The Ability of the Surprisingly Popular Method to predict Football Games

Mattia Mantovani, 467792

Erasmus University – MSc Behavioural Economics

Supervisor: Tong Wang

Second Assessor: Han Bleichrodt

August 11th, 2021

Abstract

In this thesis, our purpose is to demonstrate that the 'Surprisingly Popular Algorithm' (SPA) has the ability to foresee a future predictive setting and that it is the most accurate crowdsourcing method as compared to two other methods known as the democratic and the confidence-weighted method. The SPA method is based on peoples' evaluation of other people's judgements, thus relying on the metacognition of the crowd to combine them for making accurate collective decisions. The experimental design is created via six weekly surveys sent out to our participants who need to predict the winners of 80 Italian Football League (Serie A) games as well as state their confidence and the percentage of other participants who would agree with them. Our results show that the SPA can be applicable in situations involving European Football predictions. Over the 80 games, the SPA predicted more games and proved to have significantly higher average correct predictions per matchday as compared to the other two methods. Future research can focus on improving the accuracy of the experimental design with a larger number of games as well as with the same participants across all surveys.

1. Introduction

Crowdsourcing is an important phenomenon in economics. It requires the collection of work, opinions, and information from a group of people. These opinions are often compared to individual judgements with the goal to seek which is the most accurate in knowing facts or predicting events. In fact, as the wisdom of crowd concept states, when more judgements are collected and aggregated it often does lead to more accurate decisions and predictions as compared to individual judgements.

There are several crowdsourcing methods that are used by companies as well as individuals. One of the many advantages it presents to companies, is the fact that it allows different people with different skills to come up with new ideas and test these new ideas on customers to learn about active feedback, gaining a lot of learnings and ultimately saving time and money. For individuals, on the other hand, it might help broaden their thinking view, thus improving individuals' judgements in the future.

The 'surprisingly popular algorithm' (SPA) is one of the most recent developed crowdsourcing methods, established by Prelec, Seung & McCoy (2017). It is calculated in a way to obtain the correct answer to a single question, even when the majority believe in the wrong answer. The SPA has shown encouraging results, however there is still a limited amount of literature on the strategy, which brings us to test whether it can be used as a strong and accurate measure for making decisions, explaining behaviors, as well as eventually making predictions on future events, such as in the sports field.

The main focus for this thesis is applied to the sports field and in particular, football. Our purpose is thus to test whether the SPA is the most accurate crowdsourcing method when predicting football games. The methods used apart from the SPA are the democratic and the confidence-weighted approach. The performance of these are evaluated using the outcome of the last 80 football games of the 2020/21 season from the Italian league Serie A. Past literature has made clear that the aggregation of judgements is more effective than those of individuals for both factual judgements, where the answer is already known, as well as with predicting future events, such as American

Football, Basketball, and US elections. The particular focus of this thesis is hence whether the SPA predicts more European football games as compared to the democratic and the confidence weighted approaches.

This is of social relevance due to the fact that it can provide additional insights to the topic of predictive judgements, by revealing differences in the process of decision-making as well as expanding the view of the SPA method. It may also help broaden the thinking view, thus improving individual's judgements in the future. Finally, the results can broaden the concept's application in different sports, as for now only Basketball and American Football have been the main focus.

We pose the following research question:

Does the 'surprisingly popular algorithm' (SPA) yield more accurate football predictions compared to the democratic and the confidence-weighted method?

The paper is organized as follows. In the second section, the literature review is presented. In the third and fourth section we present the methodology and data. In the fifth and sixth section, we will discuss the results and implications. Finally, the last section will provide concluding remarks.

2. Literature Review

This chapter provides an overview of the literature related to the crowdsourcing methods that are used in the analysis. It opens with a theoretical perspective of the wisdom of crowds as well as its modern applications and its development. Next, it includes a section where the focus is solely on the SPA method by exploring previous studies that test the ability of the SPA. Furthermore, it explains the other crowdsourcing methods, and finally, the chapter concludes with a focus on the methods of conducting the research.

2.1 Wisdom of Crowds

When it comes to problem-solving, decision-making, innovating, and forecasting, the wisdom of crowd theory assumes that a large number of people can produce better results than single experts (Navajas, Niella, Garbulsky, Bahrami & Sigman, 2018). Various studies have been published over

the years to popularize the idea of the wisdom of crowds. In 2005, James Surowiecki, a New York based writer, published a book called 'The Wisdom of Crowds', which was the first to emphasize the concept of crowd wisdom (Surowiecki, 2005). For crowds to be wise, the author recommended that individuals' properties have a variety of viewpoints. In fact, there should be a mutual reliance on each other's views and no bullying from others (Surowiecki, 2005). Moreover, all the individuals from the crowd should form their own beliefs and opinions based on their personal experience. This also means that depending on individuals expertise and interests, the wisdom of crowds can be applicable in many fields, going from biology to psychology and culture, to name a few.

Surowiecki's concept of the wisdom of crowds emerged from Aristotle's theory, who presented the idea of collective judgements (Surowiecki, 2005). As an example, Aristotle explains how a dinner is more enjoyable when a group of individuals come together as they can all provide different views to what only one individual can provide (Navajas et al., 2018).

