# "How can we generate counterfactual explanations for ML models that predict the number of contributors to DNA samples?".

## What do experts look at when determining the NOC?

## What information helps experts decide between two NOC values?

In forensic work, DNA evidence is often analyzed using *Short Tandem Repeats (STR)*. These STR are specific tracks of repeated short DNA sequences of about two to six base pairs long, that have been proven to show high variability between individuals. These parts of the DNA or *loci* have been defined by CODIS, which is the United States national DNA database. One individual can have two different amounts of repeats for a single locus; one inherited from the mother, and one from the father. These thus represent the alleles for this certain region of the DNA. These STR are measured by a process called electrophoresis, which produces an electropherogram. An example can be seen in Figure 1. We will not go into detail about the measuring process, but will provide information about how these results are interpreted. In Figure 1, we only see the locus TH01 with two clear peaks at six and nine repeats. The repeat sequence for TH01 is AATG, so one of chromosome of this individual the AATG sequence is repeated six times on that location, while on the other chromosome it is repeated nine times. The y-axis represents the quantity of information found, measured in Relative Fluorescence Units (RFU). This is also referred to as peak height.

The first step of DNA STR profile interpretation is to determine whether a sample has originated from a single source, or if the sample is a mixture [1, 2]. This is usually easily discerned by looking at the Maximum Allele Count (MAC), which is a measure of the locus with the most alleles present. If this number is bigger than 2, the sample could be considered a mixture since a single human has at most 2 alleles at a given locus; one from the mother and one from the father. Determining the exact number of contributors is difficult, since most DNA profiles are not as clear cut. There are many factors that can obscure the number of contributors.

* Allele sharing: If two donors have the same allele at a locus, we speak of allele sharing. This frequently occurs when donors are relatives, since brothers and sisters share a lot of DNA. It might be difficult to distinguish if an allele is shared between donors, or if a single donor simply is homozygous for this allele; in both cases, the peak height for that allele is higher.
* Allele drop-out: If the DNA was degraded, for example due to sunlight, some parts of the DNA might not be present in the sample to measure. It is also possible that the amount of the DNA available is so small, that the alleles fall below a certain noise filter. Because of this low quality or quantity of DNA, some allele fragments might not show up in the profile at all, which is called drop-out.

These factors can decrease the number of alleles found in a certain profile, which could lead to an underestimation of the number of contributors. There are also factors that could lead to an overestimation of alleles present in a sample:

* Stutter or drop-in: During the process of measuring the STRs, a STR fragment can “slip” from the template. This could cause the electropherogram to measure this strand to have one repeat fewer, since the slipped part of the fragment is not correctly measured. In this way, a small stutter peak is found in the profile just before the valid peak.
* Other noise: The measuring process is not perfect, so some random noise or blobs might show up in the electropherogram, that do not contain any information about the DNA.

Stutter peaks and noise are often filtered out using certain thresholds. As a result, some DNA information might also be lost due to a low-quantity donor.

In general, it is more difficult to discern the NOC, when the number of donors increases.

It is important to make a correct assumption of the number of contributors, since the following steps rely on this number to determine correct evidence in criminal cases. When an incorrect NOC is used for further analysis involving the investigation of the DNA profiles, the results are unreliable [3]. It is possible to rerun the software with a different number of contributors, but

The Maximum Allele Count approach to determine the NOC is quite simple, but it is unreliable due to the factors discussed prior. Performance in general is quite poor, especially with 3 or more contributors [1, 4, 5]. On average, when assessing mixtures between 2-5 contributors, the MAC cannot obtain correct predictions for more than 70% of samples [6]. When looking at 4-person mixtures, more than 70% of the samples are characterized as 2-, or 3-person mixtures using only the MAC approach [5].

Often experts use MAC in combination with the Total Allele Count (TAC), which measures the total number of alleles across all loci. However, this measure suffers from the same obscuring factors as the MAC.

