Heliyon 7 (2021) e08413

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

On the shape of timings distributions in free-text keystroke dynamics profiles

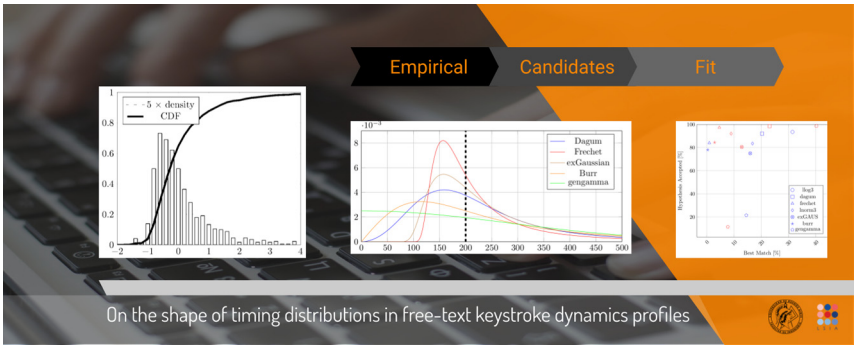
Nahuel González [a,*], Enrique P. Calot [a], Jorge S. Ierache [a], Waldo Hasperué [b]

[a] Laboratorio de Sistemas de Información Avanzados (LSIA), Facultad de Ingeniería, Universidad de Buenos Aires, Argentina
[b] Instituto de Investigación en Informática (III-LIDI), Facultad de Informática, Universidad Nacional de La Plata, Argentina

HIGHLIGHTS

- No systematic comparison of distributions to fit keystroke timings in free-text has been carried out yet.
- Most keystroke timings in free text do not follow a gaussian law.
- The three-parameter log-logistic distribution provides the best fit for hold times and flight times, over three datasets.
- Other previously considered distributions, like the log-normal, provide a good fit but not as good as the log-logistic.

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://doi.org/10.17632/sjk7kz35nh.1
Dataset link: https://doi.org/10.21227/ngv9-fa18

Keywords:
Soft biometrics
Keystroke dynamics
Free text
Probability distribution

ABSTRACT

Keystroke dynamics is a soft biometric trait. Although the shape of the timing distributions in keystroke dynamics profiles is a central element for the accurate modeling of the behavioral patterns of the user, a simplified approach has been to presuppose normality. Careful consideration of the individual shapes for the timing models could lead to improvements in the error rates of current methods or possibly inspire new ones. The main objective of this study is to compare several heavy-tailed and positively skewed candidate distributions in order to rank them according to their merit for fitting timing histograms in keystroke dynamics profiles. Results are summarized in three ways: counting how many times each candidate distribution provides the best fit and ranking them in order of success, measuring average information content, and ranking candidate distributions according to the frequency of hypothesis rejection with an Anderson-Darling goodness of fit test. Seven distributions with two parameters and seven with three were evaluated against three publicly available free-text keystroke dynamics datasets. The results confirm the established use in the research community of the log-normal distribution, in its two- and three-parameter variations, as excellent choices for modeling the shape of timings histograms in keystroke dynamics profiles. However, the log-logistic distribution emerges as a clear winner among all two- and three-parameter candidates, consistently surpassing the log-normal and all the rest under the three evaluation criteria for both hold and flight times.

* Principal corresponding author.
E-mail address: ngonzalez@lsia.fi.uba.ar (N. González).

https://doi.org/10.1016/j.heliyon.2021.e08413
Received 8 July 2021; Received in revised form 29 August 2021; Accepted 12 November 2021

1. Introduction

Whenever we type on a computer keyboard or a mobile device, our characteristic behavior leaves a trace that can be used to verify our identity. A keystroke dynamics authentication system leveraging the timings between successive key events has been proposed forty years ago [1]; since then, many improvements have been evaluated in the literature [2]. Nowadays, using a state-of-the-art neural network, it is possible to scale keystroke dynamics authentication systems to hundreds of thousands of users, with low error rates even when little training data for each of them is available [3]. Although identity verification has been the most studied application of keystroke dynamics, several others exist. For example, the analysis of keystroke timings have been used to detect deceptive intents on the part of the writer [4], to flag accounts spreading fake news about COVID [5], and even to infer the emotional state of the user [6].