Needless to say, not all collective judgements from crowds are wiser than individual judgements. In the 1990's for instance, it was observed how a group of frenetic investors participated in the stock market bubble created from the dot-com industry. These groups of individual investors using speculation bought and held internet start-ups to help these companies grow and become more profitable in the future (Navajas et al., 2018). While they were holding these, major tech firms started mass selling some of the stocks, leading to the bubble bursting, panic in the market, and these groups of individual investors to lose most of their investments.

The concept of wisdom of crowds has various number of applications in many fields, as mentioned previously. It has put a lot of development into several sets of skills such as cognition, cooperation, and coordination. In cognition, modern applications have put a lot of growth into all kinds of predictions and not only in the sports field, including forecasting stock price volatilities and election results (Hosio, Goncalves, Anagnostopoulos & Kostakos, 2016; Arrow et al., 2008). Additionally, several companies apply the idea of the wisdom of crowds in determining ideas and/or feedback from the public about their products. Netflix for instance uses crowd intelligence for large-

scale digital ratings of their products online, thus meaning that it is currently being implemented to help experts gain better ideas. This can also be applied to cooperation, as it is important for organizations as well as individuals to build an authentic community and use this to come up with unique opportunities to create a network with a potential to grow. An example is the intelligent community, which was promoted to encourage people to join and manage a community with the goal of economic development (Coe, Paquet & Roy, 2001). Finally, the wisdom of crowds also showed growth in coordination, encouraging the community to develop and collaborate knowledge. Wikipedia for instance, organizes information by collaborative sharing and self-organization by users, allowing everyone to participate and share their own knowledge (Aaltonen & Seiler, 2016).

2.2 Wisdom-of-Crowds Algorithms

'Surprisingly Popular Algorithm'

The first method, which is the one of most interest, is the 'Surprisingly Popular Algorithm' (SPA). It is a recently developed crowdsourcing technique that considers a crowd's expert minority opinion in a way that those of minority opinion can also influence the final collective choice (Prelec et al., 2017). This is because the majority is not always right. Thus, when people know that they are correct about something in the minority, they mostly know that the majority doesn't know it. The SPA can therefore integrate this by not only asking their judgement, but by additionally asking them their judgements on how much percent of participants know the correct answer (Rutchick, Ross, Calvillo & Mesick, 2020). According to the authors, the SPA is hence based on participants' evaluation of other participants judgments, relying on metacognition of the crowd. A choice by a participant for instance, is unexpectedly common when its average metacognitive judgments are lower than the actual chosen frequency.

It mostly has been used to examine factual judgements, such as the capital of a country, where the answer is already known. In this case, the participants are asked a binary choice with the wrong and right answer as well as the percentage of other participants choosing the right answer. Past literature has also tested the SPA accuracy regarding predictions with future events, for example

American Football and Basketball match winners. Also in this case it is the same format, with a binary choice by selecting the winner.

The SPA was founded by Prelec et al. (2017), who examined the ability of the method regarding factual judgements where the answers are already known. They test the US state capitals as well as other general knowledge questions and find that the SPA was effective in providing accurate crowdsourcing predictions, hence concluding that the SPA showed promising results. They additionally conclude that the method is still limited and that it is yet to be understood if it can predict situations where something is unknown.

Following Prelec et al. (2017) study, Lee, Danileiko & Vi (2018) tested the ability of the SPA to predict future events. They tested whether the method was able to predict the winners of all the 2017-2018 US National League (NFL) games as compared to the original crowdsourcing methods. They found that the effectiveness of the SPA outperforms all other crowdsourcing methods that they have tested for. The number of games predicted by the methods used by the authors is shown in Figure 1. These are the Elo method, the confidence-weighted method, the democratic mode method, media experts, and the SPA. The labelled lines indicate the number of games successfully predicted by the approaches.

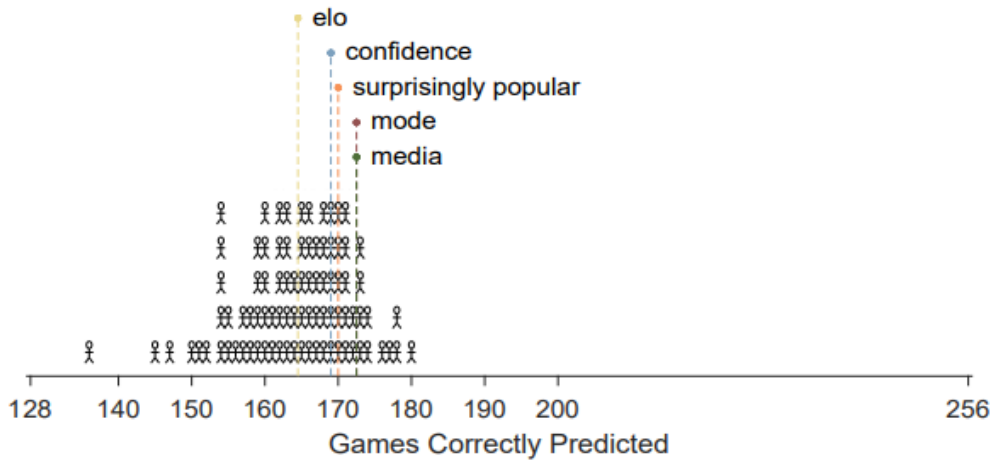

Figure 1: Number of games predicted by the crowdsourcing methods and the media experts from Lee et al. (2018, Figure 3).

