasts.

Content

1. Introduction

Forecasting in general is a vital component in the process of planning and decision making and plays a key role for strategic and business planning (Petropoulos et al., 2022). It helps companies to "i*dentify new market opportunities, anticipate future demands, schedule production more effectively, and reduce inventorie*s" (Sanders & Manrodt, 2003; p.511).

Given this general importance, improving forecasting accuracy has developed as an essential research topic with a dominant approach of combining individual forecasts (Wang et al., 2023; Makridakis et al., 2022; Petropoulos et al., 2022). The aggregation of individual forecasts or estimates is often referred to as the "wisdom of crowds" (Petropoulos et al., 2022). This concept gained popularity especially by the work of Surowiecki (2004) and is based on the idea that the aggregated estimates are often more accurate than the majority of individual estimates. This concept has been heavily investigated in different studies, showing its effectiveness in various forecasting tasks such as predicting soccer results, geopolitical forecasts or probability estimates (Fiechter & Kornell, 2021; Peeters, 2018; Yi et al. 2012).

Despite its effectiveness in these areas, the application of the wisdom of crowds concept for forecasting time series in general and with sudden shifts in underlying patterns, so called structural breaks remains largely unexplored. Existing literature rather focuses on analyzing participants' individual forecasting performance.

However, as the probability of time series being subject to such structural breaks rises with time, it has become increasingly important to account for such breaks (Ditzen et al. 2021).

These abrupt changes can be caused by economic crises, policy shifts, unexpected events like natural disasters or terrorist attacks and pose a challenge to traditional forecasting models. (Shrestha & Bhatta 2018; Cró & Martins, 2017; Salisu & Fasanya, 2013).

Ignoring such shifts can negatively affect forecasting accuracy and existing studies therefore have highlighted the necessity of developing models which account for such changes (Pesaran & Timmermann, 2004; Kalsie & Arora, 2019).

In general, models for structural break detection (like the Chow, CUSUM or the Bai and Perron test) have been successfully developed but rather focus on ex-post detection than the prediction of future values (Maheu & Gordon, 2008).

Therefore, the aim of this thesis is to investigate whether the concept of crowd wisdom is a viable method to forecast structural breaks in time series compared to individual predictions. It extends the literature on both crowd wisdom and forecasting, in particular structural break forecasting, by investigating how the aggregated estimates are influenced by factors such as trend direction and the number of structural breaks.

2. Wisdom of Crowds

The concept of the wisdom of crowds has its origin in an analysis from Galton in 1907 (Davis-Stober et al., 2014). Galton examined 787 individual judgments made from participants of a weight guessing contest. The contest was held at a livestock fair and the individuals had to guess the weight of a slaughtered and dressed ox. The best estimate was rewarded with a prize, which enabled that the judgments were made individually without communicating with other participants (Davis-Stober et al., 2014). Galton's results showed that the individual estimates were widely spread, however the average of all guesses was extremely close the true weight of the ox (Davis-Stober et al., 2014). The mean of the individual judgments reached 1,197 pounds, whereas the true value lay at 1,198 pounds (Surowiecki, 2004). The median of the guesses was 1,207 pounds, which corresponds to a deviation of just 0,8 percent. These observations lead to the conclusion of Galton that if the judgments of the individuals are unbiased and independent, the averaged judgment is more accurate than the individual judgments made. (Davis-Stober et al., 2014)

Davis-Stober et al. (2014) define a crowd as wise if "a linear aggregate of its members's judgments of a criterion value has less expected squared error than the judgments of an individual sampled randomly, but not necessarily uniformly" Davis-Stober et al., 2014, p. 80).

It is important to note that Davis-Stober et al. (2014) do not limit their definition of crowd wisdom to a specific method of aggregation, like simple averaging or weighted averaging. In this regard their

definitions deviates from previous ones, which concentrate on a comparison between the performance (described as the accuracy of estimates, meaning how close the estimates is to the true value) of the individuals versus the crowd (Davis-Stober et al., 2014). Lorenz et al. (2011) for example define a crowd as wise "*if the aggregate measure of a collection is close to the true value, even though the individual estimates are largely dispersed."* (Lorenz et al., 2011, .p 9020)

Furthermore, Davis-Stober et al. (2014) highlight that their definition allows to select the individual not only according to a uniform distribution but also according to distribution based on values like individual performance or skill of the participants.

Existing literature explains the success of crowd wisdom through eliminating biased judgments due to aggregating individual estimates and overcoming possible lacks of information of individuals (Simmons et al., 2010); (Lorenz et al., 2011). Simmons et al. (2010) for example describe that in general the individual judgments contain "signal plus noise" and that aggregation will be able to obtain the signal and eliminate that noise. Similar to this explanation, Davis-Stober et al. (2014) refers to Hogarth (1978) and Winkler (1983) who explain that aggregating individual judgments cancel out the individual errors made by the participants. Davis-Stober et al. (2014) highlight that in order for the crowd to be as wise as possible, it is important that the individual members have to be negatively correlated and not necessarily independent. For the judges to be negatively correlated a judge added to an existing group has to differ as much as possible from the other group members (Davis-Stober et al., 2014).

In terms of quantifying, crowd wisdom can be measured as the average squared error of the predictions for example (Vul & Pashler, 2008); (Davis-Stober et al., 2014).

In general the mean squared error is the average of the squared difference between an observation and a true value (Sohil et al., 2022; Hodson et al., 2021).

Generally, the MSE is smaller if the predictions of the individuals are closer to the true value and increases if the difference between predictions and true value is raising (Sohil et al., 2022). Davis-Stober et al. (2014) for example conclude that even a simple average of individuals is robust. A slightly different approach to the mean squared error is the mean absolute error (MAE) which on the other hand uses the average absolute deviations instead of squared errors. However, both of the MSE and the MAE are so

called error metrics. A major advantage of the MAE over the MSE is that it has the same dimension as the underlaying data which is analyzed. So for example if the value to be predicted is in dollar, the mean absolute error will also be denoted in dollar (Botchkarev, 2018).

Another major difference between these two measures is that the MSE assigns more weight to high errors. Due to the fact that errors are squared, few extreme errors can influence the MSE in a greater magnitude than several minor errors, meaning that the MSE is more sensitive regarding outliers inside the data (Botchkarev, 2018).

Another measure which was already used by Galton (1907) is to determine the accuracy of the crowd with the median of the predictions. Both the median and the mean are so called "measures of central tendency". In general, central tendency can be described as a specific value which is able to describe a set consisting of numerous data points (Kaur et al., 2018)

The median displays the middle value in a set of data which has to be arranged in an ascending order. Compared to the simple arithmetic mean, the median is less affected by extreme values and a skewness in the distribution of the data (Kaur et al., 2018).

For a data set containing an odd number of values the median is displayed by the center value of the ranked data points. For an even number of data points the median is calculated by averaging the two middles values of the data set (Kaur et al., 2018).

Lorenz et al. (2011) for example introduced a method on quantifying crowd wisdom which relies on the median of the estimates which they called the "wisdom of crowds indicator". This indicator considers a crowd to be wise if "if the truth lies between the most central values of all estimates" and equals zero if that all the true value lies outside of the range of all predictions made by the individuals. (Lorenz et al. (2011). Furthermore, it increases with a decreasing number of values necessary to bracket the truth.

In terms of determining the performance of the crowd compared to the individuals, the aggregated error term (for example the MSE or the MAE) can be compared between them. A possible procedure is to calculate the percentage of individuals with a higher aggregated error term than the crowd, meaning that these individuals perform worse than the collective (Davis-Stober et al., 2014;Simmons et al., 2010). Davis-Stober et al. (2014) for example used the MSE for this analysis.

It is also important to note that according to Davis-Stober et al. (2014) the concept of crowd wisdom is not subject to the method of aggregation. It is robust in terms of how the individual judgments are combined and even a simple average often is more accurate than the individual.

As quantifying crowd wisdom is essentially about aggregating and combining individuals forecasts, additional methods of combining will be presented in Chapter 3.3 in more detail.

2.1 Factors influencing crowd wisdom

In the previous experiment conducted by Galton in 1907 the key characteristics of the crowd in this experiment have been diversity and independence. According to Surowiecki (2004), the crowd consisted of skilled individuals, meaning people expected to have some sort of expertise in this field like butchers and farmers. Moreover, the crowd also contained people who had not been expected to have profound knowledge in this domain. The fact that people competed for a prize made sure to diminish their motivation to communicate with others, ultimately leading to independent and uninfluenced estimates. Afflerbach et al. (2021) highlight that in 1785 Marquis de Condorcet already concluded that for the probability of a group to reach a correct answer, competent and diverse people are a necessary requirement. Also Simmons et al. (2010) highlight independence and diversity as necessary conditions in order for a crowd to be wise. In general, if individuals possess knowledge of the other estimates inside the group and do not make their judgments independent, the wisdom of crowd effect decreases (Lorenz et al., 2011). This criterion is also emphasized by Surowiecki (2004), who explains that individuals should be able "think and act as independent as possible" (Surowiecki 2004, p. 8). Additionally, if people are allowed to communicate with each other, the likelihood that they possess similar knowledge and therefore are prone to similar errors is increasing, ultimately reducing crowd wisdom as well (Simmons et al., 2010).

Regarding the skill of the crowd, Simmons et al. (2010) elaborate that some members of the crowd should also have expertise in the relevant field. A crowd only consisting of unskilled members would lead to systematic error, meaning that individual errors would not cancel each other out and hinder the wisdom of crowd effect.

In general, in the context of crowd wisdom the level to which individual group members differ in their skills or knowledge is referred to diversity (Simmons et al., 2010; Davis-Stober et al., 2014). Not only communication between the individuals would increase the likelihood that they are subject to the same error but also if the crowd is too homogeneous in terms of skill and knowledge meaning that it is not diverse (Simmons et al., 2010). Despite the importance of the crowd to consist of members with different expertise, Davis-Stober et al. (2014) also emphaisze the importance of the trade of between skill and diversity. Whenever the necessity of having skilled members in the crowd is fullfilled, the importance of including people with different information, expertise or perspective is rising – according to Davis-Stober et al. (2014) this form of diversity is vital for a crowd to be wise.

In addition, they examined if the crowd can still be wise in the presence of judgmental biases and when the participants are able to influence each other's judgments or have knowledge of other participants estimates, violating the conditions of independence and diversity. Davis-Stober et al. (2014) conclude that in general crowd wisdom even is robust to bias on the individual level.

Existing literature also investigates the effect of crowd/group size on its performance. In general, Walter et al. (2022) conclude that crowd performance increases with its size however, this effect continuously decreases for larger sizes. Galesic et al. (2018) demonstrate that under the right circumstances smaller crowds can even be more accurate than larger ones across different tasks. In their study they observe that this effect is due to a possible decline of accuracy for lager groups regarding more complex tasks. Consequently, larger sized groups achieve higher accuracy for simpler tasks. This effect can be observed for settings in which the outcomes are often surprising like abrupt shifts in financial trends, difficult knowledge questions or elections (Galesic et al., 2018). Similar findings were presented by Kao & Couzin (2014) who also reveal that in complex settings the accuracy of smaller groups or moderately sized groups is higher compared to large ones. Kao & Couzin (2014) argue that this is due to correlated information in larger crowds as individuals tend to gain their information from correlated sources like news and from other individuals who have significant influence. However, smaller crowds are not as strongly subject to correlated information and their performance can benefit from noise being present inside the group (Kao & Couzin, 2014). Therefore they argue that the environment of which the

individuals draw their information influences crowd performance with differences regarding its magnitude of the impact for smaller groups and larger sized crowds (Kao & Couzin, 2014).

Hong et al. (2020) for example illustrate that crowd size moderates previously mentioned factors like independence and diversity of the crowd. An increasing number of individuals might light to higher diversity within the group and ultimately leading to a higher performance (Hong & Page, 2004). However, an increasing size could also lead to higher similarity and consequently reducing diversity, which was previously highlighted as an important factor for crowd performance. (Hong et al., 2020; Lamberson & Page 2012). Furthermore Lamberson & Page (2012) found that for smaller group sizes accuracy plays a more important role than diversity among the individuals.

2.2 Existing experiments on crowd wisdom & outcomes

Several papers investigated the wisdom of crowd effect in the field and shown its effectiveness in various different settings like probability estimates, ordering problems and geopolitical forecasts (Fiechter & Kornell, 2021)

Peeters (2018) for example analyze the crowd wisdom effect by showing that forecasting soccer results using crowds' valuation from an online platform (transfermarkt.de) outperform more traditional evaluations like the FIFA ranking. Furthermore, he also shows that if those crowd predictions are used in betting, that substantial monetary gains can be made and that in regard to betting, the crowd again outperforms the FIFA ranking or the ELO ranking (named after it's creator Arpad Elo). Another study by Sjöberg (2009) showed that crowds can accurately predict election outcomes. He also investigated if the forecast accuracy differs with the composition of the crowd. His results show that forecasts made by individual members of the public outperform predictions made by political scientists and journalists. This emphasizes the previously mentioned importance of a crowd to be diverse for the effectiveness of the wisdom of crowd effect.

As studies in this field often concentrate on estimating single numeric values, Yi et al. (2012) try to investigate if the performance of the crowd in more complex situations, in particular in combinatorial problems like the Euclidean traveling salesperson problem, the minimum spanning tree problem and the

spanning tree memory task. They show that even in this complex tasks in which the participants were expected to coordinate various pieces of information that the aggregation of individual answers outperforms the best individual or at least the broad majority of the individuals (Yi et al., 2012).

Mannes et al. (2014) follow a slightly different approach than just taking the simple average of the entirety of the participants. In their study they introduce the term of "the wisdom of select crowds". This term refers to the concept in which judges are ranked, for example by their accuracy regarding previous judgments. As a next step, only these top ranked judges are used for the aggregation (Mannes et al., 2014). Results of their study show that even the aggregation of a small crowd's predictions can yield to accurate and robust judgments. Similar results have been reported by Budescu & Chen (2015) who's method also relied on identifying the experts within the crowd and only aggregating those judgments. Whereas the approach of Mannes et al. (2014) relied on past performance for which the basis was the mean absolute error, Budescu & Chen (2015) developed a measure to quantify individual contribution based on which the select crowd was built. This individual contribution to the crowd was defined as "the change in the crowd's aggregated performance, measured by some merit function, with and without the target individual" (Budescu & Chen, 2015, p. 269).

3. (Judgmental) Forecasting of time series

Forecasting refers to the idea in which past but also present information is used for trying to predict future developments (Petropoulos et al., 2022).

In general, Petropoulos et al., (2022) argue that forecasting is a necessary component in the process of planning and decision making. Furthermore, they show that forecasting has a wide range of applications in a various number of fields like economics and finance (forecasting GDP, inflation, unemployment, stock returns), supply chain management (forecasting the demand of a new product or even or spare parts) and the energy sector (crude oil price forecasting, forecasting the production of photovoltaic systems etc.). Regarding organizations forecasting plays a key role for strategic and business planning. Furthermore it helps companies to "identify new market opportunities, anticipate future demands, schedule production more effectively, and reduce inventories (Sanders & Manrodt, 2003; p.511).

But even quite simple tasks in our daily routine like arriving at the office on time require at least some amount of forecasting (in this case the accurate prediction of the time needed to get to the office) (Petropoulos et al., 2022).

A special type of forecasting is called judgmental forecasting which describes the concept where forecasts (predictions of possible changes to a variable in the future) are made through human judgment (Alvarado-Valencia & Barrero, 2014a). In general judgmental forecasting was described as the starting point of the general forecasting process and as the most important component of product forecasting by O'Connor et al. (1997). Also Goodwin (2002) emphasizes the importance of judgment in terms of forecasting by elaborating that every forecasts contain judgment at some point. Even if statistical methods are used to predict a certain value, judgment is used for selecting the right method at the beginning, determining the variables to be predicted or choosing the right data (Goodwin, 2002).

Also Sroginis et al. (2023) highlight that even though statistical methods more and more improved, judgmental forecasting (especially in business forecasting and in demand planning) plays an important role.

Reimers & Harvey (2024) summarize that judgment can be involved in forecasting at three different levels. The high level refers to selecting the right model for statistical forecasts, the intermediate level is the adjustment of statistical forecasts produced by software, especially incorporating contextual information and the third (lowest) level is judgmental forecasting without any additional support (Reimers & Harvey, 2024).

Regarding product forecasting, Leitner & Leopold-Wildburger (2011) elaborate that especially individuals who are expected to make decision are often required to predict the development of time series related to the pricing of products or sales data. Successfully predicting future sales can for example help to decrease food waste and can also lead to higher sustainability of companies in the food industry for example. Bad forecasting on the other hand can lead to an unnecessary high inventory or a loss in sales (M. Lawrence et al., 2006).

Compared to other methods of forecasting, judgmental forecasting is not only cost saving but also timesaving (Leitner & Leopold-Wildburger, 2011). Another advantage of this forecasting method is that

individuals often possess additional information which is not part of statistical methods but is able to increase the accuracy of forecasts (Leitner & Leopold-Wildburger, 2011). Regarding sales forecasting, a typical example for this additional information is knowledge about the competitors actions in the market (M. Lawrence et al., 2006).

Other examples of information which is hard to quantify and incorporate in other methods than judgmental forecasting are news and rumors (Leitner & Leopold-Wildburger, 2011).

3.1 Factors influencing judgmental forecasts

The judgment of individuals in general and also in terms of forecasting can be influenced by certain aspects, which are presented in this section.

3.1.1 Noise

Alvarado-Valencia & Barrero (2014) explain that noise is usually used in forecasting to account for uncertainty and refer to Fildes et al. (2009) as they describe noise as "*unexplained uncertainty*" and as "*intrinsic and inexplicable portion of volatility after adjusting for historical and contextual data (...)"* (p. 105). In general the presence of noise diminishes the accuracy of judgmental forecasts. (Webby & O'Connor, 1996).

Adam and Ebert (1976) for example show that the forecasting accuracy decreased if the level of noise increased. Also Sanders (1992) draws the conclusion that for time series with a high level of noise, the judgmental adjustment leads to less accurate forecasts compared to statistical methods. In Addition, Andreasen and Kraus (1990) show that noise also makes it more difficult for the participants to detect trends. In terms of trended time series, the study of O'Connor et al. (1993) as well confirms that participants have difficulties in determining the trend of a time series if it contains noise.

Reimers & Harvey (2024) found that if noise in a time series is present, individuals tend to add even more noise to their forecasts which is a sign of a representativeness heuristic.

In addition, noise also plays an important role in the field of judgmental adjustment of statistical forecasts as Fildes et al. (2009) explain that forecasters try to detect certain patterns in noise even though they are just deviations that occur randomly.

Also Goodwin (2002) highlights that people usually try to determine patterns or a system behind the noise of a time series in order to try predicting this randomness, which ultimately reduces forecast accuracy.

3.1.2 Types of data visualization and their influence

Other possible influences on judgmental forecasts of individuals could arise based on the type how the data of the time series is presented – whether the different data points are displayed graphically (line, point or bar chart) or in a table. In general, Reimers & Harvey (2024) explain that in situations with high complexity or in which it is important to show how two variables are related to each other graphs are better than tables containing numeric data.

Lawrence et al. (2006) state that for trended time series forecasts based on graphs are more accurate compared to a tabular representation of the data. Also Harvey & Bolger (1996) draw the conclusion that for (linear) trended series performance is higher if the data is presented graphically to the participants compared to a tabular display. For time series without trends errors were lower with if the data was presented in tables rather than graphs (Harvey & Bolger (1996). Another study conducted by Descantis & Jarvenpaa (1989) investigated if the type of reports presented to the participants influenced their forecasts of financial statements. Results showed that data displayed as graphs slightly improves judgmental forecasts.

Lawrence et al. (1985) did not only compare judgmental forecasting to statistical approaches in general, but they also investigated the influence of the visual representation of the underlaying data. The test subjects in this study were comprised of two groups – students and experts. Results indicate that for shorter time periods the tabular display of the data leads to less accurate forecasts (based on a quarterly time frame of the data) Lawrence et al. (1985). Consequently, for short time periods a graphical display of the data should be focused on whereas for long-term time periods, the data should be presented in a

table if possible (Lawrence et al., 1985). Furthermore, the forecast accuracy of the two subject groups was quite similar and only minor differences were present when using graphs to forecast the time series. However, if the subjects were presented with a table, the experts performed better than the students (Lawrence et al., 1985).

Previous literature on the influence of data display on forecasting accuracy summarized by Theocharis et al. (2019) showed that graphical formats lead to higher accuracy than tables in terms of judgmental forecasting as the effect of trend damping was greater with a tabular display of the data (Theocharis et al., 2019).

Another study conducted by Reimers & Harvey (2024) observed if the type of the graphical display had influence of forecasts made by non-experts. The types of graphical display in their study were bar graphs, line graphs and point graphs which according to them influences "*perception, interpretation and recall of the data*" (p.45). Reimers & Harvey (2024) showed that the forecasts based on bar charts contained more noise than the ones made for data displayed as points (Reimers & Harvey, 2024). Possible reasons for this phenomenon could be that due to the nature of bar charts (distinct separation of each bar), it is more difficult for individuals to detect trends (Reimers & Harvey, 2024). Moreover, the forecasts made when using bar charts were lower overall compared to predictions made with graphs consisting of line or points, regardless if the underlaying trend had an upward or downward direction (Reimers & Harvey, 2024). Also, Tumen & Boduroglu's (2022) study showed that forecasts based on bar graphs resulted into lower performance with higher absolute errors. Furthermore, participants in their study showed a higher magnitude of trend damping (further explained in section 2.3.2) for both upward and downward trends (Tumen & Boduroglu, 2022). Harvey (1996) in his study observed a higher level of trend damping for tables than for graphs.

