Ink dating part II: Interpretation of results in a legal perspective

Agnès Koenig, Céline Weyermann

PII: S1355-0306(17)30092-8
DOI: doi: 10.1016/j.scijus.2017.08.003
Reference: SCIJUS 689

To appear in: *Science & Justice*

Received date: 24 January 2017
Revised date: 18 July 2017
Accepted date: 2 August 2017

Please cite this article as: Agnès Koenig, Céline Weyermann , Ink dating part II: Interpretation of results in a legal perspective. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Scijus(2017), doi: 10.1016/j.scijus.2017.08.003

INK DATING PART II: INTERPRETATION OF RESULTS IN A LEGAL PERSPECTIVE

Agnès Koenig[1], Céline Weyermann[1]

[1] Ecole des Sciences Criminelles, Université de Lausanne, Batochime 1015 Lausanne-Dorigny, Switzerland

Corresponding author: Agnès Koenig (agnes.koenig@unil.ch)

Abstract

The development of an ink dating method requires an important investment of resources in order to step from the monitoring of ink ageing on paper to the determination of the actual age of a questioned ink entry. This article aimed at developing and evaluating the potential of three interpretation models to date ink entries in a legal perspective: (1) **the threshold model** comparing analytical results to tabulated values in order to determine the maximal possible age of an ink entry, **(2) the trend tests** that focusing on the "ageing status" of an ink entry, and (3) **the likelihood ratio calculation** comparing the probabilities to observe the results under at least two alternative hypotheses. This is the first report showing ink dating interpretation results on a ballpoint be ink reference population.

In the first part of this paper three ageing parameters were selected as promising from the population of 25 ink entries aged during 4 to 304 days: the quantity of phenoxyethanol (PE), the difference between the PE quantities contained in a naturally aged sample and an artificially aged sample (R_{NORM}) and the solvent loss ratio (R%). In the current part, each model was tested using the three selected ageing parameters. Results showed that threshold definitions remains a simple model easily applicable in practice, but that the risk of false positive cannot be completely avoided without reducing significantly the feasibility of the ink dating approaches. The trend tests from the literature showed unreliable results and an alternative had to be developed yielding encouraging results. The likelihood ratio calculation introduced a degree of certainty to the ink dating conclusion in comparison to the threshold approach. The proposed remain quite simple to apply in practice, but should be further developed in order to yield reliable results in practice.

Keyword: Questioned document; ink dating; interpretation model; threshold; trend tests; likelihood ratio.

1 INTRODUCTION

The interpretation of ink dating results remains a difficult task. Indeed, once the analytical part is properly optimised and validated, a second stage consists in the development and evaluation of an adequate interpretation model in order to estimate the age of a questioned ink. The importance and complexity of this step should not be underestimated. Interpretation is highly dependent on the ink ageing processes. Practical constraints such as encountered in real caseworks should also be taken into account and an important limitation lies in the fact that the "source" ink or pen is almost never known in practical cases. Thus, interpretation models must be built from the knowledge of the ageing behaviour of a representative ink reference population [1-5]. The storage conditions of the document will also significantly influence the ageing results and should thus be considered in the evaluation [2, 3, 6-8]. Finally, the support properties might also significantly influence the ageing processes. However, the paper substrate properties can generally be determined and taken into account, when the document is transmitted for examination. An additional limitation resides in the fact that the available questioned ink entry might be small (i.e. a signature). Thus, models requiring low amount of samples are generally necessary to be applicable in most cases.

So far three types of interpretation models have been proposed for ink dating: **threshold models** [2, 3, 5, 9, 10], **trend tests** [2, 6, 11] and **likelihood ratios** [2, 13]. The threshold approach compares the analytical results to tabulated values in order to determine the maximal possible age of an ink entry (i.e. the ink entry is younger than x days old). The trend tests approach focuses on the "ageing status" of an ink entry, based on the fact that it is possible to differentiate an ink entry that is still ageing, from one that has stopped ageing (i.e. the ageing curve has levelled-off) [2, 6, 11]. The last approach is based on the calculation of a likelihood ratio in order to compare at least two alternative hypotheses about the age of an ink entry (i.e. the results support the hypothesis that the ink is x days old rather than y days old) [2, 13]. While each model might have advantages and disadvantages, they were hardly evaluated on representative ink populations in published studies. Quantitative data were published only for two D% thresholds values without much detail on the chosen ink population: 50 inks were used to define an eight months threshold and 30 inks were used for a 2 years threshold [[4]]. To the authors' knowledge, no other publication reported detailed results for the definition of their interpretation approaches. While likelihood ratios and trend tests were previously proposed and discussed [2, 13], no model was actually developed and tested on real data. Few earlier studies reported blind testing [4, 12] and this should ideally be performed yearly through external proficiency testing for any analysis method applied on casework specimens [13].

This study aims at evaluating and comparing the different interpretation models based on the results obtained from the analysis of ink entries from 25 ballpoint pens over 304 days. Three ageing parameters were selected as particularly promising from observations made in the first part of this article: the quantity of phenoxyethanol (PE), the difference between the PE quantities contained in a naturally aged sample and an artificially aged sample (R_{NORM}) and a solvent loss ratio (R%) previously defined in the literature [[3, 9, 10]]. The development of the interpretation methods were first discussed

in respective chapters: PE Quantity, R% and R_{NORM}. Answers were given to questions such how can a threshold be defined, how can a trend be efficiently detected, how can the probability densities be estimated? The number of true positive (i.e., feasibility of the approach) and the number of false positive (i.e., number of erroneous conclusion) were calculated and discussed. Each model was then discussed with practical considerations in mind (such as usual time range, amount of questioned sample available, availability of reference data). The most promising ageing parameter and models were discussed in view of the obtained results.

2 DATA

The different interpretation models were evaluated using three ageing parameters calculated from a population of 25 different inks. These inks were chosen because they covered a large range of ageing behaviours. They were provided by the LKA Munich that possesses a large collection of inks from several countries [14]. Ink lines aged during 4, 8, 23, 39, 52, 77, 101, 138, 165, 227, 274, and 304 days were analysed using liquid extraction followed by GC/MS. The following ageing parameters were chosen as the most promising according to the results presented in part 1 of this article [1]:

1. The PE quantity (PE_n) contained in the ink line in ng/cm.

2. The difference between the quantities of PE in a natural sample (PE_n) and in an artificially aged sample (PE_h) :

$$R_{\mathrm{NORM}} = \mathrm{PE_n} - \mathrm{PE_h} \qquad \text{Equation 1}$$

3. The so-called *solvent loss ratio* expressed in %, for which the R_{NORM} value is divided by the PE_n:

$$R\% = \frac{PE_n - PE_h}{PE_n} \times 100 \qquad \text{Equation 2}$$

The PE quantity proved easy to apply and showed a significant ageing tendency over time. Moreover, it only requires one sample of ink (i.e. 1cm). R_{NORM} values presented an ageing tendency over a longer period of time and showed reproducible results. While less reproducible, R% proved to be the only parameter able to work for ink having low initial PE quantities. However, both R_{NORM} and R% need the collection of two ink samples (i.e. 2 cm). Their potential to date inks in actual legal context will be further evaluated using the following interpretation models.

3 THRESHOLDS APPROACH

The threshold model was the first model proposed to interpret ink dating results based on solvent analysis [2, 3, 7, 9, 10]. The analytical results were used to calculate an ageing parameter and the values were then confronted to tabulated thresholds values, allowing inference of the maximal age of an ink entry. These thresholds were defined for a specific ageing parameter at a specific age in

specific conditions. They were mainly reported in the literature for the parameter R% [3, 4, 9, 10, 15], as well as for a second parameter called D% [3, 4, 15, 16]. While there was no thresholds reported for PE and R_{NORM} parameters, they can theoretically be determined. First proposed R% thresholds stated, that for values above 20%, the ink was still "fresh" [3]. More precise thresholds were later defined [9, 10]:

• if R%-values ≥ 50%, then the questioned ink entry is younger than 150 days,

• if R%-values ≥ 25%, then the questioned ink entry is younger than 300 days.

The 25% threshold was later revised as follows [9]:

• if R%-values ≥ 35%, then the questioned ink entry is younger than 18 months (549 days).

It is also generally agreed (but unpublished) that if R%-values are under the minimal threshold, then no conclusion can be drawn.

There is generally few information available about the process behind the definition of given thresholds. Only Aginsky gave examples for the 8 and 24 months thresholds related to the D% ageing parameter (not tested in this study) [4, 15]. Ink samples from a reference population, between 30 and 50 different inks on different papers stored in "normal" conditions, were analysed at different ages (6, 8, 12, 18 and 24 months). The mean (μ) and standard deviation (σ) of the obtained values at each age (t) were calculated and the following thresholds were obtained [4]:

$$Threshold\ AP_{(t\ days)} = \mu(AP_{t\ days}) + 3\sigma(AP_{t\ days}) \qquad \text{Equation 3}$$

As there is no threshold for PE quantities and R_{NORM}, they will be calculated using this method for results obtained for 138 days-old ink entries, because these are the closest measurements from 150 days (i.e. age of the already defined R ≥ 50% threshold from the literature). This method focuses on punctual threshold calculation for a given age in order to detect potential anachronism. However, the ageing parameters considered in this study all showed non-negligible variability between type of inks and this could yield false positive results [7]. In order to take this issue into consideration, the mean and standard deviation of samples lower than the threshold age were also calculated and plotted as a function of time [1]. Using these regressions, average mean and variance can be extrapolated for any chosen age and can be used to define threshold values. This allows comparison between different sets of data, even if different points in time were initially measured (i.e., our dataset can be used to extrapolate a 150 days threshold value and compare it to the one calculated by Gaudreau and Brazeau [9, 10]).

It is also interesting to optimise the definition of decision threshold values in order to minimise the rate of false positives (Figure 1), e.g. the number of results that are above the threshold, while older than 138 or 150 days depending on the threshold used, as previously done for drug profiling and ink differentiation [17, 18]. Thus, the dataset was split in two populations: the ink samples aged from 4 to 138 days (i.e. younger than 150 days old) and samples from 165 to 304 days (i.e. older than 150 days

old). The numbers of ink measurements showing a given value were then plotted as a function of the ageing parameter values (Figure 1). Ideally, both populations would show no overlapping. The decision threshold would then be defined somewhere between both population and would allow obtaining 100% of correct results. Unfortunately, this is rarely the case using real samples. Thus, the two populations of interest generally show some overlapping and the threshold must be defined by optimizing the following values (see Figure 1):

- The **rate of true positives**: number of answers below the calculated threshold for ink entries younger than 150 days,
- The **rate of false positives:** number of answers below the calculated threshold for ink entries older than 150 days,
- The **rate of true negatives**: number of answers above the threshold for ink entries older than 150 days,
- The **rate of false negatives**: number of answers above the threshold for ink entries older than 150 days.

The selection of the decision threshold will be highly dependent on the questions asked and the particular context of the forensic examination. In ink dating, this threshold should be settled to avoid false positives in order to minimize erroneous evaluation of the results. Ideally, the threshold should also minimize false negatives in order to increase the feasibility of the dating approach (i.e. increase the number of ink for which an answer can be given = true positive). However, in practice decreasing a rate of false results generally means increasing the other one and a compromise must sometimes be reached.

