2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Exploring Statistical Attributes Obtained from Fuzzy Agreement Models

Simon Miller, Christian Wagner and Jonathan M. Garibaldi
Intelligent Modelling and Analysis Group / Horizon Digital Economy Research
School of Computer Science, University of Nottingham, Nottingham, UK.
Email: s.miller@nottingham.ac.uk, christian.wagner@nottingham.ac.uk,
jmg@cs.nott.ac.uk

***Abstract*—In this paper we explore the characteristics of Type-1 Fuzzy Set agreement models based on interval data through contrasting statistical measures of the fuzzy models and the raw data respectively. We create Type-1 Fuzzy Set models using the Interval Agreement Approach, and then extract a preliminary set of attributes that encapsulate aspects of the agreement models. In order to explore what these attributes can tell us, we compare them with a set of traditional statistical measures of consensus which are applied to the raw data. Two interval-valued survey data sets are employed in this study, a synthetic data set consisting of 30 groups of 10 experts rating 25 objects which is used to provide a large example, and a real-world data set consisting of 7 groups of 4-8 cyber-security experts rating 26 security components that was collected during a decision making exercise at GCHQ, Cheltenham, UK. We show that while there are areas in which traditional methods and the attributes extracted from the Type-1 Fuzzy Set agreement models overlap, there are also attributes that do not appear to be replicated, suggesting that these attributes contain additional information about the consensus within the groups. A discussion of the results is provided, along with the conclusions that can be drawn and considerations for future work on this subject.**

***Index Terms*—Survey Data, Correlation Coefficients, Interval Agreement Approach, Agreement Modelling, Computing With Words, Type-1 Fuzzy Sets**

I. Introduction

The ability to measure the level of consensus between sets of observations or measurements has long been considered a vital tool for researchers in a wide range of fields. Quantifying consensus provides the means to make direct comparisons between experts, observations and heterogeneous data sources. For example, we may want to assess the level of agreement between different groups of experts or gauge the likelihood of a potential causal relationship, e.g., body weight and high blood pressure.

There are a wide variety of traditional models that are used to measure consensus, the majority of which measure the extent to which there is a linear or monotonic relationship between raters. Linear relationships are those that can be modelled using a straight line, and monotonic relationships are those only interested in the order in which objects have been placed (ranking). Some of the most commonly used methods are Pearson's r [1], Spearman's Rho [2], Kendall's Tau [3], Kendall's W [4], Cohen's Kappa [5] and Intraclass Correlation Coefficients (ICC) [6].

In the research presented in this paper we will obtain some attributes of consensus from Type-1 Fuzzy Set (T1FS) agreement models created using the Interval Agreement Approach IAA [7]. The IAA is a method of creating T1FSs and General Type-2 Fuzzy Sets (GT2FSs) from interval-valued survey data. The resulting sets provide a completely data-driven model that captures the variation within an individual expert over multiple surveys (intra-expert variation) and the variation between different experts (inter-expert variation) [7].

In previous work, we have shown how sets created using the IAA can be used in practical applications [8], [9], creating word and concept models for use in a Computing With Words (CWW) application [10], [11]. Here, we focus on extracting statistical information from such models (word and concept), as we believe this can further their utility and our understanding of the models. In this work, we explore the similarities/differences between the T1FS models created using the IAA, and traditional consensus measures applied to the raw data. Two data sets are used for this purpose, a synthetic data set and data from a real-world case study. In both data sets, a number of groups of experts provide ratings of a number of objects. Kendall's W and the ICC are used for comparisons involving all groups and objects as they are appropriate for measuring consensus with multiple experts/multiple objects; Standard Deviation will be used in examples looking at just one object.

There are many ways in which the T1FS models can be evaluated to gauge consensus, in this study a set of early evaluation methods have been identified in order to investigate what the T1FS models encapsulate. While they do not completely capture the characteristics of the consensus modelled in the T1FS models, they do provide a numerical representation of some aspects of the agreement (consensus) model, allowing comparison with the outputs of traditional methods.

As the T1FS models incorporate all of the data, and are a model of the agreement between experts, we can reasonably expect them to capture aspects of consensus that are not captured by traditional methods. More fundamentally, if the results produced by IAA can be completely replicated using traditional methods, then it would be more computationally efficient to use those methods in cases where only the statis-

This work was partially funded by CESG - the UK Government's National Technical Authority for Information Assurance (IA) and RCUK's Horizon Digital Economy Research Hub grant, EP/G065802/1.

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

tical summaries of the agreement, rather than the overall FS model are needed.

The following section will review relevant work in the literature, provide an overview of the techniques employed in this research and give details of the data sets that are used. Section III describes a synthetic example showing how measures of consensus can be obtained from T1FSs created using the IAA, and compares the results with traditional methods. Section IV demonstrates the proposed approach using the real world case study described, again comparing the results with traditional methods. Section V provides some discussion of the results and what they mean to the process of data analysis. Finally, Section VI presents the conclusions that can be drawn from the work, and considers directions of future study.

II. Background

In this section we will review the techniques that are employed in this research.

