

Quality evaluation of ancient digitized documents for binarization prediction

Vincent Rabeux
University of Bordeaux
LaBRi
Bordeaux
rabeux@labri.fr

Nicholas Journet
University of Bordeaux
LaBRi
Bordeaux
journet@labri.fr

Anne Vialard
University of Bordeaux
LaBRi
Bordeaux
vialard@labri.fr

Jean-Philippe Domenger
University of Bordeaux
LaBRi
Bordeaux
domenger@labri.fr

Abstract—This article proposes an approach to predict the result of binarization algorithms on a given document image according to its state of degradation. Indeed, historical documents suffer from different types of degradation which result in binarization errors. We intend to characterize the degradation of a document image by using different features based on the intensity, quantity and location of the degradation. These features allow us to build prediction models of binarization algorithms that are very accurate according to R^2 values and p-values. The prediction models are used to select the best binarization algorithm for a given document image. Obviously, this image-by-image strategy improves the binarization of the entire dataset.

I. INTRODUCTION

This paper involves quality evaluations of ancient document images. We propose a methodology based on algorithm prediction models for selecting the best algorithm for a specific task. Our approach is based on the following fact : the global quality of a document image directly impacts the result of any processing algorithm (binarization, segmentation,...). We thus propose to predict the result of an algorithm according to the type and quantity of the degradation of the processed document. We focus now on binarization prediction.

For a given binarization algorithm and a set of ground-truthed binarized images, a prediction function is built by searching a significant correlation between the algorithm performances and the quality of the images. The document image quality is measured with new dedicated features. The prediction function can then be used to predict the binarization algorithm result for any new image on which quality features have been previously computed.

To our knowledge there are no work on binarization prediction. The existing work on algorithm prediction in the field of document image analysis entails OCRs which typically use the quality features in order to create prediction models.

The first quality metrics were introduced in [1]. In this article the authors evaluate the quality of binary text documents by analysing black and white connected components. The OCR result is predicted by thresholding the quality measures (proportion of thick and broken characters). Each

document image is finally labeled as good or poor. In [2] two new measures are introduced to account for speckles and touching characters. A linear regression is used to predict the OCR performance on handwritten black and white documents. The authors of [3] complete the set of measures with new ones, which are used as inputs to a neural network to classify the images in two classes (poor or good). By reusing a script identification engine, the method proposed in [4] can select the better of two OCRs according to a classification of the text image as *broken*, *clean* or *merged*.

Other works propose strategies to select the best restoration algorithm. As in OCR prediction methods, dedicated defect features are computed on a binary image. These values are then used as inputs for different types of semi-supervised classification algorithms. The authors of [5] use the features of [3] with three new ones from [6] to select a restoration algorithm using a linear classifier.

Previous methods suffer from three main drawbacks. First, most of them require a connected component extraction and, therefore, a binarization step. These methods strongly depend on the accuracy of this preprocessing step. We believe that a better approach consist of directly analyzing the defect pixels in the initial grayscale image. Second, none of the presented articles dealing with prediction models analyze the significance of each feature.

In the following sections, we introduce different features characterizing ancient documents degradations. These features rely on a document gray levels decomposition in three different classes : ink pixels, degradation pixels and background pixels. We characterize the degradation layer by analyzing the distribution of its intensities, its quantity and its location within the image. The proposed features, dedicated to binarization evaluation, are presented in section II-B. Section III details the methodology used for creating algorithms prediction models. Prediction models of several binarization methods are then presented, all of which present very high accuracy. Finally, section III-C explains how to use the prediction models to select the best binarization algorithm for a specific image.

II. CHARACTERIZATION OF THE DEGRADATION LAYER

This section details new features used to characterize document image degradation. A first set of global features is extracted directly from grayscale histograms without spatial consideration. A second set of features characterizes the localization of the degradation.