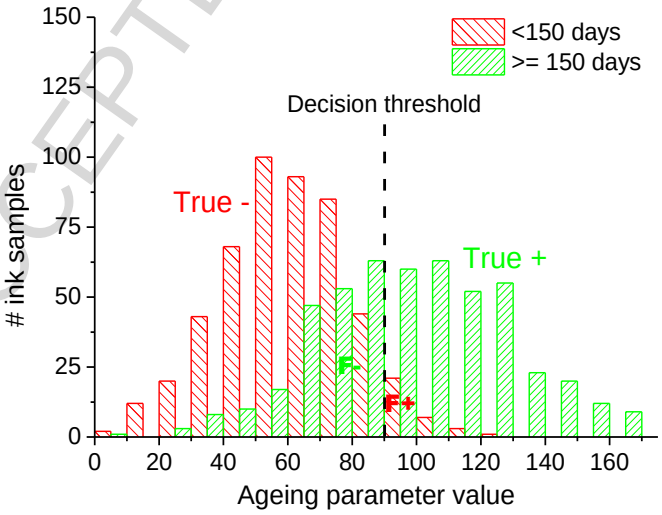

Figure 1 : Threshold definition to interpret ink dating. **True +**, true positive, **True -**: true negative, **F+**: False positive, **F-**: false negative

3.1 PE QUANTITY

For the PE quantity, no threshold values were proposed in the literature so far. Using equation 3, the following threshold value was calculated using 138 days-old-samples (μ=29 ng/cm; σ=26 ng/cm):

• *If PE quantity is ≥106 ng/cm, then the questioned ink entry is younger than 138 days.*

This threshold yielded a **false positive rate of 1 %**, i.e. there is one ink sample containing more than 106 ng/cm at 165 days. This false positive was found for an ink entry having an age close to the threshold age but slightly older. The **rate of true positive was of 12%** (indication of feasibility): i.e., these entries came from 7 different inks (Table 1). The appearance of a false positive result within the ink population used to determine the threshold showed that the method based on equation 3 may not be entirely adequate to define reliable threshold values.

The use of the following regressions modelling the mean (μ) and variance (σ^2) of PE quantity as a function of time might be more adequate:

$$\mu(\mathrm{PE}) = 17.2 + 66.2 \times \exp\left({}^{-t}/_{2.6}\right) + 39.9 \times \exp\left({}^{-t}/_{152.2}\right) \quad \text{Equation 4}$$

$$\sigma^2(\mathrm{PE}) = 339.0 + 2268.7 \times \exp\left({}^{-t}/_{98.2}\right) \quad \text{Equation 5}$$

This allowed extrapolating the average mean and standard deviation at any chosen time t [1]. A mean of 33 ng/cm and a standard deviation of 30 ng/cm (square root of 895) were obtained for 138 days, and the following threshold was defined:

• *if PE quantity is ≥123 ng/cm, then the questioned ink entry is younger than 138 days.*

This threshold is significantly higher than the previous threshold quantity. **No false positives** were detected. However, it also presented a lower rate of **true positives of 9** % (Table 1 and Figure 2) corresponding to 17 ink entries coming from 6 different inks. While this threshold slightly decreased the feasibility, it also allowed avoiding false positive results.

Empirically, the two ink entries populations (ink entries less or equal than 138 days versus ink entries older than 138 days) showed a huge overlapping up to 110 ng/cm of PE (Figure 2). If the threshold was settled at this value (i.e. at which the overlapping stopped), then no false positive was detected and the **rate of true positives reached 12%** (23 ink entries coming from 7 inks). However, in order to take into account a security margin, the threshold could be settled on higher PE quantities (e.g. 115 ng/cm), but this will in turn decrease the rate of true positive results (e.g. 11%).

Considering practical implications, a 12% true positive rate represents a relatively low rate of success, i.e. the method is working only for 7 of the 25 measured inks. By observing the results per age, it was observed that the younger was the ink sample, the larger was the number of ink presenting a positive result. Indeed, seven 4-days-old inks yielded positive results, while only two 39-days-old inks yielded positive results. Thus, documents sent early after the presumed counterfeiting would present the best

chance of success. Additionally, other threshold can be defined in order to increase the feasibility. Thus, the use of threshold for older inks would allow increasing slightly the rate of true positive and the number of ink that would give positive results. For example, the definition of the following threshold (using the empirical approach):

• *if PE quantity is ≥85 ng/cm, then the questioned ink entry is younger than 274 days.*

allowed to increase the rate of true positive up to 15%, while keeping zero false positive (Table 2). This would represent 42 ink entries made by 8 inks. This indicates that an older threshold gave better chances of success, especially for very fresh samples. However, the threshold here is only indicative as very few ink samples (n=25) were older than 274 days. On the other hand, thresholds set for inks younger than 138 days will decrease the feasibility. For example, the following threshold (empirical approach):

• *if PE quantity is ≥140 ng/cm, then the questioned ink entry is younger than 39 days.*

yielded a positive rate of 9% (Table 2).

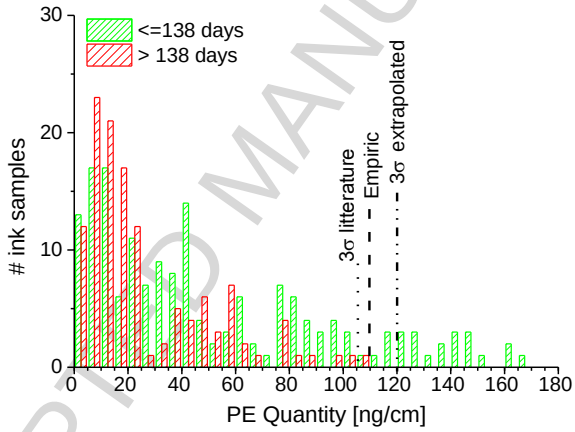

Figure 2: PE values obtained for the reference ink population (25 inks). The ink entries older than 138 days are marked in green and the ink entries younger than 138 days are marked in red. Three different thresholds were defined. The 3d threshold (literature) gave a false positive result (red box on the right of the threshold line).

3.2 R%

For R%, the following threshold reported in the literature could be evaluated on the ink population collected in this research:

• *If the R%-value ≥50%, then the ink is younger than 150 days.*

This threshold showed a high **rate of true positive of 30%** (Table 1). This represented 57 ink entries from 17 inks. Unfortunately, this threshold also led to a general rate **of false positives of 5%** (Figure 3). Indeed, R%-values above 50% were measured for 5 ink entries from 3 different inks. Those were 165, 227 and 274 days-old (Table 1). These inks were all relatively slow ageing inks and in contrary to

the PE quantity, false positives were spread over a wider time range above the threshold. This might be much more problematic in casework as slightly increasing the threshold "age" will not resolved the issue.

The presence of false positives when using published threshold can be due to several factors, such as geographical differences between the ink populations and/or differences in the ways ink samples were prepared and stored during the studies. Thus, thresholds were recalculated using the current population for 138 days-old samples. As the standard deviation calculated for the R%-values in this study was extremely high (μ = 26%, σ= 26%), the calculated R%-threshold for sample younger than 138 days, yielded a R%-value above the maximal possible value (eq. 3):

• *If the R%-value is above or equal to 105%, then the ink is younger than 138 days.*

Such a value is impossible and will never be reached in casework. The maximal possible value will never exceed 100% and values up to 73% were actually obtained in this work for 4 days-old samples [1].

By using the regressions method to evaluate the mean and standard deviation (μ= 28% , σ= 21%), the following threshold was defined [1]:

• *If the R%-value is above or equal to 91%, then the ink is younger than or equal to 138 days.*

This threshold would avoid all false positive, but would also reduce drastically the true positive rate as no inks showed such values in the whole dataset even at 4 days (Table 1). A comparable value of 91% would be obtained if a 150 days-old threshold was extrapolated. Again, such threshold would be useless to interpret ink dating results. Results demonstrated that equation 3 is inadequate to define the R% threshold values. This is due to the high variability of this ageing parameter.

While the population largely overlapped, using the empirical method allowed defining the following threshold (Figure 3):

• *If the R%-value is above or equal to 60%, then the ink is younger than to 138 days*.

No false positive were detected, and a **rate of true positives of 9% was obtained.** Thus, the feasibility is slightly lower than for the PE quantity. It corresponded to 18 ink entries coming from 10 different inks, namely less samples than the PE quantity but more inks. This was due to the fact that the R% were not completely correlated to the PE quantities and inks containing less PE could present a high R% value, especially for young samples [1].

By evaluating the dataset per sample age, it was observed that the positive samples were mainly aged of 4 days (10 inks). This number decreased quickly to 2-3 inks up to the threshold (Table 2). Considering an older age threshold with the empirical method yielded better feasibility (Figure 3):

• *If the R%-value is above or equal to 50%, then the ink is younger than or equal to 274 days* (empirical method).

A general rate of true positive of 24% would be obtained. This would correspond to 62 ink entries from 17 inks allowing comparable feasibility that the 50% threshold defined in the literature and showing no false positive results (Table 2). A younger threshold age as e.g. 39 days would lead to a true positive rate of only 6% .

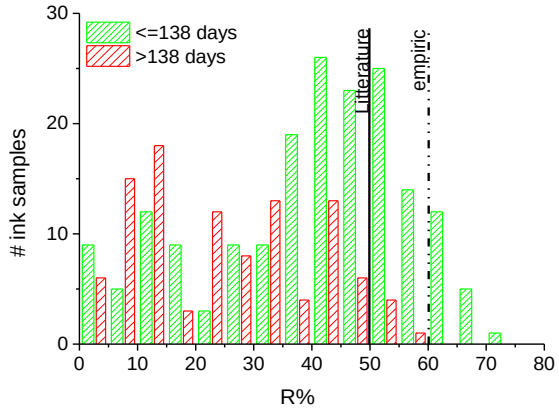

Figure 3 : R% values obtained for the reference ink population. The literature threshold gave 5 false positive results and a new empirical threshold would thus be more appropriate. However, it also reduce significantly the feasibility of this ageing parameter.

In conclusion, the 50%-threshold proposed in the literature was not adequate on the studied population and should not be transferred from one laboratory to another without further studies. These results challenge the robustness of this kind of interpretation model. It may even be necessary to re-evaluate continuously such threshold values with new ink samples, as the market is evolving and spatio-temporal representativeness is not guaranteed. This is not the first study questioning the threshold reliability [2, 7, 19] and it seems that their values were already adapted in subsequent studies [5, 9]. A previous study also reported false positive results for the R% parameter in the literature. Thus, two values of 38 and 35 % were reported for 2 different 7-years-old samples, yielding false positive results when using the 35%-threshold (less than 18 months) [12]. This demonstrates that relatively high R% value can be obtained for several years old inks.

3.3 R_{NORM}

For the R_{NORM} parameter, thresholds were defined in the same way than for the PE quantity. The following three 138 days old threshold were obtained (litterature 3σ method, regression 3σ approach, empirical):

• *If the R_{NORM}-value is above or equal to respectively 54, 52, and 45 ng/cm, then the ink is younger than to 138 days* (Figure 4).

All these thresholds would lead to no false positives in the ink population, while the rate of true positives would be slightly different: 12, 14 and 20% respectively (Table 3). This corresponded to 23, 26 and 38 samples made by 7, 7 and 10 inks. Comparatively to the other ageing parameter, the empirical method yielded the best rate of true positives, allowing better chance of success for the given population (Figure 4). 20% even represented a slightly better feasibility than for the PE quantity

and the R% in this study. Should the ink population increase, the results may change and the defined threshold must thus be considered as indications at this stage of research. .

The best chance of having a positive result was again observed for younger samples: 7 to 10 ink showed values above the thresholds at 4-days. However, older ink samples still showed interesting true positive rate (e.g. 1 to 4 inks for 101 days old samples).

The use of an older threshold allowed increasing the rate of true positive up to 37% (97 ink entries from 15 inks) without observing any false positive results

• *If the R_{NORM}-value is above or equal to 20 ng/cm, then the ink is younger than or equal to 274 days* (empirical method).

A threshold for an age of 39 days would in contrary decrease the rate of true positive to 12% (Table 2)

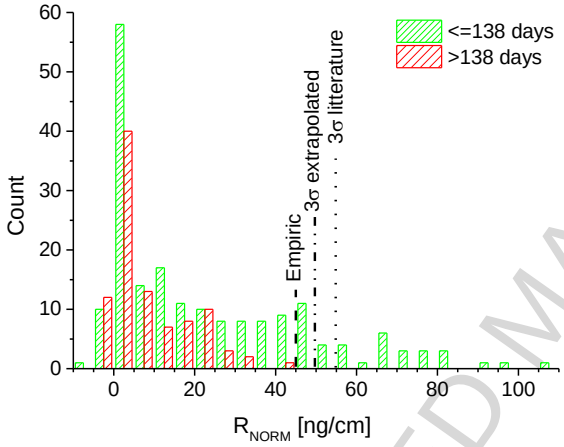

Figure 4 : PE values obtained for the reference ink population (25 inks). The red dotted line represents the 116ng/cm threshold and the red square represent the false positive area.

