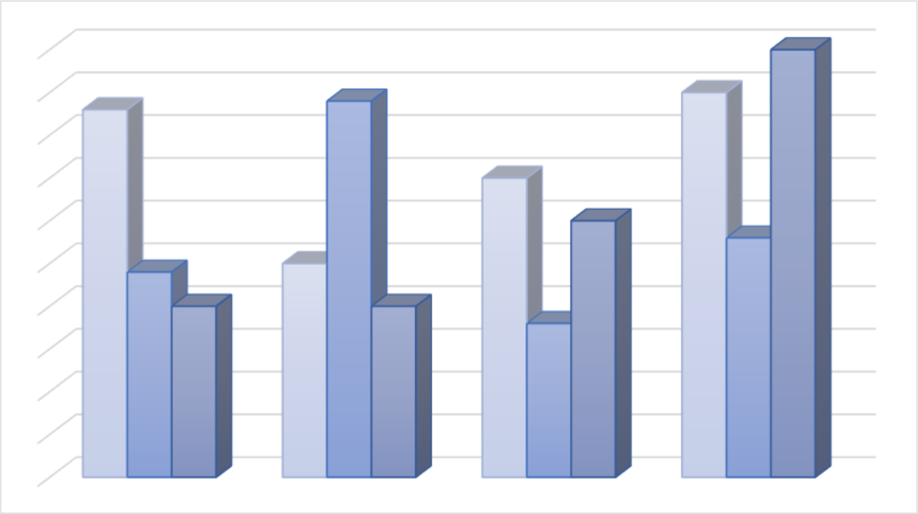
Delft University of Technology

Forecasting elections based on the aggregation of election polls

Author:
Romée Visser (4544315)



Acknowledgements

I would like to thank my supervisor Gabriela F. Nane for giving me guidance during the period of writing my thesis, but also letting me explore how to do things myself. She has given me a helping hand and advice when needed.

Abstract

Elections polls have been known to exist since 1824 [14], to fulfill the objective of what is happening and may happen. In this thesis it is investigated what the performance of election polls is and if the aggregation of polls give a better forecast than the polls themselves. The data is used of Spain during years 2012 until 2017 and The Netherlands during years 2010 until 2021. Furthermore, in this thesis the results of Spain and The Netherlands are compared. The aggregation of polls is done by using The Classical Model of Roger Cooke [2] and by using two types of Equal weighting. When using the Classical Model two different methods are used, namely a window-shift way and a cumulative way. The comparison between the aggregation of polls and the polls themselves is done by looking at the unnormalized weights and the total absolute differences. The results were quite different for the two countries. However, for both countries it appeared that method 1 was better when looking at the unnormalized weights, while method 2 was better when considering the total absolute differences. When looking at the unnormalized weights for the Spanish data the forecast when aggregating the polls, using the Item weight decision maker, was better than most of the polls' forecasts. In addition when looking at the absolute total differences of the Spanish data, the forecast when aggregating the polls, also using the Item weight decision maker, was better than most of the polls' forecasts. Similarly for the Dutch data when looking at the unnormalized weights, the forecast of the election when aggregating the polls, using the Performance-based decision maker, was better than most of the polls' forecasts. In addition, when looking at the total absolute differences, the aggregation using the Equal weight decision maker based on distribution was better than most of the polls' forecasts.

Contents

1 Abbreviations

s Empirical probability vector

p Probability vector

e Expert

m Number of calibration questions

F Cumulative distribution function (CDF) of a chi-squared distribution with 3 degrees of freedom

k Overshoot

N Total number of experts

DM Decision maker

PWDM Performance-based weight decision maker

EWDM1 Equal weight decision maker based on distribution

EWDM2 Equal weight decision maker based on quantiles

IWDM Item weight decision maker

2 Introduction

Elections are known to be held hundreds of years ago, but election polls are relatively a lot newer. However, election polls can be seen as a logic consequence of elections, as John Podhoretz claimed:

> "We're desperate for an objective rendering of what is happening and what may happen" [12].

This objective of what is happening and may happen is exactly what election polls try to fulfill. However, each poll has its own prediction. Some questions one might ask is: How well does each poll perform? And is it better to aggregate polls to get a more accurate forecast for the elections? These questions will be discussed in this thesis.

However, this thesis begins with a short introduction to the history of elections and polls. There are multiple ways to aggregate polls, so in section 4 the Classical Model will be elaborated on since it is one of the methods used to aggregate the polls. In addition, the aggregation using two types of "Equal weighting" is elaborated on. Furthermore in section 4, two methods which use the Classical Model or Equal weights to aggregate poll results are explored and compared. In sections 5 and 6 the results of Spain and The Netherlands are elaborated on. Then in section 7, the performance of each poll is analysed. In sections 8 and 9 the so called "decision makers" are compared, that is, the aggregated polls, and simultaneously method 1 and 2 are compared. Furthermore, conclusions are drawn on which decision maker is best and which method is best depending on the country. Then in section 10 it will be analysed if the forecast of the decision maker, so aggregating poll results, is better than the forecast of the polls. In addition, the different outcomes per country will be discussed and compared. Finally, some options for further research and alterations to this thesis are discussed in section 11.

2.1 History of elections and polls

As mentioned before elections are known to be held hundreds of years ago. From as early as the Medieval period to choose leaders such as the Pope and the holy Roman Emperor [18]. Also, elections are known to be one of the pillars of a democratic country. By using elections the population is able to choose an individual or a group to hold public office [18] and represent the populations needs. Although in history, not always everyone was able to vote. It were mostly white, higher class males who were allowed to vote [15]. It was not until 1919 that women in The Netherlands were able to vote [15].

Election polls have been introduced for the first time in 1824 by Raleigh Star and Wilmington American Watchman & Delaware Advertiser prior to the United States presidential elections [14]. In that year John Quincy Adams was running against Andrew, the poll correctly predicted that Jackson would win. As a consequence, polling became more popular, but still remained quite local. It was not until 1916 that the American magazine "The Literary Digest" started conducting a national survey. They did this by mailing out millions of postcards and consequently counting the answers which were returned [13]. By doing so they did predict correctly that Wilson would become president, thus making elections polls even more popular. However, the national elections polls were not always right. For example in the election of 1936 were Landon and Roosevelt were running for presidency. In that year the survey of The Literary Digest predicted that Landon would win the election based on 2.3 million survey respondents [13]. However, Roosevelt got elected instead. By that time the Literary Digest its poll was well known and well regarded, so their wrong prediction was seen as the failure of their self-proclaimed "scientific" poll [13].

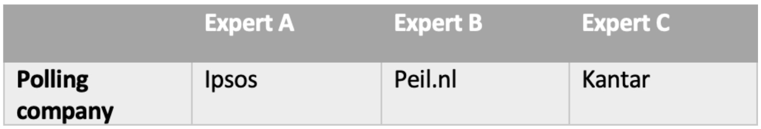
Why did the Literary Digest fail in its prediction so miserably? It is interesting to dive into that a little deeper, since the main solution for that is actually still used in the elections polls nowadays. The most obvious reason for the Literary Digest failing was that they did not have enough Roosevelt voters in their survey respondents [13]. There were more people favoring Landon participating in the survey than people favoring Roosevelt. In addition, the surveys respondents consisted more affluent Americans who were known to have more Republican sympathies at that time [13], thus favouring Landon. Nowadays, election polls try to prevent this kind of tunnel vision failures by trying to gather the most representative group for the country its population and asking their opinion. This gave also rise to different polling methods, since gathering such a representative group could be seen as quite challenging. An example of polling methods which are used nowadays are interviews or inquiries via telephone or email. However, the current election polls are still not always making perfect predictions. So it is hard to blame Literary Digest for doing so in 85 years ago.

3 The used data

The data of Spain and The Netherlands is used, these countries have a different number of election polling companies so it will be interesting to see the differences. In this section, the data will be explained in more detail. To aggregate the polls the Classical Model is used, which will be elaborated on in section 4.1. In order to use the Classical Model the data has to be "read" in a certain way, namely the polling companies will be seen as experts. Also, their predictions will be seen as "calibration questions", but this will be explained more elaborately in section 4.1. Since the data of Spain and The Netherlands is different for both countries, this section will elaborate on the given data per country.

3.1 The Netherlands

From 2010 until 2021 there were 4 elections, namely in 2010, 2012, 2017 and 2021. During this period The Netherlands had three polling companies, namely Peil.nl, Kantar and Ipsos, who forecasted the elections based on their polls. Each polling company will act as an "expert" in the Classical Model as shown in table 1 and which will be explained in section 4.1. Peil.nl is owned by Maurice de Hond, a Dutch citizen, who polls specifically in The Netherlands. While Kantar and Ipsos are international companies who poll worldwide [9]. Where for example Kantar is based in London and Ipson in Paris [16]. In the years 2010 until 2021 there are 11 parties, each party will represent a "calibration question" which will become more clear after reading section 4.1. Each polling company has expressed their estimate in the number of seats a party will receive. In The Netherlands a party with one seat or more is part of the House of Representatives [17]. All estimates have a 95% confidence interval, where for all polling companies the range of the error spans from 0 to 2 seats. Kantar and Ipsos always round up to a seat, while Peil.nl sometimes has an error of +0 seats which is because it also rounds down it seats to a whole number. In general, for Kantar and Ipsos the error for parties estimated with more than 10 seats the error is +/- 2 seats, while for parties with 10 or less estimated seats the error is +/- 1 seat. The differences in 95% confidence intervals derives from the differences in the number of people responding to each poll and the percentage of votes a party is estimated to receive. For the three companies the amount of respondents generally varies from 1400-3000 people.



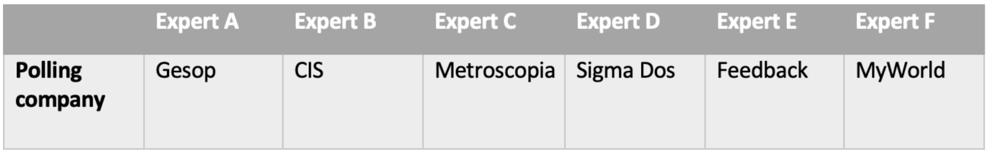
	Expert A	Expert B	Expert C
Polling company	Ipsos	Peil.nl	Kantar

Table 1: Polling companies linked to experts of The Netherlands

3.2 Spain

The data of Spain also spans from 2012 until 2021. Furthermore, it can be noted that not all data for Spain is complete from 2012 until 2021. Since for example some polling companies did not yet exist in 2012 or stopped polling at some point. From 2012 to 2017 there are six companies that have polling data, namely GESOP, Feedback, CIS, My World, Metroscopia and Sigma Dos. These companies are all Spanish companies. Furthermore, there were 7 parties from 2012 until 2017. Each polling company will act as an "expert" in the Classical Model as shown in table 2 and which will be explained in section 4.1. The elections were held in year

2012, 2015 and 2017 and this is the only time framework where so many polling companies participate in. For that reason the time framework 2012 to 2017 is used to use method 1 and 2 on, which will be explained in section 4.2 and 4.3. Since it is interesting to see if there will be different results due to the almost double amount of polling companies compared to The Netherlands. In contrast to The Netherlands, the polling companies estimate the percentage of votes a party will receive instead of seats. Similarly to The Netherlands, the data from Spain has 5%, 50% and 95% values, its value depending on which polling house does the polling. For example for the time framework 2012 to 2017, the polling companies GESOP, Feedback, CIS, My World, Metroscopia and Sigma Dos have done a polling. Where Feedback has an error of +/- 3,16 %, My World of +/- 3,2% , Metroscopia of +/- 1,7% and Sigma Dos of +/- 2,67%. As can be seen the errors differ quite a lot which makes it interesting to see the result.



	Expert A	Expert B	Expert C	Expert D	Expert E	Expert F
Polling company	Gesop	CIS	Metroscopia	Sigma Dos	Feedback	MyWorld

Table 2: Polling companies linked to experts of Spain

4 Method

As explained in the introduction the data of two countries will be analysed, namely the data of Spain and The Netherlands. This will be done through using two types of Equal weighting and the Classical Model. You might wander "why the Classical Model?". The reason for choosing the Classical Model is that it enables the aggregation of polls based on their past performance. Where it uses "expert judgement" to determine the performance of each poll. Furthermore, two methods are elaborated on, both using the Classical Model to forecast the elections of Spain and The Netherlands. In this section the Classical Model, the two types of Equal weighting and the two methods will be elaborated on.

4.1 The Classical Model

There are multiple models for making forecasts. In this thesis the Classical Model of Roger Cooke is used. It is interesting to mention that Roger Cooke has been a professor at the Delft University of Technology (TU Delft) lecturing the course Decision Theory. It was also at the TU Delft that Roger Cooke and his colleagues developed the Classical Model in the 1980s [8]. His Classical Model is using expert judgement to asses experts and make forecasts.

In the Classical Model calibration questions and unknown target questions are assessed by experts. The Classical model rates experts by assessing their performance in answering the calibration questions. The calibration questions whose uncertainty is assessed by the experts have registered true values [2]. In our data the polling companies are seen as the experts. The estimates the polling companies make for each party in each election year are seen as the calibration questions. Using these calibration questions a calibration score per expert can be computed by:

$$Cal(e) = 1 - F(2 \cdot m \cdot I(s,p)) \tag{1}$$

Where m is the number of calibration questions, s the empirical probability vector, p the probability vector and F the cumulative distribution function (cdf) of a chi-squared distribution with 3 degrees of freedom. $I(s,p)$ stands for the "relative information" or "Kullback-Leibler divergence of s and p" (see equation 15). The calibration score measures if experts' assessments are statistically accurate. The score takes a value between 0 to 1 and the higher the score the better [5]. Furthermore, an information score is computed for every question and by considering all the experts assessments for that question. In order to calculate the information score the overshoot k needs to be specified, in this case a $k = 0,1$ is used. In addition, a background distributions needs to be specified. Since the assessments of the experts are in the same order of magnitude, in other words there are no assessments that go from 3 to 3000 to 3 million, a uniform distribution is used as a background measure. The information score per expert is then computed by:

$$\begin{aligned} I(e) =& 0,05 \cdot ln(\frac{0,05}{q_5 - L^*}) + 0,45 \cdot ln(\frac{0,45}{q_{50} - q_5}) + 0,45 \cdot ln(\frac{0,45}{q_{95} - q_{50}}) \\ &+ 0,05 \cdot ln(\frac{0,05}{U^* - q_{95}}) + ln(U^* - L^*) \end{aligned} \tag{2}$$

Where q_5 is the 5% quantile and q_{50} the 50% quantile etc. L^* and U^* form the "Intrinsic range" specified in equation 17. After computing the information score per expert, the average is taken of all these information scores to get one score per expert. The information score can take any positive value, however a value bigger than 4 is not very likely [5]. The higher the score, the more informative the experts' assessments are. Then after computing the calibration and information score the performance of each expert is weighted, using the Combined score. The Combined score, or also frequently named "unnormalized weight", is computed by multiplying the calibration and information score per expert e:

$$CS(e) = Cal(e) * I(e) \tag{3}$$

Using this combined score a so called Decision Maker can be computed (see the technical appendix 11.1). The decision maker makes it possible to combine experts assessments. This means that in the case of the data from Spain and The Netherlands, the different polling companies their assessments can be combined. A decision maker represents a mathematically calculated distribution corresponding to a virtual expert. It is assumed that the real decision maker would adopt this distribution as their own.

In order to combine experts' distributions the normalized weights need to be determined, which can be computed by:

$$w_i = \frac{CS(e_i)}{\sum_{i=1}^{N} CS(e_i)} \tag{4}$$

Where $CS(e_i)$ is the combined weight of expert e_i as explained above and N is the total number of experts. So called Performance-based weights are the normalized weights and they allow us to combine experts' distributions using the probability density function (f) or the cumulative distribution function (F) (see equation 22). Since the weights are normalized the combined distributions result in a distribution as well. In order to calculate the Performance based weights a significance level α is specified at 0,05. So if the calibration score is less than 0,05 it is not taken into account in the weight. The Performance-based weights are calculated by:

$$w_{\alpha,i} = Cal(e_i) * Inf(e_i) * 1\{Cal(e_i) \geq \alpha\} \tag{5}$$

This scoring rule is an asymptotically proper scoring rule of averaging distributions [5]. This means an expert achieves his or her maximal expected weight by stating assessments in accordance with his or her true beliefs. The obtained scoring rule gives another set of weights which depends on α (see equation 21). In addition, by using this threshold α it is possible to select experts which are used in the linear pooling. Now it is possible to aggregate experts distributions using these Performance-based weights by using the equation:

$$PWDM = w_{e_1} \cdot F_1 + w_{e_2} \cdot F_2 + ... + w_{e_N} \cdot F_N \tag{6}$$

Where $PWDM$ stands for Performance-based weight decision maker and as previously described F_i is the cumulative distribution function per expert i.

Instead of using Performance-based weight it is also possible to use Item weight. Item weight is similar to the Performance-based weight. However, not the overall information score is taken into account, but for each calibration question there is an information score per expert. Together with the average of all the information scores a combined score can be obtained for each calibration question.

Another possible weighting is Equal weights. Equal weight based on distribution and Equal weight based on quantiles, where the first one is combining distributions and the second one combines 50% percentiles of the estimated value. When using Equal weights based on distribution all the experts are assigned equal "normalized weights":

$$EWDM_1 = \frac{1}{N} \cdot F_1 + \frac{1}{N} \cdot F_2 + ... \frac{1}{N} \cdot F_N \tag{7}$$

Again, N is the total number of experts and F_i the cumulative distribution function per expert i. Using Equal weight based on distribution, the uncertainty is taken into account. Since not only one value is looked at (50% value), but also the 5% and 95% value. In contrast, the Equal weight decision maker based on quantiles does not take uncertainty into account, since it only uses the 50% estimated value of an expert. It adds all the 50% estimated values and divides it by the total number of experts, which is referred to as "quantile aggregation", see equation:

$$EWDM_2 = \frac{\sum_{i=1}^{N} 50\% \text{ estimated value of } e_i}{N} \tag{8}$$

Where e_i is expert i and N the total number of experts.

In order to calculate all these different scores, the software Excalibur is used. Using Excalibur the decision makers can be computed.

One might ask: Why would you look at experts assessments so intensively? Well, validation of experts is needed since experts are often statistically inaccurate. It appeared that in 33 studies were the Classical Model was used, only less than one-third of the experts were accurate [2].

4.2 Using expert judgement method 1

In this section the first method for forecasting the elections of Spain or The Netherlands is elaborated on.

Method 1 will be a "Window shift" option for forecasting. This meaning an election year will be evaluated and then the next election year is forecasted based on the evaluated year. This will continuously be done until you have reached the last year. To make it more clear a visualization is made in figure 1. In this figure the green coloured part is the year which is evaluated and then the yellow coloured part is the subsequent year which is forecasted. So for example on top of figure 1 the year 2010 is observed and then year 2012 is forecasted, based on the performance-weighted combination of polls. Then in the middle, year 2012 is analysed and then year 2015 is forecasted based on the performance-weighted combination of polls of year 2012.

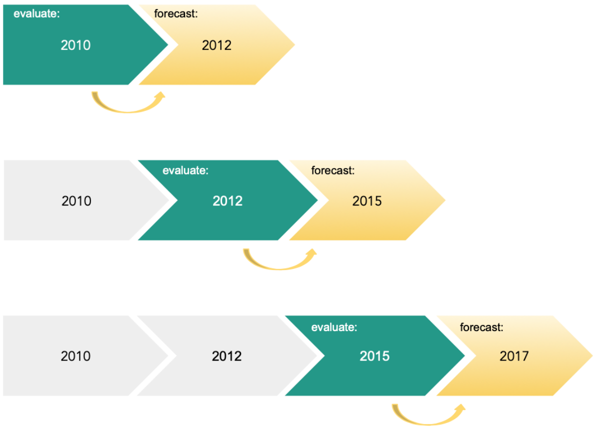


Figure 1: Window shift option for forecasting

4.3 Using expert judgement method 2

In this section the second method for forecasting the elections of Spain or The Netherlands is elaborated on.

Method 2 is a "cumulative way" of forecasting. This means that the more data there is available, the more data is used. Also, the performance-weighted combination of polls will hold more years into account. The most recent election year will use the performance-weighted combination of polls of all the previous years. To make it more clear a visualization is made in figure 3. The evaluated years are shown in green and the year to forecast in yellow. When looking at the top bar in figure 3 the year 2010 is the only year available. Based on the assumption that only the data of year 2010 is available, year 2012 is forecasted based on the performance-weighted combination of polls of year 2010. Consequently, when data is available of years 2010 and 2012, as portrayed in the middle bar in the figure, the performance-weighted combination of polls is used of years 2010 and 2012 to forecast year 2015.

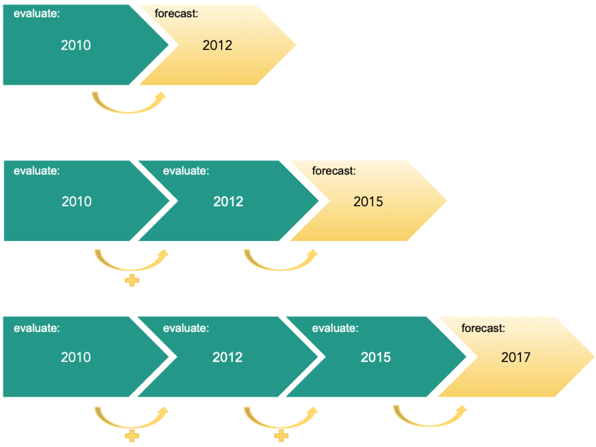


Table 3: Cumulative option for forecasting

Since the data is different for both countries, as discussed in section 3, it will be interesting to see if the outcome of the methods will be different too. This will be elaborated on in sections 8 and 9.

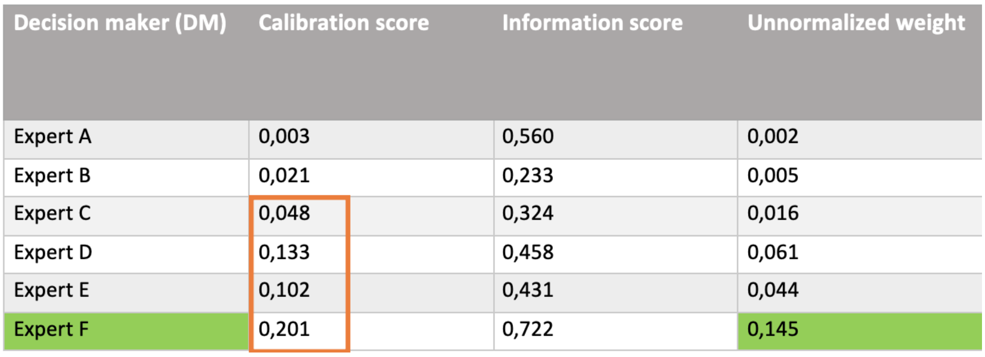
5 Results Spain

In this section the different scores and weights discussed in section 4.1 of Spain are shown and discussed. These are the calibration and information scores per expert and per decision maker, as well as the unnormalized weight for each. As discussed in 3 the polls will be referred to as experts, since the Classical Model is used. In addition, as explained in section 4.1 the software Excalibur is used to calculate the scores and the different weights. Furthermore, the decision makers are shown and discussed in this section.

5.1 Calibration and Information scores and Unnormalized weight

When using method 1 there are 7 calibration questions assessed by the experts each election year. In this subsection the calibration scores and information scores are compared in each year for each expert. When the calibration score is higher than 0,05 it is seen as statistically accurate, this is highlighted in the tables underneath with an orange box around the calibration score. Furthermore, as explained in section 4.1 the higher the information score, the more informative an expert is. In addition, the unnormalized weight or "combined score" is analysed which is the product of the calibration score and the information score.

Calibration and information scores and unnormalized weights using method 1
The table with the values of the calibration and information scores and the unnormalized weight of experts of 2012 can be seen in table 4. Expert A and B are the only experts which do not have a calibration score higher than 0,05, so they are not very statistically accurate. While the other experts all have a calibration score higher than 0,05, which means they are statistically accurate. The highest calibration score is obtained by expert F with a value of 0,201. While the lowest calibration score is obtained by expert A with a value of 0,003. When looking at the information scores, the highest information score is obtained by expert F with a value of 0,722 so it is quite informative. The lowest information score is obtained by expert B with a value of 0,233, so it is not very informative. The lowest unnormalized weights is achieved by expert A with a value of 0,002, due to the low calibration score. While the highest unnormalized weight is achieved by expert F due to its relatively high calibration and information scores.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,003	0,560	0,002
Expert B	0,021	0,233	0,005
Expert C	0,048	0,324	0,016
Expert D	0,133	0,458	0,061
Expert E	0,102	0,431	0,044
Expert F	0,201	0,722	0,145

Table 4: Calibration and information scores and unnormalized weights of experts of 2012

In figure 2 the values of table 4 are shown visually. It can be seen that expert F has the highest bars regarding the calibration and information score, shown in green. In addition, note that although expert A has relatively a high information score, due to the low calibration score of expert A it results in a low unnormalized weight.

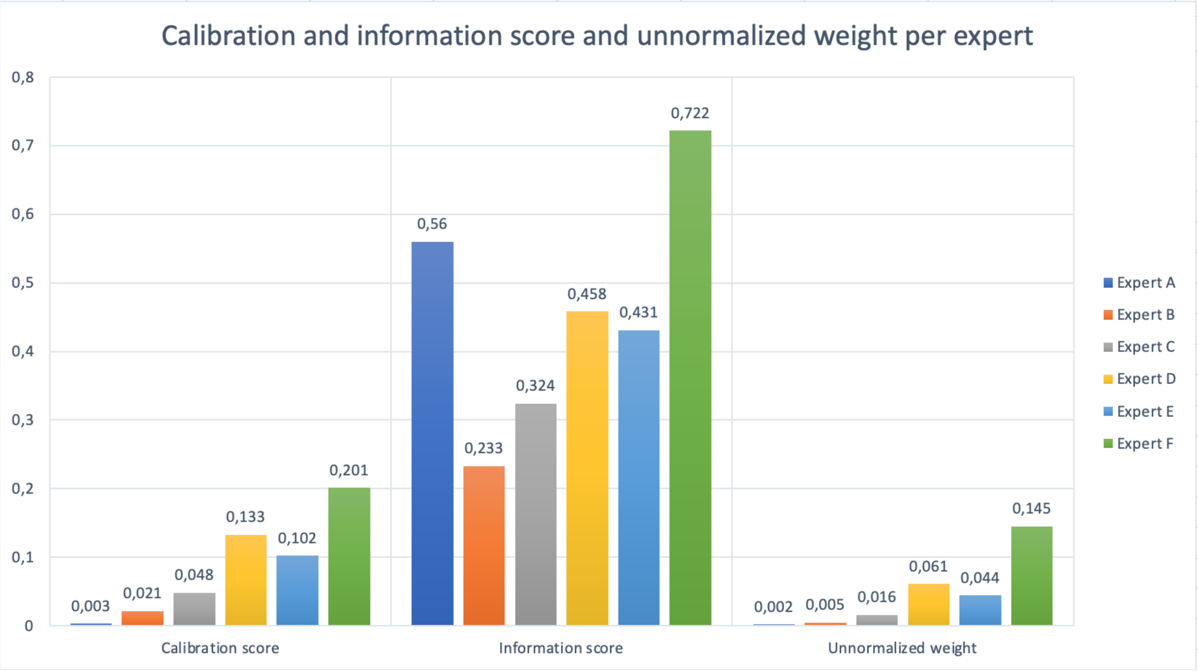
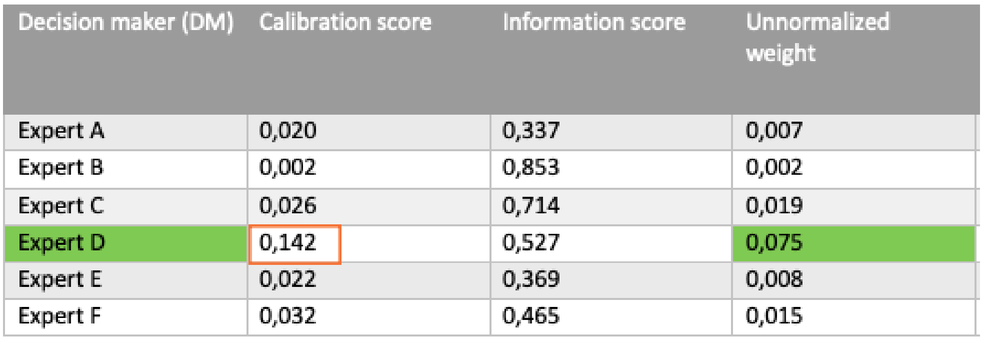


Figure 2: Calibration and information scores and unnormalized weights of experts of 2012

The table with the values of the calibration and information scores and the unnormalized weight of experts of 2015 can be seen in table 5. Expert D is the only expert who has a calibration score above 0,05 ,which means it is statistically accurate. The other experts all have a calibration score beneath 0,05, meaning they are not statistically accurate. The lowest calibration score is obtained by expert B with a value of almost 0, namely 0,002, so its statistical accuracy is very poor. When looking at the information scores, all scores are quite high. This means all experts are quite informative. The highest information score is obtained by expert B with a value of 0,853, this means expert B is the most informative expert. However, due to the higher calibration score of expert D, expert D eventually has the highest unnormalized weight which is highlighted in with a green box. All other unnormalized weights are relatively low due to the low calibration scores.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,020	0,337	0,007
Expert B	0,002	0,853	0,002
Expert C	0,026	0,714	0,019
Expert D	0,142	0,527	0,075
Expert E	0,022	0,369	0,008
Expert F	0,032	0,465	0,015

Table 5: Calibration and information scores and unnormalized weights of experts of 2015

In figure 3 the values of table 5 are shown visually. It can be seen that expert D has the highest bar regarding the calibration score and unnormalized weight. In addition, note that although expert B has the highest information score, it has a very low unnormalized weight.

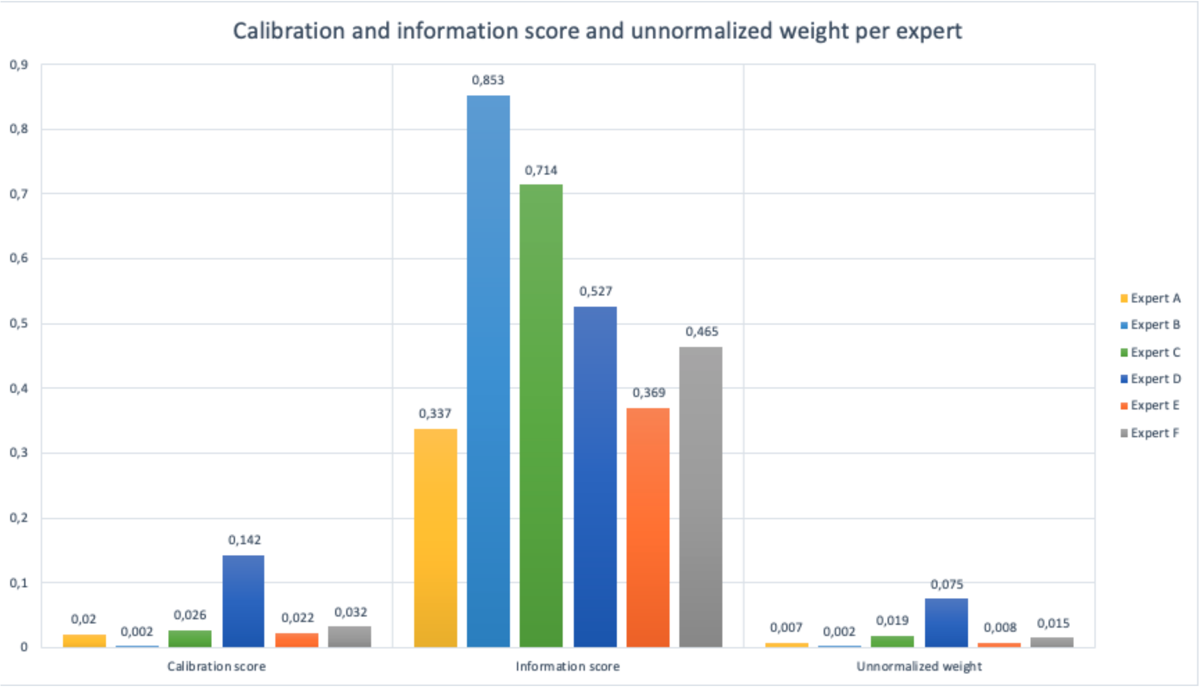
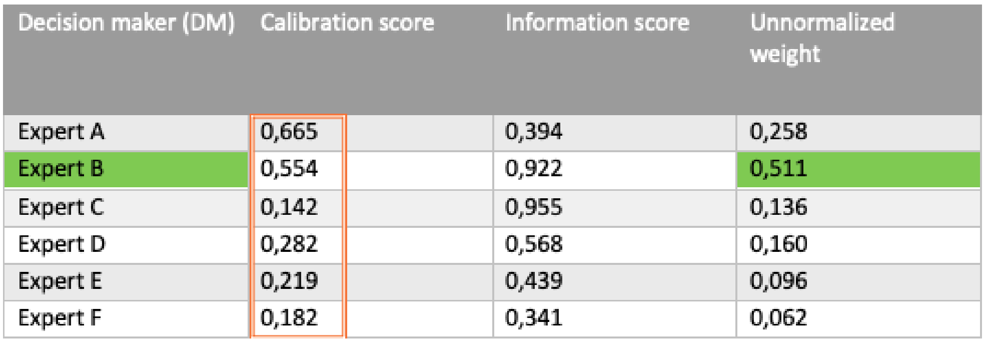


Figure 3: Calibration and information scores and unnormalized weights of experts of 2015

The table with the values of the calibration and information scores and the unnormalized weight of experts of 2017 can be seen in table 6. Note that the statistical accuracy of the experts is a lot better than in 2015. All calibration scores are above 0,05, which is shown with the orange box. The highest calibration score is obtained by expert A, with a value of 0,665, so its statistical accuracy is quite good. The lowest calibration score is obtained by expert C with a value of 0,142. When looking at the information scores, the lowest score is obtained by expert F with a value of 0,341. So expert F is comparably less informative. While expert C is comparably very informative with a score of 0,955. Due to the relatively high calibration and information score of expert B, namely 0,544 and 0,922, it results in the highest unnormalized weight with a value of 0,511.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,665	0,394	0,258
Expert B	0,554	0,922	0,511
Expert C	0,142	0,955	0,136
Expert D	0,282	0,568	0,160
Expert E	0,219	0,439	0,096
Expert F	0,182	0,341	0,062

Table 6: Calibration and information scores and unnormalized weights of experts of 2017

In figure 4 the values of table 6 are shown visually. It can be seen that almost all values are a higher compared to year 2012. Furthermore, it can be seen that the values lie relatively close sometimes. For example, expert B and C their informativeness is quite similar. In addition, it is clear from figure 4 that expert B has the best unnormalized weight and expert F the worst.

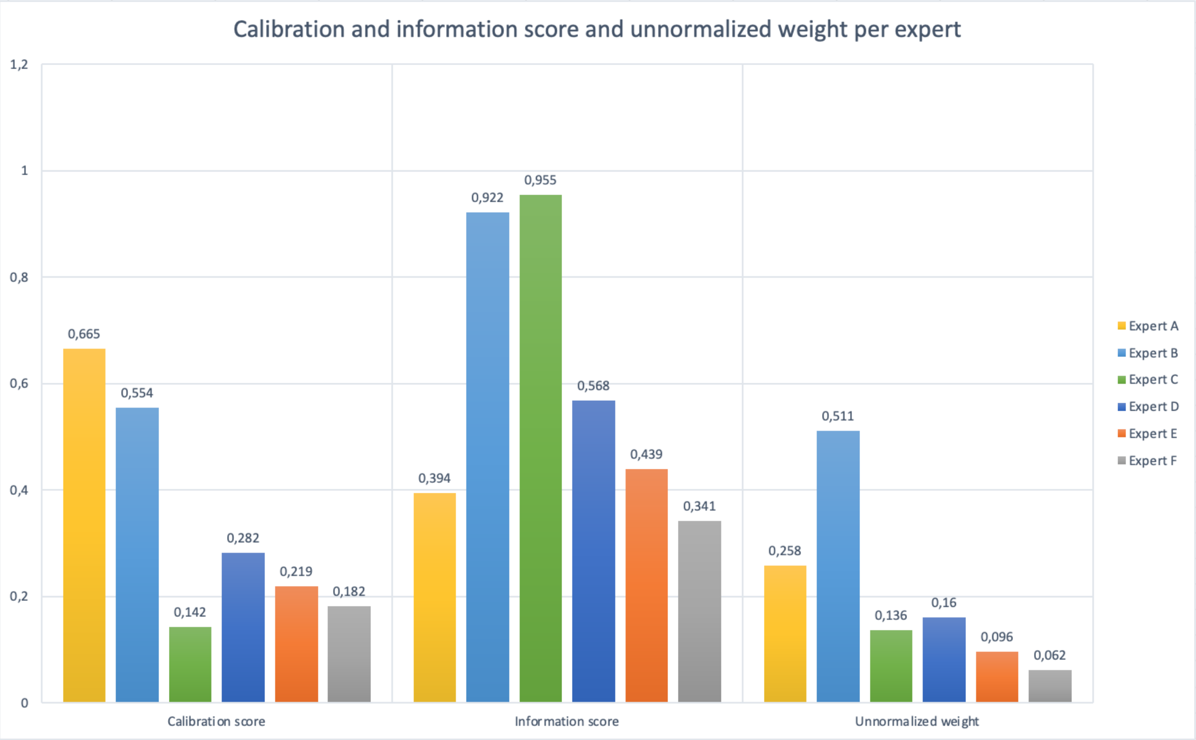
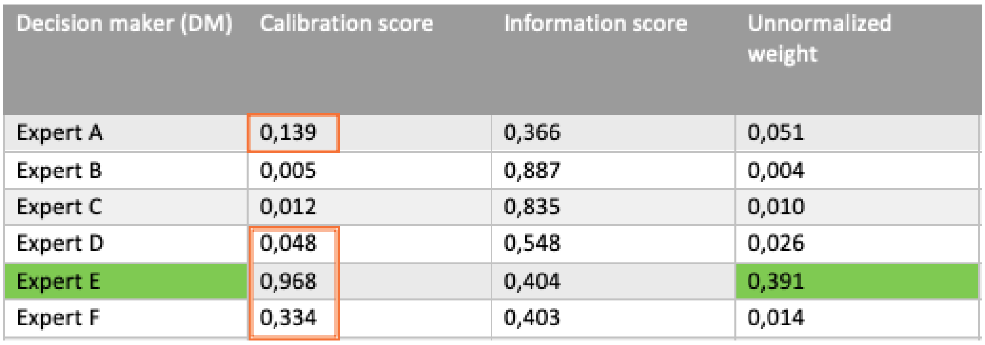


Figure 4: Calibration and information scores and unnormalized weights of experts of 2017

Calibration and information scores and unnormalized weights using method 2
When using method 2 there are 14 calibration questions assessed by the experts. The table with the values of the calibration and information scores and the unnormalized weight of experts of 2017 using method 2 can be seen in table 7. 4 out of 6 experts have calibration score which are equal to or above 0,05, which are highlighted by the orange box. Expert E has a calibration score of 0,968 which is very statistically accurate, since the highest possible score is 1. In contrast to expert B who has a score of 0,005, so whose statistical accuracy is very poor. Again, when it comes to information scores expert B and C have the highest values, respectively 0,887 and 0,835. However, due to their low calibration scores they have quite a low unnormalized weight. In contrast to expert E who has an unnormalized weight of 0,391 due to its high calibration score.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,139	0,366	0,051
Expert B	0,005	0,887	0,004
Expert C	0,012	0,835	0,010
Expert D	0,048	0,548	0,026
Expert E	0,968	0,404	0,391
Expert F	0,334	0,403	0,014

Table 7: Calibration and information scores and unnormalized weights of experts of year 2017 using method 2

In figure 5 the values of table 7 are shown visually. Expert E has clearly the highest calibration score. While when it comes to information scores expert B and C have the highest scores. When looking at the unnormalized weight in the most right part of the figure it is clear that expert E has by far the highest weight.