A. The degradation layer extraction

We assume that an ancient document can be modeled as the combination of three different layers: the text pixel layer, the background pixel layer and the degradation pixel layer. Most of the degradation (for example, bleed-through, spots, speckles, non-uniform illumination, ink loss) appears as connected components with grayscale values that differ from background and ink pixels. We distinguish the three different layers of pixels according to the pixels' gray level. Let us denote the gray level of pixel p by $g(p)$. Let \mathcal{I} be the set of ink pixels, \mathcal{D} be the set of degradation pixels and \mathcal{B} be the set of background pixels defined as follows:

- 1) $\mathcal{I} = \{p, g(p) \leq s_0\}$ ink layer
- 2) $\mathcal{D} = \{p, s_0 < g(p) < s_1\}$ degradation layer
- 3) $\mathcal{B} = \{p, g(p) \geq s_1\}$ background layer

Setting the two thresholds s_0 and s_1 can be determined using any classification algorithm. Our experiments used a 3-means clustering algorithm. Table I shows that most degradation present in a document image can be extracted using these two thresholds.

B. Global Features

We compute the following global statistic features of the grayscale histogram: mean, variance and skewness. We denote the mean of the global histogram by μ , its variance by v , and its skewness by s . The mean, variance and skewness are also computed on the three *sub-histograms* to characterize each layer distribution (ink, background and degradation):

- μ, v, s (global histogram)
- $\mu_{\mathcal{I}}, v_{\mathcal{I}}, s_{\mathcal{I}}$ (ink histogram)
- $\mu_{\mathcal{D}}, v_{\mathcal{D}}, s_{\mathcal{D}}$ (degradation histogram)
- $\mu_{\mathcal{B}}, v_{\mathcal{B}}, s_{\mathcal{B}}$ (background histogram)

The previous global features characterizing the histograms cannot precisely represent the relationship between the ink layer, the degradation layer and the background layer. Therefore, we introduce two last global features extracted from the grayscale histogram to characterize the distance between the three layers : $\mathcal{M}\mathcal{I}_{\mathcal{I}}$ and $\mathcal{M}\mathcal{I}_{\mathcal{B}}$, where $\mathcal{M}\mathcal{I}_{\mathcal{I}}$ corresponds to the distance between the average intensity of degradation pixels and the average intensity of ink pixels and, $\mathcal{M}\mathcal{I}_{\mathcal{B}}$ is the distance between the average intensity of degradation pixels and the average intensity of background pixels. (Defined for a 8bit intensity range image).

$$\mathcal{M}\mathcal{I}_{\mathcal{I}} = \frac{\mu_{\mathcal{D}} - \mu_{\mathcal{I}}}{255} \quad \mathcal{M}\mathcal{I}_{\mathcal{B}} = \frac{\mu_{\mathcal{B}} - \mu_{\mathcal{D}}}{255}$$

The gray-values of the three layers are not the only characteristics that could affect a binarization algorithm. The amount of degradation pixels is also directly correlated with the binarization performance.

We measure this performance as the relative quantity of ink and degradation pixels. We define $\mathcal{M}\mathcal{Q}$ as the following ratio : $\mathcal{M}\mathcal{Q} = \frac{\|\mathcal{D}\|}{\|\mathcal{I}\|}$.

C. Spatial deformation features

As a good binarization should preserve the shape of the objects and avoid the creation of unwanted black or white components, the location of the degradation pixels is a significant characteristic that can influence the binarization result. Figure 1 illustrates the main situations observed in real documents in which the degradation pixels spatially interfere with ink pixels.

Let S be a set of pixels. We denote the set of the 4-connected components of S by $CC(S)$. In the rest of the section, we use the following notations : $\mathcal{C}_{\mathcal{I}} = CC(\mathcal{I})$, $\mathcal{C}_{\mathcal{D}} = CC(\mathcal{D})$ and $\mathcal{C}_{\mathcal{B}} = CC(\mathcal{B})$.

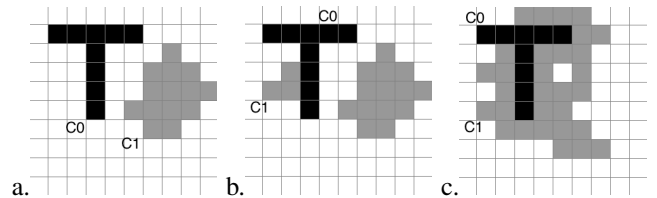


Figure 1. The different locations of a degradation component on the page: a. the degradation component is not connected to an ink component, b. a small degradation component is adjacent to an ink component, c. a large degradation component is adjacent to an ink component.