In the study of Tumen & Boduroglu (2022) using bar charts for judgmental forecasting led to a higher level of mean reversion bias (estimates are biased towards the mean of the data) compared to line and point graphs. However, Harvey & Bolger (1996) show that the use of graphs can also introduce a bias called overforecasting bias in which the participants consistently estimate higher values compared to the underlying true values. Tumen & Boduroglu (2022) summarize that using line graph often introduces

recency bias ("*forecasts were closer to the last data point compared to when the data was presented as point graphs*", p. 3266) and the use of bar graphs introduces the so called within-the-bar bias (estimates are pulled towards the x-axis) or the previously mentioned mean reversion bias.

Nevertheless, existing literature also demonstrates that bar charts can eliminate certain biases. Reimers & Harvey (2024) for example conclude that they might reduce biases like optimism bias or the damping of downward trends.

3.1.3 Information available to forecasters

Alvarado-Valencia & Barrero (2014) divide the information present in forecasting into two groups. Firstly, historical information and secondly, contextual information. Whereas historical information is defined by past values of the underlying data to be forecasted, contextual information can be referred to information which is difficult to incorporate into statistical software like certain events, promotions or reactions of a firms competitors (Alvarado-Valencia & Barrero 2014). Goodwin & Wright (1993) classify the information available to forecasters into three major groups. Next to time series information and contextual information, Goodwin & Wright (1993) also defines the label of the time series (for example if a time series is denoted as expenses or as revenue) as a distinct group of information. Such labels usually generate certain beliefs among forecasters regarding the direction of the time series (Goodwin & Wright, 1993).

Next to contextual information Sanders & Ritzman (1992) present technical knowledge as another important type of information which can influence forecasting performance. According to Sanders & Ritzman (1992), if forecasters are for example skilled in performing data analysis or possess expertise in forecasting processes in general, they have technical knowledge.

In existing literature the terms contextual information and domain knowledge are often used interchangeably, however Webby et al. (2001) explicitly distinguish between those terms. According to them, domain knowledge is derived from contextual information (= attribute of the forecasting environment) and present if the forecaster is able to "*derive the appropriate meaning from the contextual (or environmental) information*" (p. 391).

Another possible definition of domain knowledge is laid out by Lawrence et al. (2006) which define it as "*any information relevant to the forecasting task other than the timer series*" (p. 499).

In addition, Lawrence et al. (2006) highlight that there is a possibility that domain knowledge is not even available for statistical methods. As example for such knowledge, they state a sales promotion which will take place in the future (Lawrence et al., 2006). Further examples laid out by Goodwin & Wright (1993) are knowledge about the market in which a company is operating and financial insights regarding the firm.

In general, contextual information can be qualitative and quantitative (Webby & O'Connor, 1996) or as Alvarado-Valencia & Barrero (2014) refer to as soft and hard information. Information about a sales promotion taking place would be qualitative (soft) contextual information whereas qualitive (hard) information could be data regarding the weather which might influence the revenue of an ice-cream shop (Alvarado-Valencia & Barrero, 2014); Webby et al., 2001).

Lawrence et al. (2006) also explains that if a certain information can be incorporated into a statistical forecasting model this information is not classified as domain knowledge. One example laid out by Lawrence et al. (2006) is a promotion which takes place on a fixed schedule. The authors argue that this type of knowledge could be included as a seasonality in a statistic forecasting approach.

Domain knowledge may also take the form of participants being conscious about the underlaying characteristics behind a given time series (Lawrence et al., 2006). Therefore Lawrence et al. (2006) also refers to domain knowledge if forecasters possess knowledge that allows them to determine the cause why a time series is displaying a certain trend.

Regarding the influence of contextual information on forecasts made by human judgment, already Goodwin & Wright (1993) point out that several studies emphasize the positive effect of contextual information on the accuracy of forecasts.

Also, the results of the study conducted by Sanders & Ritzman (1992) show that contextual information considerably improves judgmental forecasts whereas technical knowledge did not influence the performance of the forecasts at all. In terms of judgmental forecasting in comparison with statistical approaches, Webby & O'Connor (1996) even state that domain knowledge was the "*prime determinant*

of judgmental superiority over statistical models" (p.98). Furthermore, they summarize that contextual information helps the forecasters to better detect and understand discontinuities in a given time series. Webby et al. (2001) conclude that if a great amount of domain knowledge is available, judgmental forecasting should be favored over statistical approaches.

3.2 Heuristics and biases of judgmental forecasting

In this section the thesis discusses the most common heuristics individuals use when they are expected to make forecasts. Alvarado-Valencia & Barrero (2014) point out that in existing research one goal is to determine the underlaying processes of judgmental forecasts and studies in this area show a strong focus on identifying so-called heurists and biases which might occur when forecasters produce judgmental forecasts. Heuristics are rules of thumb and in terms of forecasting are used by individuals in order to predict a certain value "*under restrictions of mental capacity, lack of information and lack of time"* (Alvarado-Valencia & Barrero, 2014, p.109). Furthermore, the use of heuristics helps to simplify a more complex task. However, these rules of thumb often lead to '*systematic and predictable errors'* (Tversky and Kahneman 1974, p.1131) which are referred to as biases and lead to less than ideal decisions (Alvarado-Valencia & Barrero 2014).

Generally, according to Harvey (2007) the heuristic used in judgmental forecasting tasks usually depends on the information used to produce the forecasts. Possible informational basis for judgmental forecasts is any relevant information to the forecasting task which the forecaster possesses or information on previous values of the forecasted variable and other variables connected to the forecasting task (Harvey 2007). Even though heuristics depend on the type of information available to the forecaster, there are typical heurists of decision making which can also be found in judgmental forecasting (Harvey, 2007; M. Lawrence & O'connor, 1995).

As explained by Harvey (2007) already Tversky and Kahneman (1974) presented three major heuristics in terms of human judgment and decision making which are: the anchoring and adjustment heuristic, the representativeness heuristic and finally the availability heuristic.

Representativeness in forecasting is present if people use the information, they possess on one variable in order to predict the value to be forecasted. Moreover, if the forecast is derived from information the forecaster has in his/her memory the so-called availability heuristic is present (Harvey, 2007).

Mechanism more closely related to judgmental forecasting are trend damping and noise modelling (Alvarado-Valencia & Barrero 2014) which will be explained in more detail in this section as well as the anchor and adjustment heuristic

Regarding the influence of heurists on forecasting performance, Alvarado-Valencia & Barrero (2014) summarize that heuristics used to form judgmental forecasts do not necessarily reduce forecasts accuracy. Furthermore, Harvey (2007) points out that the results of heuristics in general are precise enough to be deemed satisfactory.

3.2.1 Anchor and adjustment

One of the main heurists people use in judgmental forecasting and is among the most researched is the so-called anchor and adjustment heuristic (Harvey, 2007). As pointed out by M. Lawrence & O'Connor, (1995) it has increased in popularity after the work of Tversky & Kahneman (1974)

Anchoring and adjustment refers to the process in which the forecasters chose a start value at the beginning (this is referred to as the anchor) and modify it in order to forecast future values (Harvey, 2007; Goodwin & Wright, 1994; M. Lawrence & O'Connor, 1992).

As summarized by Goodwin & Wright (1994) there exist different possible values which potentially serve as an anchor. Several previous research studied which value the participants choose ranging from the time series' mean and the previous value (Goodwin & Wright, 1994). Based on the study of Harvey (2007), the anchor is the latest value of the time series presented to the forecasters. Contrary to this finding, M. Lawrence & O'Connor (1992) show that participants in their study used the long-term mean of the time series. Similar findings are presented by Goodwin & Wright (1994). In their study they also show that the anchor is equivalent to the mean of the time series. Furthermore, the adjustments made are a fraction of the means and the latest values difference. For trended time series Harvey et al. (1994) point out that the anchor is considered the previous value. People then use a fraction of the

difference between this value and the current forecast as an adjustment (Harvey et al., 1994; Goodwin & Wright, 1994).

Harvey (2007) state that in general the problem which arises with anchoring and adjustment is that the adjustment to the anchor is not sufficient enough in order to lead to optimal forecasts. For upward trends adjusted initial values tend to be lower than the true value whereas for downward trend they are usually higher than the optimal value of the series (Harvey, 2007). Epley & Gilovich (2006) for example in their study investigated possible reasons why the adjustments of the forecasters are insufficient. Results indicated that the adjustment process only lasted until the initial value laid in an interval of reasonable values (Epley & Gilovich, 2006). Furthermore, they also discovered that if the participants were provided with incentives for more accurate estimates, the bias resulting from the anchor and adjustment heuristic could be reduced (Epley & Gilovich, 2006).

Existing literature also dealt with the effects of advice on the anchoring and adjustment heuristic. Harvey (2007) summarizes that if forecasters are provided with advice they still rely too heavily on their initial anchor.

3.2.2 Trend damping

Trend damping refers to the process in which subjects usually reduce the magnitude of an upward trend downwards and vice versa – meaning that if the trend has a downward direction the trend is damped upwards (Alvarado-Valencia & Barrero, 2014).

Regarding the level of trend damping, Alvarado-Valencia & Barrero (2014) summarize that usually trend damping is higher for downward trend than for upward trends.

Harvey (2007) consider trend damping more as a specific type of the previous explained anchor and adjustment heuristic. He lays out that in this case the anchor is the latest data point of the underlaying trend of the series, and it is not adjusted sufficiently in order to correctly forecast this trend (Harvey, 2007).

Despite the potential negative influence on forecasting accuracy Armstrong (2006) point out that a forecasting model which intentionally dampens existing trends in a time series can be useful for trend

extrapolation, if the time series contains a high level of uncertainty. This means that the focus on predicting an underlying trend should be reduced with a growing level of uncertainty (Armstrong 2006). Also, Gardner & Mckenzie (1985) point out that in time series extrapolation damping trends can lead to a higher forecasting accuracy under the right circumstances. In his study Gardner & Mckenzie (1985) developed a model for trend damping and used data from over 1001 time series of the M1-Competition by Makridakis et al. (1982). Gardner & Mckenzie (1985) used a specific parameter which gradually dampens the trend over time. Results show that the proposed model by Garnder & Mckenzie (1985) could increase forecasting performance for longer time periods with linear but noisy trends.

3.2.3 Noise modelling

Noise modelling refers to the process in which the forecasters try to not only incorporate a potential underlaying signal (trend) in a time series but in addition include noise in their forecasts in an attempt to make them more reasonable (Alvarado-Valencia & Barrero, 2014). Also, Harvey (2007) highlight that in general the forecasts of the people tend to add unnecessary noise even though the estimates should only incorporate the underlaying trend of the time series. Similarly, the study of Nigel (1995) in which the participants had to make a total of six forecasts to 58 time series showed that people try to forecast noise.

Furthermore, results of existing studies in this field summarized by Harvey (2007) indicate that with an increasing level of random noise, also the magnitude of the noise modelling bias rises, meaning that people include even more noise in their estimates.

Ultimately noise modelling reduces forecasting performance and leads to less accurate estimates (Alvarado-Valencia & Barrero, 2014).

3.3 Methods to increase forecast accuracy

As laid out in the previous chapter forecasts made through human judgment are subject to various biases which in general decrease forecast accuracy. This raises the question of how estimates by individuals can be improved in order to reach more favorable results.

Therefore, this part of the thesis presents different approaches to increase forecasting performance. Combining individual as source of forecasting improvement, the use of feedback and decomposition will be presented in detail.

3.3.1 Combining Forecasts

Generally, combining forecasts can increase forecasts accuracy and existing literature shows that simple methods (e.g. the simple average) often outperform approaches with a higher complexity (Jose & Winkler, 2008). Especially with point forecasts the simple average leads to satisfactory results by utilizing the wisdom of crowd effect: the average forecast of the crowd often outperforms most of the individual forecasts (Jose et al., 2014).

In general, one advantage of the simple mean is that no combination weights must be estimated as they are all equal – this is especially beneficial if the there is no information available on how the individuals differ in their forecasting performance (Palm & Zellner, 1992). Other advantages stated by Palm & Zellner (1992) are that combining forecasts helps to decrease variance and bias, as the combination of individual forecasts "*averages out individual bias*" (p.699). Furthermore its easiness to implement and its robustness has made it to the most popular combination rule (Wang et al., 2023).

Even though complex methods of combining forecasts exist as well as newer methods which use machine learning algorithms, simple arithmetic combinations still remains its competitiveness (Wang et al., 2023; Makridakis et al., 2022). Furthermore it is commonly used as a benchmark for measuring the performance of new combining approaches (Wang et al., 2023).

However, next to the simple mean and the median, previous literature also highlights the use of a trimmed or winsorized mean in order to reduce the impact of extreme values (Jose & Winkler, 2008). In both of these two methods, the mean of the exisitng forecasts are adjsuted. Wheras winsorizing adjusts the individual forecasts in order to dampen their extreme impact, trimming refers to completely removing extreme forecasts (Jose & Winkler, 2008). Armstrong (2001) point out that combining forecasts and trimming should be used if the elimanition of large errors is crucial.

However combining forecasts can not only refer to the combination of individual estimates of one forecasting method but can also mean the combination of estimates resulting from several different forecasting models and the combination with statiscial approaches (Lee et al., 2007). The first broad study in order to investigate the performance of various forecasting methods was conducted by Makridakis et al. (1982) in which a total of 24 models was tested on 1001 time series by seven experts per method. One of the major findings of their study was that the simple average of the estimates performed better than the individual forecasting models (Makridakis et al. 1982)

Especially if time series contain structural breaks combining different forecasting models can lead to an increased accuracy (Wang et al., 2023).

3.3.2 Feedback

Lawrence et al. (2006) divide feedback into three subcategories, namely outcome-, performance- and task properties feedback. If the forecaster receives information about the last data value of the respective time series this is referred to as outcome feedback, whereas performance feedback refers to any information the individual receives regarding their accuracy or even bias of their estimates (Lawrence et al., 2006). Finally, task properties feedback is considered any statistical information the forecasters receive, including properties of the relevant time series or dependencies of the forecasting variable with other variables (Lawrence et al., 2006). By summarizing previous studies on feedback and its influence on forecasting accuracy, Lawrence et al. (2006) point out that among those three different types, the one with the lowest impact on forecasting accuracy was outcome feedback.) A possible reason for this weak impact is the difficulty for the participants to determine the cause of their forecasting error - whether bad forecasts are caused by a person's own insufficient judgment or by random variations such as random noise or unknown variables for example Lawrence et al. (2006).

The type most useful was task properties feedback (Lawerence et al., 2006). One example of a study which examines the influence of task properties provided to the participants is the one of Sanders (1997). In his study the participants received information regarding the amount of noise the time series contains and information about the time series pattern (whether it was trended, stationary or contained

seasonality)(Sanders, 1997). Results show that the performance of the individuals was higher after receiving additional task properties feedback (Sanders, 1997).

3.3.3 Decomposition

Decomposition refers to the concept of "*dividing judgmental tasks into a series of smaller and cognitively less demanding tasks, and then combining the resulting judgements*." (Lawrence et al., 2006, p. 507). One study in this field summarized by Lawrence et al. (2006) was conducted by Edmundson (1990). He found that accuracy of the forecasts could be increased if the components of time series were split into seasonal, random or trend, then forecasted separately and finally combined into one estimate (Lawerence et al., 2006).

However, Lawrence et al. (2006) also point out that existing literature also highlights possible negative effects of decomposition. Decomposition can also reduce accuracy if after dividing the initial complex tasks, the new and smaller subproblems have a higher complexity. Since after dividing the task, the forecasters have to make a higher number of judgments this can reduce the participants' willingness to engage in the forecasting task (Lawrence et al. 2006).

A study which investigated the effects of decomposition on judgment in general was conducted by MacGregor & Armstrong (1994). In their study 280 participants were faced with ten problems, leading to a total of 1078 estimates. MacGregor & Armstrong (1994). show that decomposition only improved the accuracy if the task contained "*extreme and uncertain values*" (p. 1)

4. Structural Breaks

Structural breaks can be defined as abrupt changes in time series at a specific point in time, which can be caused by regime shifts, political changes, global financial crises, or natural disasters (Shrestha & Bhatta 2018; Cró & Martins, 2017; Salisu & Fasanya, 2013).

A more general definition of structural breaks refers to any external shocks which results in a sudden shift in the underlaying parameters of a certain model (Kalis & Aurora 2019; Kim & Seok, 2018). Shrestha & Bhatta (2018) state that next to autocorrelation, stationarity, seasonality and trends, structural

breaks are frequent characteristic of time series. Also, empirical studies indicate that structural breaks occur in financial and macroeconomic time series (Pesaran & Timmermann, 2004).

Shrestha & Bhatta (2018) explain that structural breaks can occur in a trend of a time series or in in its intercept or even in both simultaneously (see figure 1 to 4).

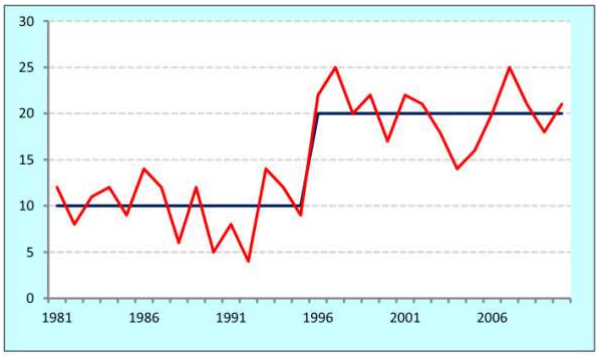

Figure 2: structural break in the intercept

(Shreshta & Bhatta, 2018, p. 76)

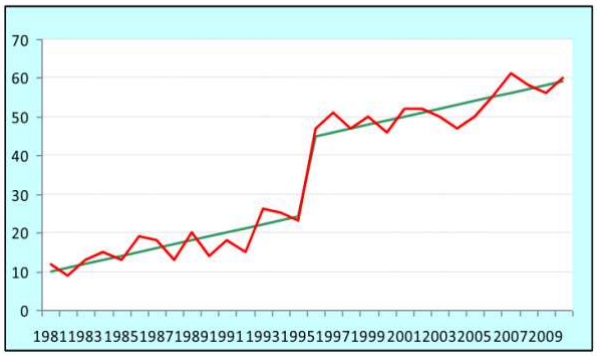

Figure 1: structural break in the intercept

(Shreshta & Bhatta, 2018, p. 76)

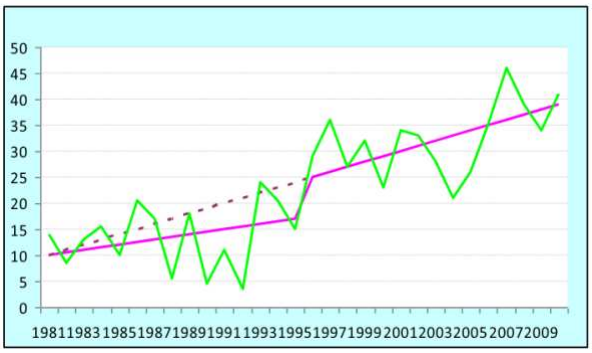

Figure 3: structural break in trend

(Shreshta & Bhatta, 2018, p. 76)

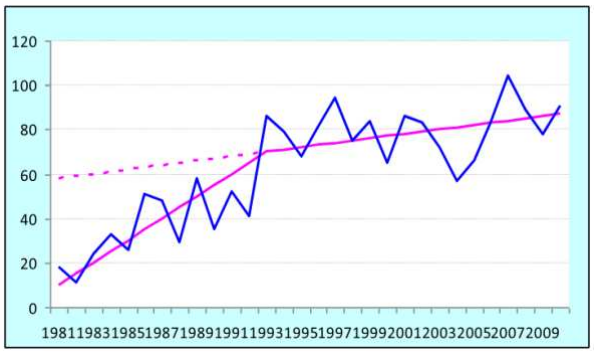

Figure 4: structural break in trend and intercept

(Shreshta & Bhatta, 2018, p. 76)

The upper graphs (Figure 1 and Figure 2) are examples for structural breaks which occur in the intercept of the respective time series whereas the third graph shows a break in trend and graph number four a break in the intercepts as well as in trend (Shrestha & Bhatta 2018).

In general, Ditzen et al. (2021) highlight that the probability of a time series being influenced by a structural break usually increases with time. As example the authors mention the financial crisis taking place in 2007-2008 and the corona pandemic in 2020 (Ditzen et al. 2021). Furthermore, Pesaran et al. (2005) point out that if a time series was subject to a break in the past, there is a high possibility that it will be influenced by another one in the future.

As the existing literature in this field has a strong focus on the ex-post identification on structural breaks in past data (Maheu & Gordon, 2008)., the section will cover some of the most important methods in order to detect structural breaks in existing time series data.