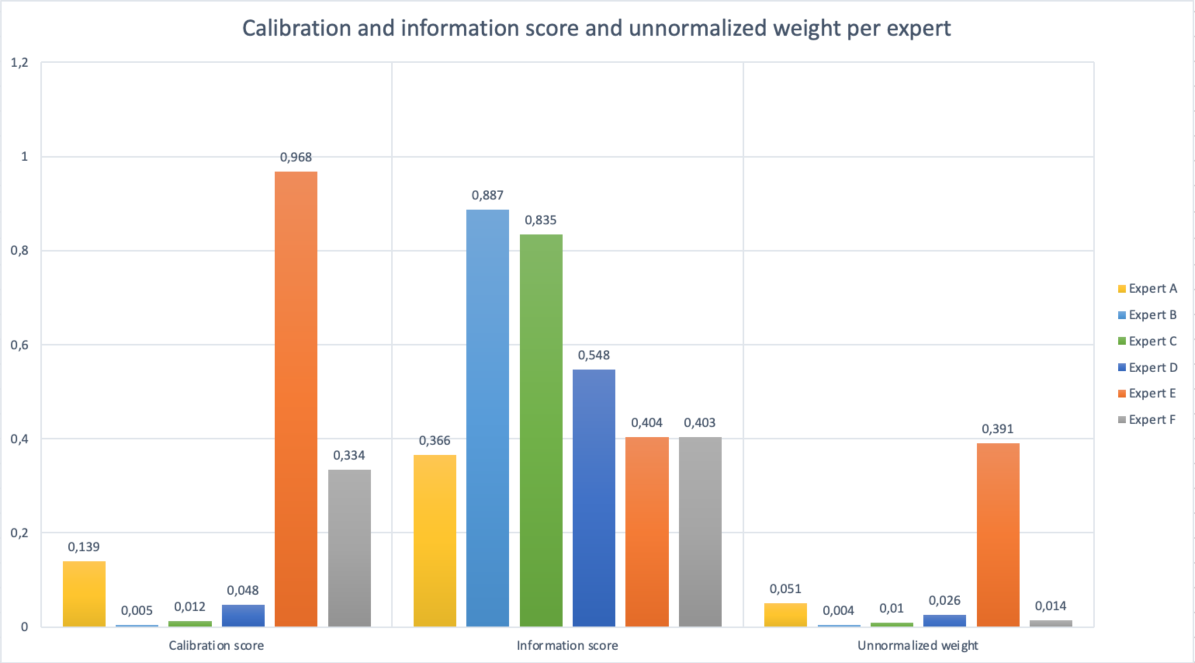


Figure 5: Calibration and information scores and unnormalized weights of experts of year 2017 using method 2

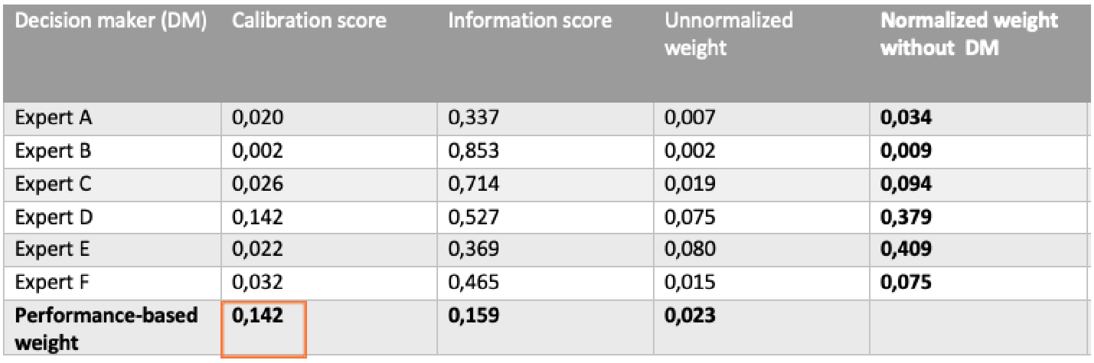
Overall, the highest calibration score is obtained by expert E in 2017 when using method 2, with a value of 0,968. While the highest information score is obtained by expert B with a value of 0,922 in year 2017 using method 1. In addition, in 2017 using method 1 the highest unnormalized weight is obtained by expert B with a value of 0,511.

5.2 Performance-based weight decision maker

In this subsection the Performance-based weight decision maker is elaborated on, which is explained more elaborately in section 4.1. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

Performance-based weight decision maker using method 1

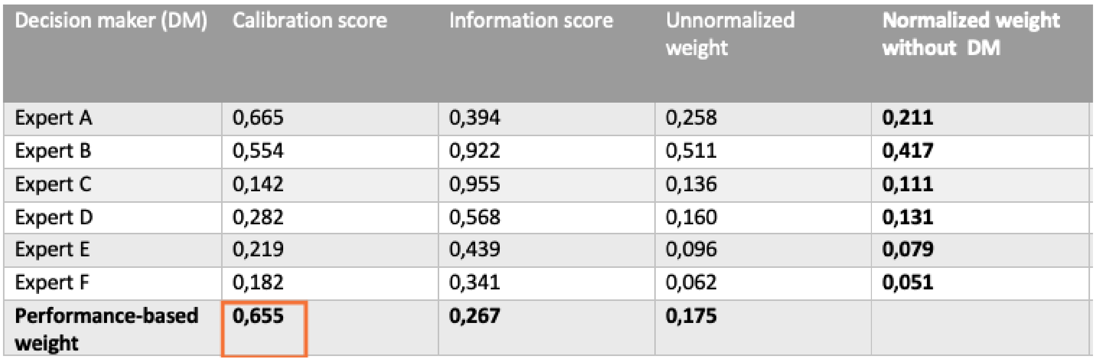
For year 2015 the outcome of computing the Performance-based decision maker is shown in table 8. Note that the calibration score of the Performance-based weight is statistically accurate, with a value of 0,142. The information score is 0,159 which does not seem very high, so it is not very informative. These lead to an unnormalized weight of 0,023 which is quite low. When looking at the normalized weight without decision maker, expert D and E have the highest weights meaning their distributions are contributing most to the distribution of the Performance-based weight decision maker. While expert B has the lowest score, with a value of 0,009, which means expert B is the least taken into account compared to the other experts.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,020	0,337	0,007	**0,034**
Expert B	0,002	0,853	0,002	**0,009**
Expert C	0,026	0,714	0,019	**0,094**
Expert D	0,142	0,527	0,075	**0,379**
Expert E	0,022	0,369	0,080	**0,409**
Expert F	0,032	0,465	0,015	**0,075**
Performance-based weight	**0,142**	**0,159**	**0,023**	

Table 8: Performance-based weight decision maker for year 2015 using method 1

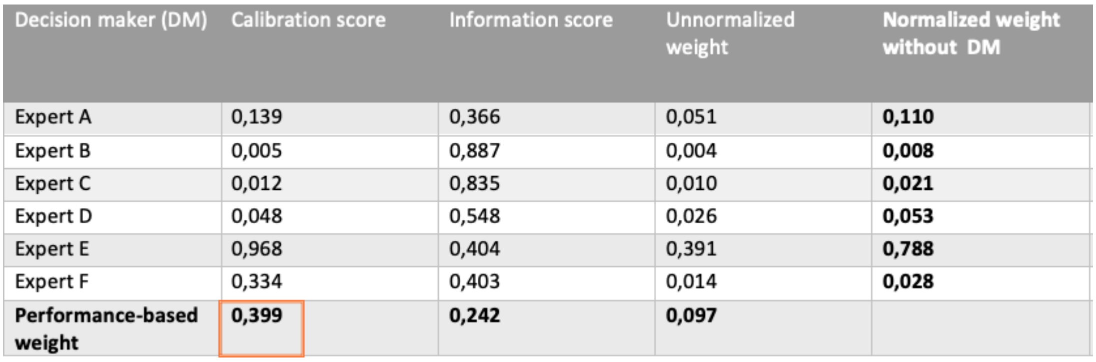
In table 10 the outcome is shown when computing the Performance-based decision maker in year 2017. Its calibration score is 0,655, which means its statistical accuracy is quite good. The information score of the Performance-based weight decision maker is 0,267, so it is more informative in 2017 than it was in 2012. However, it is still relatively not very informative. This leads to a unnormalized weight of 0,175. When looking at the unnormalized weights the highest weight is obtained by expert B with a value of 0,417. While in 2012 the lowest unnormalized weight was obtained by expert B. But in 2017 expert B is taken most into account, this is due to the high unnormalized weight of expert B in 2017. The second best expert, expert D, has unnormalized weight of 0,131. This is quite a lot lower compared to expert B, so the distribution of expert D contributes less to the distribution of the decision maker. The lowest normalized weight is obtained by expert F, with a value of 0,051.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,665	0,394	0,258	**0,211**
Expert B	0,554	0,922	0,511	**0,417**
Expert C	0,142	0,955	0,136	**0,111**
Expert D	0,282	0,568	0,160	**0,131**
Expert E	0,219	0,439	0,096	**0,079**
Expert F	0,182	0,341	0,062	**0,051**
Performance-based weight	**0,655**	**0,267**	**0,175**	

Table 9: Performance-based weight decision maker for year 2017 using method 1

Performance-based weight decision maker using method 2
For year 2017 the outcome of computing the Performance-based decision maker is shown in table 10. It can be seen that the calibration score is 0,399 so it is statistically accurate. The information score is 0,242 which is not very informative. This leads to a unnormalized weight of 0,097, which is quite low. When looking at the unnormalized weights the highest normalized weight is obtained by expert E, with a value of 0,788. This means the distribution of expert E is taken most into account. In comparison with experts B, D and F who have quite low normalized weights so their distributions are taken less into account when computing the decision maker.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,139	0,366	0,051	**0,110**
Expert B	0,005	0,887	0,004	**0,008**
Expert C	0,012	0,835	0,010	**0,021**
Expert D	0,048	0,548	0,026	**0,053**
Expert E	0,968	0,404	0,391	**0,788**
Expert F	0,334	0,403	0,014	**0,028**
Performance-based weight	**0,399**	**0,242**	**0,097**	

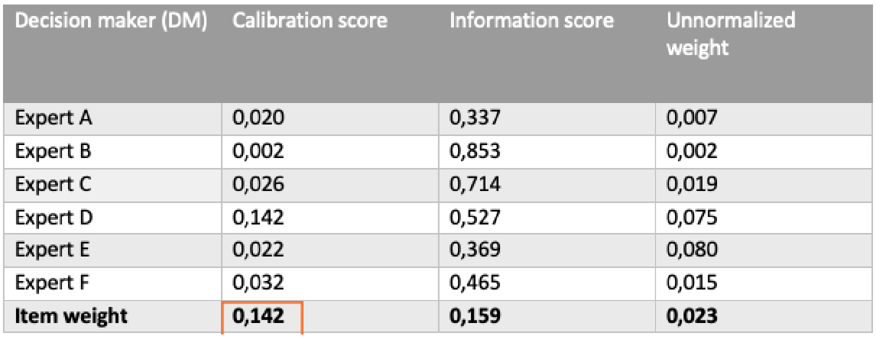
Table 10: Performance-based weight decision maker for year 2017 using method 2

Overall, all the Performance-based weight decision makers are statistically accurate. The highest calibration score is obtained in year 2017 using method 1, with a value of 0,655. In addition, the highest information score was also achieved that year using method 1, with a value of 0,267. These lead to the highest unnormalized weight of 0,175.

5.3 Item weight decision maker

In this subsection the Item weight decision maker is discussed. As mentioned in section 4.1 the Item weight is similar to the performance-based weight. However, not the overall information score is taken into account, but for each calibration question there is an information score for an expert. Together with the average of all the information scores a combined score can be obtained for each calibration question. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

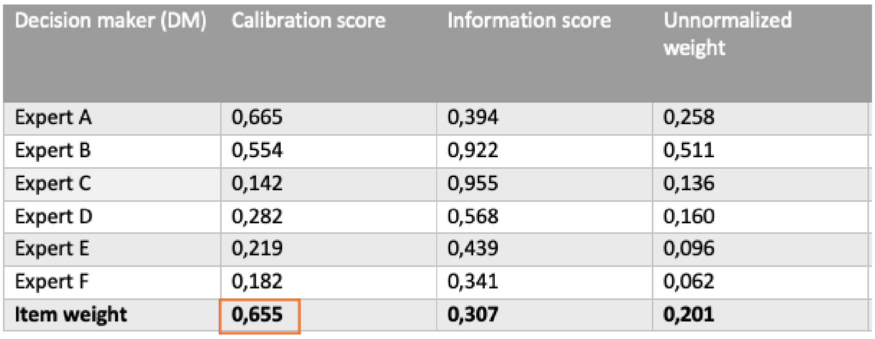
Item weight decision maker using method 1
For year 2015 the outcome of computing the Item weight decision maker is shown in table 11. It can be seen that the calibration score is statistically accurate, with a value of 0,142. The information score is 0,159, which is not very informative. This results in an unnormalized weight of 0,023, which is quite low too.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,020	0,337	0,007
Expert B	0,002	0,853	0,002
Expert C	0,026	0,714	0,019
Expert D	0,142	0,527	0,075
Expert E	0,022	0,369	0,080
Expert F	0,032	0,465	0,015
Item weight	**0,142**	**0,159**	**0,023**

Table 11: Item weight decision maker for year 2015 using method 1

In table 13 the outcome is shown when computing the Item weight decision maker in year 2017. The calibration score is 0,655, which means its statistical accuracy is quite good. Together with an information score of 0,307, which is relatively informative, an unnormalized weight is obtained of 0,201.

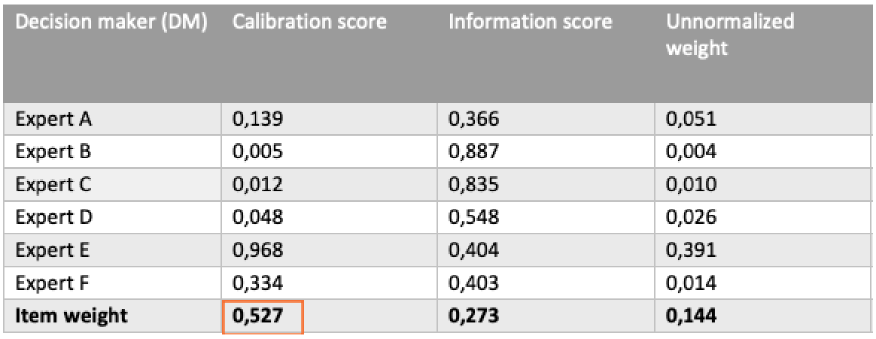


Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,665	0,394	0,258
Expert B	0,554	0,922	0,511
Expert C	0,142	0,955	0,136
Expert D	0,282	0,568	0,160
Expert E	0,219	0,439	0,096
Expert F	0,182	0,341	0,062
Item weight	**0,655**	**0,307**	**0,201**

Table 12: Item weight decision maker for year 2017 using method 1

Item weight decision maker using method 2

When using method 2 for year 2017 the outcome of computing the Item weight decision maker is shown in table 13. It can be seen that the calibration score is 0,527 which is statistically accurate. The information score is 0,273, which is not very informative. The calibration and information score result in an unnormalized weight of 0,144, which is quite high compared to all experts.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,139	0,366	0,051
Expert B	0,005	0,887	0,004
Expert C	0,012	0,835	0,010
Expert D	0,048	0,548	0,026
Expert E	0,968	0,404	0,391
Expert F	0,334	0,403	0,014
Item weight	**0,527**	**0,273**	**0,144**

Table 13: Item weight decision maker for year 2017 using method 2

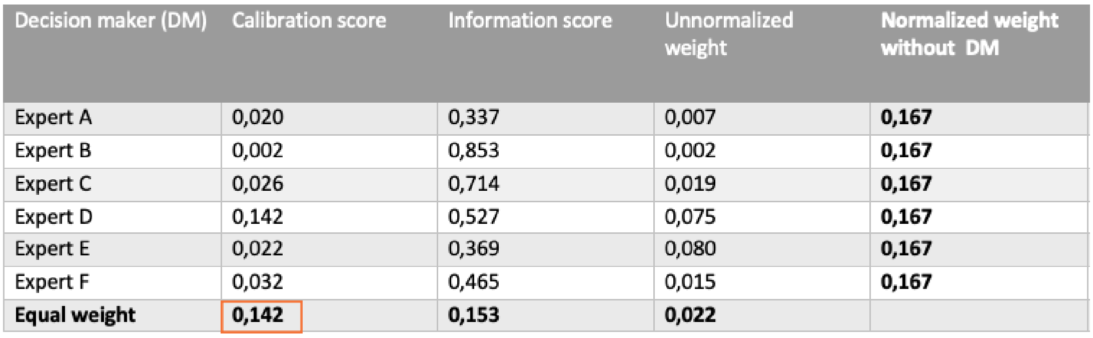
Overall, all the Item weight decision makers are statistically accurate. The highest information score is obtained in year 2017 using method 1, with a value of 0,655. In addition, the highest information score was also achieved that year using method 1, with a value of 0,307. These lead to the highest unnormalized weight of 0,201.

5.4 Equal weight decision maker based on distribution

In this subsection the results of the Equal weight decision maker based on distribution is analysed, which is combining distributions. The normalized weights for each expert are all 0,167 since the Equal weight decision maker assigns equal weight to all experts. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

Equal weight decision maker using method 1

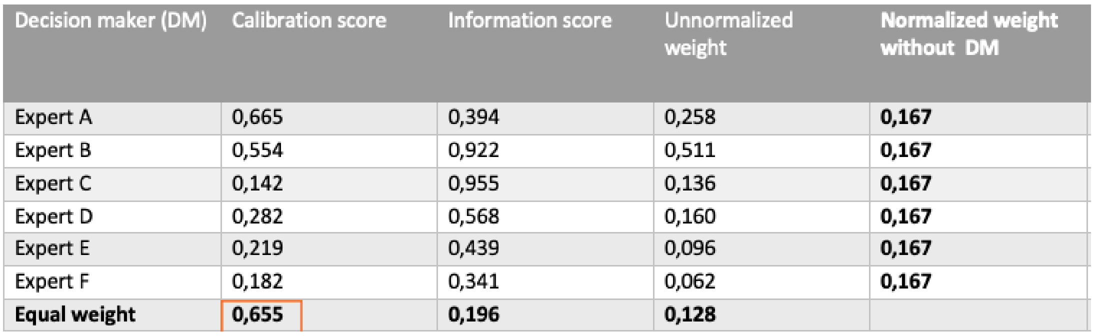
For year 2015 the outcome of computing the Equal weight decision maker based on distribution is shown in table 14. It can be seen that the calibration score is 0,142, so it is statistically accurate. The information score is 0,153 which is not very informative. This results in an unnormalized weight of 0,022.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,020	0,337	0,007	**0,167**
Expert B	0,002	0,853	0,002	**0,167**
Expert C	0,026	0,714	0,019	**0,167**
Expert D	0,142	0,527	0,075	**0,167**
Expert E	0,022	0,369	0,080	**0,167**
Expert F	0,032	0,465	0,015	**0,167**
Equal weight	**0,142**	**0,153**	**0,022**	

Table 14: Equal weight decision maker for year 2015 using method 1

In table 16 the outcome is shown when computing the Equal weight decision maker in year 2017. The calibration score is 0,655, which is quite high so its statistical accuracy is good. While the information score is 0,196, so the decision maker is not very informative. Due to the low information score, the unnormalized weight is 0,128.

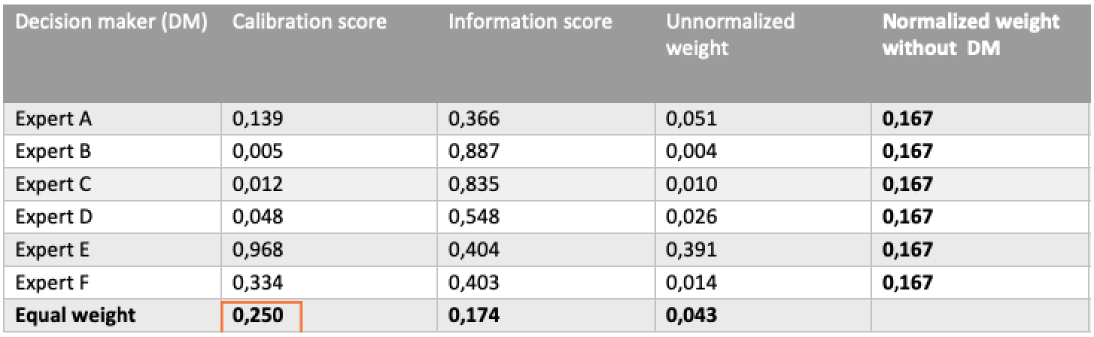


Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,665	0,394	0,258	**0,167**
Expert B	0,554	0,922	0,511	**0,167**
Expert C	0,142	0,955	0,136	**0,167**
Expert D	0,282	0,568	0,160	**0,167**
Expert E	0,219	0,439	0,096	**0,167**
Expert F	0,182	0,341	0,062	**0,167**
Equal weight	**0,655**	**0,196**	**0,128**	

Table 15: Equal weight decision maker for year 2017 using method 1

Equal weight decision maker using method 2

For year 2017 the outcome of computing the Equal weight decision maker based on distribution is shown in table 16. It can be seen that the calibration score is 0,250 this is statistically accurate. The information score is 0,174, so it is not very informative. Due to the low information score, the unnormalized weight is quite low too with a value 0f 0,043.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	0,139	0,366	0,051	**0,167**
Expert B	0,005	0,887	0,004	**0,167**
Expert C	0,012	0,835	0,010	**0,167**
Expert D	0,048	0,548	0,026	**0,167**
Expert E	0,968	0,404	0,391	**0,167**
Expert F	0,334	0,403	0,014	**0,167**
Equal weight	**0,250**	**0,174**	**0,043**	

Table 16: Equal weight decision maker for year 2017 using method 2

Overall, the highest calibration score is obtained in 2017 using method 1. In addition, the highest information score was then obtained. This leads as well to the highest unnormalized weight in 2017. All information scores of the Equal weight decision maker based on distribution are quite low compared to the experts, regardless of what year is been looked at, so it is not very informative.

5.5 Equal weight decision maker based on quantiles

In this subsection the Equal weight decision maker based on quantiles (EWDM_2) is elaborated on. As discussed in section 4.1 the Equal weight decision maker based on quantiles does not take uncertainty into account. In addition, instead of combining distributions it combines the 50% percentiles of the estimations. The values of the EWDM_2 are easily computed by equation:

$$EWDM_2 = \frac{\sum_{i=1}^{N} 50\% \text{ estimated value of } e_i}{\text{total number of experts}} \tag{9}$$

The EWDM_2 values can be found in appendix 11.2, since it does not add a lot of value to discuss the values of the EWDM_2 per election year in here. In section 8 the EWDM_2 is compared to the other decision makers.

6 Results The Netherlands

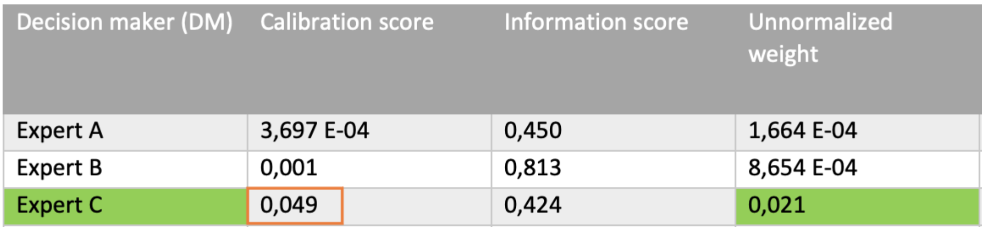
In this section the different scores and weights discussed in section 4.1 of The Netherlands are shown and discussed. These are the calibration and information scores per expert and per decision maker, as well as the unnormalized weight for each. As discussed in 3 the polls will be referred to as experts, since the Classical Model is used. In addition, as explained in section 4.1 the software Excalibur is used to calculate the scores and the different weights. Furthermore, the decision makers are shown and discussed in this section.

6.1 Calibration and Information scores and Unnormalized weight

In this subsection the calibration scores and information scores are compared in each year for each expert. When the calibration score is higher than 0,05 it is seen as statistically accurate, this is highlighted in the tables underneath with an orange box around the calibration score. Furthermore, as explained in section 4.1 the higher the information score, the more informative an expert is. In addition, the unnormalized weight or "combined score" is analysed which is the product of the calibration score and the information score.

Calibration and information scores and unnormalized weights using method 1
The table with the values of the calibration and information scores and the unnormalized weight of experts in 2010 can be seen in table 17. Expert A has a very low calibration score which means it has a poor statistical accuracy and it is relatively not very informative either

with an information score of 0,450. Expert B is more statistically accurate than expert A and it is more informative than experts A and C. Expert C has the highest calibration score, namely 0,049, but it has the lowest information score compared to the other experts, namely 0,424. When noting that the calibration score is best when closest to 1, all calibration scores are quite low. In addition, note that almost scores are lower than 0,05 except the score of expert C, this means the statistical accuracy is quite poor of expert A and B. When looking at the unnormalized weight column expert A and B have a relatively low score, namely 1,664 E-4 and 8,654 E-4. This is due to the low calibration scores. Expert C has the highest unnormalized weight, namely 0,021. However, also the unnormalized weights are overall quite low, which is the consequence of the low statistical accuracy of the experts.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	3,697 E-04	0,450	1,664 E-04
Expert B	0,001	0,813	8,654 E-04
Expert C	0,049	0,424	0,021

Table 17: Calibration and information scores and unnormalized weights of experts of 2010

In figure 6 the values of table 17 are shown visually. Now it can easily be noted that the calibration scores are very low. In contrast, the information scores are quite high, so the experts are quite informative. However, the low calibration scores lead to a low unnormalized weight, which can be seen in the most right column of figure 6.

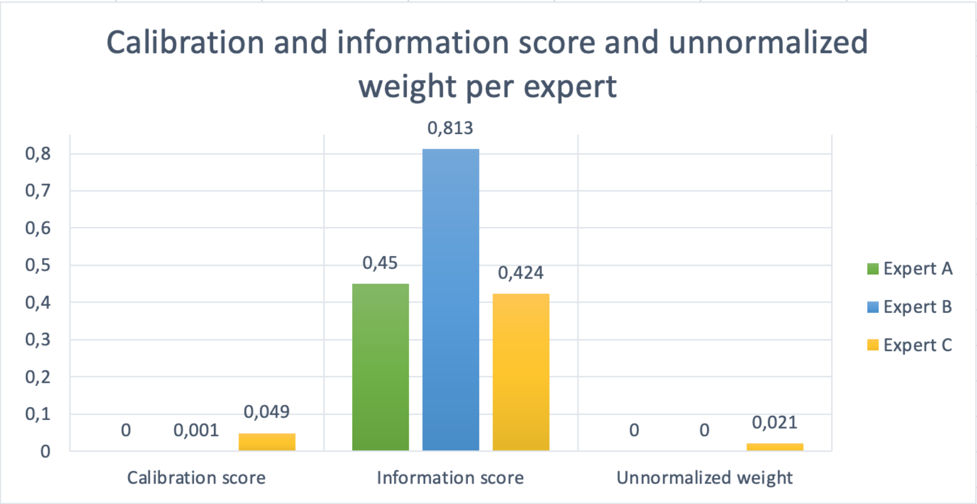
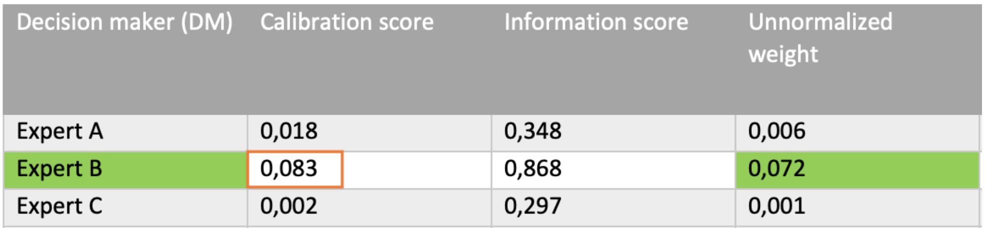


Figure 6: Calibration and information scores and unnormalized weights of experts of 2010

When using method 1 there are 11 calibration questions each election year. For year 2012 the table with the values of the calibration and information scores and the unnormalized weight of each experts are shown in table 18. When looking at expert A, its calibration score and information score is between expert B and C their scores. Expert A has a calibration score lower than 0,05, which means it is not statistically accurate. Expert B has the highest calibration score, namely 0,083, and the highest information score, namely 0,868. This means it is the most statistically accurate and most informative of the three experts. In contradiction to expert C who has the lowest calibration score and information score, namely 0,002 and 0,297, and is therefore the least statistically accurate and least informative of the three experts. The unnormalized weights are quite low, the highest unnormalized weight is obtained by expert B with a value of 0,072. If you hold the threshold of 0,05 into account for the calibration score, only expert B has significant values. However, expert B has an unnormalized weight of 0,072, which is still quite low.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,018	0,348	0,006
Expert B	0,083	0,868	0,072
Expert C	0,002	0,297	0,001

Table 18: Calibration and information scores and unnormalized weights of experts of 2012

In figure 7 the values of table 18 are shown. It can be noted that the calibration scores are higher than in year 2010. However all calibration scores are still quite low. Again, the information scores are quite high compared to the calibration scores. However, the low calibration scores lead to a low unnormalized weight, which can be seen in the most right column of figure 7.

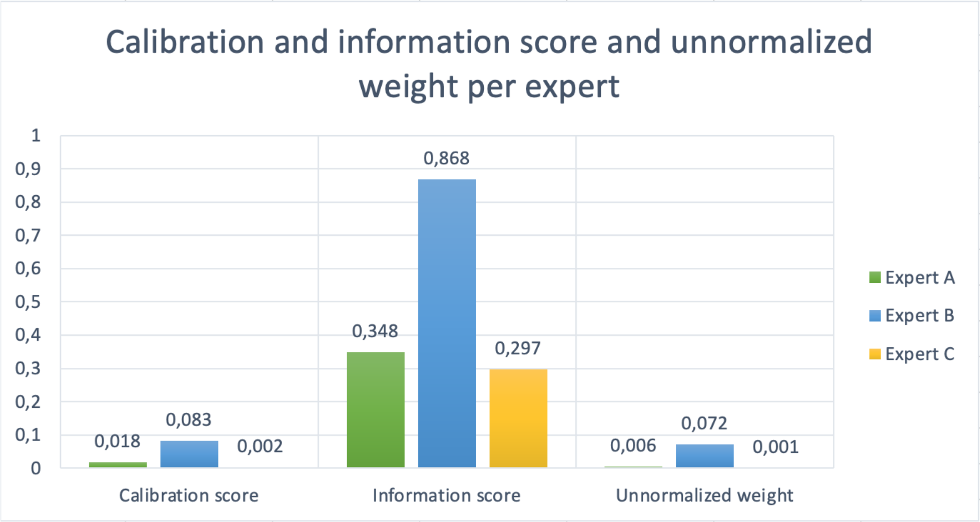
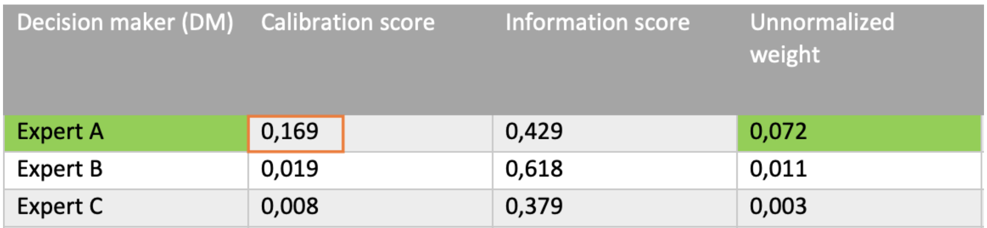


Figure 7: Calibration and information scores and unnormalized weights of experts of 2012

For year 2017 the table with the values of the calibration and information scores and the unnormalized weight of each experts can be seen in table 19. Expert A has a reasonably high calibration score which means it is quite statistically accurate. Expert B and C both have a calibration score lower than 0,05, which means they are not very statistically accurate. When looking at the information scores they range from 0,379 to 0,429 which means they are reasonably informative. The unnormalized weights however are again relatively low since the calibration scores are so low. Expert A is the only expert this year with a significant calibration score. The unnormalized weight of expert A is 0,072, which is the same as expert B had in 2012 but which is still quite low.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,169	0,429	0,072
Expert B	0,019	0,618	0,011
Expert C	0,008	0,379	0,003

Table 19: Calibration and information scores and unnormalized weights of experts of 2017

The values of 19 are visualized in figure 8. Note that the calibration score of expert A is comparably high. Together with the higher information score it can be seen that this leads to a higher unnormalized weight. While expert B has the highest information score, but due to the low calibration score the unnormalized weight shown in the most right column is relatively low.

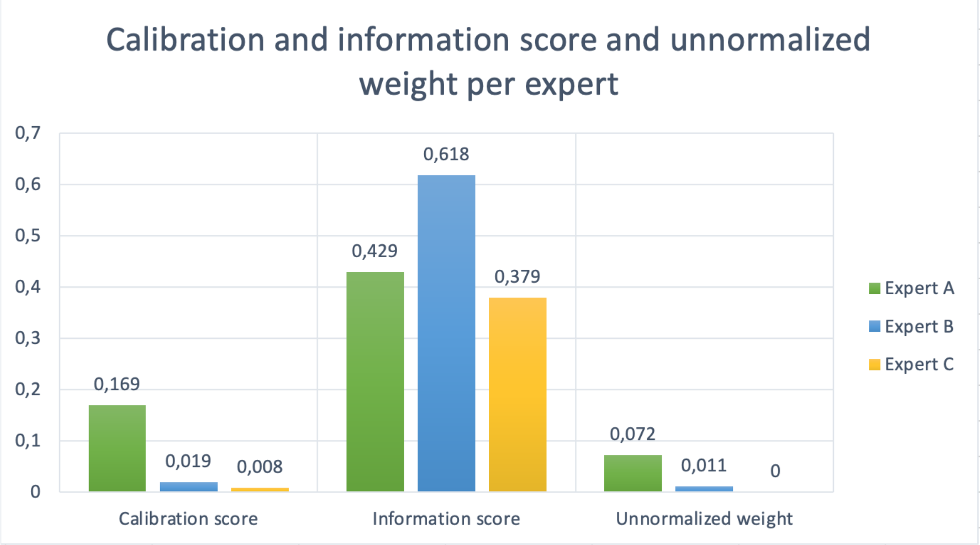
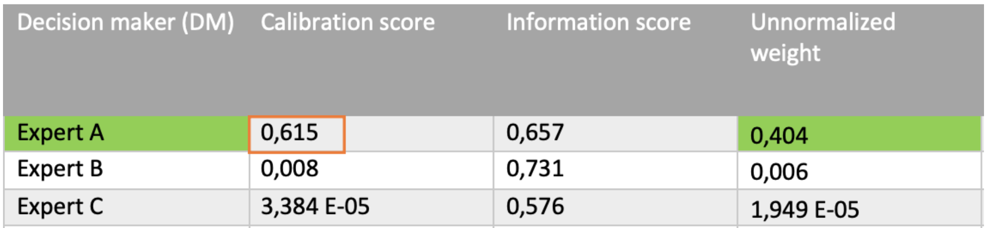


Figure 8: Calibration and information scores and unnormalized weights of experts of 2017

The table with the values of the calibration and information scores and the unnormalized weight of experts of 2021 can be seen in table 20. In 2021 the calibration scores are not very statistically accurate either for experts B and C. Both scores are well below 0,05. However, their information scores are relatively quite high so they are relatively informative. Due to the low calibration scores their unnormalized weights are still comparably low. On the contrary, the calibration score of expert A is high. The value 0,615 is relatively close to 1, which means it is statistically accurate. In addition, the information score of expert A is in 2021 reasonably high too, which means it is quite informative. The two lead to an unnormalized weight of 0,404, which is comparably high too.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,615	0,657	0,404
Expert B	0,008	0,731	0,006
Expert C	3,384 E-05	0,576	1,949 E-05

Table 20: Calibration and information scores and unnormalized weights of experts of 2021

In figure 8 the values of table 19 are shown visually. Now it is clear that expert A has the highest calibration score and a relatively high information score thus leading to the highest unnormalized weight. In addition, it can be seen that the low calibration scores of expert B and C lead to a low unnormalized weight, regardless of their high information scores.

