Precision of caracterization of paper for recycling

Gilles Pagès, Victor Reutenauer

Precision of caracterization of paper for recycling

Gilles Pagès*- Victor Reutenauer†

Novembre 2019

Contents

[1]LPSM - Sorbonne Université
[2]Fotonower

Key Word

Paper for recycling, characterization, precision, neural network, Monte-carlo, Intelligence Artificial, Correlation, Exact caracterization procedure.

Abstract

Auf deutsch

En Francais

Ce document récapitule les enjeux industriels de la caractérisation gravimétrique des mélanges papier graphique-carton ainsi que des méthodes statistiques et modèles probabilistes pour estimer la précision de ces méthodes. Dans la partie 1, apres un rappel de l'état de l'art, sont présentés différentes métriques de précisions de ces estimations de composition. Dans la partie 2 un modèle probabiliste original dans son application est proposé pour calculer ces métriques de manière efficace et une méthode et formule simplifié explique comment mettre en oeuvre cette technique. On propose une maniere fiable d'estimer les paramètres du modèle dans des cas réels avec des tests réalisés dans une usine. Dans la partie 3 sont présentés des méthodes non invasive basée sur de la reconnaissance d'image pour estimer les compositions des mélanges papier-carton. L'estimation de la précision de cette méthode sera réalisé ultérieurement.

L'intérêt du modèle est de permettre de tester les cas maximaux de correlations que l'on peut atteindre entre différentes méthodes de caractérisations. On met particulièrement en avant dans la formule 1 une mesure d'erreur entre la réalité des chargements et la mesure effectuée.

In English

This document discuss about industrial issues of gravimetric caracterisation of load of paper and board and presents statistical methods and probabilistic model to estimate the precision of these methods. In section 1, after a reminder of the state of the art, differents metrics of precision of estimation of the composition are presented. In section 2, an original model is proposed to compute the metrics efficiently and a simplified methods and formula explain how to use this technic. We propose an efficient way to estimate the parameters of the models with real case with data coming from a sorting plant. In section 3 are presented non invasive methods based on image recognition to estimate composition of load of paper for recycling. Estimation of the precision of this methods will be published lately.

The main idea of the proposed model is to test the maximum value of the correlation that we can achieved between two differents caracterization methods.

The main results is the formula 1 that links the error measured and the error of the methodology with the reality.

Thanks

I would like to thanks the many people that mades this study possible. Arnaud Dauxerre who brought us to the paper industry as long as all its colleague Thomas Krauthauf, Manfred Geistbeck, Marita Pertu, Jean Kubiak and Marc Thebaud still missing some. Employees at Fotonower Jingxuan Feng, Chengcheng Xu, Marine Colin and Stéphane Poirier worked on some aspect of the theoretical or numerical part. Former colleague that made part of the manual caracterization work, Nicolas Gueritat, Pia Chancerel and Romain Pagès. Jan Lemoux and Isabelle Margain enabled Fotonower to present its work to the paper industry.

1 Gravimetric characterisation : state of the art

1.1 Industrials issues, definition and main usage

Quality of load of paper for recycling is a key issue in the paper industry. Therefore different characterisation methodology have been developed and their statistic studied.

Among them, gravimetric characterisation of a sample is widely used in the rubbish industry and seen as the best reference available.

We present here a short description based on the norm [2] and [3].

The loose is sampled either by a mechanical way of mixing it and putting a load of it in some recipient or by simply collecting manually inside the heap to avoid selecting (although unconsciously) visually the material to be caracterized as defined in INGEDE 14 described in [8].

Then, as stated in the norm, the different elements are separated and weighted. Weight proportion are then reported.

Ademe gathered gravimetric caracterization campaign in [6] and proposed methodology for caracterization and testing its accuracy in [5].

These study are interested in the caracterization of final garbage.

We observe that the caracterization is not perfect and mainly focus on flow of truck.

The norm [1] interests us specifically because it is for only one truck. The sub-sampling procedure is introduced in it, we discuss about it in section .

1.2 Source of errors and uncertainty

The error and uncertainty in the composition of the complete load of materials comes from several reasons :

- The amount of uncharacterized material called "fines" in french. Usually of the order of one percent of the characterized sample.
- Manual error in caracterization or incomplete caracterization of some type of material inconsistency with the norm [7].
- Bias in the sample versus the complete load.
- Inhomogeneous sample due to incomplete mixing of the load.
- Small size of the samples/Size of each element of material.

The first error can be neglected based on the hypothesis that the "fine" have no bias with the rest of the sample. We can refer to [10] to have some idea on the distribution of size of the different type of material (paper/board to start with).

The second one can also be taken into account the same way we manage error in automatic detection in section 3.3. We don't take it into account in the first version of "Effective Precision Methodology" presented in section 2.2.