Pesaran & Timmermann (2004) highlight the importance of accounting for structural breaks in forecasting methods, as they might considerably influence their performance. Also Kalsie & Arora (2019) note that the identification of such breaks helps to improve estimates as well as forecasts.

4.1 Methods for detecting structural breaks in time series data

As previously explained the existing methods for structural break detection in time series data focus on the ex-post identification, especially in linear regression models (Maheu & Gordon, 2008).

Meligkotsidou & Vrontos (2008) divide the existing tests for structural breaks into three major categories. First, the so-called fluctuation tests (or '*residual based tests'* in El-Shagi & Giesen, (2013, p. 103), second the F-tests and as a third class, tests which use least squares. In general, the F-tests in the context of structural break detection tests the null hypothesis that the parameters in the forecasting model are constant against the alternative hypothesis that the data contains (at least) one structural break (Meligkotsidou & Vrontos, 2008). The seminal work of Chow (1960) was the first to propose an F-test for break detection and known break dates. The idea of Chow was than generalized in a way that for every possible break date the F-statistic was computed (Meligkotsidou & Vrontos, 2008). Consequently, the null hypothesis of parameter consistency was rejected if these F-statistics over various possible break

dates (examples are supremum, average or exponential functional) reached too high values (Zeileis et al., 2003). The first fluctuation tests used for determining structural breaks were based on cumulative sums (CUSUM, see Chapter 3.1.3)(Meligkotsidou & Vrontos, 2008).

Also El-Shagi & Giesen (2013) divide the existing structural break test into three categories, with the first two being identical to the one proposed by Meligkotsidou & Vrontos (2008). However the third category stated by El-Shagi & Giesen (2013) are testes which use Maximum Likelihood (ML) scores.. For structural break testing these scores help to evaluate parameter instability by detecting fluctuations in the scores over time, which indicates possible structural breaks in the data (Zeileis, 2005).

Generally, the tests for structural breaks have been limited to detect one single break in a time series in the beginning (Chow Test for example). Bai & Perron (1998) for example developed a method which allowed to test for more than one structural break.

4.1.1 The Chow test

According to (Kalsie & Arora, 2019) the chow test is regarded as foundational method of structural break detection in time series data. It was introduced by Gregory Chow in 1960 and in general the Chow tests evaluates whether the coefficients in a linear regression model differ across two sub-samples, or whether they remain constant over the entire dataset (Emmanuel & Maureen, 2022).

Chow (1960) explains that in economics it can be of interest to determine whether '*two sets of observations can be regarded as belonging to the same regression model"* (p. 591). For instance, Chow (1969) suggests that it could be valuable to assess whether World War II significantly influenced American consumption patterns.

Regarding time series analysis it is used to test time series data for a single structural break by using F-statistics (Kalsie & Arora, 2019).

In general, Emmanuel & Maureen (2022) point out that if a single regression line fits the underlaying data, the time series is not subject to a structural break. Contrary, if two regression lines are necessary in order to fit the data this indicates the presence of a structural break in the relevant time series. Emmanuel & Maureen (2022) summarize the steps necessary to conduct a Chow test:

The first step is to perform a regression on the whole data set and to calculate the residual sum of squares for error (RSS_e). After this initial step, two regressions have to be performed – one which only takes into account the data before the potential structural break and the other one which only inludes data after the break. Again, the residual sum of squares for error are computed for each of the separate regressions (denoted by RSS_1 and RSS_2). Step three involves calculating the F-statistic by using the following formula (k denotes the amount of predictors, n the number of observations in the time span). Emmanuel & Maureen (2022) denote the steps in the following formual (p. 16):

$$F = \frac{RSS_e - (RSS_1 + RSS_2)/k}{(RSS_1 + RSS_2)/n-2k} \sim F(k, n-2k)$$

The final step is to determine the critical values by using the F-table, where the null hypothesis states that there is no structural break in the data and the alternative hypothesis that the time series is subject to structural breaks.

In general the Chow test is limited to detect only a single structural break point in the data and the possible break point has to ben known a priori or must be specified where it might occur. Therefore other concepts like the CUSUM or more generalized versions of the Chow tests using Larange Multiplier, Likelihood Ratio and Sup Wald test have been developed (for example by Bai and Perron (1998)) in order to handle more complex scenarious, especially multiple break points and situitions in which the break point is not known beforehand (Kalsie & Arora, 2019; Pauwels et al., 2012).

4.1.2 The CUSUM test

The CUSUM test was first introduced by Brown et al. (1975) and is based on '*the cumulative sum of the recursive residuals*' (El-Shagi & Giesen, 2013, p. 104). In comparison to the previous introduced Chow test, the CUSUM test does not require knowledge about the structural break date beforehand (Turner, 2010).

Generally, the steps necessary to for this method are to first calculate the recursive residuals, secondly those are accumulated over time and finally the test statistic is compared against critical values (El-

Shagi & Giesen, 2013). Similar to the Chow test, the null hypothesis again assumes no structural break in the data whereas the alternative hypothesis suggest that the data is subject to a structural break, ultimately indicating a change in the parameters of the underlaying regression model.

In general, Otto & Breitung (2023) explain that the CUSUM tests is made of a so called "*detector statistic*" and a "*critical boundary function*" (p.2). The detector statistic calculates the standardized sum of previously mentioned recursive residuals, which are the errors between the actual value at time t and the forecasted value for time t at time t-1. For every point in the testing period this detector statistic is calculated and the null hypothesis gets rejected as soon as the boundary function and the detector statistic cross (Otto & Breitung (2023).

Pan Pang & Ting (2004) investigated conditions necessary for increasing the performance of the CUSUM. In their analysis they considered the size of the data set before and after the break as well as the change in variance expressed in percentage. Results show that the break detection achieves a higher accuracy with a rising amount of data before the break and will decrease if the amount of data after the structural break increases (Pan Pang & Ting, 2004). All in all, they highlight the importance of choosing an adequate size of the post break data set for the performance of the test (Pan Pang & Ting, 2004).

One possible limitation of the Chow test is that it has difficulties in detecting so called orthogonal structural breaks (Luger, 2001). Orthogonal structural breaks refer to changes which are uncorrelated to the mean of the regressors in the relevant model. In general, the Chow test detects changes in the main relationship between the regressors and the dependent variable, however if the change is orthogonal, the CUSUM fails to detect such changes (Luger 2011). Luger (2001) developed a modified version of the CUSUM test which is able to detect such changes.

4.1.3 The Bai and Perron structural break test

The Bai and Perron method is a series of tests, which allow to test for multiple structural breaks in linear regression models at an unknown point in time (Meligkotsidou & Vrontos, 2008; Bai & Perron, 2003). It allows to determine the number of breaks, the point of time in the time series data in which it occurs

and the respective confidence intervals for the parameters as well as the ones for the structural breaks (Mérida & Golpe, 2016).

In order to detect the breaks, the test tries to identify points in time series data in which the relationship between the dependent and independent variables changes significantly (Bai & Perron, 2003).

In general, for the Bai and Perron method the sum of all squared residuals are calculated for each segment in the data using ordinary least squares (OLS) regression (Meligkotsidou & Vrontos, 2008; Bai & Perron, 2003). The method detects structural break points in time in which the sum of the squared residuals is minimized (Bai & Perron, 2003). The next step contains a sequential hypothesis test for all segments in the data, in which at first the null hypothesis that the data contains no break is tested against the alternative that it contains at least one breaks. After that the hypothesis that the data contains m breaks is tested against the existence of m+1 break (Mérida & Golpe, 2016; Bai & Perron, 2003). This allows to evaluate whether adding an additional break significantly improves the overall model fit. For this evaluation the Bai and Perron method uses either significance tests like the F-test for example or information criteria like the Bayesian Information Criteria (BIC) or the Akaike Information Criterion (AIC) to determine the model which best fits the data (Bai & Perron, 2003; Kriegeskorte, 2015). The AIC focuses on balancing model complexity and goodness of fit. Therefore, it penalizes models with a higher number of parameters and favors models which minimize information loss (Akaike, 1974). In comparison the BIC is more conservative in regards of selecting more complex models by penalizing additional parameters more, due to the fact that it also incorporates sample size (Ferguson et al., 2019). One of the key advantages of the Bai and Perron method is that it can be used for different types of structural changes models, namely for pure structural change models (in which all parameters change at structural break points) as well as partial structural change models (in which some parameters remain constant (Bai & Perron, 2003). In general, the Bai and Perron method is widely used in different economic sectors like finance or tourism (Cró & Martins, 2017).

4.2 Structural Breaks and their presence in the economy

This section presents possible occurrences of structural breaks in the economy. Therefore, three economic sectors (Finance, Oil/gas prices, Tourism) are presented in which structural breaks can have a major impact on specific underlying variables of these sectors.

Oil and oil prices in particular was chosen because of its high impact on the economy in general as well as its major influence in terms of recessions (Mensi et al., 2014). Tourism on the other hand, is "*one of the most dynamic exporting sectors in the world economy"*and often contributes highly to the domestic economic growth (Charles et al., 2019, p.2).

4.2.1 Oil and gas prices

Mensi et al. (2014) investigated the influence of OPEC news announcement on volatility and expectations in the crude oil market and analyzed the influence of structural breaks on the volatility of crude oil prices. For their analysis they used prices (USD/barrel) of two crude oil benchmarks called Europe Brent (reference for the North Sea) and West Texas Intermediate (WTI, reference for the USA) over a period ranging from 1987 to 2012. Furthermore, they classified the announcement of OPEC into three categories: "cut", "maintain" and "increase" which refers to decisions regarding the amount of oil production.

In terms of structural break detection, they also used the Bai and Perron method, leading to the identification of five breaks for the WTI volatility as well as six for the Brent volatility. Mensi et al. (2014) explain that accounting for OPEC news announcements as well as for structural breaks helps to better understand the volatility in crude oil markets and that structural breaks reduce the persistence of the volatility in crude oil prices. In general, Mensi et al. (2014) also highlight that the structural breaks found fall in line with major events regarding the oil market. For both benchmarks WTI and Brent, the structural breaks can be linked to Russian crisis in 1998, to the Asian financial crisis in 1997 and to announcements regarding the production of oil from OPEC as well as other countries which are not part of this organization like Mexico or Russia (Mensi et al., 2014).

Similar to Mensi et al. (2014), also Salisu & Fasanya (2013) investigated oil price volatility and the influence of structural breaks using data from the WTI and Brent market of the years 2000 to 2012. Salisu & Fasanya (2013) were able to detect two structural breaks in the data. One of them is related to the conflict between Iraq and Kuwait in 1990 and the second break could be linked to global financial crises in 2008. Overall, Salisu & Fasanya (2013) emphasize the importance of modelling oil price volatility due to potential high gains and losses especially for countries with an economy dependent on oil production and exports. Furthermore, also for private investors in the oil market modelling volatility is vital as with increasing volatility, uncertainty rises which ultimately could lead to substantial losses (Salisu & Fasanya, 2013).

Another study related to prices of resources and structural breaks was carried out by Aruga (2016), who investigated if the decrease in price in the U.S market for gas had an influence on the link between the US markets and Japanese as well as the European natural gas market. Aruga (2016) found that the one of the structural breaks detected in 2006 with the use of the Bai and Perron method could be linked to the shale gas revolution. The second break discovered in 2009 coincides with the world financial crisis. Figure 5 illustrates the two break points found in the natural gas prices by Aruga (2016)

Figure 5: structural break in the natural gas price

(Fan & Xu, 2011, p. 1088)

Furthermore, Aruga (2016) found that after the structural break related to the gas revolution, the connection of the U.S gas market with the European and Japanese market no longer existed.

Fan & Xu (2011) examine the evolution of oil prices by analyzing the major influences on oil price like supply-demand, us dollar exchange rate, speculation, stock and gold market as well as geopolitics over different periods. Fan & Xu (2011) divide oil price movements into three periods by using structural break tests. Figure 6 shows the structural breaks found by Fan & Xu (2011):

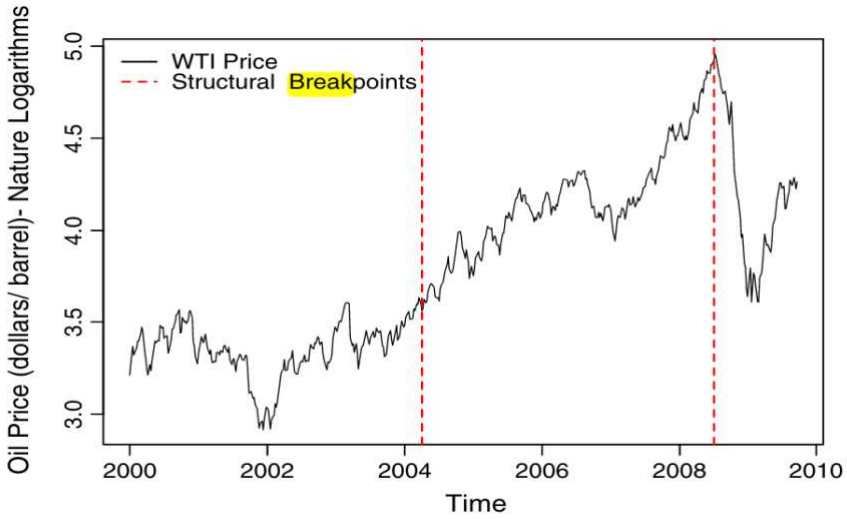

Figure 6: structural breaks in the WTI Price

(Fan & Xu, 2011, p. 1088)

Fan & Xu, (2011) describe that these break points mark two important changes in the underlaying mechanisms of the oil price market. Consequently, the first period (2000-2004, the phase before the first structural break) was characterized by relatively stable oil prices which were impacted by events like 9/11 and the Iraq war. The dynamics of supply and demand did not play an important role.

The first break point in 2004 was caused by a raising demand due the overall growing economy worldwide. Consequently, the post break period from 2004 to 2008, was dominated by an advancing world economy. Similar to Salisu & Fasanya (2013) and Aruga (2016), also Fan & Xu (2011) explain that the second structural break coincides with the global financial crisis in 2008. The following post

break period was marked by decreasing demand in oil, speculation only played a minor role, and the dynamics of demand and supply regained its importance (Fan & Xu, 2011).

4.2.2 Finance

In their literature review of structural breaks and their influence on financial time series, Andreou & Ghysels (2009) highlight that structural breaks influence key finance indicators such as volatility, financial return, equity premium and credit risk models.

One concrete analysis in this field was conducted by Meligkotsidou & Vrontos (2008) who investigated how structural breaks influence time series regarding certain hedge fund indices drawn from the Hedge Fund Composite Index and the Credit Suisse Hedge Fund Composite Index. More precisely, similar to the work of Mensi et al. (2014) and Aruga (2016) presented earlier in Chapter 4.2.1, also Meligkotsidou & Vrontos, (2008) made use of the Bai and Perron method. They detected structural breaks in so called single strategy indices like emerging markets, macro strategies and distressed securities among others (Meligkotsidou & Vrontos, 2008)

Emerging markets focuses on investing in developing countries, macro strategies try to gain from shifts in the prices of the stock markets, of currencies, commodities or interest rates by using financial instruments such as derivates and finally distressed securities strategies invests in securities of companies facing bankruptcy, restructuring or reorganization (Meligkotsidou & Vrontos, 2008). Similar to the Oil and tourism sector presented in Chapter 4.2.1 and 4.2.3, also the structural breaks found by Meligkotsidou & Vrontos (2008) coincide with major events. The break found in 2000 is related to the "*Technology and Internet bubble*" which marked a high decrease in stock prices for technology values (Meligkotsidou & Vrontos, 2008, p. 2472). Before this break the portfolios of investment fund heavily relied on technology values (Meligkotsidou & Vrontos, 2008).

By using structural break tests, Kalsie & Arora (2019) investigated the influence of the financial crisis taking place in 2008-2009 on a total of 26 macroeconomic variables. Examples for the variables which Kalsie & Arora (2019) observed are GDP growth, imports/exports of goods and services, final consumption expenditure, short term depth and real interest rate.

Furthermore, for their analysis they used data between the years 2000 to 2013 from five BRIC countries (Russia, China, India and South Africa), characterized by an emerging economy.

Kalsie & Arora (2019) highlight that including structural break in their model resulted in a more precise evaluation of the trend in the crisis period and in the ability to model the development of the variables in such periods. Ultimately, this also facilitated the detection of variables with a high sensitivity towards external shocks which could be than used in order to build a model able to indicate a developing crisis Kalsie & Arora (2019).

Another study regarding financial time series data was conducted by Caporale et al. (2020). They analyzed five European stock market indices (in Germany, the United Kingdom, France, Spain and Italy) for persistence, non-linearities and additionally accounted for structural breaks by using an extended approach based on Bai and Perron (2003). Caporale et al. (2020) discovered that the number of structural breaks in their data was between two and four. Furthermore, they point out that the majority of breaks fall in line with major events such as shutdown of US Federal government in 2013, the decision of the Federal Reserve System for increasing interest rates in 2016 and the publication of data regarding the GDP in the US, which led to an increase in the Dow Jones Index as well as European stock market indices (Caporale et al., 2020).

4.2.3 Tourism

Cró & Martins (2017) investigate the influence of structural breaks on the number of tourism arrival in 25 countries (also including Madeira Island) based on data between the years 1995 and 2014 provided by the World Bank. For their study they use the Bai and Perron method (which was introduced in Chapter 4.1.3) to detect structural breaks. After they have discovered the break dates, Cró & Martins (2017) compare theses dates with events considered as disasters and crises in the tourism sectors such as terrorist attacks, wars, natural disasters, an unstable political environment or diseases. For example, they investigate whether a relationship exists between the structural breaks in the tourist demand of Syria, Tunisia, Egypt as well as Israel and the Gulf War, the 9/11 attacks or the Arab Spring. Furthermore, they analyze the influence of financial as well as economic crises in Ireland, Greece, Cyprus, Portugal and

Spain. Cró & Martins (2017) conclude that the occurrence of tourism disaster and crises coincides with the structural break dates determined by the Bai and Perron method, leading to a lower number of tourists in the respective countries.

Similar to Cró & Martins (2017), also Mérida & Golpe (2016) used the Bai and Perron method to search for structural breaks in the tourism sector. In particular, they detected structural breaks in the relationship between the number of nights spent in accommodations in Spain and real exchanges rates in a time span ranging from 1980 to 2013.

Another study in this field was conducted by Kim & Seok (2018) who investigated possible structural breaks in GDP and tourism receipts. For their analysis, Kim & Seok used quarterly data from the Bank of Korea and Statistic Korea ranging from 1975 up to 2016. Similar to Mérida & Golpe (2016) and Cró & Martins (2017) they also found that the discovered structural break points were related to major events influencing the tourism sector. For example Kim & Seok (2018) discovered that the structural break in GDP was linked to a financial crisis taking place in Asia in 1997. The structural break in tourism receipts could be connected to a policy of the Korean government which heavily controlled outbound travel (Kim & Seok, 2018).

4.3 Literature review: The influence of structural breaks on judgmental forecasting

Not many studies have yet investigated how the presence of structural breaks influence forecasts made by individuals (Becker et al. 2009b). However, this section reviews the scarce literature on this topic and present its findings.

An early study regarding the influence of abrupt or slow changes on judgmental forecasting of time series was conducted by O'Connor et al. (1993). In their study they changed the noise level, the type of the break, the direction of the break between a total of 10 time series. Each of the time series was divided into four segments (segment 0 to segment 3). The initial segment (zero) displayed historical data to the participants, segment one was the first period which the participants had to forecast, the change was introduced in segment two and in segment three the time series returned to a more steady progression (O'Connor et al., 1993). Participants in this experiment were 17 students of the University of Hawaii.

The students have been informed beforehand that the time series might contain changes. These breaks could either be up or down and develop slowly or abrupt (O'Connor et al., 1993). In order to assess the forecast accuracy of the individuals O'Connor et al. (1993) used the absolute percentage error (APE). Results show that the error was higher for segment two which contained the break. For the third segment, the period in which the time series returned to a more stable state, APE decreased but remained higher compared to the first segment. Furthermore, O'Connor et al. (1993) point out that errors were higher for time series with a downward direction than for time series with an upward direction. Whether the change introduced into the time series was gradually or abrupt did not influence forecasting accuracy. Another finding of study was that the participants' forecasts were influenced by the latest value of the time series, which was presented to them (O'Connor et al., 1993). Moreover, the authors elaborate that the participants tended to see the break as noise which lead to the belief that the time series would ultimately return to its initial level (O'Connor et al., 1993).

Another study which analyzed judgmental forecasting performance for time series subject to structural breaks was conducted by Becker et al. (2009). Contrary to the experiment of O'Connor et al. (1993), the participants (120 students of economics and business administration in total) were not informed that the time series might contain structural breaks. In addition, they did not receive any contextual information. Another difference to the study of O'Connor et al. (1993) was that the participants only were given one value at the beginning of the forecasting task, compared to a sequence of past values. However, the first six periods were excluded from the evaluation (Becker et al. 2009). Furthermore, the participants received monetary incentives, meaning that in every period of the times series 60 cents were awarded for a forecast equaling the true value. The reward decreased by 20 cents for predictions with a discrepancy of one unit and by 40 cents for deviations of two units (Becker et al., 2009). In general, this led to a monetary gain of seven euros on average for a completion time of 30 minutes.