A. Correlation Coefficients

The measurement of correlation/consensus between multiple data sources is a practice which occurs in almost every field of research. The methods to measure correlation have been created over a very long period of time, and most of them work in a very similar way.

In previous work we have used traditional correlation coefficients to examine variation in opinion between cyber-security experts when ranking technical attacks on a proposed government system [12]. The aim of the work was to identify areas where there is an established consensus of opinion, where there is significant disagreement, to identify individuals who are consistently making judgements that strongly disagree with the norm and to show how aggregation can reduce the effects of variation. Two techniques were used to gauge consensus. The first was to calculate group mean rankings for each group of experts and the overall groups of experts and compare them with individual experts using Spearman's Rho. The second approach used Kendall's W to examine the variation within groups of experts. The first approach allows us to see how experts are dispersed around the mean ranking, giving an index of how each deviates from the mean. The second approach corresponds to comparing each expert with every other expert. Using these methods we were able to show that while there is some variation in the opinions of experts, there is a definite consensus of opinion, and that group aggregate rankings are more consistent than the individual experts' rankings.

Spearman's Rho [2], also called Spearman's Rank Correlation Coefficent, measures the statistical dependence of two sets of rankings (ordinal data). The coefficient is a number in [-1,1] that indicates the level of correlation; 1 denotes a perfect positive correlation, 0 means there is no correlation, and -1 denotes a perfect negative correlation. Kendall's W [4], also called Kendall's Coefficient of Concordance, is also used with ordinal data, and allows a correlation coefficient to be computed for more than two raters, the result being a value between 0 (no correlation) and 1 (perfect correlation). There are many similar methods for measuring correlation in data, the main variations are whether they can be applied to more than two raters, and the type of data (i.e., nominal, ordinal, interval and ratio) that they are intended to be used with. Table I gives an overview of some popular methods of measuring correlation. The Test Type column lists whether a measure makes assumptions about the distribution of the data - parametric (P) - or not - non-parametric (NP) - the Data Type column specifies the type(s) of data that each measure can be used with (interval, ratio, ordinal or nominal), the Corr. Type column describes the kind of relationship the method measures between raters (linear or monotonic) and the Raters column lists whether the method can be applied to just two raters or a group of two or more raters.

TABLE I
Correlation Coefficients - Summary

	Test Type	**Data Type**	**Corr. Type**	**Raters**
Pearson's r	P	Int./Rat.	Lin.	2
Spearman's Rho	NP	Ord.	Mono.	2
Kendall's Tau	NP	Ord.	Mono.	2
Kendall's W	NP	Ord.	Mono.	2+
Cohen's Kappa	NP	Nom.	Lin.	2
ICC	NP	Int./Rat./Ord.	Lin./Mono.	2+

Although all of these methods work in a similar way, each is tailored to work in specific circumstances and measure correlation in a set way. For example Pearson's r is used to measure correlation between ratings that are interval or ratio data when a normal distribution can be assumed; Spearman's Rho is used to measure correlation between ordinal (ranked) data and does not assume that there is a normal distribution; Kendall's W is like Spearman's Rho but can be used with more than two sets of ratings.

Intraclass Correlation Coefficients (ICC) are a set of six measures that are used to measure the reliability of raters in a variety of circumstances [13]. Table II describes each of the six forms as described in [13]. The columns show each form of ICC from ICC(1,1) to ICC(3,k), the rows describe what type of test each is (e.g., 'One Way' or 'Two Way'), and the characteristics of the data they are used with (e.g., 'Fixed Judges' or 'Random Judges').

TABLE II
Six forms of ICC

	1,1	**2,1**	**3,1**	**1,k**	**2,k**	**3,k**
One Way	●			●		
Two Way		●	●		●	●
Fix Jdg			●			●
Rand Jdg	●	●		●	●	
Abs Agree	●	●		●	●	
Cons			●			●
Ind Rating	●	●	●			
Mean Rating				●	●	●

To decide which of these measures is appropriate, a series of questions should be asked. 1) Is the order of the measures (ratings) important? If it isn't a one way model is used, if it is, a two way model is used. 2) Are the judges a random sample from a larger population, or a fixed population? 3) Do differences in the judges mean ratings matter? i.e., if judge a consistently gives higher ratings than judge b is it important (absolute agreement), or are we just interested in the order they have placed the targets in (consistency)? 4) Are the ratings from individual judges, or are they the mean of multiple judges ratings? Using these questions a suitable measure can be selected.

In this study we extract consensus values from T1FSs created using the IAA [7] with synthetic and real-world data sets that contain experts' ratings of a set of objects. For comparison, Kendall's W [4] and ICC [13] are used with the raw data to measure the level of consensus. As we have seen, these methods appropriate when the data contains multiple judges rating multiple objects, as is the case in this work. This allows us to explore the nature of the values extracted from the T1FSs, and to inform our understanding of what they represent.

