Application of the hypothesis analysis method using Cohen's Kappa index to measure the agreement between leather sorters

Dr. Patricia Casey [1,*]***, Eng. Gustavo Altobelli*** [1]***, Tech. Pablo Pignatelli*** [1]

[1] Fonseca Tannery (Curtiembres Fonseca S.A.), Buenos Aires, Argentina.

Abstract: One of the more sensitive activities carried out in a tannery, because of the impact the results could cause on the customer and on the tannery's business is, certainly, the sorting or grading of the hides in the different stages (wet blue or wet white, crust and finished). The surface defects affect the aesthetics appearance of leather and leather goods as well as the usable area. If the grade assigned to a hide by a sorter is lower than the one it really is, according to the agreement with the customer it will cause economical damage to the tannery. On the other hand, if the grade assigned is higher than the one it really is, the customer will be disappointed and claim. The detection of defects on the surface of natural materials such as leather is a difficult task because of the variety of shapes and textures, as well as in the quality and quantity of defective areas. The most common method applied in the leather industry to sort or grade hides is visual inspection. Notwithstanding the sorting or grading criteria could be well defined and agreed between the tannery and the customer, we can always wonder: if we change from one sorter to another, which will be the concordance level existing between them? In other words, how far the sorters will coincide on their evaluation? Will the result be a real or random one? Due to the variability between two or more sorters which is traditionally recognized as an important source of mistakes, from our point of view, it would be very useful to transform an attribute analysis into a variable one. The tanneries that supply leather to the automotive companies have to certify their Quality System according to QS 9000 or ISO/TS 16949. One of the requirements in that norm and technical specification is to apply the Automotive Industry Action Group (AIAG) reference Manual "Measurement Systems Analysis" which describes the study of the measurement systems by attributes. In this paper, we describe the practical way on which the agreement between sorters is measured and the application of the hypothesis analysis method, cross tabulation, calculating Kappa coefficient initially proposed by Cohen, which is a statistical concordance rating measure for qualitative (categorical items) between two sorters.
Key words: automotive leather; grading; surface defects; sorting criteria; Cohen's Kappa index.

1 Introduction

One of the more sensitive activities carried out in a tannery, because of the impact the results could cause on the customer and on the tannery's business is, certainly, the sorting or grading of the hides in the different stages (wet blue or wet white, crust and finished).

The surface defects affect the aesthetics appearance of leather and leather goods as well as the usable area.

Depending on the final article produced and whether it will be used for shoe, garment, upholstery, etc, the type and quantity of natural defects acceptable or unacceptable by a customer change substantially.

Notwithstanding methods and machines to identify natural superficial defects on hides have been developed to sort hides automatically [1] [2] [3] [4], there are inconveniences for the application of them.

* Corresponding author, Phone: +54 11 4001-4013. E-mail: pcasey@fonseca.com.ar

According to Fraunhofer Chalmers research Centre of Industrial Mathematics, the problem of automatically classifying a hide according to quality may be divided into two steps:

1- Automatic detection of all relevant defects such as scratches, insect bites, etc., while avoiding spurious detections caused by natural irregularities of the hide (i.e. veins).
2- The result of detection algorithm is a lot of defects describing their type, intensity/severity, and position on the hide. All of this is taken into account by a grading algorithm to select quality class.

So the detection of defects on the surface of natural materials such as leather, is a difficult task because of the variety of shapes and textures, as well as in the quality and quantity of defective areas.

The most common method applied in the leather industry to sort or grade hides is visual inspection carried out by well trained and highly experienced people (sorters).

The leather grades are identified with numbers or letters. The lower the number or the letter is, the better the selection is.

So if the grade assigned to a hide by a sorter is lower than the one it really is, according to the agreement with the customer it will cause economical damage to the tannery. On the other hand, if the grade assigned is higher than the one it really is, the customer will be disappointed and claim.

Notwithstanding the sorting or grading criteria could be well defined and agreed between the tannery and the customer, we can always wonder: if we change from one sorter to another, which will be the concordance level existing between them? In other words, how far the sorters will coincide on their evaluation? Will the result be a real or random one?

Due to the variability between two or more sorters which is traditionally recognized as an important source of mistakes, from our point of view, it would be very useful to transform an attribute analysis into a variable one.

The tanneries that supply leather to the automotive companies have to certify their Quality System according to QS 9000 or ISO/TS 16949. One of the requirements in that norm and Technical Specification is to apply the Automotive Industry Action Group (AIAG) reference Manual "Measurement Systems Analysis" which describes the study of the measurement systems by attributes.

In this paper, we describe the practical way on which agreement between sorters is measured and the application of the hypothesis analysis method, cross tabulation, calculating Kappa coefficient initially proposed by Cohen, which is a statistical concordance rating measure for qualitative (categorical items) between two sorters.

2 Definition of Kappa index or coefficient and Effectiveness

2.1 Kappa measures the concordance between the evaluations carried out by two appraisers when both classify the same object.

The first evidence of Cohen's Kappa in print can be attributed to Galton (1892) [5]. The seminal paper introducing Kappa as a new technique was published by Jacob Cohen in the Journal Educational and Psychological Measurement in 1960 [6].