Regarding the construction of the time series, Becker et al. (2009) developed three different types of time series. In the first category each series contained one structural break, whereas in the second and third category the time series where subject to two breaks. In the second category, the first break (upward) occurred during a downward direction and the second one (decrease) during an upward trend.

In Addition, the two structural breaks in the second category were characterized by the same magnitude of the increase. The values in this category returned to the initial level after the last break.

In the third category the second break was characterized through an increase twice as large as the first one (Becker et al., 2009). For this type of time series, both structural breaks had an upward movement. For the analysis of the forecasting performance, they used the mean absolute error (MAE).

Results show that for the first type of time series (containing only one change) the variance was higher for the period after the first break, but Becker et al. (2009) explain that it quickly decreased to the previous level before the structural break. For the time series with two structural breaks, the participants had difficulties in predicting the right date of the break. Moreover, the adaption to the new progression of the time series required more periods than compared to category one. Also, half of the participants forecasted three structural breaks, although the series only contained two.

Overall, Becker et al. (2009) highlight that averaging the forecasts lead to a lower MAE than 90 percent of the participants (109 out of 120), which suggests using combined forecasts to increase accuracy. The average forecasts also were able to predict the two break dates in type one and two correctly, whereas as mentioned previously the individuals struggled to correctly predict the corresponding dates.

Also Durbach & Montibeller (2018) investigated how forecasts of time series are influenced by structural breaks. They refer to them as "*shocks" – "unexpected events that are extreme in magnitude, rare in frequency and short-lived in duration*" (p.1). The authors divided their analysis into four stages. The first one focused on how participants behave prior to the break, the second one dealt with their behavior when the break was happening, the third category was dedicated to an abrupt recovery and finally the last category was devoted to the period after the recovery (Durbach & Montibeller, 2018). All of the shocks were characterized by an abrupt downward motion (Durbach & Montibeller, 2018).

Furthermore, Durbach & Montibeller (2018) stated that the aim of their experiment was to analyze how "*learning about stage one, then experiencing stage two, affect judgments about stage three and four*" (p.2). The participants (96 students in total, from three different universities and programs) were expected to point forecast a total of 24 monthly data points and to imagine forecasting data of shares. Besides this information, no context was provided to the participants, and they were not informed that

the time series are subject to structural breaks. Similar to Becker et al. (2009) the students in this experiment received an incentive for the lowest overall error. Results of this study show that participants tend to underestimate the likelihood that a structural break reoccurs as well as the magnitude of the recovery period. The students showed a bias towards returning to the previous mean of the time series rather than to adjust to new highs in the recovery period. In general, the forecasts by the participants were lower in magnitude for the shock itself and the following recovery period. Furthermore, Becker et al. (2009) highlight that most of the participants expected the time series to quickly recover after the shock. This could be derived from the fact that only 10% of the forecasts after the break were as low as the break itself or lower.

5. Methodology & Research Design

5.1 Display of the data

As laid out in Chapter 3.2.2 the type of data visualization can influence the forecasting performance of participants. The previously reviewed literature on this topic shows that a majority of studies draw the conclusion that in general performance is higher if the data is presented graphically compared to tables (Harvey & Bolger, 1996; Lawrence et al., 2006; Theocharis et al., 2019). For example Harvey & Bolger (1996) argue that if the data is presented graphically people perform better as they underestimate trends in a lesser magnitude than when they are presented with tabular data. Furthermore, Lawrence et al. (2006) explain that especially for trended time series forecasts based on graphs are more accurate compared forecasts based on tabular data.

Therefore, the method of choice for the experiment of this thesis was to graphically present the time series to the participants.

5.2 Generating the time series

The participants are presented with a total of 12 different time series. The range of values for each time series ranges between 0-200.

In general, the experiment uses artificially generated data using Microsoft Excel in which, the x-axis of the time series will be displayed as 'Period'. The y-axis for each series is labeled as 'Value'. Generally, the approach of this experiment follows the one of O'Connor et al. (1993) described in Chapter. 4.3. Therefore, also in this experiments, time series with different characteristics are presented to the participants.

Time series which do not contain structural breaks are also added to the experiment, to be able to compare forecast accuracy between time series containing structural breaks and time series without such breaks. The experiment also investigates if crowd performance differs if the participants are faced with a trended change or a sudden change (structural break).

In addition, the experiment investigates if the type of the initial direction, the direction of the break or the direction after the break influences forecast accuracy and crowd wisdom. It also examines if the number of breaks impacts the performance and therefore four time series are introduced containing two structural breaks. Furthermore, the experiment evaluates forecast accuracy between time series subject to noise versus time series without random noise added.

The time series used in the experiment (zero or one structural break) are as followed:

Series type	Structural Break	Number of structural breaks	Initial Direction	Direction 2 (direction of the Break or change)	Direction 3 (direction after the Break)
5	Yes	1	Downward	Upward	Downward
6	Yes	1	Downward	Upward	Upward
7	No	None	Upward	Downward	-
9	No	None	Downward	Upward	-
10	Yes	1	Upward	Downward	Downward
12	Yes	1	Upward	Downward	Upward

Time series with noise:					
1	Yes	1	Upward (with noise)	Downward	Downward (with noise)
3	Yes	1	Downward (with noise)	Upward	Upward (with noise)

In order to limit the scope of the experiment and the time participants need to invest to complete it, the amount of time series with noise is set to two. Random values with a mean of zero and a standard deviation of zero are introduced using the Random number Generator of Excel. These random values are multiplied by 10 and without changing the original characteristic of each time series (downward or upward shaped) these values are added to the time series to introduce noise.

In general, the time series with two structural breaks will be displayed as followed:

Series type	Structural Break	Number of structural breaks	Initial Direction	Direction 2 (direction of the Break or change)	Direction 3 (direction after the 1st Break)	Direction 4 (direction of the 2nd break)	Direction 4 (after the 2nd break)
2	Yes	2	Downward	Upward	Upward	Downward	Downward
4	Yes	2	Upward	Downward	Downward	Upward	Upward
8	Yes	2	Upward	Downward	Upward	Downward	Downward
11	Yes	2	Downward	Upward	Downward	Upward	Upward

Segments:

The time series containing two structural breaks are divided into four segments, whereas all the other types have three different segments.

Segment 0 (periods 1-10):

For each time series Segment 0 contains 10 historical data points displaying the initial direction/trend of the series. For downward time series the values will be decreased by 10 for each period and for upward trended series the values will increase by the same amount.

Segment 1 (periods 11-15):

In Segment 1 the initial direction/trend will be continued, and values will decrease/increase by the same magnitude as in Segment 0. Furthermore, the participants have to forecast the values starting with Segment 1.

Segment 2 (periods 16-20)

Continuing with Segment 2 the structural break will be introduced in period 16.

In order to simulate a sudden shift in values, the value of the previous period is either multiplied (for simulating upward breaks) or divided (simulating downward breaks) by a number between 1.5 and 3 and rounded to an integer. After the break the time series is either upward or downward trended again (values increase/decrease by 10 for each period).

Segment 3 (periods 21-25):

For time series consisting of four segments, the second break will be modelled in period 21 and follows the logic mentioned in Segment 2 except that for this segment values only decrease or increase by a value of two after the structural break.

After creating the different time series, there order was randomized with the goal to prevent the participants from detecting certain patterns.

The following graphs display the 12 time series described above:

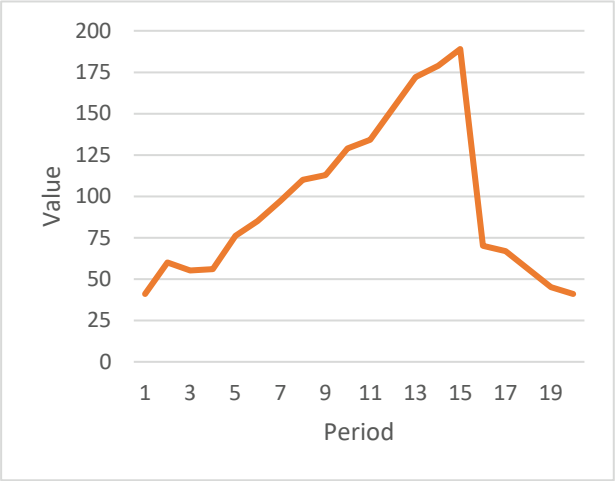

Figure 7: Time Series 1

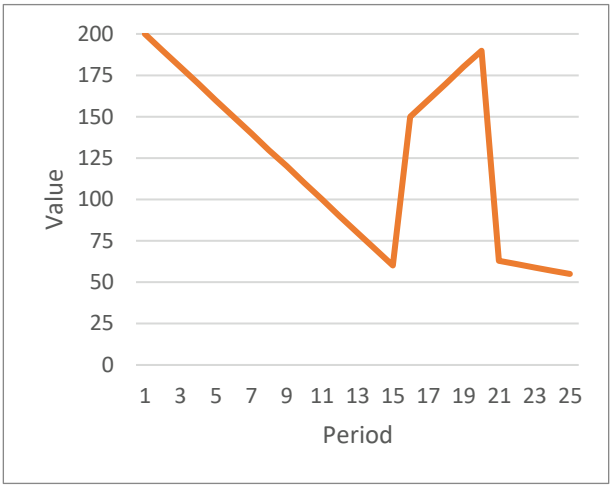

Figure 8: Time Series 2

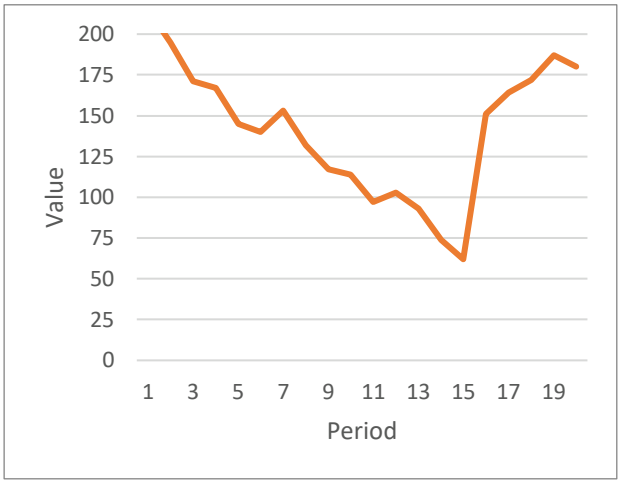

Figure 9: Time Series 3

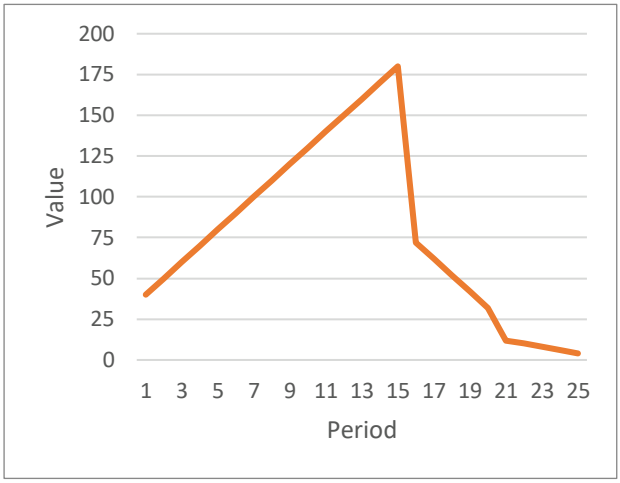

Figure 10: Time Series 4

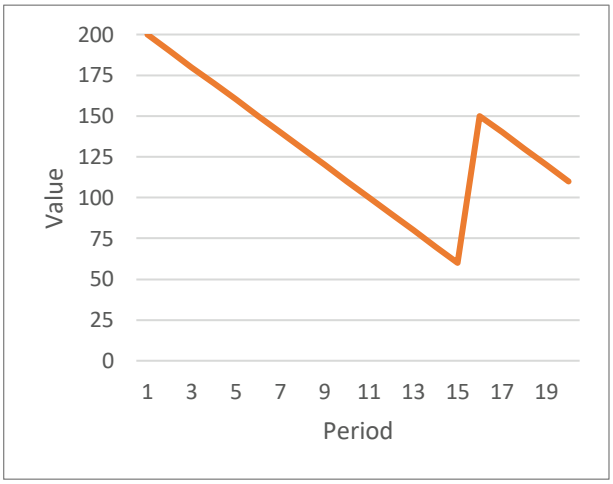

Figure 12: Time Series 5

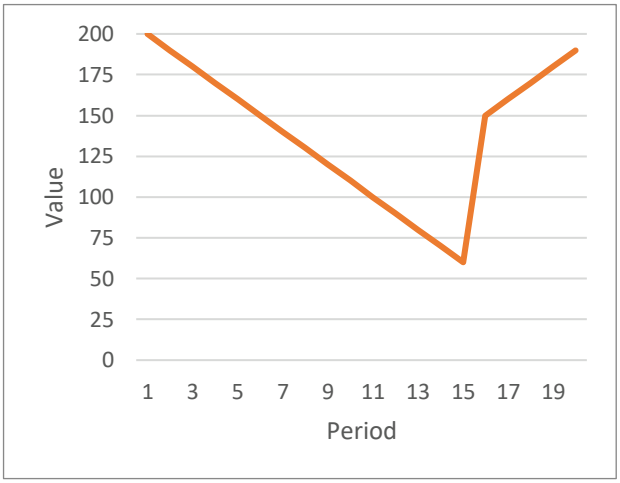

Figure 11: Time Series 6

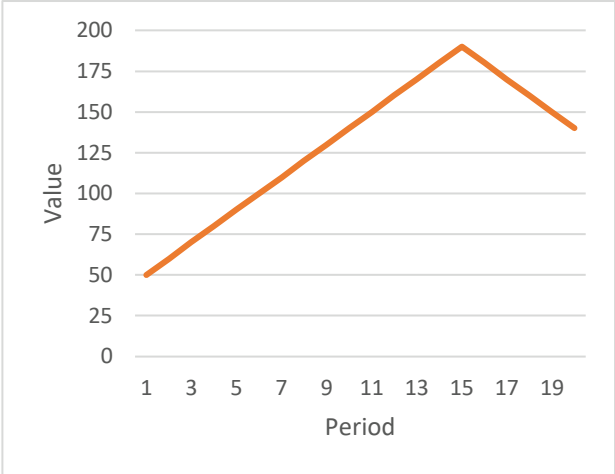

Figure 14: Time Series 7

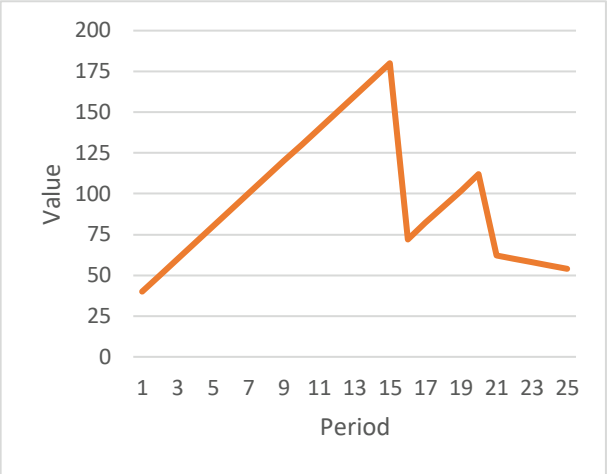

Figure 13: Time Series 8

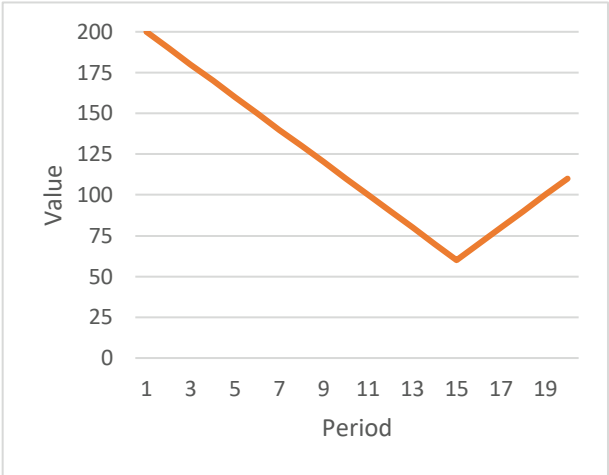

Figure 15: Time Series 9

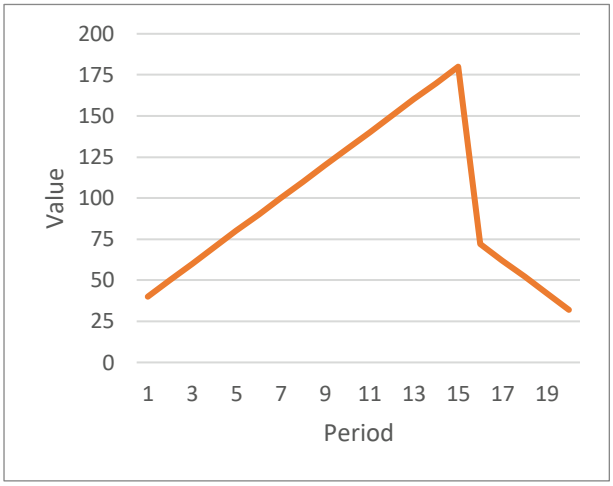

Figure 16: Time Series 10

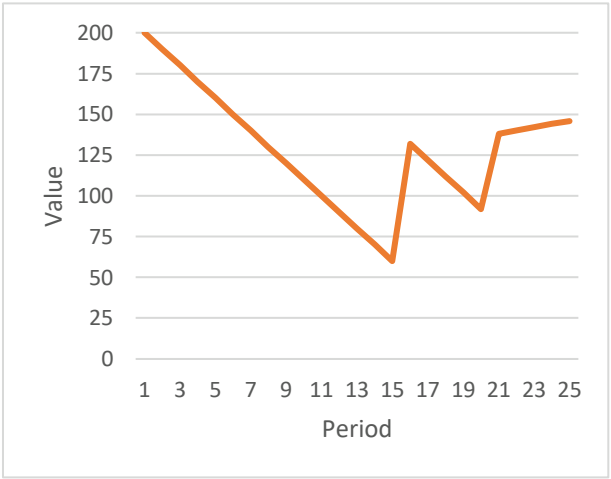

Figure 18: Time Series 11

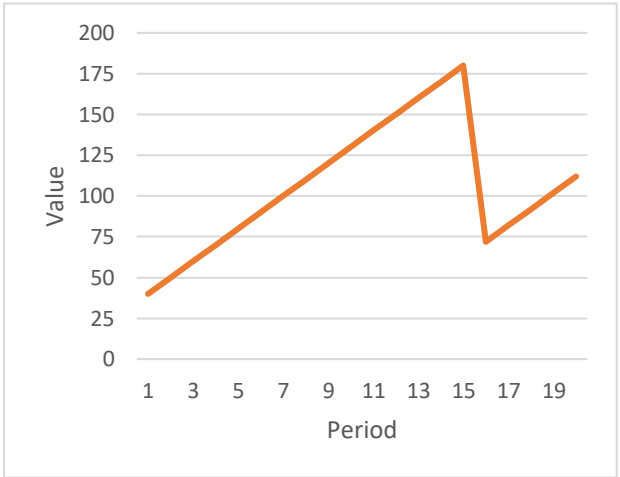

Figure 17: Time Series 12

5.3 Process of the study

The tool to carry out the online experiment is "SoSci Survey" – a tool for creating online surveys provided by the University of Vienna. For the previous generated time series each segment represents one question. The participants were informed of the importance to complete the experiment and forecast the values without consulting or communicating with others. In order to examine how diverse the participants were in terms of background, the following demographic data was collected from the participants:

- Gender (male, female, diverse, other)
- Age (18-25, 26-35, 36-45, 46-55, 56-65, >65)
- educational background (none, compulsory school, apprenticeship graduation, A-levels, university degree (bachelor),
- employment status (employed, unemployed, self-employed, student, retired),

Similar to O'Connor et al. (1993) the participants were informed that some of the series might contain structural breaks. Additionally, they were provided with a short definition of the term. Furthermore, they received a notification that for their forecasts the possible range of values is 0-200. However, following the definition of Alvarado-Valencia & Barrero (2014) presented in Chapter 3.1.3 no contextual information was provided to them, as the time series were entirely artificial without any connection to real world data. Moreover, generic labels were chosen over specific labels like "expenses, revenue" etc. for the time series, as these could great certain beliefs among participants about how the time series progresses, which might influence their forecasts (Goodwin & Wright, 1993; see Chapter 3.1.3). Regarding possible mechanism for improving forecasting accuracy outlined in Chapter 3.3, participants received feedback by presenting them the right direction after forecasting each period. Another approach presented was decomposition – dividing a more complex task into smaller less demanding tasks. Instead of asking the participants to forecast each time series at once, the forecasting task was decomposed into several smaller segments.

For each time series the participants had to point forecast a total of 10-15 values (5 for each Segment, starting with Segment 1). After forecasting one segment, the true values of it were shown to the participants. Additionally, after forecasting each series, the time series with the correct values was displayed. In general, the forecasts were made via text input and for each series type at once.