B. The Interval Agreement Approach (IAA) [7], [9]

We proceed by giving a brief overview of the IAA. In previous work we showed how T1 and zSlices based GT2FSs can be constructed from interval-valued survey responses using the IAA [7], [9]. The IAA consists of two steps. In the first, interval-valued data from multiple surveys of a single user is used to compute a T1FS that captures and models the opinion of an individual. The degree of membership to the T1FS represents the intra-user agreement, that is, how much agreement there is over multiple surveys for the same user. In the second step, a number of T1FSs (each representing multiple surveys of a single participant) are combined to create a zSlices based GT2FS. In this set, as before, the primary domain captures the agreement that the users have with themselves over repeated surveys. The secondary domain represents the inter-user agreement, that is, the level of agreement between multiple users. The GT2FSs capture both the intra- and inter-user agreement in two distinct domains. Note that it is possible to 'swap' the modelling of the intra-user uncertainty in the primary membership and the inter-user uncertainty modelling in the secondary membership. This is useful in some applications where, for example, only one sample is available per user but many different users have been sampled. The FSs produced using the IAA use all available data and are solely determined by the data. There is no pre-processing of the data, and no assumptions are made regarding the distribution of the data.

In this paper, we will use the IAA to construct T1FSs that capture and represent the inter-expert variation in both a synthetic data set and a real-world data set. We extract characteristics (e.g., fuzziness and height) of the resulting T1FS models in order to compare and contrast them with traditional statistical consensus analysis methods applied directly to the raw data.

C. Data

Two sets of data are used in this study, a synthetic data set, and a real-world data set collected during a survey exercise.

The synthetic data set consists of 30 groups of experts, each containing 10 experts. Each simulated expert has interval-valued ratings of 25 objects. This synthetic example has been created to provide a large scale example (avoiding the statistical pitfalls with small sample sizes), that is of the same format as the real-world data set, which is small in comparison. The intervals were generated around random start points, with end points selected from a normal distribution of values centred on zero. This ensures that smaller intervals were more likely, and that there is some coherence in the ratings provided by the 'experts'. The data set we have generated is a useful resource for creating and evaluating models, and making comparisons between methods/models. The data set we have used can be found at *http://ima.ac.uk/resources/WCCI2014_NumExample.zip*. In this paper we show an initial application of the data set.

The real-world data set was collected in a survey exercise involving 39 cyber security experts from seven groups drawn from government and commercial backgrounds. The study formed part of collaborative research conducted by GCHQ - The UK's signals intelligence and information assurance agency, and the University of Nottingham. The experts took part in an exercise that was designed to elicit expert opinion on the difficulty of technical attacks, and the difficulty of compromising specific security components within those attacks.

A scenario was presented to the experts that detailed a proposed UK government system that included details of the system, its components and a series of possible attacks. Experts were then asked to rank the attacks in order of difficulty and rate the difficulty of compromising/bypassing each of the security components. In this research it is the security component ratings that we are interested in.

Experts' ratings were elicited using a novel approach that allowed the capturing of both their opinion of the difficulty of compromising/bypassing a component and their certainty in their answer. Answers were expressed by drawing an ellipse over a scale 0 to 1, the intersection points of which were used to create intervals. The width of the interval indicates the level of uncertainty the expert has expressed, the wider the interval is the more uncertain the expert is. We believe that using a single pen stroke to denote uncertainty intervals is more intuitive and easier to perform than drawing separate lines for each end point of an interval. Figure 1 shows two example answers with a) less uncertainty and b) more uncertainty. Figure 2 shows a subset of the data, Group A's ratings of Component 1.

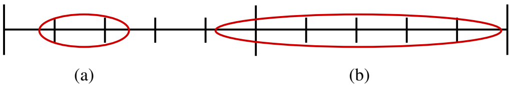


Fig. 1. Interval responses, where 'a' is a less uncertain response and 'b' is a more uncertain response

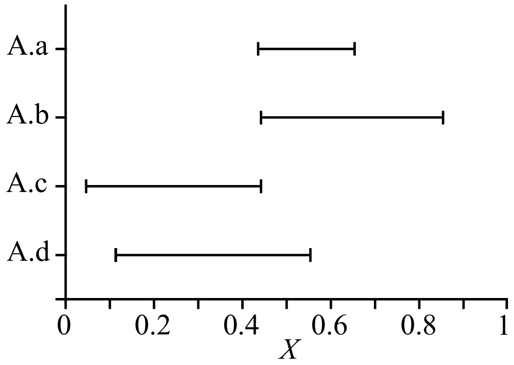


Fig. 2. Group A's interval ratings of the difficulty of compromising Component 1

Further details of the case study and the subsequent analysis can be found in [12].

III. Synthetic Example

First, we will demonstrate the extraction of attributes of consensus using the synthetic example described in the previous section. Using the IAA, T1FSs were created for each group, for each item being rated. Figure 3 shows an example of the sets created, in this case for Group 1's rating of Object 1. Table III shows the raw interval data used to create the T1FS shown in Fig. 3. One of the effects of using the IAA is that where there are non-overlapping intervals in the raw data, non-convex fuzzy sets are created as can be seen in Fig. 3. This is intended, as stated previously the IAA does not make assumptions about the distribution of the created fuzzy sets and is solely data-driven. Further discussion on this can be found in [7] and [9].

