Agreeing to disagree: graphics for comparing expert classifications

Unwin, Antony
Augsburg University, Department of Computer Oriented Statistics and Data Analysis
Universitaetsstrasse 14
86135 Augsburg, Germany
unwin@math.uni-augsburg.de

Hoegg, Tanja (2nd author)
Augsburg University, Department of Computer Oriented Statistics and Data Analysis
Universitaetsstrasse 14
86135 Augsburg, Germany
tanja.hoegg@gmx.de

Pilhoefer, Alexander (3rd author)
Augsburg University, Department of Computer Oriented Statistics and Data Analysis
Universitaetsstrasse 14
86135 Augsburg, Germany
alexander.pilhoefer@gmx.de

Introduction

Harry S. Truman is supposed to have asked for a one-armed economist, because he was fed up with economists telling him "On the one hand..., and then on the other hand". Experts may not always agree (and some not even with themselves, apparently), but quite often they do, more or less, and when they provide ratings it is interesting to consider how this 'more or less' can be measured and judged statistically. The term 'agreement' may refer to agreement between continuous measures as well as to agreement between ordinal or nominal classifications, but in this paper we concentrate on classifications. A number of agreement measures have been suggested, sometimes together with tests for assessing the significance of the agreement. Since equal numerical values of a statistic can arise from quite different data, it is always sensible to display the data in a plot. Perhaps surprisingly, there have not been many suggestions made for visualising the agreement (or lack of it) between raters. In this paper we look at a dataset on the credit ratings of countries, as judged by different financial agencies, and evaluate the agreement between the agencies numerically and visually. We use this example to comment on measures of agreement and to make some recommendations for graphic displays.

Rating Measures

The simplest measure of agreement between two raters is *diag*, the proportion of cases where the same classification is made. As so often with simple and obvious measures, this has its disadvantages (it depends on the number of possible ratings and takes no account of the orderings of ratings). Cohen's *kappa* attempts to get round these disadvantages and is defined as

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)}$$

where $p(a)$ is the relative observed agreement (i.e. *diag*) and $p(e)$ is the agreement that would have been observed by chance given the marginal distributions of the raters. κ takes no account of any ordering in the ratings, but weighted κ does. It sums the differences between ratings, weighted by the distance (i.e. 0 for agreement, 1 for a difference of 1, 2 for a difference of 2, ...), divides this sum by an equivalent sum for the expected scores based on chance agreement, and subtracts the resulting ratio from 1. Asymptotic distributions and tests are available for the κ coefficients. You will also find Scott's Pi (which is like Cohen's κ , but calculates the expected values differently), Fleiss's kappa, which is a generalisation of Scott's Pi to more than 2 raters and Light's kappa, which averages bivariate kappas between multiple raters. Kappa-type measures are discussed in statistical modelling terms in de Mast [2]. The Rand Index could also be used for comparing two raters, though this compares two partitions (so there could be different numbers of rating classes used by two raters) rather than two classifications with the same classes. There is also an adjusted version of the Rand Index, which is corrected for chance agreement. The R package *irr* provides a fuller list of measures. Several other R packages calculate some of the measures as well (κ is particularly popular) though hardly any of these packages use graphics in their examples to illustrate the agreement they are assessing and not all offer statistical tests or confidence intervals. Quite different patterns can give rise to the same statistic values and it is not only the level of disagreement, but how reliable that measure is and also what kind of disagreement there is, that should be of interest. The measures alone are of most value, when you have several raters you wish to compare. In this they are like correlation coefficients, useful when there are many pairs of raters or variables, a poor summary for one pair of raters or variables.

Sovereign Credit Ranking Dataset

Financial rating agencies grade the creditworthiness of firms and countries. The country ratings of three of these agencies (Moody's, Fitch, S&P) for 123 sovereign countries were published in the Guardian newspaper's Datablog in April 2010 [1]. There are two measures from each agency, an alphabetic code with 16 categories and an outlook with three (negative, stable, positive). Not all agencies rated every country and for the purposes of this paper, we shall restrict ourselves to the 77 countries with ratings and outlooks from each agency. To give a flavour of the data, consider the ratings for Ireland in April 2010. Moody's said Aa2, equivalent to AA for the other agencies, (51 countries were rated worse) with a stable outlook, Fitch said AA- (49 countries rated worse) and stable, while S&P said AA (49 countries rated worse) and negative. κ values for the three pairs of outlook ratings are The weighted versions of κ give about the same values, but the respective approximate standard errors are one third to a half smaller. Using a common interpretation suggested by Landis and Koch [6] (If kappa is less than 0, "No agreement", if 0-0.2, "Slight agreement", if 0.2-0.4, "Fair agreement", if 0.4-0.6, "Moderate agreement", if 0.6-0.8, "Substantial agreement", if 0.8-1.0, "Almost perfect agreement".), you would conclude that the levels of agreement are 'slight', 'slight', and 'fair'. (It should be noted that there is no particular justification for this scale, it was just a suggestion by the two authors.) It is instructive to look at the three tables leading to these results. (The data for the first agency named are in the rows and the data for the second agency in the columns.)