Table 1 : 138 days old thresholds per ageing parameter calculated with different methods: 1) 3σ method proposed in literature [4, 15], 2) 3σ method using extrapolated mean and standard deviation, 3) Empirical method. For each threshold, the general rate of true positives (V+), false positives (F+) and also the number of reference ink yielding results above the threshold for each age was calculated (in green true positives and in red false positives).

AP	Thres. value < 138 days (*)	Met.	Total results		# of inks yielding positive results per age											
			F+ (%)	T+ (%)	4	8	23	39	52	77	101	138	165	227	274	304
PE [ng/cm]	106	1	1	12	7	6	4	2	2	2	1	0	1	0	0	0
	123	2	0	9	6	2	2	2	2	2	1	0	0	0	0	0
	110	3	0	12	7	6	3	2	2	2	1	0	0	0	0	0
R% [%]	50*	Litt	5	30	17	9	6	3	8	4	6	4	1	2	2	0
	105	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	90	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	60	3	0	9	10	2	3	0	0	0	2	1	0	0	0	0
R_{NORM} [ng/cm]	54	1	0	12	7	6	4	3	2	0	1	0	0	0	0	0
	52	2	0	14	7	7	4	3	3	1	1	0	0	0	0	0
	45	3	0	20	10	7	6	3	6	2	4	0	0	0	0	0

Table 2 : Thresholds calculated for all the measured agesusing the empirical method. The true positive rate is indicated in green, as the threshold definition method is the empirical method, the rate of false positive is always 0%.

Threshold age [days]	Threshold value (True positive rate %)					
	PE Quantity [ng/cm]		R% [%]		R_{NORM} [ng/cm]	
≤ 4	170	(0%)	70	(4%)	85	(12%)
≤ 8	155	(6%)	70	(2%)	85	(6%)
≤ 23	155	(4%)	65	(8%)	75	(12%)
≤ 39	**140**	**(9%)**	**65**	**(6%)**	**70**	**(12%)**
≤ 52	130	(10%)	65	(5%)	60	(16%)
≤ 77	130	(9%)	65	(4%)	60	(13%)
≤ 101	110	(13%)	65	(4%)	45	(22%)
≤ 138	**110**	**(12%)**	**60**	**(9%)**	**45**	**(20%)**
≤ 165	105	(11%)	60	(8%)	35	(26%)
≤ 227	90	(15%)	55	(14%)	25	(32%)
≤ 274	**85**	**(15%)**	**50**	**(24%)**	**20**	**(37%)**

3.4 PRACTICAL CONSIDERATIONS

Decision thresholds are very straightforward and easy to apply in casework to estimate the maximal possible age of an ink. Indeed, the forensic expert simply compares the obtained value for the questioned ink with the tabulated threshold values. Theoretically, it is universal and case unrelated as it can be applied in every caseworks without needing additional data. However, several drawbacks were also highlighted in this study. For example, proposed thresholds from the literature yielded false positive results in this study, showing the limits of such interpretation models.

The first complexity lies in the definition of the decision threshold values. One has to minimize error rate. Thus, ideally the number of false positive results should be equal to zero. On the other hand, in order for the method to be useful in a majority of casework, the number of true positive should also be as high as possible. This means, that the number of false negative results should be close to zero. While the first criteria could be easily met, setting the threshold in order to avoid false positives also meant drastically reducing the number of true positive with a maximum of 20% of measured samples for which the 138 days old threshold gave a useful answer, corresponding to 10 inks out of 25 (40%)for the R_{NORM} parameter. .

Thus, while practical, decision threshold showed risks of false positive results that should be taken into account, as well as limitations in terms of feasibility. In order to be applicable, their definition should be based on large ink population and they should be evaluated regularly through blind testing. Finally, they should also be continuously updated over time and space to remain representative of the ink entries that may be encountered in caseworks. Conclusion should also be formulated using probabilities (or mention of possible error due to outliers). Indeed, one can never exclude a false positive result and certainty does not exist in (forensic) science [20, 21].

The selection of adequate threshold values may help decrease the risk to encounter false positive results. Logically, such results will generally appear for ages close to the threshold value (e.g. for samples slightly older than 138 days). Thus, one possibility to decrease the risk of false positive results would be to insert a time gap between the document date and the threshold age used to interpret the results. This could be done as follows, considering the defence and accusation typical proposition:

- The document was made 274 days ago or earlier (t = 274 days)
- The document was made more recently (e.g. t≤138 days)

Thus, we define two threshold values to evaluate the results by plotting the PE quantity of both ink populations (see example for R_{NORM} in Figure 5):

- the ink samples younger than or equal to 138 days (younger)

- the ink samples aged of 274 days old (older)

It can be seen that the overlapping of the two populations is thus reduced (Figure 5) as ink sample between 138 and 274 days are excluded (165 and 227 days) from the interpretation model. This should actually decrease the probability to encounter false positive near the threshold values by introducing an error margin. In the example given in Figure 2, there is two possibilities to define a threshold. The more conservative would be to use the samples of 138 days to calculate the threshold (45 ng/cm). The number of true positive would remain the same as the 138 days old threshold, namely 20%, but the risk of encountering one value close to the threshold would drastically decrease. The risk of error related to the conclusion will subsequently remain small.

A second possibility could be considered and would consist of defining a threshold between the 138 days old threshold (45 ng/cm) and the moment the number of false positive reaches zero with the 274 days old samples (25 ng/cm). Thus, for example a threshold at 30 ng/cm could be settled and this would increase the number of true positive to 27 %. The conclusion could take the form of: “An ink entry having a R_{NORM} value of more than 30 ng/cm support the hypothesis that the ink is younger than 274 days old.” The expert could add that no ink samples older than 274 days showed values above 25 ng/cm on the studied populations (here ballpoint pen entries made with 25 ballpoint pen inks and stored in a file at 23 ± 1°C). The uncertainty related to the conclusion will be higher than the previous threshold but smaller than the threshold of 274 days (table 2). By introducing the notion of alternatives hypothesis and uncertainty, the proposed solution may tend to the development of a likelihood ratio method in order to attribute a probability to the conclusion (see below).

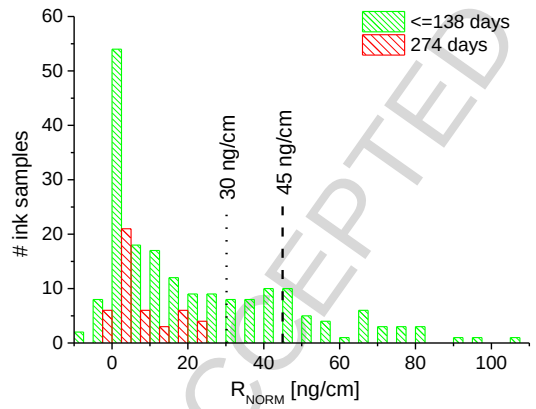

Figure 5 : Distribution of PE quantities for two populations of ink samples. The samples younger than 138 days (green) and the samples older than 274 days old (red).

4 TRENDS APPROACH

The trend test models are not based on the absolute value of the ageing parameter, but on their ageing kinetics [2, 6, 7, 11]. The aim of this model is to determine if the questioned ink is still ageing at the time of the first analysis, or if the ageing already stopped (i.e., no trend can be detected anymore). If no trend is detected, then no conclusion can be drawn. Thus, this type of interpretation requests several analyses of the questioned ink entry over time. The first analysis is carried out when the

questioned document is received in the laboratory. Then, it is repeated every two weeks in order to obtain chronological data to apply the model. Generally five measurement were advised [6, 11], but no less than four should be performed [22]. The first trend test proposed was the Neumann test for an ageing parameter called V% obtained through sequential extraction and analysis using thermodesorption GC/MS [2, 6, 11]:

$$PG = \frac{1}{(n-1)\cdot\sigma^2}\sum_{i=1}^{i=n-1}(AP - AP_{i+1})^2 \qquad \text{Equation 6}$$

, where n is the number of measurements, σ is the standard deviation measured from the data, and AP_1, AP_2,...AP_n are the chronologically ordered measurements for the ageing parameter. The obtained result PG is then confronted to a critical statistical value $X_{n,p}$ depending on the number of data points considered (e.g. n=5) and a given confidence interval (e.g. p=99.5%) [2, 11]. If the obtained PG value is smaller than the critical $X_{5,99.5\%}$ value (0.8204), the measurements indicate a significant trend, and are thus considered to be still ageing. According to the literature, this would mean that the ink is younger than 6 months old [2, 6, 11]. However, a recent study showed that this approach actually yielded false negatives, i.e. the test did not detect trends when there was visually one. This was due to the small quantity of data points considered as well as their high variations [6]. The use of seven points instead of five showed better results, but also showed limits in the application of the Neumann test for ink dating.

Another model based on the calculation of a slope was also proposed [6]. A linear slope was calculated between the data points (AP_i) and its significance was determined using a T-test:

$$-m = -\frac{n\sum_{i=1}^{n}(AP_i) - n\sum_{i=1}^{n}t_i\sum_{i=1}^{n}(AP_i)}{n\sum_{i=1}^{n}t_i^2 - \left(\sum_{i=1}^{n}t_i\right)^2} \qquad \text{Equation 7}$$

, where *AP* represented the ageing parameter value that could be the PE quantity, the R_{NORM} value, or the R%. and *t* represented the age of the different samples used in the calculation. For the trend test, the significance of the slope was determined using the following test:

$$t-test = \frac{m}{S_m} \qquad \text{Equation 8}$$

, where S_m is the error calculated on the slope.

As the dataset were not acquired to have two weeks intervals between each samples, the trend tests could not be tested as proposed in the literature. However, in order to evaluate the capacity of such tests, the Neumann and slope tests were applied on three different time ranges of the ageing curves, each containing 4 datapoints. The first time range was constituted of samples aged 8, 39, 77 and 101

days (Δ=93 days), the second time range contained sample aged 77, 101, 138 and 165 days (Δ=88 days), and the third time range contained samples aged of 227, 274 and 304 days (Δ=73 days). Logically, the first time range (8-101 days) should show an ongoing ageing for most inks, while the third (227-304 days) should level-off and show fewer trend or none. The slope tests were applied to all time ranges while the Neumann test could not be applied to the last time range due to the low sample number (n=3). Then, only the slope test results are shown and discussed here. The Neumann tests results calculated for the two first wide ranges yielded comparable results and conclusions.

4.1 PE QUANTITY

While slope values tended to decrease as a function of time (Figure 9A), the actual number of slope detected remained comparable over time (see table 3). In fact, 10 inks showed a statistical trend for the first time range of samples (8-101 days), while 10 inks and 11 inks presented trends for the second and third time range respectively (Table 4). While interesting, these results highlight that a significant amount of trends were still detected after 227 days, namely more than six months. Thus, this model cannot be applied in the time frame measured in this study (i.e., ca. 1 year).

Moreover, the results per ink showed inconsistencies (Table 4). While 7 inks presented correct trends behaviour, namely trends only for the first time range (O, Z), for the first and second time ranges (S, T, X), or for all time range (H, P), other presented illogical trend detection (Table 3). Indeed, three inks presented a trend for the first and third ranges, but not for the middle range (inks G, I, W). Some inks presented no trend for the first time range, but one or both for the subsequent ranges (A, B, E, J, K, L, M, Y). Finally, seven inks presented no trends at all (C, D, N. Q, U. V). While these results seemed to be correct for most of these inks, Ink U and N did not show trends using the Slope or Neuman tests, while the PE quantity clearly decreased in the first few months (Figure 6). A relatively high standard deviation may explain these results [6].

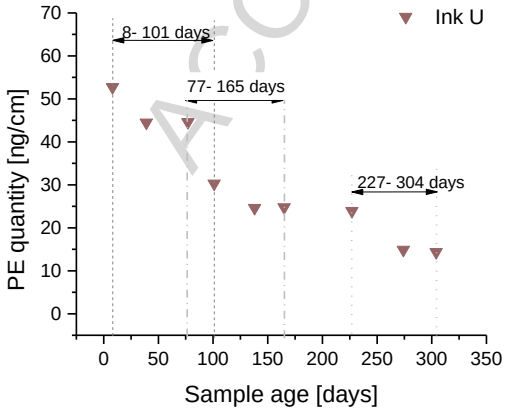

Figure 6 : Samples of ink U used for the slope test.