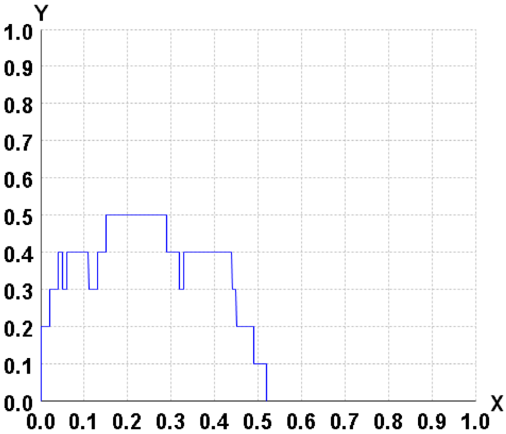


Fig. 3. T1FS created for synthetic Group 1's interval ratings of Object 1

The T1FSs created are a model of the agreement of the experts' individual interval ratings of each object. Here, we will explore the use of attributes of the models to indicate characteristics of groups of experts. In this initial study, we have identified a set of preliminary attributes that can be obtained from the fuzzy agreement models. Table IV shows a list of the attributes extracted, and their values when computed for the T1FS shown in Fig. 3.

TABLE III
Raw data for Group 1's interval ratings of Object 1

Expert	L	R
1	0.31	0.49
2	0.27	0.27
3	0.13	0.32
4	0.15	0.31
5	0	0.05
6	0.02	0.29
7	0	0.11
8	0.33	0.45
9	0.06	0.44
10	0.04	0.52

TABLE IV
T1FS Attributes for Group 1, Object 1

Ave Abs Dev	Fuzziness	Area	Core Size
0.25	0.75	0.19	0.14
Support Size	**Height**	**Spread**	
0.52	0.52	0.25	

Each of these values provides information about the group of experts. The *Average Absolute Deviation* tells us how much deviation from the centroid there is (on average) in the T1FS, *Fuzziness* employs the measure of fuzziness presented in [14] to quantify the vagueness of each T1FS, *Area* is the area under the curve of a T1FS, *Core Size* is the distance between the left most point of the maximum μ value of the T1FS and the right most point, *Support Size* measures the width of the set. *Height* indicates how much agreement there is within the group, where they all agree the height will be 1, lower values indicate less agreement, and *Spread* is an alternative measure of the distribution of the set that produces values closer to 1 when Height is 1 and Support Size is close to 0, and closer to 0 as the Support Size approaches 1. Spread is calculated using (1), where F is a T1FS agreement model.

$$Spread(F) = Height * (1 - SupportSize) \qquad (1)$$

Traditionally, a measure such as Standard Deviation would be used to measure variation between experts. Table V shows the Standard Deviations of the minimum, mean and maximum of the intervals for Group 1, Object 1. These three values have been computed as calculating the Standard Deviation for just one value (e.g., the mean) will give a very limited picture of the variation in the intervals for each expert. In contrast, the attributes of the T1FS models are a result of combining all of the intervals together using the IAA. The model produced gives us more information than looking at the individual intervals, as it shows where there is agreement/discord within the group. We are then able to extract attributes of the agreement model that take into account all of the data contained in the intervals, and how they relate to one another.

TABLE VII
COMPARISON OF ATTRIBUTES FOR GROUPS 1-30, OBJECTS 1-25 WITH KENDALL'S W AND ICC

	Mean Ave Abs Dev	**Mean Fuzziness**	**Mean Area**	**Mean Core Size**	**Mean Support Size**	**Mean Height**	**Mean Spread**
Kendall's Max	**-0.52**	-0.10	**-0.38**	-0.13	**-0.43**	**-0.57**	0.07
Kendall's Mean	-0.14	0.21	0.11	-0.22	-0.35	-0.32	0.19
Kendall's Min	-0.12	0.05	0.05	-0.28	-0.26	-0.24	0.11
ICC1 Max	**-0.52**	-0.10	**-0.38**	-0.13	**-0.43**	**-0.57**	0.07
ICC1 Mean	-0.14	0.21	0.11	-0.22	-0.35	-0.32	0.19
ICC1 Min	-0.12	0.05	0.05	-0.28	-0.26	-0.25	0.11
ICC2 Max	**-0.52**	-0.10	**-0.38**	-0.13	**-0.43**	**-0.57**	0.07
ICC2 Mean	-0.14	0.21	0.11	-0.22	-0.35	-0.32	0.19
ICC2 Min	-0.12	0.05	0.05	-0.28	-0.26	-0.25	0.11
ICC3 Max	**-0.52**	-0.10	**-0.38**	-0.13	**-0.43**	**-0.57**	0.07
ICC3 Mean	-0.14	0.21	0.11	-0.22	-0.35	-0.32	0.19
ICC3 Min	-0.12	0.05	0.05	-0.28	-0.26	-0.24	0.11
ICC1k Max	**-0.53**	-0.10	**-0.39**	-0.13	**-0.44**	**-0.58**	0.08
ICC1k Mean	-0.14	0.22	0.12	-0.23	-0.35	-0.31	0.19
ICC1k Min	-0.12	0.06	0.06	-0.29	-0.26	-0.25	0.11
ICC2k Max	**-0.53**	-0.10	**-0.39**	-0.13	**-0.44**	**-0.58**	0.08
ICC2k Mean	-0.14	0.22	0.12	-0.23	-0.35	-0.31	0.19
ICC2k Min	-0.12	0.06	0.06	-0.29	-0.26	-0.25	0.11
ICC3k Max	**-0.53**	-0.10	**-0.39**	-0.13	**-0.44**	**-0.58**	0.08
ICC3k Mean	-0.14	0.22	0.12	-0.23	-0.35	-0.31	0.19
ICC3k Min	-0.12	0.06	0.06	-0.29	-0.26	-0.25	0.11