Besides this, Lee et al. (2018) test whether there are significant differences between the self-assessed participants. They found that the least knowledgeable participants of American Football predicted 150 games out of 256, whilst the most knowledgeable predicted between 155 and 175 games. In specific, the participants who rated themselves as *'Extremely Knowledgeable'* predicted 170 games in total. This performance outperformed the least knowledgeable participants as well as 69 media experts, and was only inferior to 19 media experts and similar to other six. The authors thus not only show that the SPA outperforms the majority of predictions, but also show that the participants who are self-assessed experts outperform the least knowledgeable ones. They even do better than experts, hence showing that the SPA method can be a widely used algorithm to forecast NFL games, especially when using self-assessed experts.

A different study by Rutchick et al. (2020) also tests the ability of the SPA method to provide accurate crowdsourcing predictions. They focus on three studies regarding future outcomes of American Football games, Basketball games, and midterm US elections. In their first one, they used samples without self-assessed experts who rate themselves most knowledgeable and found no evidence that the SPA is superior to the other crowdsourcing methods, which are the democratic mode method, the confidence-weighted method, and media experts. In their following two studies, they use self-assessed experts, and test whether the SPA method is particularly effective with aggregate judgements of these experts, whose knowledge should be superior. They find that the SPA was most effective when predictions were made by the most objectively self-assessed experts. This result is therefore consistent with the previous study conducted by Lee et al. (2018) as well as other studies who test the power of statistical groups using the crowdsourcing methods, such as Mannes, Soll & Larrick (2014). The conclusions are also similar to the previously stated papers, that is, the SPA is most effective when the assessments of the most knowledgeable are taken into account, hence asserting that the method, in general, outperforms individual judgements.

Past literature has proven that the SPA is a successful and promising method in multiple occasions (Cui et al., 2019). Some authors however criticized the technique, stating that it is limited

and that it has only been effective on occasions where the correct answer has already been established. Thus, ideally there is a need to find out under what conditions the SPA is most effective (Rutchick et al., 2020).

Democratic Approach (Modal judgement)

The democratic crowdsourcing approach is a straightforward method and relies on a group of individuals voting on a specific topic to conclude (McCallister, 2016). This method must be performed with equality and fairness to obtain accurate predictions. This means that the participants need to all be given the same information as well as answer the question in the same circumstances. The final decision must rely on the choice that the participants choose most often, that is the majority. This works best when there are larger groups, as obtaining 100% approval on issues with this kind of sample size is always more difficult (McCallister, 2016).

The typical crowd predictions are done using the democratic vote procedure. However, there are several limitations when using this strategy as it is biased for rather superficial information at the expense of knowledge not widely shared, which is why different algorithms have been tested for crowdsourcing predictions (Prelec et al., 2017). Nonetheless, Lee et al. (2018) and Rutchick et al. (2020) find similar crowdsourcing predictions with this method compared to the SPA.

Confidence-Weighted Approach

The confidence-weighted crowdsourcing approach is similar to the democratic approach. In this case, the decisions of participants are multiplied by the degree of confidence they have in the choice and divided by the total confidence, so it acts as a weight. Thus, these choices and ratings of confidence are combined to determine the optimal overall crowd decision, meaning that individual confidence ratings play a significant role and can improve the real-group performance. In the case of football predictions, this approach hence yields more appropriate predictions when participants are defined as self-assessed experts. Once again, as the democratic method, this approach also finds similar and accurate crowdsourcing predictions (Lee et al., 2018; Rutchick et al., 2020).

2.3 Algorithms vs. Individual Judgements

Over the ages, algorithms have been consistently performing well in various task predictions. Most people, however, often mistrust their advice. Efendić, Van de Calseyde & Evans (2020) study the purpose to determine the effect on individuals' trust from predictions by algorithms under varied time differences; if they trusted after short while or after a long time prediction. They conducted seven studies and found that algorithm predictions that were slowly generated were seen to be less accurate. Therefore, individuals are unwilling to trust algorithms that take a long time, whilst they are willing to trust the faster generated predictions. Football predictions using algorithms also replicate the same idea (Efendić et al., 2020). It is hence essential to note that when using algorithms for predictions, one should consider time to impact the results significantly as well as use self-assessed experts so that the projections are likely to be more effective (Rutchick et al., 2020).

From the above literature, the wisdom of crowd theory proves that aggregation of judgements is more effective in decision making than those of individuals, which has been made clear on many occasions, such as in NFL predictions, US. Elections, and Basketball predictions. Particularly, other crowdsourcing methods have limitations as they cannot make accurate predictions where the majority is wrong, whilst the SPA yielded more effective even in these cases. This thesis will hence focus solely on European football to determine whether the SPA method can predict accurately and better than the democratic and the confidence-weighted approaches.

3. Methodology & Data

3.1 Experimental Design

To test whether the SPA method can predict European football games accurately and better than the democratic and the confidence weighted method, a methodological design is developed. This approach consists in weekly surveys that asks participants to predict several football games. With the data of the surveys, the final choice of the SPA, the democratic, and the confidence-weighted approach are calculated.