The third one, the bias can also be studied once stated. This is done for example for surfacic scanning of material on conveyor which is for example ineffective to detect glass mixed with paper since almost all of it goes to the bottom of the conveyor.

The last two are the problem tackled in this study. The inhomogeneous sample is not an issue, our methodology apply only for same inhomogeneity of material. We do not introduce homogeneity measure that could explain how to compare two mixing procedure. Here are two example :

Some usual sampling procedure

- For one load of a truck : X times using a mecanical loader to mix the material before selecting a sample.
- Picking manually a sample by gathering Y different sub-sample around the load by putting the hand inside the heap (to avoid a bias of manual selection).

Our calibration and precision methodology of the caracterization methods must be applied for one specific mixing and sampling methodology with fixed procedure.

The methodology proposed in 2 gives correlation of a complete caracterization of the load with the chosen caracterization procedure. Without using or (simple) formula it would be untractable either in production of for a test since it would require to characterized the whole load. The best that can be approached is a visual dynamic characterization procedure on top of a conveyor. This is not the subject of this study.

We propose in the next paragraph 1.3.1, 1.3.2 and 1.3.3 three different error metrics between two different caracterization methodology.

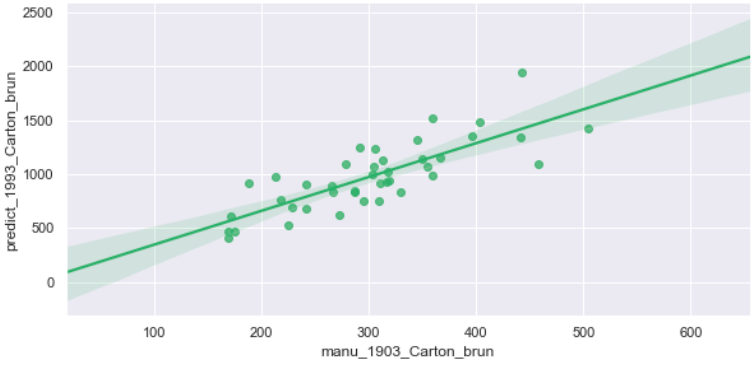

Figure 1: *Correlation graph*

1.3 Several error metrics

These error metrics are here to estimate the difference between two caracterization methods of a set of load among for example : INGEDE 7 (visual), INGEDE 14 (weight), QUALIPAPIA (introduced in section 3) or any caracterization methods of bales (that are outside the scope of this study) or even the exact composition of a load. In some case, we can also consider two independent exectution of the same caracterisation methods. For INGEDE 7 visual, we can for example, without considering security issues, apply it to one side of the heap and the other side. It is of course easier to consider multiple INGEDE 14. The main results of this study is the relationship between the error metrics for different comparison presented in section 2.2.

An exact caracterization methods could be used by an industrialization of a tools as the one developped at Darmstad Technical University presented here [11].

We uses the variable defined in table 1 and introduce most of them when they are used in equation.

1.3.1 Confusion matrix

The most intuitive and business-oriented way of considering the precision of a caracterization method is the confusion matrix, either between the same methodology applied two times independently or between two different methodology.

The confusion matrix which on example can be found in figure 3.

In column are classes on a given measurement methods m and in rows the classes either exact (the complete load), or of a reference methods n.

This matrix as the property of having sum on column equal to one, or 100%, and in a cell of row m_0 and column n_0 : c_{m_0,n_0}, is the conditionnal expectancy of having the method m a result m_0 and a result n_0.

$$c_{m_0,n_0} = \mathbf{E}\left[p_U^n \in n_0 | p_U^m \in m_0\right]$$

Or in natural language, if the measurment method m gives a class m_0, c_{m_0,n_0} is the probability that the methods n gives a class n_0.

1.3.2 Correlation

We use the notation listed in table 1.

The correlation is simply :

$$\rho^{(i,j)} \frac{\mathbb{E}\left[p_A^{(i)} p_A^{(j)}\right] - \mathbb{E}\left[p_A^{(i)}\right]\mathbb{E}\left[p_A^{(j)}\right]}{\sqrt{\left(\mathbb{E}\left[p_A^{(i)} p_A^{(i)}\right] - \mathbb{E}\left[p_A^{(i)}\right]\mathbb{E}\left[p_A^{(i)}\right]\right)\left(\mathbb{E}\left[p_A^{(j)} p_A^{(j)}\right] - \mathbb{E}\left[p_A^{(j)}\right]\mathbb{E}\left[p_A^{(j)}\right]\right)}}$$

We usually display the data in a chart in figure 1.

1.3.3 Mean square error, absolute or relative

Mean square error or MSE can be better appreciated by taking its square root in order to have something homogeneous to the quantity studied and become the Square Root Mean Square Error or RMSE.