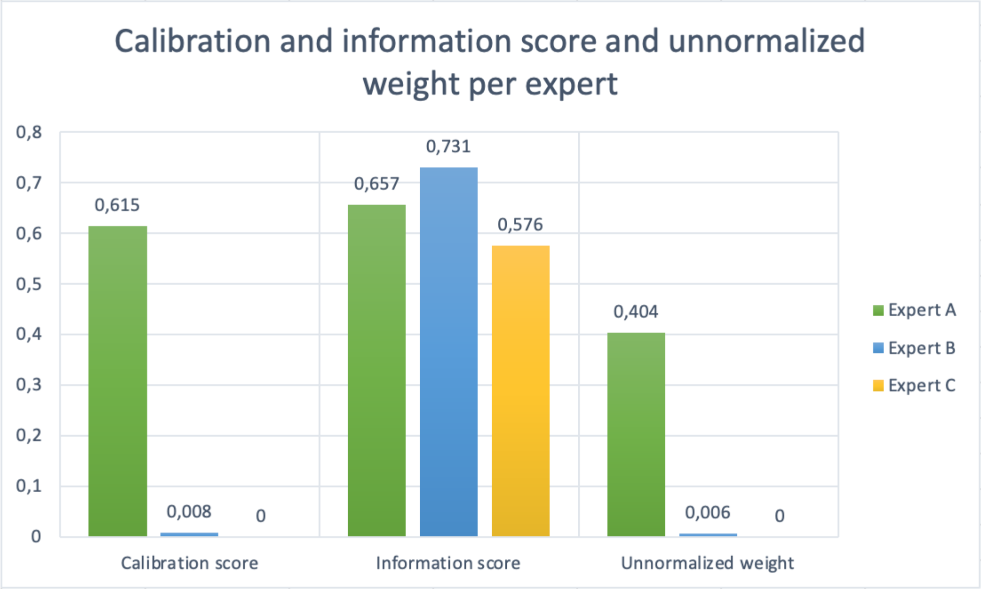
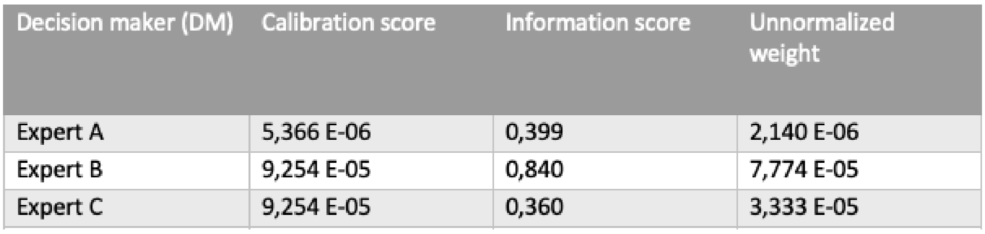


Figure 9: Calibration and information scores and unnormalized weights of experts of 2021

Calibration and information scores and unnormalized weights using method 2
Now the calibration scores and information scores and unnormalized weights are compared in each year for each expert when using method 2. When using method 2 there are 22 calibration questions assessed by the experts when forecasting year 2017. When forecasting year 2021 there are 33 calibration questions assessed by the experts. For year 2017 the table with the values of the calibration and information scores and the unnormalized weight of experts are shown in table 21. Note that all calibration scores are quite low, they are all below 0,05 which means their statistical accuracy is poor. The information score of expert B is relatively high, which means its informativeness is reasonable. However, due to the low calibration score expert B has a low unnormalized weight as well. All in all, the unnormalized weights are very low.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	5,366 E-06	0,399	2,140 E-06
Expert B	9,254 E-05	0,840	7,774 E-05
Expert C	9,254 E-05	0,360	3,333 E-05

Table 21: Calibration and information scores and unnormalized weights of experts of 2017 using method 2

The values of table 22 are visually shown in figure 11. Now it is clear that all experts had very low calibration scores. This leads to very low unnormalized weights regardless of their higher informativeness.

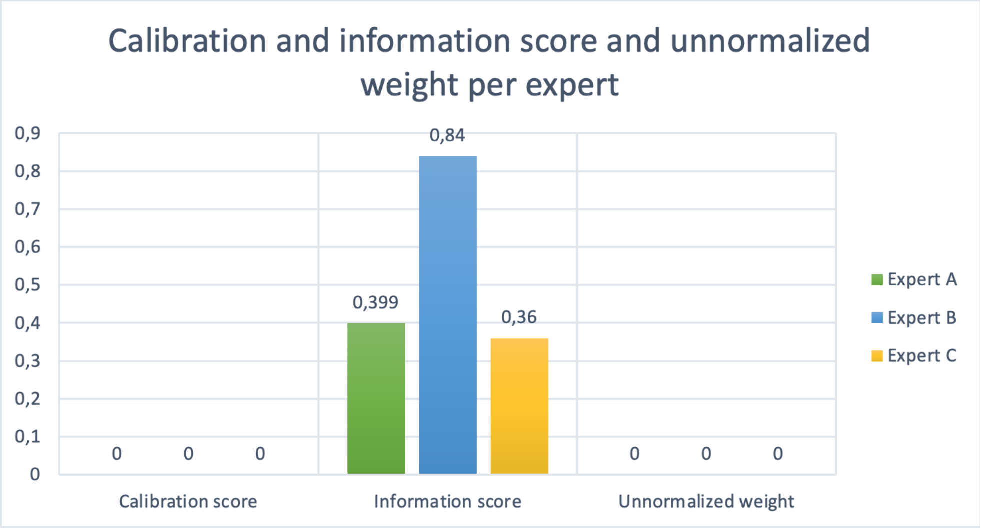
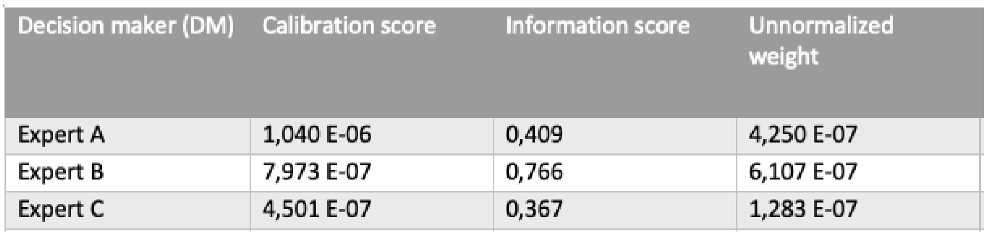


Figure 10: Calibration and information scores and unnormalized weights of experts of 2017 using method 2

The table with the values of the calibration and information scores and the unnormalized weight of experts of year 2021 can be seen in table 22. The informativeness of the experts is relatively high which can be seen in the third column of table 22. However, just as in year 2017 using method 2 all the calibration scores are quite low, they are all below 0,05 which means their statistical accuracy is poor. The information score of expert B is relatively high, but due to the low calibration score expert B has a low unnormalized weight as well. All in all, the unnormalized weights are very low.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	1,040 E-06	0,409	4,250 E-07
Expert B	7,973 E-07	0,766	6,107 E-07
Expert C	4,501 E-07	0,367	1,283 E-07

Table 22: Calibration and information scores and unnormalized weights of experts of 2021 using method 2

In figure 11 the values of table 22 are shown visually. Now it is clear that all experts had very low calibration scores. Thus leading to very low unnormalized weights regardless of their higher informativeness.

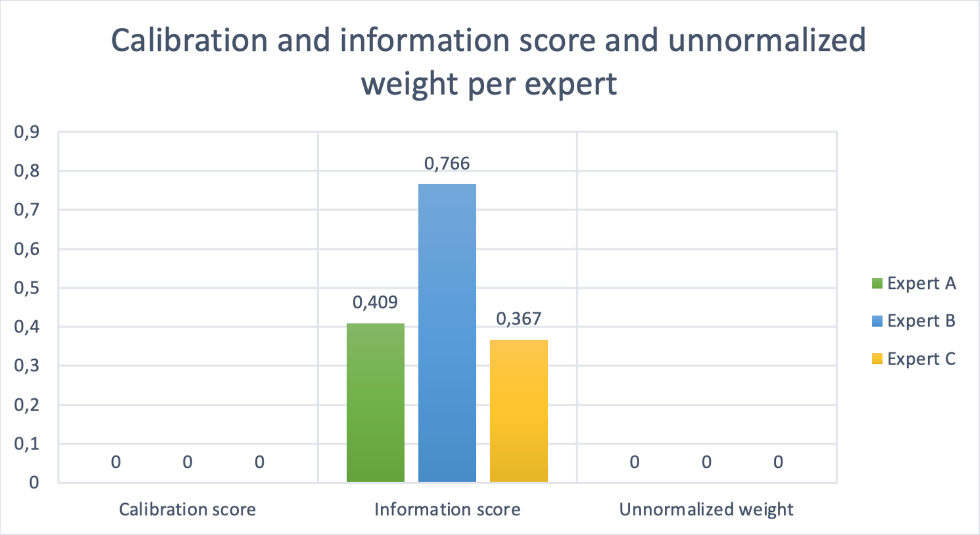


Figure 11: Calibration and information scores and unnormalized weights of experts of 2021 using method 2

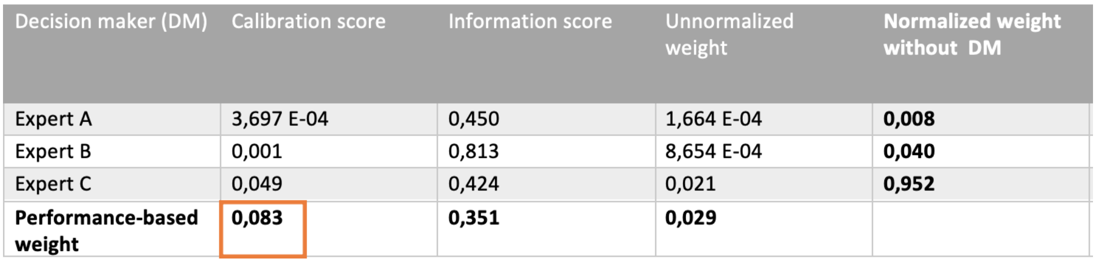
Overall, the highest calibration score is obtained by A in 2021 using method 1 and also the highest unnormalized weight is obtained by expert A in 2021 with a value of 0,404. The highest information score is obtained by expert B in 2012 using method, with a value of 0,868. It can be noted that using method 1 there is only one expert each year which is significant and this expert is not the same for each year. However, when using method 2 none of the experts have calibration scores higher than 0,05. As a result of the low calibration scores all the unnormalized weights are very low too when using method 2.

6.2 Performance-based weight decision maker

In this subsection the Performance-based weight decision maker is elaborated on, which is explained more elaborately in section 4.1. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

Performance-based weight decision maker using method 1

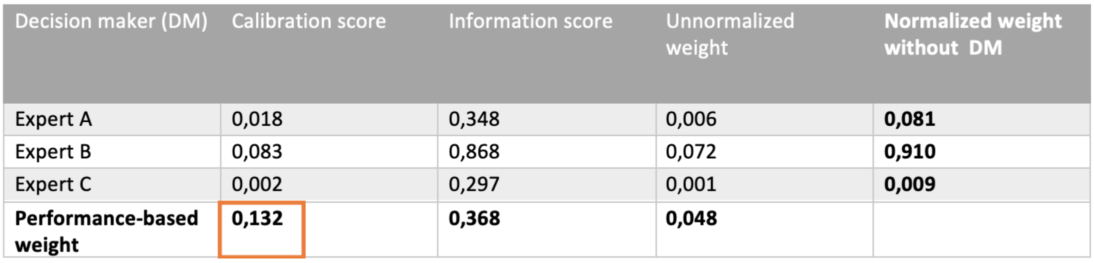
In tables 23, 24 and 25 the outcomes are shown when computing the Performance-based decision maker. When looking at the year 2012, shown in table 23, the calibration score of the performance-based decision maker is 0,083, which is above the significance level 0,05. This means it is statistically accurate. The information score is 0,351, which means it is quite informative. These scores lead to the unnormalized weight with a value of 0,029. The highest normalized weight, which can be seen in the most right column, is obtained by expert C with a value of respectively 0,952. This means the distribution of expert C contributes more to the distribution of the decision maker compared to the other experts who have quite low normalized weights, namely 0,008 and 0,040.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	3,697 E-04	0,450	1,664 E-04	**0,008**
Expert B	0,001	0,813	8,654 E-04	**0,040**
Expert C	0,049	0,424	0,021	**0,952**
Performance-based weight	**0,083**	**0,351**	**0,029**	

Table 23: Performance-based weight decision maker for year 2012 using method 1

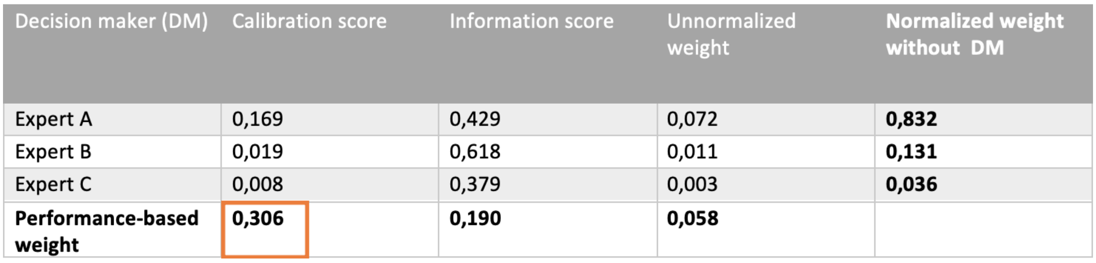
When looking at the year 2017, the calibration score of the performance-based decision maker is 0,132, so statistically accurate. The information score is 0,368, so it is relatively informative. The information and calibration score lead to an unnormalized weight of 0,048. The highest normalized weight without decision maker is obtained by expert B with a value of respectively 0,910, so the distribution of expert B contributes most to the distribution of the decision maker. This is in contrast with expert A and C who have quite low normalized weights, respectively 0,081 and 0,009.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	0,018	0,348	0,006	**0,081**
Expert B	0,083	0,868	0,072	**0,910**
Expert C	0,002	0,297	0,001	**0,009**
Performance-based weight	**0,132**	**0,368**	**0,048**	

Table 24: Performance-based weight decision maker for year 2017 using method 1

When looking at the year 2021, shown in table 25, the calibration score of the Performance-based decision maker is 0,306, so statistically accurate, and the information score is 0,190. Leading to an unnormalized weight with a value of 0,058, which is the highest compared to 2010 and 2012. In 2021 the highest normalized weight is obtained by expert A with a value of respectively 0,832, so the distribution of expert A contributes the most to the distribution of the decision maker. In addition, the normalized weight of expert B is 0,131, so the distribution of expert B is also contributing a little to the distribution of the Performance-based decision maker. This is in contrast with expert C who has a very low normalized weight of 0,036.



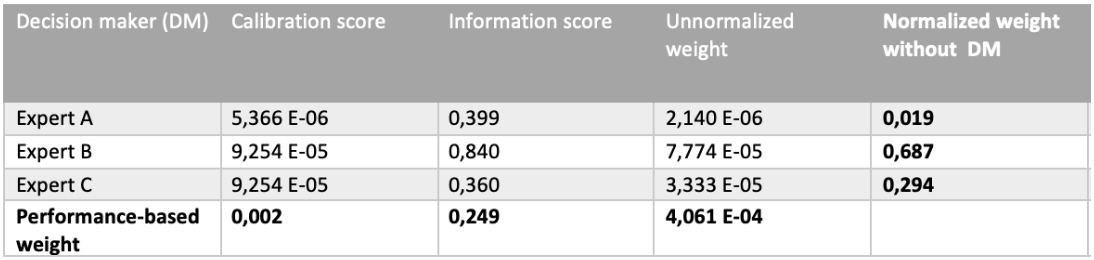
Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	0,169	0,429	0,072	**0,832**
Expert B	0,019	0,618	0,011	**0,131**
Expert C	0,008	0,379	0,003	**0,036**
Performance-based weight	**0,306**	**0,190**	**0,058**	

Table 25: Performance-based weight decision maker for year 2021 using method 1

All in all, the highest Performance-based weight is achieved in year 2021 with an unnormalized weight of 0,058. Furthermore, since strong weights reward good expertise in uncertainty quantification [5], this means expert A is rewarded most in 2021, expert B in 2017 and expert C in 2012.

Performance-based weight decision maker using method 2

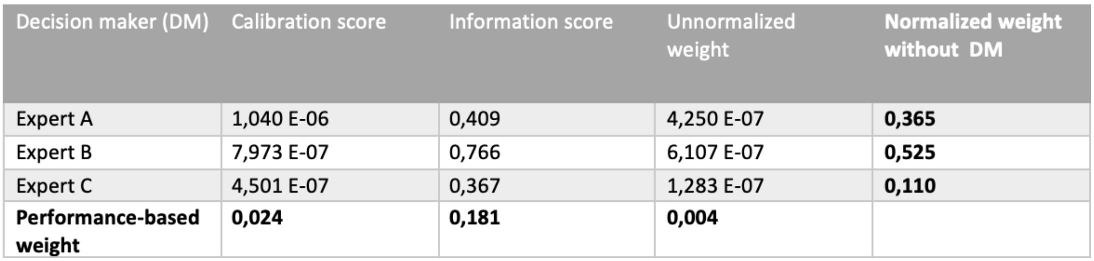
In tables 26 and 27 the outcome of the Performance-based weights are shown when using method 2. When looking at the year 2017, the calibration score of the performance-based decision maker is 0,002, which is higher than any of the experts. In addition, the information score is 0,249, which is lower than the experts. So the performance-based decision maker is more statistically accurate and less informative compared to the experts. However, the calibration score is still lower than the significance level 0,05, so its statistical accuracy is quite poor. The unnormalized weight is quite low, namely 4,061 E-04 which could be expected due to the low calibration score. In addition, it can be see that in the case of the Performance-based weight expert B has the highest normalized weight without decision maker, so the distribution of expert B is contributing most to the distribution of the decision maker compared to the other experts.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	5,366 E-06	0,399	2,140 E-06	**0,019**
Expert B	9,254 E-05	0,840	7,774 E-05	**0,687**
Expert C	9,254 E-05	0,360	3,333 E-05	**0,294**
Performance-based weight	**0,002**	**0,249**	**4,061 E-04**	

Table 26: Performance-based weight decision maker for years 2017 using method 2

When looking at the year 2021 using method 2, the calibration score of the performance-based decision maker is 0,024 and the information score 0,181. Again, the calibration score is higher than the experts their calibration scores and the information score is lower than the experts their information scores. However, also for this case the calibration score is lower than 0,05, so again not very statistical accurate. This leads to a low unnormalized weight, namely 0,004. Here again the normalized weight without decision maker of expert B is the highest, which means expert B will get rewarded most in uncertainty quantification.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	1,040 E-06	0,409	4,250 E-07	**0,365**
Expert B	7,973 E-07	0,766	6,107 E-07	**0,525**
Expert C	4,501 E-07	0,367	1,283 E-07	**0,110**
Performance-based weight	**0,024**	**0,181**	**0,004**	

Table 27: Performance-based weight decision maker for year 2021 using method 2

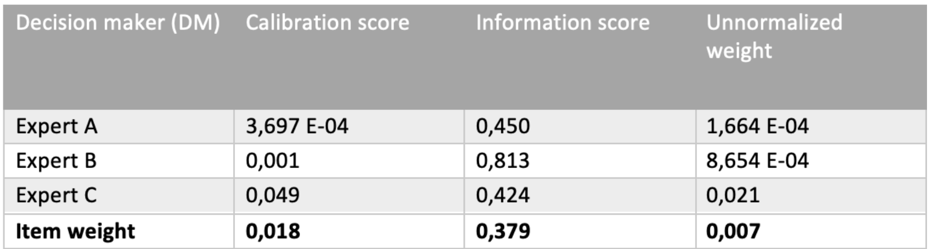
Expert B has the highest normalized weight in both year 2017 as years 2021. Strong weights reward good expertise in uncertainty quantification [5], so this means expert B is rewarded most. Furthermore, the highest unnormalized weight of the Performance-based weight is achieved in 2021 with a value of 0,004. Compared to method 1 this is quite low, since when using method 1 in 2021 the value of the Performance-based unnormalized weight was 0,058.

6.3 Item weight decision maker

In this subsection the Item weight decision maker is discussed. As mentioned in section 4.1 the Item weight decision maker is similar to the Performance-based weight decision maker. However, not the overall information score is taken into account, but for each calibration question there is an information score for an expert. Together with the average of all the information scores a unnormalized score can be obtained for each calibration question. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

Item weight decision maker using method 1

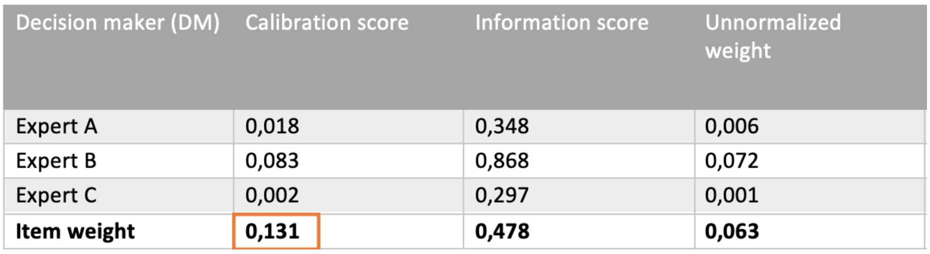
In tables 28, 29 and 30 the outcomes are shown. When looking at the year 2010, shown in table 28, the calibration score of the Item weight decision maker is 0,018, so its statistical accuracy is quite poor. The information score is 0,379, which means it is relatively informative. The unnormalized weight is 0,007 which is quite low, because of the lower calibration score.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	3,697 E-04	0,450	1,664 E-04
Expert B	0,001	0,813	8,654 E-04
Expert C	0,049	0,424	0,021
Item weight	**0,018**	**0,379**	**0,007**

Table 28: Item weight decision maker for year 2012 using method 1

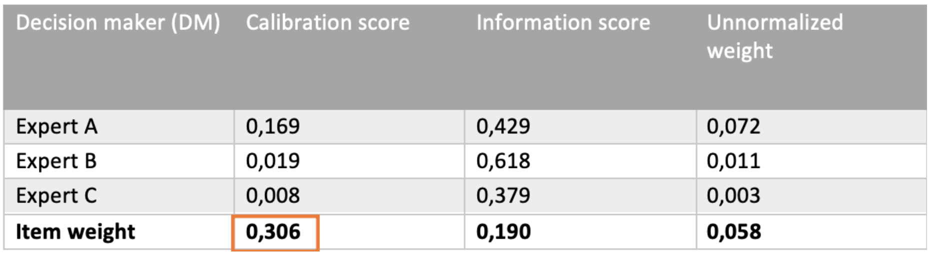
When looking at the year 2017, the calibration score of the Item weight decision maker is 0,131, so it is statistically accurate. The information score is 0,478, so it is reasonably informative. The unnormalized weight results in 0,063 which is a reasonably higher compared to year 2012.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,018	0,348	0,006
Expert B	0,083	0,868	0,072
Expert C	0,002	0,297	0,001
Item weight	**0,131**	**0,478**	**0,063**

Table 29: Item weight decision maker for year 2017 using method 1

When looking at the year 2021, shown in table 30, the calibration score of the Item weight decision maker is 0,306, so it is statistically accurate. The information score is 0,190 which is the lowest information score compared to years 2012 and 2017, meaning it is least informative. The calibration and information score lead to an unnormalized weight of 0,058 which is quite low because of the low information score.



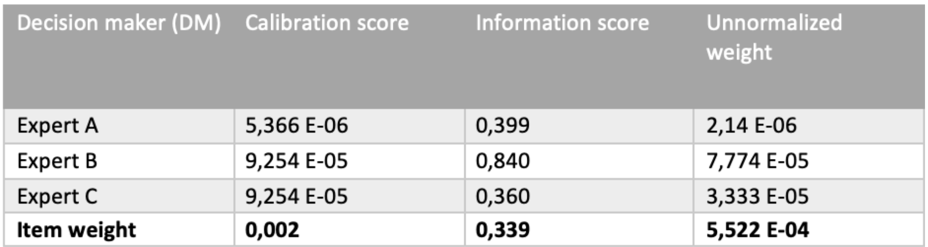
Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	0,169	0,429	0,072
Expert B	0,019	0,618	0,011
Expert C	0,008	0,379	0,003
Item weight	**0,306**	**0,190**	**0,058**

Table 30: Item weight decision maker for year 2021 using method 1

Overall, the highest unnormalized weight is obtained in year 2017 with a value of 0,063, which is still reasonably low. Furthermore. since strong weights reward good expertise in uncertainty quantification [5], expert A is rewarded most in 2021, expert B in 2012 and expert C in 2012.

Item weight decision maker using method 2

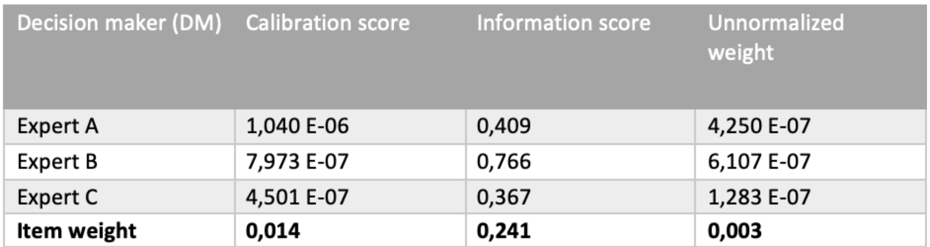
In tables 31 and 32 the outcomes of the Item weight decision maker are shown using method 2. When looking at the year 2017, the calibration score of the Item weight decision maker is 0,002 and the information score 0,339. Again both are quite low which means it is not very informative nor statistically accurate. This leads to a very low unnormalized weight of 5,522 E-04.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	5,366 E-06	0,399	2,14 E-06
Expert B	9,254 E-05	0,840	7,774 E-05
Expert C	9,254 E-05	0,360	3,333 E-05
Item weight	**0,002**	**0,339**	**5,522 E-04**

Table 31: Item weight decision maker for years 2017 using method 2

When looking at the year 2021 using method 2, the calibration score of the Item weight decision maker is 0,014 and the information score 0,241. Also here, both are quite low which means it is not very informative nor statistically accurate. This leads to an unnormalized weight of 0,003 for the Item weight decision maker.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Expert A	1,040 E-06	0,409	4,250 E-07
Expert B	7,973 E-07	0,766	6,107 E-07
Expert C	4,501 E-07	0,367	1,283 E-07
Item weight	**0,014**	**0,241**	**0,003**

Table 32: Item weight decision maker for years 2021 using method 2

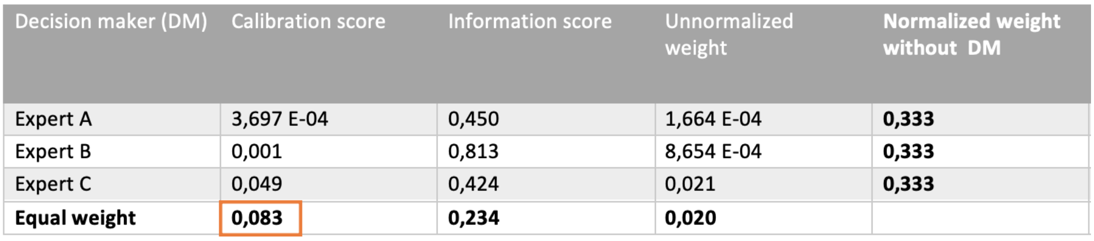
All in all, the unnormalized weights of the Item weight decision maker are very low using method 2. The highest unnormalized weight is obtained in 2021 with a value of 0,003 compared to 0,058 in 2021 when using method 1.

6.4 Equal weight decision maker based on distribution

In this subsection the results of the Equal weight decision maker based on distribution is analysed, which is combining distributions. The normalized weights for each expert are all 0,333 since the Equal weight decision maker assigns equal weight to all experts. The calibration scores with a value higher than 0,05 are highlighted with an orange block.

Equal weight decision maker using method 1

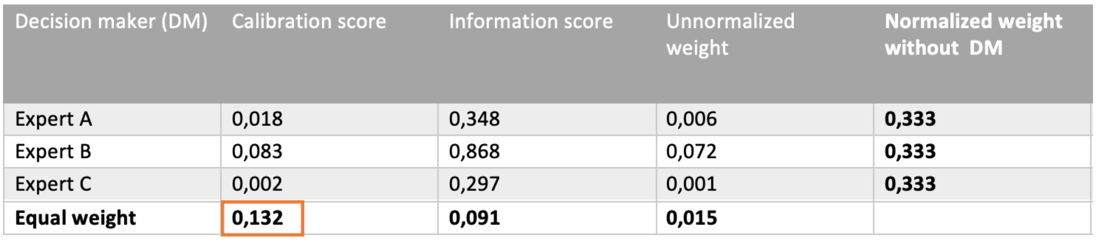
In tables 33, 34 and 35 the results of the Equal weight decision maker are shown. When looking at the year 2012, shown in table 33, the calibration score of the Equal weight decision maker is 0,083 which is statistically accurate. The information score is 0,234, together with the calibration score this leads to an unnormalized weight of 0,020.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	3,697 E-04	0,450	1,664 E-04	**0,333**
Expert B	0,001	0,813	8,654 E-04	**0,333**
Expert C	0,049	0,424	0,021	**0,333**
Equal weight	**0,083**	**0,234**	**0,020**	

Table 33: Equal weight decision maker for year 2012 using method 1

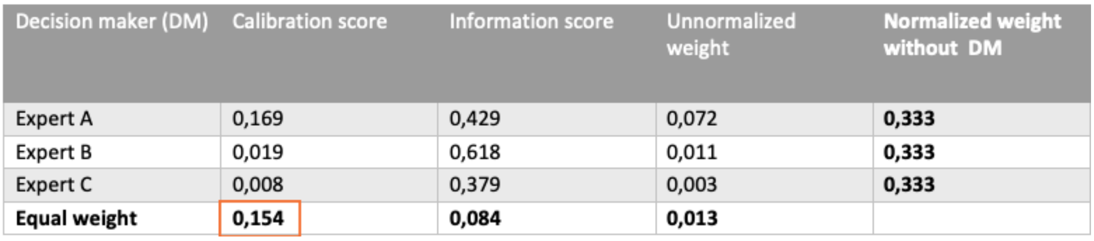
When looking at the year 2012, the calibration score of the Equal weight decision maker is 0,132, which is statistically accurate. The information score is 0,091, so the decision maker is not very informative. Leading to an unnormalized weight of 0,015 for the Equal weight decision maker.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	0,018	0,348	0,006	**0,333**
Expert B	0,083	0,868	0,072	**0,333**
Expert C	0,002	0,297	0,001	**0,333**
Equal weight	**0,132**	**0,091**	**0,015**	

Table 34: Equal weight decision maker for year 2017 using method 1

When looking at the year 2017, shown in table 35, the calibration score of the Equal weight decision maker is 0,154, which is statistically accurate. The information score is 0,084, which means it is not very informative. The unnormalized weight is quite low, namely 0,013, because of the lower informativeness.

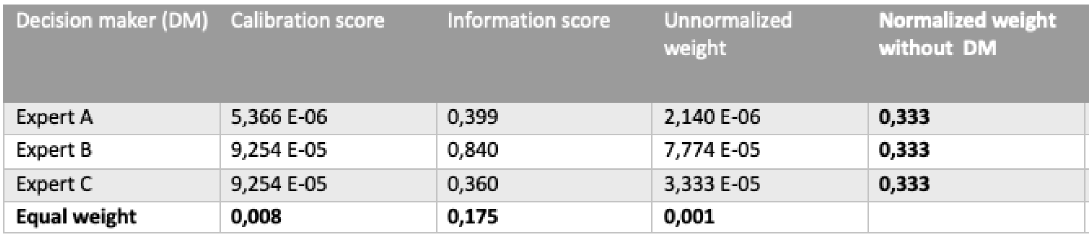


Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	0,169	0,429	0,072	**0,333**
Expert B	0,019	0,618	0,011	**0,333**
Expert C	0,008	0,379	0,003	**0,333**
Equal weight	**0,154**	**0,084**	**0,013**	

Table 35: Equal weight decision maker for year 2021 using method 1

Equal weight decision maker using method 2

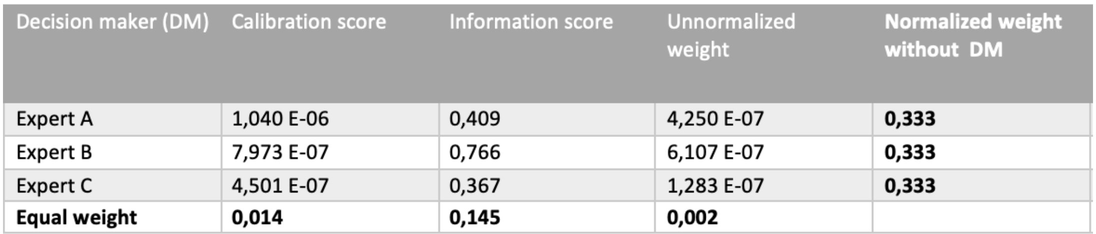
When looking at the year 2017 using method 2, the calibration score of the Equal weight decision maker is 0,008. This calibration score is below the significance level of 0,05, which means it is not very statistically accurate. The information score is 0,175 which is reasonably low too, meaning it is not very informative. The lower information score results in a low unnormalized weight of 0,001.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	**Normalized weight without DM**
Expert A	5,366 E-06	0,399	2,140 E-06	**0,333**
Expert B	9,254 E-05	0,840	7,774 E-05	**0,333**
Expert C	9,254 E-05	0,360	3,333 E-05	**0,333**
Equal weight	**0,008**	**0,175**	**0,001**	

Table 36: Equal weight decision maker for year 2017 using method 2

When looking at the year 2021 using method 2, the calibration score of the Equal weight decision maker is 0,014 and the information score 0,145. Again both are quite low which means it is not very informative nor statistically accurate. In addition, the lower information and calibration score lead to a low unnormalized weight of 0,002 for the Equal weight decision maker.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight	Normalized weight without DM
Expert A	1,040 E-06	0,409	4,250 E-07	**0,333**
Expert B	7,973 E-07	0,766	6,107 E-07	**0,333**
Expert C	4,501 E-07	0,367	1,283 E-07	**0,333**
Equal weight	**0,014**	**0,145**	**0,002**	

Table 37: Equal weight decision maker for years 2021 using method 2

Overall, all unnormalized weights of the Equal weight decision maker based on distribution are relatively low due to low calibration scores. The highest unnormalized weight using method 2 was in 2021 with a value of 0,002. While when using method 1, the unnormalized weight of the decision maker is 0,013. All information scores of the Equal weight decision maker based on distribution are quite low compared to the experts, regardless of what year it been looked at, so it is not very informative.

6.5 Equal weight decision maker based on quantiles

In this subsection the Equal weight decision maker based on quantiles (EWDM_2) is elaborated on. As discussed in section 4.1 the Equal weight decision maker based on quantiles does not take uncertainty into account. In addition, instead of combining distributions it combines the 50% percentiles of the estimations. The values of the EWDM_2 are easily computed by equation:

$$EWDM_2 = \frac{\sum_{i=1}^{N} 50\% \text{ estimated value of } e_i}{\text{total number of experts}} \tag{10}$$

The EWDM_2 values can be found in appendix 11.3, since it does not add a lot of value to discuss the values of the EWDM_2 per election year in here. In section 9 the EWDM_2 is compared to the other decision makers.

7 How well did the polls perform?

In this section the performance of the polls is visualised and analysed. As discussed in section 3, The Netherlands has three polling companies and Spain has six polling companies. One can compare how well each poll did by looking at the total absolute differences of their estimated values per party per year, but also by looking at the unnormalized weights. The absolute total difference of an election year is calculated by the following formula:

$$\text{Absolute total difference} = \sum_{i=1}^{m} \mid (\text{realization } - 50\% \text{ percentile of calibration question } i) \mid \tag{11}$$

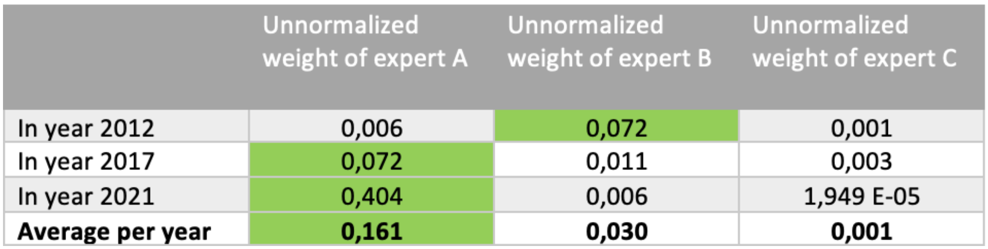
Where m is the total number of calibration questions. The absolute total difference of "all years" is calculating by adding the total differences of the elections years. When looking at the total absolute difference it is desired to be as low as possible, since that means the estimated values lie more close to the realizations of the elections. However, note that this does not hold any uncertainty into account. Since only the differences between the estimated values and realizations are looked at. When looking at the unnormalized weights uncertainty is taken into account. Since not only the estimated values are considered, but the 5% and 95% values are taken into account as well.

7.1 Dutch polls

In this subsection the different Dutch polling companies are assessed. This is done by looking at the unnormalized weights and by looking at the total absolute differences of the forecasted values of each expert with the realization values. The values shown in the tables when looking at the total absolute differences are given in the number of seats a party will receive. The best value in each row is highlighted with a green box.

Comparing experts when looking at the unnormalized weight

From section 6.1 it can be concluded that all highest unnormalized weights were obtained using method 1. In table 38 the unnormalized weights are compared for each expert. When looking at the unnormalized weights per election year, the forecast of expert B was best in election year 2012. While expert A was best in the forecast of election years 2017 and 2021. In addition, when looking at the average of the unnormalized weights per expert, expert A has the highest weight, namely 0,161.



	Unnormalized weight of expert A	Unnormalized weight of expert B	Unnormalized weight of expert C
In year 2012	0,006	0,072	0,001
In year 2017	0,072	0,011	0,003
In year 2021	0,404	0,006	1,949 E-05
Average per year	**0,161**	**0,030**	**0,001**

Table 38: Differences between unnormalized weights

In figure 12 the values of table 38 are shown visually. Now one can see more clear that expert A has the highest unnormalized weight in year 2021 (shown in grey). In addition, expert A has the highest unnormalized weight in 2017 (shown in orange). Furthermore, the highest average, shown by the yellow line, is obtained by expert A as well. Lastly, it can be seen quite easily that expert C has the lowest unnormalized weights, since the bars are almost zero for every election year.

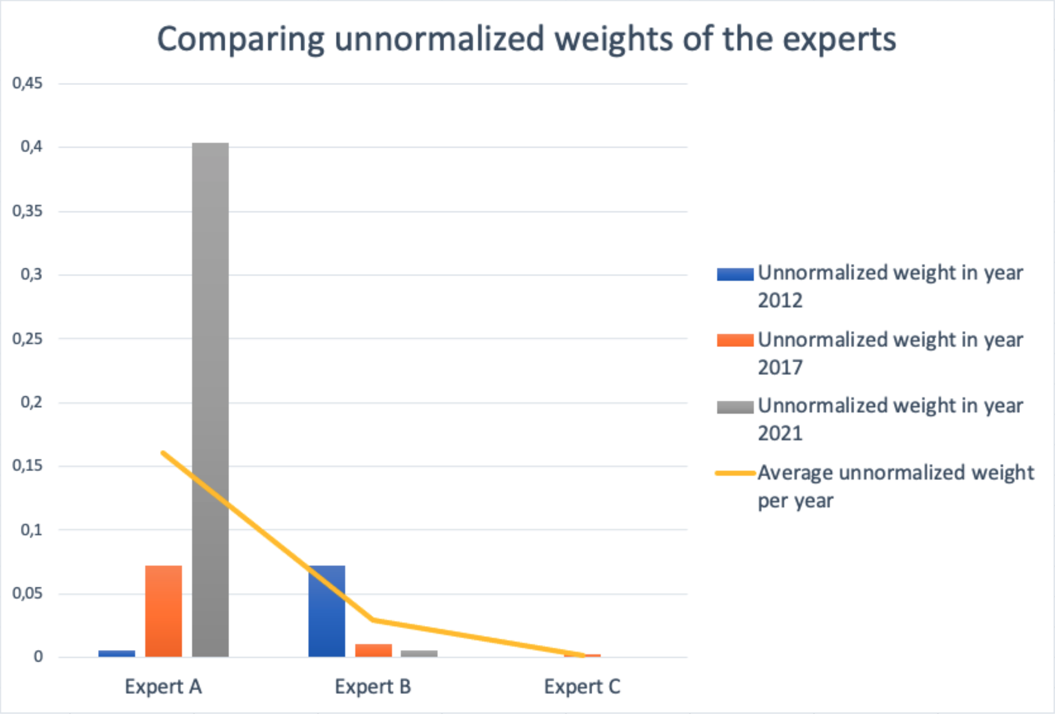


Figure 12: Differences between estimated values and realizations per expert per year

Comparing experts when looking at the absolute total differences
Remember that when looking at the absolute total differences, one wants the values to be as low as possible. In figure 13 it can be seen that expert A (the green line) lies relatively low compared to expert B and C. Meaning its estimates of the elections generally lie more close to the realizations of the elections. Furthermore, it can bee seen that expert A has the lowest value when looking at "all years", this means that when looking at all the estimations made, expert A has forecasted the elections the most accurate to the realization values compared to expert B and C.