Pairs of Agencies	κ	**se(κ)**
Moody and Fitch	0.167	0.166
Moody and S&P	0.016	0.167
Fitch and S&P	0.330	0.138

On the face of it the agreement looks pretty good, but that is because the agencies all rated almost 60% of the countries (46 out of 77) as 'Stable'. Whether the values of Cohen's κ do justice to

Table 1: Moody and Fitch

	Neg	Stable	Pos
Neg	2	3	0
Stable	7	55	4
Pos	1	4	1

Table 2: Moody and S&P

	Neg	Stable	Pos
Neg	1	4	0
Stable	10	51	5
Pos	0	5	1

Table 3: Fitch and S&P

	Neg	Stable	Pos
Neg	6	4	0
Stable	5	52	5
Pos	0	4	1

the tables is a moot point, especially when you observe that only one of the κ's would be regarded as significantly different from zero. The poor performance of κ in such situations is well known [4] and the example demonstrates how important it is to look at the data, either in a table or in a graphic display.

Visualization of Ratings

In the literature there are few plots drawn for agreement data. One specific suggestion is Bangdiwala's Observer Agreement Chart [3] shown in Fig. 1 for the outlook ratings from Moody's and Fitch.

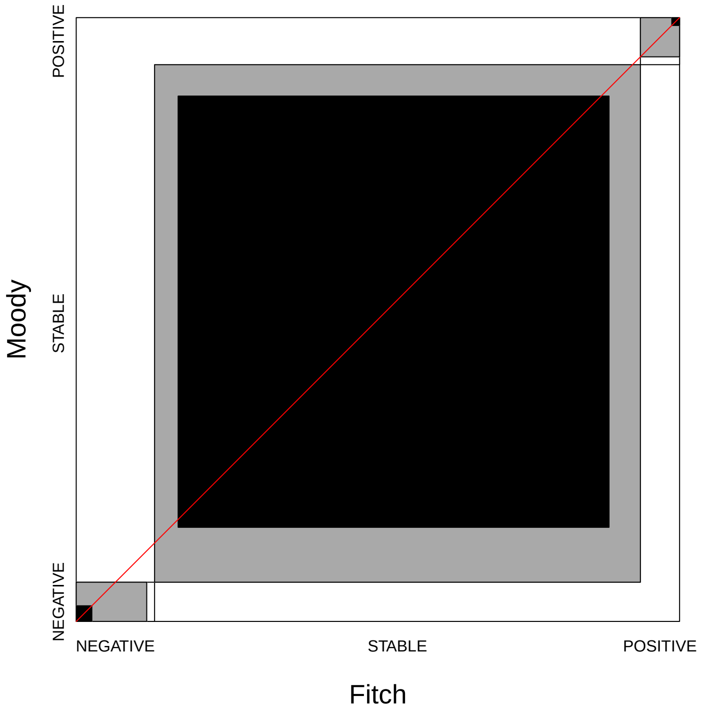

Figure 1: Bangdiwala's Observer Agreement Chart for the Moody's and Fitch ratings.

The black rectangles are proportional to the squares of the observed data, the white rectangles show the maximum possible agreement given the marginal totals. Notice the implicit marginal ratings distributions along the horizontal and vertical axes obtained by extending the rectangle sides. The grey rectangles express partial agreement by including contributions from off-diagonal cells. This plot may not make good use of the space available, as it is concentrated on the diagonal, which is fine here because most of the data are on the diagonal. It also takes a bit of getting used to.

An alternative approach is to first of all draw the distributions of the individual raters (since it is good to have them explicitly) and then to draw some kind of area plot for the data table, either multiple barcharts, conditioning on one of the raters, or a fluctuation diagram [5]. All these plots are shown for the Moody's and Fitch ratings in Fig. 2. The plots so far highlight the heavy influence of the 'Stable' rating, something the statistics rather obscured. Neither the statistics nor the plots shown up till now consider more than two raters at once and that is the next challenge.