Example of 'Time Series 3' (Segment 2) the participants had to forecast in the tool 'SoSci Survey:

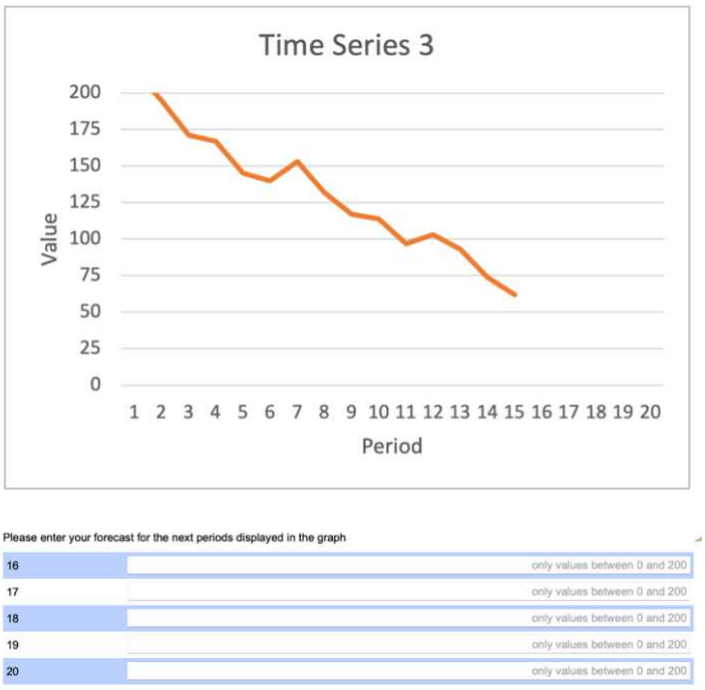

Figure 19: Time Series 3 in the tool "SoSci Survey"

5.4 Determining crowd wisdom & forecast accuracy

In order to determine the forecast accuracy of the individual participants the absolute errors were calculated over the 12 time series for every forecast:

$$Absolute\ error_{i,p,T} = |\, y_{p,T} - z_{i,p,T}|$$

- $y_{p,T}$ ……true value at period p for time series T.
- $z_{i,p,T}$….. predicted value by participant i at period p for time series T.

Secondly, the individual Mean Absolute Error (MAE) and the median of the absolute errors was calculated for each individual over the 12 time series. MAE was chosen as error metric as it is less sensitive to outliers than the MSE, meaning that it is less influenced by a small number of extreme errors (Botchkarev, 2018; see Chapter 1.1).

For the crowd forecasts, the predictions of all the individuals were aggregated for each data point by using the mean as well as the median.

This was regarded as the crowd prediction for the respective period of the time series. Similar to the individual error, the crowd error term was calculated by the absolute deviation of the true value. Based on the mean as aggregation method, for one period (x) this can be denoted as:

$$Absolute\ Error\ (crowd)x = \frac{1}{n}\sum_{i=1}^{n} Forecast_{i,x} - True\ Value_x$$

- $Forecast_{i,x}$ …..forecast of the i-th individual for period x
- $True\ Value_x$ …the true value for period x
- n……number of forecasts (individuals in the crowd)

Whereas with the median as aggregation method:

$$Absolute\ Error\ (crowd)x = \ Median\ (Forecast_{1,x,}, Forecast_{,x}, \dots, Forecast_{n,x}\ ) - True\ Value_x$$

- $Forecast_{i,x}$ …..forecast of the i-th individual for period x
- $True\ Value_x$ …the true value for period x
- n……number of forecasts (individuals in the crowd)

These (absolute) crowd errors were also aggregated through the MAE as well as the median over different segments and for the whole time series. This functioned as indicators for crowd performance. The procedure allowed to compare the forecasting performance of the individuals to the one of the crowd based on segments and the whole time series. In particular, this comparison was expressed by the percentage of individuals who displayed a higher MAE than the one of the crowd - referred to as

percentage beat by the crowd in this thesis. This term as well as the process is based on the approach described by Davis-Stober et al. (2014).

For this thesis a crowd is considered as wise if is able to beat more than 50% of all the individuals. The analysis of crowd wisdom concentrates on comparing the performance on the initial trend of the time series, the first break, the period after the first break, the second break and finally the period after the second break. Furthermore, it analysis how the direction of the structural break or the direction of the trend influenced crowd wisdom.

Dividing the analysis into these segments, allows to investigate how crowd wisdom changes after different stages of the time series, for example how the crowd behaves after the first break, how it is influenced by the second break and how structural breaks influence crowd predictions in general compared to periods which are not subject to sudden changes.

Furthermore, in order to examine how crowd wisdom is influenced by group size, smaller subgroups of the participants were created randomly. After selecting a random subgroup of three, five, nine and fifteen participants the performance was compared to the one of the whole crowd for each time series based on the MAE.

All of the above processes and calculations were modelled by using the programming language Python. The code used is laid out in Appendix 1.

6. Results

A total of 71 individuals participated in the experiment. The majority of them was female (44) whereas 27 were male. The greatest number of individuals were in the age of 26 to 35 followed by the group of individuals aged between 36 to 45.

Regarding the employment status more than half of the participants stated that they are currently employed, whereas only 14 participants were students. Retirees were the least represented group with 8 out of 71 participants.

With respect to the highest level of education the highest number of participants had a bachelor's degree followed by a master's degree. The third biggest group was represented by subjects who completed the A-levels. Tables 1 to 4 display the demographic data of the participants in detail:

Table 1: Gender of participants

GENDER	Nr. of participants
Female	44
Male	27
non-binary	0
other	0
not answered	0

Table 2: Age distribution

AGE	Nr. of participants
18-25	5
26-35	27
36-45	20
46-55	7
56-65	6
>65	6
not answered	0

Table 3: Employment status

EMPLOYMENT	Nr. of participants
employed	47
unemployed	0
student	14
retired	8
other	2
not answered	0

Table 4: Highest level of education

EDUCATION	Nr. of participants
none	0
compulsory school	0
apprenticeship	6
A Levels	16
bachelor's degree	24
master's degree	22
doctorate	1
other	2
not answered	0

6.1 Overall crowd performance

As we can see from Table 1, the percentage beat by the crowd was above 50% for all 12 time series. Up until Time Series 5 the mean absolute errors are decreasing. Overall, the crowd performance was highest for time series containing two structural breaks (74.0%), followed by time series which contained only a smooth change in trend and no break (73.9%) on average. The forecasts of the crowd had the lowest MAE for Time Series 9 with no structural break. The lowest Median of 10 was in Time Series 7. The performance of the crowd was worst if the time series was subject to one structural break and additionally contained noise – for which the crowd also displayed the highest forecasting errors and the lowest percentage of individuals beaten. However, in that case it was still able to perform better than 65.5% of the individual subjects.

The MAE was the highest in Time Series 1 with one structural break and the lowest in Time Series 9 (no structural break), which was the time series with highest percentage beat by the crowd. The median peaked for Time Series 2 which contained two structural breaks.

If we compare the four time series with one structural break and without noise with the four time series containing two structural breaks, the percentage beat by the crowd on average is greater for time series with two structural breaks, 69.0% compared to 74.0% respectively.

The least percentage was reached in Time Series 4 (containing two structural breaks) with 59.2% of individuals displaying a higher forecasting error than the crowd.

Table 5: overall crowd performance

Time Series	MAE of the crowd	Median of absolute errors of the crowd	% beat by the crowd (based on MAE)	Number of breaks
Crowd size (nr. of participants)	71	71	71	
Time Series 1	70.2	71	63.4%	1
Time Series 2	69.4	98	73.2%	2
Time Series 3	57.3	64	67.6%	1
Time Series 4	56.5	50	59.2%	2
Time Series 5	38.5	38	74.7%	1
Time Series 6	55.0	60	67.6%	1
Time Series 7	25.2	10	60.6%	0
Time Series 8	39.6	53	78.9%	2
Time Series 9	14.0	15	87.3%	0
Time Series 10	55.3	52	70.4%	1
Time Series 11	30.2	37	78.9%	2
Time Series 12	34.3	29	63.4%	1
Ø MAE:				
0 structural breaks:		19.6		
1 structural break (no noise):		45.8		
1structural break (with noise):		63.8		
1 structural break (total):		51.8		
2 structural breaks:		48.9		
Ø percentage beat by the crowd:				
0 structural breaks:		73.9%		
1 structural break (no noise):		69.0%		
1 structural break (with noise):		65.5%		
1 structural break (total):		70.4%		
2 structural breaks:		74.0%		

TS 1 & TS 3 contained noise

Regarding the overall predictions for structural breaks by the crowd, it was observed that in many cases the crowd anticipated a structural break in the right period, but often forecasted the opposite direction of the structural break. This could be seen in Time Series 1 (Figure 20), Time Series 3 (Figure 22), Time Series 5 (Figure 24), Time Series 6 (Figure 25) and Time Series 9 (Figure 28). The same could be observed for the first break in Time Series 11 (Figure 30).

Furthermore, results show that the crowd damped the magnitude of the correctly forecasted breaks. For Time Series 8 (Figure 27) the crowd correctly anticipated the structural break in period 21, however with a less magnitude compared to the true data. Also, the second break for Time Series 11 was anticipated by the crowd, however again by a smaller magnitude than the true values.

For the Time Series 10 (Figure 29) the crowd completely failed to forecast the structural break and for Time Series 12 (Figure 31) the crowd prediction in period 16 rather showed a change in trend than the actual structural break.

Another major observation by comparing the true values with the aggregated crowd forecast is, that crowd also damped both upward and downward trends. This was especially evident in the upward initial trend of Time Series 4 (Figure 23) Time Series 8, Time Series 10 as well as in the downward initial trend of Time Series 5 and Time Series 6

Results show that for various downward trends the crowd correctly forecasted several periods but expected the trend to stop earlier. This could be observed for the initial period (11-15) in Time Series 2 and Time Series 9.

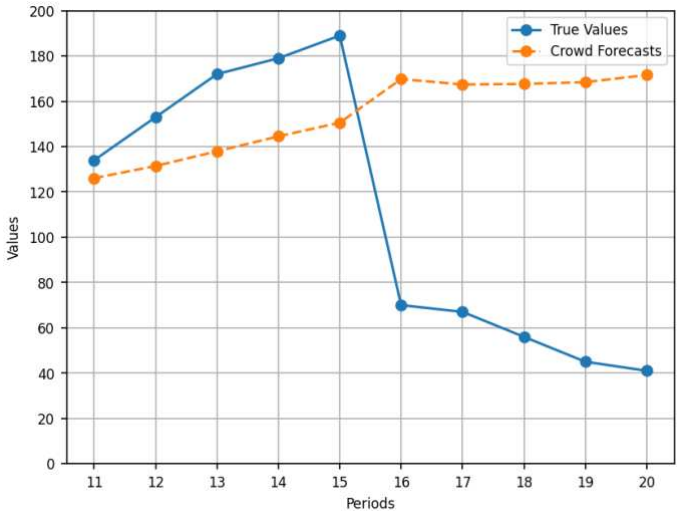

Figure 20: Time Series 1 - True values vs. crowd predictions

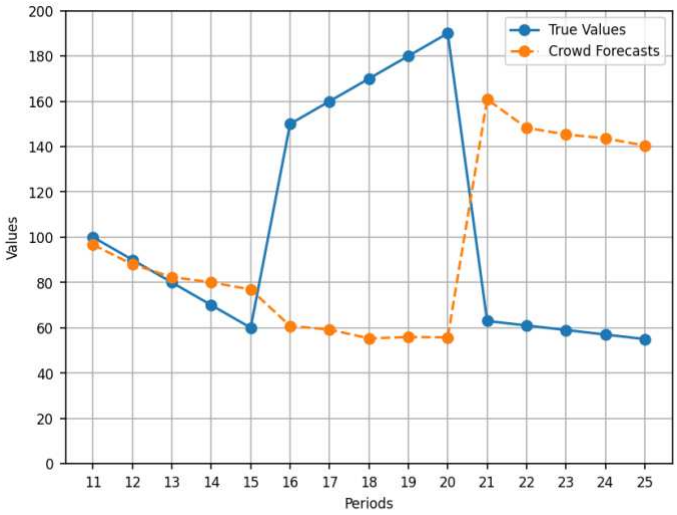

Figure 21: Time Series 2 - True values vs. crowd predictions

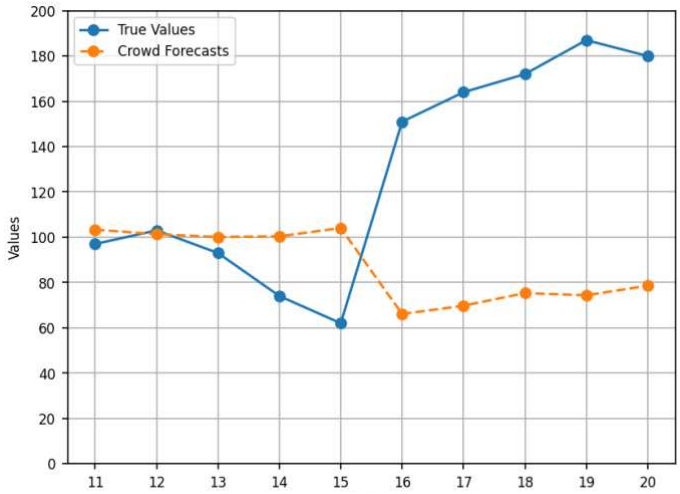

Figure 22: Time Series 3 - True values vs. crowd predictions

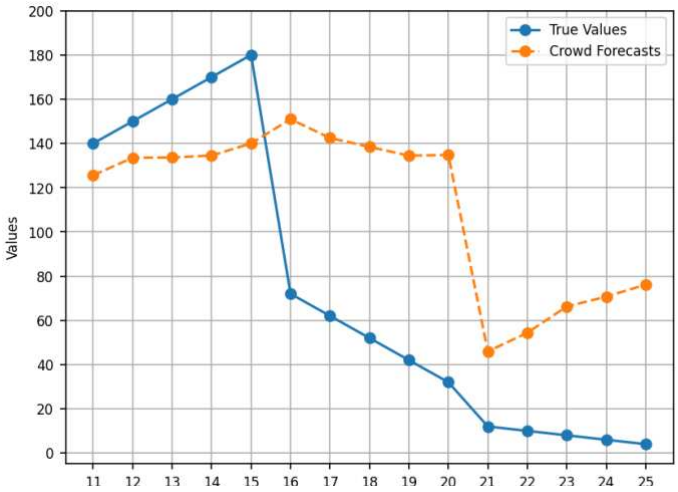

Figure 23: Time Series 4 - True values vs. crowd predictions

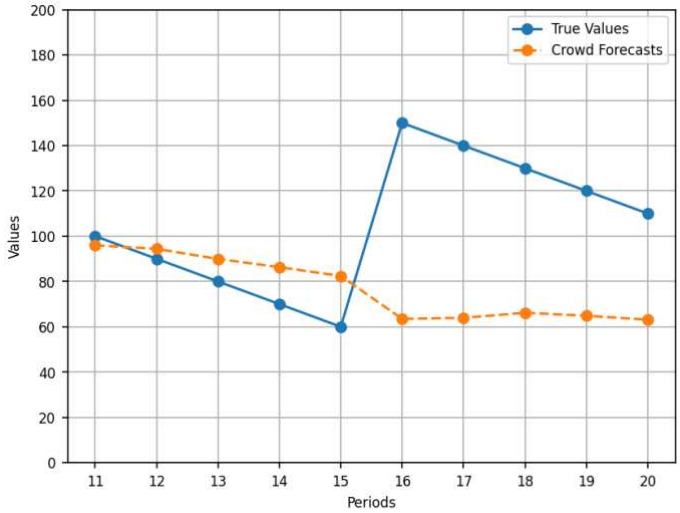

Figure 25: Time Series 5 - True values vs. crowd predictions

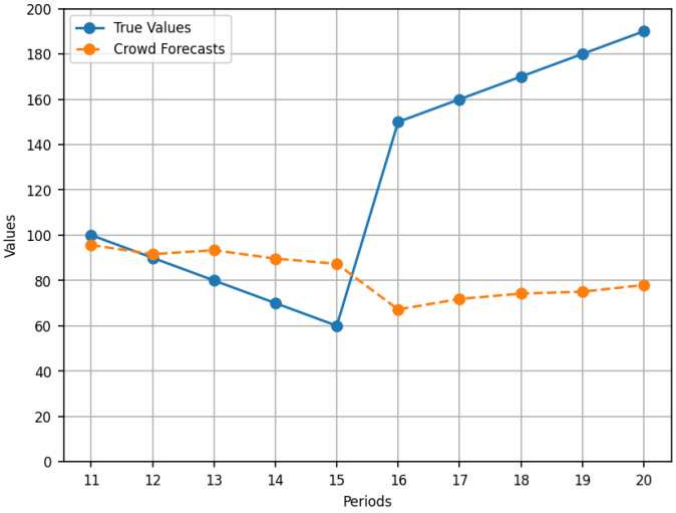

Figure 24: Time Series 6 - True values vs. crowd predictions

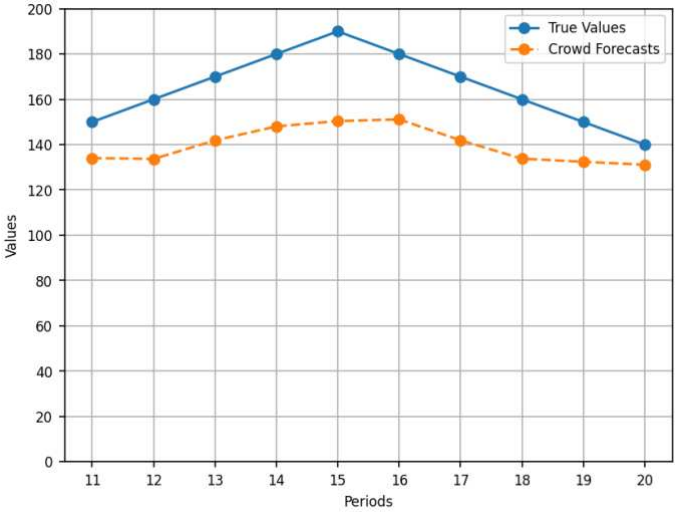

Figure 26: Time Series 7 - True values vs. crowd predictions

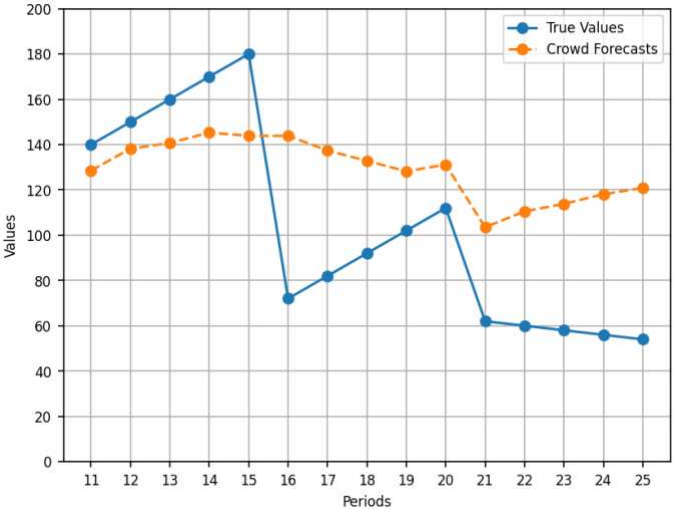

Figure 27: Time Series 8 - True values vs. crowd predictions

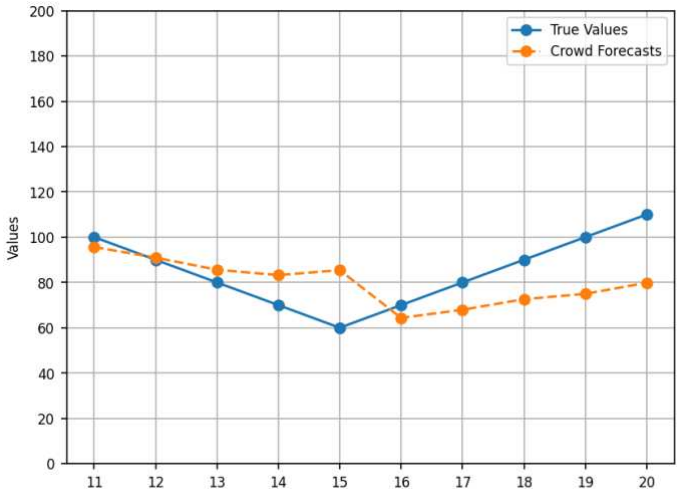

Figure 29: Time Series 9 - True values vs. crowd predictions

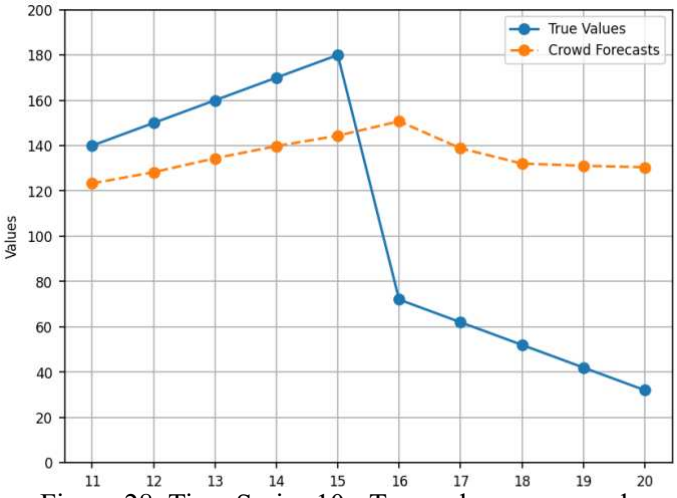

Figure 28: Time Series 10 - True values vs. crowd predictions

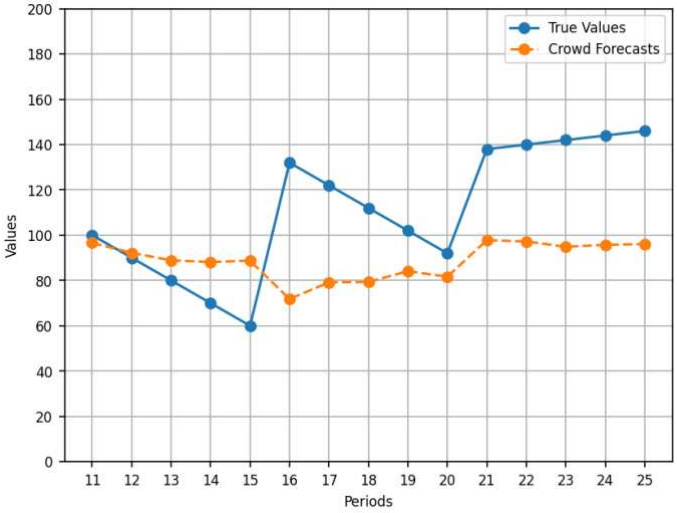

Figure 31: Time Series 11 - True values vs. crowd predictions

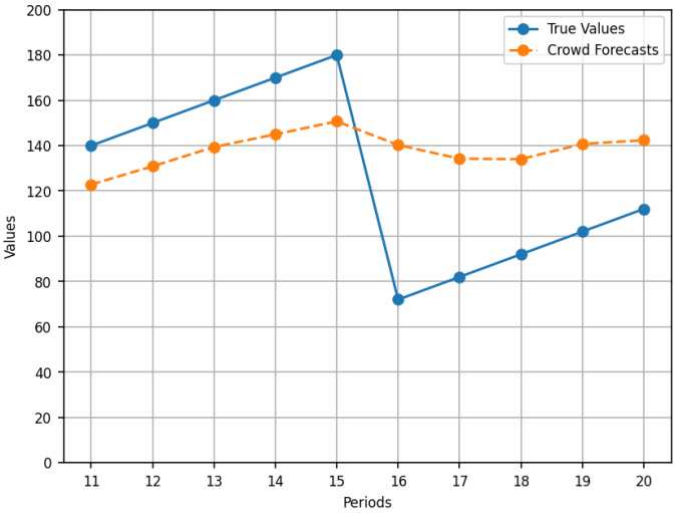

Figure 30: Time Series 12 - True values vs. crowd predictions

Following this general analysis the following sections analyze the different periods of the time series in more detail, especially in terms of forecasting error and crowd performance compared to those of the individual participants.

