Learning Algorithms for Evaluating Forensic Glass Evidence

Soyoung Park and Alicia Carriquiry

CSAFE - Iowa State University

October 24, 2018

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Acknowledgments

- We are grateful to Dr. Peter Weis for generously sharing knowledge, data and standards.
- We wish to thank Dr. Joann Buscaglia and Dr. Tatiana Trejos for their input.
- ISU database of glass fragments can be downloaded from: **www.github.com/CSAFE-ISU/AOAS-2018-glass-manuscript**.
- LA-ICP-MS measurements obtained by Dr. David Peate, University of Iowa.
- Work funded by Carriquiry's endowed President's Chair in Statistics at ISU.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Partial list of literature cited - I

- Parker and Holford, 1968. *Applied Statistics*
- Curran et al., 1997. *Science and Justice* (I, II, III)
- Curran et al., 2000. *Forensic Interpretation of Glass Evidence*
- Breiman, 2001. *Machine Learning*
- Koons and Buscaglia, 2001. *J of Forensic Sci.*
- Curran, 2003. *Int Stat Review*.
- Aitken and Lucy, 2004. *Applied Statistics*

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Partial list of literature cited - II

- Campbell et al., 2009. *Science and Justice*
- Zadora, 2009. *J of Foren Sci*
- Weis et al., 2011, *J of Anal Atom Spectrometry*
- Hepler et al., 2012. *Foren Sci Int*
- ASTM-E2330-2013, 2013. *ASTM International*
- Trejos et al., 2013, *J of Anal Atom Spectrometry*
- ASTM-E2927-2016, 2016. *ASTM International*
- Park and Carriquiry, 2018. *Annals of Appl Stat*

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Outline

- Setting up the problem
- Data
- Interpretation methods recommended in ASTM-E2330 and ASTM-E2927.
- Approach we propose.
- Comparison of methods.
- Final thoughts.

Some background information

- Glass evidence may arise when a glass object is broken during the commission of a crime.
- Small fragments can transfer to the perpetrator.

Do the fragments on the suspect come from the broken glass object at the scene?

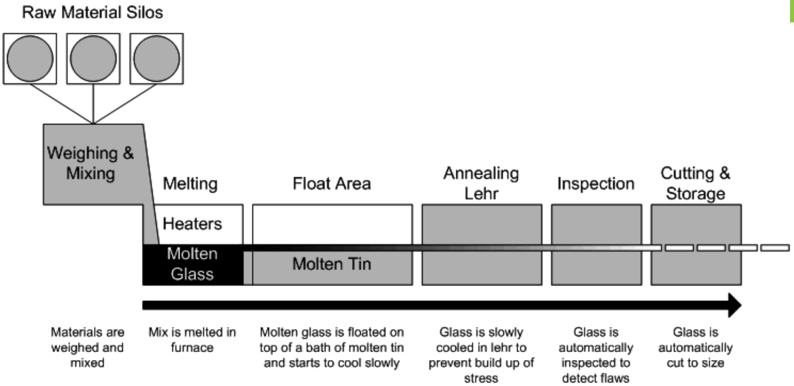
- Two related questions:
  - What is the degree of similarity between fragments on the suspect and the broken glass?
  - Is the degree of similarity *typical* among fragments from the same source?

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Glass manufacture

- Glass is made by melting together sand, soda ash, dolomite, limestone and sodium sulfate at high temps.
- Manufacturers also add *cullet* (reclycled broken glass) to the mixture.
- Float glass is produced by floating the molten mixture on a bed of liquified tin as it cools down.
- The ribbon of glass is then cut and processed for transportation.

Production of float glass (Tangram Tech)



csafe Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Properties of glass

- Physical: color, thickness, coating.
- Optical: refractive index.
- Chemical: concentration of elements in glass.
- **Elemental concentrations** can be measured using different technologies. Here we focus on LA-ICP-MS.
- Typically, 18-20 elements used to characterize glass in forensic applications.

Measurements and datasets

- Focus on the concentration (in ppm) of 18 elements in glass: Li, Na, Mg, Al, K, Ca, Ti, Mn, Fe, Rb, Sr, Zr, Ba, La, Ce, Nd, Hf, Pb.
- Method: LA-ICP-MS (details in github repository for ISU data).
- **BKA database** (Weis et al., 2011):
  - Sixty two samples of float glass with different origin, one fragment per sample, six replicate measurements.
  - One sample from VA, 34 fragments, six replicates. One fragment was measured on 11 consecutive days.