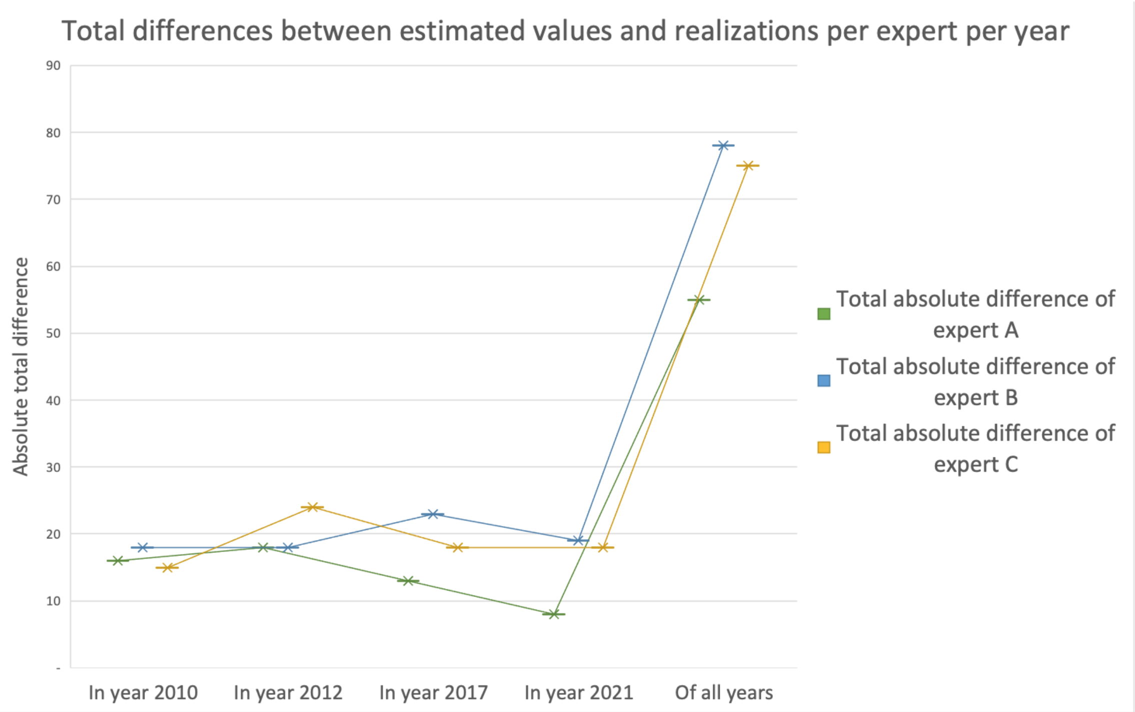
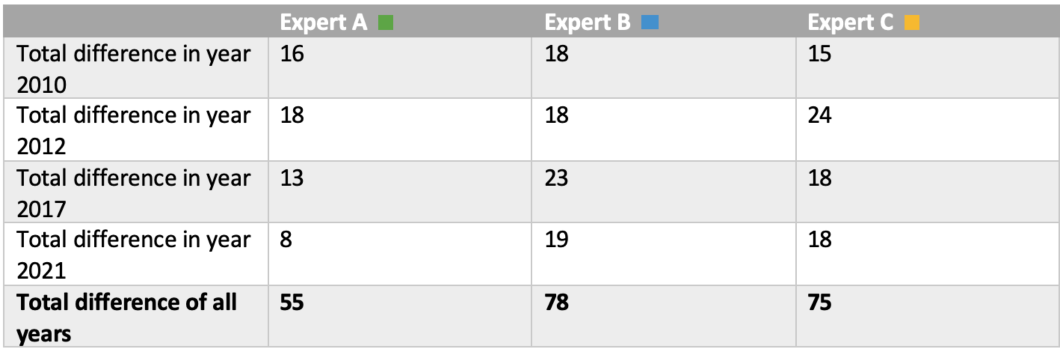


Figure 13: Differences between estimated values and realizations per expert per year

What is described above, is also shown in table 39, where it can be seen that expert A has the lowest absolute total difference when looking at all years, namely with a value of 55 seats. While expert B has a total difference of 78 seats and expert C of 75 seats. When comparing the values of expert A with the other experts, it can be noted that expert C has a lower total difference in year 2010. This means expert C was better in forecasting the election in 2010 than expert A. In year 2012 expert B was as good as expert A in forecasting the elections, they both have a absolute total difference of 18 seats. In 2017 and 2021 expert A had the lowest absolute total differences in each year, thus being the best in forecasting those years. Furthermore, it can be seen from this table that none of the experts show a steady improvement from 2010 onwards. However, from expert A it can be said that since 2012 it does show an improvement in forecasting the elections, since its absolute total differences decreases each year.

All in all, expert A can be seen as the best expert in forecasting the elections when looking at the absolute total difference of all years.



	Expert A 🟩	Expert B 🟦	Expert C 🟨
Total difference in year 2010	16	18	15
Total difference in year 2012	18	18	24
Total difference in year 2017	13	23	18
Total difference in year 2021	8	19	18
Total difference of all years	**55**	**78**	**75**

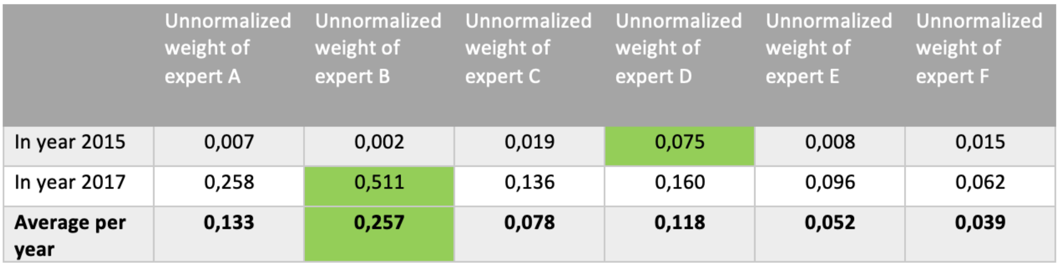
Table 39: Differences between estimated values and realizations per expert per year

7.2 Spanish polls

In this subsection the different Spanish polling companies are assessed. This is done by looking at the unnormalized weights and by looking at the total absolute differences of the forecasted values of each expert with the realization values. The values shown in the tables when looking at the total absolute differences are given in the the percentage of number of votes a party will receive. The best value in each row is highlighted with a green box.

Comparing experts when looking at the unnormalized weight

From section 5.1 it can be concluded that all highest unnormalized weights were obtained using method 1. In table 40 the unnormalized weights are compared for each expert. When looking at the unnormalized weights per election year, expert D was best in forecasting the elections of 2015. While in 2017 expert B was best. In addition when looking at the average unnormalized weight per year, expert B was best with an unnormalized weight of 0,257.



	Unnormalized weight of expert A	Unnormalized weight of expert B	Unnormalized weight of expert C	Unnormalized weight of expert D	Unnormalized weight of expert E	Unnormalized weight of expert F
In year 2015	0,007	0,002	0,019	0,075	0,008	0,015
In year 2017	0,258	0,511	0,136	0,160	0,096	0,062
Average per year	**0,133**	**0,257**	**0,078**	**0,118**	**0,052**	**0,039**

Table 40: Differences between unnormalized weights

In figure 14 the values of table 40 are shown visually. Now one can see more clear that expert B achieved the highest unnormalized weight in 2017 (shown in orange), while expert D in 2015 (shown in blue). In addition, the highest average of unnormalized weights, shown by the grey line, is obtained as well by expert B.

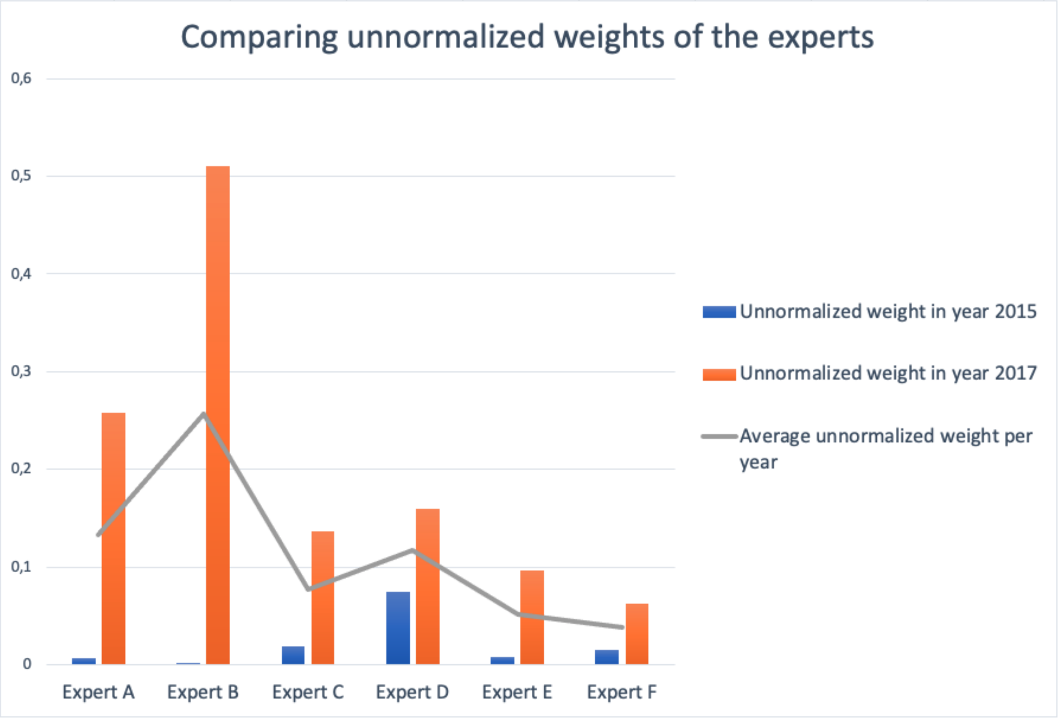


Figure 14: Differences between estimated values and realizations per expert per year

Comparing experts when looking at the absolute total differences

In figure 15 the total differences between the estimated values in each election year and the realizations of the election year can be seen. Furthermore, the total difference of all the years is showed in the most right part of figure 15. It can be seen that expert B lies relatively high each year compared to the others, this means it has done quite poor in predicting the elections. When looking at the figure it seems that expert A and expert E reach the lowest values in respectively years 2015 and 2017. Regarding "all years" they are also reasonably close, however expert E achieves the lowest total difference meaning it has done relatively well predicting the elections.

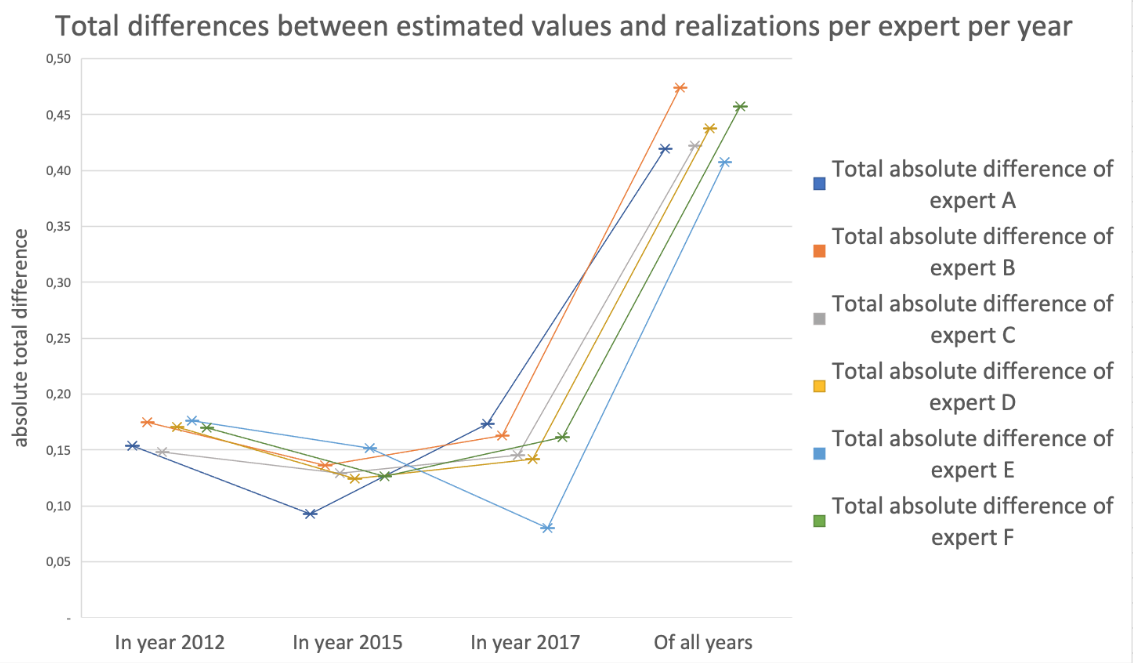
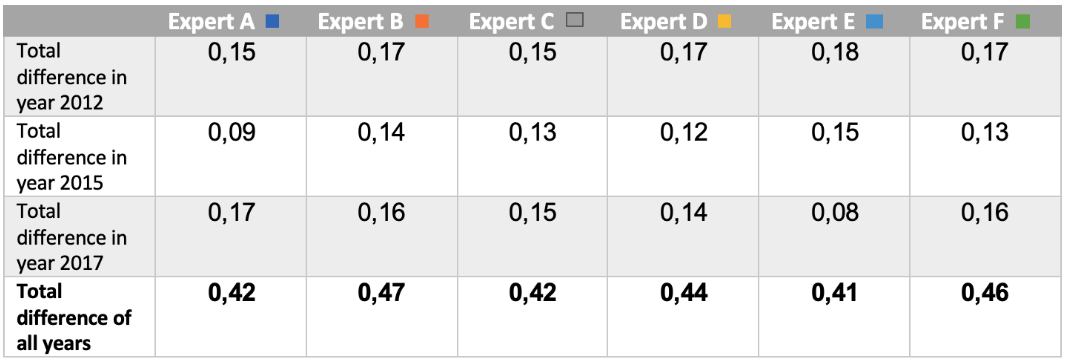


Figure 15: Differences between estimated values and realizations per expert per year

What is described above, is also shown in table 41, where it can be seen that expert E has the lowest total difference with the realization when looking at all years, namely with a value of 0,41. While expert B has the highest total difference with a value of 0,47. Furthermore, it can be noted that expert E is the only expert who shows an improvement in forecasting since year 2012, since the absolute total difference per year decreases. When only looking at the absolute total difference in 2012 expert A and expert C are the best in forecasting. When looking at year 2015, expert A was the best in its forecast and in year 2017 expert E was the best in its forecast.

Overall, expert E is the best in forecasting when looking at all the years.



	Expert A 🟦	Expert B 🟧	Expert C ⬜	Expert D 🟨	Expert E 🟦	Expert F 🟩
Total difference in year 2012	0,15	0,17	0,15	0,17	0,18	0,17
Total difference in year 2015	0,09	0,14	0,13	0,12	0,15	0,13
Total difference in year 2017	0,17	0,16	0,15	0,14	0,08	0,16
Total difference of all years	**0,42**	**0,47**	**0,42**	**0,44**	**0,41**	**0,46**

Table 41: Differences between estimated values and realizations per expert per year

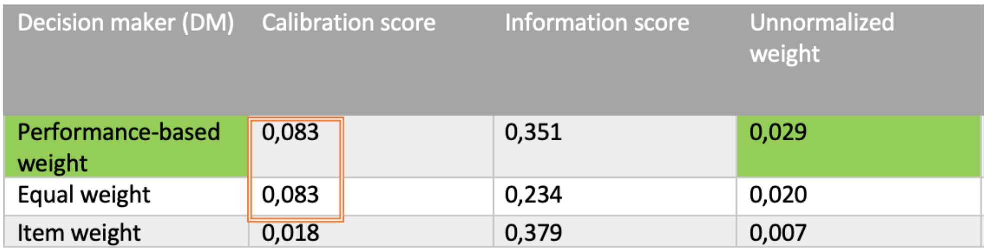
8 Comparing decision makers and method 1 and 2 for The Netherlands

In this section the decision makers of the Dutch data are compared. As discussed in 4.1, there are four decision makers used in this thesis namely the Performance-based decision maker, the Equal weight decision maker based on quantiles and distribution and the Item weight. In addition, the results of method 1 and 2 are compared and evaluated. As mentioned before, there are 11 calibration questions each election year using method 1. While when using method 2 there are 22 calibration questions for election year 2017 and 33 calibration questions for election year 2021.

8.1 Comparing decision makers and method 1 and 2 looking at the unnormalized weights

All decision makers have forecasted values for the elections years. When comparing the decision makers one can look at the unnormalized weights to compare the different decision makers, which are shown in the most right columns in the tables. When looking at the unnormalized weight the uncertainty is taken into account. So not only the estimates are looked at, but also the 5% and 95% values are taken into account. The values which are surrounded by an orange box are significant, when holding a significance level of 0,05 into account.

Comparing decision makers of method 1
As shown in table 42 there are two decision makers with calibration scores higher than 0,05, namely the Performance-based weight decision maker and the Equal weight decision maker. Both have a value of 0,083. However, when looking at the information score the Performance-based weight decision maker is more informative than the Equal weight decision maker. Resulting in an overall higher unnormalized weight for the Performance-based weight decision maker, shown in green in table 42, meaning it is slightly better. In figure 16 this is shown visually in a barplot.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,083	0,351	0,029
Equal weight	0,083	0,234	0,020
Item weight	0,018	0,379	0,007

Table 42: Comparing decision makers for year 2012 using method 1

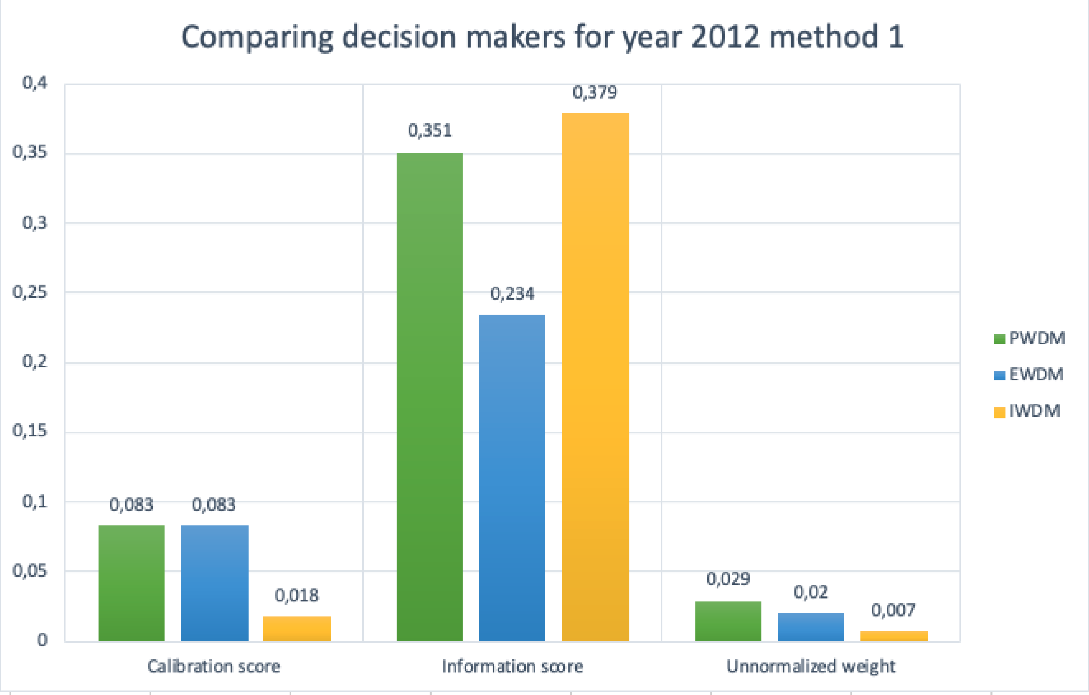
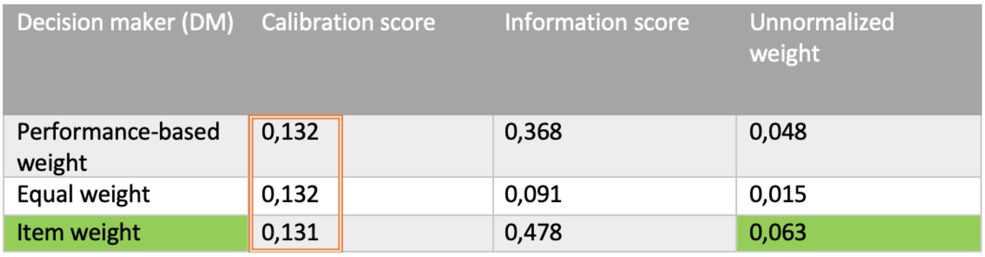


Figure 16: Comparing decision makers for year 2012 using method 1

When looking at table 43, it can be seen that all three decision makers have a calibration score higher than 0,05. This means they are statistically accurate. When comparing the information scores of the three, it is clear that the Item weight is the most informative leading to an overall higher unnormalized weight. In figure 17 this is shown visually in a barplot.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,132	0,368	0,048
Equal weight	0,132	0,091	0,015
Item weight	0,131	0,478	0,063

Table 43: Comparing decision makers for year 2017 using method 1

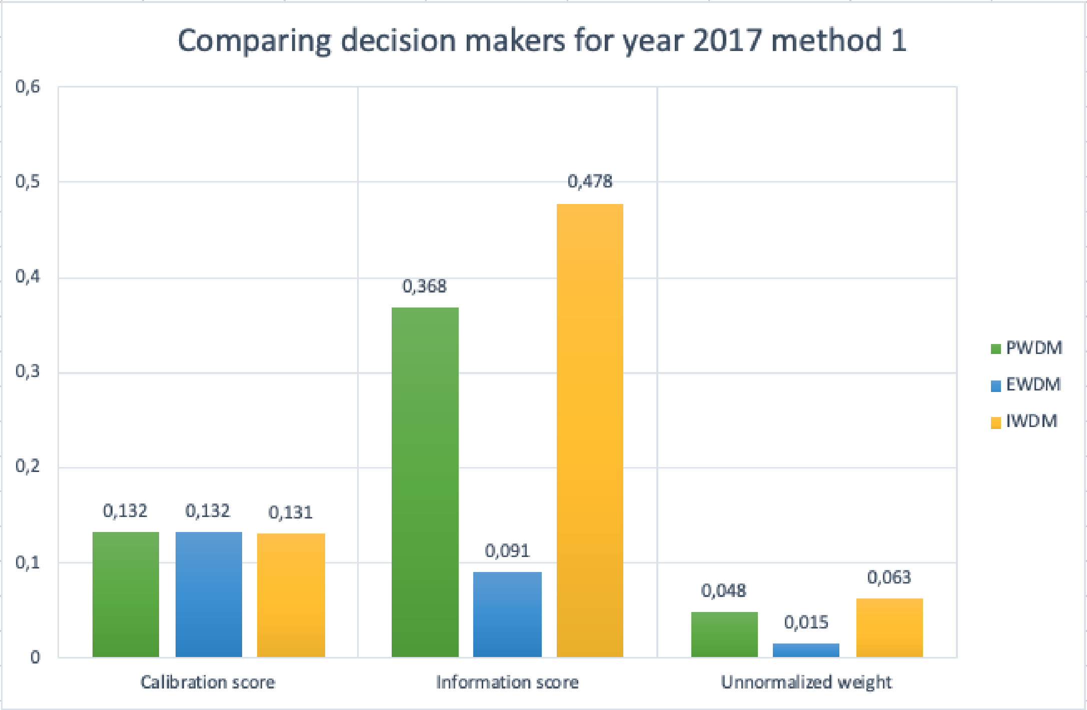
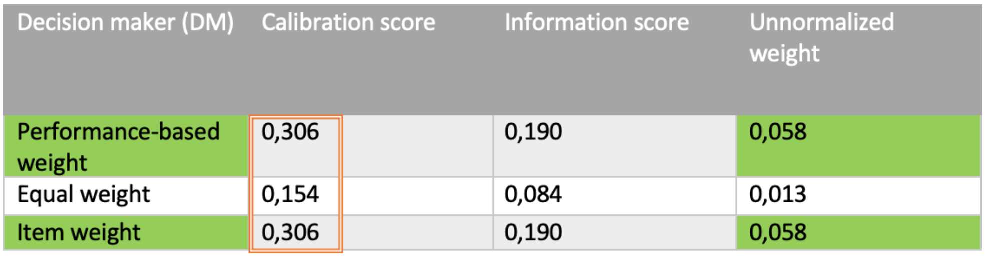


Figure 17: Comparing decision makers for year 2017 using method 1

As shown in table 44 all three experts have a calibration score higher than 0,05, which means they are all statistically accurate. When looking at the information score the Performance-based weight decision maker and the Item weight decision maker have the same highest values thus being evenly informative and leading to the same unnormalized weight. In figure 18 this is shown visually in a barplot.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,306	0,190	0,058
Equal weight	0,154	0,084	0,013
Item weight	0,306	0,190	0,058

Table 44: Comparing decision makers for year 2021 using method 1

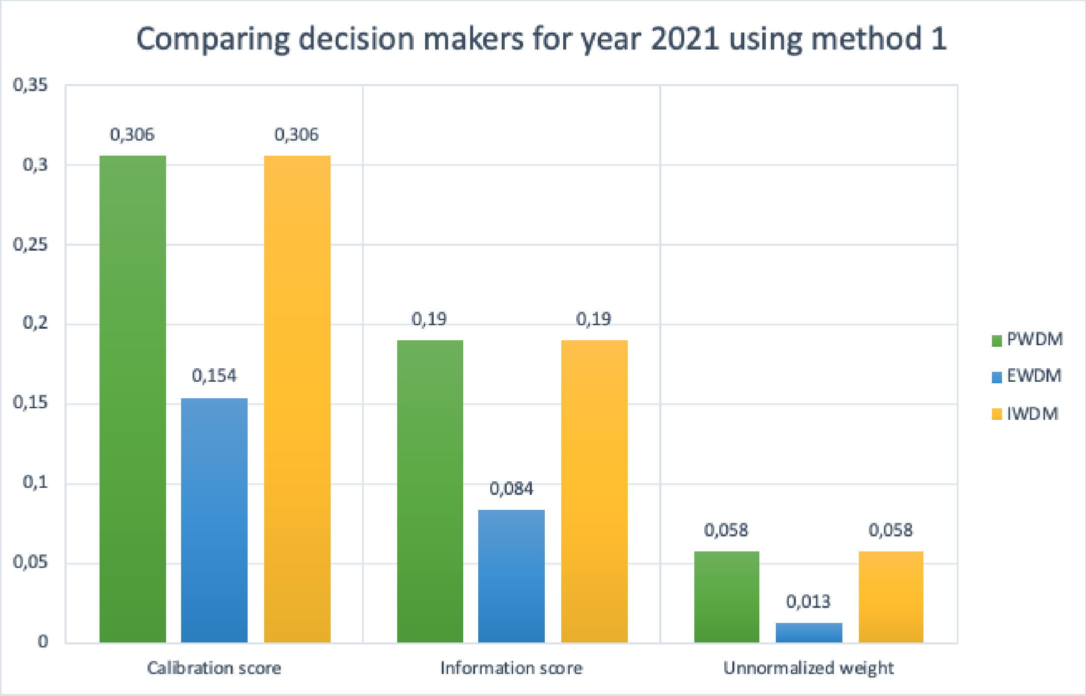
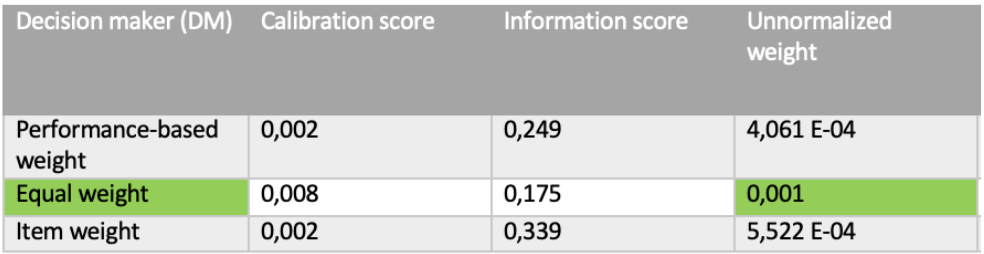


Figure 18: Comparing decision makers for year 2021 using method 1

To conclude this subsection, when using method 1 the Performance-based weight decision maker and the Item weight decision maker are better than the Equal weight decision maker. In addition, it can be noted that in all years the Performance-based weight decision maker has a significant calibration score while the other decision makers do not in year 2012. The highest unnormalized weight was achieved by the Item weight decision maker in year 2017.

Comparing decision makers of method 2

When using method 2 the decision makers of 2017 are shown in table 45. Note that now 22 calibration questions have been assessed and there are no decision makers with calibration scores higher than 0,05. Which means none of them are statistically accurate when holding on to the value significance level of 0,05. When looking at the information score the Item weight decision makers is the most informative compared to the others. However the Equal weight decision maker has overall a higher unnormalized weight, due to its higher calibration score. Consequently, the Equal weight decision maker is shown in green in table 45, meaning it is slightly better than the other decision makers. In figure 19 this is shown visually in a barplot.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,002	0,249	4,061 E-04
Equal weight	0,008	0,175	0,001
Item weight	0,002	0,339	5,522 E-04

Table 45: Comparing decision makers for years 2017 using method 2

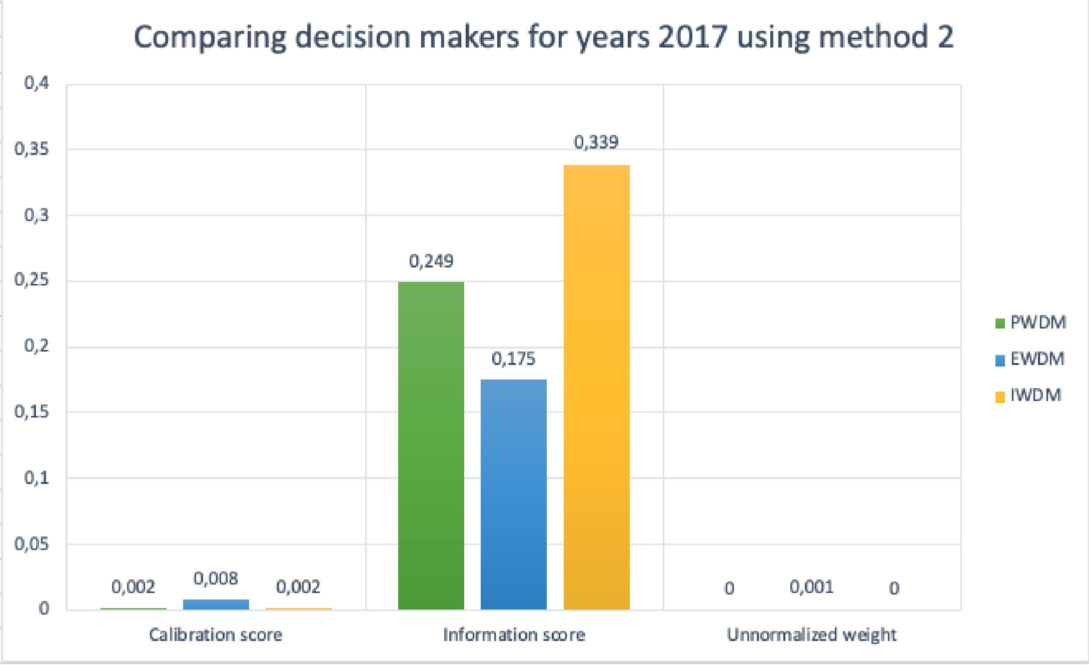
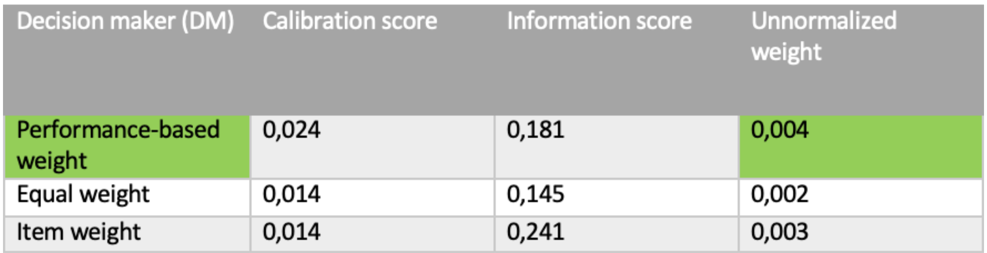


Figure 19: Comparing decision makers for year 2021 using method 2

In table 46 the decision makers are compared for election year 2021. Note that now 33 calibration questions have been assessed. When looking at table 46 it can also be noted that none of the decision makers have a higher calibration score than 0,05. Which means they are not statistically accurate when considering the value 0,05. The Performance-based weight decision maker has the highest calibration score with a relatively high information score compared to the others, thus leading to the highest unnormalized weight, namely with a value of 0,004. The weight of the Performance-based decision maker is still quite low. In figure 20 this is shown visually in a barplot.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,024	0,181	0,004
Equal weight	0,014	0,145	0,002
Item weight	0,014	0,241	0,003

Table 46: Comparing decision makers for years 2021 using method 2

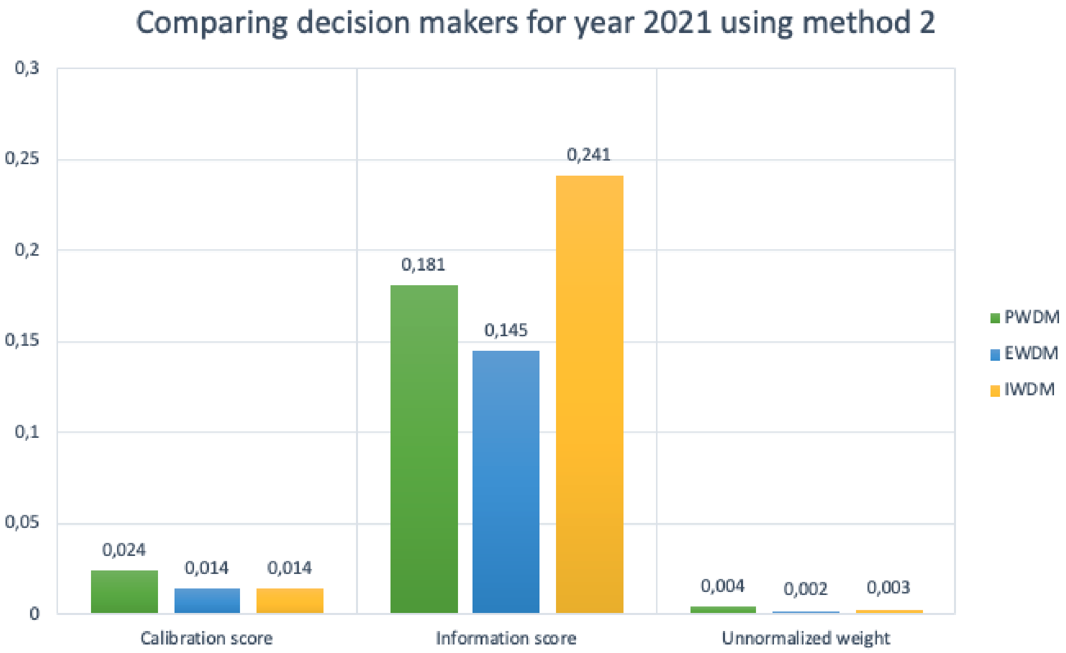
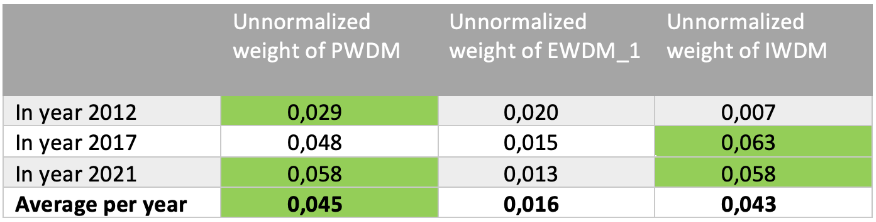


Figure 20: Comparing decision makers for year 2021 using method 2

Conclusion best decision maker and method when looking at the unnormalized weights

Overall, all decision makers performed better when using method 1 than using method 2. Since when using method 1 all unnormalized weights, regardless of what decision maker was used, are higher than those of when using method 2. From this we conclude that method 1 is better in forecasting the elections when looking at the unnormalized weights. Now in table 47 the decision makers' unnormalized weights per election year are compared using method 1. The highest value per election year is shown in green. It can be seen that in election year 2012, the Performance-based decision maker was best in forecasting the election, since it had the highest unnormalized weight with a value of 0,029. In 2017 the Item weight decision maker had the highest unnormalized weight, namely 0,063. The Performance-based decision maker and Item weight decision maker both have the same highest unnormalized weight with a value of 0,058. When looking at the average unnormalized weight per year, the Performance-based weight decision maker had the highest unnormalized weight.

All in all, the Performance-based weight decision maker based on distribution is the best decision maker for Spain when using method 1.



	Unnormalized weight of PWDM	Unnormalized weight of EWDM_1	Unnormalized weight of IWDM
In year 2012	0,029	0,020	0,007
In year 2017	0,048	0,015	0,063
In year 2021	0,058	0,013	0,058
Average per year	**0,045**	**0,016**	**0,043**

Table 47: Comparing decision makers based on their unnormalized weights using method 1

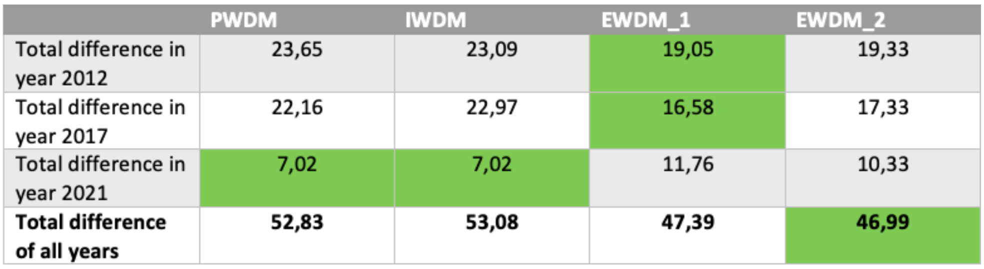
8.2 Comparing decision makers and method 1 and 2 by looking at the total absolute difference

In this subsection the decision makers are compared by looking at the total difference between the forecast of the decision maker and the realization of the election years. The absolute total difference of an election year is calculated by the following formula:

$$\text{Absolute total difference} = \sum_{i=1}^{m} |\ (\text{realization} \ \ - 50\% \text{ percentile of calibration question } i)\ | \tag{12}$$

Where m is the total number of calibration questions. The absolute total difference of "all years" is calculating by adding the total differences of the elections years. When looking at the total absolute difference it is desired to be as low as possible, since that means the estimated values lie more close to the realizations of the elections. When looking at the absolute difference the uncertainty is not taken into account, only the estimates and realizations are looked at. This enables us to see which method performed best and which decision maker. In tables 48 and 49 the decision makers are abbreviated, for the list of abbreviations see section 1. The values are given in the number of seats a party will receive.