Keystroke dynamics is a soft biometric trait. The general idea behind its analysis, which remains the same for authentication and other tasks, is to model how the legitimate user would have typed under certain conditions. The training data for those models generally consists of the past observations of keystroke timings, and possibly other features like pressure and acceleration, when available. For example, a baseline method to verify whether a password has been typed by the legitimate user or an impostor is, for each key, to average all the observations in his or her profile from previous logins, and then to calculate a scaled distance to the keystroke timings vector of the current login attempt [7].

Which distribution or distributions do these timings obey? Though clearly a central element for the accurate modeling of the behavioral pattern of the user, this question has not been explicitly addressed except by a few authors. Usually the approach has been to presuppose normality of the underlying variables [8] or to assume the masking effect produced by smoothing distance formulas, or other metrics employed, will make deviations or biases in the individual terms disappear in the whole [9]. By the fact that previous methods appear to work rather well using only two parameters (mean and variance) and presupposing normality for keystroke timing models, it seems evident that this founding postulate is not far from reality.

On the other hand, careful consideration of the individual shapes of keystroke timing distributions could lead to improvements in the error rates of current methods or possibly inspire new ones, as long as the distributions can be modeled in a simple way with a few parameters. Else, the models would suffer from overfitting and unnecessary complexity would be added to implementations. Joking about a similar topic, John von Neumann has been quoted asserting that *"with four parameters I can fit an elephant, and with five I can make him wiggle his trunk"* [10]. Following his witty remark, the candidate distributions in this study were restricted to those with two or three parameters.

With the advent of free-text keystroke dynamics analysis, the problem of fitting keystroke timing profiles has grown in importance. It can be expected that passwords and short fixed texts, consistently typed in a row with a rather stable cadence, would produce normally distributed timing profiles. But free text involves pauses and hesitations of many different kinds. Thinking, looking at the keyboard, resting, external interruptions, etc., occur invariably however short the sample might be, skewing the distribution, changing its shape, and adding heavy tails. Considering that most distance metrics and classification methods are sensitive to discrepancies between the assumed model and the empirical data, it is puzzling that a systematic study of histogram shapes was not an early step in the discipline. Not long ago a systematic comparison of a large number of candidates has been carried out [11, 12] but, unfortunately, it is restricted to fixed text.

1.1. Contribution

The aim of this paper is to evaluate several candidate distributions and rank them according to their merit for fitting the timing histograms of free-text keystroke dynamics profiles. As will be discussed in section 2, similar studies have been conducted using passwords and fixed texts as source material, but the difference between those and a free text typing task justifies an experiment for the latter. To the best of our knowledge, no systematic comparison of distributions for fitting the timing histograms of free-text keystroke dynamics profiles can be found in the literature. The main contributions are

- Evaluating seven distributions with two parameters and seven more with three parameters for the task of fitting timing histograms of free-text keystroke dynamics profiles, using three criteria: best match percentage, Akaike information criterion, and hypothesis rejection rate.
- Showing, based on the results of the experiment, that the three-parameter log-logistic distribution, which has not been considered before in the literature of free-text keystroke dynamics, provides the best fit for flight times.
- Making the source and results datasets publicly available to help research in the topic, as well as to encourage replication and validation of these results.

1.2. Organization

The rest of the article is organized as follows. Section 2 reviews past approaches to modeling the timing distributions. Section 3 describes the experimental setting, including the problem statement, methodological guidelines, candidate distributions, evaluation details, and presentation of results, together with a description of source datasets, tools, and the availability of the resulting dataset. Section 4 discusses the results and future lines of research derived from this work. The last section summarizes the conclusions.

2. Previous studies

Even though they share many similarities, the difference between passwords and free text is significant enough to demand divergent approaches when dealing with keystroke dynamics. For example, one of the earliest attempts at user verification using features derived from keystroke timings noticed a fivefold increase in error rates when methods that performed well for passwords were applied to free text [13]. This motivated the introduction by other authors of specialized methods for the task, starting with the R metric [14, 15] that required samples of around 800 characters, and lately reaching low error rates when scaled to thousands of users using a neural network specialized for free text [3]. The difference between passwords and free text can be justified by considering the decision interval that precedes the motor process; while passwords are typed straightforwardly, during composition or transcription of longer texts the user introduces pauses to think or read the next set of words, hesitates, and is subject to involuntary interruptions in the flow of the task. Thus, the distribution of the keystroke timings are expected to differ significantly.