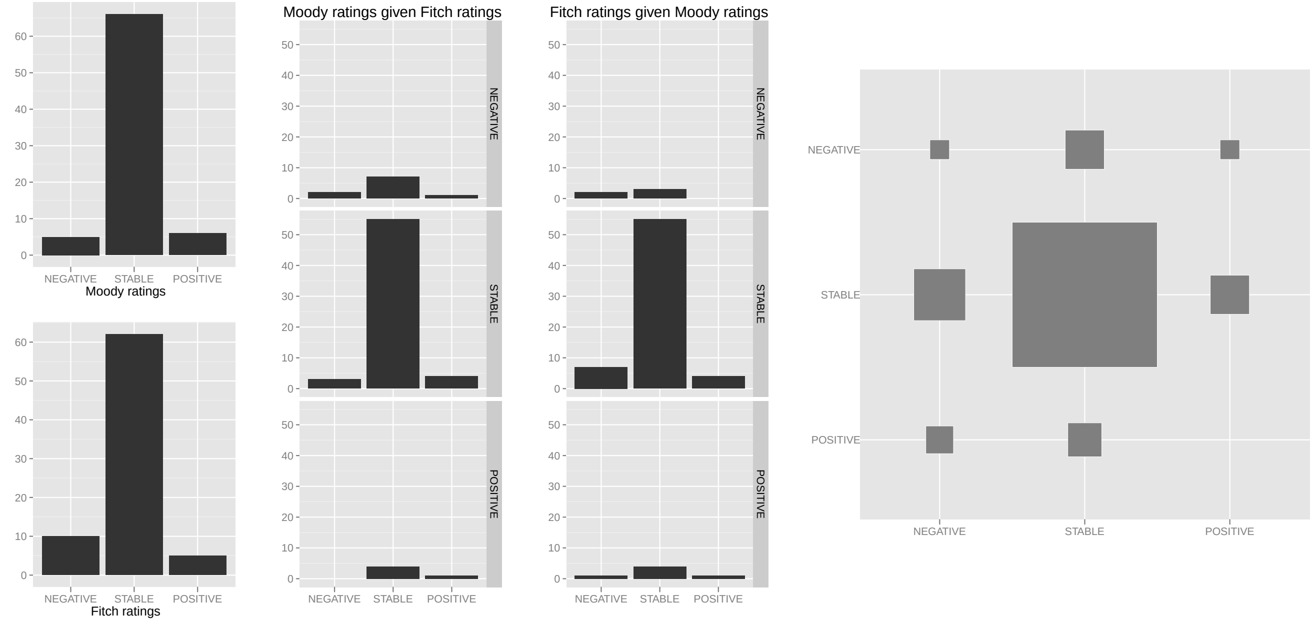

Figure 2: Barcharts for the Moody's and Fitch ratings (on the far left). Multiple barcharts of the Moody's ratings conditioned on the Fitch ratings (left of centre) and of the Fitch ratings conditioned on the Moody's ratings (right of centre). A fluctuation diagram of the Moody's and Fitch ratings. The barcharts show that the rating distributions are similar with the vast majority of ratings being 'Stable'. The multiple barcharts show that there is an association between the ratings, though it is difficult to assess because of the preponderance of 'Stable' ratings. The fluctuation diagram shows that apart from the agreement on 'Stable' ratings the two sets are almost symmetrically spread around.

More than two raters

Of the many measures for comparing two raters, there are few which have been generalised to more. Fleiss's κ and Light's κ are both available in the R package *irr* and give values for the outlook ratings in the credit rating dataset of 0.174 and 0.171 respectively. Although the values hardly differ, the p-values reported do: for Fleiss's κ we get $p < 0.001$ and for Light's κ $p = 0.62$. Presumably quite different null hypotheses are being tested! Fleiss's κ is tested against a null hypothesis of 0. The null hypothesis used for Light's κ is not stated and possibly it is testing whether the individual κ's are significantly different from one another. (Both κ's may also be calculated in the *psy* package. It uses bootstrapping to generate confidence intervals, whereas *irr* uses approximate tests.)