Kappa is defined as follows:

$$\mathbf{K = (\mathit{po} - \mathit{pe}) / (1 - \mathit{pe})}$$

Where: *po* is the relative observed agreement among raters

pe is the hypothetical probability of agreement by chance, using the observed data to calculate the probabilities of each observer randomly assigning each category.

1 – *pe* represents the margin of possible agreement not due to hazard

If K = 1 , it means the concordance is perfect, the raters are in complete agreement.

If K = 0 , it means that there is no agreement between the raters (other than what should be expected by chance).

Kappa does not take into account the level of agreement between both appraisers, it only analyzes if there is concordance or not.

This analysis indicates if all the appraisers show good agreement between each other but does not tell us how well the measurement system sorts good parts from bad ones. So for this analysis we use the Effectiveness.

A general rule applied on many working fields indicates the concordance is from good to excellent if Kappa > 0.75, while values of Kappa < 0.40 indicate a poor concordance.

On the bibliography there are other K tables available, such as the ones of Landis and Koch:

Kappa index	**Agreement Level**
< 0	No Agreement, less then chance
0 – 0,20	Slight
0,21 – 0,40	Fair
0,41 – 0,60	Moderate
0,61 – 0,80	Substancial
0,81 – 1,00	Almost Perfect

2.2 Kappa calculation

To calculate Kappa index, the Cross Tabulation Method is applied.

The values identified as observed, correspond to the evaluations carried out with each characteristic (it could be approved [A] or rejected [R]).

If the “standard”, for example, assigns “grade 1” to a hide and sorter #1 assigns the same grading to the same hide, the letter “A” is written on the table for this sorter.

If the “standard”, for example, assigns “grade 1” and sorter #1 assigns the same hide “grade 2”, the letter “R” is written for this sorter.

The expected count is the hypothetical probability of agreement only by chance.

Cross Tabulation			Sorter # 1		Total
			Rejected	Approved	
Sorter #2	Rejected	Observed Expected	33 10.92	6 28.08	**39**
	Approved	Observed Expected	9 31.08	102 79.92	**111**
		Total	**42**	**108**	**150**

#	Sorter #1			Sorter #2			Sorter #3		
	C1	C2	C3	D1	D2	D3	E1	E2	E3
1	R	R	A	A	R	A	A	R	R

<u>Expected Values</u>

For example, total approved by sorter #1 (**108**) multiplied by total approved by sorter #2 (**111**), divided by total evaluated (**150**). Expected value: **79.92**.

Cross Tabulation			Sorter #1		Total
			Rejected	Approved	
Sorter #2	Rejected	Observed Expected	33 10,92	6 28,08	**39**
	Approved	Observed Expected	9 31,08	102 79,92	**111**
		Total	**42**	**108**	**150**

X

Cross Tabulation			Sorter #1		Total
			Rejected	Approved	
Sorter #2	Rejected	Observed Expected	33 10,92	6 28,08	**39**
	Approved	Observed Expected	9 31,08	102 79,92	**111**
		Total	**42**	**108**	**42**

Diagonal indicates the coincidences, with which we calculate **Kappa**.

po = (102 + 33) / 150 = 0,90 ***pe*** = (79,92 + 10,92) / 150 = 0,61

Kappa index = (0,90 – 0,61) / (1 – 0,61) = **0,75** ***Efectiveness*** = (102 + 33) x 100 / 150 = **90 %**

2.3 Efectiveness definition of the Measurement System

Effectiveness = Number of correct decisions / Total opportunities for a decision

3 Control of sorting criteria and acceptance criterias

Periodically at the tannery, the concordance on sorting criteria of the workers who carry out the sorting of wet blue, wet white, crust or finished hides to avoid discrepancies between them, is evaluated practically.

The workers who carry out this activity are qualified ones.

To carry out this work, the supervisor separates 50 hides. Each hide is identified by a number on the flesh side, from #1 to #50.

The sorter evaluates each hide according to the criteria that should be applied (which varies from wet blue, to crust or finished leather and also depending on the final article) and writes on a form the selection or grade assigned according to his opinion.

Once the sorter finishes the operation, he delivers the form to the supervisor and the next sorter repeats the operation in the same way.

The supervisor calculates the percentage of hides each grade according to each sorters opinion; then compares the results of both sorters and calculates the percentage of difference between them, taking as a standard the most experienced sorter on the work.

If the differences, in all selections or grades, are equal or less than 10%, it is considered that the two sorters maintain the same selection criteria.

If the difference between both sorters is higher than 10%, they repeat the exercise.

Due there is no measurement instrument or device to let us know if the grade assigned by a sorter is correct or not, the supervisor´s decision is considered as “standard”.

To do this work, the sorting operation has been done by three sorters and each one repeated the operation three times.

4 Experimental

4.1 Control of sorting criteria

As we explained on item 3, 50 wet blue hides and 50 finished hides have been separated. The standard (supervisor) sorted them and wrote the selection on a form. Then sorters #1, #2 and #3 re-sorted the finished hides individually. Sorters #4, #5 and #6 carried out the same exercise on wet blue hides.