TABLE V
STANDARD DEVIATIONS FOR GROUP 1, OBJECT 1

Min	**Mean**	**Max**
0.13	0.12	0.16

TABLE VI
COMPARISON OF ATTRIBUTES FOR GROUPS 1-30, OBJECT 1 WITH STANDARD DEVIATION

	Ave Abs Dev	**Fuzziness**	**Area**	**Core Size**
Min Std Dev	0.12	-0.09	-0.22	0.27
Mean Std Dev	0.09	-0.33	-0.32	0.36
Max Std Dev	*0.44*	-0.13	0.18	0.19
	Support Size	**Height**	**Spread**	
Min Std Dev	0.27	0.28	-0.18	
Mean Std Dev	*0.42*	*0.42*	*-0.40*	
Max Std Dev	***0.60***	***0.59***	***-0.62***	

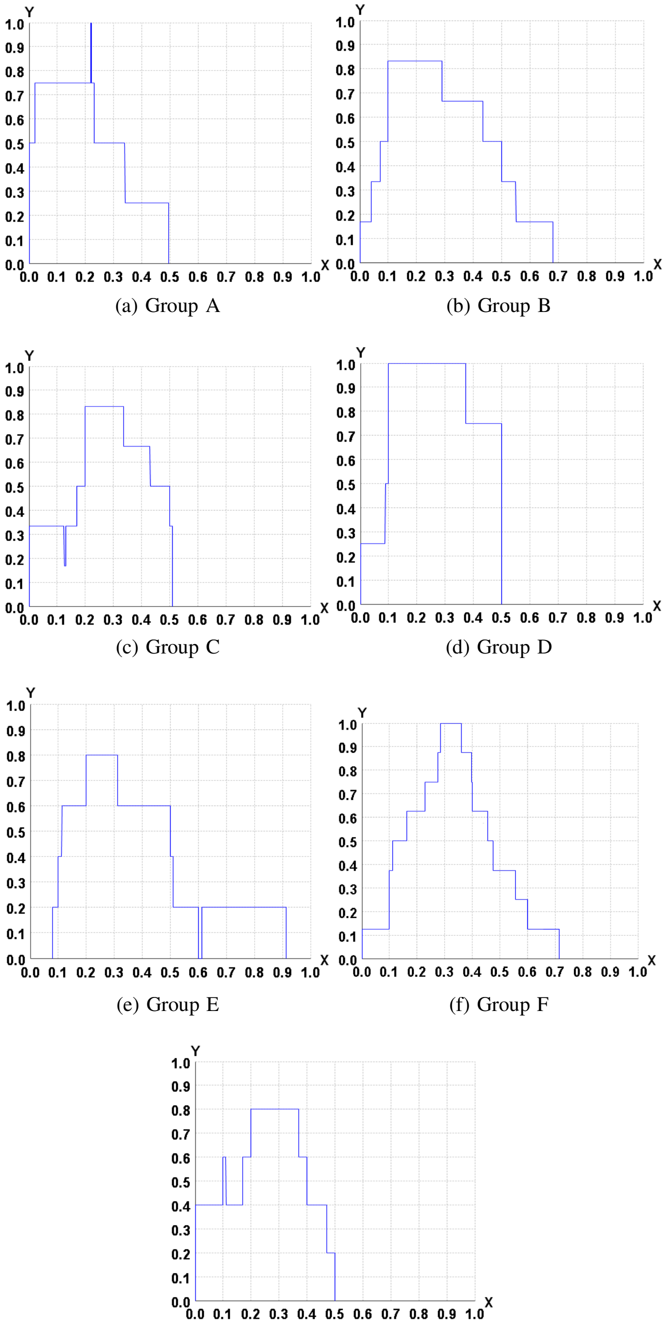
A. Comparison

Each attribute informs us about the agreement/coherence of the group of experts. In order to explore potential relationships between the attributes of the T1FS models, and the values produced using the Standard Deviation, we now compare the attributes of the T1FSs for all 30 groups (for Object 1) with the Standard Deviations of the minimum, mean and maximum of the experts' intervals. Table VI shows the Pearson's Correlation Coefficients for each of the attributes when compared with the Standard Deviations of the minimum, mean and maximum of the intervals from each group of experts. It should be noted that Pearson's Correlation Coefficient is being used to compare the attributes of the T1FS models, and the outputs of the traditional methods when applied to the raw data. It is *not* being compared with the attributes of the T1FS models in this or any of the subsequent examples, as it is not suitable for use on more than 2 sets of ratings. Values shown in *italics* denote that the p-value is less than 0.05 and those in **bold** have a correlation coefficient that indicates strong correlation (>= 0.5).