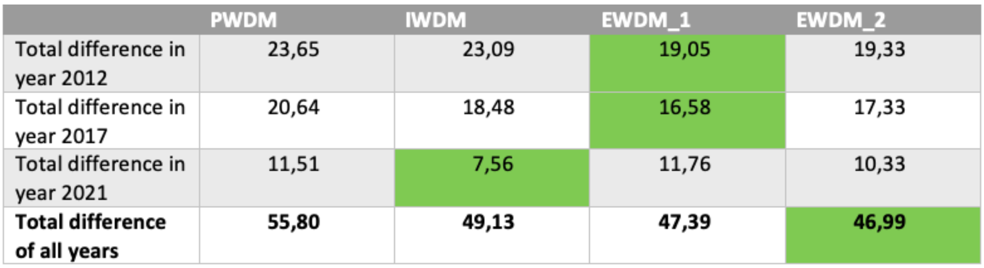
Comparing method 1 by looking at the total absolute difference
In table 48 the total difference between the forecasted values of the decision makers and the realizations using method 1 are shown. It can be seen from this table that in year 2012 and 2017 the Equal weight decision maker based on distribution (EWDM_1) is best in forecasting the elections of 2012 and 2017. In year 2021 the Performance-based weight decision maker and Item weight decision maker were best in forecasting the elections of 2021. When looking at the overall best score, so looking at the "total difference of all years" the Equal weight decision maker based on quantiles (EWDM_2) was best in forecasting, since it has the lowest absolute total difference namely 46,99.



	PWDM	IWDM	EWDM_1	EWDM_2
Total difference in year 2012	23,65	23,09	19,05	19,33
Total difference in year 2017	22,16	22,97	16,58	17,33
Total difference in year 2021	7,02	7,02	11,76	10,33
Total difference of all years	**52,83**	**53,08**	**47,39**	**46,99**

Table 48: Total difference between forecast of decision makers and realization using method 1

Comparing method 2 by looking at the total absolute difference
For method 2 the absolute total differences between the forecasted values of the decision makers and the realizations are shown in 49. It can be seen that in year 2012 and 2017 the Equal weight decision maker based on distribution (EWDM_1) is best in forecasting the elections of 2012 and 2017. While in year 2021 the Item weight decision maker is best in forecasting elections. Since the Equal weight decision maker based on quantiles (EWDM_2) has relatively low scores each year it is the best forecaster when looking at all years, with a value of 46,99.



	PWDM	IWDM	EWDM_1	EWDM_2
Total difference in year 2012	23,65	23,09	19,05	19,33
Total difference in year 2017	20,64	18,48	16,58	17,33
Total difference in year 2021	11,51	7,56	11,76	10,33
Total difference of all years	**55,80**	**49,13**	**47,39**	**46,99**

Table 49: Total difference between forecast of decision makers and realization using method 2

Conclusion best decision maker and method when looking at the total absolute differences
To conclude this subsection, when looking at the absolute total difference of all years, there is no difference in the best decision maker between method 1 and 2. Since both for method 1 and method 2 the Equal weight decision maker based on quantiles (EWDM_2) has the lowest total difference of all years. Furthermore, when adding the total differences of all years for method 1 and method 2 respectively get the values 200,29 and 199,31. It appears that method 2 has a lower score, this means method 2 was better for forecasting these elections than method 1. It is quite remarkable that this Equal weight decision maker came out best in forecasting the elections since it does not hold the performance of each expert into account. In addition, according to Roger Cooke the Equal weight decision maker is known to be less informative and less statistically accurate than the performance-based combination of the experts [2]. Also, when noting that using method 2 the experts performed quite poor when considering the calibration scores, it is remarkable that the results of method 2 were more accurate in forecasting the elections than when using method 1, where the experts had higher values for the calibration questions (see section 6). The reason for this could be that one expert is underestimating and the other one is overestimating. So when you aggregate them, you get a better forecast. Now, since the Equal weight decision maker based on quantiles came out best in its forecasts, the just described scenario is probably the case.

9 Comparing decision makers and method 1 and 2 for Spain

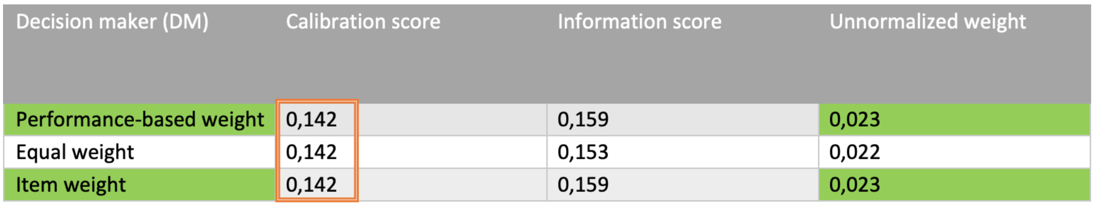
In this section the decision makers of the Spanish data are compared. As discussed in 4.1, there are four decision makers used in this thesis namely the Performance-based decision maker, the Equal weight decision maker based on quantiles and distribution and the Item weight. In addition, the results of method 1 and 2 are compared and evaluated. As mentioned before, there are 7 calibration questions per election year when using method 1. While when using method 2 there are 14 calibration questions.

9.1 Comparing decision makers and method 1 and 2 looking at the unnormalized weights of Spain

All decision makers have forecasted values for the elections years. When comparing the decision makers one can look at the unnormalized weights to compare the different decision makers, which are shown in the most right columns in the tables. When looking at the unnormalized weight the uncertainty is taken into account. So not only the estimates are looked at, but also the 5% and 95% values are taken into account. The values which are surrounded by an orange box are significant, when holding a significance level of 0,05 into account.

Comparing decision makers of method 1 looking at the unnormalized weights
In table 50 the decision makers are shown for year 2015 using method 1. As can be seen in this table, all decision makers have calibration scores higher than 0,05. When looking at

the information scores the Performance-based weight decision maker and Item weight decision maker are the most informative, both have a value of 0,159. Resulting in an overall higher unnormalized weight for the Performance-based weight and Item weight decision makers, shown in green in table 50 meaning these are slightly better. In addition, this is shown visually in figure 21.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,142	0,159	0,023
Equal weight	0,142	0,153	0,022
Item weight	0,142	0,159	0,023

Table 50: Comparing decision makers for year 2015 using method 1

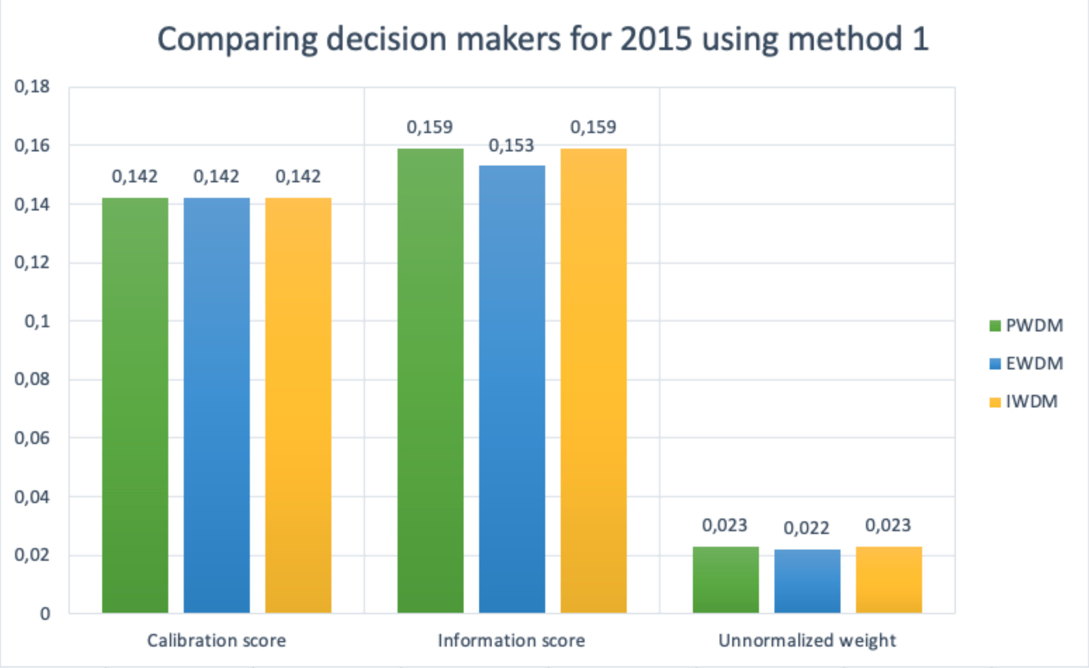
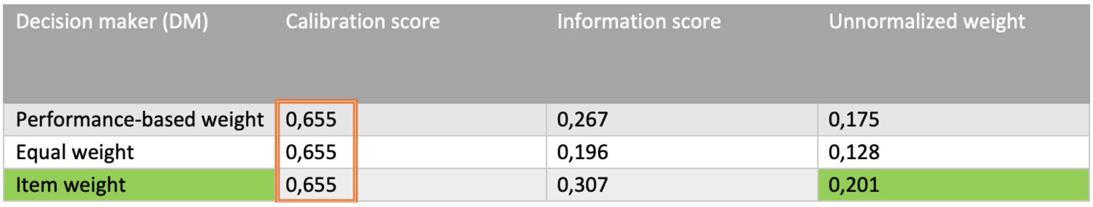


Figure 21: Comparing decision makers for year 2015 using method 1

In table 51 the decision makers are shown for year 2017 using method 1. Again, note that all calibration scores are statistically accurate and all quite high. Now looking at the information scores in 2017, the Item weight decision maker has the highest score, namely 0,307, so it is the most informative. Since all calibration scores are the same this leads to the highest unnormalized weight for the Item weight decision maker, namely one with a value of 0,201. In addition, this is shown visually in figure 22.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,655	0,267	0,175
Equal weight	0,655	0,196	0,128
Item weight	0,655	0,307	0,201

Table 51: Comparing decision makers for year 2017 using method 1

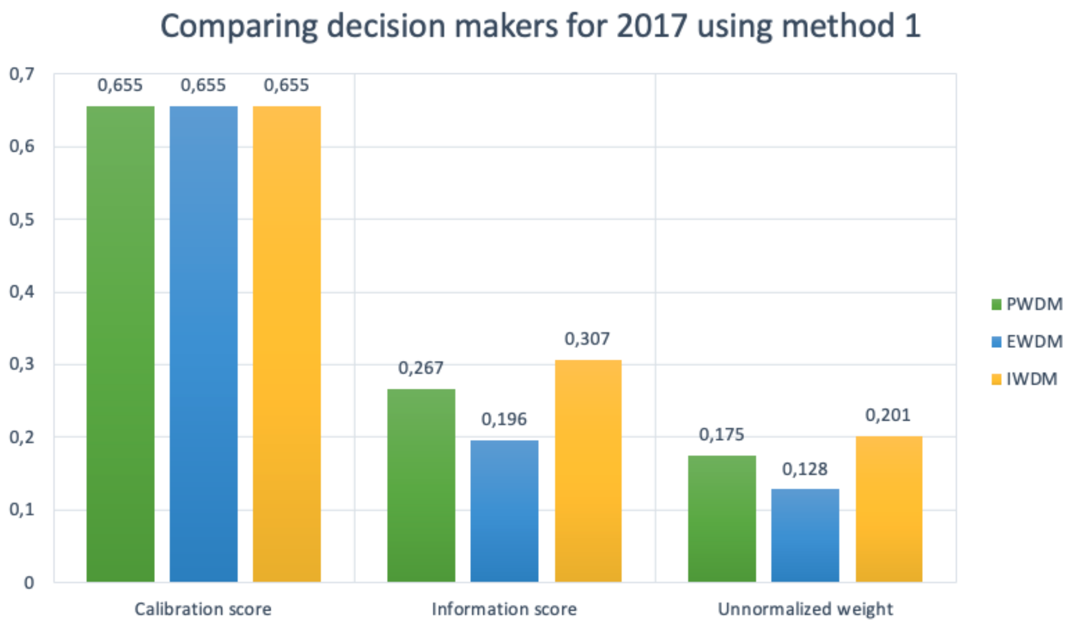
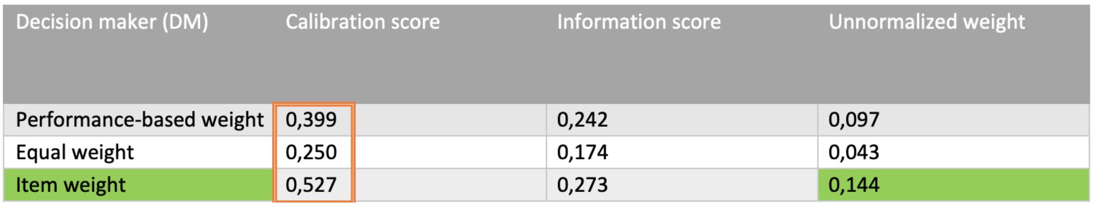


Figure 22: Comparing decision makers for year 2017 using method 1

Comparing decision makers of method 2 looking at the unnormalized weights

In table 52 the decision makers are shown using method 2. Note that now 14 calibration questions have been assessed and that all calibration scores are higher than 0,05, thus being statistically accurate. The Item weight decision maker has the highest calibration score compared to the other decision makers. In addition, it has the highest information score, so being the most informative. This leads to the highest unnormalized weight of 0,144. The lowest calibration and information score is seen by the Equal weight decision maker with respectively the values 0,250 and 0,174. The values of table 52 are shown visually in figure 23.



Decision maker (DM)	Calibration score	Information score	Unnormalized weight
Performance-based weight	0,399	0,242	0,097
Equal weight	0,250	0,174	0,043
Item weight	0,527	0,273	0,144

Table 52: Comparing decision makers for year 2017 using method 2

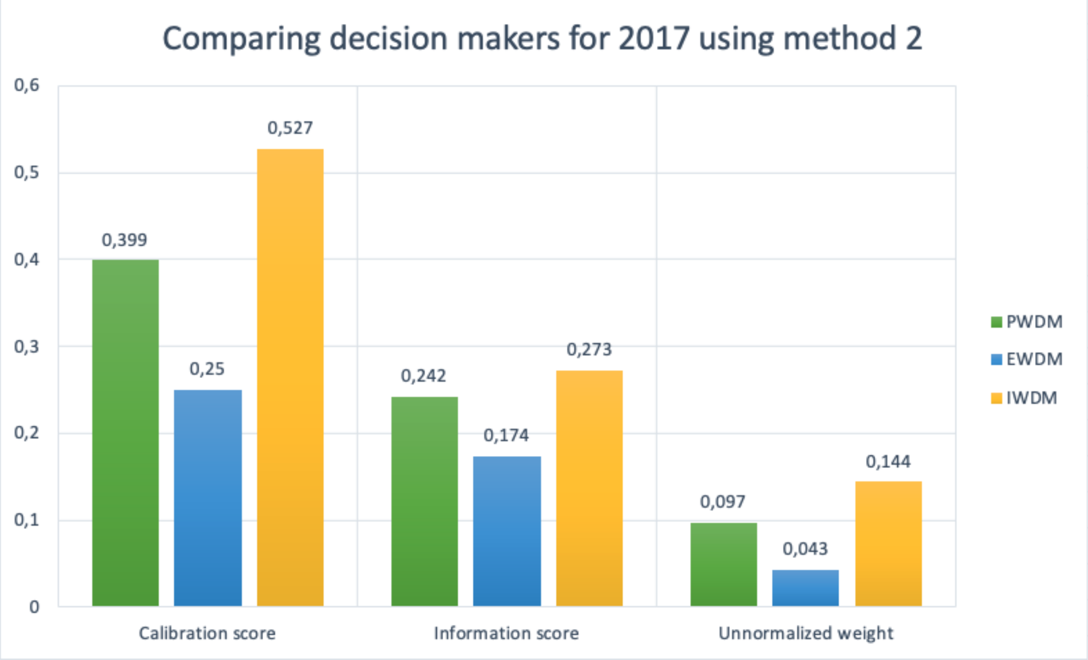
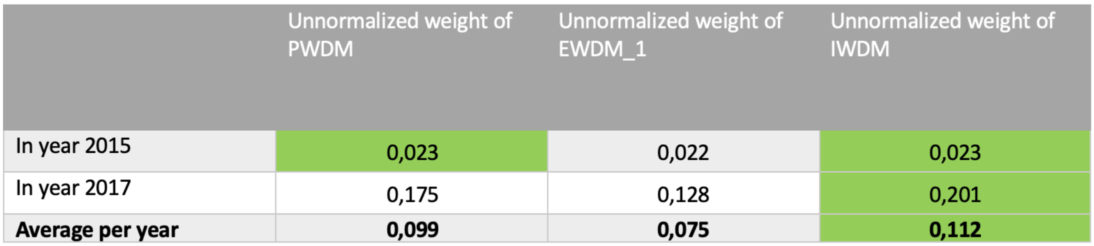


Figure 23: Comparing decision makers for year 2017 using method 2

Conclusion best decision maker when looking at the unnormalized weights
Overall, all decision makers performed better when using method 1 than using method 2. Since all unnormalized weights when using method 1, regardless of what decision maker is used, are lower than those of when using method 2. From this we conclude that method 1 is better in forecasting the elections when looking at the unnormalized weights. Now in table 53 the decision makers' unnormalized weights per election year are compared using method 1. The highest value per election year is shown in green. It can be seen that in election year 2015, the Performance-based decision maker and the Item weight decision maker have the highest unnormalized weights, namely both 0,023. in year 2017 the Item weight decision maker has the highest unnormalized weight. When looking at the average unnormalized weight per year the Item weight decision maker has the highest score with a value of 0,112. All in all, the Item weight decision maker is the best decision maker for Spain when using method 1.



	Unnormalized weight of PWDM	Unnormalized weight of EWDM_1	Unnormalized weight of IWDM
In year 2015	0,023	0,022	0,023
In year 2017	0,175	0,128	0,201
Average per year	**0,099**	**0,075**	**0,112**

Table 53: Comparing decision makers based on their unnormalized weights using method 1

9.2 Comparing decision makers and method 1 and 2 by looking at the total absolute difference

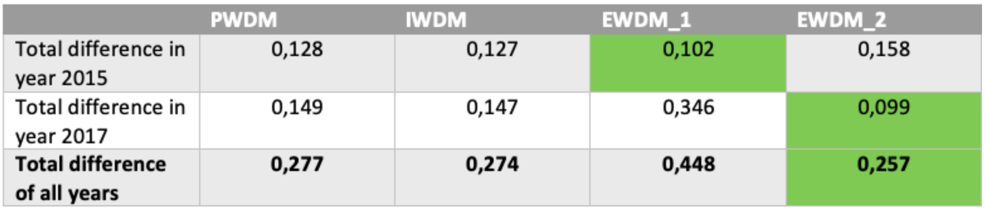
When comparing the decision makers one can look at the total absolute difference between the forecast of the decision maker and the realization of the election years. The absolute total difference of an election year is calculated by the following formula:

$$\text{Absolute total difference} = \sum_{i=1}^{m} \mid (\text{realization } - 50\% \text{ percentile of calibration question } i) \mid \tag{13}$$

Where m is the total number of calibration questions. The absolute total difference of "all years" is calculating by adding the total differences of the elections years. When looking at the total absolute difference it is desired to be as low as possible, since that means the estimated values lie more close to the realizations of the elections. When looking at the absolute difference the uncertainty is not taken into account, only the estimates and realizations are looked at. This enables us to compare the methods and see which method performed best and which decision maker. In tables 54 and 55 the abbreviations of the decision makers are mentioned in the first row. The abbreviations of the decision makers are explained in 1. The values are given in the percentage of number of votes a party is estimated to receive.

Comparing method 1 by looking at the total absolute difference

In table 54 the total difference between the forecasted values of the decision makers and the realizations using method 1 are shown. It can be noted that in year 2015 the Equal weight decision maker based on distribution (EWDM_1) has the lowest total difference, thus forecasting the best in that election year. When looking at year 2017 the Equal weight decision maker based on quantiles (EWDM_2) has performed best. In addition, when looking at the total difference of all years the Equal weight decision maker based on quantiles (EWDM_2) performed best with a value of 0,257.

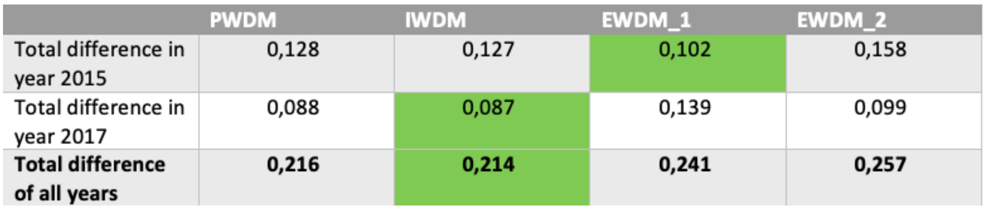


	PWDM	IWDM	EWDM_1	EWDM_2
Total difference in year 2015	0,128	0,127	0,102	0,158
Total difference in year 2017	0,149	0,147	0,346	0,099
Total difference of all years	**0,277**	**0,274**	**0,448**	**0,257**

Table 54: Total difference between forecast of decision makers and realization using method 1

Comparing method 2 by looking at the total absolute difference

For method 2 the differences between the forecasted values of the decision makers and the realizations are shown in 55. In this table it can be seen that in year 2015 the Equal weight decision maker based on distribution (EWDM_1) performed best. In 2017 the Item weight decision maker had the best forecast. Lastly, when looking at all years the Item weigh decision maker performed best as well, with a value of 0,214.



	PWDM	IWDM	EWDM_1	EWDM_2
Total difference in year 2015	0,128	0,127	0,102	0,158
Total difference in year 2017	0,088	0,087	0,139	0,099
Total difference of all years	**0,216**	**0,214**	**0,241**	**0,257**

Table 55: Total difference between forecast of decision makers and realization using method 2

Conclusion best decision maker and method when looking at the total absolute differences
Overall, when looking at the best decision maker method the Equal weight decision maker based on quantiles (EWDM_2) is best for method 1 while the Item weight decision maker was best for method 2. When looking at the "total difference of all years" row the value of the Item weight decision maker is lower than that of the Equal weight decision maker based on quantiles, namely 0,214 instead of 0,257. Thus when looking at the total difference of all years the Item weight decision maker using method 2 is best for forecasting the elections. In addition, when adding the total differences of all years for method 1 and method 2 respectively get the values 1,256 and 0,928. It appears that method 2 has a lower score, this means method 2 was better for forecasting these elections than method 1. The fact that the Item weight decision maker using method 2 was best in forecasting the elections seems logical, since the Item weight decision maker takes the performance of each experts into account, so experts who performed worse are taken less into account in making a forecast of the election. In contrast to the Performance-based decision maker, the Item weight decision maker takes the information scores per expert into account instead of the average, so it distinguishes which expert is more informative per some calibration question. By not taking the average the lower information scores do not have to be taken into account, thus the Item weight decision maker leads to a higher information score in this case.

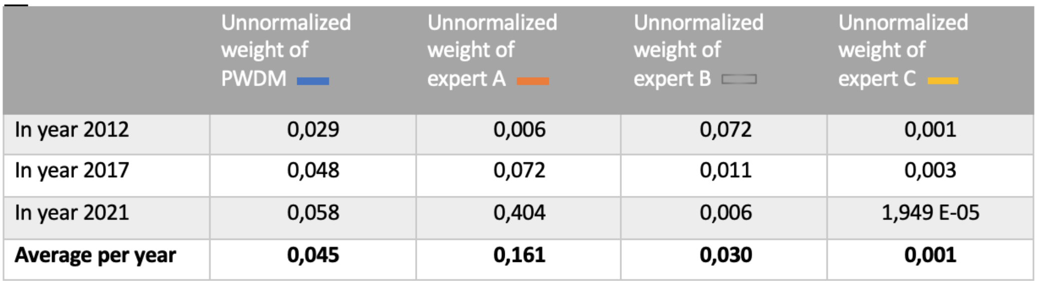
10 Conclusion

In this section, it is evaluated if the aggregation of polls give a better forecast than the polls themselves. Since, when forecasting the elections one would publish only one decision maker with the forecast of the election, an overall “best decision maker” is chosen to be compared with the experts.

10.1 Comparing the best decision maker with the experts for The Netherlands

In this subsection the "best decision maker" is compared with the experts of The Netherlands. First by looking at the unnormalized weights and then by comparing them by looking at the total absolute differences.

Comparing the best decision maker with the experts looking at the unnormalized weights
When looking at the unnormalized weights it is concluded in subsection 8.1 that method 1 is better than method 2 for forecasting the elections. In addition, from subsection 8.1 it is concluded that the Performance-based weight decision maker is the best decision maker for The Netherlands. In table 56 the unnormalized weights of the Performance-based decision maker and the experts are shown per election year. Note that it is different per year whether the decision maker performs better or not. When looking at year 2012, only expert B performs better than the decision maker. While in years 2017 and 2021 only expert A performs better than the decision maker. In addition, when looking at the average unnormalized weight per year only expert A is better than the decision maker.



	Unnormalized weight of PWDM 🟦	Unnormalized weight of expert A 🟧	Unnormalized weight of expert B ▭	Unnormalized weight of expert C 🟨
In year 2012	0,029	0,006	0,072	0,001
In year 2017	0,048	0,072	0,011	0,003
In year 2021	0,058	0,404	0,006	1,949 E-05
Average per year	**0,045**	**0,161**	**0,030**	**0,001**

Table 56: Comparing Performance-based decision maker with experts

In figure 24 the unnormalized weights of the Performance-based weight decision maker and the best expert per election year are shown visually. It can be seen that the unnormalized weights of expert A (orange line) from 2017 onwards are above the unnormalized weights of the Performance-based decision maker (blue line). In contrast, the decision makers blue line is above the other experts lines when looking at the figure from 2017 onwards, meaning the Performance-based decision makers' unnormalized weights were better starting from year 2017.

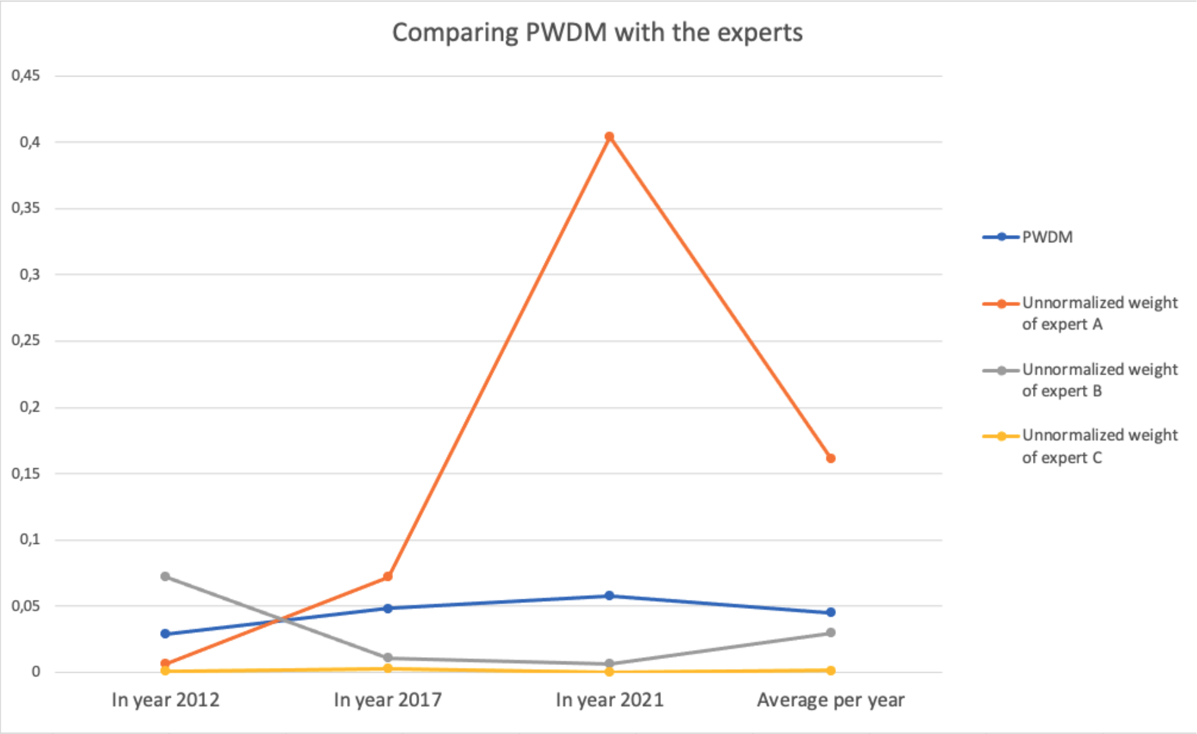


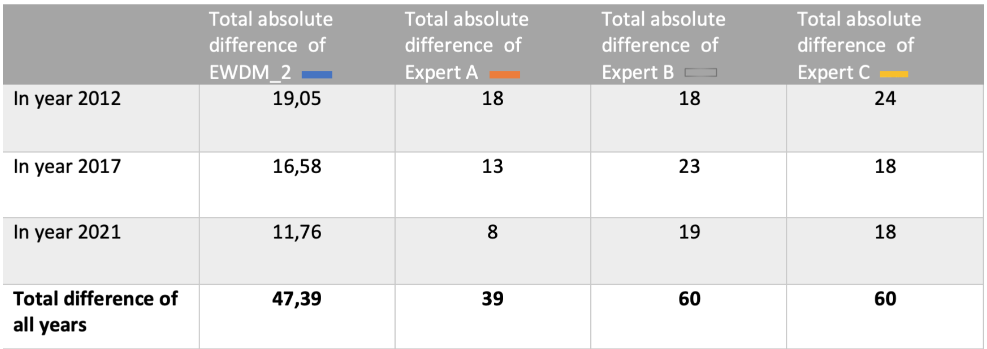
Figure 24: Comparing Performance-based decision maker with experts

All in all, when considering the unnormalized weights of the Performance-based weight decision maker and the experts, the aggregation of the polls is better than 2 out of 3 experts in forecasting the elections. However as discussed, each year there is one expert which performed better when forecasting the elections than the decision maker.

Comparing the best decision maker with the experts looking at the absolute total difference

When looking at the total absolute differences, it could be concluded from subsection 8.2 that method 2 is the best method for forecasting the elections, since the total difference of all years added was lower than for method 1. The total absolute difference is desired to be as low as possible, since this means the estimate lies more close to the realization. Furthermore, regarding the data of The Netherlands the Equal weight decision maker based on quantiles (EWDM_2) scored best with a score of 46,99.

In table 57 the total absolute difference to the realization per election year can be seen for the Equal weight decision maker based on quantiles and the experts. Note that in year 2012 the decision maker scored worse than expert A and B, since the absolute total difference is desired to be as low as possible. While in 2017 and 2021 it only scored worse than expert A. In addition, when looking at the total difference of all years when looking at all years it only scored worse than expert A. Consequently, expert A makes a better overall forecast than the Equal weight decision maker based on quantiles. This is quite surprising since, as discussed in section 4 the Equal weight decision maker based on quantiles is in general not a very good decision maker since it does not hold the performance of the experts into account.



	Total absolute difference of EWDM_2 🟦	Total absolute difference of Expert A 🟧	Total absolute difference of Expert B ▭	Total absolute difference of Expert C 🟨
In year 2012	19,05	18	18	24
In year 2017	16,58	13	23	18
In year 2021	11,76	8	19	18
Total difference of all years	**47,39**	**39**	**60**	**60**

Table 57: Comparing EWDM_2 with experts of The Netherlands

In figure 25 the values of table 57 are shown. It can be seen that expert A, shown in orange, lies below all values of the Equal weight decision maker based on quantiles shown in blue. In addition, it can be seen that the decision maker (blue line) lies almost below all the other experts their values.

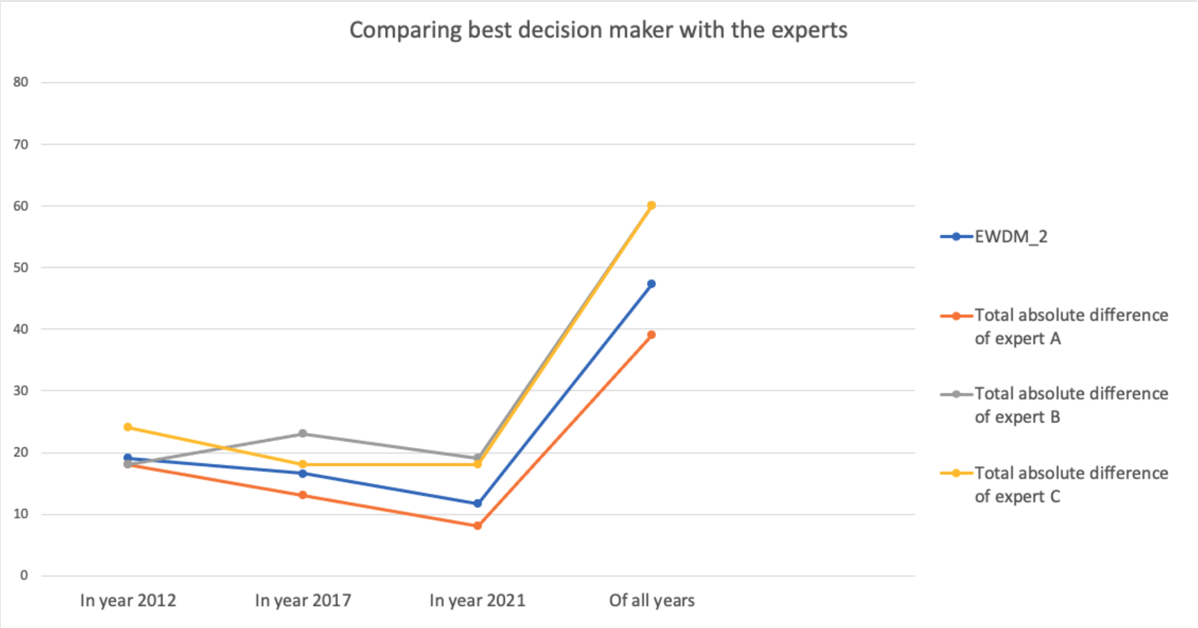


Figure 25: Comparing EWDM_2 with experts of The Netherlands

All in all, as explained in the last paragraph of subsection 8.1 it is quite surprising that the Equal weight decision maker based on quantiles was the best decision maker for the Dutch data.

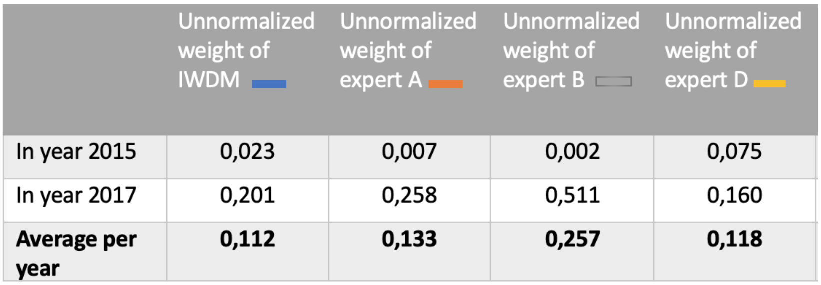
A possible reason for this could be that most of the experts are under and overestimating in their forecasts. However, it is clear from figure 25 that expert A is not, since expert A has lower total absolute differences for each election year than the Equal weight decision maker based on quantiles. When regarding the question if the aggregation of the polls is better or worse in forecasting the elections than the experts, it seems that expert A is better in forecasting the elections than the aggregation of the polls. However, experts B and C are overall worse than the decision maker. So the aggregation of polls is better than 2 out of 3 polls.

10.2 Comparing the best decision maker with the experts for Spain

In this subsection the "best decision maker" is compared with the experts of Spain. First by looking at the unnormalized weights and then by comparing them by looking at the total absolute differences. In the figures in this section the three best experts with the highest average of unnormalized weights or lowest total differences have been chosen to be visualised, since when visualising all experts it would get chaotic. The table with the values of the best decision maker and all the experts can be seen in the appendix 11.4.

Comparing the best decision maker with the experts looking at the unnormalized weights

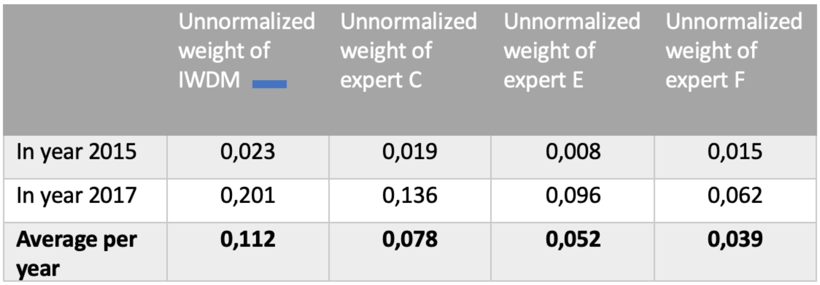
When looking at the unnormalized weights it is concluded in subsection 9.1 that method 1 is better than method 2 for forecasting the elections. In addition, from subsection 9.1 it is concluded that the Item weight decision maker is the best decision maker. In table 58 the unnormalized weights of the Item weight decision maker and the three best expert are shown. It can be seen that in year 2015 only expert D was better than the decision maker. While in 2017 experts A and B have better unnormalized weights than the decision maker.



	Unnormalized weight of IWDM 🟦	Unnormalized weight of expert A 🟧	Unnormalized weight of expert B ▭	Unnormalized weight of expert D 🟨
In year 2015	0,023	0,007	0,002	0,075
In year 2017	0,201	0,258	0,511	0,160
Average per year	**0,112**	**0,133**	**0,257**	**0,118**

Table 58: Comparing IWDM with the three best experts

When considering the other experts, the Item weight decision maker is better than all the experts. This can be seen in table 59, where all values of the decision maker are higher than of the experts C, E and F.



	Unnormalized weight of IWDM 🟦	Unnormalized weight of expert C	Unnormalized weight of expert E	Unnormalized weight of expert F
In year 2015	0,023	0,019	0,008	0,015
In year 2017	0,201	0,136	0,096	0,062
Average per year	**0,112**	**0,078**	**0,052**	**0,039**

Table 59: Comparing IWDM with the other experts

In figure 26 the Item weight decision maker is compared with the three best experts when looking at the highest averages of unnormalized weights. It can be seen that, the decision maker lies between the experts their values depending on what year you look at.