Databases (cont'd)

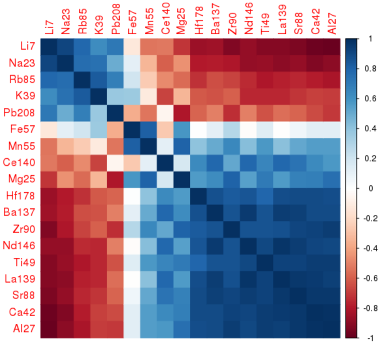
- **FIU database** (Almirall et al., circa 2002):
  - One hundred twelve samples of architectural float glass, different origin.
  - One fragment per sample, three replicate measurements.
- Fourteen elements in common with Weis et al. database.

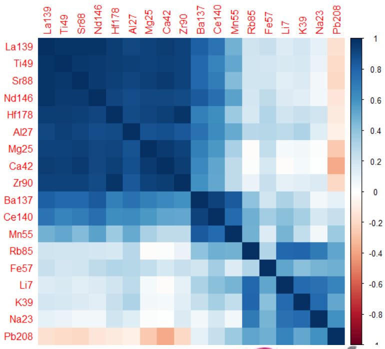
csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Datasets (cont'd)

- **ISU database** (Park and Carriquiry, 2018):
  - Thirty one samples of architectural float glass from manufacturer A and 17 samples from manufacturer B.
  - Twenty four fragments per sample.
  - Twenty one fragments replicated five times, three fragments replicated 20 times.
- Total: 1,152 fragments, 7,920 measurement vectors.
- Same elements as in BKA analyses.

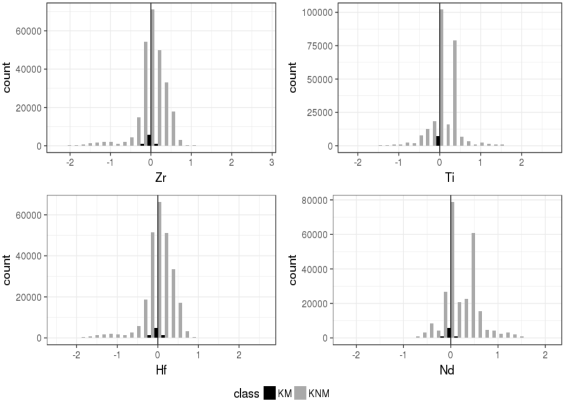
Some simple statistics



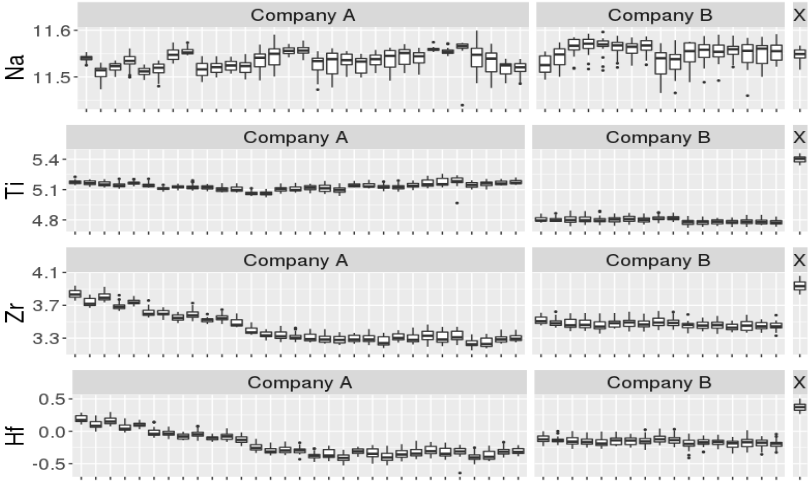


csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Some simple statistics (cont'd)



Some simple statistics (cont'd)



Analysis and interpretation of data

- Three main types of approaches:
  - **Interval-based, univariate**: Weis et al., 2011, Trejos et al., 2013. Recommended in ASTM-E2330-12 and ASTM-E2927-16.
  - Multivariate, parametric: Parker and Holford (1968), Curran and collaborators (1997, 2000, 2003, 2009...).
  - **Non-parametric, multivariate**: Zadora (2009), Park and Carriquiry (2018).