Let $c_{\mathcal{I}} \in \mathcal{C}_{\mathcal{I}}$ be an ink component and $c_{\mathcal{D}} \in \mathcal{C}_{\mathcal{D}}$ be a degradation component. We denote the predicate returning true if $c_{\mathcal{I}}$ and $c_{\mathcal{D}}$ are connected by $SG(c_{\mathcal{I}}, c_{\mathcal{D}})$:

$$SG(c_{\mathcal{I}}, c_{\mathcal{D}}) = \exists (p_{\mathcal{I}}, p_{\mathcal{D}}) \in c_{\mathcal{I}} \times c_{\mathcal{D}} \mid p_{\mathcal{I}} \text{ and } p_{\mathcal{D}} \text{ are 4-connected}$$

We distinguish three different cases that can produce different types of binarization errors:

- 1) If $c_{\mathcal{I}}$ and $c_{\mathcal{D}}$ are not connected (figure 1.a), the original character will not be altered by the binarization process. If this configuration occurs numerous times, the binarization can lead to a document image highly degraded by many small black spots between characters. Let $\mathcal{C}_{\mathcal{M}\mathcal{A}}$ be the set of degradation components that are not connected to any ink component:

$$\mathcal{C}_{\mathcal{M}\mathcal{A}} = \{c_{\mathcal{D}} \in \mathcal{C}_{\mathcal{D}} \mid \forall c_{\mathcal{I}} \in \mathcal{C}_{\mathcal{I}}, SG(c_{\mathcal{I}}, c_{\mathcal{D}}) = false\}$$

The relative quantity of non-connected ink and degradation components is measured by $\mathcal{M}\mathcal{A}$:

$$\mathcal{MA} = \frac{\|C_{\mathcal{MA}}\|}{\|C_I\|}$$

- 2) If c_I and c_D are connected (Figure 1.b), the original character may be altered by the binarization: degraded pixels may be misclassified as ink pixels. Let $C_{\mathcal{MS}}$ be the set of all ink components that are connected to at least one degradation component:

$$C_{\mathcal{MS}} = \{c_I \in C_I \mid \exists c_D \in C_D, SG(c_I, c_D)\}$$

The feature \mathcal{MS} is defined as the ratio between the number of ink components that may be expended by at least one degradation component and the total number of ink components:

$$\mathcal{MS} = \frac{\|C_{\mathcal{MS}}\|}{\|C_I\|}$$

- 3) \mathcal{MSG} measures the possible extent of ink component deformation using the number of known ink components that may be modified by the binarization process. It is defined as the mean area of the pairs of components that satisfy SG over the mean area of all ink components:

$$\mathcal{MSG} = \frac{\text{Average}_{\{(c_I, c_D) \mid SG(c_I, c_D)\}} (\|c_I\| + \|c_D\|)}{\text{Average}_{c_I \in C_I} (\|c_I\|)}$$

The higher \mathcal{MSG} is, the more likely it is that the document has large spots around ink components. Combined with other features (for example, \mathcal{MI}_I), \mathcal{MSG} helps predict whether the spots lead to binarization errors.

Given all of the previously defined features, each document image is characterized by a vector of dimension 18. An example is given in Table I which shows the degradation extraction and the values of the proposed features on one document image. The analysis of these values indicates that it may be preferable to use Sauvola's method to binarize this image. Indeed, the values of \mathcal{MI}_I and \mathcal{MI}_B are low meaning that a global thresholding method like Otsu's is likely to fail to correctly classify the pixels. The value of \mathcal{MSG} is also high : there are large spots around the characters. Window-based method have, most of the time, better results on this kind of documents. This hypothesis is confirmed with the f-score of Otsu's and Sauvola's methods. On this image, Otsu makes a score of 0.4 and Sauvola of 0.7.

III. PREDICTING BINARIZATION METHODS ACCURACY

The measures introduced in this paper characterize a document's quality. In this paper we focus on a use case that is rarely presented in the state of the art : the prediction of binarization methods accuracy.