Results of practical control of sorting criteria:

Finished:	Sorter #1	5%
	Sorter #2	6%
	Sorter #3	8%
Wet blue:	Sorter #4	8%
	Sorter #5	9%
	Sorter #6	9%

4.2 Kappa and Efectiveness calculation

With the results of the three times sorting, we carried out Cross Tabulation, and calculated Kappa index and Efectiveness:

Results obtained on sorting finished hides

Cross Tabulation			Sorter # 1		Total
			Rejected	Approved	
Sorter #2	Rejected	Observed	33	6	**39**
		Expected	10.92	28.08	
	Approved	Observed	9	102	**111**
		Expected	31.08	79.92	
		Total	**42**	**108**	**150**

Cross Tabulation			Sorter # 1		Total
			Rejected	Approved	
Sorter #3	Rejected	Observed	18	6	**24**
		Expected	5.28	18.72	
	Approved	Observed	15	111	**126**
		Expected	27.72	98.28	
		Total	**33**	**117**	**150**

Cross Tabulation			Sorter # 2		Total
			Rejected	Approved	
Sorter #3	Rejected	Observed	12	12	**24**
		Expected	4.80	19.20	
	Approved	Observed	18	108	**126**
		Expected	25.20	100.08	
		Total	**30**	**120**	**150**

	Kappa Index			Efectiveness		
Sorter #	1	2	3	1	2	3
1	-	0.75	0.55	-	90%	86%
2	0.75	-	0.32	90%	-	80%
3	0.55	0.32	-	86%	80%	-

Results obtained on sorting wet blue hides

Cross Tabulation			Sorter # 4		Total
			Rejected	Approved	
Sorter #5	Rejected	Observed	15	12	**27**
		Expected	4.32	22.68	
	Approved	Observed	9	114	**123**
		Expected	19.68	103.32	
		Total	**24**	**126**	**150**

Cross Tabulation			Sorter # 4		Total
			Rejected	Approved	
Sorter #6	Rejected	Observed	18	15	**33**
		Expected	5.28	27.72	
	Approved	Observed	6	111	**117**
		Expected	18.72	98.28	
		Total	**24**	**126**	**150**

Cross Tabulation			Sorter # 6		Total
			Rejected	Approved	
Sorter #6	Rejected	Observed	18	15	**33**
		Expected	5.94	27.06	

	Approved	Observed Expected	9 21.06	108 95.94	**117**
		Total	**27**	**123**	**150**

	Kappa Index			**Efectiveness**		
Sorter #	4	5	6	4	5	6
4	-	0.50	0.55	-	86%	86%
5	0.50	-	0.50	86%	-	84%
6	0.55	0.50	-	86%	84%	-

5 Conclusions

Sorting of finished hides:

Using the practical method to control sorting criteria, the three sorters are approved.

Taking into acount Kappa and Efectiveness results, we can learn more:

- Sorter #1 and sorter #2 are the couple that has the higher concordance and also has the highest effectiveness percentage, so the probability of agreement only by chance is small.
- Notwithstanding sorter #3 compared with sorter #1 has an acceptable effectivity percentage, the concordance on the selection criteria is moderate.
- The discrepancy on selection criteria between sorters #2 and #3 is greater than the discrepancy both individually have with sorter #1.

This means if we have to choose a team according to the above mentioned results, the best couple is sorter #1 and sorter #2; then sorter #1 and sorter #3; but we will not work with sorter #2 and #3 together.

Sorting of wet blue hides:

Using the practical method to control sorting criteria, the three sorters are approved, but very close to the tolerance limit.

Taking into acount Kappa and Efectiveness results:

- On the three cases: sorter #4 and #5, sorters #4 and #6; and sorter #5 and #6, the concordance of selection criteria (Kappa) is similar and according to the results, Moderate. The effectiveness of the measurement system results are also similar.
- Taking into account the level of complexity to detect the defects on the wet blue is higher than on finished hides, the level of disagreement or non concordance material could be higher.

We can conclude that Kappa analysis is an interesting tool, to choose the best sorting couples.

References

[1] Hoang K., W. Wen, A. Nachimuthu, X. L. Jiang. Achieving automation in leather surface inspection. Computers In Industry, 34, pp.43-54, October 1997.

[2] Pölzleitner W., A. Niel. Automatic inspection of leather surfaces. Society of Photo optical Instrumentation Engineers, 2347, pp.50-58, November 1994.

[3] Puzicha J., M. Buhmann, Y. Rubner, C. Tomasi. Empirical Evaluation of Dissimilarly Measures for Color and Texture. Proceedings ICCV, pp.1165-1173, 1999.

[4] Roever D., W. Wen, H. Kaebernick, K. Hoang. Visual Inspection System for Leather Hide. US Patent 6 157 730, 2000.

[5] Galton, F. (1892).Finger prints. Macmillan, London.

[6] Cohen, Jacob (1960). A coefficient of agreemnt for nominal scales. Educational and Psychological Measurement, Vol. 20, N°1, pages 38-46.