Two possibilities for visualising several ratings together are fluctuation diagrams and parallel coordinate plots. Fig. 3 shows a fluctuation diagram for the Moody's, Fitch and S&P ratings. This is a straightforward generalisation of the one for two raters shown in Fig. 2, but now there are 27 possible combinations instead of 9. While the plot does highlight rating combinations that occur very often (obviously the one in the middle for the 46 countries rated stable by each agency), it is difficult to grasp the overall picture. The parallel coordinate plot shown in Fig. 4 is easier to generalise, though again it is most informative when there is appropriate structure in the dataset. There is one vertical axis for each rating agency (in the order Moody's, S&P, Fitch) and the ratings are plotted by category on the axis. Each country is represented by a polyline linking the ratings it received. By sorting the countries to minimise line crossings and by ordering the axes suitably, as has been done in this display, more information can be revealed [7]. The single country (Bulgaria) with different outlook ratings from all three agencies is noticeable. As well as the big main group in the middle (all stable), other groups can also be picked out.

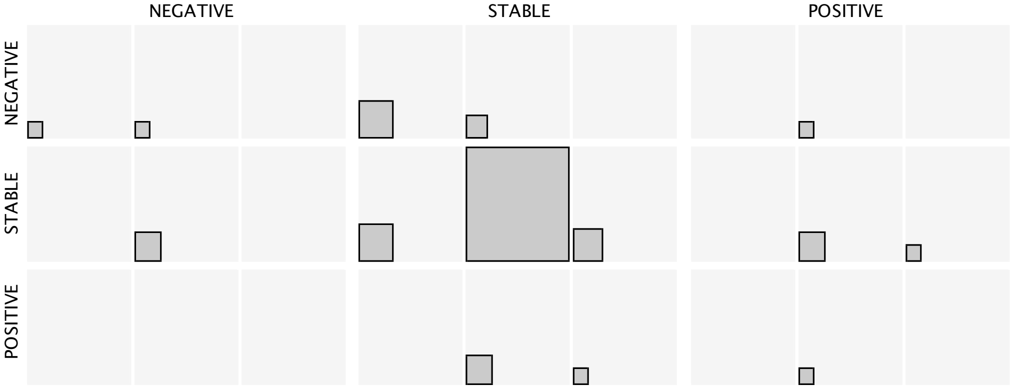

Figure 3: A fluctuation diagram of the ratings of the three agencies. The Moody's ratings are on the horizontal axis, the Fitch ratings on the vertical axis and the S&P ratings are on the horizontal axis nested conditionally within each Moody's rating. The large cell in the middle represents the 46 countries given a stable outlook by all three agencies.

The outlook ratings have been discussed at length here to illustrate the various possibilities and to emphasise that graphics tell us much more about the data than any measure of agreement. As well as being difficult to interpret and having poor statistical properties, it is unreasonable to expect them to summarise a complex structure of agreements in one single number. Before leaving the outlook ratings it has to be acknowledged that they are dependent on the main ratings. A lower main rating with a positive outlook could be regarded as equivalent to a higher main rating with a negative outlook. This will not be pursued further here, but we now move to an examination of the main ratings.

Ratings with many categories

In the dataset all three agencies used equivalent 16-step scales. Fitch and S&P go from $B-$ (worst) to AAA (best) and Moody's from $B3$ to Aaa. There are worse ratings possible, but none were used in this dataset. Barcharts for the rating distributions of the three agencies show that each used some categories more than others, but that there were no dramatic differences. The top rating was given most often by all three and there was close agreement on this, with Moody's and S&P selecting the same 16 countries, while Fitch gave its top ranking to 14 of these while lowering the other two, New Zealand and Australia, by one grade. κ and weighted κ values for the three pairs of ratings are (The approximate standard errors are all around 0.06 for the κ's and 0.16 for the weighted κ's.) The weighted κ values are much higher than the unweighted ones, because they take account of the ordering of the ratings, a far more important issue for a scale of 16 steps than for one of 3 steps (as