Looking at the table, we can see that Support Size, Height and Spread closely correspond with the Standard Deviations of the mean and the maximum, and the Average Absolute Deviation is closely correlated with the Standard Deviation of the maximum. This suggests that these values measure similar (though not identical) characteristics of the *group* of experts. The correlation between the remaining values and the traditional methods ranges from weak to moderate. From this we can deduce that these values capture something that is different to the Standard Deviation. Also shown is that there are considerable differences in correlations with Standard Deviations for each interval value. The reason for this may be that each Standard Deviation value is capturing the variation in a separate, and distinct, part of the data. As stated previously, the T1FS models encapsulate all of the data in an agreement model, and therefore the attributes obtained from the models are impacted by the whole data set.

In this example we have focused on one object, and the Standard Deviations of the experts' interval ratings of that object. In the next example we extend this to look at experts' ratings of a set of 25 objects, and compare the outputs of traditional methods applied to the raw data with the attributes we have extracted from our T1FS models. Table VII shows the results.

As in the previous example, it can be seen that some of the values extracted match those produced using traditional methods more closely than others. In this case the Mean



(a) Group A

(b) Group B

(c) Group C

(d) Group D

(e) Group E

(f) Group F

(g) Group G

Fig. 4. T1FSs for Group A-G's assessments of component (hop) 26

Average Absolute Deviation, Mean Area, Mean Support Size, and Mean Height produce results that are closely correlated with the outputs of traditional methods when applied to the maximums of the intervals. As before, the remaining attributes vary from weak to moderate correlation with the traditional methods, and there is considerable differences between correlations with traditional methods applied to the minimum, mean and maximum of the interval data. It is interesting that Spread is not strongly correlated with any of the methods used, as it was strongly correlated with the Standard Deviation of the maximums of intervals in the previous example. It is possible that while this attribute shared some characteristics with the Standard Deviation, it is different to Kendall's W and the ICC.

In view of the examples shown in this section, it can be seen that while there are some similarities between the attributes obtained from the T1FS models and the traditional methods used, as expected, there are also many differences. Attributes including Fuzziness and Core Size never show better than moderate correlation with Standard Deviation, Kendall's W or the ICC, which implies that they measure other aspects of agreement.

In the next section, data from a real-world case study is used to demonstrate the acquisition of the identified attributes in a practical example.

IV. Case Study

In this section we explore the use of T1FS models created with the IAA for a real-world case study to derive characteristics of the ratings of groups of cyber-security experts. Data collected from technical experts at GCHQ is used to create a set of T1FS models that capture the opinions of 7 groups of experts on the overall difficulty of compromising each of 26 security components. As an example, Fig. 4 shows the T1FSs created using all 7 groups' ratings of Component 26.

In some groups (i.e., Figs. 4b, 4c, 4e and 4g) it can be seen that there is no area where all experts are in agreement (i.e., the height is smaller than 1). Where $\mu = 1$ all experts are in agreement. It is clear however, that there is a consensus of opinion within all groups that this security component is relatively easy to compromise/bypass. It can also be seen that some groups have produced sets with a larger support size, that is, they are wider. This indicates that there is either more variation in opinion within these groups, or the there are higher levels of uncertainty in the experts' assessments. Also of note is the number of levels within each group, which reflects the number of individuals within each group. For example, there are 8 members of Group F, and 4 members of Group D.

TABLE VIII
Comparison of Attributes for Groups A-G, Component 1 and Standard Deviations of Intervals

Value	**Ave Abs Dev**	**Fuzziness**	**Area**	**Core Size**
Min	**0.51**	-0.45	-0.18	0.08
Mean	0.25	***-0.84***	**-0.57**	-0.05
Max	0.06	***-0.92***	**-0.73**	0.01

Value	**Support Size**	**Height**	**Spread**
Min	**0.51**	**0.61**	-0.45
Mean	**0.72**	**0.73**	**-0.69**
Max	**0.65**	**0.64**	**-0.65**

A. Comparison

As before, initially we will focus on T1FS models produced for a single component. A set of values has been extracted