This section presents a unified methodology that is able to predict most types of binarization methods (for example, adaptive thresholding, clustering, entropic, document dedicated). Our methodology is evaluated on 11 binarization methods used in document analysis. The methods are referenced in the text by their author's names : Bernsen ; Kapur ; Kittler ; Li ; Ridler ; Sauvola ; Otsu (these 7 methods are described in [7]); Sahoo [8]; Shanbag [9]; White [10]; Shijian [11].

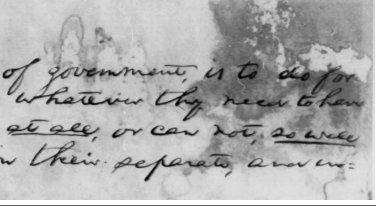
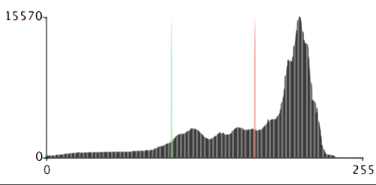

For this specific use case we follow a methodology based on a step-wise linear regression. This methodology can be applied to all types of binarization methods and therefore is first detailed in a general context. The following sub-section (III-A) presents this methodology and the dataset we used to train and validate our prediction models. In sub-section III-B, we analyze the accuracy of 3 binarization prediction models that are highly used in document images : Otsu, Sauvola and Shijian. In this article, the accuracy of the last 10 methods is not presented. However, all prediction models are used to create a process that selects the binarization algorithm that is the most suited for an image. The accuracy of this process (and therefore the accuracy of the prediction models) is analyzed in sub-section III-C.

A. Predicting algorithms results with a step wise multivariate linear regression.

To create the prediction model, we use a multivariate step wise linear regression [12], followed by a repeated random sub-sampling validation (cross validation). This over all process can be divided in several steps :

- 1) Features and F-scores computation: The 18 proposed features are computed for each image. We also run the binarization algorithm on the overall dataset and measure its accuracy relative to the ground truth. In the following section, these f-scores are called ground truth f-scores.
- 2) Generation of the predictive model : This step consists of applying a step wise multivariate linear regression to the overall dataset, allowing us to select the most significant features for predicting the given binarization algorithm. The output of this step is a linear function that gives a predicted f-score value for any image, for one binarization algorithm, knowing the selected features.
- 3) Evaluation of model accuracy: The R^2 value indicates the proportion of variability in a dataset that is accounted for by the statistical model and provides a measure of how well the model predicts future outcomes. The best theoretical value for R^2 is 1. Moreover, a p-value is computed for each selected feature indicating its significance. We choose to keep the model only if $R^2 > 0.7$ and if a majority of p-values are lower than 0.1.

Table I
EXAMPLE ON AN IMAGE FROM THE DIBCO DATASET : EXTRACTION OF THE DEGRADATION LAYER AND FEATURES VALUES.

Image					GrayScale Histogram					3-mean clusters							
																	
$\mathcal{M}I_I$	$\mathcal{M}I_B$	$\mathcal{M}Q$	$\mathcal{M}A$	$\mathcal{M}S$	$\mathcal{M}S\mathcal{G}$	s_i	s_g	s_b	v_i	v_g	v_b	μ_i	μ_g	μ_b	s	v	μ
0.2	0.1	0.3	0.05	0.2	3.6	-0.4	-0.05	-0.5	741	392	161	66	135	199	-1.25	2065	171

- 4) Model validation using Cross-Validation : the training of a prediction model and its accuracy measurement is done a several times (in our experiments : 100 times) on different subsets of a data-set :
 - a) The overall set of images is randomly split (90% is used as training set and 10% as validation set).
 - b) A prediction model is trained on the 90%.
 - c) On the validation set, the accuracy of the model is measured by two values. The R^2 and the slope coefficient of the validation regression, which also needs to be the closest to 1.
 - d) These two metrics are averaged on all splits.
- 5) The averaged metrics allows to statistically validate the prediction model.

In this experiment we use a merge of the well known datasets named DIBCO¹ and H-DIBCO².

B. Prediction models of commonly used binarization methods in document analysis systems

Otsu's binarization method: The selected most significant measures are : $\mathcal{M}I_I$, v_i , v_b , μ_b , μ and v . This can be explained by the fact that Otsu's binarization method is based on a global grayscale histogram thresholding. That is why measures such as $\mathcal{M}I_I$, μ and v are significant and have such low p-values. The estimated coefficients are presented in table II. By repeating 100 times a random sub-sampling validation gives a mean slope coefficient of 0.989 and a mean R^2 of 0.987. This cross validation step estimates that the predictive model will perform in practice.