The football games that will be used in the weekly surveys represent only one tournament. That is the Italian Football League, known as the Serie A. There are 20 teams in the league, meaning that there are ten games on each matchday that are usually played over the weekend. Due to time constraints and already being more than halfway done throughout the Serie A league of 2020-2021, there are only eight matchdays left, which represent a total of 80 games. There are two cases where there will be two matchdays in one week, meaning that there will be ten games played over the week and ten over the weekend. Thus, two of the surveys include 20 games to be predicted by the participants. In total, this means that we will hand out six weekly surveys with a total of 8 matchdays and 80 games to be predicted.

The survey is created on Qualtrics with several characteristics that need to comply to make it as efficient as possible. It starts off with a small introduction that adds what it is about as well as who to contact if the participants have any questions. Before participants give their predictions, we also ask them three simple demographic questions. These are the gender, age, and how much knowledge they rate themselves regarding Italian football. This is done before the predictions so that if some subjects quit in the middle of the survey, it is still possible to use their data to run the whole analysis. After the demographics, we ask participants for each of the ten games to predict the winner, rate their confidence, and how much percent of other participants share their choice. This is needed to calculate the final choice for each of the three crowdsourcing methods. Finally, the order of the games is randomized in order to have a comparable crowd size across all games even if some subjects quit the survey in the middle.

The structure of the survey is shown in the Appendix. It is presented in the same way for each of the six surveys to facilitate the process for subjects. As the demographics are asked in each survey, this also facilitates the recruitment process since subjects can differ each week. Additionally, there are two options to predict the winner as all current evidence on the SPA is with binary choices; one being the home team wins and the other being a combined option of a draw or the away team to win. The reasoning behind this is because a draw is a positive result for the away team, thus meaning that

they play for this result as well.

The recruiting process is divided into two main target groups that are engaged on social media accounts and groups. The first target group involves Italian friends as well as other friends from the many years living in Milan as well as other countries. In this case, those that are football lovers, meaning they watch the games and/or are updated about the Italian football league, are recruited using social media platforms. WhatsApp, Instagram, and Facebook are the main examples, with WhatsApp being the most utilized and most important platform. The second target group involves social media groups. These involve several groups on WhatsApp, Facebook, and Reddit with the sole purpose of discussing the Serie A. These groups on WhatsApp and Facebook, are various Italian fantasy football league groups of maximum 12 people that entertain discussions on the league, as already mentioned. Additionally, the group on Reddit has around 15000 people interested in the league and discussing about it: Reddit.com/r/seriea/. The survey is always sent with a standard message from myself that is meant to recruit the right audience, ergo, football lovers. Overall, we will hence be relying on only one incentive: interest in the topic.

Finally, benchmark predictions are also collected for all the 80 games using the following source: fivethirtyeight.com. It gives club soccer predictions based on an algorithm on current and historical data results for all teams, and hence does not rely on human judgements. This provides a state-of-the-art benchmark as compared to the other three crowdsourcing methods, which are human judgement benchmarks.

3.2 Method

The analysis is done in several sections with the ultimate goal to have enough proof to answer the research question. First, we show illustrative examples regarding the SPA method and how we calculated some of the predictions. We then analyze the overall performance of the SPA compared to the other crowdsourcing methods as well as the fivethirtyeight algorithm. This is done via a histogram that shows the total number of games predicted out of 80 by all methods for the whole sample as well as for only the self-assessed experts. We include two panels for only the self-assessed experts to see

whether there is a difference in overall performance when taking only them into account. Finally, we further show the number of games predicted in each matchday by the methods as well as a the statistical significance of the findings.

3.3 Descriptive Statistics

Following the data collection, the responses were extracted to an Excel spreadsheet and cleaned accordingly. In Table 1 below, the descriptive statistics for the demographic questions are shown. These are presented in frequencies for each of the categorical variable's gender, age, and the knowledge the participants rated themselves as.

Table 1

Frequency Statistics of the Categorical Variables Gender, Age, and Knowledge for each survey.

Categorical Variables	**Survey 1**	**Survey 2**	**Survey 3**	**Survey 4**	**Survey 5**	**Survey 6**
Observations (N)	83	49	53	41	29	37
Gender						
Male	73	39	47	36	26	33
Female	9	7	4	3	3	4
Non Binary/Third Gender	0	3	1	0	0	0
Prefer not to say	1	0	1	2	0	0
Age (years)						
18-24	66	37	39	28	19	24
25-34	6	2	5	1	1	2
35+ years	11	10	9	12	9	11
Knowledge						
Extremely knowledgeable	28	15	11	8	11	13
Very knowledgeable	24	17	18	17	7	12
Moderately knowledgeable	18	8	15	10	7	6
Slightly knowledgeable	8	7	6	4	3	6
Not knowledgeable at all	5	2	3	2	1	0

In the table above, the number of observations are shown for every survey. These include all responses that answered at least one of the predictions, meaning that also partial responses were recorded. Only for participants who have not answered any of the predictions and only the demographic questions were excluded from Table 1 and hence the analysis. We see that for all surveys, the highest distribution of participants are males from ages 18-24, who rate themselves as extremely or very knowledgeable. This shows that the recruitment process was done accurately, targeting mainly football lovers who self-assess themselves as experts.