For the absolute error, we use :

$$E_A^{(i,j)} = p_A^{(i)} - p_A^{(j)}$$

$$\sqrt{\mathbb{E}\left[\left(E_A^{(i,j)}\right)^2\right] - \mathbb{E}\left[\left(E_A^{(i,j)}\right)\right]^2}$$

And the relative error will be the same formula by taking :

$$R_A^{(i,j)} = \frac{p_A^{(i)} - p_A^{(j)}}{p_A^{(i)}}$$

1.4 Ideal load of mixed paper and board

The model presented in section 2.1 is based on a real physical case (totally intractable in an industrial process of transporting paper and board).

We consider an ideal case of load of mixed paper and board with the following properties :

- all in bullet
- bullet will either be full paper or full board
- bullets will have same size and weight

In this case the model proposed in section 2.1.1 is the exact representation of the reality.

2 Tractable methodology to estimate the precision of a gravimetric caracterisation method

This model enable to manipulate various data of caracterisation coming from sample of different way.

By observing the simplest error metrics presented in section 1.3.3, we calibrate our model by MSE-matching, or maximizing the log-likelihood and then estimate the most business-oriented error metrics that are the confusion matrix.

A first order approximation, exact when the size of sample are the same for each caracterization is presented in the most simple way in section 2.2.

2.1 Theoretical settings

Pr. Shauble in [10] studies statistically the size of different type of material in mixed paper and board. For a sake of simplicity and efficiency we uses a simpler model where all element have the same size. Our methodology can be expanded to multiple size or even continuous distribution of size nearer to the reality. Despite that fact we have the strong belief that this most simpler model is the best first order estimation of the precision of a caracterization procedure.

2.1.1 Implicit independent size Model

Having in mind the ideal load presented in section 1.4, we will calibrate the value of any risk metrics computed on multiple load as the one presented here : 1.3 into an implicit size of independent element of the material considered. This is totally equivalent to considering the intrisic standard deviation of the model or the correlation with the real proportion of unwanted material in the load in 1.3.3 but can in some way be more easy to understand since it can be linked to some physical ideal case described in section 1.4. These ideal case are not tractable in a production process of the industry but can be physically achieved.

As in the survey industry (without using quota methodology), we consider the choice of each asked individual as independent.

Comparison with smile modelling Extending this methodology to a model where the implicit size depends on the proportion of unwanted material is under this real probability the same idea that modelling smile under the risk neutral probability in financial industry.

In fact instead of maximizing the log-likelihood we could also find the same MSE as shown in section 2.1.4.

2.1.2 Binomial or gaussian and poisson approximation of the model

On considère NMC réalisations qui peuvent être : des fractions de balles, des chargements de camions ou des matieres passant sur un convoyeur pour un interval de temps court. Pour chacun de ces ensembles on considère qu'elles ont un taux P de contaminant ou de distribution de deux différentes types de matières et on effectue $NNMC$ tirage de taille S pour chacune d'entre-elles. P et S peuvent être aléatoires avec par exemple une distribution uniforme.

Using Binomial law L'observation P^O du taux de contaminant ou répartition des matières suit une loi binomiale de parametres p et $\frac{S}{s}$.

Using CLT for gaussian approximation We will simply apply the Central Limit theorem or uses the known binomial law to find this implicit independent size and then uses it to compute others risk metrics.

Lorsque S tend vers l'infini, $\sqrt{\frac{S}{s}}\left(P^O - P\right)$ converge vers une loi gaussienne de parametre $0, P\left(1-P\right)$

On simule donc p^O avec une réalisation gaussienne $P, P(1-P)\frac{s}{S}$

Poisson approximation For small S a Poisson approximation might be more accurate.

2.1.3 Formula of the metrics in the implicit independent size model

We refer to the metrix defined in section 1.3

Proof are computed in annex 6.3.

Correlation $Cor(p_{gravi}, p_{quali}) = \frac{1}{\sqrt{1+\frac{1}{N_{gravi}}(\frac{6}{M}-4)}\sqrt{1+\frac{1}{N_{quali}}(\frac{6}{M}-4)}}$

Confusion matrix

Mean square error, absolute or relative

Monte-Carlo simulation of complex risk metrics For any risk metrics, it can be easier to implement a Monte-Carlo simulation methods, instead of computing exact or approximated formula of any quantity in the model. This has been done to produce the confusion matrix in the figure 3.

2.1.4 Estimation or calibration of the implicit independent size

The main idea of this study is to observe multiple sampling or sub-sampling in order to estimate the way paper and board are mixed and how a sample represent a complete load.

Proof are in annex 6.3

Regression or MSE-matching formula The most intuitive way is analog to a regression of the size of the sample on the precision of the methods, and base on matching the mean square error in the model and in the observed data.

$$s^* = S\frac{\sum_{0\leqslant i\leqslant NMC}\left(p_{i,0}-p_{i,1}\right)^2}{\sum_{0\leqslant i\leqslant NMC}\left(p_{i,0}+p_{i,1}\right)\left(2-p_{i,0}-p_{i,1}\right)}$$

We propose numerical test of the model in annex 6.4 to observe the convergence.