6.1.1 Initial Trend

Looking at the results for the initial trend (periods 11 to 15), the crowd performed better if the initial trend of the time series was in a downward direction compared to an upward movement. The percentages of individuals beat were 64.1% versus 36.6% on average. This shows that the individual participants were able to outperform the collective for upward initial trends. Furthermore, the forecasting error was lower for downward directions overall.

Time Series 3 which contained noise has the highest error compared to other time series with a downward initial trend. Also, for time series with an upward trend, Time Series 1 which contained noise was the one with the highest errors.

For time series with an upward initial, trend Time Series 7 which contained no break, reached the highest percentage beat by the crowd.

Table 8: Crowd performance for the initial trend

Time Series	Number of breaks	Initial trend	MAE	% beat by crowd
Crowd size (nr. of participants)			71	71
Time Series 2	2	0	6.9	73.2
Time Series 3	1	0	16.7	69.0
Time Series 5	1	0	11.4	56,3
Time Series 6	1	0	13.3	62,0
Time Series 9	0	0	9.9	69,0
Time Series 11	2	0	12.3	54,9
Time Series 1	1	1	27.3	35,2
Time Series 4	2	1	26.5	36,6
Time Series 7	0	1	28.4	40,9
Time Series 8	2	1	20.7	36,6
Time Series 10	1	1	26.0	39,4
Time Series 12	1	1	22.3	31.0

0 = Down; 1 = UP

Ø MAE for downward initial trend:	11.8
Ø MAE for upward initial trend:	25.2
Ø percentage beat for downward initial trend:	64,1%
Ø percentage beat for upward initial trend:	36,6%

TS 1 & TS 3 contained noise

6.1.2 First structural break

For the first structural break in period 16, the errors on average were lower for downward breaks with a MAE of 79.6 compared to 80.7 for upward structural changes. Moreover, the percentage of individuals beat by the crowd was higher compared to breaks with an upward direction.

Overall, the percentage of individuals beat by the crowd is higher than 50% for all time series except for Time series 7 which contained no break.

The structural break in Time Series 1 displayed the highest forecasting error followed by the break in Time Series 2 which had a upward direction.

The structural break in time series 1 has the highest percentage followed by the break in Time Series 8 which also had an upward trend.

Also considering the time series with a smooth change in trend, the highest percentage beat by the crowd overall was reached in Time Series 9 with 87,3%, which contained a smooth upward change in trend. This time series also had the lowest error in the period of the change compared to all other time series. Furthermore, it can be observed that up until Time Series 4 the forecasting errors for the structural breaks were decreasing – from a MAE of 99.8 to 62.0

Table 11: Crowd performance for the 1st structural break

Time Series	Number of breaks	Initial Trend	Direction 1st break/change	Absolute error (1st break)	Percentage beat by crowd (%)
Crowd size (nr. of participants)				71	71
Time Series 1	1	1	0	99.8	83.1
Time Series 8	2	1	0	72.0	77.5
Time Series 4	2	1	0	79.0	62.0
Time Series 12	1	1	0	68.3	62.0
Time Series 10	1	1	0	78.7	59.2
Time Series 7	0	1	0	28.9	42.3
Time Series 9	0	0	1	5.6	87.3
Time Series 2	2	0	1	89.2	71.8
Time Series 5	1	0	1	86.5	69.0
Time Series 6	1	0	1	82.8	64.8
Time Series 3	1	0	1	84.9	63.4
Time Series 11	2	0	1	60.2	59.2
0 = Down; 1 = UP					
MAE for downward break:			79.6		
MAE for upward break:			80.7		
Ø percentage beat for downward break:			64,3%		
Ø percentage beat for upward break:			67,2%		
TS 1 & TS 3 contained noise					

6.1.3 Period after the first structural break or change in trend

Similar to the initial trend we can observe that the errors for downward trends in the period after the first structural break or change in trend (periods 17 to 20) are slightly lower than those for upward trends with a MAE of 65.7 and 69.6 respectively. However, whereas in the initial trend the average percentage of individuals beat by the crowd was higher if the trend was downward, this changed that for the trend after the first break. For the periods after the first break or change the average percentage of individuals beat by the crowd is slightly lower for if the trend is downward (66.7% versus 67.9% for upward trends). If the errors in the initial trend are compared with the ones trend after the first break, it is evident that the errors were higher in the periods after the 1st break compared to the initial trend. On average the MAE was three times higher for downward trends and six times higher for upward trends.

The highest errors were in Time Series 2 in which the trend after the break was upward, followed by Time Series 1 were the trend was downward.

Furthermore, the percentage of individuals beat by the crowd increased compared to the initial trend for the majority of the time series. It only decreased for Time Series 2, Times Series 3 and Time Series 6.

Table 14: Crowd performance in the period after the first break/change in trend

Time Series	Number of structural breaks	Trend after 1st break/change	MAE Initial trend	MAE after 1st break/change	% beat by crowd Initial trend	% beat by crowd after 1st break/change
Time Series 1	1	0	27.3	116.5	35.2	71.8
Time Series 4	2	0	26.5	90.5	36.6	54.9
Time Series 5	1	0	11.4	60.5	56.3	62.0
Time Series 7	0	0	28.4	20.2	40.9	77.5
Time Series 10	1	0	26.0	86.1	39.4	60.2
Time Series 11	2	0	12.3	20.2	54.9	73.2
Time Series 2	2	1	6.9	118.5	73.2	63.4
Time Series 3	1	1	16.7	101.3	69.0	62.0
Time Series 6	1	1	13.3	100.2	62.0	54.9
Time Series 8	2	1	20.7	35.4	36.6	84.5
Time Series 9	0	1	9.9	21.1	69.0	76.1
Time Series 12	1	1	22.3	40.8	31.0	66.2
Number of Participants			71	71	71	71

0 = Down; 1 = UP

Ø MAE for downward trends after 1st break/change:	65.7
Ø MAE for upward trends after 1st break/change:	69.6
Ø percentage beat for downward trends after 1st break/change:	66.6%
Ø percentage beat for upward trends after 1st break:	67.9%
Ø increase in MAE compared to initial trend (DOWN):	210.9%
Ø increase in MAE compared to initial trend (UP):	507.4%

TS 1 & TS 3 contained noise

6.1.4 Second break

In the second structural break (period 21) the percentage beat by the crowd decreased for all four time series. Furthermore, also the absolute errors in the second break were smaller compared to the first break except for Time Series 2. In this time series direction of the second break differed from the direction of the 1st break and errors increased whereas in all other cases the errors decreased compared to the first structural break. For Time Series 4, the crowd was only able to beat 38.0% (see Table 9).

Table 17: crowd performance for the second structural break

Time Series	Number of breaks	Direction 1st break/change	Direction 2nd break	Absolute error 1st break	Absolute error 2nd break	% beat by crowd 1st break	% beat by crowd 2nd break
Crowd size (nr. of participants)				71	71	71	71
Time Series 2	2	1	0	89.2	97.9	71.8	70.4
Time Series 4	2	0	0	79.0	33.9	62.0	38.0
Time Series 8	2	0	0	72.0	41.5	77.5	64.8
Time Series 11	2	1	1	60.2	40.2	59.2	54.9

6.1.5 Period after the second break

Compared to the trend after the first structural break, the MAE increased in two of the four time series (Time Series 8 and Time Series 11) for the period after the second break (periods 22-25). However, the performance of the crowd measured by individuals with a higher MAE, decreased in the period after the second break – in three out of four time series the percentage of individuals beaten by the crowd declined.

Table 20: crowd performance in the period after the second structural break

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				71	71	71	71
Time Series 2	2	1	0	118.5	86.4	63.4	66.2
Time Series 4	2	0	0	90.5	59.4	54.9	46.5
Time Series 8	2	0	0	35.4	58.8	84.5	59.2
Time Series 11	2	1	1	20.2	47.0	73.2	52.1

6.2 Performance of small crowds

As summarized in Table 11, for smaller crowds results showed that for most time series the MAE is constantly decreasing with an increasing number of individuals within the crowd. Only for Time Series 1 the MAE stays constant over the different crowd sizes ranging from three to nine participants. Furthermore, for Time Series 4 the MAE slightly increased with crowd size. For this time series the crowd consisting of three randomly selected participants was able to slightly outperform the whole crowd with a MAE of 56.4 compared to 56.5. Also, for Time Series 12 the crowd with 15 participants was able to achieve a slightly lower MAE than the 71 participants. Similarly, the randomly selected 15 participants for Time Series 8 were able to perform equally to the 71 participants with a MAE of 39.6 for both crowds. Both Time Series 4 and Time Series 8 contained two structural breaks.

As the errors decreased with increasing group size, it could be observed that consequently the percentage of individuals beaten by the crowd increased with group size. However, for Time Series 2, Time Series 3 and Time Series 12 the changes of the MAE over the group sizes from 5 to 9 were too small in order to influence the percentage beat by the crowd. Consequently, the percentage stayed constant for the mentioned time series and crowd sizes. These results are summarized in Table 12.

Table 23: overall performance of small crowds (MAE)

Time Series	overall MAE of the crowd	MAE of select crowd				Number of breaks
Crowd size (nr. of participants)	71	3 (Random)	5 (Random)	9 (Random)	15 (Random)	
Time Series 1	70.2	70.3	70.3	70.3	70.2	1
Time Series 2	69.4	71.9	70.8	70.1	70.0	2
Time Series 3	57.3	60.1	58.8	58.2	57.8	1
Time Series 4	56.5	56.4	56.5	56.5	56.8	2
Time Series 5	38.5	42.2	41.1	39.6	39.2	1
Time Series 6	55.0	58.7	57.3	56.3	55.6	1
Time Series 7	25.2	30.6	27.7	25.9	26.0	0
Time Series 8	39.6	41.8	40.3	40.0	39.6	2
Time Series 9	14.0	22.1	19.0	16.5	15.2	0
Time Series 10	55.3	55.5	56.3	55.3	55.2	1
Time Series 11	30.2	34.1	32.3	31.3	30.7	2
Time Series 12	34.3	36.4	34.7	34.3	34.2	1

Table 26: overall performance of small crowds (% beat of individuals)

Time Series	% beat by the crowd (based on MAE)					Number of breaks
Crowd size (nr. of participants)	71	3 (Random)	5 (Random)	9 (Random)	15 (Random)	
Time Series 1	63.4%	62.0%	62.0%	62.0%	63.4%	1
Time Series 2	73.2%	69.0%	71.8%	71.8%	71.8%	2
Time Series 3	67.6%	64.8%	67.6%	67.6%	67.6%	1
Time Series 4	59.2%	59.2%	59.2%	57.8%	56.3%	2
Time Series 5	74.7%	69.0%	71.8%	74.7%	74.7%	1
Time Series 6	67.6%	64.8%	64.8%	66.2%	67.6%	1
Time Series 7	60.6%	53.5%	59.2%	60.6%	60.6%	0
Time Series 8	78.9%	73.2%	77.5%	78.9%	78.9%	2
Time Series 9	87.3%	73.2%	83.1%	85.9%	85.9%	0
Time Series 10	70.4%	69.0%	66.2%	70.4%	70.4%	1
Time Series 11	78.9%	67.6%	70.4%	71.8%	76.1%	2
Time Series 12	63.4%	59.2%	63.4%	63.4%	63.4%	1

6.2.1 Initial Trend

On average the crowd consisting of 15 participants had the same MAE of 25.2 as the whole crowd for initial trends with an upward direction.

For downward initial trends, the crowd with 71 individuals was able to achieve the lowest MAE on average. Also, for the individual time series none of the smaller crowds could outperform the whole crowd in the downward initial trends. Furthermore, the individual time series, the crowd with five participants was able to achieve a slightly lower MAE for the initial trend in Time Series 1 (27.0 vs. 27.3). However, the difference was too small in order to affect the percentage beat by the crowd and was 35.2% for both groups. For Time Series 8, the group of 15 people could reach a lower MAE than the 71 participants.

On average the whole crowd could reach the highest percentages of individuals beaten for both upward and downward initial trends. However, especially for upward initial trends the differences regarding this percentage were only minor with a range of values between 35.4 and 36.6.

Results are summarized in Table X.

Table 29: Performance of small crowds in the initial trend

Time Series	Number of breaks	Initial trend	MAE					% beat by crowd				
Crowd size (nr. of participants)			71	3	5	9	15	71	3	5	9	15
Time Series 2	2	0	6.9	13.2	11.7	9.6	8.4	73.2	46.5	54.9	64.8	67.6
Time Series 3	1	0	16.7	22.0	19.8	18.6	17.5	69.0	54.9	60.6	60.6	64.8
Time Series 5	1	0	11.4	18.1	15.4	13.8	12.5	56.3	40.9	45.1	50.7	50.7
Time Series 6	1	0	13.3	20.3	18.6	15.9	15.1	62.0	49.3	56.3	57.8	57.8
Time Series 9	0	0	9.9	18.3	15.3	13.0	11.8	69.0	40.9	50.7	54.9	60.6
Time Series 11	2	0	12.3	18.5	16.4	14.6	13.5	54.9	46.5	46.5	49.3	52.1
Ø Downward initial trend			11.8	18.4	16.2	14.3	13.1	64.1	46.5	52.4	56.4	58.9
Time Series 1	1	1	27.3	27.6	27.0	27.5	27.3	35.2	32.4	35.2	33.8	35.2
Time Series 4	2	1	26.5	27.0	26.7	27.0	26.7	36.6	36.6	36.6	36.6	36.6
Time Series 7	0	1	28.4	29.8	29.2	28.9	28.6	40.9	39.4	39.4	40.9	40.9
Time Series 8	2	1	20.7	21.0	21.6	21.2	20.3	36.6	36.6	36.6	36.6	36.6
Time Series 10	1	1	26.0	27.1	26.0	26.6	26.1	39.4	38.0	39.4	38.0	38.0
Time Series 12	1	1	22.3	23.5	22.7	22.2	22.4	31.0	29.6	31.0	31.0	31.0
Ø Upward initial trend			25.2	26.0	25.5	25.6	25.2	36.6	35.4	36.4	36.2	36.4

6.2.2 First structural break

Contrary to the overall performance described in Chapter 6.2, the errors for the first break did not clearly diminish with an increasing group size.

For the first structural break the crowd with five participants was able to outperform the whole crowd of 71 participants and reached the lowest error and highest percentage of individuals beaten for structural breaks with a downward direction. Crowds with nine randomly selected individuals reached the same average errors as the whole crowd of 71 participants, followed by the crowd with 15 participants for which the errors only differed by 0.1.

Overall, the errors for downward breaks were more dispersed than for upward structural breaks. For those breaks the forecasting errors were nearly the same over the different crowd sizes and only had a range of 80.4 to 80.7 on average. For upward structural beaks, the whole crowd performed the worst with a MAE of 80.7, closely followed by the group of 80.6. Groups with three, five and nine participants reached a smaller MAE of 80.4. Due to the fact that the differences in the MAE were only minor the

percentage beat by the crowd only were similar for upward structural breaks over the different crowd sizes. The crowd with five participants lead to the lowest percentage of 64.8.

The performance over the different crowd is summarized in the following table:

Table 31: Performance of small crowds for the first break/change

Time Series	Number of breaks	Initial Trend	Direction 1st break/change	Absolute error (1st break/change)					Percentage beat by crowd (%)				
Crowd size (nr. of participants)				71	3	5	9	15	71	3	5	9	15
Time Series 1	1	1	0	99.8	100.2	100.2	99.6	99.7	83.1	76.1	76.1	83.1	83.1
Time Series 8	2	1	0	72.0	72.8	70.3	72.4	71.8	77.5	74.6	77.5	74.6	77.5
Time Series 4	2	1	0	79.0	80.6	77.3	79.2	79.1	62.0	62.0	73.2	62.0	62.0
Time Series 12	1	1	0	68.3	69.4	67.8	68.0	68.8	62.0	62.0	63.4	62.0	62.0
Time Series 10	1	1	0	78.7	78.6	78.1	78.7	79.2	59.2	59.2	59.2	59.2	59.2
Time Series 7	0	1	0	28.9	31.6	30.3	28.3	29.2	42.3	26.7	26.7	42.3	42.3
Ø Downward break (excl. TS 7)				79.6	80.3	78.7	79.6	79.7	68.8	66.8	69.9	68.2	68.8
Time Series 9	0	0	1	5.6	15.5	12.3	9.2	7.6	87.3	74.7	76.1	85.6	85.9
Time Series 2	2	0	1	89.2	89.7	89.6	89.3	89.1	71.8	71.8	71.8	71.8	71.8
Time Series 5	1	0	1	86.5	86.4	86.7	84.6	86.2	69.0	69.0	69.0	69.0	69.1
Time Series 6	1	0	1	82.8	82.0	83.8	82.8	82.8	64.8	64,8	64.8	64.8	64.8
Time Series 3	1	0	1	84.9	83.7	86.0	84.6	84.9	63.4	63.4	59.2	63.4	63.4
Time Series 11	2	0	1	60.2	60.4	59.5	60.5	59.9	59.2	59.2	59.2	59.2	59.2
Ø Upward break (excl. TS 9)				80.7	80.4	80.4	80.4	80.6	65.6	65.6	64.8	65.6	65.7

0 = Down; 1 = UP

6.2.3 Period after the first structural break or change in trend

Similar to the overall performance, the MAE for the period after the first structural break, is decreasing with group size. The lowest MAE for downward facing periods was achieved by the crowd consisting of 71 participants and the highest MAE was displayed by the group of three participants with a MAE of 65.7 and 69.9 respectively. For upward periods the MAE is also decreasing with group size, however the differences between the crowd size of nine participants and the one with 15 are only minor with a MAE of 69.8 and 69.9 on average, leading to a slightly higher percentage of individuals beaten with 67.9% compared to 67.6%

Overall, results also show that the magnitude of the increase between the errors of the initial trend and the trend after the first break or change rose with crowd size.