4. Results

The section will begin with the presentation of two examples of predictions of the SPA method. This is done to show how those may diverge from the prediction provided by other methods such as the democratic and the confidence-weighted one, as well as how the SPA method may, at times, fail. Then, we go on to present the distribution of correct answers, out of the 80 Serie A football games considered, by our survey participants. This is presented for the full sample, then only for the subset who self-rated themselves as *Extremely Knowledgeable,* and finally for those who rated themselves as *Extremely* and *Very Knowledgeable* together. Moreover, we show how the collective judgements compare to the benchmark provided by the state-of-the-art website, fivethirtyeight. Table 3 will summarise the number of correctly predicted games by the three methods and the benchmark, broken down by matchday (from 1 to 8). This is done for the total representation of the performance of the SPA method as compared to the others, and to show that the method is comparable to that of the fivethirtyeight predictions and at most times better than the other crowdsourcing methods. Finally, Table 4 checks the statistical significance of our findings.

4.1 Two Examples of Predictions of the SPA method

The two games considered to illustrate how the predictions of the SPA method can diverge from those of the democratic and confidence-weighted methods are shown in Table 2.

Table 2

Predictions of the different methods for Cagliari-Parma and Sassuolo-Lazio

Cagliari 4 - Parma 3			Sassuolo 2 - Lazio 0		
Method	**Metric**	**Prediction**	**Method**	**Metric**	**Prediction**
Democratic(Mode)	50.65%	Cagliari	Democratic(Mode)	68%	Lazio
Confidence-Weighted	68.83%	Cagliari	Confidence-Weighted	54.96%	Lazio
Surprisingly Popular	56.68%	Parma	Surprisingly Popular	69.72%	Sassuolo

The metrics are shown in terms of the home teams, Cagliari and Sassuolo. For the democratic method, the percentage of people who picked the two teams; for the confidence-weighted method, the confidence weight; and finally the SPA method based on the metacognitive judgement of respondents.

Regarding Cagliari-Parma, the respondents' judgements were particularly split as only 50.65% believed Cagliari was going to win. However, since having a democratic majority is enough for the democratic method, Cagliari is the predicted winner of the game. With respect to the confidence-weighted method, after weighting respondents' choices by the reported confidence in their choice, Cagliari resulted as the team with the higher confidence score with 68.83%. For the SPA method, the expected agreement on Cagliari (56.68%) is larger than the percentage of people who actually picked Cagliari (50.65%), implying that many of the people who selected Parma, believed very few would agree with them, whereas Cagliari backers were more confident of the agreeability of their choice. This resulted in predicting Parma as the winner of the game, which did not turn out to be the case.

Regarding Sassuolo-Lazio, the vast majority of respondents selected Lazio as the winner of the game with 68% favouring the Roman team. Even after re-weighting the answers by the reported confidence respondents had in their answers, Lazio emerged as the predicted winner. Nonetheless, even if by a very thin margin (1.72%), this game provides an interesting example of the SPA method giving the correct answer when the other methods failed to do so. In fact, the SPA score of 69.72% for the Lazio victory, being higher than the number of people who actually picked Lazio, elects Sassuolo as the predicted winner for the SPA method, which turned out to be the case with a score of

2-0. Similarly to the previous case, this is a reflection of the fact that those who selected Lazio were convinced that most would agree with them, whereas those who backed Sassuolo were aware of the fact that their choice would turn out to be less popular.

4.2 Distributions of Correct Answers

In Figure 2 there are three panels that show the overall distributions of the correct predictions for the 80 games by the crowdsourcing methods and the fivethirtyeight benchmark. Panel A shows the full sample of respondents, with the SPA method performing as second-best totalling 57/80 correct game predictions. Only the fivethirtyeight benchmark forecasted better with 58 (red vertical line) correct predictions. In Panel A, both the democratic and the confidence-weighted method correctly predicted only 55 games, less than the SPA method.

Moving on to Panel B, this one includes only those who self-rate themselves as *Extremely Knowledgeable.* The results are slightly more surprising. Both the democratic and the confidence-weighted method only predicted 51 out of 80 games correctly. The SPA on the other hand performed as well as the fivethirtyeight algorithm, something that does not hold for the full sample.