Correlation matching As shown in figure 4, we can simply calibrate the model by matching the correlation computed in the model presented in 2.1.3 and the correlation computed from the data, 2.2.2.

Log-likelihood Log-likelihood of a realization of a random variable is an additive amount that is simply the logarithm of the density of the realisation. When observing independent realization of the same random variable we can find its law.

When the gaussian approximation is valid, e.g for small value of independent size model, we use the notation listed in table 1 and the procedure of sub-sampling defined in section 2.2.1.

For a sub-sampling realization p^O, with the notation of **??** using for the variance the notation $V = p\left(1-p\right)\frac{s}{S}$, the log-likelihood will be :

$L = -\frac{\left(p^o - p\right)^2}{2V} - \frac{1}{2}\log\left(V\right) - \frac{1}{2}\log\left(2\pi\right)$

The value of the parameter $\hat{s}$ of the calibrated model is

$$\hat{s} = \arg\max_{s>0}\left[\mathbb{E}\left[-\frac{\left(p^o - p\right)^2}{2V} - \frac{1}{2}\log\left(V\right) - \frac{1}{2}\log\left(2\pi\right)\right]\right]$$

This will be a double sum on all sample of a load on all sub-sample.

This is broadly equivalent to having the same MSE since it is equivalent asymptotically.

Since the optimized quantity is a convex function of s it can be optimized very easily.

2.2 Tractable methodology

Here start the main result of this study.

2.2.1 Sub-sampling or multiple sampling procedure

We refer to the sub-sampling defined in [1] or remained in section 1.1.

Sub-sampling means, having selected a sample, to split it in different sub-sample, this is what has been done to gather the data presented in section 2.3.

The multiple sampling procedure can imply bias since it is not always possible to keep the same procedure, for example when it is destructive as for a bale, or if a mixing process by a bulldozer change the homogenousity of the heap.

Whereas in the sub-sampling procedure, mixing the sample is required and more homogenous it is before sub-sampling, the better it is.

2.2.2 Computation of the correlation

For a bi-sampling procedure, sub-sampling with only two subsample, the formula to compute the correlation of the two independent measure of the same heap are the following. As listed in table 2, we note these estimators $p_i = p_i^g(P)$ and $p' = p_i^{g'}(P)$, for several $(i \leqslant N)$ different heap or load.

$$\rho^{g,g'} = \frac{1/N \sum p_i p'_i - 1/N \sum p_i 1/N \sum p'_i}{\sqrt{\left(1/N \sum p_i^2 - (1/N \sum p_i)^2\right)\left(1/N \sum p'^2_i - (1/N \sum p'_i)^2\right)}}$$

2.2.3 Correlation between the real proportion of the heap and the measure

Then the correlation of this sampling procedure with the real proportion p is :

$$\boxed{\rho^{g,p} = \sqrt{\rho^{g,g'}}} \tag{1}$$

And the more general equation for different methodology is :

$$\boxed{\rho^{g,i} = \rho^{g,p}\rho^{p,i}} \tag{2}$$

A proof can be found in annex 6.5.

2.2.4 Efficient formula for the confusion matrix

For the sake of simplicity we will not derive here the formula linking the correlation between two measure with the confusion matrix but will only says that it is linked to the gaussian copula.

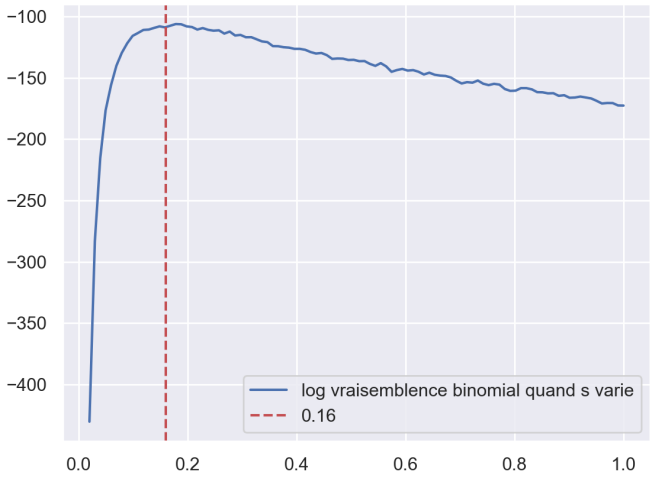

Figure 2: *Maximising log-likelihood*

2.3 Numerical Example

We made a practical study on the mixed paper and board coming out from a sorting plant during summer 2019.

2.3.1 Description of the data

The data are 80 caracterizations of bales with 20 classes computed.

For three of the classes a sub-sampling procedure of usually 4 differents sub-sample were done.

2.3.2 Statistical calibration of the model

In our case we found that the implicit independent size was $0.16Kg$ as shown in figure 2.