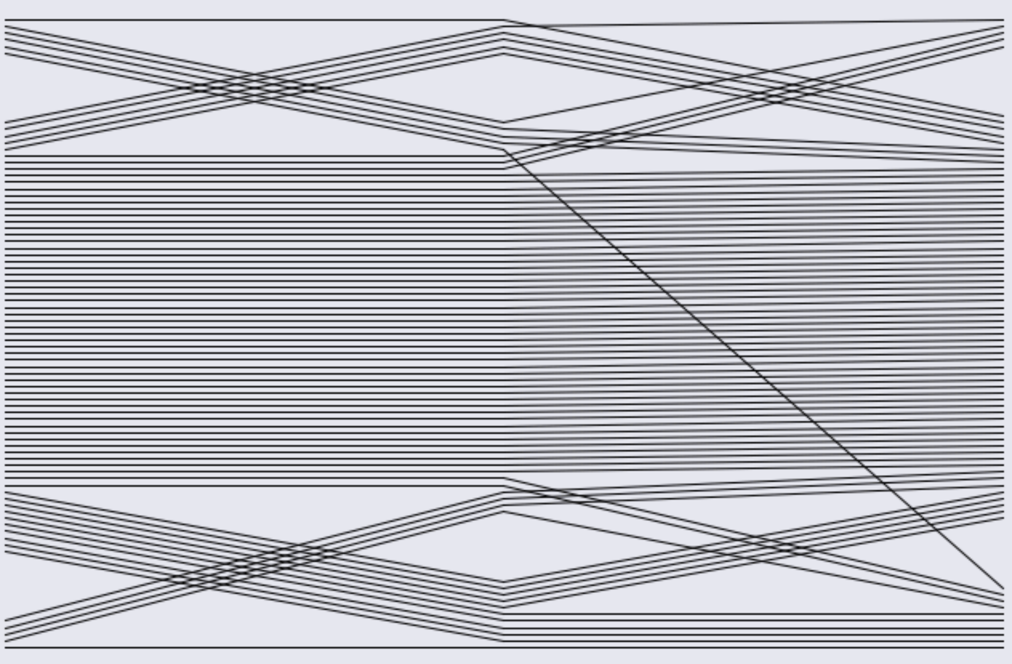

Figure 4: A cpcp (a parallel coordinate plot for continuous and categorical data) of the outlook ratings of the three agencies with the axis for Moody's to the left, the axis for S&P in the middle and the axis for Fitch to the right. The three ratings of each of the 77 countries are linked by a corresponding pair of lines. It is also clear that the differences between the ratings are at most of one level with the exception of the line crossing from top to bottom (Bulgaria).

Pairs of Agencies	κ	**weighted**(κ)
Moody and Fitch	0.49	0.90
Moody and S&P	0.48	0.86
Fitch and S&P	0.40	0.84

with the outlook ratings). The high agreement values are hardly a surprise. It would be astonishing if any country was rated substantially differently by two agencies and the maximum difference in the dataset between any two ratings is of three steps, which arises for seven countries. The typical overall pattern for two raters can be seen in Fig. 5 which is a fluctuation diagram of the Moody's and Fitch ratings. The few cases that are rated differently can be picked out.

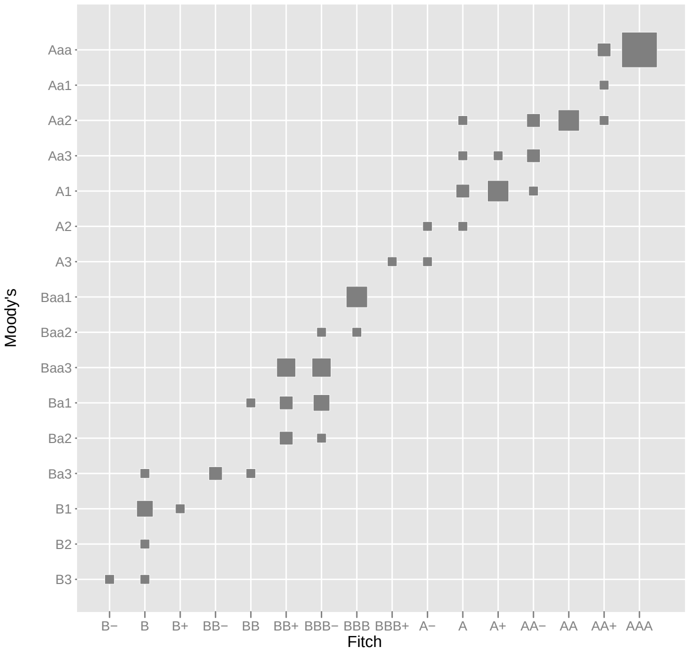

Figure 5: A fluctuation diagram of the full ratings of Moody's and Fitch. The strong agreement between the agencies is obvious. Where there are slightly bigger differences can be seen with a little more effort.

The huge number of possible combinations of three ratings each with 16 categories makes a visualisation of all in a fluctuation diagram unthinkable. However, it is also obvious that only a few of the 4096 possible combinations can arise, given that there are just 77 countries in total (in fact 53 combinations arose with 44 occurring only once, and one, all three raters giving a country the highest rating, 14 times). A parallel coordinate plot is a possible alternative, as illustrated in Fig. 6. The block of equally rated countries at the top is clearly visible, as are some of the countries labelled differently by the agencies. Note that for this plot the axes have been ordered and the cases ordered within each rating step to improve the clarity of the representation. This form of plot can easily be extended to more raters. It benefits from using interactivity to query cases of interest and to link cases to other plots of the data. Lines can be selected to identify countries by querying or linking to a barchart. Alternatively selections can be made in other graphics of the dataset and highlighted in the parallel coordinate plot.