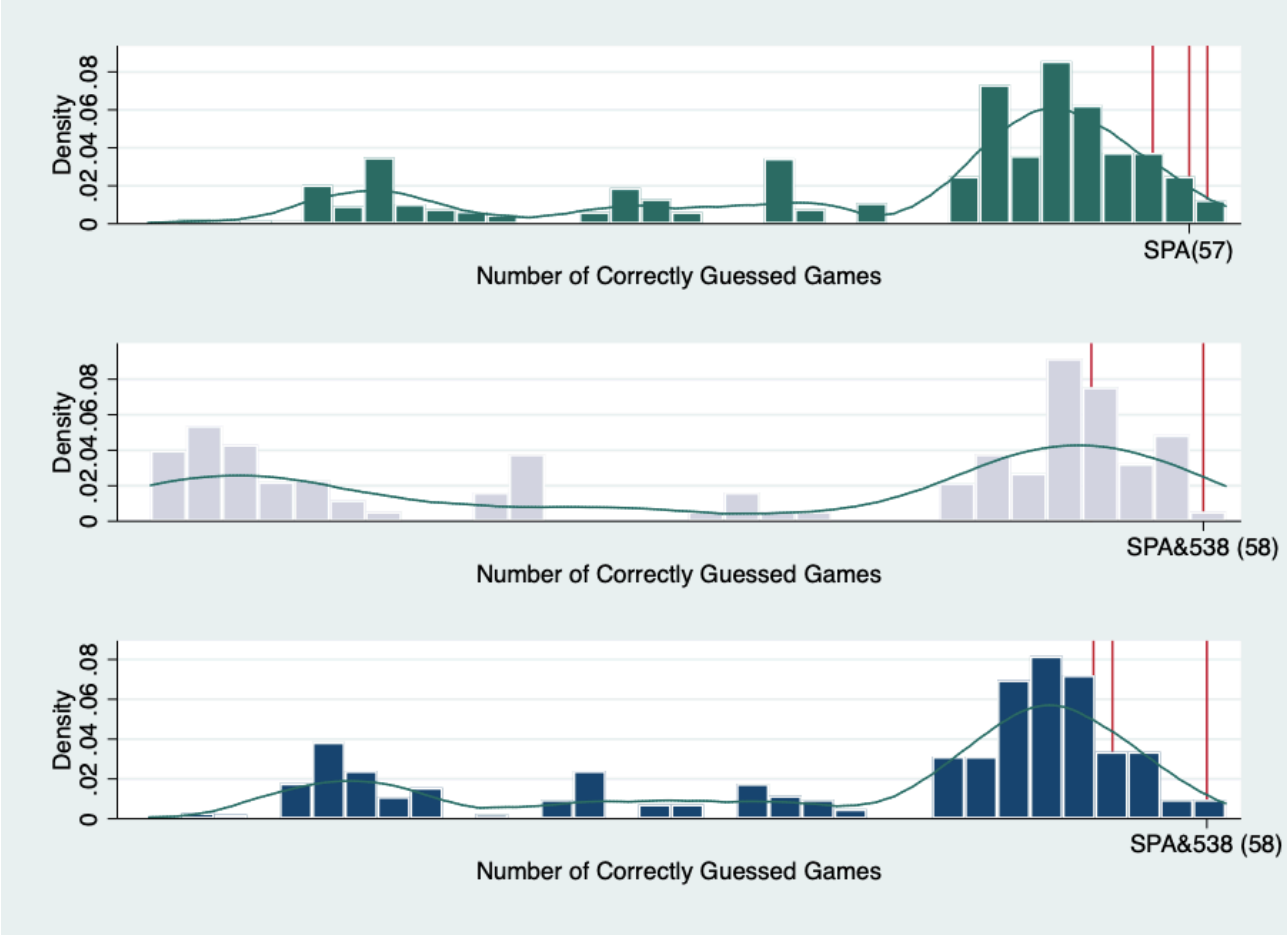

Figure 2 – The figure shows the overall distributions of correct predictions out of 80 games by the crowdsourcing methods and the 538 benchmark. It is divided into 3 panels: the first one starting from the top displays the correct predictions by the crowdsourcing methods and the 538 benchmark using the full sample of participants, the second one uses only those who self-rate themselves as extremely knowledgeable, and the third one those who self-rate as either extremely or very knowledgeable. In each of the panels the number of correct predictions out of 80 by the SPA method is highlighted, with the other vertical lines representing the two remaining methods or the 538 benchmark.

The disappointing performance of the democratic and confidence-weighted method, which are obviously highly correlated, for the extremely knowledgeable, may be attributed to the fact that this subset of respondents may be more sophisticated than the full sample of respondents. This means that they may be more prone to select unpopular answers, which reflects in poor performance for the

first two methods as often the underdog teams end up losing. It does however improve the performance of the SPA method as it includes people's metacognitive judgements in its predictions.

Panel C includes the participants who self-rated themselves as *Very Knowledgeable* and *Extremely Knowledgeable*. Here, the performance of the democratic and confidence-weighted method slightly improves, moving from 51 to 52 and from 51 to 53 respectively. The SPA method instead, matches the performance of the 538 website, with both predicting 58 games out of the 80 games.

4.3 Correct Predictions by Matchday

Table 3

Breakdown of correct forecasts by matchday for each of the methods as well as the 538 benchmark

Serie A 2020/2021					
Matchday 1			Matchday 2		
Method	Correct	Accuracy	Method	Correct	Accuracy
Democratic (Mode)	7	70%	Democratic (Mode)	6	60%
Confidence-Weighted	7	70%	Confidence-Weighted	7	70%
Surprisingly Popular	7	70%	Surprisingly Popular	7	70%
Benchmark(*538*)	8	80%	Benchmark(*538*)	7	70%
Matchday 3			Matchday 4		
Method	Correct	Accuracy	Method	Correct	Accuracy
Democratic (Mode)	7	70%	Democratic (Mode)	8	80%
Confidence-Weighted	7	70%	Confidence-Weighted	8	80%
Surprisingly Popular	7	70%	Surprisingly Popular	8	80%
Benchmark(*538*)	7	70%	Benchmark(*538*)	9	90%
Matchday 5			Matchday 6		
Method	Correct	Accuracy	Method	Correct	Accuracy
Democratic (Mode)	8	80%	Democratic (Mode)	7	70%
Confidence-Weighted	7	70%	Confidence-Weighted	7	70%
Surprisingly Popular	8	80%	Surprisingly Popular	7	70%
Benchmark(*538*)	7	70%	Benchmark(*538*)	7	70%
Matchday 7			Matchday 8		
Method	Correct	Accuracy	Method	Correct	Accuracy
Democratic (Mode)	5	50%	Democratic (Mode)	7	70%
Confidence-Weighted	5	50%	Confidence-Weighted	7	70%
Surprisingly Popular	5	50%	Surprisingly Popular	8	80%
Benchmark(*538*)	5	50%	Benchmark(*538*)	8	80%

The Table above displays the number of games correctly predicted as well as the accuracy over the total of ten games per matchday. They are broken down by the crowdsourcing methods, democratic, confidence-weighted, and SPA, alongside the benchmark provided by the 538 website.