3 Neural network and image recognition for non invasive caracterisation of loose

3.1 Neural network

Neural network are object invented around year 1950' which after long years of gestation have find real application during the year 2010' thanks the cheap computing power of Graphical Processor Unit (GPU) and the large amount of data among them photos made available on the internet thanks ADSL during the year 2000' and smartphone and 3G/4G in the following decade.

3.2 Qualipapia

Qualipapia ia a recent developement of a technology to compute caracterization of load of paper for recycling thanks photo of the surface of an heap of loose paper.

3.3 Debiasing detection errors

As every machine learning methodology, it contains some errors. First order error or false positive, when we believe having found something but which is wrong. And second order error or false negative of tracked element that we have missed.

Thus. in the same spirit of the computation debiaising of counting Varroa, bee mites thanks deep learning presented in the internal report of Fotonower [12], we propose here to unbiased the surfacic detection of unwanted material.

Notation

We consider a classification or detection of **Paper/Unwanted**:

- P_p is the precision of paper
- R_p is the recall of paper
- P_{NP} is the precision of unwanted material
- R_{NP} is the recall of unwanted material
- NP_T is the exact number of paper element (or square centimeter)
- NP_D is the number of detected paper
- NP_{DP} is the number of paper detected as paper
- NNP_T is the exact number of unwanted element
- NNP_D is the number of unwanted detected as unwanted
- NNP_{DNP} is the number of unwanted
- E_r is the real proportion of unwanted material
- E_m is the measured proportion of unwanted material
- E_{dbm} is the unbiased proportion of unwanted material
- C is a constante $C = \frac{P_{NP} * R_P}{P_P * R_{NP}}$

Unbiasing formula

$$E_{dbm} = \frac{C * E_m}{(1 - E_m) + C * E_m}$$

A demonstration can be found in section 6.6.

3.4 Precision of this method

Precision-recall graph are the main tools to estimate the quality of a neural network detection method.

We provide confusion matrix in some case in figure 3.

3.5 From surfacic to weight characterization

Since it is based on private data, the transformation from surfacic to weight caracterization and detection is not in the scope of this study, but it can be achived quite straightforwardly in a caracterization campaign or based on theoretical density of material.

4 Conclusion

We have proposed in this study an original model to compute precision metrics of caracterization methods for one single load.

This is mainly based on estimation of correlation of double sampling procedure.

A simple formula has been proposed to link the correlation for different comparison of methodology.

Business-wised metrics as confusion between the quality of a load and the real quality of loads are then proposed to be computed in simple way.

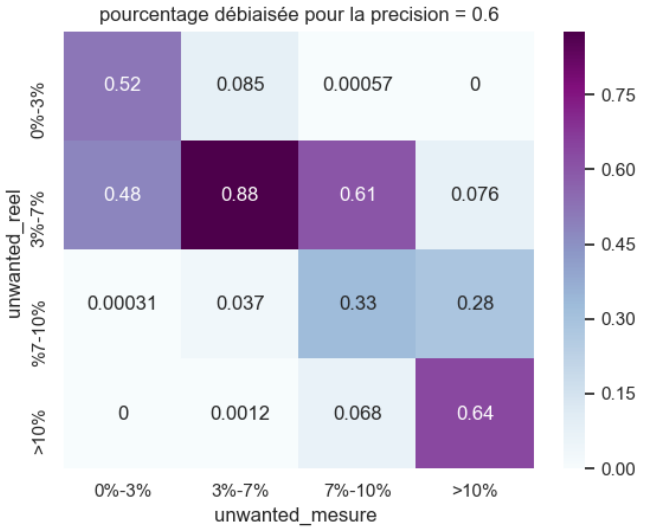

Figure 3: *Confusion matrix thanks unbiasing*

5 Bibliography

References

[1] Xp x30-474 : Déchets ménagers et assimilés - constitution d'un échantillon ponctuel sur une benne de déchets ménagers et assimilés collectés sélectivement. 2010.

[2] Nf x30-408 : Déchets ménagers et assimilés - méthode de caractérisation – analyse sur produit brut. 2013.

[3] Nf x30-466 : Déchets ménagers et assimilés - méthode de caractérisation – analyse sur produit sec. 2013.

[4] Louis-Rose S. Benedikt, P. Déchets ménagers et assimilés panorama des normes et documents normatifs existants. July 2013.

[5] Louis-Rose S. Benedikt, P. Guide pour la realisation de campagnes de caracterisation des dechets menagers : Conseils et methodes a destination des collectivites. April 2014.

[6] Follet-S. Naquin P. Bonnet, J. Actualisation du guide de mise en oeuvre d'une campagne locale de caractérisation des déchets ménagers et assimilés rapport intermédiaire : Analyse documentaire et état de l'art et analyse statistique. November 2013.