Table 34: Performance of small crowds after the first break/change (crowd size: 3)

Time Series	Number of structural breaks	Trend after 1st break/change	MAE Initial trend	MAE after 1st break/change	% beat by crowd Initial trend	% beat by crowd after 1st break/change
Crowd size (nr. of participants)			3	3	3	3
Time Series 1	1	0	27.6	116.4	32.4	71.8
Time Series 4	2	0	27.0	89.8	36.6	56.3
Time Series 5	1	0	18.1	61.0	40.9	62.0
Time Series 7	0	0	29.8	31.7	39.4	60.6
Time Series 10	1	0	27.1	86.5	38.0	60.6
Time Series 11	2	0	18.5	33.8	46.5	62.0
Time Series 2	2	1	13.2	119.0	46.5	63.4
Time Series 3	1	1	22.0	100.8	54.9	62.0
Time Series 6	1	1	20.3	99.6	49.3	56.3
Time Series 8	2	1	21.0	42.3	36.6	71.8
Time Series 9	0	1	18.3	28.3	40.9	64.8
Time Series 12	1	1	23.5	43.1	29.6	66.2

0 = Down; 1 = UP

Ø MAE for downward trends after 1st break/change:	69.9
Ø MAE for upward trends after 1st break/change:	72.2
Ø percentage beat for downward trends after 1st break/change:	62.2%
Ø percentage beat for upward trends after 1st break:	64.1%
Ø increase in MAE compared to initial trend (DOWN):	183.3%
Ø increase in MAE compared to initial trend (UP):	298.3%

TS 1 & TS 3 contained noise

Table 36: Performance of small crowds after the first break/change (crowd size: 5)

Time Series	Number of structural breaks	Trend after 1st break/change	MAE Initial trend crowd	MAE after 1st break/change crowd	% beat by crowd Initial trend	% beat by crowd after 1st break/change
Crowd size (nr. of participants)			5	5	5	5
Time Series 1	1	0	27.0	116.9	35.2	71.8
Time Series 4	2	0	26.7	91.5	36.6	54.9
Time Series 5	1	0	15.4	60.9	45.1	62.0
Time Series 7	0	0	29.2	26.4	39.4	70.4
Time Series 10	1	0	26.0	86.1	39.4	60.6
Time Series 11	2	0	16.4	30.1	46.5	62.0
Time Series 2	2	1	11.7	117.7	54.9	63.4
Time Series 3	1	1	19.8	101.7	60.6	62.0
Time Series 6	1	1	18.6	101.2	56.3	53.5
Time Series 8	2	1	21.6	38.0	36.6	81.7
Time Series 9	0	1	15.3	25.4	50.7	71.8
Time Series 12	1	1	22.7	41.6	31.0	66.2

0 = Down; 1 = UP

Ø MAE for downward trends after 1st break/change:	68.7
Ø MAE for upward trends after 1st break/change:	70.9
Ø percentage beat for downward trends after 1st break/change:	63.6%
Ø percentage beat for upward trends after 1st break:	66.4%
Ø increase in MAE compared to initial trend (DOWN):	196.4%
Ø increase in MAE compared to initial trend (UP):	331.5%

TS 1 & TS 3 contained noise

Table 37: Performance of small crowds after the first break/change (crowd size: 9)

Time Series	Number of structural breaks	Trend after 1st break/change	MAE Initial trend crowd	MAE after 1st break/change crowd	% beat by crowd Initial trend	% beat by crowd after 1st break/change
Crowd size (nr. of participants)			9	9	9	9
Time Series 1	1	0	27.5	116.6	33.8	71.8
Time Series 4	2	0	27.0	90.7	36.6	54.9
Time Series 5	1	0	13.8	60.6	50.7	62.0
Time Series 7	0	0	28.9	23.1	40.9	73.2
Time Series 10	1	0	26.6	86.1	38.0	60.6
Time Series 11	2	0	14.6	28.2	49.3	67.6
Time Series 2	2	1	9.6	118.2	64.8	63.4
Time Series 3	1	1	18.6	101.4	60.6	62.0
Time Series 6	1	1	15.9	100.2	57.8	54.9
Time Series 8	2	1	21.2	36.2	36.6	84.5
Time Series 9	0	1	13.0	22.3	54.9	76.1
Time Series 12	1	1	22.2	40.6	31.0	66.2

0 = Down; 1 = UP

Ø MAE for downward trends after 1st break/change:	67.6
Ø MAE for upward trends after 1st break/change:	69.8
Ø percentage beat for downward trends after 1st break/change:	65.0%
Ø percentage beat for upward trends after 1st break:	67.9%
Ø increase in MAE compared to initial trend (DOWN):	199.3%
Ø increase in MAE compared to initial trend (UP):	388.6%

TS 1 & TS 3 contained noise

Table 38: Performance of small crowds after the first break/change (crowd size: 15)

Time Series	Number of structural breaks	Trend after 1st break/change	MAE Initial trend crowd	MAE after 1st break/change crowd	% beat by crowd Initial trend	% beat by crowd after 1st break/change
Crowd size (nr. of participants)			15	15	15	15
Time Series 1	1	0	27.3	116.6	35.2	71.8
Time Series 4	2	0	26.7	90.6	36.6	54.9
Time Series 5	1	0	12.5	60.5	50.1	62.0
Time Series 7	0	0	28.6	20.9	40.9	77.5
Time Series 10	1	0	26.1	86.4	38.0	60.6
Time Series 11	2	0	13.5	26.3	52.1	71.8
Time Series 2	2	1	8.4	118.9	67.6	63.4
Time Series 3	1	1	17.5	101.3	64.8	62.0
Time Series 6	1	1	15.1	100.6	57.8	53.5
Time Series 8	2	1	20.3	35.5	36.6	84.5
Time Series 9	0	1	11.8	21.5	60.6	76.1
Time Series 12	1	1	22.4	41.3	31.0	66.2

0 = Down; 1 = UP

Ø MAE for downward trends after 1st break/change:	66.9
Ø MAE for upward trends after 1st break/change:	69.9
Ø percentage beat for downward trends after 1st break/change:	66.4%
Ø percentage beat for upward trends after 1st break:	67.6%
Ø increase in MAE compared to initial trend (DOWN):	208.2%
Ø increase in MAE compared to initial trend (UP):	433.7 %

TS 1 & TS 3 contained noise

6.2.4 Second break

Similar results to the ones of the whole crowed presented in Chapter 6.1 were achieved by the smaller crowds. Also, for all crowd sizes (9 to 15) the percentage beat by the crowd decreased for all four time series. Furthermore, the absolute errors in the second break were smaller compared to the first break, except for Time Series 2, in which the direction of the second break differed from the direction of the 1st break, errors increased whereas in all other cases the errors decreased compared to the first structural break. This could also be observed for crowd consisting of the 71 participants (see chapter 6.1.4).

The absolute errors for three and five participants overall were equal, only for Time Series 2 the crowd consisting of five individuals could reach lower absolute errors. The crowd of 15 participants displayed higher errors for each structural break compared to the crowd of nine participants. On average the crowd of 15 individuals reached the same level of MAE (53.6) than the crowd of five individuals. The average errors of the nine randomly selected participants were as low as for the whole crowd of 71 participants. Overall, the differences in the MAE were so small that the percentage of individuals beat by the crowd stayed constant over the different crowd sizes.

Table 19 to Table 22 shows the performance of the different crowd sizes for the second structural breaks.

Table 39: performance of small crowds for the 2nd break (crowd size: 3)

Time Series	Number of breaks	Direction 1st break/change	Direction 2nd break	Absolute error 1st break	Absolute error 2nd break	% beat by crowd 1st break	% beat by crowd 2nd break
Crowd size (nr. of participants)				3	3	3	3
Time Series 2	2	1	0	89.7	99.3	71.8	70.4
Time Series 4	2	0	0	80.6	34.3	62.0	38.0
Time Series 8	2	0	0	72.8	42.5	74.6	64.8
Time Series 11	2	1	1	60.4	40.0	59.2	54.9
Ø					54.0		

Table 42: Performance of small crowds for the 2nd break (crowd size: 5)

Time Series	Number of breaks	Direction 1st break/change	Direction 2nd break	Absolute error 1st break	Absolute error 2nd break	% beat by crowd 1st break	% beat by crowd 2nd break
Crowd size (nr. of participants)				5	5	5	5
Time Series 2	2	1	0	89.6	97.7	71.8	70.4
Time Series 4	2	0	0	77.3	34.3	73.2	38.0
Time Series 8	2	0	0	70.3	42.5	77.5	64.8
Time Series 11	2	1	1	59.5	40.0	59.2	54.9
Ø					53.6		

Table 45: Performance of small crowds for the 2nd break (crowd size: 9)

Time Series	Number of breaks	Direction 1st break/change	Direction 2nd break	Absolute error 1st break	Absolute error 2nd break	% beat by crowd 1st break	% beat by crowd 2nd break
Crowd size (nr. of participants)				9	9	9	9
Time Series 2	2	1	0	89.3	97.9	71.8	70.4
Time Series 4	2	0	0	79.2	33.9	62.0	38.0
Time Series 8	2	0	0	72.4	41.4	74.6	64.8
Time Series 11	2	1	1	60.5	40.3	59.2	54.9
Ø					53.4		

Table 48: Performance of small crowds for the 2nd break (crowd size: 15)

Time Series	Number of breaks	Direction 1st break/change	Direction 2nd break	Absolute error 1st break	Absolute error 2nd break	% beat by crowd 1st break	% beat by crowd 2nd break
Crowd size (nr. of participants)				15	15	15	15
Time Series 2	2	1	0	89.1	98.4	71.8	70.4
Time Series 4	2	0	0	79.1	34.0	62.0	38.0
Time Series 8	2	0	0	71.8	41.6	77.5	64.8
Time Series 11	2	1	1	59.9	40.3	59.2	54.9
Ø					53.6		

6.2.5 Period after the second break

As previously described for the whole crowd of 71 participants in Chapter 6.1.5, the MAE increased in two of the four time series (Time Series 8 and Time Series 11) compared to the trend after the first structural break. For Time Series 2 and Time Series 4 the MSE declined. This could also be observed for all other crowd sizes. Furthermore, the performance of the crowd measured by individuals with a higher MAE, decreased in the period after the second break – in three out of four time series the percentage of individuals beaten by the crowd declined and was constantly below 50% for Time Series 4.

On average, the crowd consisting of five participants was able to achieve the lowest MAE over the second structural breaks with 62.7, closely followed by the whole crowd with 62.9 and the crowd of nine participants with a MAE of 63.1. Five participants and the whole crowd were able to beat the same percentage of individuals, namely 56.0%.

Three and fifteen Randomly selected participants performed equally with a MAE of 64.1 on average.

Results are summarized in the Tables 23 to 27.

Table 51: Performance of the whole crowd after the 2nd break (crowd size: 71)

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				71	71	71	71
Time Series 2	2	1	0	118.5	86.4	63.4	66.2
Time Series 4	2	0	0	90.5	59.4	54.9	46.5
Time Series 8	2	1	0	35.4	58.8	84.5	59.2
Time Series 11	2	0	1	20.2	47.0	73.2	52.1
Ø					62.9		56.0

Table 53: Performance of small crowds after the 2nd break (crowd size: 3)

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				3	3	3	3
Time Series 2	2	1	0	119.0	87.0	63.4	64.8
Time Series 4	2	0	0	89.8	61.0	56.3	43.7
Time Series 8	2	1	0	42.3	60.4	71.8	59.2
Time Series 11	2	0	1	33.8	47.8	62.0	52.1
Ø					64.1		55.0

Table 55: Performance of the small crowds after the 2nd break (crowd size: 5)

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				5	5	5	5
Time Series 2	2	1	0	117.7	85.6	63.4	66.2
Time Series 4	2	0	0	91.5	59.6	54.9	46.5
Time Series 8	2	1	0	38.0	58.3	81.7	59.2
Time Series 11	2	0	1	30.1	47.4	62.0	52.1
Ø					62.7		56.0

Table 57: Performance of the small crowds after the 2nd break (crowd size: 9)

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				9	9	9	9
Time Series 2	2	1	0	118.2	86.6	63.4	64.8
Time Series 4	2	0	0	90.7	59.8	54.9	46.5
Time Series 8	2	1	0	36.2	59.1	84.5	59.2
Time Series 11	2	0	1	28.2	46.7	67.6	52.1
Ø					63.1		55.7

Table 59: Performance of the small crowds after the 2nd break (crowd size: 15)

Time Series	Number of breaks	Direction after 1st break/change	Direction after 2nd break	MAE after 1st break crowd	MAE after 2nd break crowd	% beat by crowd after 1st break	% beat by crowd after 2nd break
Crowd size (nr. of part.)				15	15	15	15
Time Series 2	2	1	0	118.9	87.0	63.4	64.8
Time Series 4	2	0	0	90.6	61.0	54.9	43.7
Time Series 8	2	1	0	35.5	60.4	84.5	59.2
Time Series 11	2	1	1	26.3	47.8	71.8	52.1
Ø					64.1		55.0

Finally, the major results for the crowd consisting of all 71 participants can be summarized as:

- **Overall, the crowd outperformed the individuals for all 12 time series.**

  The percentage of individuals beaten was above 59% for all 12 time series

- **The crowd struggled in predicting the right direction and magnitude of structural breaks**

  Several structural breaks were anticipated correctly by the crowd - however, it damped the magnitude of the shifts and often forecasted the opposite direction.

- **The crowd was worse in forecasting upward trends and upward structural breaks**

  The forecasting performance of the crowd was worse for upward initial trends (periods without a structural break or a smooth change in trend), displaying higher errors than for upward initial trends. Additionally, the percentage beaten by the crowd stayed below 40% for upward initial trends over all crowd sizes and time series. Generally, forecasting errors were also lower for downward structural breaks than for upward structural breaks.

- **Crowd performance was highest for time series containing two structural breaks**

- **Errors were higher in periods after a structural break, compared to periods with no structural break happening prior**

- **If the second structural break followed the direction of the first, crowd performance diminished**

- **Individuals were able to outperform the crowd for upward initial trends.**

  For upward initial trends the percentage of individuals beaten by the crowd was consistently below 50%

The major results for smaller crowds (3,5,9,15 participants) are:

- **Generally, the MAE decreased if the crowd size increased.**

  Exceptions were Time Series 1 with a constant MAE across smaller crowds ranging from three to nine participants and Time Series 4 with a slight increase of the MAE with crowd size.

- **For specific time series and trends, smaller crowds were able to perform equivalent to or better than the whole crowd**

  For different section in some of the 12 time series, smaller crowds were able to achieve a slightly lower MAE than the whole crowd or the same level of error. For example, in the initial trend of Time Series 1 the crowd consisting of five participants was able to achieve a lower MAE than the whole crowd. For Time Series 4 the group with three participants also had a slightly lower MAE than the whole crowd. Additionally, the crowd of 71 participants was outperformed by the crowd with 15 participants for Time Series 12.

- **Minor differences in the MAE over different group sizes did not impact the percentage of individuals beaten by the crowd**

  Often the changes in the MAE for the different crowds were so small that the percentages of individuals beaten by the crowd displayed no change. For example, Time Series 1 with five participants showed no change in the percentage of individuals beaten compared to 71 participants.

- **Downward trends consistently had lower errors compared to upward trends across smaller crowds and the whole crowd**

- **Structural breaks were equally or better forecasted by smaller crowds**

  Regarding the first structural break of the time series, the crowd of five participants achieved a lower MAE for downward structural breaks than the whole crowd. Furthermore, also if the first structural break was upward, smaller crowds (3, 5 and 9 participants) performed slightly better than the whole crowd (80.4 vs. 80.7). However, these minor differences lead to similar percentages of individuals beaten over the different crowd sizes. Looking at the second structural breaks, the group of nine participants achieved the same MAE than the whole crowd.

7. Discussion & Conclusion

Several papers have investigated the effectiveness of the wisdom of crowd concept in different settings - ranging from probability estimates, ordering problems and geopolitical forecasts, forecasting soccer results and solving more complex problems like the Euclidean traveling salesperson problem (Fiechter & Kornell, 2021; Peeters, 2018; Yi et al. 2012).

This master thesis extends existing wisdom of crowds' literature by showing that the forecasting of time series subject to abrupt changes (structural breaks), is another effective application of the wisdom of crowd concept. By determining the median as well as the mean absolute error of 71 participants over 12 time series, findings reveal that the crowd was able to beat the individual estimates in every time series, with a percentage of at least 59 percent of the individuals displaying a higher MAE than the crowd. This exceeds the previously selected threshold of 50 percent for this experiment in order for a crowd to be considered as wise.

These results align with Becker et al. (2009) who found that averaging forecasts leads to higher forecasting accuracy and in their experiment to a lower MAE than 90 percent of the participants (109 out of 120).

Despite outperforming induvial forecasts**,** the crowd in general struggled in predicting the right direction as well as the magnitude of structural breaks. For several breaks, the crowd forecasted the opposite direction. Additionally, their forecasts were less extreme than the true values. This observed damping of

trends and break points aligns with previous research on individual forecasts conducted by Tumen & Boduroglu's (2022), O'Connor et al. (1997) and Becker et al. (2009)

While O'Connor et al. (1993) observed that the errors of individuals were higher for time series with a downward direction than for time series with an upward direction, the opposite could be observed in this experiment. Crowds were worse in forecasting trends with an upward direction as well as upward structural breaks. Consistent with O'Connor et al. (1993), results also showed that the errors were higher in the periods after a structural break compared to the initial period without a break or change happening, again highlighting the difficulty of adapting to abrupt changes.

Regarding the analysis of smaller crowds (3,5,9 and 15 participants) the findings align with Walter et al. (2022), by showing that crowd performance increases with crowd size. In particular, the MAE decreased with increasing number of participants with only minor exceptions in two out of 12 time series presented (Time Series 1 and Time series 4). However, it is also notable that overall these differences in the MAE over the crowd sizes were minor and did not affect the percentage of individuals beaten by the crowd in several cases.

As outlined in Chapter 2.1., Galesic et al. (2018) point out that smaller crowds can outperform lager ones especially in settings with surprising outcomes such as abrupt shifts in financial trends. This observation was partially supported in the present study, as smaller crowds reached an equal or even lower MAE on structural breaks than the crowd consisting of all 71 participants.

In summary, the presented findings suggest that relying on the wisdom of crowd effect is the preferable method over using individual estimates for forecasting time series in general and in particular for those subject to structural breaks.

Consequently, the results also have practical relevance for areas in which forecasting still relies on (individual) human judgment (e.g product, sales forecasting). Also, organizations in volatile markets (e.g financial markets, tourism or oil sector) could benefit from approved forecasting accuracy by integrating crowd-based forecasting techniques instead of relying on individual predictions.

8. Limitations & Future Research

Overall, the presented experiment and results were based on 71 participants without collecting information regarding their expertise on forecasting.

Previous research like Simmons et al. (2010) and Davis-Stober et al. (2014) presented in Chapter 2.1 suggests that crowd wisdom is influenced by the skill of the participants. Future studies could analyze if weighting forecasts based on the participants' prior forecasting experience improves overall crowd performance especially for time series with structural breaks. Even though the findings are based on a relatively small sample size of 71 participants, it is arguable if a greater more diverse crowd would have reached a higher performance. Walter et al. (2022) point out that crowd performance increases with its size, however, this effect continuously decreases for larger sizes.

To improve forecasting accuracy and to dampen possible negative effects of typical biases like noise modelling, anchoring or trend damping (see Chapter 3.2), methods such as feedback and task decomposition (presented in Chapter 3.3) were used for the experiment. Participants received feedback by presenting them the right direction after forecasting each period. Instead of asking them to forecast one time series at once, the forecasting task itself was divided into several smaller periods.

Despite those mechanism, no form of incentive for low forecasting errors was provided to the participants. A monetary incentive for accurate forecasts could have also been a tool in order to reduce overall forecasting error and improve crowd performance.

While the results of the experiment showed the effectiveness of the wisdom of crowd over individual forecasting performance, newer techniques of forecasting recently emerged with the use of machine learning and deep neural networks, which have also been proven useful for structural break forecasting (Kaushik et al., 2021; Marcondes & Marçal, 2023). Especially in light with the general difficulty for the crowd to correctly predict structural breaks shown in this experiment, future research could explore creating models that combine those two approaches in order to significantly improve structural break forecasting.

Particularly recent studies suggest that the future in forecasting lies in the combination of statistical and machine learning generated forecasts and the incorporation of human judgment (Makridakis et al., 2020a; Petropoulos et al., 2022; Kourentzes et al., 2018) which highlights the necessity of further research in this area by evaluating new approaches or adapting existing ones.