Another approach is to look at either a frequency table or barchart of the combinations that do occur. That can be insightful, especially if there are only a few such concentrations, but does not take account of the ordering of the categories.

MS diagnosis

The credit rating dataset provides examples where agreement is very strong. It would be mis-

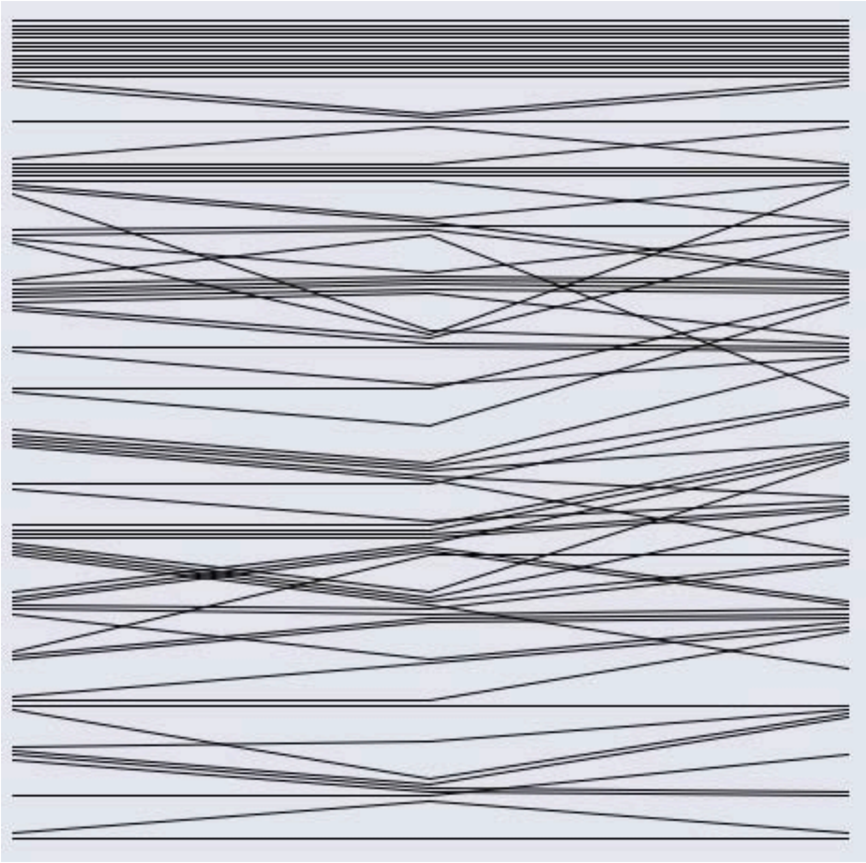

Figure 6: A cpcp of the full ratings by the three agencies with the axis for Moody's to the left, the axis for Fitch in the middle and the axis for S&P to the right. Lowest ratings are at the bottom, highest at the top.

leading to publish an article on agreement amongst experts and not include any examples exhibiting weak agreement. In the R package *vcd* there is a dataset entitled MSPatients, which records how two neurologists classified their patients. Fig. 7 shows the ratings of the Winnipeg neurologist in the rows and of the New Orleans neurologist in the columns for the group of Winnipeg patients, firstly in a Bangdiwala plot and then in a fluctuation diagram. The κ and weighted κ values are 0.21 and 0.38 respectively.