Interval-based methods

- Proceeding one element at a time, do:
  1. Obtain three replicate measurements from ≥ 3 fragments from K source and fragment from Q sample.
  2. From K fragments, compute mean and SD, and construct interval
$$\text{mean} \pm 4 \times \max(\text{SD}, 0.03\text{mean})$$
- If mean concentrations in Q fall inside the intervals *for all elements*, declare Q to be indistinguishable from the known source.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Interval-based methods (cont'd)

- Good attributes of standard interval-based method:
  - Easy to implement.
  - Does not depend on external, reference population data.
- Limitations:
  - Relies only on case work measurements, so does not address **significance of similarity between K and Q**.
  - Statistically inefficient: ignores dependencies among elements.
  - Probability that K and Q are deemed indistinguishable *increases* with noise in the measurements.

Interval-based methods (cont'd)

- The interval based method can be expressed as a score:

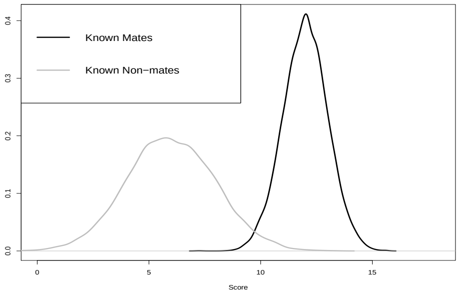
$$S_{ASTM\,i} \text{ for element }_i = \left|\frac{\text{mean in K} - \text{mean in Q}}{\max(\text{SD}, 0.03\text{mean})}\right|.$$

  And $S_{ASTM} = \max(S_{ASTM})_i$.
- **Decision rule: if $S_{ASTM} \leq 4$, samples are indistinguishable**.
- Weis et al. (2011) version: Use a fixed *relative standard deviation* (FRSD) computed from 90 mean concentrations of element obtained from DGG 1.
- If FRSD < 0.03mean $\longrightarrow$ FRSD $= 0.03$mean.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Supervised learning algorithms- General idea

- Develop a score that quantifies the similarity between two fragments of glass.
- Use the score to compare all possible pairs of fragments from a large and "representative" sample of glass fragments.



Learning algorithms - General idea (cont'd)

- **IF** (big if) the "training data" are extensive and representative and **IF** the distribution of scores is different for known mates and known non-mates, **THEN** the practitioner working on a case would:
    1. Compute the similarity score for her pair of known and question fragments. Answers the question: *How similar are these two fragments?*
    2. Compare the value of the score to the reference score distributions. Answers the question: *Is the value typical if my samples have a common source or is it alike scores observed when fragments have a different source?*

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Learning algorithms – Pros

- Advantages:
  - Quantifies the degree of similarity between two fragments.
  - Enables calculation of *probative value of the evidence*.
  - Permits calculation of "error rates".
  - Once trained, algorithm can be used to compare any pair of fragments as long as they are of the same type as the fragments in the training data.
  - Statistically more appealing – exploits dependencies across elemental concentrations.

Learning algorithms – Cons

- Limitations:
  - Performance of algorithm is **critically dependent** on training data.
  - Black box type of approach.
  - Selection of reference population of non-mated pairs requires some thought.
  - Algorithm needs to be re-trained as new population measurements become available.

Learning algorithms – How

- Given pairs of fragments for which we know "ground truth", the algorithm learns which combinations of *feature values* are associated with pairs of fragments that are known mates and known non-mates.
- Presented with a new pair (not in the training dataset), the algorithm determines whether the corresponding feature values suggest whether they are mates or not.

Learning algorithms – Features

- *Features* are measurements determined by us because we believe that they can help us classify fragments into mates/non-mates.
- We defined 18 features as follows:
  - For each fragment, take logs of all concentrations and average over replicates to get 18 mean concentrations.
  - Compute the differences of 18 means for two fragments. These differences are the features we use to classify pairs of fragments as mated or non-mated.

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Learning algorithms – Which

- There are many different supervised learning algorithms: logistic regression, support vector machines, random forests, Bayesian classification and regression trees,....
- We present results obtained using random forests. BART results were similar. (Breiman, 2001; Chipman et al., 2010).
- RFs are not difficult to implement but one must be mindful of:
  - Imbalances in the training data.
  - Independence (or close to) of units in training and testing datasets.

Implementation of RF

- Training and validation dataset: 28 panes from companies A and B produced on different dates plus the 62 panes from different sources from BKA data.
  - 7,705 pairs of mated fragments and 260,573 pairs of non-mated fragments.
- Testing dataset: 20 panes from companies A and B produced on different dates plus the pane from BKA with multiple fragments analyzed.
  - 5,590 mated fragments and over 123,805 non-mated pairs.
- Down-sampling of majority class and 10-fold validation.
- We report only Out-Of-Bag errors.