The log–normal distribution was first proposed by Montalvão et al. [16] as a good fit for the flight times in their datasets. Instead of using it directly, a transformation was applied to the empirical data to approximate and simplify its cumulative distribution function; the overall effect does not differ much from assuming lognormality. As their main objective was improving the performance of algorithms that do not incorporate intervals distribution equalization, no attempt was made at justifying the choice beyond the empirical fit and the error rate reduction which is successfully achieved. However, the authors pioneered the idea of setting aside the assumption of normality in the field of keystroke dynamics analysis.

From that milestone onwards, the log-normal distribution has been a common choice for modeling flight times whenever the shape of the histograms is considered. Monaco et al. have shown lowered error rates in comparison with other anomaly detectors while testing a partially observable hidden Markov model that used the log-normal as the density function for time intervals [17]. The distribution was also proved useful as the underlying model for a spoofing attack with partial information by the same authors [18]. A plausible explanation for the resemblance to a log-normal of the empirical shape of distributions resulting from human dynamics (not specific for typing tasks) is offered by Barabási [19], deducing it from the hypothesis that human subjects execute tasks using a decision-based queueing process with priorities.

Assuming that users type with a normally distributed pure motor delay and an exponentially distributed decision time between keys, the idea of fitting flight times with an exgaussian distribution is natural. Chukharev-Hudilainen [20] used the scale parameter of the exgaussian distributions fitting user flight times to detect pauses and linguistic hesitation, and to shed light on the psycholinguistic processes underlying the typing task. As he notes, there is a rich and long-established literature in psychology about using an exgaussian to fit response times, which are very similar to flight times in both their empirical distributions and their theoretical model of occurrence. For example, see [23] or [24], where the values of the parameters estimated from the measured response times were used to infer task conflicts.

Comparisons between candidate distributions are sparse in the literature, where most of the time the problem is ignored. Two counterexamples are [21] and [22], where a log-normal is compared against Benford and Zipf's power laws and an exponential. To our best knowledge, this is the first systematic attempt to compare several distributions for fitting keystroke dynamics timing profiles when the text is not short and fixed, as in a password or a passphrase. Attempting to overcome the limitations in existing datasets, Migdal and Rosenberger [11, 12] have carried out a detailed comparison of almost twenty candidate distributions for the generation of synthetic datasets using statistical models; the Gumbel distribution provided the best overall fit. Our approach differs in the target tasks that were considered and the evaluation criteria; while theirs, using the GREYC dataset [25], represents short fixed texts like usernames and passwords that the user has typed repeatedly, ours is focused on free text composition and transcription tasks. Unsurprisingly, the results differ between the two studies but the Gumbel distribution still performs adequately.

A structured literature review was carried out to find out more recent studies on the topic. The academic databases Google Scholar, Microsoft Academic, and Scopus were queried with the software *Publish or Perish 7* [26], using the mandatory keywords *keystroke dynamics*, *distributions*, and *lognormal* to restrict the search. The objective of including the last keyword was to cut down on the number of results and concentrate on the most relevant ones; the lognormal being the first distribution evaluated for the purpose of fitting timings histograms in free text keystroke dynamics profiles [16] beyond the gaussian, it is expected to be used as a base case in any comparison. Publications from 2017 onwards were considered and sorted by relevance. A total of 15 studies meeting the aforementioned criteria were found in Google Scholar, 6 in Microsoft Academic, and 10 in Scopus, with several overlaps. Those only mentioning but not dealing directly with keystroke dynamics were manually filtered.

Most of the remaining studies consider the lognormal and other related distributions in a more general setting that includes not only keystroke dynamics but also touch-screen biometrics [27]. Going beyond authentication, [28] and [29] employ the sigma-lognormal model of rapid human movements to detect the age group of users based on their interaction with a touch screen, while [30] leverages different distributions to discriminate a human user from a bot. No other systematic comparison of distributions for the task of fitting keystroke timings histograms was found other than the aforementioned [21], [22], and [11, 12]. A historical summary of distributions evaluated for the purpose is shown in Table 1.