Globally, these results showed that no decision can be reached for sample under one year old due to the results inconsistencies. Observations will probably remain problematic after 1 year due to the specimen variations and this approach should thus be avoided for ink dating interpretation without thorough (including blind) testing on a large population of inks.

4.2 R%

For R% values, very few trends were detected on the three different time ranges. 16 inks showed no positive trends at all, while 9 inks presented only one statistical trends, unfortunately not only for the first range (3 inks) but also for the second and third range (2 and 4 inks respectively) (Table 4).

Even the mean slope values calculated for the ink population showed no decreasing as a function of time (Figure 9B). Moreover, the appearance of a trend seemed random and thus, hardly reliable (Figure 7).

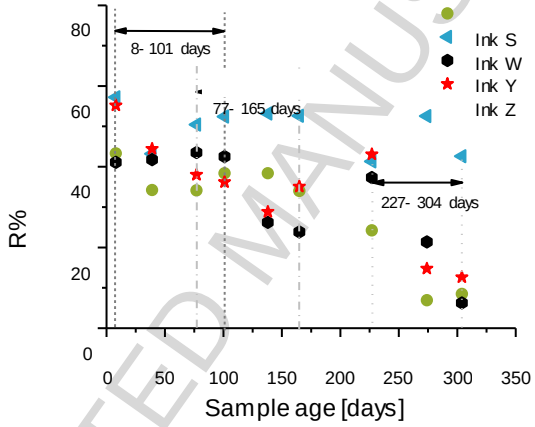

Figure 7– R% as a function of time for 4 selected inks: ink W showed no positive trends at all, ink Z showed a trend only in the 1st time range (8-101 days), ink Y showed a trend only for the 2nd and 3rd time range (77-165 days) and ink S showed a trend only for the 3rd time range (227-304 days).

The Neumann trend test presented similar issues and detected even less trends in comparison to the slope test. The inadequacy of the trend tests for the R% can be explained by the high variability of the obtained values for this parameter. The RSD can reach up to 40% for one ink (i.e. intra-ink RSD) [7]. Thus, the trend test model should also be avoided for the R% parameter.

4.3 R_{NORM}

For R_{NORM} values, 7 inks presented a trend for the first time range (8-101 days), it raised to 9 inks for the second time range (77-165 days) and decreased to 5 inks for the last time range (165-304 days) (Table 4). The mean slope values tended to slow down as a function of time (Figure 9C).

7 inks presented logical trends (trends observed for the first only or first and second time ranges or all), 10 inks presented no trend at all and 8 inks presented illogical trends. six presented only one trend in the second (inks J, Q, X, Y) or in the third time range (inks A, I), two inks presented a trend for the second and third time range but not for the first one (ink B, S) (Table 4). While most of these irregularities could be explained by the data variability, ink S and J did actually present a visible slope for the first time range that was not detected by the trend tests (Figure 8). In fact, both inks presented a R_{NORM}-value at 8 days way above the values of 39, 77, and 101 days preventing both trend tests to detect the trend as it increased the variance of the data. This phenomenon was previously reported in the literature and highlight difficulties for the application of such tests for ink dating interpretation [6]. This showed a lack of reliability and robustness of such models.

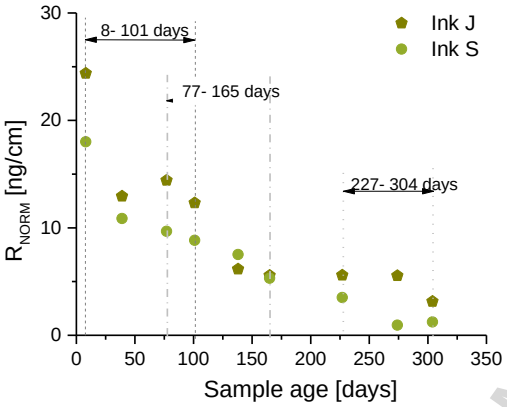

Figure 8 : R_{NORM} value as a function of time for 2 selected inks : ink S an J that showed no positive trends For the first time range but trend for the 2nd and 3rd. , ink Z showed a trend only in the 1st time range (8-101 days)

4.4 PRACTICAL CONSIDERATIONS

Trend tests are more complicated to apply in practice than decision thresholds. They necessitate several analyses over time and thus, require more samples and more time. The ageing parameters R_{NORM} and R% already necessitate the destructive analysis of two samples of 1 cm for their calculation. If 5 samples are needed over several weeks, this represents 10 cm of ink in total. As a questioned signature can be rather limited in casework, such method would rarely be possible, and would thus reduce its feasibility in practice. While theoretically universal and case unrelated, it was not possible to see reliable decrease of trends over time for any of the tested ageing parameters. In fact both trend tests previously proposed in the literature, the slope test and the Neumann test (results not shown), cannot be considered reliable for ink dating interpretation of the tested ageing parameters.

In fact, there are two main issues in the application of such tests: the detection of the trend is actually independent of the age of the sample and very dependent on the variability of the data. An alternative idea in order to apply such tests can be based on the fact that, indeed, ink ageing occur very quickly the first weeks after deposition of the ink on paper. Then, the ageing does level off over time (see data

presented in the first part of this article [1]). Thus, one could define a decision threshold based on the mean slope decrease as observed in Figure 9.

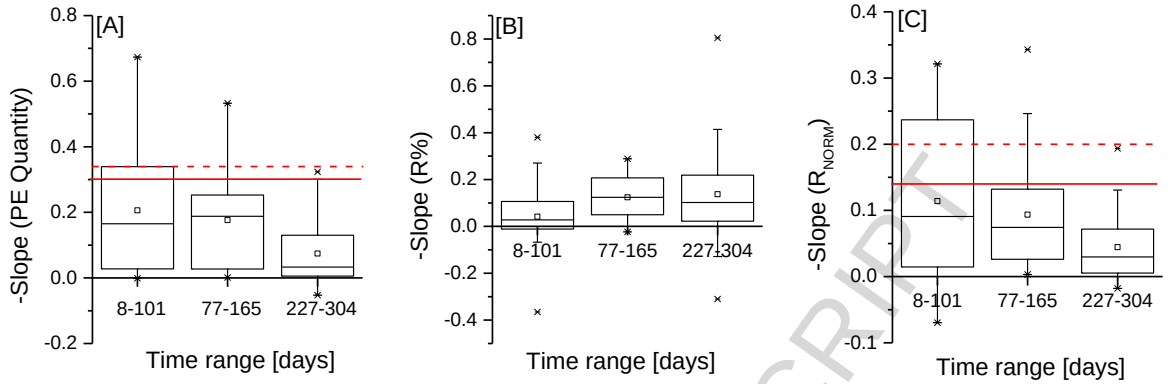

Figure 9 : –Boxplot of the –slope calculated for the three different time range for A. PE Quantity, B. R% and C. $R_{NORM.}$ the red lines represent a decision threshold based on a mean decrease of the measured slope over time. The maximal value obtained for the third ageing range is clearly below some of the values obtained for the two first ranges, allowing the definition of such a threshold for the PE quantity (A) and the R_{NORM} (C).

For PE Quantity, a threshold could be defined at a -slope of 0.34 (see red line in Figure 9A):

- If the –slope value was above 0.34 ng/cm was obtained, then the ink is less than 165 days old-

This would yield 0 false positive with a rate of true positive of 24%, corresponding to 12 samples and 9 inks (Table 3). Decreasing the threshold to 0.30 would then increase the feasibility with a true positive rate of 28%, corresponding to 14 samples and 9 inks. However, this would also yield 2 false positives (Table 3).

For R_{NORM}, a threshold could be defined at 0.20 (see red line in Figure 9C):

- If a –slope value above 0.20 was obtained, then the ink is less than 165 days old-

This would yield 0 false positive with a rate of true positive of 22%, corresponding to 11 samples and 8 inks. Decreasing the threshold to 0.14 would then increase the feasibility with a true positive rate of 34%, corresponding to 17 samples and 11 inks. However, this would also yield 1 false positive (Table 3).

For R%, such an approach is not feasible, as no mean decrease of the slopes were detected on the measured intervals (Figure 9B).

Table 3 : 165 days threshold defined for the PE Quantity and R_{NORM}. The true and false positive rate are indicated. No thresholds could be defined for the R%.

Ageing parameter	Defined thresholds -slope (empirically)	True positive rate (%)	False positive rate
PE Quantity	0.34	24	0
	0.30	28	8%
R_{NORM}	0,20	22	0
	0,14	34	4%

While further study would be needed in order to determine if smaller time ranges (e.g. 5 measurements every two weeks [citation]) would also reliably lead to a continuous diminution of the obtained slopes, such an approach would clearly be much more reliable and less sensitive to variability of the results than previously proposed approaches.

Table 4 : Summary of slope trend detection for three different time range and ageing parameters. Green = trend detected, red= no trend detected.

AP	Time range [days]	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	S	T	U	V	W	X	Y	Z	#ink
PE quantity	8-101																										10
	77-165																										10
	227- 304																										11
R%	8-101																										3
	77-165																										2
	227- 304																										4
R_{NORM}	8-101																										7
	77-165																										9
	227- 304																										5

5 LIKELIHOOD RATIO APPROACH - LR

The last approach evaluated in this work is based on the calculation of likelihood ratios and has been largely developed and discussed in forensic science, including for age estimation [2, 19, 23-26]. The strength of an observation (i.e., obtained results) is evaluated in regard of two different hypotheses and the case context [27-29]. Again, no information about the absolute age of an ink entry is given. However, the uncertainty of the results can be taken into account and expressed in the results in terms of probability/odds. Until now, such a model has never been tested for ink dating using real reference data, only subjective probabilities were used to show the potential of this interpretation approach [2, 19, 30]. Different models were proposed in the literature, including Baysenets, for different ageing problematics such as evaluating the time since discharge [23], the moment of deposition of fingermarks [24], or the age of living people [25, 26]. Each model has to take into account specificities such as the type of hypotheses (e.g. punctual times versus intervals), the type of data (e.g. continuous or discreet, uni or multivariate) and the type of ageing processes (e.g. regression fits and factors influencing transfer and influence storage). Thus, the development of a likelihood approach model will be specific to each problematic. For example, while the person at the origin of a fingermark will generally be identified before dating is performed [31]; the pen at the origin of a writing trace will rarely be known. Thus, the composition of the ink transferred on paper can hypothetically come from any ballpoint pen available on the market and the model must take this factor into account as “undetermined” factor. On the other hand, while a fingermark can be found outside and suffer from any environmental condition [31], a document is generally stored inside under relatively controlled conditions such as an office in a file, plastic folder or envelope. The type of substrate is generally known and the *grammage* can be measured. While the pressure cannot be directly determined, there is a correlation with the line width that can optically be determined.

The development of a likelihood ratio model will go through several steps before it can be applied in practical caseworks. First, two alternative hypotheses concerning the age of the questioned ink entry must be defined. Those propositions should be determined at the reception of the documents, before carrying any examinations. The defence hypothesis (t_d) will generally state that the questioned document is authentic and subsequently, the date printed or written on the document will be used to calculate the interval between the stated creation of the document and the analysis of the ink entries. The second hypothesis (t_a) will be the counterfeiting hypothesis and will be defined by the complainant or accusation. This second interval will usually be smaller than for t_d. It will be a suggested counterfeited or falsification time, and the context of the case might allow a more or less precise determination of t_a.

The second step will necessitate the result obtained in the casework (q), as well as results obtained from a representative reference population at time t_d and t_a. In this study, the different ageing parameters were considered in order to evaluate their suitability.

In order to calculate a likelihood ratio, the expert will then assess the probability to obtain the observed results (q) if the document **is** fraudulent (i.e. t_a is true) in comparison to the probability to obtain the same results (q) if the document **is** authentic (i.e. t_d is true):

$$LR = \frac{P(q|t_a)}{P(q|t_d)} \quad \text{Equation 9}$$

Generally, when the LR =1, the evidence is inconclusive. When the LR>1, the results support t_a, and when the LR<1, it supports t_d. A verbal scale can additionally be used to translate the obtained LR value into verbal conclusion. For example, the following scale could be used to verbally communicate the obtained results [32]:

- LR= 1: The results support neither propositions (the results does not bring new information)
- LR = 1-10: The results support proposition t_a rather than t_d (the support is qualified weak or limited)
- LR= 10-100: The results support the proposition t_a rather than t_d (this support is qualified moderate)
- LR= 100-1000: The results support the proposition t_a rather than t_d (this support is qualified strong)
- LR= 1000-10000: The results support the proposition t_a rather than t_d (this support is qualified very strong)
- LR> 10000: The results support extremely strongly the proposition ta rather than td.