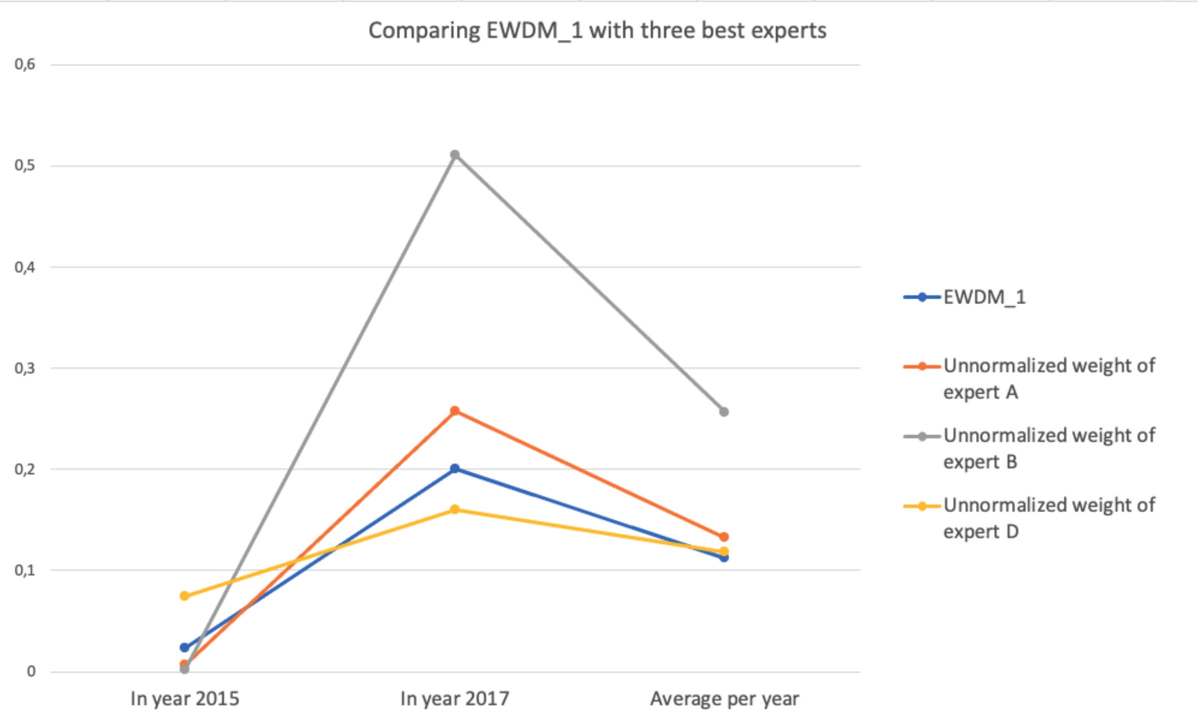


Figure 26: Comparing EWDM_1 with the three best experts

So when considering the unnormalized weights, the aggregation of the polls is better in forecasting the elections for 5 out of 6 experts in 2015. In 2017 the aggregation of polls is better than 4 out of 6 experts their forecasts. While when looking at the average of unnormalized weight per year, the decision maker is better than 3 out of 6 experts.

Comparing the best decision maker with the experts looking at the absolute total difference

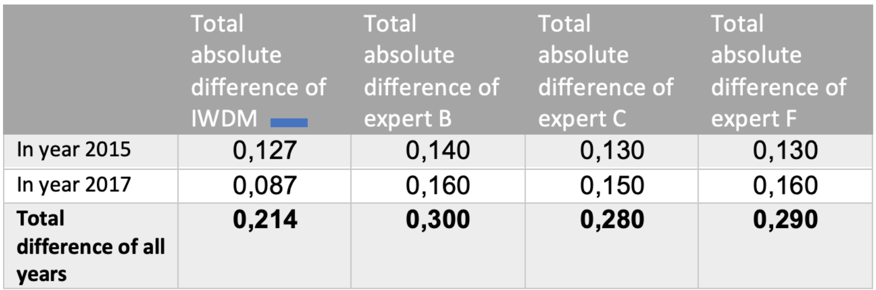
From subsection 9.2 the conclusion was drawn that method 2 is overall the best method for forecasting the elections, since the total difference of all years added was lower than for method 1. Furthermore, regarding the Spanish data the Item weight decision maker forecasted the elections best. In figure 27 the Item weight decision maker and the experts their estimates are compared. The total absolute difference is desired to be as low as possible, since this means the estimate lies more close to the realization.

In table 60 the absolute total difference to the realization per election year is portrayed of the three best experts and the Item weight decision maker. It can be seen that in year 2015 the Item weight decision maker has forecasted the election worse than experts A and D, since its total difference is higher. While in 2017 only expert E was better in forecasting the election than the decision maker. When considering all years the decision maker is better than all experts, since its total absolute difference of all years is the lowest.

	Total absolute difference of IWDM 🟦	Total absolute difference of expert A 🟧	Total absolute difference of expert D ▭	Total absolute difference of expert E 🟨	Total absolute difference of expert F
In year 2015	0,127	0,090	0,120	0,150	0,130
In year 2017	0,087	0,170	0,140	0,080	0,160
Total difference of all years	**0,214**	**0,260**	**0,260**	**0,230**	**0,290**

Table 60: Comparing IWDM with three best experts of Spain

In table 61 the Item weigh decision maker is compared with the other experts who are not visualized in figure 27. Note that all values of the Item weight decision maker are better than experts B, C and F.



	Total absolute difference of IWDM ▬	Total absolute difference of expert B	Total absolute difference of expert C	Total absolute difference of expert F
In year 2015	0,127	0,140	0,130	0,130
In year 2017	0,087	0,160	0,150	0,160
Total difference of all years	**0,214**	**0,300**	**0,280**	**0,290**

Table 61: Comparing IWDM with the other experts

In figure 27 the values of table 60 are shown visually. Note that the decision maker shown by the blue line lies relatively low, meaning it has relatively low total absolute differences. However, in year 2015 the orange line is below the decision maker and in year 2017 the yellow line. When looking at the total absolute difference of all years added the value of the decision maker is relatively a lot lower than the experts.

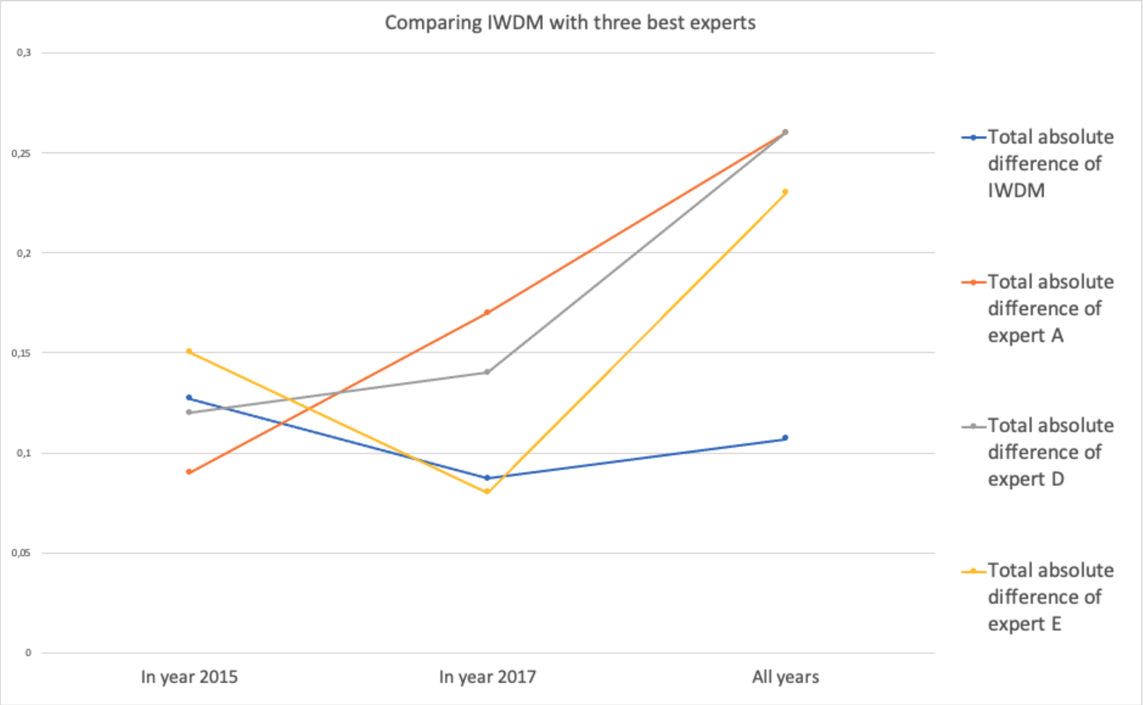


Figure 27: Comparing IWDM with three best experts of Spain

To conclude, in election years 2015 and 2017 the Item weight decision maker is better than 5 out of 6 experts in forecasting the elections. When looking at the absolute total difference of all years added the decision maker is better than all experts. So, it can be concluded that the forecast when aggregating the polls (using the Item weight decision maker) is better than at least 5 out of 6 experts in forecasting the elections. This makes sense, since the Item weight decision maker takes the performance of each experts into account. In contrast to the Performance-based decision maker it takes the information scores per expert into account instead of the average, so it distinguishes which expert is more informative per some calibration question. By not taking the average the lower information scores do not have to be taken into account, thus the Item weight decision maker leads to a higher information score in this case.

10.3 Comparing The Netherlands and Spain

In this section the outcomes of The Netherlands and Spain are compared. Note that Spain has double the amount of polling companies than The Netherlands. Regarding the Classical Model this means it has double the amount of experts. Furthermore, the Dutch polling companies forecasted the amount of seats of 11 parties while the Spanish polling companies forecasted the amount of votes a party will receive of 7 parties.

From subsections 10.1 and 10.2 it can be concluded that for both countries method 1 was better in its forecast when looking at the unnormalized weights. While method 2 was better in its forecast when looking at the absolute total differences. For The Netherlands the Performance-based decision maker was best in forecasting the elections when looking at unnormalized weights. While for Spain the Item weight decision maker was best. When looking at the absolute total differences, the Equal weight decision maker based on quantiles was the best in forecasting the elections. While for Spain the Item weight decision maker was best. Furthermore, for both countries it appeared that, regardless of how you compare, the decision maker was better in forecasting the elections than most of the polls. However, there was always at least one polls which performed better than the decision maker. Meaning the aggregation of the polls is better than most polls, but not all.

11 Discussion

In this section some possible improvements and further research possibilities for this thesis are discussed.

First of all, since by using Excalibur the data had to be copied by hand. This could have caused typing errors or other small mistakes. Although, big typing errors were obvious during evaluating the results since this led to strange results, however the smaller errors could have been missed. It would be best to import the data, so nothing has to be copied by hand, but then another software has to be used. When this research would be expanded it would be possible to write a code in the software R Studio, where large data sets can be imported.

Secondly, the Dutch data was portrayed in "number of seats", which led to less logical results compared to the Spanish data since small alterations had to be made to the data-set in order to put it in Excalibur. For example it was impossible to fill in 0 for the 5% value and 0 for the 50% value in Excalibur, since the 5% value had to be lower than the 50% value. However, in the Dutch data set the number of seats was estimated sometimes at 0 (which is the 50% value) so the 5% value would also be 0 since there cannot be a negative number of seats, but this resulted in an error. This was solved by changing the 50% value into 0,01 seat, but that is of course not possible in real life. In the future it would be best to get the data in a percentage of the number of votes a party will receive instead of the number of seats. Also, it would be more clear if the data of The Netherlands and Spain would be portrayed in the same parameter, so both for example in the number of votes a party will receive.

In addition, for future research more data would be valuable for looking at the different results between method 1 and 2. In my opinion it would make most sense if method 2 would be the best in forecasting the elections in Spain and The Netherlands regardless of how you compare them (by unnormalized weight or absolute total difference). Since, method 2 uses more calibration questions so more about each expert its performance is known. However, when looking at the unnormalized weight it appeared that method 1 was better in forecasting the elections than method 2. This might be due to the relatively low amount of comparable elections years (at most 2 with the Dutch data), so this would be interesting to look at in further research.

Furthermore, there could be looked at more countries than The Netherlands and Spain. For example, the United Kingdom has a British Polling Council which is an association of polling organisations. They publish election polls and are known to value transparency in polling [1].

Since, they have quite a large database of polls, this would be useful and interesting to look at in further research.

Lastly, it is important to note that there are other aggregation methods to aggregate polls. In this thesis the Classical Model by Roger Cooke has been used. The Classical Model uses mathematical aggregation of experts assessments. Other methods using for example behavioral aggregation or quantile aggregation could also be investigated in further research. In addition, there are "mixed methods" such as the Delphi method which could be interesting to look at too [5].

References

[1] BPC (n.d.). About The BPC *British Polling Council.* June 26th, 2021, from https://www.britishpollingcouncil.org

[2] Cooke, R. (2018). Expert Elicitation: Using the Classical Model to Validate Experts' Judgments. *Research Gate.* May 26th, 2021, from https://www.researchgate.net/publication/323724149_Expert_Elicitation_Using_the_Classical_Model_to_Validate_Experts%27_Judgments

[3] De Hond, M (n.d.). Informatie over Peil.nl *Peil.nl.* May, 12th, 2021, from https://home.noties.nl/peil/informatie/

[4] De Hond, M. (2021). Prognose TK2021: Lage opkomst, versplintering zet verder door *Peil.nl.* May 15th, 2021, from https://www.maurice.nl/peilingen/2021/03/16/prognose-tk2021-lage-opkomst-versplintering-zet-verder-door/

[5] DelftX (n.d.). Decision Making Under Uncertainty: Introduction to Structured Expert Judgment. *edX.* April 21st, 2021, from https://learning.edx.org/course/course-v1:DelftX+SEJ1x+1T2021/block-v1:DelftX+SEJ1x+1T2021+type@sequential+block@e1bbe4fbcbde4c419b12a4bf2c741adc/block-

[6] Louwerse, T. (n.d.). Methode Peilingwijzer *Peilingwijzer.* May 8th, 2021, from https://peilingwijzer.tomlouwerse.nl/p/methode.html

[7] Lighttwist Software (n.d.). Excalibur standalone *Lighttwist Software.* 09 June, 2021, from https://lighttwist-software.com/excalibur/

[8] Nane, G. et al. (2021). The Classical Model: The Early Years. *Springer.* May 26th, 2021, from https://link.springer.com/chapter/10.1007/978-3-030-46474-5_9

[9] Nasard, A. (n.d.). About Kantar *Kantar.* June 19th, 2021, from https://www.kantar.com/north-america/about

[10] NOS (2021). Nieuwe update exitpoll Ipsos: VVD grootste partij, D66 na grote winst tweede partij *publisher.* May 8th, 2021, from https://nos.nl/collectie/13860/artikel/2373044-nieuwe-update-exitpoll-ipsos-vvd-grootste-partij-d66-na-grote-winst-tweede-partij

[11] Panagopoulos, C. (2009). Preelection Poll Accuracy in the 2008 General Elections. *Reseach Gate.* March 24th, 2021, from https://www.researchgate.net/publication/229686301

[12] Podhoretz, P. (n.d.). Polls quotes. *BrainyQuote.com.* June 3th, 2021, from https://www.brainyquote.com/quotes/john_podhoretz_484784?src=t_polls

[13] Squire, P. (2012). Why the 1936 Literary Digest Poll Failed (p.23-54). *Cambridge University Press.* May 25th, 2021, from https://www.cambridge.org/core/journals/social-science-history/article/president-landon-and-the-1936-literary-digest-poll/E360C38884D77AA8D71555E7AB6B822C

[14] Tankard, J. (1972). Public Opinion Polling By Newspapers in the Presidential Election Campaign of 1824. *Journalism Quarterly.* May 25th, 2021, from https://journals.sagepub.com/doi/abs/10.1177/107769907204900219

[15] Thomsen, N. (2007). Women's Rights (p.350). *Infobase publishing.* May 25th, 2021

[16] Truchot, D. (jaar). About us *Ipsos.* June 19th, 2021, from https://www.ipsos.com/en/about-us

[17] Tweede Kamer (n.d.). The House of Representatives at work *Tweede Kamer Der Staten Generaal.* June 23th, 2021, from https://www.houseofrepresentatives.nl/how-parliament-works/house-representatives-work

[18] Webb, P. (2003). Election. *Britannica*. May 25th, 2021, from https://www.britannica.com/topic/election-political-science

Appendix

11.1 Technical Appendix

s Empirical probability vector

p Probability vector

m Number of calibration questions

F Cumulative distribution function (CDF) of a chi-squared distribution with 3 degrees of freedom

k overshoot

N total number of experts

Absolute total difference:

$$\text{abs. total diff.} = \sum_{i=1}^{m} \mid (\text{realization} - 50\% \text{ percentile of calibration question } i) \mid \tag{14}$$

Kullback-Leibler divergence of s and p:

$$I(s,p) = s_1 \cdot ln(\frac{s_1}{p_1}) + s_2 \cdot ln(\frac{s_2}{p_2}) + s_3 \cdot ln(\frac{s_3}{p_3}) + s_4 \cdot ln(\frac{s_4}{p_4}) \tag{15}$$

Calibration score:

$$Cal(e) = 1 - F(2 \cdot m \cdot I(s,p)) \tag{16}$$

Intrinsic range:

$$[L^*, U^*] = [L - 0,1 \cdot (U - L), U + 0,1 \cdot (U - L)] \tag{17}$$

Information score:

$$\begin{aligned} I(e) =& 0,05 \cdot ln(\frac{0,05}{q_5 - L^*}) + 0,45 \cdot ln(\frac{0,45}{q_{50} - q_5}) + 0,45 \cdot ln(\frac{0,45}{q_{95} - q_{50}}) \\ &+ 0,05 \cdot ln(\frac{0,05}{U^* - q_{95}}) + ln(U^* - L^*) \end{aligned} \tag{18}$$

Unnormalized weight or combined score:

$$CS(e) = Cal(e) * I(e) \tag{19}$$

Normalized weights:

$$w_i = \frac{CS(e_i)}{\sum_{i=1}^{N} CS(e_i)} \tag{20}$$

Performance-based scoring rule:

$$w_{\alpha,i} = Cal(e_i) * Inf(e_i) * 1\{Cal(e_i) \geq \alpha\} \tag{21}$$

Performance-based Decision Maker:

$$PWDM = w_{e_1} \cdot F_1 + w_{e_2} \cdot F_2 + ... + w_{e_N} \cdot F_N \tag{22}$$

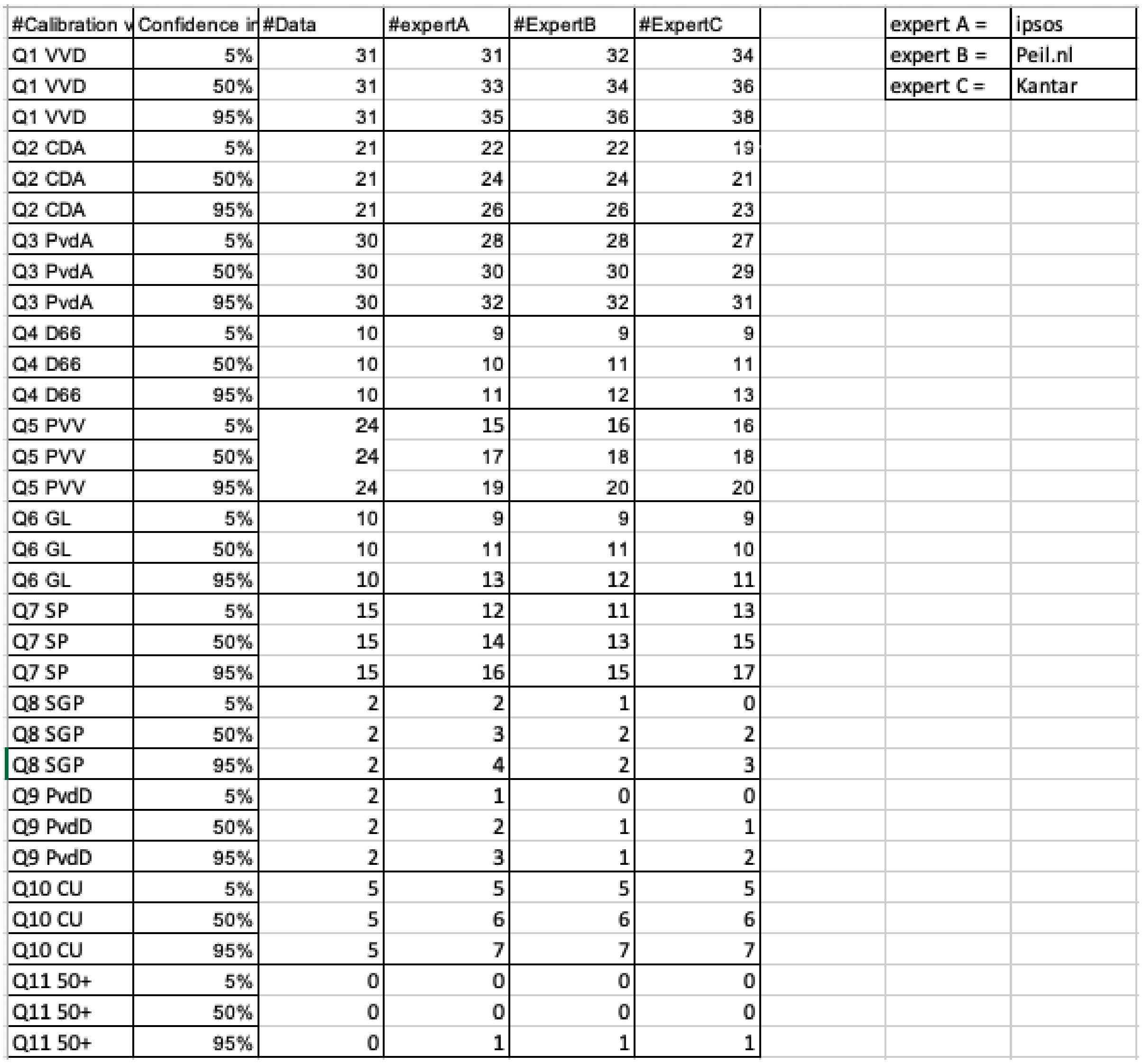
Equal weight Decision Maker based on distribution:

$$EWDM_1 = \frac{1}{N} \cdot F_1 + \frac{1}{N} \cdot F_2 + ... \frac{1}{N} \cdot F_N \tag{23}$$

Equal weight Decision Maker based on quantiles:

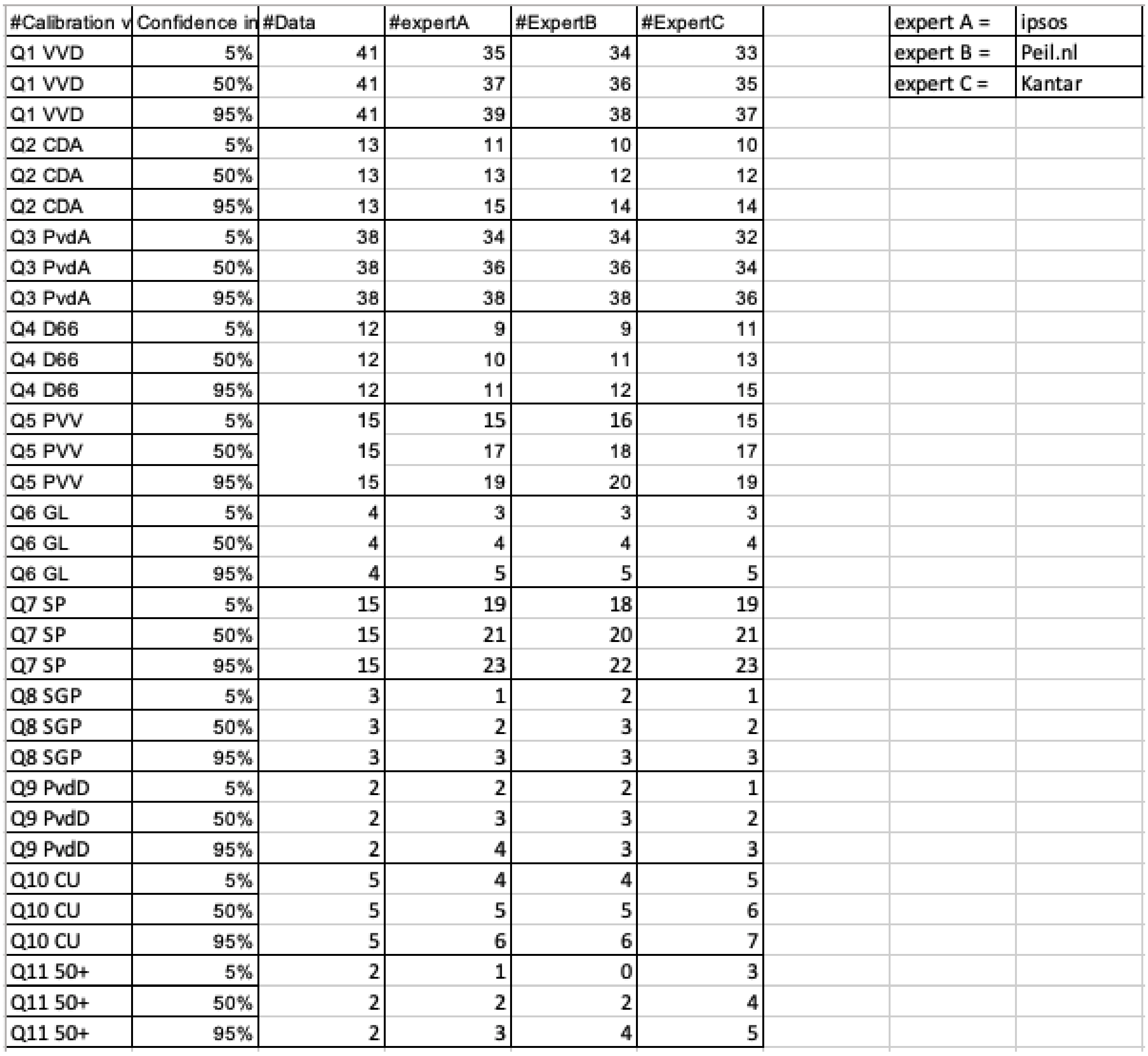
$$EWDM_2 = \frac{\sum_{i=1}^{N} 50\% \text{ estimated value of } e_i}{\text{total number of experts}} \tag{24}$$

11.2 Appendix A: data of The Netherlands



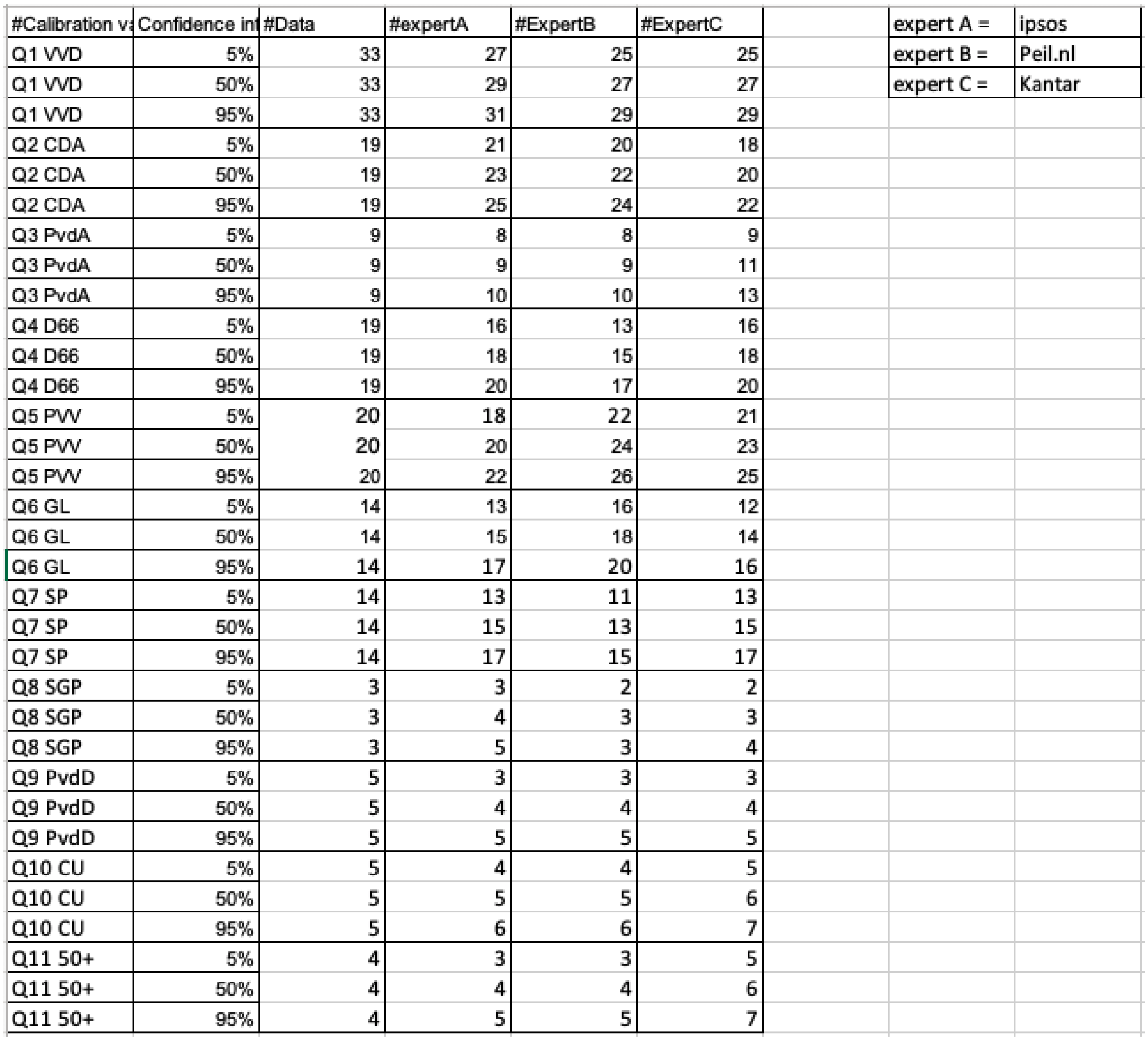
#Calibration v	Confidence ir	#Data	#expertA	#ExpertB	#ExpertC		expert A =	ipsos
Q1 VVD	5%	31	31	32	34		expert B =	Peil.nl
Q1 VVD	50%	31	33	34	36		expert C =	Kantar
Q1 VVD	95%	31	35	36	38			
Q2 CDA	5%	21	22	22	19			
Q2 CDA	50%	21	24	24	21			
Q2 CDA	95%	21	26	26	23			
Q3 PvdA	5%	30	28	28	27			
Q3 PvdA	50%	30	30	30	29			
Q3 PvdA	95%	30	32	32	31			
Q4 D66	5%	10	9	9	9			
Q4 D66	50%	10	10	11	11			
Q4 D66	95%	10	11	12	13			
Q5 PVV	5%	24	15	16	16			
Q5 PVV	50%	24	17	18	18			
Q5 PVV	95%	24	19	20	20			
Q6 GL	5%	10	9	9	9			
Q6 GL	50%	10	11	11	10			
Q6 GL	95%	10	13	12	11			
Q7 SP	5%	15	12	11	13			
Q7 SP	50%	15	14	13	15			
Q7 SP	95%	15	16	15	17			
Q8 SGP	5%	2	2	1	0			
Q8 SGP	50%	2	3	2	2			
Q8 SGP	95%	2	4	2	3			
Q9 PvdD	5%	2	1	0	0			
Q9 PvdD	50%	2	2	1	1			
Q9 PvdD	95%	2	3	1	2			
Q10 CU	5%	5	5	5	5			
Q10 CU	50%	5	6	6	6			
Q10 CU	95%	5	7	7	7			
Q11 50+	5%	0	0	0	0			
Q11 50+	50%	0	0	0	0			
Q11 50+	95%	0	1	1	1			

Table 62: Data of The Netherlands 2010



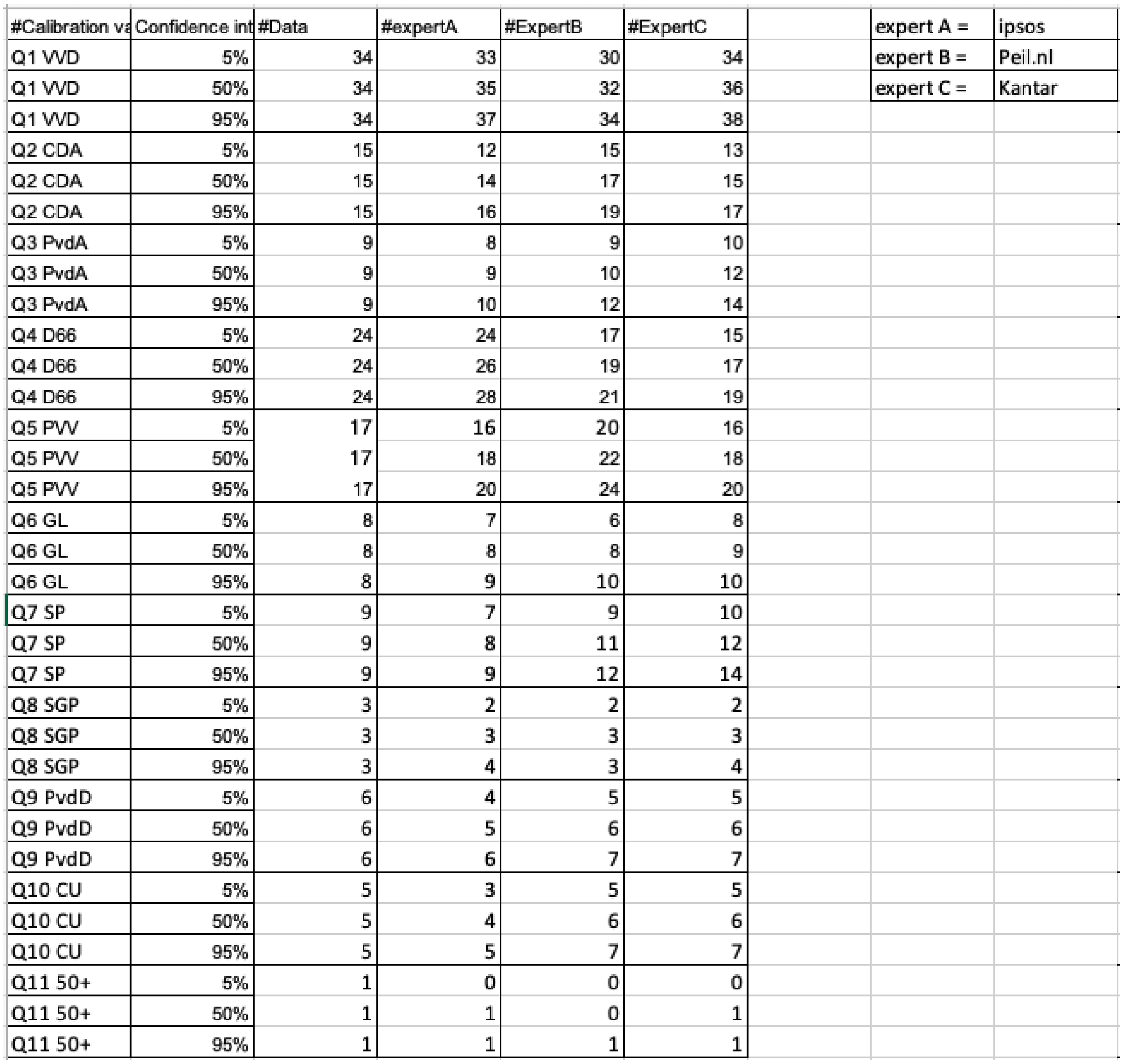
#Calibration v	Confidence in	#Data	#expertA	#ExpertB	#ExpertC		expert A =	ipsos
Q1 VVD	5%	41	35	34	33		expert B =	Peil.nl
Q1 VVD	50%	41	37	36	35		expert C =	Kantar
Q1 VVD	95%	41	39	38	37			
Q2 CDA	5%	13	11	10	10			
Q2 CDA	50%	13	13	12	12			
Q2 CDA	95%	13	15	14	14			
Q3 PvdA	5%	38	34	34	32			
Q3 PvdA	50%	38	36	36	34			
Q3 PvdA	95%	38	38	38	36			
Q4 D66	5%	12	9	9	11			
Q4 D66	50%	12	10	11	13			
Q4 D66	95%	12	11	12	15			
Q5 PVV	5%	15	15	16	15			
Q5 PVV	50%	15	17	18	17			
Q5 PVV	95%	15	19	20	19			
Q6 GL	5%	4	3	3	3			
Q6 GL	50%	4	4	4	4			
Q6 GL	95%	4	5	5	5			
Q7 SP	5%	15	19	18	19			
Q7 SP	50%	15	21	20	21			
Q7 SP	95%	15	23	22	23			
Q8 SGP	5%	3	1	2	1			
Q8 SGP	50%	3	2	3	2			
Q8 SGP	95%	3	3	3	3			
Q9 PvdD	5%	2	2	2	1			
Q9 PvdD	50%	2	3	3	2			
Q9 PvdD	95%	2	4	3	3			
Q10 CU	5%	5	4	4	5			
Q10 CU	50%	5	5	5	6			
Q10 CU	95%	5	6	6	7			
Q11 50+	5%	2	1	0	3			
Q11 50+	50%	2	2	2	4			
Q11 50+	95%	2	3	4	5			