Table 1. History of distributions evaluated for the task.

Authors	Year	Distributions
Montalvão et al. [16]	2006	lognormal
Chukharev-Hudilainen [20]	2014	exgaussian
Monaco et al. [21]	2015	Benford
Iorliam et al. [22]		Zipf
Migdal & Rosenberger [12]	2019	arcsine, beta, betaprime, chi, chisquare, erlang, exponential, gamma, gumbel, laplace, logistic, normal, lognormal, rayleigh, raised cosine, Student's t, uniform, triangular

3. Experimental setting

3.1. Problem statement

The main objective of this study is to compare several candidate distributions in order to rank them according to their merit for fitting timing histograms in keystroke dynamics profiles. Note that the problem is different and more complex than simply fitting one set of observations to a distribution, because the histogram for every alphanumeric key should be modeled. Thus the search is for a distribution that provides the best fit more often, or fits better on average, or is rejected less often, across the collection of all profiles. For a precise definition of the latter term please refer to subsection 3.6.

Additionally, merit should not be addressed only as the minimization of a certain measure of the difference between the model and empirical data. Several additional concerns not having to do with the purely statistical problem of fitting are relevant when the models are meant to be used in real biometric systems. To begin with, the number of parameters in the candidate distributions should be as small as possible. Not only to avoid the theoretical nuisances of overfitting and hindering generalizability, but also because the estimation of additional parameters requires more observations to be reliable. Thus, choosing highly parameterized distributions as models lengthens the training required before a biometric system can achieve a low error rate. This is a practical concern for continuous keystroke dynamics verification and spoofing attacks like [18], which become more effective when fewer keystrokes are needed to bootstrap reliable profiles.

Ideally, a statistical model should provide both predictive and explanatory power [31]. The former one was already addressed in the form of fitting quality. The latter one, though at first glance not immediately relevant for the practical issues of biometric verification, can nonetheless find applications in the extended field of keystroke dynamics analysis. After all, the observed timing histogram is the result of neural and motor processes which, if correctly modeled, can shed light on the physical and emotional state of the user. In the same way that the mean flight time is highly correlated with typing proficiency and its variance can help to classify the emotional state of the user [32] or detect cognitive impairment [33], the parameters of the highest ranking distributions should offer some kind of insight into the underlying processes.

3.2. Methodological guidelines

Several common pitfalls have proliferated in keystroke dynamics studies, and biometrics in general, to such an extent that they have

motivated specialized literature addressing and trying to prevent them. Killourhy and Maxion [34] emphasize the importance of conducting comparative experiments in contrast to one-off evaluations — where a new technology and a new dataset are evaluated together — and strengthening conclusions with statistical inference. Jain et al. have proposed a set of guidelines [35] for best practices in biometrics research; even though they are aimed at the evaluation of biometric recognition systems and thus do not extrapolate directly to this experiment, the rules that apply were followed.

More explicitly, to make sure this experiment provides value to the research community, it was made sure that:

- *The experiment is comparative*. Three publicly available datasets are used for this study. The candidate distributions include common cases previously used in the literature as a baseline.
- *The experiment is replicable*. The datasets and the base tools are free and openly available. No additional data or tool is needed to replicate our results.
- *The conclusions are generalizable*. The raw data comes from different studies, was captured under different circumstances, with different users and scopes, and by different authors. The datasets are extensive and at least LSIA corresponds to real operational data.
- *The conclusions are founded on statistical inference*. Whenever possible, confidence intervals and other techniques are used to show significance.

With respect to the last item, there is no standard set of guidelines specific for biometric studies according to the best of our knowledge. To make up for it, [36] which is aimed at the medical research community is followed here.

3.3. Datasets

The dataset LSIA is the same as previously used in [37], which is a longer and harsher version than the one used in [38]. It consists of typing sessions recorded during daily work in a healthcare environment for more than a year where the users, mostly doctors, worked rotating shifts and on duty.

The dataset KM was used in [39] to evaluate if a free writing or a transcription task would produce similar enough profiles to be used interchangeably with the purpose of simplifying future data acquisitions, as volunteers were found more prone to contribute transcribed texts than composing original ones. The task separation for this study was kept, but merging the different groups in which the users were split.