Each step of the process will be developed and discussed using a case scenario, as the likelihood ratio approach is a case based approach. The true and false positive rate, as well as the true and false negative rate were additionally calculated from the reference population in order to discuss the feasibility of the model.

5.1 CASE SCENARIO AND HYPOTHESES

The following case scenario will be used to illustrate the model. While hypothetical, this is a typical case scenario as can be encountered in practice. It is also an ideal case, as the document is received within 1-2 months of the falsification hypothesis (t_a):

The tax office is reviewing the tax form sent by M. Jones. An invoice for suspicious expenses made on August 10th 2015 is missing. Thus, on July 8th 2016 the tax office ask M. Jones to send the invoice as complementary information. They receive the document printed with a signature on it, on July 28th 2016 and suspect a fraud. The questioned document is then sent for examination to a forensic laboratory and analysed on August 7th 2016.

The following alternative hypotheses can be formulated:

- **Hypothesis t_a**: the accusation will formulate the hypothesis that the document was created when it was asked to be sent to the tax office, namely between the moment the tax office sent the request

and the moment they received the paper. Then the document would be between 11 and 30 days old (i.e., time intervals calculated between the 8th or 28th of July and the 7th of August 2016).

- **Hypothesis t_d:** the defence will argue that the document is authentic and the age of the ink entry is 364 days old, i.e. the time gap between the document creation date (10th of August 2015) and the analysis time (7th of August).

In the framework of the case scenario, it will be assumed that the expert has the choice to calculate one of the three different ageing parameters considered in this study: PE quantity, R% and R_{NORM}.

5.2 PROBABILTY ESTIMATION AND LR CALCULATION

The critical step of the whole process lies in the estimation of the probability to obtain the observation (q) knowing the age (*t*) used in the LR calculation.

$$P(q|t) \qquad \text{Equation 10}$$

This necessitates adequate reference data depending on the case context, especially the age of the ink entries as defined by the alternative hypotheses. As in practice the reference population cannot be adapted specifically to each case circumstances, the reference data will essentially consist of different inks analysed at different ages (as many as possible to be representative of inks that may be encountered in practice). For example in this work, ink entries up to 1 year of age made with 25 different blue and black ballpoint pens were analysed.

Previously proposed densities estimation assumed that the data followed a normal distribution [23]. However, the distribution of the different ageing parameters considered in this study at the different ages (t) did not follow normal distributions. This was tested with two different normality tests (shapiro wilk [33] and kolgorov-smirnof tests [34]). Several distributions such as the lognormal and exponential distribution were tested but were found to be unsuitable even in combination with data pre-treatments (e.g. square root, inverse). Thus, the non-parametric kernel density estimation (KDE) was used [35] to evaluate the density of probability for each ageing parameter at the different ages:

$$P(q|t_1) = \hat{f}_{h_{t_1},t_1}(q) = \frac{1}{N \times h_{t_1}} \sum_{N=1}^{N} K\left(\frac{x - x_i}{h_{t_1}}\right) \qquad \text{Equation 11}$$

, where *K(x)* is the Kernel, generally a function such as a statistical law (e.g. Gaussian) and h_{t_1} is the bandwidth that was assigned to the dataset. This factor will influence the smoothing of the density curve. The selection of the bandwidth represents a critical point in the application of this method, because it will determine the precision of the density curve [35-38]. On the other hand, the selection of the kernel function generally has little influence on the resulting density. Thus, the following common normal distribution was considered in this study:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \text{Equation 12}$$

In order to estimate the best bandwidth value for the questioned dataset, several bandwidth estimation methods exist, all using different calculations and giving different bandwidths. As there is no ideal procedure to select the optimal bandwidth [36], and considering the small quantity of data per age in this study, three different bandwidth estimation methods were tested: the *rule-of-thumb* (ROT) proposed by Silvermann [35], the *Sheather and Jones* method (SJ) [38] and the *biased cross validation method* (BCV) proposed by Scott and Terrell [37]. The probabilities of the three alternative results for each ageing parameter values were compared for each parameter. In addition, the obtained densities of probability were also compared with the dataset profile (histogram).

The kernel density estimation is an empirical method, meaning it is possible to estimate the density only for the data and age at disposal. In our case scenario, given the two hypotheses $t_{a:}$ the ink entry is 11 to 30 days old and $t_{d,}$ i.e. the ink entry is 364 days old, the expert must select adequate reference data from the available set to estimate the probabilities, but there is no sample having 11, 30 or 364 days.

To estimate the probability of finding a results under the hypotheses t_a and t_d, the kernel density estimations should be applied on samples of the age of t_a and t_d. If these ages are not available in the dataset, the probabilities will have to be estimated on close ages. It is advised to choose an ink age slightly older for t_a and slightly younger for t_d to remain conservative and in favour of the defence (i.e., obtained LR values will be minimised). In this scenario, 39 days were selected for the accusation hypothesis and 304 days for the defence. Indeed, 39 days old samples will contain less or equal mean quantities of PE, RPA and R% values than 11 to 30 days old inks, while 304 days-old samples will contain more or equal mean quantities than 364 days (i.e. the ageing parameters decrease over time until the ageing level off). The following reformulated hypotheses are thus considered conservative and in favour of the accused:

- **Hypothesis of the accusation t_a**: the document was created after it was required by the tax office with less than 39 days ago.

- **Hypothesis of the defence t_d:** the document was created over 304 days ago as specified by the date on the document.

5.3 PE QUANTITIES

The estimation of the probabilities for the different PE quantities using kernel estimations and three different bandwidth selectors were calculated for 39 and 304 days (Figure 10).

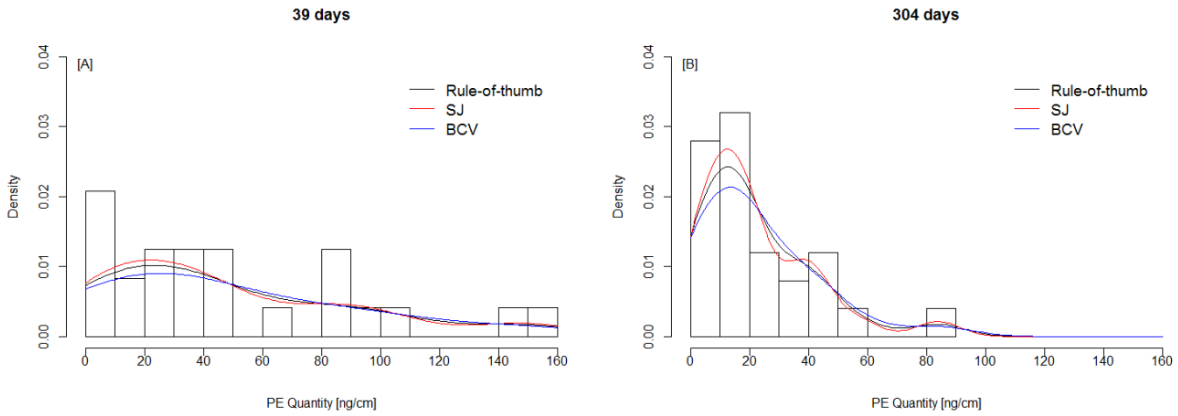

Figure 10 : Density of Probability for the PE quantities at 39 (A) and 304 days (B)

The probability densities calculated for 39 days old samples showed few differences between the three bandwidth selectors whatever the PE quantities measured. For example, for the different PE quantities up to 150 ng/cm, RSD values obtained for the densities were under 12%. Thus, the bandwidth calculation did not have a significant influence on the resulting probabilities. On the contrary, the probabilities measured for 304 days-old-samples showed more variations, especially for lower PE quantities (Figure 10). By example, RSD values between bandwidth selectors reached 8 and 12% for 20 and 90 ng/cm, respectively. For larger quantities, the probability densities are extremely low (close to 0 actually). Thus, we observe also much larger variations with a RSD of 173% for 150 ng/cm. In this study, the SJ bandwidth estimation method returned the smallest bandwidth to calculate the kernel densities. As a consequence, the density curves were less smoothed, more accurate, but also proved to be more sensitive to missing data than the other methods, i.e. range of values containing no inks such as 60 - 80 ng/cm for 304 days old samples or extreme values, i.e. values above the highest PE quantity measured for the reference population (>160 ng/cm for 39 days and > 90 ng/cm for 304 days). Thus, the use of higher bandwidths would be more adequate. The BCV method generally smoothed and flattened the density curves and minimized the probability of the most frequent values of this study. It always estimated the highest bandwidth. The ROT method generally tended to give an average density curve showing low-profiled maxima and minima.

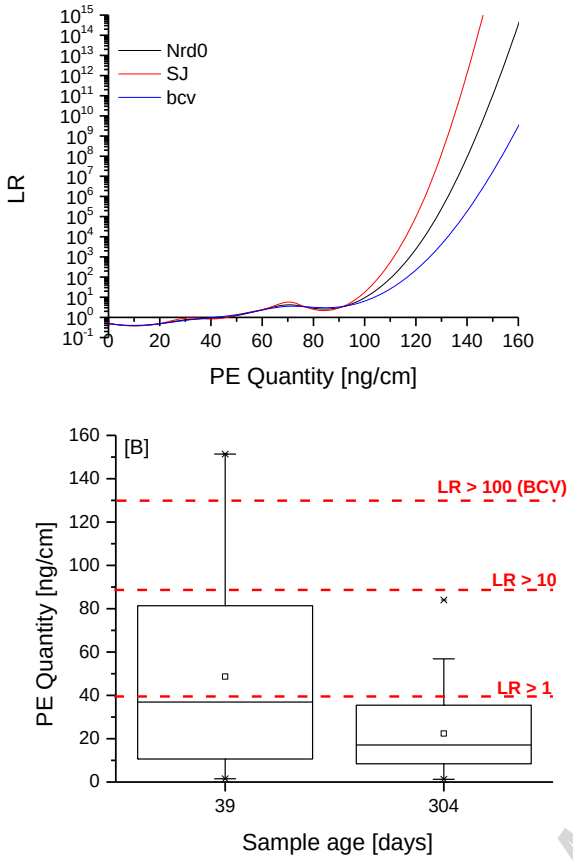

Figure 11 : A. LR curves for the PE quantity calculated using probabilities calculated with different bandwidth selectors (ROT, SJ, and BCV) for the two alternative hypotheses, ta = 39 days and td= 304 days. B. Boxplot of the PE quantity obtained for the 39 and 304 days old samples from the reference population.. Indications of LR intervals are drawn in red.

The calculation of LR values for PE quantities from 0 to 160 ng/cm (=q in eq. 7) allowed to compare to the different bandwidth selector used in the probability estimation (Figure 11). Main differences were observed for PE quantities above ca 100 ng/cm. The LR values calculated with the SJ bandwidths estimation always were the highest and the LR using the BCV were always the smallest. Thus, it must be kept in mind that the choice of a bandwidth selector will possibly influence significantly the results (especially when using a small reference population). For application in casework, it would be advised first to increase the reference population and also to use the bandwidth minimizing the LR values in order to remain conservative (*in dubio pro reo*). In this study, the LR resulting from BCV bandwidth probabilities always gave the lowest values.

Up to 40 ng of PE per cm ink, LR values under 1 were generally obtained. For example, it was 2.5 more probable to observe a result of 10 ng/cm of PE in the questioned ink, if the document was 304 days old rather than 39 days old (using the BCV bandwidth). Indeed, results below 40 ng/cm weakly supported that the document was authentic rather than falsified (i.e., 1/LR = 1-10, very low variations due to bandwidth estimator). While it would interesting to be able to support that an ink entry is old, this support remained very weak. Thus, it tended to confirm the general consensus in ink dating, that scientists can only conclude that an ink entry is fresh (young) if high quantities of PE are detected.