For the sake of clarity, it is important to mention that *Accuracy* in this setting is measured as the percentage of correctly guessed games out of the total of ten games per matchday. From the Table it is interesting to notice that in three occasions, that is matchdays 3, 6, and 7, all the methods display the same level of accuracy. In particular, matchday 7 is of interest because it is the one where all the methods predicted the least games. This is probably attributable to the fact that towards the end of the season the teams face different incentives, resulting in games to become more unpredictable. In two occasions, matchday 1 and 4, the benchmark 538 outperformed all the other competing methods, especially for matchday 4 where the algorithm was able to correctly predict nine out of ten games. Regarding matchday 8, the SPA method and 538 website jointly outperformed the two other competing methods with eight against seven correct predictions. Finally, in one occasion only (matchday 5), the SPA method outperformed the 538 benchmark with eight against seven correct predictions.

It is important to note that, in general, the number of correct predictions doesn't vary substantially across the methods, with mostly variations of one correct game. However, it is important to corroborate the hypothesis of this thesis that the SPA methods consistently ranks as either first- or second-best prediction method.

Table 4

Differences in average predictions per matchday between SPA and other methods

Differences in Mean Predicted Games			
Method	**Difference**	**t-stat**	**p-value**
Democratic(Mode)	-0.625	-2.06	0.043
Confidence-Weighted	-0.583	-1.92	0.058
Benchmark (*538*)	0.042	0.14	0.891
F-stat	2.83		

Table 4 above is a robustness check to formally check the statistical significance of our findings. We performed this by pulling together in a dataset the number of correctly predicted games by the full sample, the ones that rated themselves as '*Extremely Knowledgeable*', and the ones that rated themselves as '*Extremely or Very Knowledgeable*'. Then, a categorical variable was created

with the number of games predicted correctly by each method, and finally we compared using a linear regression how the average of the correctly predicted games by each of the eight matchdays differed across the methods with the SPA. We note that the average correct predictions are slightly lower for the democratic and the confidence-weighted methods as compared to the SPA method. The t statistic for the democratic method is significant to the 5% level, thus indicating that there are significant differences in average correct predictions. For the confidence-weighted method, the t statistic is slightly lower, yet still significant and suggesting that the SPA outperforms the method on average. Regarding the benchmark, the average is slightly higher for the fivethirtyeight algorithm, however not significant, indicating that the average correct predictions are roughly similar. The overall f statistic is 2.83, thus significant to the 5% level, hence meaning that there are differences between all methods put together and the SPA method. The results are thus in line with the previous findings found from the histogram as it concludes that the SPA does outperform the democratic and the confidence-weighted methods, while it is roughly the same as the fivethirtyeight predictions.

5. Discussion

Our results show that with regards to the human judgements, the SPA method outperformed the democratic method as well as the confidence-weighted method over the total of eighty football games. Additionally, in comparison to the state-of-the-art benchmark, fivethirtyeight, the SPA had similar overall predictions. Both of these findings suggest promising outcomes for applying the SPA method to European Football predictions and hence confirm what Lee et al. (2018) and Rutchick et al. (2020) found. We also confirm the effectiveness of the most knowledgeable participants, who outperformed the least knowledgeable ones with regards to the SPA method. This can be noted from Figure 2, where the number of games correctly predicted by the SPA method increases when using the sample of only self-assessed experts, hence meaning that those were more accurate. In fact, from Panel B and Panel C, when only including the experts, the correct predictions were equal to the ones from fivethirtyeight. These findings hence assert that the SPA can outperform the majority of

predictions, whether they are human individual judgements, collective judgements, or state-of-the-art judgements. This suggests that it can be a widely used algorithm to forecast football games, especially when using self-assessed experts.

Despite these results, there are several limitations when evaluating the crowdsourcing methods based on only eighty games from one season. With a total of 760 games in one season, it's clear that the purpose of this thesis can be viewed as a motivating demonstration of the ability and applicability of the SPA method when making future predictions. Moreover, the survey presented limitations that could have likely caused distortions in the results section with regards to the tables and the histogram. The fact that the participants self-rate their knowledge may cause problems and inaccuracies regarding the overall predictions of the games. In fact, self-rating can distort the true level of expertise and most likely caused the predictions of only the extremely and very knowledgeable participants to give lower correct predictions as compared to the full sample with regards to the democratic and confidence-weighted methods. Besides this, the participants of the surveys differed from week to week, leading to the three panels of Figure 2 being imprecise.