**nC-tool [7]:** Estimates the NOC by simulations performed on the TAC. This achieves better results than using the MAC only, obtaining correct predictions for roughly 76% of 2-5 person mixtures [6].

In 2019, a Machine Learning (ML) model was created that derived the NOC with an accuracy of roughly 82% [6]. This Random Forest (RF) model was trained based on 590 profiles of 2-5 person mixtures, obtained from **TODO:find how many**  donors. The data used for training was not the original electrophoresis results, but consists of 19 features such as the MAC, locus-specific information, and other statistical features of the data.

Allowing stutter peaks to be counted as alleles [8].

How relatives influence the LR [9]

More contributors, more likely to be estimated to have fewer NOC [10].

**Decision Tree [11]**

Derive NOC from

* Decision tree with

Tested various ML approached (RF / MLP / LDA), showing similar performance to the RF19 model. They obtained very high performance (96%) with a RFC 35 model.

Difference with [2] is “Benschop et al. used 1174 unique donors to construct 590 profiles [20], whereas the PROVEDIt dataset only had 26 unique donors within the 766 profiles used”

This means that the classifiers probably overfit to certain donors.

“In conclusion, the decision tree method for NoC assignment has been shown to be over 77% accurate, with increasing performance with improved stutter and artefact filters”

They used a decision tree to classify peaks as stutter or allele.

**Background STR mixture interpretation [12]**

Information about how statistical analysis is done to determine the LR with the Hd and Hp. Showing that the LR is still the de-facto standard method.

“The peak height information is of benefit for analyzing mixed profiles.”

“The effect of incorrect estimation of the number of donors (caused by allele sharing) to the LR value was examined by Benschop (…) and was illustrated to exert a great effect on the LR” [3]

**Background about NFI-used software for LR calculation DNAStatistX**

Shows the importance of correct NOC estimations: under-assigned number of contributors can cause the model the fail calculating the LR because the observed peaks cannot be well explained.

Also includes the NOC model + the generic RF11 model (with a lower accuracy of ~

Also includes the LoCIM method for inferring the major contributor.

# Notes metrics

* Accurate w.r.t. true effects of variables
* Speed?

## Counterfactuals [13]

“MOC returns a Pareto set of counterfactuals that represents different trade-offs between our proposed objectives, and which are constructed to be diverse in feature space.”

* Low number of feature changes (sparse explanations)
* Close to nearest observed data points (plausible explanations)

## ICE / PD plots /

ICE paper about PD plots: “Note that the approximation here is twofold: we estimate the true model with fˆ, the output of a statistical learning algorithm, and we estimate the integral over xC by averaging over the N xC values observed in the training set.”

“Visually, ICE plots disaggregate the output of classical PDPs. Rather than plot the target covariates’ average partial effect on the predicted response, we instead plot the N estimated conditional expectation curves: each reflects the predicted response as a function of covariate xS, conditional on an observed xC.”

Dependence between features must be visualized in explanations

The quality of explanations is sometimes evaluated by performing a quantitative evaluation of a user study. Users are asked to perform a certain task and the explanations help support this task. How well and how fast the humans can accomplish the task is measured as accuracy and efficiency respectively [14, 15]. Subjectively, users were asked for their preference of explanation type in a 1 versus 1 fashion and asked to provide reasons.

Explanations should have few features, as humans pick usually just a few reasons. They should be specific to the problem at hand, and every instance should be explained in the same deterministic way [16]. Deterministic, or consistent feature attributions [17].

Exploration using several visual aids [18].