Sauvola's binarization method: The selected measures are : $\mathcal{M}I_B$, $\mathcal{M}Q$, $\mathcal{M}A$, μ , s , s_i , v_i . The estimated coefficients are presented in table III. It is no surprise that $\mathcal{M}A$ is selected for this binarization method. Indeed window based methods are sensitive to small noise components without any real ink information. The cross validation by repeating 100 times a random sub-sampling validation gives a mean slope coefficient of 1.0007 and a mean R^2 of 0.99.

¹<http://users.iit.demokritos.gr/~bgat/DIBCO2009/>

²<http://users.iit.demokritos.gr/~bgat/H-DIBCO2010/>

Table II

OTSU PREDICTION MODEL : ALL MEASURES ARE SIGNIFICANT ($p - value < 0.1$), THE MODEL IS ALSO LIKELY TO PREDICT CORRECTLY FUTURE UNKNOWN IMAGES GIVEN THAT THE R^2 MEASURES AND ADJUSTED R^2 MEASURE ARE HIGHER THAT 0.9.

	Feature coef.	Std. Error	p-value
Intercept	1.187e + 00	1.604e - 01	< 0.0001
$\mathcal{M}I_I$	1.244e + 00	2.042e - 01	< 0.0001
v_i	2.422e - 02	1.534e - 02	< 0.1
v_b	-4.336e - 02	1.095e - 02	< 0.01
μ_b	-2.662e - 02	3.585e - 03	< 0.0001
μ	2.445e - 02	3.296e - 03	< 0.0001
v	3.262e - 04	5.326e - 05	< 0.0001

These results allows us to conclude that, as with Otsu's prediction model, this model is accurate and can be used in practice.

Table III

SAUVOLA PREDICTION MODEL : ALL MEASURES ARE SIGNIFICANT ($p - value < 0.1$), THE MODEL IS ALSO LIKELY TO PREDICT CORRECTLY FUTURE UNKNOWN IMAGES GIVEN THAT THE R^2 EQUALS 0.8 AND ADJUSTED R^2 EQUALS 0.77.

	Feature coef.	Std. Error	p-value
(Intercept)	1.61+00	1.9-01	< 0.0001
$\mathcal{M}I_B$	1.19	4.3e-01	< 0.01
$\mathcal{M}Q$	-1.1	2.6e-01	< 0.0005
$\mathcal{M}A$	2.3e-01	1.22e-02	< 0.05
μ	-4.56e-03	9.54e-04	< 0.0001
s	7.709e-02	2.334e-02	< 0.0001
s_i	1.431e-01	3.255e-02	< 0.0001
v_i	4.264e-04	8.307e-05	< 0.0001

Shijian binarization method: The Shijian binarization method is also very accurate. Indeed, the 100 cross validations give a mean R^2 of 0.99. The selected variables and their estimated coefficients are presented in table IV.

C. Automatic and optimal selection of binarization methods

Given a document image, a binarization method and its prediction model, we can compute all of the features required by the model and use them as inputs. The result is the predicted accuracy of this specific binarization method

Table IV
SHIJIAN PREDICTION MODEL : THE MODEL IS LIKELY TO PREDICT CORRECTLY FUTURE UNKNOWN IMAGES GIVEN THAT THE R^2 EQUALS 0.86 AND ADJUSTED R^2 EQUALS 0.82.

	Feature coef.	Std. Error	p-value
(Intercept)	1.068e+00	1.093e-01	< 0.0001
ML_B	-7.971e-01	3.003e-01	< 0.05
MA	3.162e-02	6.469e-03	< 0.0001
MSG	-3.276e-02	2.846e-03	< 0.0001
var	-1.389e-04	5.131e-05	< 0.0001
s_i	3.882e-02	2.219e-02	< 0.0001
s_g	1.328e-01	3.597e-02	< 0.001
μ_i	-4.004e-04	4.387e-04	< 0.5