The dataset PROSODY was used in [40] to study cues of deceptive intent reflected in typing pattern variations. It contains deceptive and truthful writing, as well as copying tasks where the user agrees or disagrees with the content. Gay marriage and gun control are included as controversial and opinionated topics to induce strong emotional responses in the writers and a more neutral topic of restaurant reviews balances the set. Such a rich source of diversity deserves not to be crippled, especially considering that each subset has enough samples to be significant. Thus, results for each of its categories of topic and task are reported.

Every dataset used here was captured in a different environment and setting, with a different set of users. Varying tasks were requested, ranging from simple copying to free composition. Emotional dimensions are explicitly considered in the last dataset and at least LSIA includes the effect of stress and fatigue. The diversity of sources is expected to help avoid single dataset caveats [41] such as selection and capture bias. Main dataset characteristics like user and session count are shown in Table 2.

Table 2. Main dataset characteristics.

Dataset		Task	Users	Profiles $N \geq 20$	Profiles $N \geq 40$
LSIA		Free text	42	5167	2777
KM		Free text	20	1644	854
		Transcription		1551	840
PROSODY	GAY	Copy 1	400	13188	8911
		Copy 2		12909	8772
		Fake Essay		13697	9233
		True Essay		15362	9959
	GUN	Copy 1	400	13996	9380
		Copy 2		13408	9148
		Fake Essay		13991	9374
		True Essay		16026	10380
	REVIEW	Copy 1	500	13447	9850
		Copy 2		13638	9902
		Fake Review		14890	10483
		True Review		15937	10909

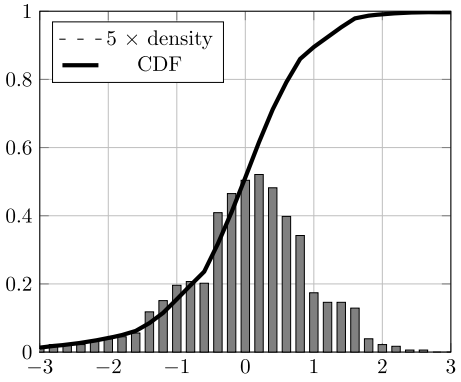


Fig. 1. A typical histogram for hold times, taken from dataset KM.

3.4. Availability of source and results data

A dataset containing CSV files with the timing features (hold times and flight times) of every keypress in the three source datasets — grouped by dataset, user, task, virtual key code, and feature — is made available both at the laboratory website, at IEEE DataPort [42], and as a Mendeley Data repository [43]. No additional data is needed to replicate the results here reported or to evaluate additional distributions that were not considered in this study.

The files in the dataset are named using the following convention: DATASET-TASK-USER-FEATURE-VK, and organized in folders according to their dataset and task. Due to the number of files being greater than a hundred thousand, they are packaged in the DISTRIBUTIONS.zip file. Five files, which are also included inside the package, are added to exemplify the naming convention. For example, KM-transcribed-USERs019-FT-VK32.csv contains the timing observations for the flight time (FT) of the space key (VK32, virtual key code 32) when pressed by the user s019 in the dataset KM, while he is carrying out a transcription task. Each file contains a single timing value per line, in miliseconds, for an observation of the corresponding feature, virtual key, and user.

3.5. Candidate distributions

It has been recognized for a while that hold times (down-up) and flight times (down-down) are expected to show different histogram shapes, the former being rather similar to a normal variable while the latter presents fatter tails and positive skew. Figs. 1 and 2 show the density and cumulative distribution of hold times and flight times for the space key of the user *s019* from dataset KM, grouped in buckets of 0.2 standard deviations around the mean, to exemplify these assertions.

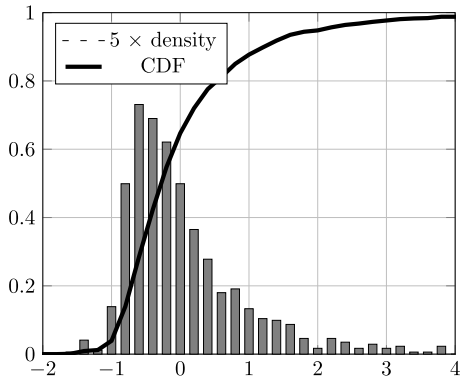


Fig. 2. A typical histogram for flight times, taken from dataset KM.