The literature about explainable machine learning on tabular data often refers to domains where such tools are used in decision support settings. Domain experts are provided with an ML model which has automated their task such as determining whether or not someone is granted a loan, or whether a patient is at risk for developing cancer. The explanations are then required to help the experts determine if the prediction is trustworthy. Sometimes these explanations do not seem to correspond well with intuition. This can be caused by various underlying issues, but this is not often made clear to the user. Examples of such issues are

* The features listed should not contribute to the prediction in the way that they are shown
  + Maybe the feature is highly correlated with another feature, which causes the model to assign all contribution to one, therefore skewing perception.
  + Maybe the feature value is underrepresented in the training set, which causes the model to make decisions on little data, with as a result, a poor generalization
* The model is uncertain

Show prototypes and criticisms

Research into XAI has shown the need for comparison and evaluation of methods [19-25], and the recent interest in the implementation of counterfactual explanations [19-23]. Although there are a few key components highlighted by these surveys, they also mention that the evaluation must be done specifically to certain applications [23]. One could specify a specific goal to be achieved by the explanations which should be tested [24]. Also the relevance of explanations to a certain audience [25].

## Counterfactual explanations

1. Benschop, C., *PowerPlex Fusion 6C Profile analysis & interpretation.* 2020.

2. Heidebrecht, B.J., *Mixture Interpretation (Interpretation of Mixed DNA Profiles with STRs Only)*, in *Encyclopedia of Forensic Sciences: Second Edition*. 2013. p. 243-251.

3. Benschop, C.C.G., et al., *The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures.* Forensic Science International: Genetics, 2015. **19**: p. 92-99.

4. Haned, H., et al., *Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count?* Journal of Forensic Sciences, 2011. **56**(1): p. 23-28.

5. Paoletti, D.R., et al., *Empirical analysis of the STR profiles resulting from conceptual mixtures.* Journal of Forensic Sciences, 2005. **50**(6): p. 1361-1366.

6. Benschop, C.C.G., et al., *Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach.* Forensic Science International: Genetics, 2019. **43**: p. 102150.

7. Benschop, C., A. Backx, and T. Sijen, *Automated estimation of the number of contributors in autosomal STR profiles.* Forensic Science International: Genetics Supplement Series, 2019. **7**.

8. Gill, P., et al., *Interpretation of complex DNA profiles using empirical models and a method to measure their robustness.* Forensic Science International: Genetics, 2008. **2**(2): p. 91-103.

9. Dørum, G. and T. Egeland, *Likelihood ratios for complex mixtures with relatives.* Forensic Science International: Genetics Supplement Series, 2013. **4**(1): p. e61-e62.

10. Coble, M.D., et al., *Uncertainty in the number of contributors in the proposed new CODIS set.* Forensic Science International: Genetics, 2015. **19**: p. 207-211.

11. Kruijver, M., et al., *Estimating the number of contributors to a DNA profile using decision trees.* Forensic Science International: Genetics.

12. Tao, R., et al., *Separation/extraction, detection, and interpretation of DNA mixtures in forensic science (review).* International Journal of Legal Medicine, 2018. **132**.

13. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature – PPSN XVI*. 2020. Cham: Springer International Publishing.

14. Kim, B., O. Koyejo, and R. Khanna. *Examples are not enough, learn to criticize! Criticism for Interpretability*. in *NIPS*. 2016.

15. Ribeiro, M.T., S. Singh, and C. Guestrin. *Anchors: High-Precision Model-Agnostic Explanations*. in *AAAI*. 2018.

16. Lakkaraju, H., et al., *Interpretable & Explorable Approximations of Black Box Models.* 2017.

17. Lundberg, S. and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*. 2017.

18. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.

19. Du, M., N. Liu, and X. Hu, *Techniques for interpretable machine learning.* Communications of the ACM, 2020. **63**(1): p. 68-77.

20. Adadi, A. and M. Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).* IEEE Access, 2018. **6**: p. 52138-52160.

21. Barredo Arrieta, A., et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.* Information Fusion, 2020. **58**: p. 82-115.

22. Carvalho, D.V., E.M. Pereira, and J.S. Cardoso, *Machine learning interpretability: A survey on methods and metrics.* Electronics (Switzerland), 2019. **8**(8).

23. Gilpin, L.H., et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018.

24. Lipton, Z.C., *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.* Queue, 2018. **16**(3).

25. Murdoch, W.J., et al., *Definitions, methods, and applications in interpretable machine learning.* Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(44): p. 22071-22080.