However, low PE quantities do not indicate that an ink is old because a significant number of inks presented very low PE quantities in their formulas (i.e. even 1 day after deposition). Moreover, the probabilities of finding low amount of PE in fresh sample is probably underestimated in the reference population chosen for this study (i.e. all inks contained PE, while previous studies tend to show that up

Table 5 : Overview of obtained LR values for samples from the reference population given the two alternatives propositions t1: 39 days and t2: 304 days (BCV bandwidth selection). As can be seen lower LR values were more frequently obtained (i.e. weak support). However, such values also yielded false positive results, while higher LR values never yielded false positive results (i.e., moderate and strong support). The ageing parameter R_{NORM} yielded the most promising results with no false positive value and a globally higher feasibility (i.e., number of inks for which a interpretable results was obtained).

	# of ink (% of samples)					
Ageing parameter	PE Quantity ng/cm (BCV)		R% (BCV)		R_{NORM} (BCV)	
	True positive	False positive	True positive	False positive	True positive	False positive
Support t_a LR >1	8 32%	2 8%	17 60%	6 24%	9 36%	0
Support t_a LR >10	2 8%	0	0	0	8 4%	0
Suppot t_a LR >100	2 8%	0	0	0	7 28%	0

to 20% of inks might not contain this compound [39].

For results between **40 and 100 ng of PE per cm of ink**, obtained LR values were above 1 but well below 10 (with very small variations due to bandwidth estimator). Thus, results supported weakly that the questioned ink entry was 39 days old rather than 304 days old. This can be explained by the fact that such quantities can also be found in older samples, yielding the risk of false positive (i.e. LR slightly above 1 indicating an ink younger than 39 days, while the ink is actually 304 days). In fact, two ink samples of 304 days contained 57 and 84 ng PE per cm (see Figure 14B), yielding LR values of 1.7 and 3.1, respectively (using BCV Bandwidth) (Table 5).

Finally, **above 100 ng PE per cm of ink**, the obtained LR values increased significantly with non-negligible differences between selected bandwidth. The LR rapidly reached values of ca.100 (around 110-130 ng/cm), ca. 1000 (around 115-135 ng/cm) and ca. 10'000 (around 120-140 ng/cm). Such values are however very rarely found in 39 days old samples (2 inks with quantities above 100 ng/cm) and never found in the 304-days-old samples (Table 5). This explains these extremely high values. While a larger reference population might yield different results, the observed tendencies should remain the same.

For the two selected alternative hypotheses, the LR approach actually showed comparable results to the threshold approach discussed above. Results can be interpreted with some confidence, only when the obtained PE quantity in the questioned ink entry reaches a certain quantity (i.e., ca. 100 ng/cm or above).

5.4 R%

The Kernel probabilities obtained for the three different R% also showed small differences as a function of the bandwidth selectors (Figure 12) Main differences appeared for the 39 days-old samples as the probabilities calculated for the R% values reached up to 22% (R%~45%). While the SJ and ROT methods yielded similar densities curves for 39 and 304 days old samples, the BCV methods presented smoother curves that minimized the probabilities for the most frequent R% values, especially for the 39 days old ink population (Figure 12). The probabilities obtained for the 304 days old samples were comparable between bandwidth up to R=50%, the RSD then increased up to 57% (at R=60%). Again the probabilities tended to vary more for high and somehow extreme values, i.e. values not encountered in the ink sample population of both ages. In contrary to PE quantities, SJ method did not always give the smaller bandwidth. In fact, the ROT gave the smallest bandwidth for the 304 days dataset, BCV in contrary always gave the larger bandwidth.

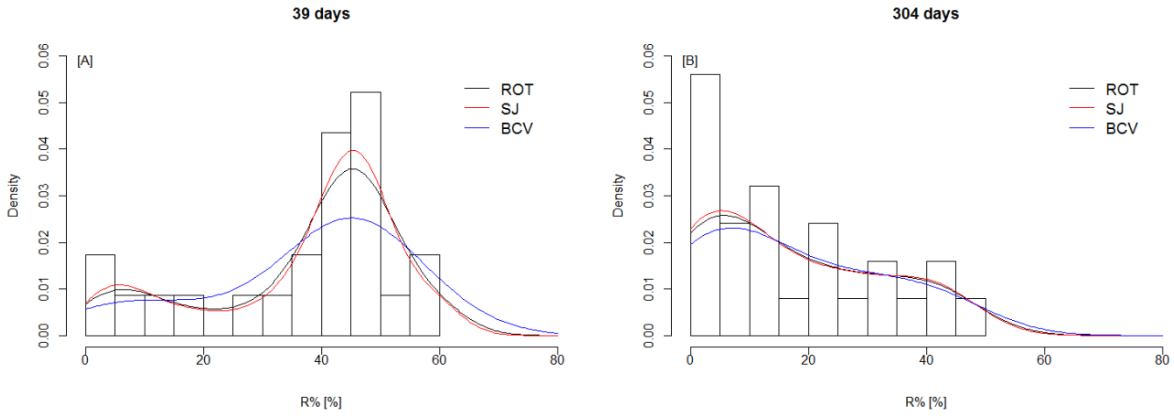

Figure 12 : Density of Probability for the PE quantities at 39 (A) and 304 days (B)

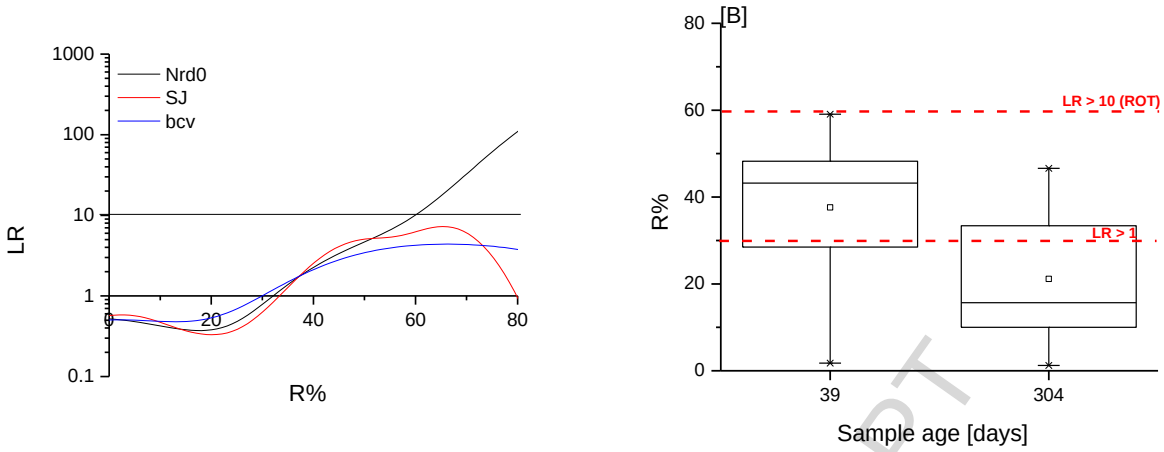

Figure 13 - A. LR curves obtained for the alternative propositions t_a= 39 days and t_d = 304 days using probabilities calculated with different bandwidth selectors (ROT, SJ, and BCV). B. Boxplot of the R% values obtained for the 39 and 304 days old samples from the reference population. Indications of LR intervals are drawn in red.

The LR values obtained for this ageing parameter yielded very low LR values even for the highest R%-values encountered ($R\%_{max}$:59%), and for all three kernel estimation.

For **R% value under ca. 30%**, the LR obtained showed small differences between the three different kernel calculations and the values were comprised between 0.3 and 1. This weakly supports the hypothesis that the document is authentic (i.e., t_d = 304 days). However, the observation of such LR values can be considered as inconclusive since they are very close to one (Figure 13).

R% values above 30% tended to support the counterfeit hypothesis (i.e, t_d = 39 days) (Figure 13 A). Significant differences were observed between the different Kernel calculations, especially after 50%. However, the obtained LR generally remained low, particularly for the SJ and BCV kernel estimators, as LR values never reached a LR of 10. Using the SJ bandwidth, the LR even tended to decrease after a value of 67%. This translates a problem in the model using this bandwidth selector, since the probability of finding values above 67% for 304 days old sample should not reach the probability of finding these values for 39 days old samples. For these models the risks of false positive is high. For example, using BCV, 6 inks yielded false positive results (Table 5). The LR calculated with the kernel probabilities using ROT bandwidth was the only one that yielded LR values above 10 when R% values exceeded 60%. However such high values were not observed in the 39-days-old samples of this study. The choice of the better bandwidth selector between BCV and ROT is not easy for this ageing parameter. However, results will not be greatly influenced and BCV gave the smallest LR..

Thus, given the two selected alternative hypotheses and the reference population of this study, the use of the R% instead of the PE quantity would generally lead to very limited conclusions i.e. weak support, as well as a higher risk of false positive results. This showed the difficulty to discriminate between relatively fresh and old samples using this ageing parameter (ca. 1 month vs 1 year). Further research using a larger reference population and older samples will be necessary to study the full potential and actual limitations of the R% parameter for ink dating.

5.5 R_{NORM}

The probabilities calculated for R_{NORM} results showed little differences between the different bandwidth estimators for the probabilities of 39 days old ink samples. The obtained RSD values were generally below 5%, except for value ≤ 5 ng/cm for which the RSD reached 13%. Thus, the different kernel density curves using the three bandwidth estimators did not influence much the probabilities for this dataset (Figure 14A).

However, dissimilarities were observed for the 304 days-old-samples (Figure 14B). High variations of probabilities calculated with the different bandwidth were observed for the highest values. For example, R_{NORM}= 60 ng/cm yielded a RSD of 173%. This value was actually never encountered in this population. Moreover, significant difference in the estimated probabilities were also found for the lower R_{NORM} values, up to 20% for values below 25 ng/cm.. The density of probabilities obtained using the SJ method tended to be closer to real values (red curve in Figure 14B) and this yielded the lowest bandwidth of the study. For the other bandwidth selectors maxima at 3 and 16 ng/cm were also observed but were less marked (blue and black curves in Figure 14B). The BCV returned again the highest bandwidth and presented the flattest curves. For 304 days old samples, the bandwidth selector method clearly had an influence on the probability density estimations and as a consequence on the LR values obtained with these densities (Figure 15A).

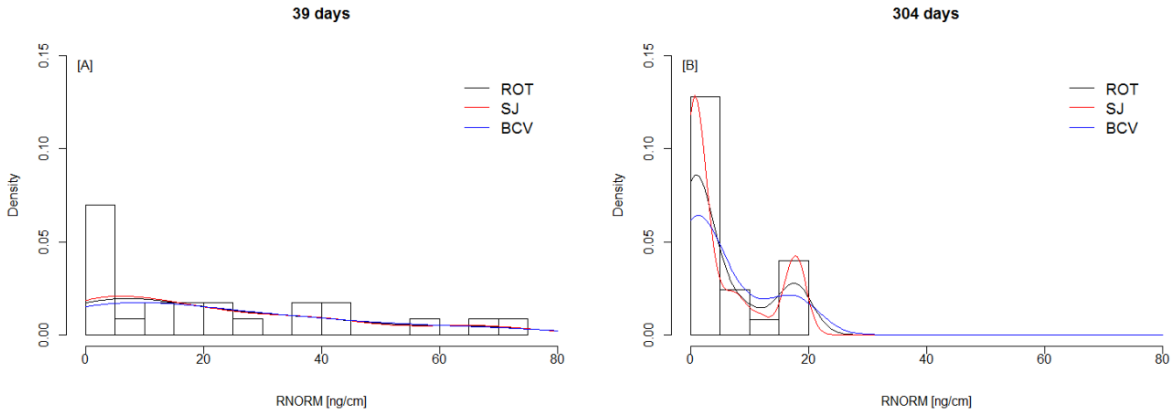

Figure 14 : Density of Probability for the PE quantities at 39 and 304 days

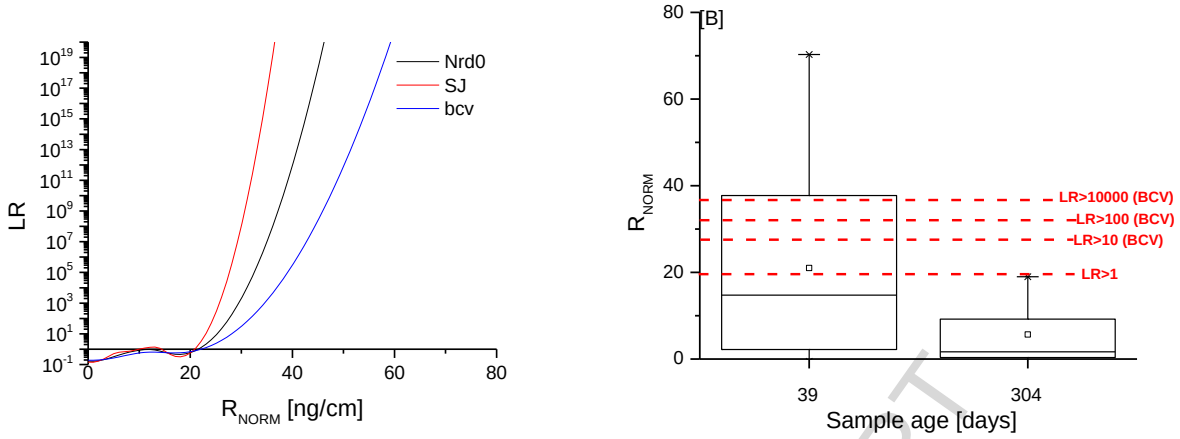

Figure 15 :- LR curves obtained for the hypotheses ta: the document is 39 days old and td : the document is 304 days old using probabilities calculated with different bandwidth selectors (ROT, SJ, and BCV). B. Boxplot of the RNORM values in the 39 days old and 304 days old populations.