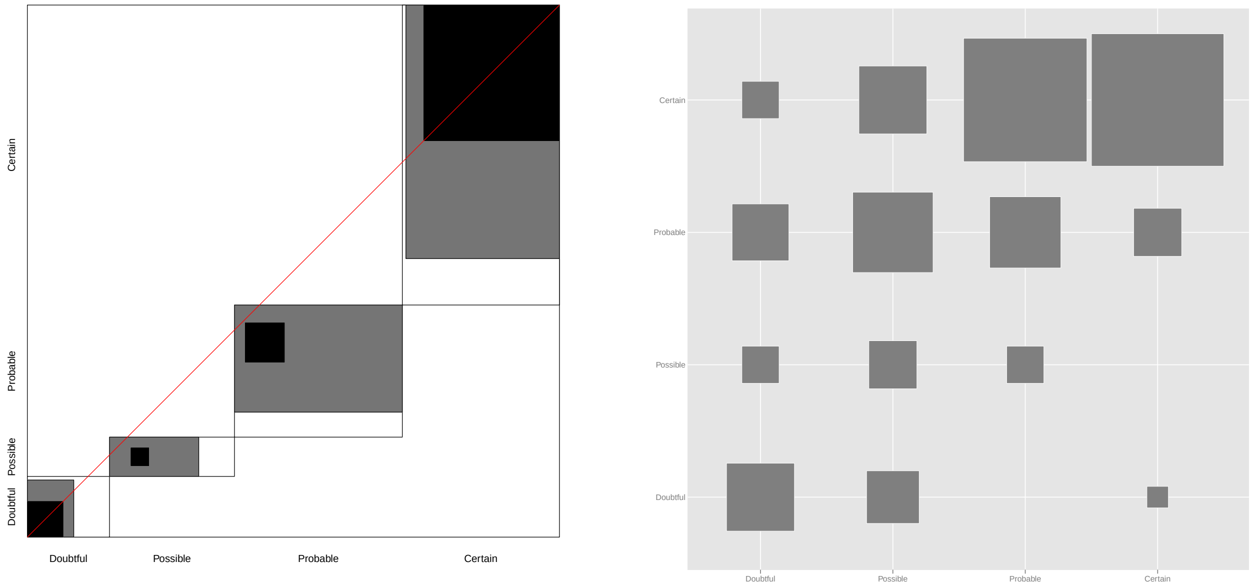

Figure 7: MS ratings of two neurologists compared in a Bangdiwala plot on the left and in a fluctuation diagram on the right. The Winnipeg neurologist's ratings are in the rows and the New Orleans neurologist's ratings in the columns. The levels are in order: Doubtful, Possible, Probable, Certain.

Fig. 7 emphasises the difference between the two kinds of plot. In the Bangdiwala plot the cell positions depend on the marginal distributions, whereas in the fluctuation diagram each cell has a fixed grid position whatever the marginal distributions. The Bangdiwala plot only shows the raw data for the diagonal entries and the cell size is proportional to the square of its count, while the fluctuation diagram shows all the raw data and each cell size is directly proportional to the count. Both plots make clear the asymmetry of the ratings, that the Winnipeg neurologist rates the patients as more likely to be afflicted than the New Orleans neurologist. The Bangdiwala plot aims to emphasise the level of agreement and displays the components of the associated agreement statistic. The fluctuation diagram shows the pattern of agreement. An alternative is to look at the conditional patterns of agreement as in the middle two displays in Fig. 2. The barcharts on the top of Fig. 8 show that unless the Winnipeg neurologist's rating is certain, the New Orleans rating is only weakly associated with it. The lower barcharts show that a rating of certain or probable from the New Orleans neurologist means the Winnipeg rating is likely to be certain. For other New Orleans ratings there is little association between them and the Winnipeg ratings.