Table 63: Data of The Netherlands 2012



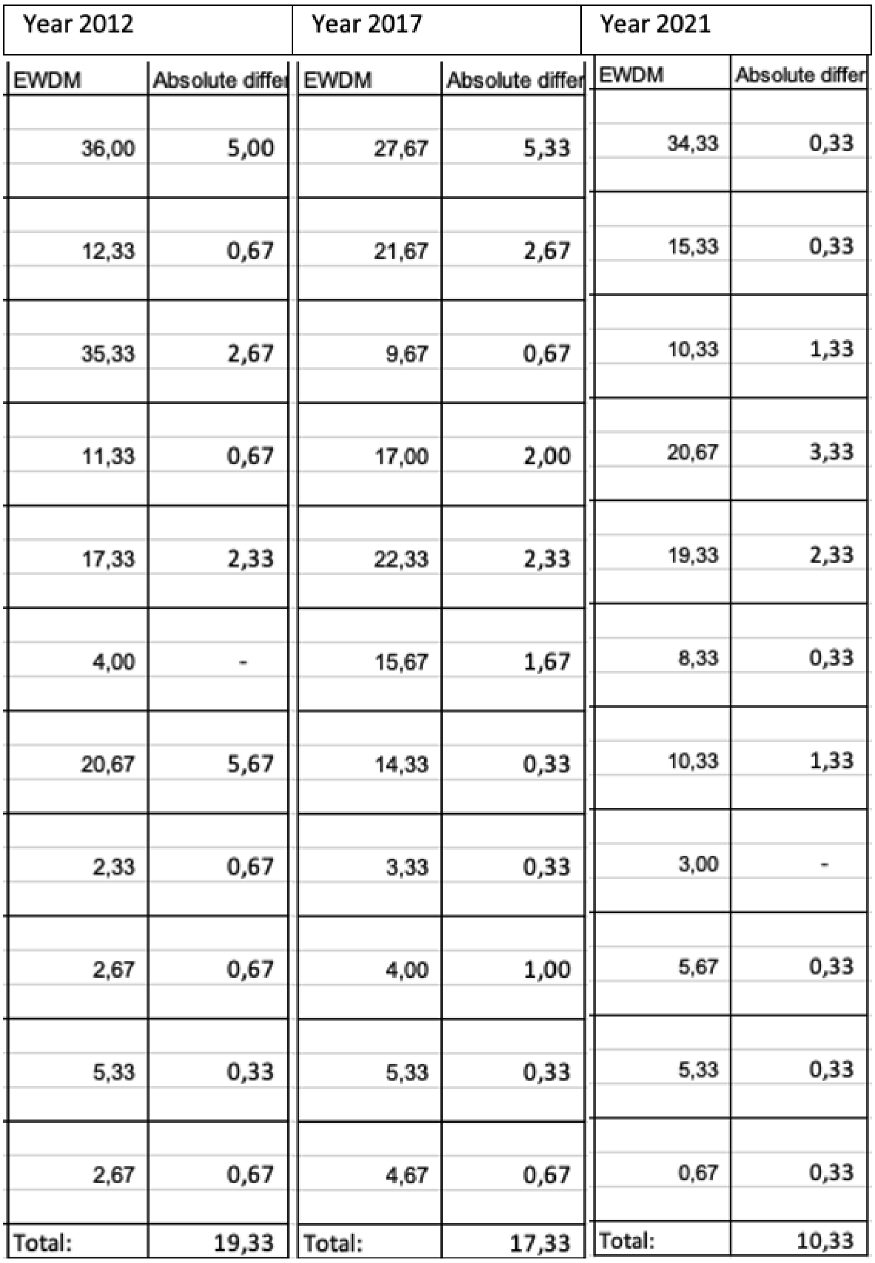
#Calibration va	Confidence int	#Data	#expertA	#ExpertB	#ExpertC		expert A =	ipsos
Q1 VVD	5%	33	27	25	25		expert B =	Peil.nl
Q1 VVD	50%	33	29	27	27		expert C =	Kantar
Q1 VVD	95%	33	31	29	29			
Q2 CDA	5%	19	21	20	18			
Q2 CDA	50%	19	23	22	20			
Q2 CDA	95%	19	25	24	22			
Q3 PvdA	5%	9	8	8	9			
Q3 PvdA	50%	9	9	9	11			
Q3 PvdA	95%	9	10	10	13			
Q4 D66	5%	19	16	13	16			
Q4 D66	50%	19	18	15	18			
Q4 D66	95%	19	20	17	20			
Q5 PVV	5%	20	18	22	21			
Q5 PVV	50%	20	20	24	23			
Q5 PVV	95%	20	22	26	25			
Q6 GL	5%	14	13	16	12			
Q6 GL	50%	14	15	18	14			
Q6 GL	95%	14	17	20	16			
Q7 SP	5%	14	13	11	13			
Q7 SP	50%	14	15	13	15			
Q7 SP	95%	14	17	15	17			
Q8 SGP	5%	3	3	2	2			
Q8 SGP	50%	3	4	3	3			
Q8 SGP	95%	3	5	3	4			
Q9 PvdD	5%	5	3	3	3			
Q9 PvdD	50%	5	4	4	4			
Q9 PvdD	95%	5	5	5	5			
Q10 CU	5%	5	4	4	5			
Q10 CU	50%	5	5	5	6			
Q10 CU	95%	5	6	6	7			
Q11 50+	5%	4	3	3	5			
Q11 50+	50%	4	4	4	6			
Q11 50+	95%	4	5	5	7			

Table 64: Data of The Netherlands of the experts in 2017



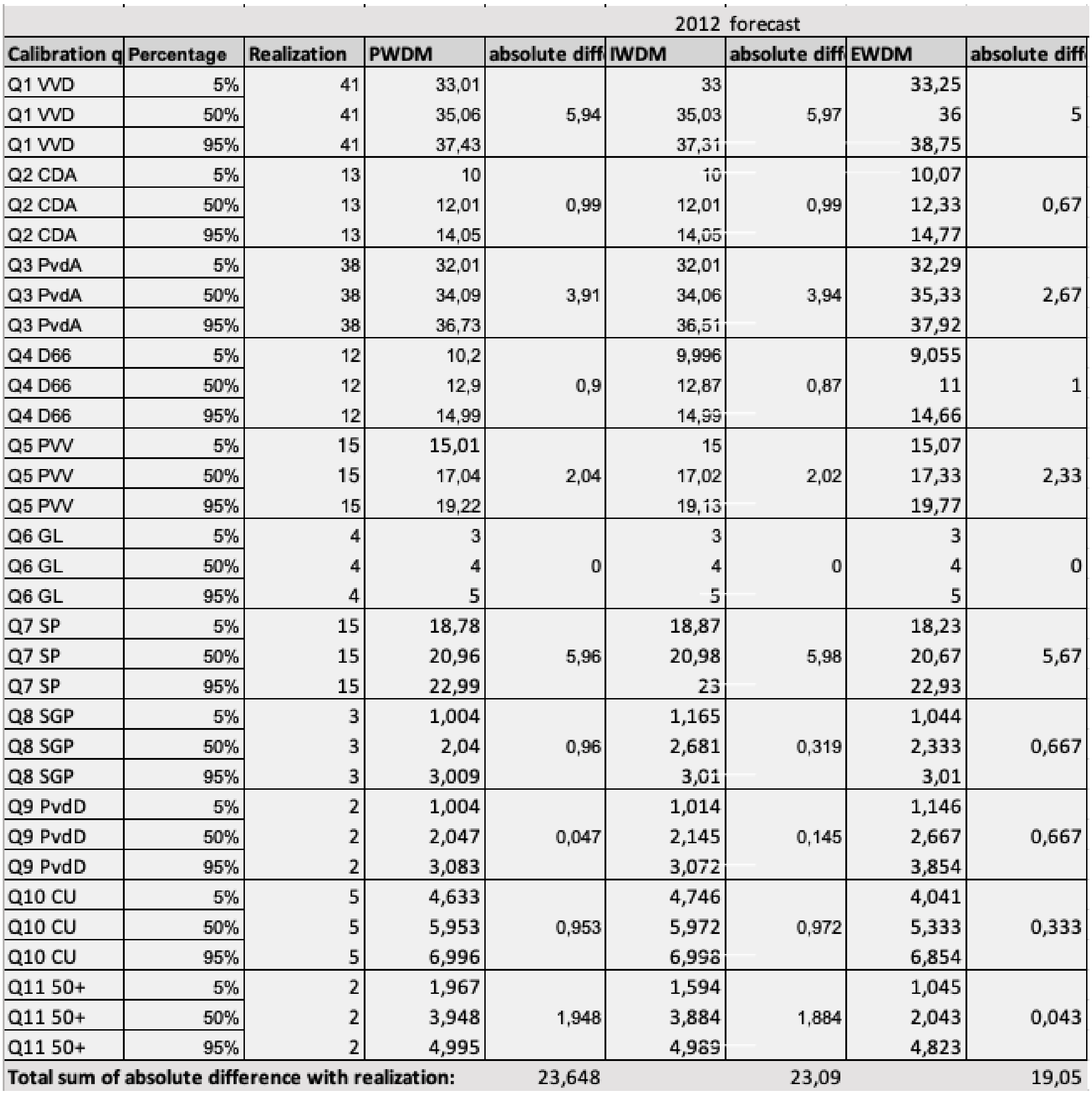
#Calibration va	Confidence int	#Data	#expertA	#ExpertB	#ExpertC		expert A =	ipsos
Q1 VVD	5%	34	33	30	34		expert B =	Peil.nl
Q1 VVD	50%	34	35	32	36		expert C =	Kantar
Q1 VVD	95%	34	37	34	38			
Q2 CDA	5%	15	12	15	13			
Q2 CDA	50%	15	14	17	15			
Q2 CDA	95%	15	16	19	17			
Q3 PvdA	5%	9	8	9	10			
Q3 PvdA	50%	9	9	10	12			
Q3 PvdA	95%	9	10	12	14			
Q4 D66	5%	24	24	17	15			
Q4 D66	50%	24	26	19	17			
Q4 D66	95%	24	28	21	19			
Q5 PVV	5%	17	16	20	16			
Q5 PVV	50%	17	18	22	18			
Q5 PVV	95%	17	20	24	20			
Q6 GL	5%	8	7	6	8			
Q6 GL	50%	8	8	8	9			
Q6 GL	95%	8	9	10	10			
Q7 SP	5%	9	7	9	10			
Q7 SP	50%	9	8	11	12			
Q7 SP	95%	9	9	12	14			
Q8 SGP	5%	3	2	2	2			
Q8 SGP	50%	3	3	3	3			
Q8 SGP	95%	3	4	3	4			
Q9 PvdD	5%	6	4	5	5			
Q9 PvdD	50%	6	5	6	6			
Q9 PvdD	95%	6	6	7	7			
Q10 CU	5%	5	3	5	5			
Q10 CU	50%	5	4	6	6			
Q10 CU	95%	5	5	7	7			
Q11 50+	5%	1	0	0	0			
Q11 50+	50%	1	1	0	1			
Q11 50+	95%	1	1	1	1			

Table 65: Data of The Netherlands of the experts in 2021



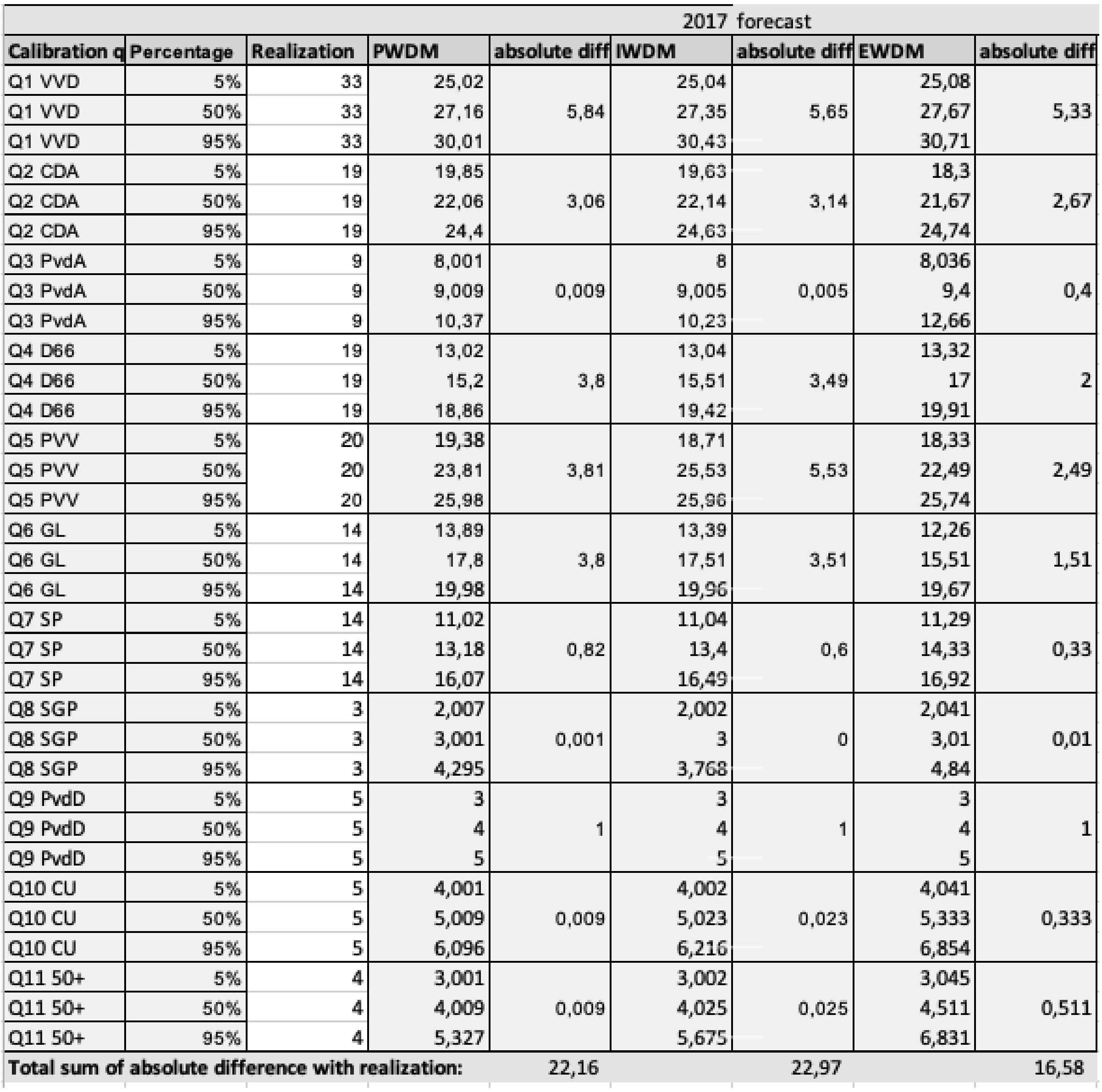
Year 2012		Year 2017		Year 2021	
EWDM	Absolute differ	EWDM	Absolute differ	EWDM	Absolute differ
36,00	5,00	27,67	5,33	34,33	0,33
12,33	0,67	21,67	2,67	15,33	0,33
35,33	2,67	9,67	0,67	10,33	1,33
11,33	0,67	17,00	2,00	20,67	3,33
17,33	2,33	22,33	2,33	19,33	2,33
4,00	-	15,67	1,67	8,33	0,33
20,67	5,67	14,33	0,33	10,33	1,33
2,33	0,67	3,33	0,33	3,00	-
2,67	0,67	4,00	1,00	5,67	0,33
5,33	0,33	5,33	0,33	5,33	0,33
2,67	0,67	4,67	0,67	0,67	0,33
Total:	19,33	Total:	17,33	Total:	10,33

Table 66: Data of EWDM_2 forecast in 2012, 2015 and 2017



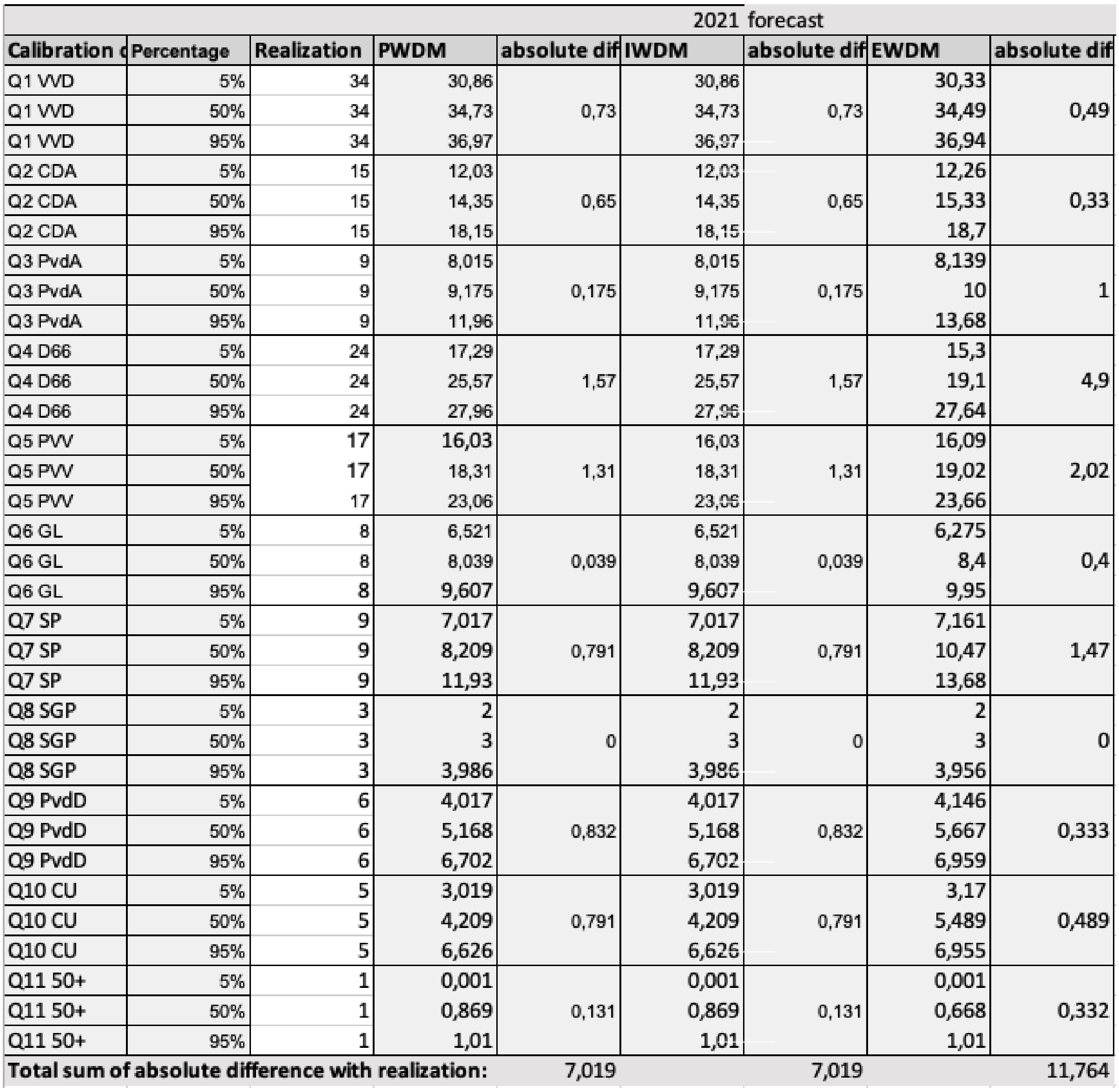
			2012 forecast					
Calibration q	Percentage	Realization	PWDM	absolute diff	IWDM	absolute diff	EWDM	absolute diff
Q1 VVD	5%	41	33,01		33		33,25	
Q1 VVD	50%	41	35,06	5,94	35,03	5,97	36	5
Q1 VVD	95%	41	37,43		37,31		38,75	
Q2 CDA	5%	13	10		10		10,07	
Q2 CDA	50%	13	12,01	0,99	12,01	0,99	12,33	0,67
Q2 CDA	95%	13	14,05		14,05		14,77	
Q3 PvdA	5%	38	32,01		32,01		32,29	
Q3 PvdA	50%	38	34,09	3,91	34,06	3,94	35,33	2,67
Q3 PvdA	95%	38	36,73		36,51		37,92	
Q4 D66	5%	12	10,2		9,996		9,055	
Q4 D66	50%	12	12,9	0,9	12,87	0,87	11	1
Q4 D66	95%	12	14,99		14,99		14,66	
Q5 PVV	5%	15	15,01		15		15,07	
Q5 PVV	50%	15	17,04	2,04	17,02	2,02	17,33	2,33
Q5 PVV	95%	15	19,22		19,13		19,77	
Q6 GL	5%	4	3		3		3	
Q6 GL	50%	4	4	0	4	0	4	0
Q6 GL	95%	4	5		5		5	
Q7 SP	5%	15	18,78		18,87		18,23	
Q7 SP	50%	15	20,96	5,96	20,98	5,98	20,67	5,67
Q7 SP	95%	15	22,99		23		22,93	
Q8 SGP	5%	3	1,004		1,165		1,044	
Q8 SGP	50%	3	2,04	0,96	2,681	0,319	2,333	0,667
Q8 SGP	95%	3	3,009		3,01		3,01	
Q9 PvdD	5%	2	1,004		1,014		1,146	
Q9 PvdD	50%	2	2,047	0,047	2,145	0,145	2,667	0,667
Q9 PvdD	95%	2	3,083		3,072		3,854	
Q10 CU	5%	5	4,633		4,746		4,041	
Q10 CU	50%	5	5,953	0,953	5,972	0,972	5,333	0,333
Q10 CU	95%	5	6,996		6,998		6,854	
Q11 50+	5%	2	1,967		1,594		1,045	
Q11 50+	50%	2	3,948	1,948	3,884	1,884	2,043	0,043
Q11 50+	95%	2	4,995		4,989		4,823	
Total sum of absolute difference with realization:				23,648		23,09		19,05

Table 67: Data of decision makers' forecast of The Netherlands in 2012 using method 1



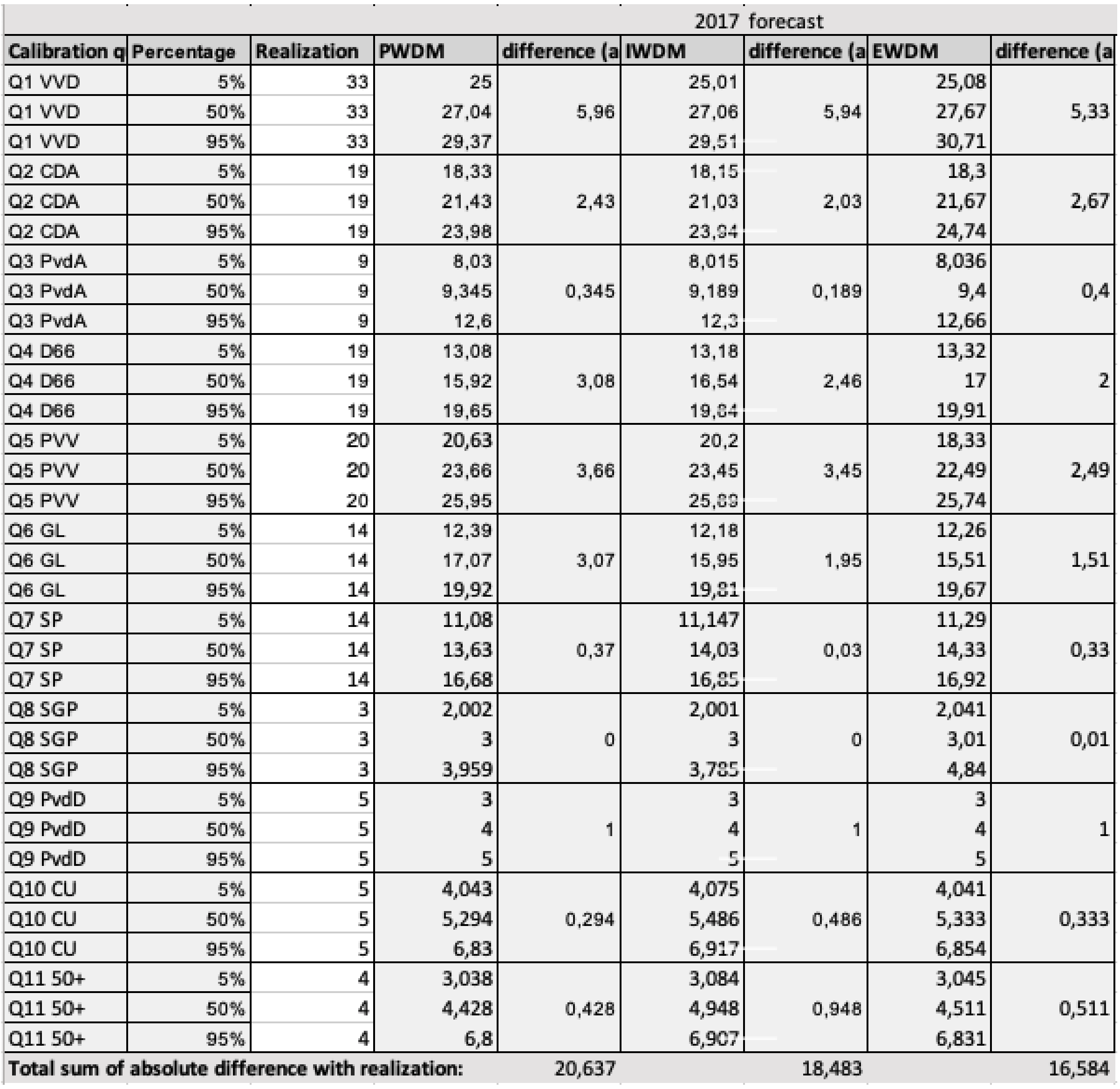
				2017 forecast				
Calibration q	Percentage	Realization	PWDM	absolute diff	IWDM	absolute diff	EWDM	absolute diff
Q1 VVD	5%	33	25,02		25,04		25,08	
Q1 VVD	50%	33	27,16	5,84	27,35	5,65	27,67	5,33
Q1 VVD	95%	33	30,01		30,43		30,71	
Q2 CDA	5%	19	19,85		19,63		18,3	
Q2 CDA	50%	19	22,06	3,06	22,14	3,14	21,67	2,67
Q2 CDA	95%	19	24,4		24,63		24,74	
Q3 PvdA	5%	9	8,001		8		8,036	
Q3 PvdA	50%	9	9,009	0,009	9,005	0,005	9,4	0,4
Q3 PvdA	95%	9	10,37		10,23		12,66	
Q4 D66	5%	19	13,02		13,04		13,32	
Q4 D66	50%	19	15,2	3,8	15,51	3,49	17	2
Q4 D66	95%	19	18,86		19,42		19,91	
Q5 PVV	5%	20	19,38		18,71		18,33	
Q5 PVV	50%	20	23,81	3,81	25,53	5,53	22,49	2,49
Q5 PVV	95%	20	25,98		25,96		25,74	
Q6 GL	5%	14	13,89		13,39		12,26	
Q6 GL	50%	14	17,8	3,8	17,51	3,51	15,51	1,51
Q6 GL	95%	14	19,98		19,96		19,67	
Q7 SP	5%	14	11,02		11,04		11,29	
Q7 SP	50%	14	13,18	0,82	13,4	0,6	14,33	0,33
Q7 SP	95%	14	16,07		16,49		16,92	
Q8 SGP	5%	3	2,007		2,002		2,041	
Q8 SGP	50%	3	3,001	0,001	3	0	3,01	0,01
Q8 SGP	95%	3	4,295		3,768		4,84	
Q9 PvdD	5%	5	3		3		3	
Q9 PvdD	50%	5	4	1	4	1	4	1
Q9 PvdD	95%	5	5		5		5	
Q10 CU	5%	5	4,001		4,002		4,041	
Q10 CU	50%	5	5,009	0,009	5,023	0,023	5,333	0,333
Q10 CU	95%	5	6,096		6,216		6,854	
Q11 50+	5%	4	3,001		3,002		3,045	
Q11 50+	50%	4	4,009	0,009	4,025	0,025	4,511	0,511
Q11 50+	95%	4	5,327		5,675		6,831	
Total sum of absolute difference with realization:				22,16		22,97		16,58

Table 68: Data of decision makers' forecast of The Netherlands in 2017 using method 1



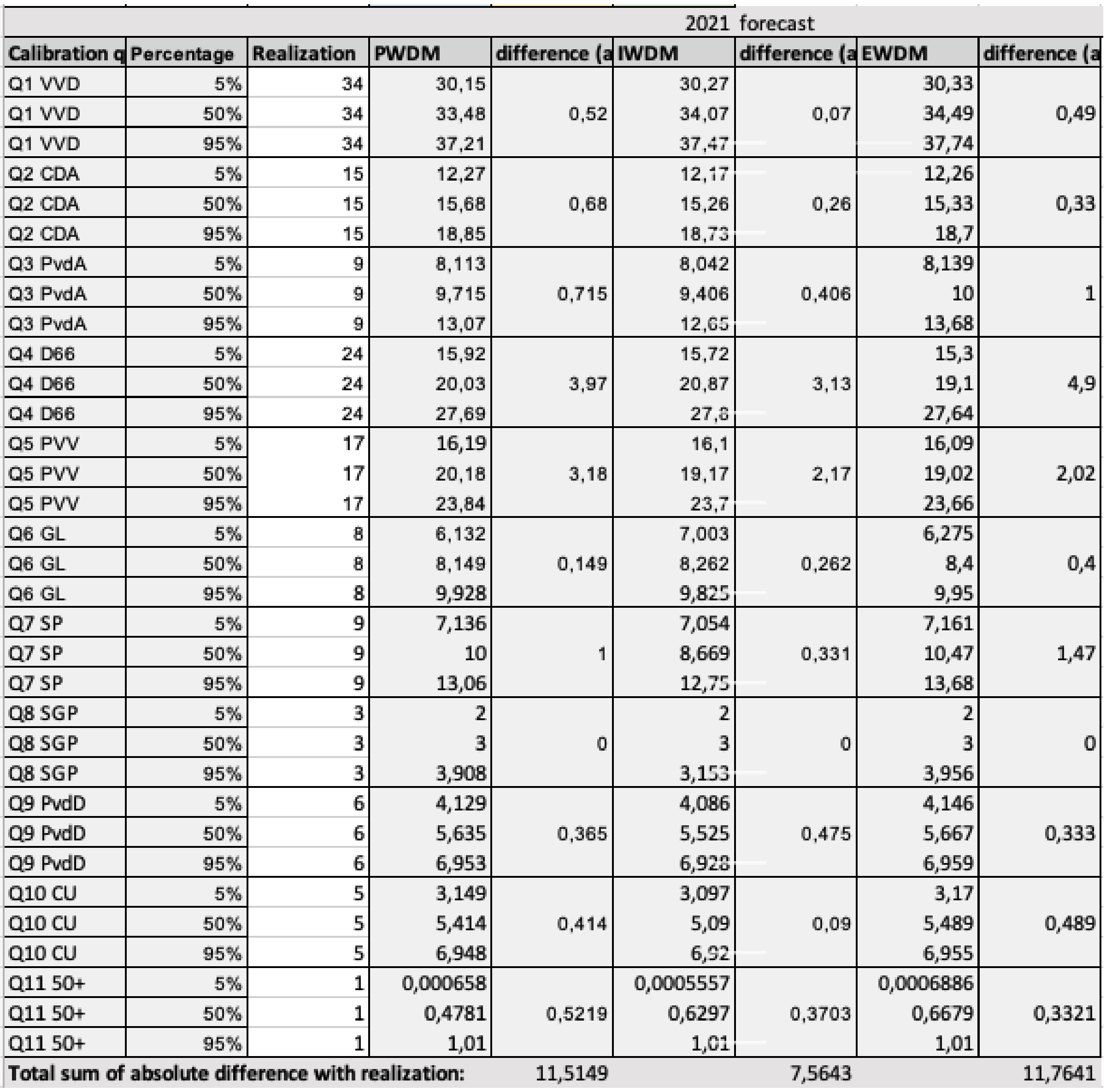
					2021 forecast			
Calibration q	Percentage	Realization	PWDM	absolute dif	IWDM	absolute dif	EWDM	absolute dif
Q1 VVD	5%	34	30,86		30,86		30,33	
Q1 VVD	50%	34	34,73	0,73	34,73	0,73	34,49	0,49
Q1 VVD	95%	34	36,97		36,97		36,94	
Q2 CDA	5%	15	12,03		12,03		12,26	
Q2 CDA	50%	15	14,35	0,65	14,35	0,65	15,33	0,33
Q2 CDA	95%	15	18,15		18,15		18,7	
Q3 PvdA	5%	9	8,015		8,015		8,139	
Q3 PvdA	50%	9	9,175	0,175	9,175	0,175	10	1
Q3 PvdA	95%	9	11,96		11,96		13,68	
Q4 D66	5%	24	17,29		17,29		15,3	
Q4 D66	50%	24	25,57	1,57	25,57	1,57	19,1	4,9
Q4 D66	95%	24	27,96		27,96		27,64	
Q5 PVV	5%	17	16,03		16,03		16,09	
Q5 PVV	50%	17	18,31	1,31	18,31	1,31	19,02	2,02
Q5 PVV	95%	17	23,06		23,06		23,66	
Q6 GL	5%	8	6,521		6,521		6,275	
Q6 GL	50%	8	8,039	0,039	8,039	0,039	8,4	0,4
Q6 GL	95%	8	9,607		9,607		9,95	
Q7 SP	5%	9	7,017		7,017		7,161	
Q7 SP	50%	9	8,209	0,791	8,209	0,791	10,47	1,47
Q7 SP	95%	9	11,93		11,93		13,68	
Q8 SGP	5%	3	2		2		2	
Q8 SGP	50%	3	3	0	3	0	3	0
Q8 SGP	95%	3	3,986		3,986		3,956	
Q9 PvdD	5%	6	4,017		4,017		4,146	
Q9 PvdD	50%	6	5,168	0,832	5,168	0,832	5,667	0,333
Q9 PvdD	95%	6	6,702		6,702		6,959	
Q10 CU	5%	5	3,019		3,019		3,17	
Q10 CU	50%	5	4,209	0,791	4,209	0,791	5,489	0,489
Q10 CU	95%	5	6,626		6,626		6,955	
Q11 50+	5%	1	0,001		0,001		0,001	
Q11 50+	50%	1	0,869	0,131	0,869	0,131	0,668	0,332
Q11 50+	95%	1	1,01		1,01		1,01	
Total sum of absolute difference with realization:				7,019		7,019		11,764

Table 69: Data of decision makers' forecast of The Netherlands in 2021 using method 1



					2017 forecast			
Calibration q	Percentage	Realization	PWDM	difference (a	IWDM	difference (a	EWDM	difference (a
Q1 VVD	5%	33	25		25,01		25,08	
Q1 VVD	50%	33	27,04	5,96	27,06	5,94	27,67	5,33
Q1 VVD	95%	33	29,37		29,51		30,71	
Q2 CDA	5%	19	18,33		18,15		18,3	
Q2 CDA	50%	19	21,43	2,43	21,03	2,03	21,67	2,67
Q2 CDA	95%	19	23,98		23,94		24,74	
Q3 PvdA	5%	9	8,03		8,015		8,036	
Q3 PvdA	50%	9	9,345	0,345	9,189	0,189	9,4	0,4
Q3 PvdA	95%	9	12,6		12,3		12,66	
Q4 D66	5%	19	13,08		13,18		13,32	
Q4 D66	50%	19	15,92	3,08	16,54	2,46	17	2
Q4 D66	95%	19	19,65		19,84		19,91	
Q5 PVV	5%	20	20,63		20,2		18,33	
Q5 PVV	50%	20	23,66	3,66	23,45	3,45	22,49	2,49
Q5 PVV	95%	20	25,95		25,88		25,74	
Q6 GL	5%	14	12,39		12,18		12,26	
Q6 GL	50%	14	17,07	3,07	15,95	1,95	15,51	1,51
Q6 GL	95%	14	19,92		19,81		19,67	
Q7 SP	5%	14	11,08		11,147		11,29	
Q7 SP	50%	14	13,63	0,37	14,03	0,03	14,33	0,33
Q7 SP	95%	14	16,68		16,85		16,92	
Q8 SGP	5%	3	2,002		2,001		2,041	
Q8 SGP	50%	3	3	0	3	0	3,01	0,01
Q8 SGP	95%	3	3,959		3,785		4,84	
Q9 PvdD	5%	5	3		3		3	
Q9 PvdD	50%	5	4	1	4	1	4	1
Q9 PvdD	95%	5	5		5		5	
Q10 CU	5%	5	4,043		4,075		4,041	
Q10 CU	50%	5	5,294	0,294	5,486	0,486	5,333	0,333
Q10 CU	95%	5	6,83		6,917		6,854	
Q11 50+	5%	4	3,038		3,084		3,045	
Q11 50+	50%	4	4,428	0,428	4,948	0,948	4,511	0,511
Q11 50+	95%	4	6,8		6,907		6,831	
Total sum of absolute difference with realization:				20,637		18,483		16,584

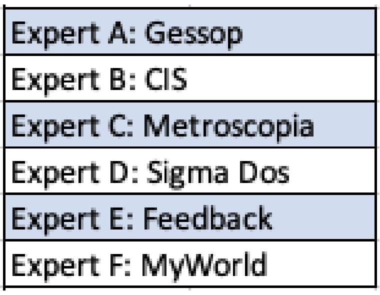
Table 70: Data of decision makers' forecast of The Netherlands in 2017 using method 2



					2021 forecast			
Calibration q	Percentage	Realization	PWDM	difference (a	IWDM	difference (a	EWDM	difference (a
Q1 VVD	5%	34	30,15		30,27		30,33	
Q1 VVD	50%	34	33,48	0,52	34,07	0,07	34,49	0,49
Q1 VVD	95%	34	37,21		37,47		37,74	
Q2 CDA	5%	15	12,27		12,17		12,26	
Q2 CDA	50%	15	15,68	0,68	15,26	0,26	15,33	0,33
Q2 CDA	95%	15	18,85		18,73		18,7	
Q3 PvdA	5%	9	8,113		8,042		8,139	
Q3 PvdA	50%	9	9,715	0,715	9,406	0,406	10	1
Q3 PvdA	95%	9	13,07		12,65		13,68	
Q4 D66	5%	24	15,92		15,72		15,3	
Q4 D66	50%	24	20,03	3,97	20,87	3,13	19,1	4,9
Q4 D66	95%	24	27,69		27,8		27,64	
Q5 PVV	5%	17	16,19		16,1		16,09	
Q5 PVV	50%	17	20,18	3,18	19,17	2,17	19,02	2,02
Q5 PVV	95%	17	23,84		23,7		23,66	
Q6 GL	5%	8	6,132		7,003		6,275	
Q6 GL	50%	8	8,149	0,149	8,262	0,262	8,4	0,4
Q6 GL	95%	8	9,928		9,825		9,95	
Q7 SP	5%	9	7,136		7,054		7,161	
Q7 SP	50%	9	10	1	8,669	0,331	10,47	1,47
Q7 SP	95%	9	13,06		12,75		13,68	
Q8 SGP	5%	3	2		2		2	
Q8 SGP	50%	3	3	0	3	0	3	0
Q8 SGP	95%	3	3,908		3,153		3,956	
Q9 PvdD	5%	6	4,129		4,086		4,146	
Q9 PvdD	50%	6	5,635	0,365	5,525	0,475	5,667	0,333
Q9 PvdD	95%	6	6,953		6,928		6,959	
Q10 CU	5%	5	3,149		3,097		3,17	
Q10 CU	50%	5	5,414	0,414	5,09	0,09	5,489	0,489
Q10 CU	95%	5	6,948		6,92		6,955	
Q11 50+	5%	1	0,000658		0,0005557		0,0006886	
Q11 50+	50%	1	0,4781	0,5219	0,6297	0,3703	0,6679	0,3321
Q11 50+	95%	1	1,01		1,01		1,01	
Total sum of absolute difference with realization:				11,5149		7,5643		11,7641