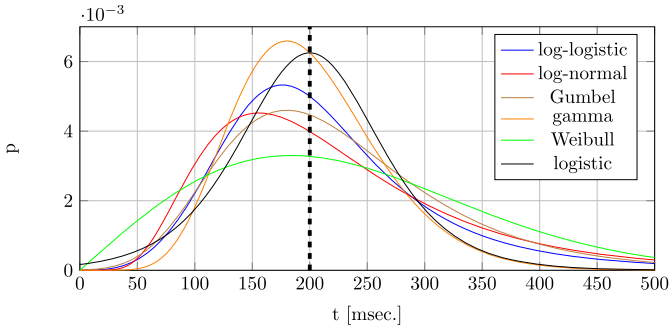


Fig. 3. Candidate distributions with two parameters.

A visual inspection of histograms from different users and datasets is enough to convince oneself that they provide a rather accurate description of empirical data.

As candidates for evaluation, seven well-known distributions with two parameters and seven more with three parameters were picked from [44]. Three formal requisites were being right tailed, with positive skew and infinite support, at least in the positive real numbers. A general shape resemblance to observed profiles was an additional subjective criterion. An attempt was made to represent different families of distributions as much as possible, preferring those with shape parameters that controlled the skewness, to account both for hold times and flight times. Finally, practical considerations restricted the choice to those having an implementation in R, compatible with the package `fitdistrplus`.

The chosen two-parameter candidates were log-normal (lnorm), logistic (log), log-logistic (llog), gamma, Weibull, and Gumbel; a normal distribution was also included as a comparison baseline. The terms in parentheses show the abbreviations used for tables, which match the corresponding R function. Fig. 3 exemplifies the first six, which have been fit to have their mean at 200 ms and a variance consistent with empirical examples of keystroke timing histograms as those made available with the results dataset.

The chosen three-parameter candidates were exgaussian (exGAUS), translated log-normal (lnorm3), translated log-logistic (llog3) Burr, Frechet, generalized gamma (gg), and Dagum. Once again, the terms in parentheses show the abbreviations used for tables, which match the corresponding R function. Fig. 4 exemplifies five of them, under the same conditions as in the previous one. Translated log-normal and log-logistic are not shown, as they have the same shape as their two-parameter counterparts but include an additional threshold parameter with can shift them left or right.

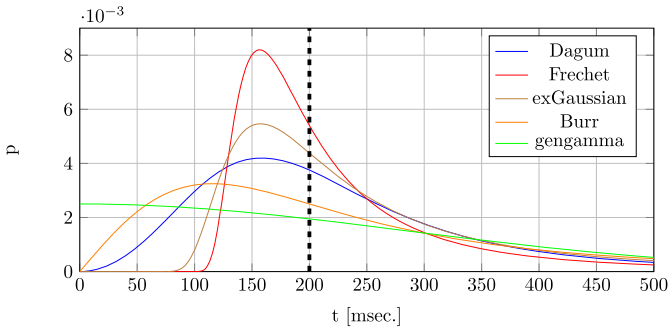


Fig. 4. Candidate distributions with three parameters.

3.6. Evaluation

Each considered dataset contains several typing sessions for each user, consisting of a sequence of keystrokes where its hold times (down-up) and flight times (down-down) were recorded alongside other relevant information. All of the latter was ignored except for the name of the typing task. The rationale for this action was to observe how different tasks influence the best-fitting distributions. Keystrokes were grouped on a per-user basis, packing them independently of their sessions. Thus, a profile was built for each dataset, task, user, virtual key code, and feature, consisting of a set of timing values. The temporal evolution of individual cadences was not considered, forming instead a single histogram for each timing profile.

All candidate distributions were evaluated against each alphanumeric profile with enough keystrokes (20 for two parameters and 40 for three), truncating them to 100 samples at most for performance considerations. Their parameters were estimated using maximum likelihood estimation and the resulting log-likelihood was corrected using Akaike's additional terms for parameter count and small sample bias, the latter only if required. Then a hypothesis verification was applied using Anderson-Darling goodness of fit test. The latter is better suited to the task at hand because it is more sensitive towards the tails than to the bulge of the distributions, for example in comparison with Kolmogoroff-Smirnoff test. An additional advantage is that R's implementation seamlessly compensates for repetitions in the sample set, a cumbersome artifact of recording keystroke timings with a discrete clock.