The resulting LR showed the highest values of the study raising drastically after a R_{NORM} value of 20 ng/cm (Figure 15A).

For R_{NORM} values under 20 ng/cm, the calculated LR ranged between ca 0.1 and 1, the different LR calculation using different bandwidth gave similar results. Observing such R_{NORM} values would be up to 9 times more probable if the document was authentic than if the document was 39 days old (i.e. weak support for the defence hypothesis, t_d=304 days). While limited, such conclusions did not yield any false positive results (Table 5). This can be explained by the fact that all 304-days-old samples from this study possessed R_{NORM} values below 20 ng/cm.

For R_{NORM} values above 20ng/cm, the LR increased very quickly above 1 (Figure 18). Significant differences were observed between the different bandwidth selectors as a function of the R_{NORM} value. For example, a R_{NORM} value of 25 ng/cm yielded a LR of 3.1 for BCV, 11 for SJ and 340 for the SJ method. All three LR supported the hypotheses of the antedated document (i.e. 39 days old), but the strength of the evidence changed hugely from weak to strong support. This showed the importance of the selection of this parameter in the calculation of the probabilities. The SJ bandwidth selector always gave the highest LR-values and BCV the smaller LR-values. Even using the conservative BCV selector, several inks from this study presented a R_{NORM} value at 39 days allowing to reach a high LR. Thus one ink presented a LR slightly above 10, and 7 presented a LR above 10'000, none were in-between (see boxplots in Figure 18B and Table 5).

Again, the R_{NORM} ageing parameter proved to be the best ageing parameter, yielding significant LR-values for up to 36% of the inks and no false positive (BCV method). In general, this ageing parameter was the most promising for ink dating purposes.

5.6 PRACTICAL CONSIDERATIONS

The likelihood ratio approach allowed comparing the potential of each ageing parameter. While the R% gave the lowest LR values, the R_{NORM} yielded the highest. This tends to confirm previous observations made for the threshold and slope approaches. While the use of the PE quantity could be preferred in cases where little question ink is available, the R_{NORM} should clearly be preferred when possible.

While the chosen case scenario represented an ideal case, other alternative will also be encountered in practice. Thus, it is important to assess the results that can possibly encountered depending on the alterative propositions in order to decide if ink dating is still feasible or not in a specific case (and for how many ink formulation a positive result could be obtained). Such a process is called pre-evaluation [40, 41]. Indeed, the expert can wonder if the estimation of the age of an ink entry is still possible even when the maximum possible age of the counterfeit document (t_a) is different than the one presented in this study or if the document age (t_d) is younger. The feasibility was tested by considering different alternative case scenario. This was done using the most promising ageing parameter (R_{NORM}). The BCV method was used for the following LR calculation. It generally gave the smallest LR from the three bandwidths.

Alternative 1: *The document was sent and analysed at a different time (t_a = 8 days, 101, 165 days or 274 days)*

The LR curves for the hypotheses t_a=8, 39, 101 and 165 were very similar, while the curve for the hypothesis t_a=274 days differed significantly for R_{NORM} values above 30 ng/cm (Figure 16A). In fact, the LR values increased slower as a function of the quantity. These observations showed that LR values were highly correlated to the document age hypothesis (t_d) and less to the maximal presumed counterfeited date (t_a), except when this date was close to t_d.

While LR values were quite similar, the chance of detecting an antedated ink entry tended to vary as a function of the counterfeiting hypothesis (t_a). In fact, 12 inks of the 8 days-old-ink population would present R_{NORM} values supporting the right hypothesis, 9 of them with LR above 100 (Table 6). In comparison, only 9 samples supported the right hypothesis of for t_a= 39 days, 7 of them yielded LR above 100 (Table 6). As expected, the number of ink samples presenting a sufficient R_{NORM} value to reach a LR of 100 decreased as a function of the age of t_a and for t_a= 274 days, the maximum LR obtained was only 3 (Table 6 and Figure 16B). While a global decrease of the true positive rate as a function of the t_a hypothesis was observed for all ageing parameters, including PE and R% (results not shown). The R_{NORM} parameter remained the most promising. While a LR above 100 could still be detected after 165 days using this parameter, the chance of success is much higher if the document is sent quickly to the laboratory after it is contested, confirming observations from the threshold approach.

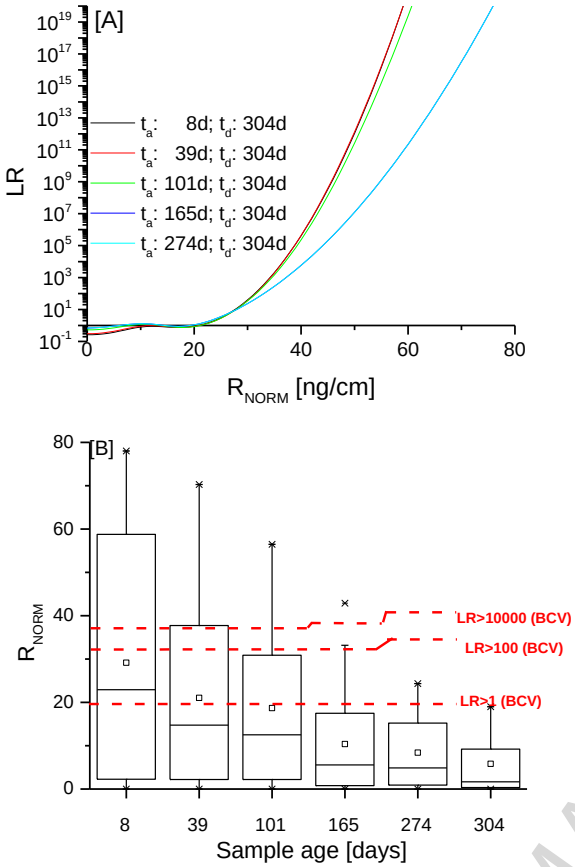

Figure 16 : A) Obtained LR for different hypotheses pairs, t_a : 8, 39, 101, 165, 274 days and t_d : 304 days. B) Boxplot of R_{NORM}

Table 6 : Summary of the results for different hypotheses t1: 8, 39, 101, 165 days and t2: 304 days, the ageing the number of sample inside the reference population having a particular LR.

Hypotheses Pair t_a / t_d [days]	8 / 304		39/304		101 / 304		165 / 304		274 / 304	
Type of answer:	True positive	False positive	True positive	False positive	True positive	False positive	True positive	False positive	True positive	False positive
# of ink sample that: Support ta >1	12 48%	0	9 36%	0	8 32%	0	8 32%	2 (LR<2)	10	4 (LR<2)
Support ta >10	10 40%	0	8 4%	0	6	0	3 12%	0	0	0
Support ta >100	9 36%	0	7 28%	0	5 20%	0	2 8%	0	0	0

values in the considered populations. The red lines represent the range of LR values.

Alternative 2 : *The document age is different, and the time range between the counterfeiting and the document age diminished*

Different document ages (t_d) of were also considered in addition to 304 days: 101 and 165 days (while the maximal presumed counterfeit age remained t_a=39 days old). In contrary to the t_a hypothesis, changing the t_d hypothesis drastically influenced the obtained LR curve (Figure 17A). In fact, LR values above 10 (moderate support) would be reached at a R_{NORM} value of 28, 56 and 83 ng/cm for t_d= 304, 165 and 101 days, respectively. Values LR above 100 (strong support) would be reached at a R_{NORM} values in ng/cm of 32 (t_d=304 days), 63 (td = 165 days) and 99 (t_d = 101 days) (figure 17). This indicated that LR values were more correlated to the age of the document than the supposed counterfeited age.

Again, LR values tended to decrease when the time range between the two alternatives decreased. This was also observed through a decrease in the true positive rate as a function of the gap between t_a and t_d. For t_d= 165 days, it was still possible to support the hypothesis of counterfeiting as two 39-days-old ink samples yielded LR above 100, one sample reached a LR of ca. 10 and eight samples yielded samples between 1 and 10 (Table 7 and Figure 17 B and C). However, for t_d=101 days, no samples yielded LR values above 2 (Table 7 and Figure 17B).

Thus, a large time interval between the hypotheses (at least during the first year after the document creation) will significantly increase the feasibility. However, even considering t_a=8 days and t_d= 304 days (i.e., largest time interval studied in this work), the maximal measured true positive rate represented 10 inks yielding LR values above 10 (Table 7). Two additional inks yielded LR values between 1 and 10, and for such values the risk of false positive could not be excluded in this study. Fortunately, no false positives were observed for LR above 10 in this study.

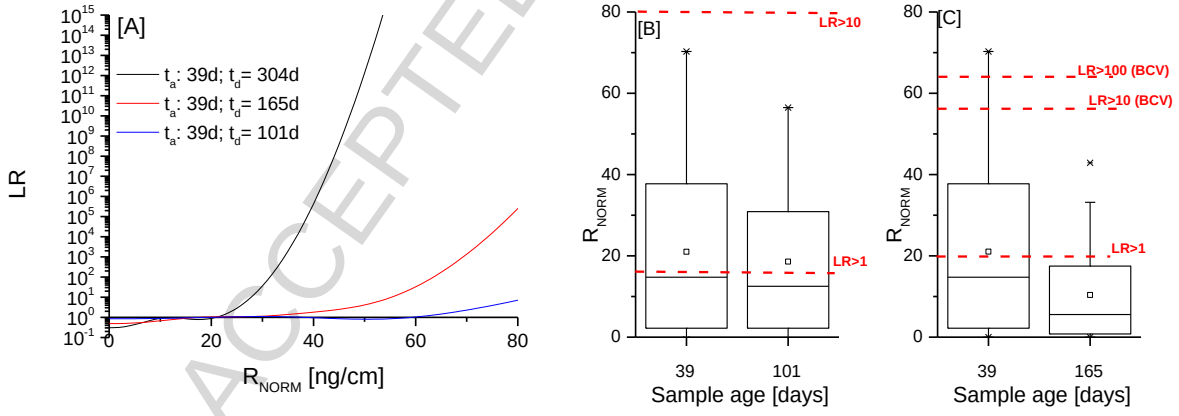

Figure 17 : A. LR curves for different hypotheses pair, ta : 39 days, td= 101, 165 and 304 days.**B and C** Boxplots of the R_{NORM} values in the 39 and 101 days old population (B) and the 39 and 165 days old population and the indication of the value of the LR according the hypotheses linked to the populations

Table 7: Summary of the results for different hypotheses t_a:39 days days and t_d: 101, 165 and 304 days. For each hypotheses pair, the number of reference ink having a specific LR was reported. It gave the number of true and false positive for each hypotheses pair.