[7] European comitee for standardization. Paper and board - european list of standard grades of paper and board for recycling. *European comitee for standardization*, 2014.

[8] Andreas M Faul. Quality requirements in graphic paper recycling. *Cellulose Chemistry & Technology*, 44(10):451, 2010.

[9] E Frank. Zum verhalten bunter druckfarben beim deinken: eine auswertung von ingede-forschungsarbeiten. *Wochenblatt für Papierfabrikation*, 124(19):834–837, 1996.

[10] Anke Gottschling, Tobias Krebs, and Samuel Schabel. Sampling of paper and board for recycling : model helps to determine suitable sample sizes. *Professional Papermaking*, 12(2):42–47, December 2017.

[11] Tobias Krebs. Ap messsystem - video presentation - https://www.youtube.com/watch?v=07niczdc6ui, 2018.

[12] Chencheng Xu and Kevin Zagalo. Comptage de varroas. May 2019.

Lexic	Definition	Traduction	Übersetzung
Load	Complete load of opened bales or truck of loose material or specific amount of material on a conveyor for a fixed time or amount	Chargement	Laden
Sample	Part of a load that should be selected through a specific methods as the one listed in 1.2	Echantillon	Probe
Caracterization procedure	Process to identify manually or with a specified method all the element of a sample.	Procédure de caracterization	Charakterisierung Prozess
Calibration of the precision of the caracterization procedure.	Way to estimate the precision of a characterisation procedure of a sample of a load by having multiple sample of the same load, goal of this study.	Calibration de la précision d'une méthode de caractérisation.	Kalibrierung der Genauigkeit eines Charakterisierung Prozess

Table 1: *Definition of terminology*

6 Annex

6.1 Lexic

Lexic can be found in table 1.

6.2 Definition of variable

Definition of notation and their translation can be found in table 2.

6.3 Proof of the law of the metrics in the independent implicit size model

Correlation

Proof of formula given in 2.1.3.

Hypothesis On suppose que:

- La taille implicite s soit 200g
- Le taux p suit une loi uniforme $\mathcal{U}(0, 0.1)$
- La taille de tirage S_g soit 40kg par la méthode Gravimetric, d'où le nombre d'objets N_{gravi} est $\frac{40kg}{200g} = 200$
- La taille de tirage S_q soit 100kg pour le projet Qualipapia, d'où le nombre d'objets N_{quali} est $\frac{100kg}{200g} = 500$
- Le tas pèse 20 tonnes
- Le taux de erreur réelle p suit une loi uniforme $\mathcal{U}(0, M)(M < 1)$
- Le taux de erreur observée p_I, $p_I|p$ suit une loi normal $\mathcal{N}(p, \frac{p(1-p)}{N})$
- Les deux méthodes (Gravimetric/Qualipapia) sont indépendantes pour un même tas, c'est à dire pour un p fixé, $p_{gravi}|p$ et $p_{quali}|p$ sont indépendants.

Variable	Definition	Dimension	Section
$p^{(m)}(P)$	Proportion of material P for caracterization method m	Without dimension	2.1.4
$p_i^{(m)}(P)$	Proportion of material P for caracterization method m for sample i	Without dimension	2.2.2
S_i	Size of a sample	Kg	
s	Implicit size of independent element of a load	Kg	
$NMC_{i,}$	Number of sub-sample for sample i	No dimension	
$S_{i,j}$	Size of a sub-sample j of sample i	Kg	
P	Paper	No Dimension	6.6
NP	Non Paper	No Dimension	6.6
Pr	Precision of a detection method, proportion of the element correctly detected among the element detected.	No Dimension	3
R	Recall of a detection method, proportion of element detected among all the element to detect.	No Dimension	3
$\mathbb{E}$	Expectancy against a probability of observed event or in a probabilistic model	No Dimension	

Table 2: *Definition of variable*

Proof $\mathbb{E}(p) = M/2$
$Var(p) = M^2/12$
$\mathbb{E}(p^2) = Var(p) + (\mathbb{E}(p))^2 = M^2/3$
$\mathbb{E}(p_I) = \mathbb{E}(\mathbb{E}(p_I|p)) = \mathbb{E}(p) = M/2$
$Var(p_I) = \mathbb{E}(p_I^2) - (\mathbb{E}(p_I))^2 = \mathbb{E}(\mathbb{E}(p_I^2|p)) - (\mathbb{E}(p))^2$
$\mathbb{E}(\mathbb{E}(p_I^2|p)) = \mathbb{E}(Var(p_I|p) + (\mathbb{E}(p_I|p))^2) = \mathbb{E}(\frac{p(1-p)}{N} + p^2)$