Setting up the comparison

- Focus on results from both interval-based methods and RF, using only the 5,590 + 123,805 fragments in the test dataset.
- To implement Trejos et al. (2013) and Weis et al. (2011) approaches we:
  - Selected a Q fragment at random. Averaged concentrations over five reps.
  - Selected three K fragments at random from either the same or a different pane as Q. Averaged concentrations over 15 reps.
  - Computed S_{ASTM} (as recommended in ASTM-E2927 and ASTM-E2330) and as modified in Weis et al. (2011).
  - Counted how many pairs of fragments were correctly classified.

Setting up the comparison (cont'd)

- For each fragment Q, we randomly selected 30 comparison sets, for a total of 15,300 comparisons of known mates and 150,060 comparisons of known non-mates.

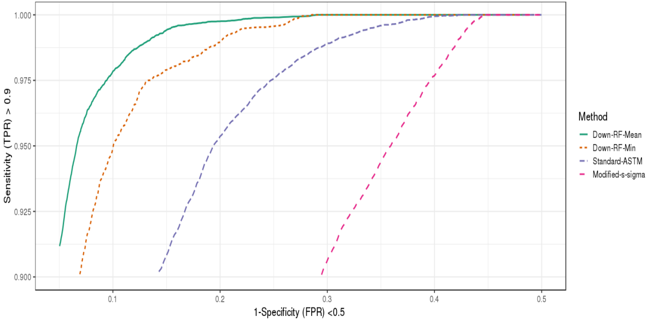
Two possibilities for the RF

- In the RF, we make pairwise comparisons.
- With three fragments (or more) from K, we can construct pairs in different ways:
  - Combine three K fragments into an “average” fragment (akin to ASTM approach).
  - Compute three similarity scores and pick one.
- Thresholds:
  - Interval-based methods: pair is mated if score is less than 4.
  - For RF, classify pair as mated if RF score is larger than 0.5.

Results

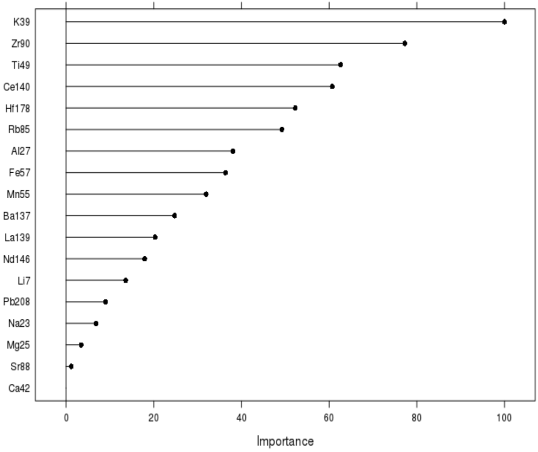
Model	AUC	EER	Opt. Threshold	FPR	FNR
RF-Mean	0.984	0.061	0.590	0.076	0.037
RF-Min	0.975	0.080	0.330	0.101	0.049
ASTM	0.954	0.122	3.300	0.142	0.0984
Weis et al.	0.899	0.204	12.961	0.298	0.096

ROC curves



csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

Feature importance



csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

When methods disagree...

Method	False Negatives	False Positives
Standard ASTM	92	7635
Random Forest	9	2299

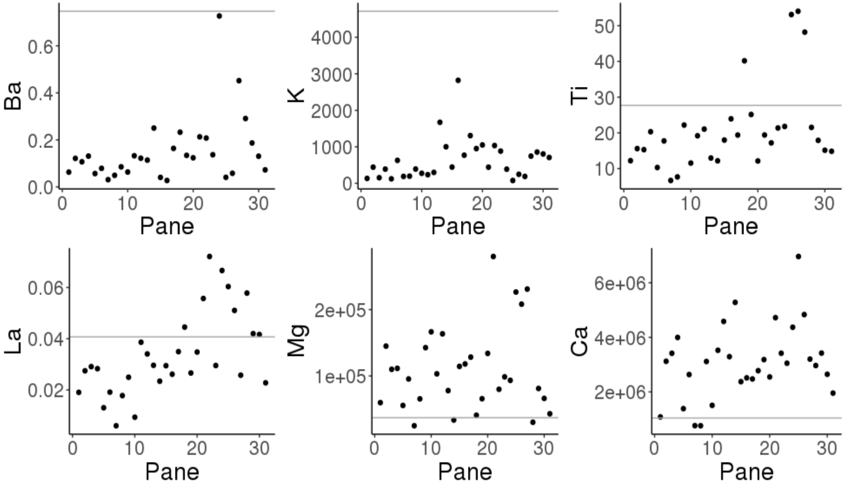
Elements driving false positives: K (28%), Li (19%), Zr (14%), Ba (9%), Ti (7%).