Hypotheses Pair t_a / t_d [days]	39/304		39/165		39/101	
Ink sample age [days]	True positive	False positive	True positive	False positive	True positive	False positive
Support ta >1	9	0	11	6 (LR≤2)	7 (LR ≤2)	3 (LR<2)
Support ta >10	8	0	3	0	0	0
Support ta >100	7	0	2	0	0	0

While very informative, the proposed model remains preliminary and it must be kept in mind that the LR values obtained in this work are only estimations and not accurate values. In fact, the reference data used in this study were constituted of only 25 ink formulations chosen for their ageing behaviour type and not according to their occurrence in the ballpoint pen ink market. The ideal reference population should indeed contain more inks in order to be statistically representative of formulations found on the market.

Despite this, the preliminary model developed in this work clearly showed the potential of an interpretation based on the calculation of likelihood ratios. Results tended to show similarities to the threshold approach, with added numerical information in two forms:

- A differentiation is made between weak and strong support depending on the obtained ageing parameter value (see verbal scale above).
- The formulation of two alternatives hypotheses will also influence obtained LR values (influence of the case scenarios.

6 CONCLUSION

Three different models of interpretation were evaluated in this work. The reference population consisted of ink entries made from 25 different ballpoint pens over 304 days. The potential of three different ageing parameters were more particularly tested: the PE quantity, R% and R_{NORM}.

The **threshold approach** proved to be very easy to implement. It is however prone to false positive results and one should be careful in the selection of appropriate thresholds for ink dating purposes. One solution would be to define a “conservative” threshold value with a security interval between the defined threshold and the conclusion. Thus, a threshold could be defined for 138 days old ink entries (i.e. R_{NORM} = 45 ng/cm) and be used to conclude that an ink is less than 274 days:

- If a R_{NORM} value above 45 ng/cm is detected in the questioned ink, then the results support that the ink is less than 274 days old.

This solution allows reducing significantly the risk of false positive results. Globally, the feasibility (i.e. true positive rate) remained quite low with a maximal value of 10 inks (40%) obtained for R_{NORM} in order to conserve 0 false positive.

The **trend tests approach** as previously proposed in the literature, clearly yielded unreliable results and should not be applied in practice. Such approaches are based on statistical tests and tabulated data in order to evaluate if a set of data present a (descending) trend. However, the high variability of the results and the small quantity of data points makes such tests unreliable for ink dating purposes, at least on the three testes ageing parameters. An alternative was proposed in order to define a threshold based on the reference data. It was thus possible to define promising thresholds values for the PE Quantity and the R_{NORM}, with a maximal number of 9 inks (36%) that would give positive results (R_{NORM}). Such tests will remain inapplicable for the R%, as this parameter did not present a mean descending trend on the tested ink population.

The **likelihood ratio approach** proved to be a very promising interpretation model. While slightly more complicated to apply than thresholds, it also yielded added information about the strength of the evidence, the actual feasibility and the potential of each ageing parameter. It was thus demonstrated that R% yielded globally lower LR values, while the R_{NORM} parameter again yielded the highest LR values. R_{NORM} yielded LR values above 100 for 9 inks (strong support) and showed that below a LR of 10, no conclusion should be given because the risk of false positives was non negligible.

Some rules could be highlighted concerning the ink dating method. (1) The choice of the ageing parameter influences the chance of detecting an antedated ink entry. The R_{NORM} parameter allowed detecting an antedated ink entry for more inks, independently of the interpretation model. (2) The supposed “counterfeiting” age (hypothesis of the accusation) is particularly important. A quick reaction between the supposed counterfeit date and the analysis date will yield higher chances to successfully detect the fraud. (3) Ink dating is a time sensitive matter, thus a large time interval between the two alternatives hypotheses also increased the feasibility of the ink dating processes. Ideally, a counterfeiting hypothesis of several days (from 4 to 39 days) versus the date of the document (304 days or more) represent ideal case scenarios as demonstrated in this study.

Further studies should now focus on the acquisition of more representative data in order to tune these interpretation models for their proper use in practice. A larger ink reference population should be selected in order to be statistically representative of the ballpoint pen inks found on the market. Moreover, these inks should be stored under different conditions as usually encountered in practice over longer time intervals (analysis every two weeks up to 2-3 years). Generally, slow ageing inks and influence factors slowing down the ageing parameter decreases should be more particularly studied, as these specific conditions might actually yield false positive results in the considered interpretation models.

7 RÉFÉRENCES

1. Koenig, A. and C. Weyermann, *Ink dating part 1 : Statistical description of selected ageing parameters in a ballpoint ink reference population.* Science and Justice, 2017. **Submitted**.
2. Weyermann, C., et al., *Minimum requirements for application of ink dating methods based on solvent analysis in casework.* Forensic Science International, 2011. **210**(1-3): p. 52-62.
3. Aginsky, V., *Dating and characterizing writing, stamp, pad, and jet printer inks by gas chromatography/mass spectrometry.* International Journal of Forensic Document examiners, 1996. **2**(2): p. 103-116.
4. Aginsky, V.N., *Current Methods for Dating Ink On Document*, in *60th annual Conference of the American Society of Questionned Document Examiner*. 2002: San Diego.
5. Bugler, J.H., H. Buchner, and A. Dallmayer, *Age determination of ballpoint pen ink by thermal desorption and gas chromatography-mass spectrometry.* Journal of Forensic Sciences, 2008. **53**(4): p. 982-988.
6. Koenig, A., et al., *Ink dating using thermal desorption and gas chromatography/mass spectrometry: Comparison of results obtained in two laboratories.* Journal of Forensic Sciences, 2015. **60**(s1): p. S152-S161.
7. Koenig, A., S. Magnolon, and C. Weyermann, *A comparative study of ballpoint ink ageing parameters using GC/MS.* Forensic Science International, 2015. **252**: p. 93-106.
8. Weyermann, C., et al., *A GC/MS study of the drying of ballpoint pen ink on paper.* Forensic Science International, 2007. **168**(2-3): p. 119-127.
9. Gaudreau, M. and V. Aginsky, *Essentials of the solvent loss ratio method*, in *68th Annual Conference of the American Society of Questionned Document Examiners (ASQDE)*. 2010: Victoria, USA.
10. Gaudreau, M. and L. Brazeau, *Ink dating, a solvent loss ratio method*, in *6th Annual Conference of the American Society of Questioned Document Examiners*. 2002: San Diego, California.
11. Bügler, J.H. *Method validation for age determination of ballpoint inks*. in *5th Annual Conference of the European Document Experts Working Group (EDEWG)*. 2008. Bunratty, Ireland.
12. Aginski, V.N., *Ink aging testing-Do preceding indentation examinations affect ink aging parameters.* Journal of the American Society of Questionned Document Examiners, 2014. **17**(2): p. 49-63.
13. Brunelle, R.L. and A.A. Cantu, *Training Requirements and Ethical Responsibilities of Forensic Scientists Performing Ink Dating Examinations.* Journal of Forensic Sciences, 1987. **32**(6): p. 1502-1506.
14. Bügler, J.H., M. Graydon, and B. Ostrum. *The parctical use of the Munich ink reference collection in daily casework*. in *6th European Document Examiners Working Group (EDEWG) Conference*. 2010. Dubrovnik, Croatia.
15. Aginsky, V.N. *Current Methods for Dating Ink on Documents*. in *65th Annual Conference of the American Society of Questioned Document Examiners*. 2007. Boulder, Colorado.
16. Aginsky, V.N., *Measuring ink extractability as a function of age - why the relative aging approach is unreliable and why it is more correct to measure ink volatile components than dyes.* International Journal of Forensic Document examiners, 1998. **4**(3): p. 214-230.

17. Esseiva, P., et al., *Illicit drug profiling, reflection on statistical comparisons.* Forensic Science International. **207**(1): p. 27-34.

18. Weyermann, C., et al., *Statistical discrimination of black gel pen inks analysed by laser desorption/ionization mass spectrometry.* Forensic Science International, 2012. **217**(1-3): p. 127-133.

19. Weyermann, C., B. Schiffer, and P. Margot, *A logical framework to ballpoint ink dating interpretation.* Science & Justice, 2008. **48**(3): p. 118-125.

20. Aitken, C.G.G. and F. Taroni, *Uncertainty in Forensic Science*, in *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2005, John Wiley & Sons, Ltd. p. 1-34.

21. Evett, I.W., *Expert evidence and forensic misconceptions of the nature of exact science.* Science and Justice, 1996. **36**(2): p. 118-122.

22. Neumann, J., et al., *The mean square successive difference.* Annals of mathematical statistics, 1941. **12**: p. 153-162.

23. Gallidabino, M., et al., *Estimating the time since discharge of spent cartridges: A logical approach for interpreting the evidence.* Science and Justice, 2013. **53**(1): p. 41-48.

24. Girod, A., et al., *Aging of target lipid parameters in fingermark residue using GC/MS: Effects of influence factors and perspectives for dating purposes.* Science & Justice, 2016. **56**(3): p. 165-180.

25. Sironi, E., et al., *Probabilistic graphical models to deal with age estimation of living persons.* International Journal of Legal Medicine, 2016. **130**(2): p. 475-488.

26. Sironi, E., V. Pinchi, and F. Taroni, *Probabilistic age classification with Bayesian networks: A study on the ossification status of the medial clavicular epiphysis.* Forensic Science International, 2016. **258**: p. 81-87.

27. Aitken, C.G.G. and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2005: John Wiley & Sons, Ltd.

28. Evett, I.W., *A Bayesian approach to the problem of interpreting glass evidence in forensic science casework.* Journal of the Forensic Science Society, 1986. **26**(1): p. 3-18.

29. Robertson, B. and G.A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom.* 1995, Chichester, UK; : John Wiley & Sons.

30. Koenig, A., et al., *Ink dating: How to interpret the results?*, in *in: 6th European Academy of Forensic Science (EAFS) conference*. 2012: The Hague, Netherland.

31. Girod, A., et al., *Fingermark age determinations: Legal considerations, review of the literature and practical propositions.* Forensic Science International. **262**: p. 212-226.

32. Marquis, R., et al., *Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings.* Science and Justice, 2016.

33. Royston, P., *Approximating the Shapiro-Wilk W-test for non-normality.* Statistics and Computing, 1992. **2**(3): p. 117-119.

34. Conover, W.J., *Practical Nonparametric Statistics*. 1971: John Wiley & Sons.

35. Silverman, B.W., *Density estimation for statistics and data analysis / B.W. Silverman*. Monographs on statistics and applied probability ; [26]. 1986, London ; New York: Chapman and Hall.

36. Heidenreich, N.-B., A. Schindler, and S. Sperlich, *Bandwidth selection for kernel density estimation: a review of fully automatic selectors.* AStA Advances in Statistical Analysis, 2013. **97**(4): p. 403-433.

37. Scott, D.W. and G.R. Terrell, *Biased and unbiased cross-validation in density estimation.* Journal of the American Statistical Association, 1987. **82**(400): p. 1131-1146.

38. Sheather, S.J. and M.C. Jones, *A reliable data-based bandwidth selection method for kernel density estimation.* Journal of the Royal Statistical Society, Series B, 1991. **53**: p. 683–690.

39. LaPorte, G.M., et al., *The identification of 2-phenoxyethanol in ballpoint inks using gas chromatography/mass spectrometry-relevance to ink dating.* Journal of Forensic Sciences, 2004. **49**(1): p. 155-159.

40. Cook, R., et al., *A model for case assessment and interpretation.* Science and Justice. **38**(3): p. 151-156.

41. Jackson, G., et al., *The nature of forensic science opinion-a possible framework to guide thinking and practicce in investigation and in court proceedings.* Science and Justice. **46**(1): p. 33-44.

8 ACKNOWLEDGEMENTS:

The authors wish to thank the Dr. Matteo Gallibadino and M. Sironi for their help and discussions about the likelihood ratio model. They also acknowledge Dr. J. Bügler and Ms Linden from the Landeskriminalamt of Munich for sharing their collection of ballpoint pen inks as well as for their precious help and availability.

The authors finally wish to thank the Swiss National foundation for its support in the frame of this research (n°PP00P1_123358 and PP00P1_150742).

Highlights (second part):

- Comparison of three interpretation models on the same ink reference population
- Evaluation of the models using three promising ink dating ageing parameters
- Proposition of a new approach to exploit “trends” in the interpretation
- Development of a preliminary likelihood ratio model