Please note that performance metrics typically used to rank methods for intrusion detection and identity verification, like precision, accuracy, FAR, FRR, ERR, etc., are not suited for this context, where keystroke timing distributions in free text are being fitted. In their place, it is natural to consider the number of times a distribution achieves the best fit against the other candidates, and verify the results using hypothesis testing, which is a standard tool when fitting sample data. From several available tests, the Anderson-Darling has been shown to be a top performer even with small sample sizes [45] and has the advantage of working well with most empirical distributions and candidate laws [46]. Finally, the Akaike information criterion was chosen for its ability to compensate for the number of parameters in the distributions.

3.7. Tools

The statistical software tool `R` [47] was employed for most of the complex calculations of this study. Package `fitdistrplus` [48] provided the core functionality for fitting and `ADGofTest` [49] the implementation of Anderson-Darling goodness of fit test, while packages `actuar` [50], `brms` [51], `distr` [52], `FAdist` [53], `gamlss.dist` [54] and `qualityTools` [55] provided the candidate distributions. Additional modules like dataset files parsing, table building, and general module gluing were done in C#.

Table 3. Results by dataset.

Dataset		Task	Two parameters			Three parameters		
			Best HT/FT	Avg. HT/FT	Least rej. HT/FT	Best HT/FT	Avg. HT/FT	Least rej. HT/FT
LSIA		Free text	**gumbel**/llogis	llogis/llogis	llogis/llogis	**lnorm3**/llog3	**dagum**/llog3	llog3/llog3
KM		Free text	logis/llogis	logis/llogis	logis/llogis	llog3/llog3	llog3/llog3	**lnorm3/dagum**
		Transcription	**gumbel**/llogis		logis/**lnorm**	**lnorm3**/llog3	**dagum**/llog3	**dagum/lnorm3**
PROSODY	GAY	Copy 1	logis/llogis	logis/llogis	logis/llogis	llog3/llog3	llog3/llog3	llog3/llog3
		Copy 2						
		Fake Essay						
		True Essay						llog3/**dagum**
	GUN	Copy 1	logis/llogis	logis/llogis	logis/llogis	llog3/llog3	llog3/llog3	llog3/llog3
		Copy 2						
		Fake Essay						
		True Essay						
	REVIEW	Copy 1	logis/llogis	logis/llogis	logis/llogis	llog3/llog3	llog3/llog3	llog3/llog3
		Copy 2						
		Fake Essay						
		True Essay						

Table 4. Distributions merit for hold times ordered by best match count, two-parameter distributions.

Dataset		Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
LSIA		Free text	22.51%	12.07%	26.85%	11.36%	8.52%	12.29%	6.39%
KM		Free text	24.59%	7.03%	14.05%	8.67%	7.49%	29.74%	8.43%
		Transcription	19.76%	6.9%	16.43%	9.29%	12.86%	24.76%	10%
PROSODY	GAY	Copy 1	35.65%	10.02%	18.26%	7.01%	11.49%	11.63%	5.94%
		Copy 2	35.3%	10.18%	17.92%	7.77%	10.45%	12.55%	5.83%
		Fake Essay	33.98%	10.67%	17.4%	8.33%	10.76%	12.74%	6.12%
		True Essay	35.81%	9.36%	17.04%	9.08%	10.19%	12.91%	5.63%
	GUN	Copy 1	33.72%	9.08%	17.53%	8.25%	11.91%	13.27%	6.23%
		Copy 2	33.44%	8.91%	18.24%	8.69%	11.55%	13.36%	5.82%
		Fake Essay	34%	9.23%	17.31%	8.72%	11.66%	13.57%	5.51%
		True Essay	34.54%	8.78%	15.97%	8.99%	11.09%	14.11%	6.52%
	REVIEW	Copy 1	33.04%	9.52%	18.6%	8.49%	12.38%	11.79%	6.18%
		Copy 2	32.08%	10.16%	18.44%	9.43%	12.78%	11.04%	6.07%
		Fake Review	31.97%	10.4%	16.61%	9.91%	12.27%	13.01%	5.83%
		True Review	32.62%	10.01%	17.17%	9.97%	11.98%	12.4%	5.85%

Table 5. Distributions merit for flight times ordered by best match count, two-parameter distributions.