for this specific image. Table V presents some f-score statistics obtained from binarizing the DIBCO dataset. The first line corresponds to the best theoretical f-scores (having the ground truth, we know for each image the binarization method that will provide the best f-score). The second line corresponds to the f-scores obtained using only Shijian’s method. The last line corresponds to the f-scores obtained using our automatic binarization selection. We analyse the accuracy of our binarization method selection algorithms in several ways. First, the method has a slightly better (2%) mean accuracy than using only Shijian’s method. Importantly, note that our algorithm has a higher global accuracy (the standard deviation equals 0.04). Last, the worst binarization result of our method is much higher than Shijian’s (56%). Second, we compared our method with the optimal selection that we can compute from the ground truth. The results are very similar, indicating that the prediction models are accurate enough to select the best binarization method for each image (70% perfect match). The mean error of our method is 0.009 (standard deviation equals 0.02), and, the worst error equals 0.06.

Table V
BINARIZATION OF THE DIBCO DATASET. COMPARISON BETWEEN THE BEST THEORETICAL F-SCORE (COMPUTED FROM THE GROUND TRUTH), F-SCORES OBTAINED USING ONLY SHIJIAN’S METHOD AND F-SCORES OBTAINED FROM OUR AUTOMATIC SELECTION.

F-Score	Mean	Std. Dev.	Min	Max
Optimal selection	0.913	0.04	0.77	0.96
Shijian	0.891	0.12	0.21	0.95
Automatic selection	0.906	0.04	0.77	0.96

IV. CONCLUSION AND RESEARCH PERSPECTIVES

This paper presented 18 features that characterize the quality of a document image. These features are used a in step-wise multivariate linear regression to create prediction models for 11 binarization methods. Repeated random sub-sampling cross-validation shows that 10 of 11 models are very accurate and can be used to automatically choose the best binarization method. Moreover, given the step-wise approach of the linear regression, these models are not over

parameterized. One of our future research goals is to apply the same methodology to predict OCR error rates.

ACKNOWLEDGEMENTS

We would like to thanks the DIBCO team for providing datasets. This work was completed within the DIGIDOC project financed by the ANR (*Agence Nationale de la Recherche*)

REFERENCES

- [1] L. Blando, J. Kanai, and T. Nartker, “Prediction of ocr accuracy using simple image features,” in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 319–322.
- [2] M. Cannon, J. Hochberg, P. Kelly, and J. White, “An automated system for numerically rating document image quality,” in *Proceedings of SDIUT, 1997*, pp. 162–167.
- [3] J. Gonzalez, J. Kanai, and T. Nartker, “Prediction of ocr accuracy using a neural network,” *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*, vol. 29, pp. 356–370, 1998.
- [4] V. Ablavsky, J. Pollak, M. Snorrason, and S. M.R., “Ocr accuracy prediction as a script identification problem,” in *Proceedings of the 2003 Symposium on Document Image Understanding Technology*, D. Doermann, Ed., 2003, pp. 135–142.
- [5] A. Souza, M. Cheriet, S. Naoi, and C. Suen, “Automatic filter selection using image quality assessment,” in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 2003, pp. 508–512.
- [6] M. Cannon, J. Hochberg, and P. Kelly, “Quality assessment and restoration of typewritten document images,” *International Journal on Document Analysis and Recognition*, vol. 2, no. 2, pp. 80–89, 1999.
- [7] P. Stathis, E. Kavallieratou, and N. Papamarkos, “An evaluation survey of binarization algorithms on historical documents,” in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.
- [8] J. Sauvola and M. Pietikäinen, “Adaptive document image binarization,” *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.
- [9] A. Shanbhag, “Utilization of information measure as a means of image thresholding,” *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 5, pp. 414–419, 1994.
- [10] J. White and G. Rohrer, “Image thresholding for optical character recognition and other applications requiring character image extraction,” *IBM Journal of Research and Development*, vol. 27, no. 4, pp. 400–411, 1983.
- [11] S. Lu, B. Su, and C. L. Tan, “Document image binarization using background estimation and stroke edges,” *International journal on document analysis and recognition*, vol. 13, no. 4, pp. 303–314, 2010.
- [12] M. Thompson, “Selection of variables in multiple regression: Part i. a review and evaluation,” *International Statistical Review/Revue Internationale de Statistique*, pp. 1–19, 1978.