On a donc:
$Var(p_I) = \mathbb{E}(\frac{p(1-p)}{N} + p^2) - (\mathbb{E}(p))^2 = \frac{\mathbb{E}(p-p^2)}{N} + Var(p) = \frac{M/2-M^2/3}{N} + M^2/12$
$\mathbb{E}(p_I p) = \mathbb{E}(\mathbb{E}(p_I p|p)) = \mathbb{E}(p\mathbb{E}(p_I|p)) = \mathbb{E}(p^2) = M^2/3$
$Cov(p, p_I) = \mathbb{E}(p_I p) - \mathbb{E}(p)\mathbb{E}(p_I) = M^2/12$
$Cor(p, p_I) = Cov(p, p_I)/\sqrt{Var(p)Var(p_I)} = \frac{1}{\sqrt{1+\frac{1}{N}(\frac{6}{M}-4)}}$

Avec la même méthode de calcul, on peut avoir aussi la corrélation de p_{gravi} et p_{quali}:
$Cor(p_{gravi}, p_{quali}) = \frac{1}{\sqrt{1+\frac{1}{N_{gravi}}(\frac{6}{M}-4)}\sqrt{1+\frac{1}{N_{quali}}(\frac{6}{M}-4)}}$

Figure La corrélation est une fonction croissante en N (nombre d'objets), on peut tracer la fonction pour le maximum du taux 10% dans le modele, c'est représenté sur la figure 4.

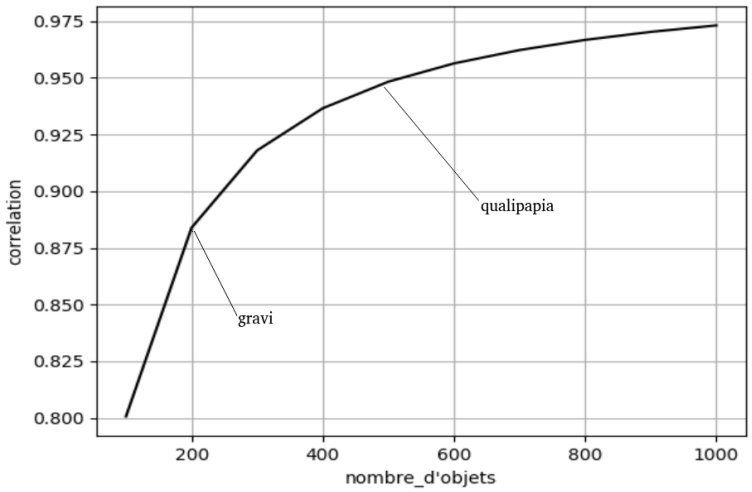

Figure 4: *Correlation pour M = 10%*

Proof of the estimation formula of the parameter

Proof of formula proposed in 2.1.4.

For the MSE

We observe NMC realizations of sample with NMC_i realization of sub-sample for each sample $i \leqslant NMC$ of size and percentage :

$$(p_{i,j}, S_{i,j})_{0 \leqslant j \leqslant NMC_i, 0 \leqslant i \leqslant NMC}$$

Thanks the Central Limit Theorem, we approximate the law of each sub-sample conditionnal to the sample with a gaussian using $p_i^o = \frac{\sum_{0 \leqslant j \leqslant NMC_i} p_{i,j} S_{i,j}}{\sum_{0 \leqslant j \leqslant NMC_i} S_{i,j}}$

$$p_{i,j} \sim \mathcal{N}\left(p_i^o, p_i^o\left(1 - p_i^o\right)\frac{s}{S_{i,j}}\right)$$

where s is the implicit independent size.
The empirical (observed) variance of sub-sample of a same sample are :

$$\sum_{0\leqslant j\leqslant NMC_i} p_{i,j}^2 - \left(p_i^o\right)^2$$

Therefore the estimator s^* of the implicit independent size that matchs the variance of the model and the empiric over all the sample follow the equation :

$$\sum_{0\leqslant i\leqslant NMC} \frac{1}{NMC_i} \sum_{0\leqslant j\leqslant NMC_i} p_{i,j}^2 - \left(p_i^o\right)^2 = \sum_{0\leqslant i\leqslant NMC} p_i^o\left(1-p_i^o\right)\frac{s^*}{S_{i,j}}$$

$$s^* = \frac{\sum_{0\leqslant i\leqslant NMC} \frac{1}{NMC_i} \sum_{0\leqslant j\leqslant NMC_i} p_{i,j}^2 - \left(p_i^o\right)^2}{\sum_{0\leqslant i\leqslant NMC} \frac{p_i^o\left(1-p_i^o\right)}{S_{i,j}}}$$