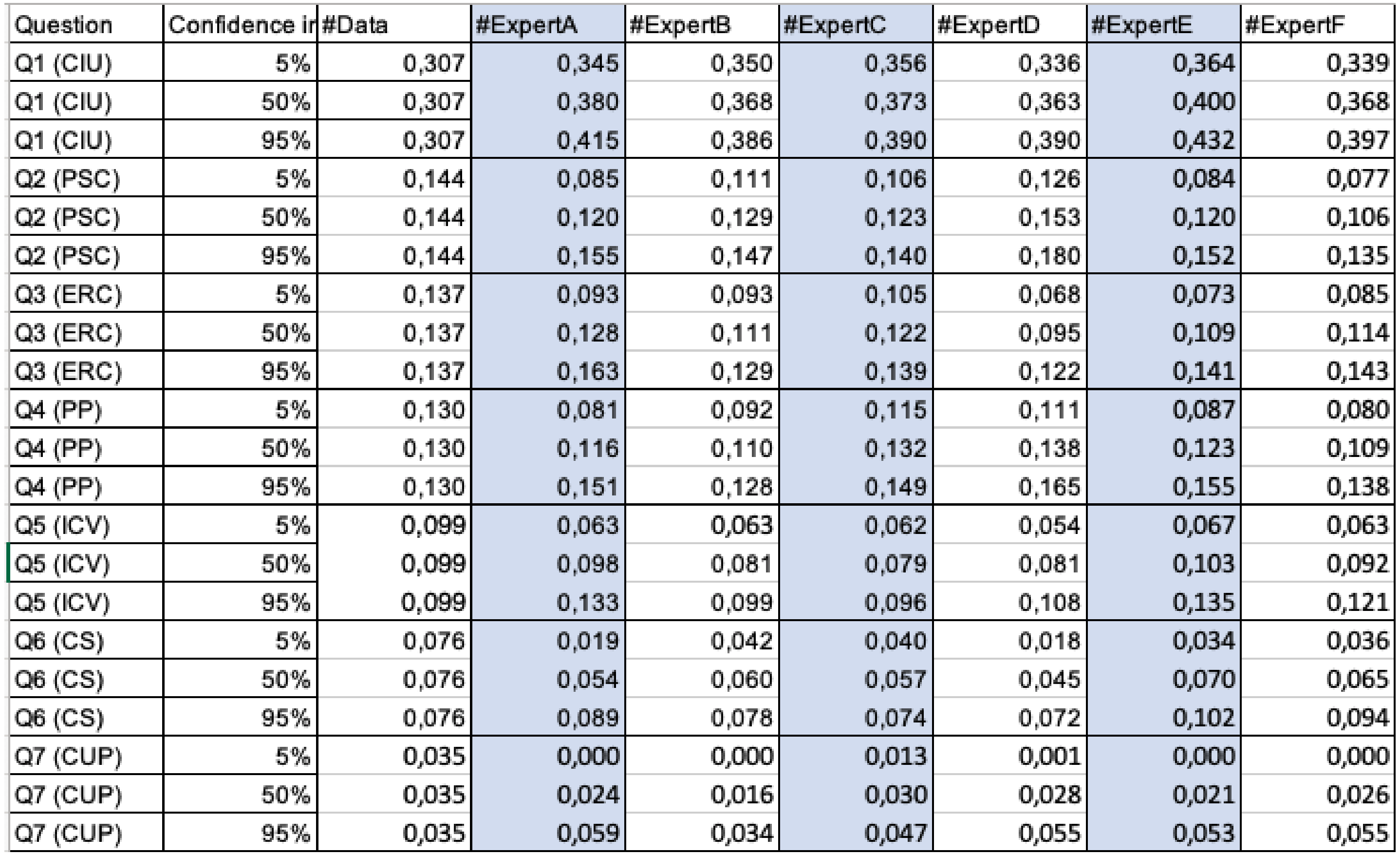
Table 71: Data of decision makers' forecast of The Netherlands in 2021 using method 2

11.3 Appendix B: data of Spain



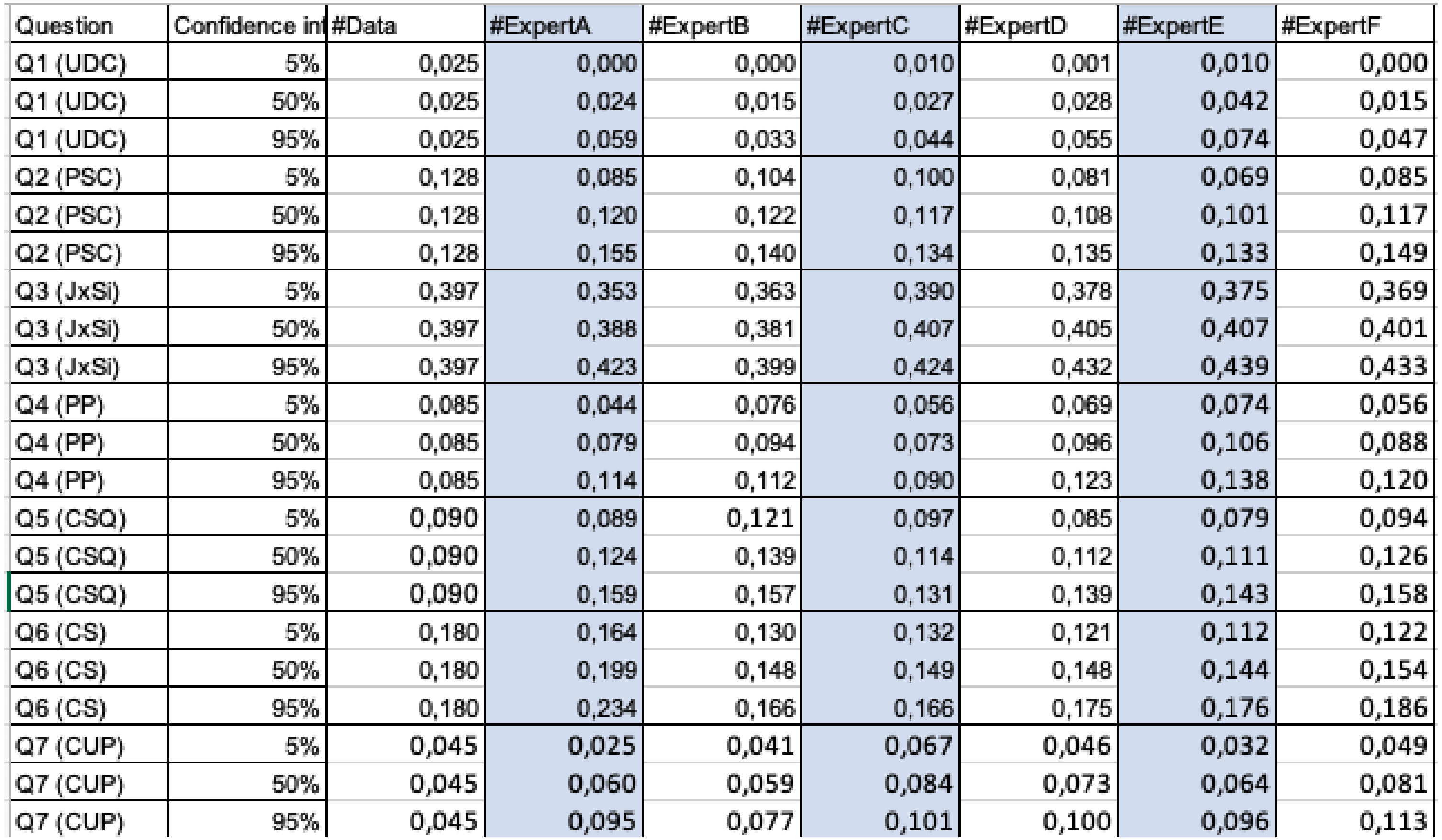
Expert A: Gessop
Expert B: CIS
Expert C: Metroscopia
Expert D: Sigma Dos
Expert E: Feedback
Expert F: MyWorld

Table 72: Experts of Spain



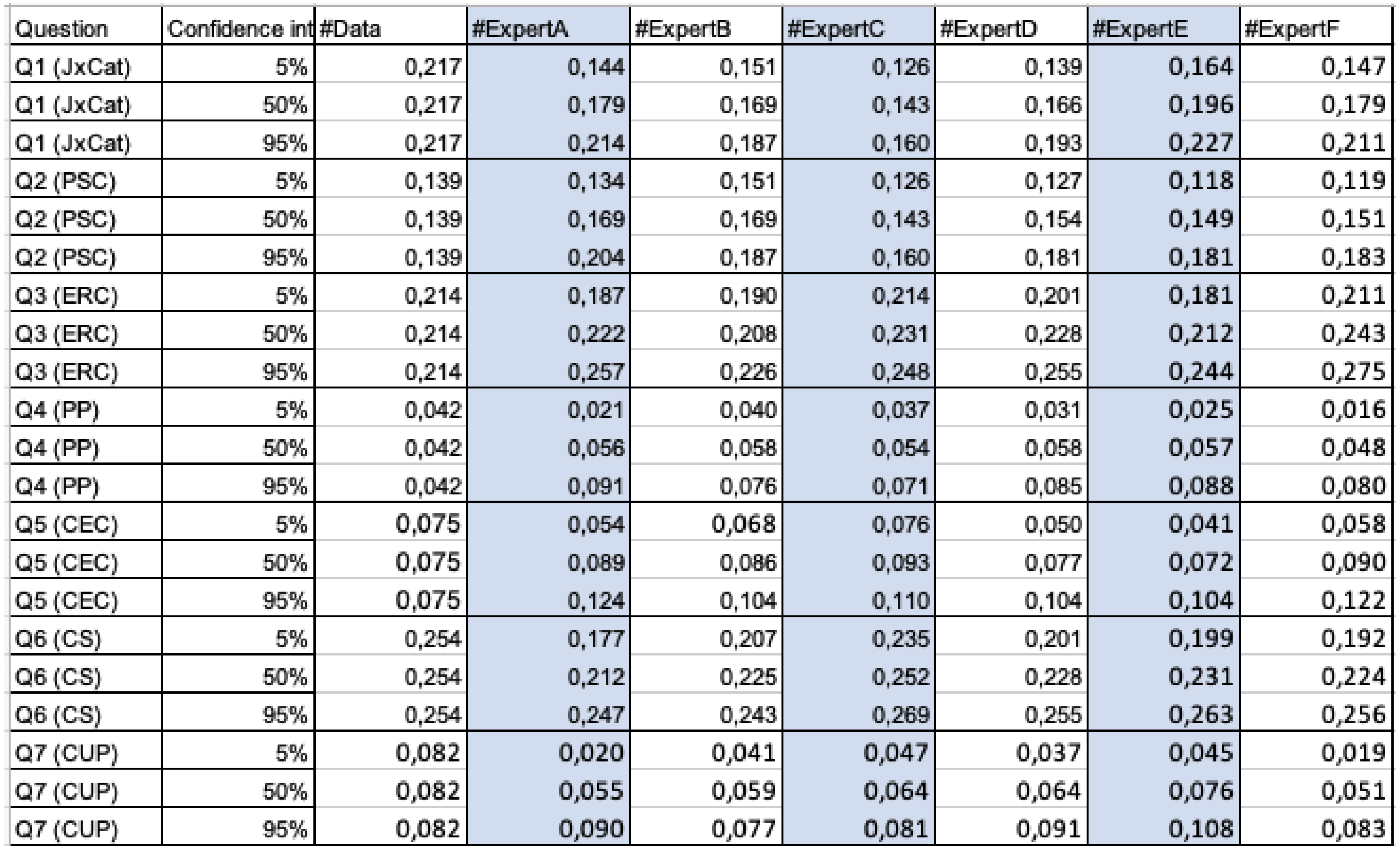
Question	Confidence ir	#Data	#ExpertA	#ExpertB	#ExpertC	#ExpertD	#ExpertE	#ExpertF
Q1 (CIU)	5%	0,307	0,345	0,350	0,356	0,336	0,364	0,339
Q1 (CIU)	50%	0,307	0,380	0,368	0,373	0,363	0,400	0,368
Q1 (CIU)	95%	0,307	0,415	0,386	0,390	0,390	0,432	0,397
Q2 (PSC)	5%	0,144	0,085	0,111	0,106	0,126	0,084	0,077
Q2 (PSC)	50%	0,144	0,120	0,129	0,123	0,153	0,120	0,106
Q2 (PSC)	95%	0,144	0,155	0,147	0,140	0,180	0,152	0,135
Q3 (ERC)	5%	0,137	0,093	0,093	0,105	0,068	0,073	0,085
Q3 (ERC)	50%	0,137	0,128	0,111	0,122	0,095	0,109	0,114
Q3 (ERC)	95%	0,137	0,163	0,129	0,139	0,122	0,141	0,143
Q4 (PP)	5%	0,130	0,081	0,092	0,115	0,111	0,087	0,080
Q4 (PP)	50%	0,130	0,116	0,110	0,132	0,138	0,123	0,109
Q4 (PP)	95%	0,130	0,151	0,128	0,149	0,165	0,155	0,138
Q5 (ICV)	5%	0,099	0,063	0,063	0,062	0,054	0,067	0,063
Q5 (ICV)	50%	0,099	0,098	0,081	0,079	0,081	0,103	0,092
Q5 (ICV)	95%	0,099	0,133	0,099	0,096	0,108	0,135	0,121
Q6 (CS)	5%	0,076	0,019	0,042	0,040	0,018	0,034	0,036
Q6 (CS)	50%	0,076	0,054	0,060	0,057	0,045	0,070	0,065
Q6 (CS)	95%	0,076	0,089	0,078	0,074	0,072	0,102	0,094
Q7 (CUP)	5%	0,035	0,000	0,000	0,013	0,001	0,000	0,000
Q7 (CUP)	50%	0,035	0,024	0,016	0,030	0,028	0,021	0,026
Q7 (CUP)	95%	0,035	0,059	0,034	0,047	0,055	0,053	0,055

Table 73: Data of Spain of the experts in 2012



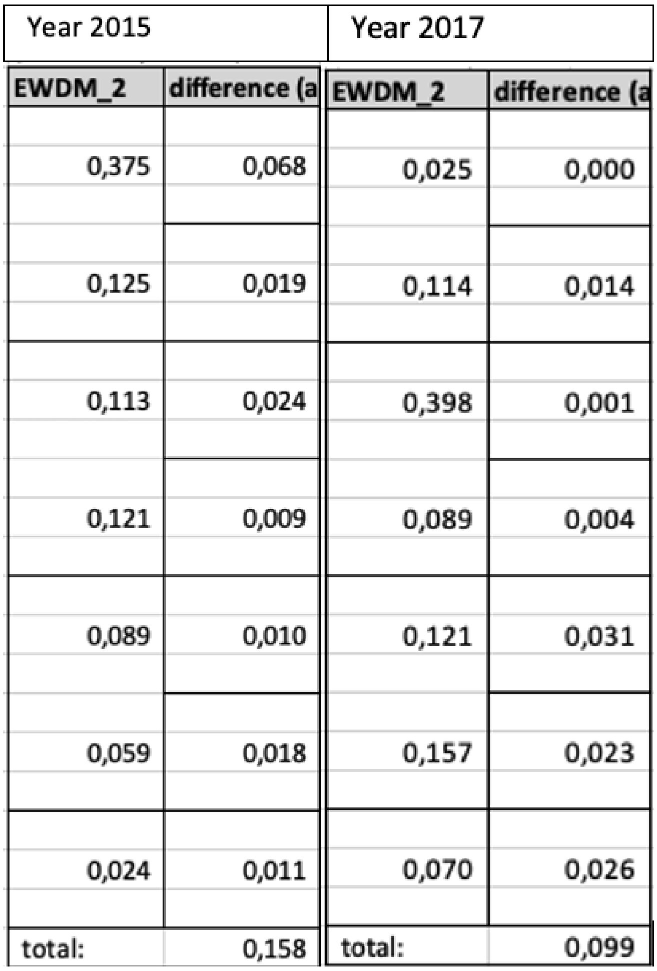
Question	Confidence int	#Data	#ExpertA	#ExpertB	#ExpertC	#ExpertD	#ExpertE	#ExpertF
Q1 (UDC)	5%	0,025	0,000	0,000	0,010	0,001	0,010	0,000
Q1 (UDC)	50%	0,025	0,024	0,015	0,027	0,028	0,042	0,015
Q1 (UDC)	95%	0,025	0,059	0,033	0,044	0,055	0,074	0,047
Q2 (PSC)	5%	0,128	0,085	0,104	0,100	0,081	0,069	0,085
Q2 (PSC)	50%	0,128	0,120	0,122	0,117	0,108	0,101	0,117
Q2 (PSC)	95%	0,128	0,155	0,140	0,134	0,135	0,133	0,149
Q3 (JxSi)	5%	0,397	0,353	0,363	0,390	0,378	0,375	0,369
Q3 (JxSi)	50%	0,397	0,388	0,381	0,407	0,405	0,407	0,401
Q3 (JxSi)	95%	0,397	0,423	0,399	0,424	0,432	0,439	0,433
Q4 (PP)	5%	0,085	0,044	0,076	0,056	0,069	0,074	0,056
Q4 (PP)	50%	0,085	0,079	0,094	0,073	0,096	0,106	0,088
Q4 (PP)	95%	0,085	0,114	0,112	0,090	0,123	0,138	0,120
Q5 (CSQ)	5%	0,090	0,089	0,121	0,097	0,085	0,079	0,094
Q5 (CSQ)	50%	0,090	0,124	0,139	0,114	0,112	0,111	0,126
Q5 (CSQ)	95%	0,090	0,159	0,157	0,131	0,139	0,143	0,158
Q6 (CS)	5%	0,180	0,164	0,130	0,132	0,121	0,112	0,122
Q6 (CS)	50%	0,180	0,199	0,148	0,149	0,148	0,144	0,154
Q6 (CS)	95%	0,180	0,234	0,166	0,166	0,175	0,176	0,186
Q7 (CUP)	5%	0,045	0,025	0,041	0,067	0,046	0,032	0,049
Q7 (CUP)	50%	0,045	0,060	0,059	0,084	0,073	0,064	0,081
Q7 (CUP)	95%	0,045	0,095	0,077	0,101	0,100	0,096	0,113

Table 74: Data of Spain of the experts in 2015



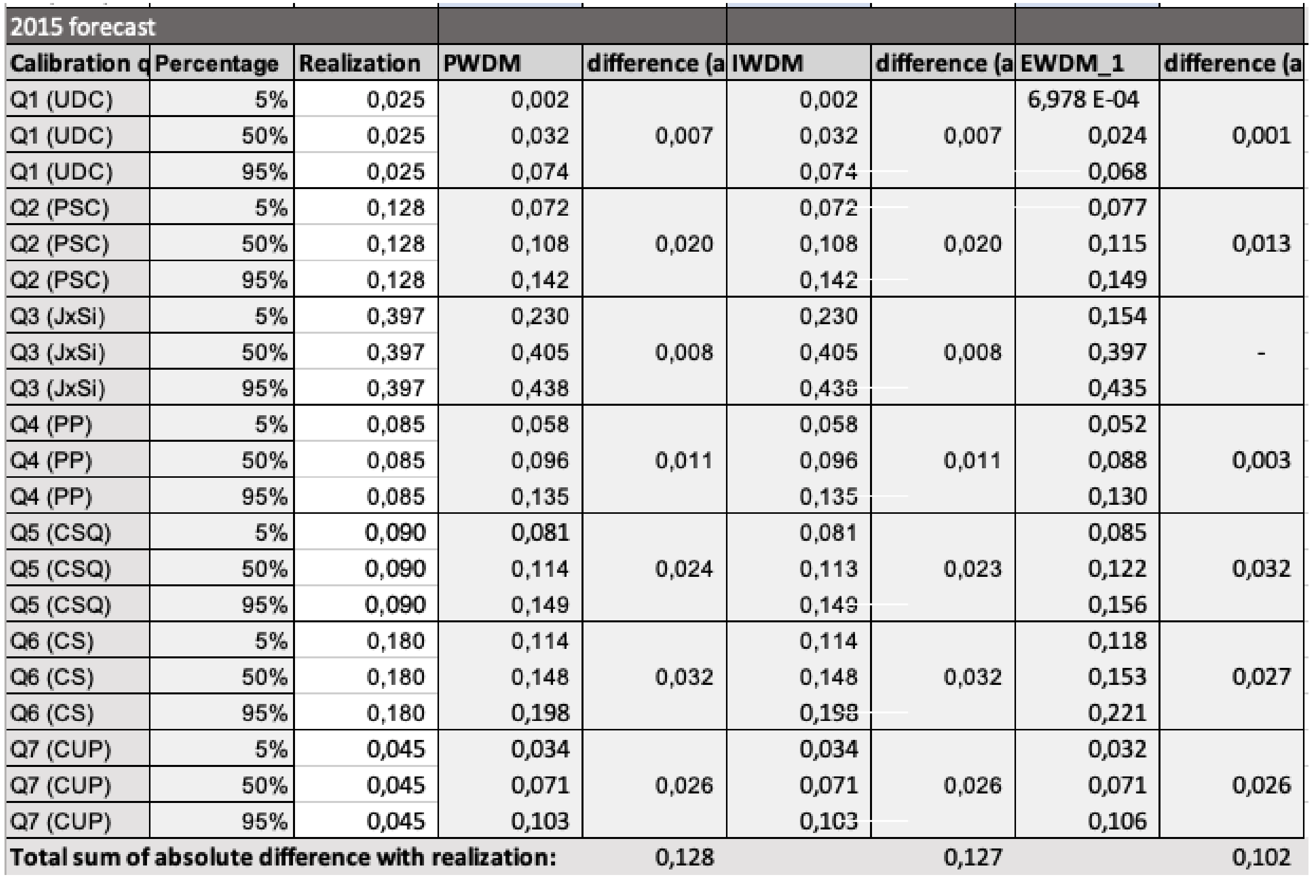
Question	Confidence int	#Data	#ExpertA	#ExpertB	#ExpertC	#ExpertD	#ExpertE	#ExpertF
Q1 (JxCat)	5%	0,217	0,144	0,151	0,126	0,139	0,164	0,147
Q1 (JxCat)	50%	0,217	0,179	0,169	0,143	0,166	0,196	0,179
Q1 (JxCat)	95%	0,217	0,214	0,187	0,160	0,193	0,227	0,211
Q2 (PSC)	5%	0,139	0,134	0,151	0,126	0,127	0,118	0,119
Q2 (PSC)	50%	0,139	0,169	0,169	0,143	0,154	0,149	0,151
Q2 (PSC)	95%	0,139	0,204	0,187	0,160	0,181	0,181	0,183
Q3 (ERC)	5%	0,214	0,187	0,190	0,214	0,201	0,181	0,211
Q3 (ERC)	50%	0,214	0,222	0,208	0,231	0,228	0,212	0,243
Q3 (ERC)	95%	0,214	0,257	0,226	0,248	0,255	0,244	0,275
Q4 (PP)	5%	0,042	0,021	0,040	0,037	0,031	0,025	0,016
Q4 (PP)	50%	0,042	0,056	0,058	0,054	0,058	0,057	0,048
Q4 (PP)	95%	0,042	0,091	0,076	0,071	0,085	0,088	0,080
Q5 (CEC)	5%	0,075	0,054	0,068	0,076	0,050	0,041	0,058
Q5 (CEC)	50%	0,075	0,089	0,086	0,093	0,077	0,072	0,090
Q5 (CEC)	95%	0,075	0,124	0,104	0,110	0,104	0,104	0,122
Q6 (CS)	5%	0,254	0,177	0,207	0,235	0,201	0,199	0,192
Q6 (CS)	50%	0,254	0,212	0,225	0,252	0,228	0,231	0,224
Q6 (CS)	95%	0,254	0,247	0,243	0,269	0,255	0,263	0,256
Q7 (CUP)	5%	0,082	0,020	0,041	0,047	0,037	0,045	0,019
Q7 (CUP)	50%	0,082	0,055	0,059	0,064	0,064	0,076	0,051
Q7 (CUP)	95%	0,082	0,090	0,077	0,081	0,091	0,108	0,083

Table 75: Data of Spain of the experts in 2017



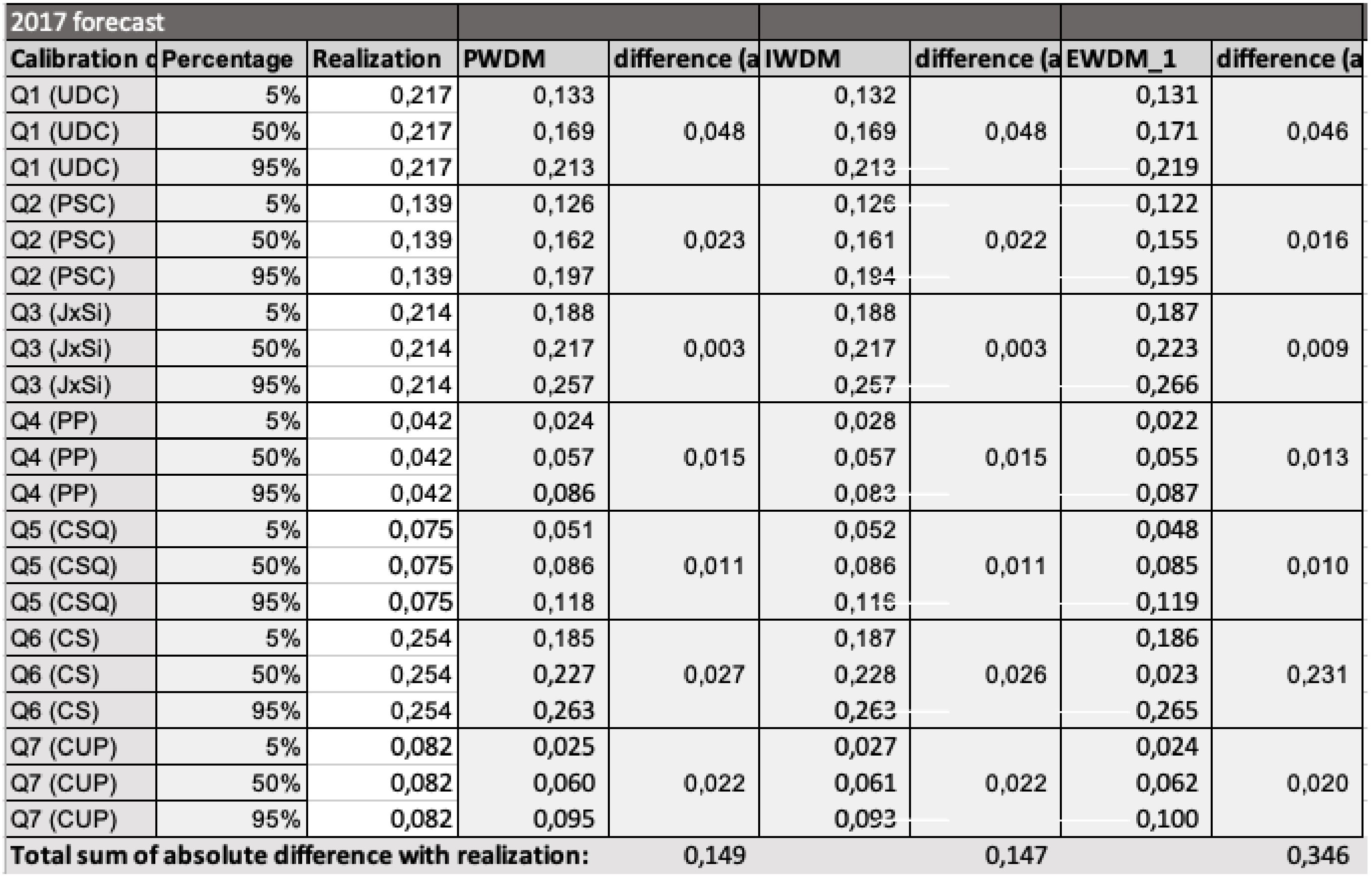
Year 2015		Year 2017	
EWDM_2	difference (a	EWDM_2	difference (a
0,375	0,068	0,025	0,000
0,125	0,019	0,114	0,014
0,113	0,024	0,398	0,001
0,121	0,009	0,089	0,004
0,089	0,010	0,121	0,031
0,059	0,018	0,157	0,023
0,024	0,011	0,070	0,026
total:	0,158	total:	0,099

Table 76: Data of EWDM_2 forecast in 2015 and 2017



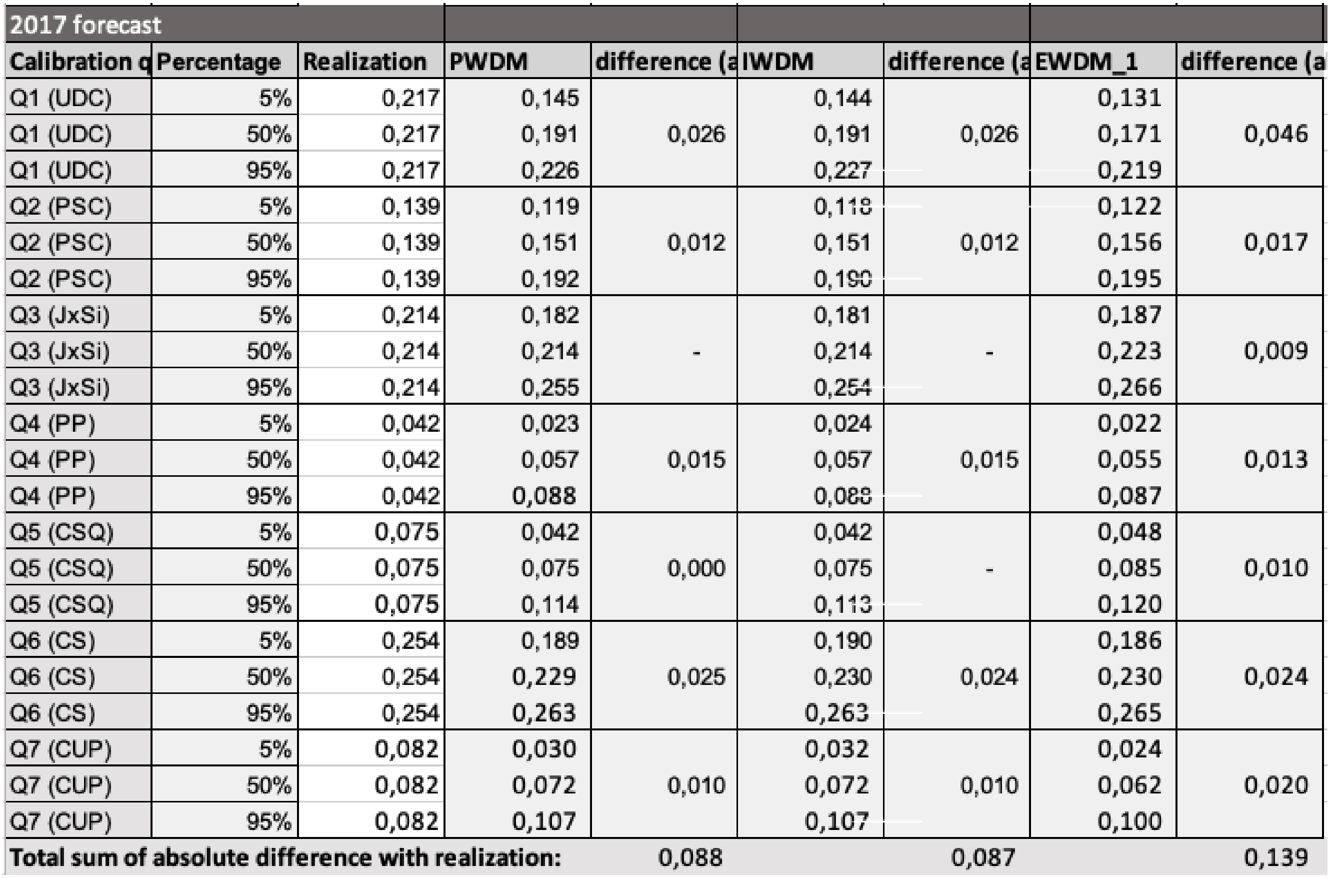
2015 forecast								
Calibration q	Percentage	Realization	PWDM	difference (a	IWDM	difference (a	EWDM_1	difference (a
Q1 (UDC)	5%	0,025	0,002	0,007	0,002	0,007	6,978 E-04	0,001
Q1 (UDC)	50%	0,025	0,032		0,032		0,024	
Q1 (UDC)	95%	0,025	0,074		0,074		0,068	
Q2 (PSC)	5%	0,128	0,072	0,020	0,072	0,020	0,077	0,013
Q2 (PSC)	50%	0,128	0,108		0,108		0,115	
Q2 (PSC)	95%	0,128	0,142		0,142		0,149	
Q3 (JxSi)	5%	0,397	0,230	0,008	0,230	0,008	0,154	-
Q3 (JxSi)	50%	0,397	0,405		0,405		0,397	
Q3 (JxSi)	95%	0,397	0,438		0,438		0,435	
Q4 (PP)	5%	0,085	0,058	0,011	0,058	0,011	0,052	0,003
Q4 (PP)	50%	0,085	0,096		0,096		0,088	
Q4 (PP)	95%	0,085	0,135		0,135		0,130	
Q5 (CSQ)	5%	0,090	0,081	0,024	0,081	0,023	0,085	0,032
Q5 (CSQ)	50%	0,090	0,114		0,113		0,122	
Q5 (CSQ)	95%	0,090	0,149		0,149		0,156	
Q6 (CS)	5%	0,180	0,114	0,032	0,114	0,032	0,118	0,027
Q6 (CS)	50%	0,180	0,148		0,148		0,153	
Q6 (CS)	95%	0,180	0,198		0,198		0,221	
Q7 (CUP)	5%	0,045	0,034	0,026	0,034	0,026	0,032	0,026
Q7 (CUP)	50%	0,045	0,071		0,071		0,071	
Q7 (CUP)	95%	0,045	0,103		0,103		0,106	
Total sum of absolute difference with realization:				0,128		0,127		0,102

Table 77: Data of decision makers' forecast of Spain in 2015 using method 1



2017 forecast								
Calibration c	Percentage	Realization	PWDM	difference (a	IWDM	difference (a	EWDM_1	difference (a
Q1 (UDC)	5%	0,217	0,133		0,132		0,131	
Q1 (UDC)	50%	0,217	0,169	0,048	0,169	0,048	0,171	0,046
Q1 (UDC)	95%	0,217	0,213		0,213		0,219	
Q2 (PSC)	5%	0,139	0,126		0,126		0,122	
Q2 (PSC)	50%	0,139	0,162	0,023	0,161	0,022	0,155	0,016
Q2 (PSC)	95%	0,139	0,197		0,194		0,195	
Q3 (JxSi)	5%	0,214	0,188		0,188		0,187	
Q3 (JxSi)	50%	0,214	0,217	0,003	0,217	0,003	0,223	0,009
Q3 (JxSi)	95%	0,214	0,257		0,257		0,266	
Q4 (PP)	5%	0,042	0,024		0,028		0,022	
Q4 (PP)	50%	0,042	0,057	0,015	0,057	0,015	0,055	0,013
Q4 (PP)	95%	0,042	0,086		0,083		0,087	
Q5 (CSQ)	5%	0,075	0,051		0,052		0,048	
Q5 (CSQ)	50%	0,075	0,086	0,011	0,086	0,011	0,085	0,010
Q5 (CSQ)	95%	0,075	0,118		0,116		0,119	
Q6 (CS)	5%	0,254	0,185		0,187		0,186	
Q6 (CS)	50%	0,254	0,227	0,027	0,228	0,026	0,023	0,231
Q6 (CS)	95%	0,254	0,263		0,263		0,265	
Q7 (CUP)	5%	0,082	0,025		0,027		0,024	
Q7 (CUP)	50%	0,082	0,060	0,022	0,061	0,022	0,062	0,020
Q7 (CUP)	95%	0,082	0,095		0,093		0,100	
Total sum of absolute difference with realization:				0,149		0,147		0,346

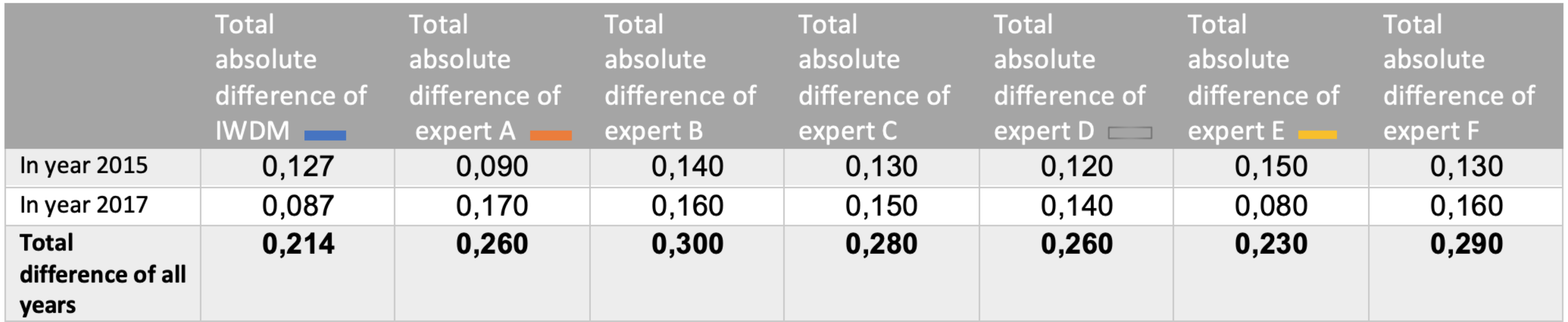
Table 78: Data of decision makers' forecast of Spain in 2017 using method 1



2017 forecast								
Calibration q	Percentage	Realization	PWDM	difference (a	IWDM	difference (a	EWDM_1	difference (a
Q1 (UDC)	5%	0,217	0,145	0,026	0,144	0,026	0,131	0,046
Q1 (UDC)	50%	0,217	0,191		0,191		0,171	
Q1 (UDC)	95%	0,217	0,226		0,227		0,219	
Q2 (PSC)	5%	0,139	0,119	0,012	0,118	0,012	0,122	0,017
Q2 (PSC)	50%	0,139	0,151		0,151		0,156	
Q2 (PSC)	95%	0,139	0,192		0,190		0,195	
Q3 (JxSi)	5%	0,214	0,182	-	0,181	-	0,187	0,009
Q3 (JxSi)	50%	0,214	0,214		0,214		0,223	
Q3 (JxSi)	95%	0,214	0,255		0,254		0,266	
Q4 (PP)	5%	0,042	0,023	0,015	0,024	0,015	0,022	0,013
Q4 (PP)	50%	0,042	0,057		0,057		0,055	
Q4 (PP)	95%	0,042	0,088		0,088		0,087	
Q5 (CSQ)	5%	0,075	0,042	0,000	0,042	-	0,048	0,010
Q5 (CSQ)	50%	0,075	0,075		0,075		0,085	
Q5 (CSQ)	95%	0,075	0,114		0,113		0,120	
Q6 (CS)	5%	0,254	0,189	0,025	0,190	0,024	0,186	0,024
Q6 (CS)	50%	0,254	0,229		0,230		0,230	
Q6 (CS)	95%	0,254	0,263		0,263		0,265	
Q7 (CUP)	5%	0,082	0,030	0,010	0,032	0,010	0,024	0,020
Q7 (CUP)	50%	0,082	0,072		0,072		0,062	
Q7 (CUP)	95%	0,082	0,107		0,107		0,100	
Total sum of absolute difference with realization:				0,088		0,087		0,139

Table 79: Data of decision makers' forecast of Spain in 2017 using method 2

11.4 Appendix C: Comparing best decision maker and all experts of Spain



	Total absolute difference of IWDM 🟦	Total absolute difference of expert A 🟧	Total absolute difference of expert B	Total absolute difference of expert C	Total absolute difference of expert D ▭	Total absolute difference of expert E 🟨	Total absolute difference of expert F
In year 2015	0,127	0,090	0,140	0,130	0,120	0,150	0,130
In year 2017	0,087	0,170	0,160	0,150	0,140	0,080	0,160
Total difference of all years	**0,214**	**0,260**	**0,300**	**0,280**	**0,260**	**0,230**	**0,290**

Table 80: Total absolute differences of best decision maker and all experts

	Unnormalized weight of EWDM_1 🟦	Unnormalized weight of expert A 🟧	Unnormalized weight of expert B ⬜	Unnormalized weight of expert C	Unnormalized weight of expert D 🟨	Unnormalized weight of expert E	Unnormalized weight of expert F
In year 2015	0,023	0,007	0,002	0,019	0,075	0,008	0,015
In year 2017	0,201	0,258	0,511	0,136	0,160	0,096	0,062
Average per year	**0,112**	**0,133**	**0,257**	**0,078**	**0,118**	**0,052**	**0,039**

Table 81: Unnormalized weights of best decision maker and all experts