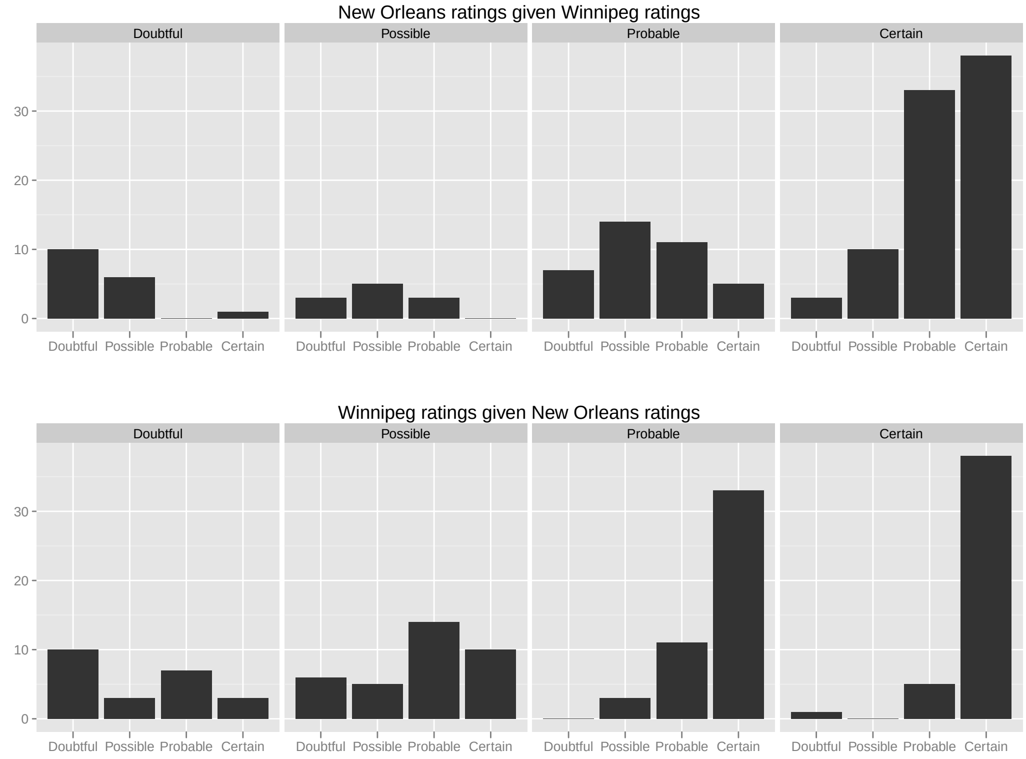

Figure 8: Conditional barcharts for the MS neurologist ratings of the Winnipeg patients.

Software

The majority of the graphics and analyses were carried out using R, especially making use of the graphics package *ggplot2* [9]. The multivariate fluctuation plot in Fig. 3 was drawn in the interactive statistical graphics software Mondrian [8] and the identification and counting of the countries in the examples was also done using Mondrian. It's easier and more flexible to query interactively than to write whatever R code might be necessary. Interactive tools are useful for reordering axes and categories too, both tasks which can be needed when evaluating data involving multiple raters.

Summary

Applications comparing ratings have been mostly in medical fields in the past. Nowadays they turn up in other fields as well and the issue of financial ratings has become progressively more important. Given the number of different rating agencies, the need to compare ratings is obvious. Measures of rating agreement fall down on three counts: they attempt to summarise all the information about agreement in a single number, they are not easy to interpret and they have poor statistical properties. Graphic displays are potentially much more informative. We recommend using a range of graphics including barcharts of the individual rater distributions, multiple conditional barcharts, fluctuation diagrams, and cpcp plots. Just as one number cannot possibly summarise a dataset, one graphic alone is rarely sufficient to reveal all the relevant information.

In the example used in this paper the agreements are very close and it is the few larger disagreements that are probably of primary interest. In other applications there can be a much broader spread of disagreement and it will be the patterns of disagreement that are of more interest. Fluctuation diagrams are an effective way of identifying which disagreements arise, at least for two raters, and give a better overall picture than Bangdiwala plots or cpcp plots. For an overview of several raters cpcp plots are the only real alternative and when coupled with interactive tools can offer many insights, even when the patterns of disagreement are complicated.

All countries have been treated equally in the analyses in this paper, whether they are a big country such as China or a small country such as Estonia. It would be interesting to look at the dataset together with weightings for the size of the countries. Whether the weighting should be population, GDP, or a measure of the size of the debt, we leave to another analysis.

References

[1] Guardian Datablog. How Fitch, Moody's and S&P rate each country's credit rating. URL `http:/www.guardian.co.uk/news/datablog/2010/apr/30/credit-ratings-country-fitch-moodys-standard`, 2010.

[2] J. de Mast. Agreement and kappa-type indices. *American Statistician*, 61(2):148–153, 2007.

[3] M. Friendly. *Visualizing Categorical Data.* SAS, Cary, N.C., 2000.

[4] K.L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61:29–48, 2008.

[5] H. Hofmann. Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1):11–26, 2000.

[6] J.R. Landis and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[7] A. Pilhoefer and A. Unwin. New approaches in visualization of categorical data R-package extracat. *(submitted to Journal of Statistical Software)*, 2011.

[8] M. Theus. Mondrian. URL `http://stats.math.uni-augsburg.de/Mondrian/`, 2005.

[9] H. Wickham. *ggplot2: Elegant graphics for data analysis.* useR. Springer, 2009.

RÉSUMÉ (ABSTRACT)

Do finanical agencies agree on the rating of countries' creditworthiness? Do doctors always agree on diagnoses of patients? Do professors agree in their evaluations of students? Comparing ratings of experts is difficult if they are not in full agreement. Various statistics have been suggested, but they do not tell the full story. They also tend to concentrate on pairs of experts, although often there may be three or even more. Graphics offer additional insight and this paper reviews several visualization possibilities and makes some recommendations.