In the simple case where all sub-sample have same size S and where $NMC_i = 2$ it simplifies to :

$$\begin{aligned} s^* &= \frac{\sum_{0\leqslant i\leqslant NMC} p_{i,0}^2 + p_{i,1}^2 - 2\left(p_i^o\right)^2}{2\sum_{0\leqslant i\leqslant NMC} \frac{p_i^o\left(1-p_i^o\right)}{S}} \\ &= S\frac{\sum_{0\leqslant i\leqslant NMC} 2p_{i,0}^2 + 2p_{i,1}^2 - \left(p_{i,0}+p_{i,1}\right)^2}{\sum_{0\leqslant i\leqslant NMC}\left(p_{i,0}+p_{i,1}\right)\left(2-p_{i,0}-p_{i,1}\right)} \\ &= S\frac{\sum_{0\leqslant i\leqslant NMC}\left(p_{i,0}-p_{i,1}\right)^2}{\sum_{0\leqslant i\leqslant NMC}\left(p_{i,0}+p_{i,1}\right)\left(2-p_{i,0}-p_{i,1}\right)} \end{aligned}$$

6.4 Numerical validation of the asymptotic convergence of log-likelihood and MSE-matching

TODO

Simulated data

Observed data

6.5 Proof of the formula linking the correlation between two measure and between the measure and the reality

. We prove here the main equation 1 of this study from section 6.5

We consider the real proportion of a load being a random variable P, we consider two random variable Q and G conditional representing two independent measure, eventually with the same methods. We suppose independence and that the measure are not biaised :

- $Q|P \coprod P$, $G|P \coprod P$ and $Q|P \coprod G|P$
- $\mathbb{E}\left[Q|P\right] = P$ and $\mathbb{E}\left[G|P\right] = P$

First we have :

$$Var\left[Q\right] = \underbrace{\mathbb{E}\left[Var\left[Q|P\right]\right]}_{=W_P} + \underbrace{Var\left[P\right]}_{=V_P}$$

and

$$Covar\left[G,Q\right] = Var\left[P\right]$$

and

$$Covar\left[G,P\right] = Var\left[P\right]$$

Then

$$Corr\left[G,P\right] = \frac{V_P}{\sqrt{\left(V_P\right)}\sqrt{\left(V_P+W_G\right)}}$$

$$Corr\left[G,Q\right] = \frac{V_P}{\sqrt{(V_P + W_Q)}\sqrt{(V_P + W_G)}}$$

which proves the results.

6.6 Proof of the unbiasing formula of surfacic detection

Proof of formula found in 3.3.

- P_p is the precision of paper
- R_p is the recall of paper
- P_{NP} is the precision of unwanted material
- R_{NP} is the recall of unwanted material
- NP_T is the exact number of paper element (or square centimeter)
- NP_D is the number of detected paper
- NP_{DP} is the number of paper detected as paper
- NNP_T is the exact number of unwanted element
- NNP_D is the number of unwanted detected as unwanted
- NNP_{DNP} is the number of unwanted
- E_r is the real proportion of unwanted material
- E_m is the measured proportion of unwanted material
- E_{dbm} is the unbiased proportion of unwanted material
- C is a constante $C = \frac{P_{NP} * R_P}{P_P * R_{NP}}$

Par la définition de précision:

$$P_p = \frac{NP_{DP}}{NP_D} \tag{1}$$

$$P_{NP} = \frac{NNP_{DNP}}{NNP_D} \tag{2}$$

Par la définition de Rappel:

$$R_p = \frac{NP_{DP}}{NP_T} \tag{3}$$

$$R_{NP} = \frac{NNP_{DNP}}{NNP_T} \tag{4}$$

Par la définition de Pourcentage non-papier:

$$E_r = \frac{NNP_T}{NP_T + NNP_T} \tag{5}$$

$$E_m = \frac{NNP_D}{NP_D + NNP_D} \tag{6}$$

Par (1) et (3), nous avons:

$$NP_{DP} = P_P * NP_D = R_P * NP_T \tag{7}$$

Par (2) et (4), nous avons:

$$NNP_{DNP} = P_{NP} * NNP_D = R_{NP} * NNP_T \tag{8}$$

Nous pouvons calculer par (7) et (8):

$$NP_T = \frac{P_P * NP_D}{R_P}$$

$$NNP_T = \frac{P_{NP} * NNP_D}{R_{NP}}$$

$$\frac{NNP_T}{NP_T} = \frac{P_{NP} * R_P}{P_P * R_{NP}} * \frac{NNP_D}{NP_D} \tag{9}$$

Nous considérons $C = \frac{P_{NP} * R_P}{P_P * R_{NP}}$ une constante quand précision et rappel sont fixés, par (6), nous avons:

$$\frac{NNP_D}{NP_D} = \frac{E_m}{1 - E_m} \tag{10}$$

Alors, par (9) et (10):

$$\frac{NNP_T}{NP_T} = C * \frac{E_m}{1 - E_m}$$

Alors par (5):

$$E_r = \frac{C * \frac{E_m}{1-E_m}}{1 + C * \frac{E_m}{1-E_m}}$$

$$E_r = \frac{C * E_m}{(1 - E_m) + C * E_m}$$

Donc:

$$E_{dbm} = \frac{C * E_m}{(1 - E_m) + C * E_m}$$