csafe
Center for Statistics and Applications in Forensic Evidence
ForensicStats.org

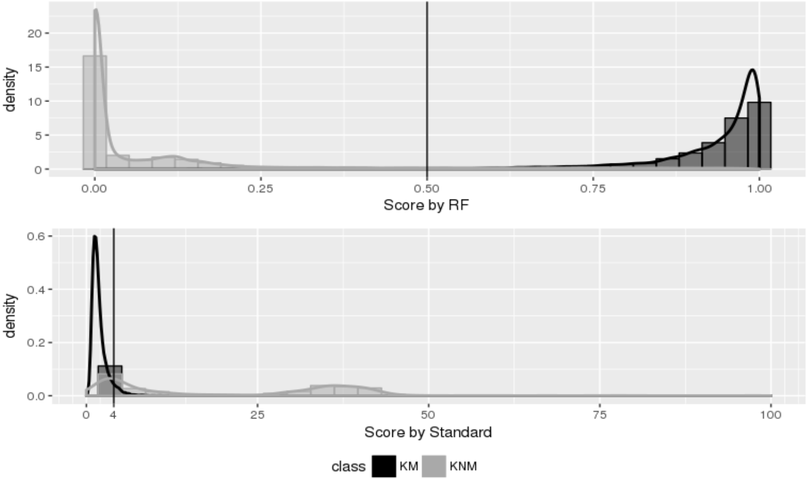
Within and between pane variances

- In these data, interval-based methods produce a high proportion of FPs (falsely declaring that two fragments are mates).
- This occurs because for some elements, the within-pane variance in concentration is *larger* than the between-pane variance.
- When this occurs, intervals get larger.

Within and between variances (cont'd)



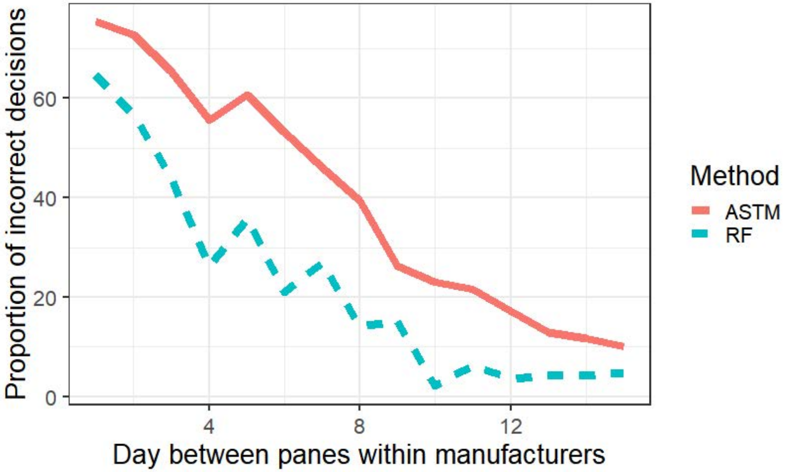
Distributions of scores



Increasing number of K fragments

	Standard 4 - σ			
Error	3 controls	6 controls	9 controls	12 controls
FNR	0.0559	0.0176	0.0067	0.0042
FPR	0.1866	0.1948	0.2017	0.2043
	Modified 4 - σ			
Error	3 controls	6 controls	9 controls	12 controls
FNR	0.4482	0.4303	0.4184	0.4203
FPR	0.0628	0.0646	0.0662	0.0674

Effect of manufacture date



Some final thoughts

- It is possible to do better than the current standards.
- Little progress will occur unless more data become PUBLICLY available. **Not sharing data impedes progress.**
- From a statistical viewpoint, glass problem is similar to bullet lead problem:
  - What do we mean by "same source"?
  - Is it possible to draw conclusions beyond "cannot exclude"?
- Interpretation sections in ASTM standards seem to be premature and can be misleading for lay persons such as jurors.

THANKS

alicia@iastate.edu
sypark@iastate.edu