TABLE IX
COMPARISON OF ATTRIBUTES FOR GROUPS A-G, COMPONENTS 1-26 WITH KENDALL'S W AND ICC

	Mean **Ave Abs Dev**	**Mean** **Fuzziness**	**Mean** **Area**	**Mean** **Core Size**	**Mean** **Support Size**	**Mean** **Height**	**Mean** **Spread**
Kendall's Max	-0.30	0.37	0.06	0.25	***-0.78***	***-0.88***	**0.73**
Kendall's Mean	-0.40	0.32	-0.05	0.27	***-0.81***	***-0.86***	***0.78***
Kendall's Min	***-0.78***	0.36	**-0.53**	0.40	***-0.85***	**-0.73**	***0.89***
ICC1 Max	-0.18	0.41	0.23	**0.52**	***-0.79***	***-0.88***	**0.73**
ICC1 Mean	-0.31	0.18	0.04	0.41	***-0.77***	***-0.80***	**0.75**
ICC1 Min	**-0.65**	0.14	-0.45	0.18	**-0.70**	**-0.62**	***0.76***
ICC2 Max	-0.19	0.37	0.22	0.49	***-0.79***	***-0.87***	**0.73**
ICC2 Mean	-0.33	0.18	0.02	0.41	***-0.77***	***-0.80***	***0.76***
ICC2 Min	**-0.68**	0.15	-0.47	0.22	**-0.72**	**-0.62**	***0.78***
ICC3 Max	-0.16	0.18	0.20	0.34	**-0.70**	***-0.80***	**0.65**
ICC3 Mean	-0.40	0.11	-0.10	0.35	**-0.73**	**-0.73**	**0.74**
ICC3 Min	-0.75	0.16	**-0.57**	0.33	**-0.71**	**-0.55**	***0.79***
ICC1k Max	0.29	-0.03	**0.50**	0.05	-0.27	-0.48	0.17
ICC1k Mean	-0.03	-0.39	0.05	-0.14	-0.22	-0.30	0.23
ICC1k Min	**-0.53**	-0.20	**-0.59**	-0.24	-0.24	-0.14	0.32
ICC2k Max	0.28	-0.11	0.46	-0.04	-0.23	-0.44	0.15
ICC2k Mean	-0.05	-0.40	0.02	-0.15	-0.20	-0.27	0.22
ICC2k Min	**-0.56**	-0.21	**-0.63**	-0.18	-0.25	-0.12	0.34
ICC3k Max	0.22	-0.34	0.30	-0.25	-0.11	-0.29	0.07
ICC3k Mean	-0.14	-0.45	-0.15	-0.21	-0.12	-0.14	0.16
ICC3k Min	**-0.65**	-0.19	**-0.74**	-0.01	-0.25	-0.04	0.36

from the models for each group in order to infer some statistical properties of each model. Table VIII shows the Pearson's Correlation Coefficient for each value when compared with the standard deviation of each interval endpoint and the means of the intervals.

The results show that there is some variation in the level of correlation between the attributes of the T1FS model and the Standard Deviations of the intervals. Support Size and Height closely match the Standard Deviations of the minimum, mean and maximums of the intervals, suggesting that they capture similar characteristics of the data. Average Absolute Deviation, Fuzziness, Area and Spread agree to some extent with the Standard Deviations, but not all, indicating that there are some differences in what they represent. Finally, Core doesn't match any of the Standard Deviations at all, from this we can assume that it is capturing something that is not captured by the Standard Deviation, and possibly, any of the other attributes.

In the next set of comparisons, models of all 26 security components will be used. This allows us to examine the consensus of the experts' ratings of components using both traditional methods and using values extracted from T1FS agreement models. Table IX shows the Pearson's Correlation Coefficient for the attributes of the T1FS models, and values computed when Kendall's W and ICC are applied to the minimum, mean and maximum of the intervals the experts provided.

As in the previous example, there is some variation in the level of correlation between the outputs of the traditional methods and the attributes extracted from the T1FS agreement models. Mean Support Size, Mean Height and Mean Spread are all strongly correlated with the outputs of Kendall's W and ICC1-3 suggesting that they measure similar properties, but are not strongly correlated with ICC1k-3k which are used when the values entered are mean values as opposed to individual values. Mean Absolute Average Deviation, Mean Fuzziness, Mean Area, and Mean Core Size occasionally show strong correlation, but on the whole they do not match the outputs of the traditional methods. This tells us that the characteristics captured in these attributes cannot be replicated with the traditional methods used.

V. DISCUSSION

In this work we have shown that there are attributes of T1FS agreement models that closely match those produced by selected traditional statistical methods, suggesting that they capture similar properties and could be used for the same types of application. We have also shown that there exist attributes that capture information that cannot be replicated using the selected traditional statistical methods. These attributes are perhaps more interesting as they illustrate that the IAA generated T1FS models capture more characteristics of the uncertainty in expert opinion than the selected statistical methods. The process of combining intervals to produce T1FS models provides information about the consensus of opinion that cannot be gained through the traditional statistical methods. The IAA discards none of the raw data, as such, no information is lost. Traditional statistical methods typically discard most of the raw data to produce a single value. For some applications this may be appropriate, however, when a more complete model of agreement and uncertainty is required the IAA created models offer an ideal solution. What the additional information tells us about the groups of experts' agreement, and how/where it can be applied will be the subject of future research.

In this initial study we have chosen standard deviation, Kendall's W and ICC for comparison. Clearly, there are many other traditional statistical methods that we could compare against, however, they all involve taking a data set and significantly reducing it to produce a measure of a particular characteristic of the raw data, discarding most of the information present in the data.