While there is room for improvement in accuracy, this thesis shows that people are able to provide accurate predictions in a future predictive setting and not only with questions were the answer is already known. This means that people are also able to provide accurate meta-cognitive judgments, with the SPA method capturing the people's knowledge who have insights to a surprise winner for the individual games. It is especially the case for the experts, who are probably more aware of surprise winners and thus predict more games. Therefore, if the accuracy with regards to the experimental design improves, the SPA method can be used as a real-world benchmark.

6. Conclusion

The SPA method is based on the peoples' evaluation of other people's judgements, thus relying on the metacognition of the crowd to combine them to make accurate group decisions. In this thesis, the ability of the SPA was tested using weekly participants that predicted in total eighty

football games. The method was compared to two other human collective crowdsourcing methods, democratic and confidence-weighted method, as well as to a state-of-the-art benchmark, fivethirtyeight. The results were proven to be promising in forecasting a future predictive setting, as football games, and hence suggests that the SPA can be applicable in many situations involving future predictions. Over the eighty games with the full sample, the SPA was able to predict more games as compared to the other two methods, and one game less as compared to the benchmark. When using only the self-assessed experts from the sample, the SPA was as accurate as the state-of-the-art benchmark, and again more accurate than the crowdsourcing methods. These findings were proven when formally checking the statistical significance of the methods, where it was noted that the average correct predictions per matchday were significantly higher for the SPA method as compared to the crowdsourcing methods, while slightly lower, but not statistically significant, with regards to the fivethirtyeight predictions. We thus confirm that the SPA method outperformed the other methods.

Future research can focus on reducing the limitations to provide more accurate overall predictions and less distortions with regards to the tables and figures. That is, increasing the number of games throughout a whole season as well as using the same individuals throughout each of the forecasts. Also, to avoid distortions in the results of the crowdsourcing methods, it is best to use only experts with regards to the field, so that to gain the full effect of the predictive ability and whether they can be used as an accurate real-world benchmark.

References

Aaltonen, A., & Seiler, S. (2016). The cumulative growth in user-generated content production: evidence from Wikipedia. *Management Science*, *62*(7), 2054-2069.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O. ... & Zitzewitz, E. (2008). The promise of prediction markets. *Science-new york then Washington-*, *320*(5878), 877.

Coe, A., Paquet, G., & Roy, J. (2001). E-governance and smart communities: a social learning challenge. *Social science computer review*, *19*(1), 80-93.

Cui, Q., Tang, C., Xu, G., Wu, C., Shi, X., Liang, Y., Chen, L., Lee, H. P., & Huang, H. (2019). Surprisingly Popular Algorithm-Based Comprehensive Adaptive Topology Learning PSO. *2019 IEEE Congress on Evolutionary Computation (CEC)*. https://doi.org/10.1109/cec.2019.8790002

Efendić, E., Van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, *157*, 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Hosio, S., Goncalves, J., Anagnostopoulos, T., & Kostakos, V. (2016, July). Leveraging the wisdom of the crowd for decision support. In *Proceedings of the 30th International BCS Human-Computer Interaction Conference 30* (pp. 1-12).

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*(4), 322–333. https://ideas.repec.org/a/jdm/journl/v13y2018i4p322-333.html

Lee, M., Danileiko, I., & Vi, J. (n.d.). *Corrigendum for "Testing the Ability of the Surprisingly Popular Method to Predict NFL Games."* http://journal.sjdm.org/18/18331/Corrigendum.pdf

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.* doi:10.1037/a0036677

McCallister, J. (2016). Democratic Decision-Making Style: Definition & Overview. In *Study.com*. https://study.com/academy/lesson/democratic-decision-making-style-definition-lesson-quiz.html

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126-132.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532-535.

Rutchick, A. M., Ross, B. J., Calvillo, D. P., & Mesick, C. C. (2020). Does the "surprisingly popular" method yield accurate crowdsourced predictions? *Cognitive research: principles and implications*, *5*(1), 1-10. https://doi.org/10.1186/s41235-020-00256-z

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Appendix

Survey Questions

1. Hello! In this survey I will ask you to predict the winners for this weekend's Serie A football games. I'm interested in people's judgement processes and predictive ability. It takes approximately 5 minutes to complete.

   If you have any questions at any time about the survey, you may contact me, Mattia Mantovani, at mattia.mantovani@hotmail.com

   Thank you very much for participating in my survey. Click the next button to get started!

2. Demographics
   a. What is your gender?
      i. Male
      ii. Female
      iii. Non-binary / Third gender
      iv. Prefer not to say
   b. What is your age?
      i. 18-24
      ii. 25-34
      iii. 35+ years
   c. What would you rate your knowledge on Italian football?
      i. Extremely knowledgeable
      ii. Very knowledgeable
      iii. Moderately knowledgeable
      iv. Slightly knowledgeable
      v. Not knowledgeable at all

3. Example Game Question (There are ten games to predict in four surveys, and 20 games to predict in two surveys)

   Milan – Juventus

a) Who will win the match?
   a. Milan
   b. Juventus or Draw

b) Rate your confidence for the above choice. (Using a Qualtrics slider from 50 representing a pure guess to 100 representing completely sure)

c) Give an estimate on how many participants (in %) will choose your same result. (Using a Qualtrics slider from 0 to 100)

Finish Survey