9. Literature

Afflerbach, P., Van Dun, C., Gimpel, H., Parak, D., & Seyfried, J. (2021). A Simulation-Based Approach to Understanding the Wisdom of Crowds Phenomenon in Aggregating Expert Judgment. *Business & Information Systems Engineering*, *63*(4), 329–348. https://doi.org/10.1007/s12599-020-00664-x

Alvarado-Valencia, J. A., & Barrero, L. H. (2014a). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, *36*, 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Alvarado-Valencia, J. A., & Barrero, L. H. (2014b). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior*, *36*, 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Andreou, E., & Ghysels, E. (2009). Structural Breaks in Financial Time Series. In T. Mikosch, J.-P. Kreiß, R. A. Davis, & T. G. Andersen (Hrsg.), *Handbook of Financial Time Series* (S. 839–870). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71297-8_37

Aruga, K. (2016). The U.S. shale gas revolution and its effect on international gas markets. *Journal of Unconventional Oil and Gas Resources*, *14*, 1–5. https://doi.org/10.1016/j.juogr.2015.11.002

Bai, J., & Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica*, *66*(1), 47–78. https://doi.org/10.2307/2998540

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1–22. https://doi.org/10.1002/jae.659

Becker, O., Leitner, J., & Leopold-Wildburger, U. (2009). Expectation formation and regime switches. *Experimental Economics*, *12*(3), 350–364. https://doi.org/10.1007/s10683-009-9213-0

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for Testing the Constancy of Regression Relationships Over Time. *Journal of the Royal Statistical Society: Series B (Methodological)*, *37*(2), 149–163. https://doi.org/10.1111/j.2517-6161.1975.tb01532.x

Budescu, D. V., & Chen, E. (2015). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*, *61*(2), 267–280. https://doi.org/10.1287/mnsc.2014.1909

Caporale, G. M., Gil-Alana, L. A., & Poza, C. (2020). Persistence, non-linearities and structural breaks in European stock market indices. *The Quarterly Review of Economics and Finance*, *77*, 50–61. https://doi.org/10.1016/j.qref.2020.01.007

Charles, A., Darné, O., & Hoarau, J.-F. (2019). How resilient is La Réunion in terms of international tourism attractiveness: An assessment from unit root tests with structural breaks from 1981-2015. *Applied Economics*, *51*(24), 2639–2653. https://doi.org/10.1080/00036846.2018.1558349

Cró, S., & Martins, A. M. (2017). Structural breaks in international tourism demand: Are they caused by crises or disasters? *Tourism Management*, *63*, 3–9. https://doi.org/10.1016/j.tourman.2017.05.009

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. https://doi.org/10.1037/dec0000004

Durbach, I., & Montibeller, G. (2018). Predicting in shock: On the impact of negative, extreme, rare, and short lived events on judgmental forecasts. *EURO Journal on Decision Processes*, *6*(1–2), 213–233. https://doi.org/10.1007/s40070-017-0063-2

El-Shagi, M., & Giesen, S. (2013). Testing for Structural Breaks at Unknown Time: A Steeplechase. *Computational Economics*, *41*(1), 101–123. https://doi.org/10.1007/s10614-011-9271-1

Emmanuel, O., & Maureen, T. (2022). *CHOW TEST FOR STRUCTURAL BREAK: A CONSIDERATION OF GOVERNMENT TRANSITION IN NIGERIA FORM MILITARY TO CIVILIAN DEMOCRATIC GOVERNMENT*. *1*.

Epley, N., & Gilovich, T. (2006). The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fan, Y., & Xu, J.-H. (2011). What has driven oil prices since 2000? A structural change perspective. *Energy Economics*, *33*(6), 1082–1094. https://doi.org/10.1016/j.eneco.2011.05.017

Ferguson, J. M., Taper, M. L., Zenil-Ferguson, R., Jasieniuk, M., & Maxwell, B. D. (2019). Incorporating Parameter Estimability Into Model Selection. *Frontiers in Ecology and Evolution*, *7*, 427. https://doi.org/10.3389/fevo.2019.00427

Fiechter, J. L., & Kornell, N. (2021). How the wisdom of crowds, and of the crowd within, are affected by expertise. *Cognitive Research: Principles and Implications*, *6*(1), 5. https://doi.org/10.1186/s41235-021-00273-6

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, *5*(1), 1–15. https://doi.org/10.1037/dec0000059

Goodwin, P. (2002a). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, *30*(2), 127–135. https://doi.org/10.1016/S0305-0483(01)00062-7

Goodwin, P. (2002b). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega*, *30*(2), 127–135. https://doi.org/10.1016/S0305-0483(01)00062-7

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*(2), 147–161. https://doi.org/10.1016/0169-2070(93)90001-4

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, *13*(1), 5–24. https://doi.org/10.1080/13546780600872502

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgemental forecasting. *International Journal of Forecasting*, *12*(1), 119–137. https://doi.org/10.1016/0169-2070(95)00634-6

Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean Squared Error, Deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12), e2021MS002681. https://doi.org/10.1029/2021MS002681

Hong, H., Ye, Q., Du, Q., Wang, G. A., & Fan, W. (2020). Crowd characteristics and crowd wisdom: Evidence from an online investment community. *Journal of the Association for Information Science and Technology*, *71*(4), 423–435. https://doi.org/10.1002/asi.24255

Jose, V. R. R., Grushka-Cockayne, Y., & Lichtendahl, K. C. (2014). Trimmed Opinion Pools and the Crowd's Calibration Problem. *Management Science*, *60*(2), 463–475. https://doi.org/10.1287/mnsc.2013.1781

Jose, V. R. R., & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*(1), 163–169. https://doi.org/10.1016/j.ijforecast.2007.06.001

Kalsie, A., & Arora, A. (2019). Structural break, US financial crisis and macroeconomic time series: Evidence from BRICS economies. *Transnational Corporations Review*, *11*(3), 250–264. https://doi.org/10.1080/19186444.2018.1475087

Kao, A. B., & Couzin, I. D. (2014). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133305. https://doi.org/10.1098/rspb.2013.3305

Kaushik, R., Jain, S., Jain, S., & Dash, T. (2021). Performance evaluation of deep neural networks for

forecasting time-series with multiple structural breaks and high volatility. *CAAI Transactions on Intelligence Technology*, *6*(3), 265–280. https://doi.org/10.1049/cit2.12002

Kim, S.-E., & Seok, J. H. (2018). Tourism and Economic Growth in Korea: Focusing on the Structural Break. *Korea International Trade Research Institute*, *14*(4), 177–188. https://doi.org/10.16980/jitc.14.4.201808.177

Kriegeskorte, N. (2015). Crossvalidation. In *Brain Mapping* (S. 635–639). Elsevier. https://doi.org/10.1016/B978-0-12-397025-1.00344-4

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, *22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

Lawrence, M., & O'Connor, M. (1992). Exploring judgemental forecasting. *International Journal of Forecasting*, *8*(1), 15–26. https://doi.org/10.1016/0169-2070(92)90004-S

Lawrence, M., & O'connor, M. (1995). The anchor and adjustment heuristic in time-series forecasting. *Journal of Forecasting*, *14*(5), 443–451. https://doi.org/10.1002/for.3980140504

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*(3), 377–390. https://doi.org/10.1016/j.ijforecast.2007.02.006

Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – A review of the literature. *European Journal of Operational Research*, *213*(3), 459–469. https://doi.org/10.1016/j.ejor.2011.01.006

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

MacGregor, D. G., & Armstrong, J. S. (o. J.). *Judgmental Decomposition: When Does It Work?*

Maheu, J. M., & Gordon, S. (2008). Learning, forecasting and structural breaks. *Journal of Applied Econometrics*, *23*(5), 553–583. https://doi.org/10.1002/jae.1018

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153. https://doi.org/10.1002/for.3980010202

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 competition: Background, organization, and implementation. *Special Issue: M5 competition*, *38*(4), 1325–1336. https://doi.org/10.1016/j.ijforecast.2021.07.007

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

Marcondes Pinbto, J., & Marçal, E. F. (2023). Reinforcement Learning: An Artificial Intelligence Approach to Forecasting when There Are Structural Breaks. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4358446

Meligkotsidou, L., & Vrontos, I. D. (2008). Detecting structural breaks and identifying risk factors in hedge fund returns: A Bayesian approach. *Journal of Banking & Finance*, *32*(11), 2471–2481. https://doi.org/10.1016/j.jbankfin.2008.05.007

Mensi, W., Hammoudeh, S., & Yoon, S.-M. (2014). How do OPEC news and structural breaks impact returns and volatility in crude oil markets? Further evidence from a long memory process. *Energy Economics*, *42*, 343–354. https://doi.org/10.1016/j.eneco.2013.11.005

Mérida, A., & Golpe, A. A. (2016). Tourism-led Growth Revisited for Spain: Causality, Business Cycles and Structural Breaks. *International Journal of Tourism Research*, *18*(1), 39–51. https://doi.org/10.1002/jtr.2031

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, *9*(2), 163–172. https://doi.org/10.1016/0169-2070(93)90002-5

O'Connor, M., Remus, W., & Griggs, K. (1997). Going Up-Going Down: How Good Are People at Forecasting Trends and Changes in Trends? *Journal of Forecasting*, *16*(3), 165–176. https://doi.org/10.1002/(SICI)1099-131X(199705)16:3<165::AID-FOR653>3.0.CO;2-Y

Otto, S., & Breitung, J. (2023). Backward CUSUM for Testing and Monitoring Structural Change with

an Application to COVID-19 Pandemic Data. *Econometric Theory*, *39*(4), 659–692. https://doi.org/10.1017/S0266466622000159

Palm, F. C., & Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting*, *11*(8), 687–701. https://doi.org/10.1002/for.3980110806

Pan Pang, K., & Ting, K. M. (2004). Improving the Centered CUSUMS Statistic for Structural Break Detection in Time Series. In G. I. Webb & X. Yu (Hrsg.), *AI 2004: Advances in Artificial Intelligence* (Bd. 3339, S. 402–413). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30549-1_36

Pauwels, L. L., Chan, F., & Mancini Griffoli, T. (2012). Testing for Structural Change in Heterogeneous Panels with an Application to the Euro's Trade Effect. *Journal of Time Series Econometrics*, *4*(2). https://doi.org/10.1515/1941-1928.1141

Peeters, T. (2018). Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, *34*(1), 17–29. https://doi.org/10.1016/j.ijforecast.2017.08.002

Pesaran, M. H. (2023). *Forecasting Time Series Subject to Multiple Structural Breaks*.

Pesaran, M. H., & Timmermann, A. (2004). How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting*, *20*(3), 411–425. https://doi.org/10.1016/S0169-2070(03)00068-2

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Salisu, A. A., & Fasanya, I. O. (2013). Modelling oil price volatility with structural breaks. *Energy Policy*, *52*, 554–562. https://doi.org/10.1016/j.enpol.2012.10.003

Sanders, N. R. (1992). Accuracy of Judgmental Forecasts: A Comparison. *Omega, 20*(3), 511-522.

Sanders, N. R., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, *5*(1), 39–52. https://doi.org/10.1002/bdm.3960050106

Sanders, N. R. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega*, *25*(2), 135–144. https://doi.org/10.1016/S0305-0483(96)00057-6

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, *31*(6), 511–522. https://doi.org/10.1016/j.omega.2003.08.007

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. W. (2010). Intuitive Biases in Choice vs. Estimation: Implications for the Wisdom of Crowds. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1553935

Sjöberg, L. (2009). Are all crowds equally wise? A comparison of political election forecasts by experts and the public. *Journal of Forecasting*, *28*(1), 1–18. https://doi.org/10.1002/for.1083

Sohil, F., Sohali, M. U., & Shabbir, J. (2022). An introduction to statistical learning with applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, *6*(1), 87–87. https://doi.org/10.1080/24754269.2021.1980261

Sroginis, A., Fildes, R., & Kourentzes, N. (2023). Use of contextual and model-based information in adjusting promotional forecasts. *European Journal of Operational Research*, *307*(3), 1177–1191. https://doi.org/10.1016/j.ejor.2022.10.005

Theocharis, Z., Smith, L. A., & Harvey, N. (2019). The influence of graphical format on judgmental forecasting accuracy: Lines versus points. *FUTURES & FORESIGHT SCIENCE*, *1*(1), e7. https://doi.org/10.1002/ffo2.7

Tumen, C., & Boduroglu, A. (o. J.). *Judgmental Time Series Forecasting: A systematic analysis of graph format and trend type*.

Turner, P. (2010). Power properties of the CUSUM and CUSUMSQ tests for parameter instability. *Applied Economics Letters*, *17*(11), 1049–1053. https://doi.org/10.1080/00036840902817474

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases*. *185*.

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518–1547. https://doi.org/10.1016/j.ijforecast.2022.11.005

Webby, R., & O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, *12*(1), 91–118. https://doi.org/10.1016/0169-2070(95)00644-3

Webby, R., O'Connor, M., & Lawrence, M. (2001). Judgmental Time-Series Forecasting Using Domain Knowledge. In J. S. Armstrong (Hrsg.), *Principles of Forecasting* (Bd. 30, S. 389–403). Springer US. https://doi.org/10.1007/978-0-306-47630-3_17

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The Wisdom of the Crowd in Combinatorial Problems. *Cognitive Science*, *36*(3), 452–470. https://doi.org/10.1111/j.1551-6709.2011.01223.x

Zeileis, A., Kleiber, C., Krämer, W., & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, *44*(1–2), 109–123. https://doi.org/10.1016/S0167-9473(03)00030-6

Zeileis, A. (2005). A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals. Econometric Reviews, 24(4), 445–466. https://doi.org/10.1080/07474930500406053

List of Tables

List of Figures

Appendix 1: Python Code

This program was used to analyze forecasting accuracy of the crowd for the 12 time series. The true values and forecasted values are loaded from separate csv-file. The periods as well as the segments of the respective time series are identified to extract the true values and the forecasts for each time series. For the analysis of forecast accuracy, the program calculates the Mean Absolute Error (MAE).

The MAE is calculated for the entire crowd (all 71 participants) as well as for randomly selected participants based on a specified group size. This group size can changed within the code.

For each subgroup, forecasts are averaged, and compared to true values in order to compute the MAE of the subgroup. In order to guarantee robustness this process is performed multiple times, according to the number of iterations set. The program also determines the percentage of the participants whose individual forecasts were less accurate than forecasts made by the crowd or subgroup.

```
import pandas as pd
```

```
import numpy as np
from sklearn.metrics import mean_absolute_error

# Loading the forecasting data stored in a csv file
df_forecast = pd.read_csv('Forecasts_total.csv', delimiter=';')

# Load true values stored in a seperate csv file
true_values_df = pd.read_csv('truevalues.csv', delimiter=';')

# Allows the set the periods which should be analized
analysis_periods = range(11, 26)

# Function which detects the right segment for a given period
def detect_segment_for_period(time_series_id, period):
    for seg in ['Seg1', 'Seg2', 'Seg3']:
        col_name = f"True_TS{time_series_id}_{seg}: {period}"
        if col_name in true_values_df['Period'].values:
            return seg
    raise ValueError(f"Period {period} not found in any segment for Time Series ID {time_series_id}.")

# Function which extracts the true values for the available periods
def get_true_values_for_periods(time_series_id):
    available_periods =
true_values_df[true_values_df['Period'].str.contains(f"True_TS{time_series_id}_Seg")]
    if available_periods.empty:
        raise KeyError(f"No true values found for Time Series ID {time_series_id}.")

    filtered_periods = available_periods['Period'].str.extract(r": (\d+)$").astype(int)
    filtered_periods = filtered_periods[filtered_periods[0].isin(analysis_periods)].values.flatten()

    true_values = available_periods[available_periods['Period'].str.endswith(tuple(map(str,
filtered_periods)))]
    return true_values['value'].tolist(), filtered_periods.tolist()

# Function which determines the forecast columns dynamically
```

```
def get_forecast_columns_for_periods(time_series_id, filtered_periods):
    forecast_columns = [
        col for col in df_forecast.columns
        if col.startswith(f"Forecast_TS{time_series_id}") and
        int(col.split(":")[1]) in filtered_periods
    ]
    return forecast_columns

# Function which aligns the arrays for calculations
def align_arrays(array1, array2):
    min_length = min(len(array1), len(array2))
    if len(array1) != len(array2):
        print(f"Warning: Arrays aligned to minimum length {min_length}.")
    return array1[:min_length], array2[:min_length]

# Function which calculates the MAE
def calculate_mae_via_absolute_errors(df, forecast_columns, true_values):
    forecasts = df[forecast_columns].values

    if len(forecasts) > 0:
        mean_forecasts = np.mean(forecasts, axis=0)

            # Aligns the true_values and the forecasts to the same length by calling the created
            function above
        true_values, mean_forecasts = align_arrays(true_values, mean_forecasts)

        absolute_errors = np.abs(mean_forecasts - true_values)
        mae = np.mean(absolute_errors)
        return mae
    else:
        return None

# Function which calculate the Median Absolute Error
def calculate_median_via_absolute_errors(df, forecast_columns, true_values):
    forecasts = df[forecast_columns].values
```

```
    if len(forecasts) > 0:
        median_forecasts = np.median(forecasts, axis=0)

        # Align true_values and forecasts to the same length
        true_values, median_forecasts = align_arrays(true_values, median_forecasts)

        absolute_errors = np.abs(median_forecasts - true_values)
        median_error = np.median(absolute_errors)
        return median_error
    else:
        return None

# Function to calculate the percentage of participants beaten based on MAE
def calculate_percentage_beaten(df, forecast_columns, true_values, reference_error):
    higher_error_count = 0
    total_participants = len(df)
    for _, row in df.iterrows():
        forecast = row[forecast_columns].values
        forecast, true_values_aligned = align_arrays(forecast, true_values)
        individual_absolute_errors = np.abs(forecast - true_values_aligned)
        individual_error = np.mean(individual_absolute_errors)
        if individual_error > reference_error:
            higher_error_count += 1
    percentage_beaten = (higher_error_count / total_participants) * 100
    return higher_error_count, percentage_beaten

# Parameters which can be set fort he analysis
group_sizes = [3, 5, 9, 15]  # List of group sizes to analyze
n_iterations = 1000  # Number of Monte Carlo iterations
time_series_ids = range(1, 13)  # Assuming time series IDs range from 1 to 12
total_participants = len(df_forecast['ID'].unique())

# DataFrame which stores the results
results = []
```

```
# This for loop performs the analysis for each time series
for ts_id in time_series_ids:
    try:
        true_values, filtered_periods = get_true_values_for_periods(ts_id)
        if not filtered_periods:
            print(f"No valid periods for Time Series ID {ts_id}. Skipping.")
            continue
        forecast_columns = get_forecast_columns_for_periods(ts_id, filtered_periods)

        df_forecast_filtered = df_forecast.dropna(subset=forecast_columns)
        if df_forecast_filtered.empty:
            print(f"No forecast data available for Time Series ID {ts_id}. Skipping.")
            continue

        # MAE
        crowd_mae = calculate_mae_via_absolute_errors(df_forecast_filtered, forecast_columns,
true_values)

        # Median Absolute Error
        crowd_median = calculate_median_via_absolute_errors(df_forecast_filtered, forecast_columns,
true_values)

        # percentage of participants beaten by the crowd (MAE)
        crowd_count, percentage_beaten_by_crowd = calculate_percentage_beaten(
            df_forecast_filtered, forecast_columns, true_values, crowd_mae
        )

        # Add overall crowd results (MAE and Median)in the output
        results.append({
            'Group Size': 'All',
            'Time Series ID': ts_id,
            'Metric': 'MAE',
            'Error (Crowd)': round(crowd_mae, 2) if crowd_mae is not None else "N/A",
            'Individuals Beaten by Crowd (Count)': crowd_count,
```

```
            'Individuals Beaten by Crowd (%)': f"{percentage_beaten_by_crowd:.2f}%",
        })
        results.append({
            'Group Size': 'All',
            'Time Series ID': ts_id,
            'Metric': 'Median',
            'Error (Crowd)': round(crowd_median, 2) if crowd_median is not None else "N/A",
            'Individuals Beaten by Crowd (Count)': "N/A",
            'Individuals Beaten by Crowd (%)': "N/A",
        })

        # Subgroup Analysis
        for group_size in group_sizes:
            group_maes = []
            group_medians = []
            for _ in range(n_iterations):
                best_group = np.random.choice(df_forecast_filtered['ID'].unique(), size=group_size,
replace=False)
                df_group = df_forecast_filtered[df_forecast_filtered['ID'].isin(best_group)]

                group_mae = calculate_mae_via_absolute_errors(df_group, forecast_columns, true_values)
                if group_mae is not None:
                    group_maes.append(group_mae)

                group_median = calculate_median_via_absolute_errors(df_group, forecast_columns,
true_values)
                if group_median is not None:
                    group_medians.append(group_median)

            if group_maes and group_medians:
                avg_group_mae = np.mean(group_maes)
                std_dev_group_mae = np.std(group_maes)
                avg_group_median = np.mean(group_medians)

                results.append({
```

```
                'Group Size': group_size,
                'Time Series ID': ts_id,
                'Metric': 'MAE',
                'Error (Crowd)': round(avg_group_mae, 2),
                'Std Dev': round(std_dev_group_mae, 2),
            })
            results.append({
                'Group Size': group_size,
                'Time Series ID': ts_id,
                'Metric': 'Median',
                'Error (Crowd)': round(avg_group_median, 2),
                'Std Dev': "N/A",
            })

    except Exception as e:
        print(f"Error for Time Series ID {ts_id}: {e}")

# Saves the results to a CSV file
results_df = pd.DataFrame(results)

# Sorts the results by Time Series ID, Metric, and Group Size for clarity
if not results_df.empty:
    results_df = results_df.sort_values(by=['Time Series ID', 'Metric', 'Group Size'])
    results_df.to_csv('16_Analysis_With_MAE_and_Median.csv', index=False)
    print("Analysis complete. Results saved as 'Analysis_With_MAE_and_Median.csv'.")
else:
    print("No results to save.")
```