We have selected a preliminary set of attributes in order to make direct comparisons with the selected traditional methods. In reality, there are an almost infinite number of attributes and combinations of attributes that we could extract from the model, and those selected certainly do not encapsulate all of the characteristics of the models.

There are many other attributes that we could extract. For example, skew and kurtosis are particularly appropriate for the T1FS models produced in this research. The *skewness* of a distribution measures its asymmetry. In the case of a T1FS agreement model it may be that agreement is greater toward one end of the scale. If agreement is greater at the lower end of the scale (and the tail is longer on the right hand side) it is called *positive skew*, and if the reverse is true it is called *negative skew*. Measuring skewness can inform us about where the weight of opinion lies. *Kurtosis* is a measure of the peakedness of distribution, that is, how high the peak is in comparison with the rest of the distribution. In the terms of a T1FS agreement model, kurtosis provides a measure of the point with the greatest agreement, and how this compares with agreement over the rest of the scale. A distribution with a high peak and *heavy* tails has *positive* kurtosis (also termed *leptokurtic*) and a distribution with a low peak and *light* tails has *negative* kurtosis (also termed *platykurtic*). Measuring kurtosis tells us whether agreement is focused and high, or spread out over the scale and low. Both of these measures represent useful attributes of the agreement of a group of experts, and as such, will be of interest in future work.

When considering the consensus of opinion between groups of experts on multiple objects/security components we have currently employed the mean values of attributes extracted from T1FS agreement models. While this does give us an overview of each group, it also removes information, creating one value to represent many values. A more satisfactory method of analysing such data would be the use GT2FSs created using the IAA. This involves combining each group's T1FSs agreement model of a particular object/component to create an overall model incorporating all groups, rather than inspecting each group's T1FS model in isolation [7]. As with the T1FS agreement models, attributes can be extracted from the GT2FS agreement models that tell us about the consensus within the groups. Unlike the T1FS models, a GT2FS agreement model can also provide information about the consensus between groups.

VI. Conclusions and Future Work

In this paper we have shown that T1FS agreement models contain attributes such as Support Size, Height and Spread, that measure similar characteristics of group agreement to standard deviation, Kendall's W and the ICC. This infers that these attributes can be used in similar applications to the traditional methods. We have also shown that the T1FS models contain additional information such as Core Size and Fuzziness that cannot be completely replicated using the selected traditional methods. This highlights the advantage of using T1FS models over statistical methods in that no information is discarded, so an almost infinite number of descriptive attributes can be derived.

In addition to this we have also contributed a significant synthetic data set consisting of 30 groups, each containing 10 simulated experts with interval ratings of 25 objects. This data set is a freely available resource for use in creating, evaluating and comparing models and methods. The data set can be found at *http://ima.ac.uk/resources/WCCI2014_NumExample.zip*

This research represents the beginning of a course of study. In future work we will consider what these preliminary attributes can tell us about the agreement modelled in the T1FSs, and extract additional attributes including skew and kurtosis. We will also extend our study to include the use of GT2FS agreement models, in order to obtain information about both intra- and inter-expert agreement.

References

[1] K. Pearson, "Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia." *Proceedings of the Royal Society of London*, vol. 59, no. 353-358, pp. 69–71, 1895.

[2] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[3] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[4] M. G. Kendall and B. B. Smith, "The problem of m rankings," *The annals of mathematical statistics*, vol. 10, no. 3, pp. 275–287, 1939.

[5] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[6] J. A. Harris, "On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large," *Biometrika*, vol. 9, no. 3/4, pp. 446–472, 1913.

[7] S. Miller, C. Wagner, and J. Garibaldi, "Constructing general type-2 fuzzy sets from interval-valued data," in *Proceedings of the 2012 IEEE International Conference on Fuzzy Systems*, Brisbane, Australia., June 2012.

[8] C. Wagner, S. Miller, and J. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets," in *Proceedings of the 2013 IEEE International Conference on Fuzzy Systems*, Hyderabad, India., July 2013, pp. 1–9.

[9] C. Wagner, S. Miller, J. Garibaldi, D. Anderson, and T. Havens, "From interval-valued data to general type-2 fuzzy sets," *Fuzzy Systems, IEEE Transactions on*, vol. In Press, 2014, DOI: 10.1109/TFUZZ.2014.2310734.

[10] L. Zadeh, "Fuzzy logic = computing with words," *Fuzzy Systems, IEEE Transactions on*, vol. 4, no. 2, pp. 103 –111, may 1996.

[11] J. Mendel and D. Wu, *Perceptual Computing: Aiding People in Making Subjective Judgments*, ser. IEEE Press Series on Computational Intelligence. Wiley, 2010. [Online]. Available: http://books.google.co.uk/books?id=f1iYFtsOh1UC

[12] S. Miller, S. Appleby, J. Garibaldi, and U. Aickelin, "Towards a more systematic approach to secure systems design and analysis," *International Journal of Secure Software Engineering*, vol. 4, no. 1, pp. 11–30, 2013.

[13] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[14] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*. Prentice Hall New Jersey, 1995.