Dataset		Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
LSIA		Free text	48.94%	28.71%	8.33%	5.99%	5.33%	1.97%	0.73%
KM		Free text	78.69%	16.39%	3.04%	1.17%	0.47%	0.23%	NONE
		Transcription	70.48%	18.81%	6.67%	1.9%	0.71%	0.95%	0.48%
PROSODY	GAY	Copy 1	65.38%	18.14%	6.77%	4.38%	3.21%	1.71%	0.4%
		Copy 2	66.67%	18.42%	5.82%	3.93%	3.22%	1.46%	0.48%
		Fake Essay	70.68%	17.92%	4.74%	2.91%	2.34%	1.16%	0.26%
		True Essay	71.82%	16.44%	4.47%	3.34%	2.49%	1.07%	0.38%
	GUN	Copy 1	65.31%	18.79%	6.16%	4.45%	3.14%	1.75%	0.41%
		Copy 2	65.38%	18.74%	5.97%	3.97%	3.2%	2.11%	0.64%
		Fake Essay	68.43%	19.27%	5.13%	3.24%	2.62%	0.99%	0.32%
		True Essay	70.94%	18.23%	4.29%	2.97%	2.08%	1.15%	0.35%
	REVIEW	Copy 1	64.51%	18.64%	6.47%	4.88%	3.28%	1.71%	0.51%
		Copy 2	64.37%	18.38%	6.9%	4.74%	3.36%	1.78%	0.47%
		Fake Review	69.94%	17.86%	4.84%	3.38%	2.48%	0.98%	0.52%
		True Review	70%	18.75%	4.13%	3.74%	2.14%	0.96%	0.28%

3.8. Presentation of results

Results are summarized in three ways. The first set of tables, spanning from Table 4 to Table 7, was built by counting how many times each candidate distribution provides the best fit and ranking them in order of success. Tables 4 and 5 show detailed results for hold times and flight times respectively, for two-parameter distributions; both are represented, together with their confidence intervals, in Fig. 5. Similarly,

Table 6. Distributions merit for hold times ordered by best match count, three-parameter distributions.

Dataset		Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
	LSIA	Free text	19.26%	16.63%	0.33%	24.03%	21.65%	NONE	18.11%
KM		Free text	34.34%	18.69%	0.76%	17.93%	19.95%	0.25%	8.08%
		Transcription	29.49%	20.79%	1.12%	18.54%	20.22%	NONE	9.83%
PROSODY	GAY	Copy 1	31.64%	19.38%	0.89%	16.82%	16.64%	0.19%	14.45%
		Copy 2	33.88%	21.03%	0.43%	15.88%	14.24%	0.34%	14.2%
		Fake Essay	30.19%	19.88%	1.33%	17.93%	14.87%	0.4%	15.41%
		True Essay	31.82%	20.67%	1.36%	15.1%	15.32%	0.62%	15.1%
	GUN	Copy 1	34.22%	20.84%	0.69%	14.98%	14.29%	0.43%	14.55%
		Copy 2	34.19%	20.83%	0.61%	15.88%	14.15%	0.47%	13.87%
		Fake Essay	33.35%	21.06%	0.81%	15.71%	13.7%	0.43%	14.94%
		True Essay	33.36%	22.05%	1.09%	14.99%	13.59%	0.32%	14.61%
	REVIEW	Copy 1	32.75%	19.79%	1.05%	15.21%	14.33%	0.06%	16.81%
		Copy 2	31.81%	20.34%	1.06%	15.93%	14.98%	0.37%	15.51%
		Fake Review	30.23%	20.33%	0.4%	16.23%	15.87%	0.27%	16.67%
		True Review	31.4%	21.65%	1.1%	15.94%	14.84%	0.2%	14.88%

Table 7. Distributions merit for flight times ordered by best match count, three-parameter distributions.

Dataset		Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
	LSIA	Free text	29.45%	12.43%	5.02%	14.72%	28.77%	1.62%	8%
KM		Free text	26.65%	27.92%	7.36%	7.87%	10.66%	11.17%	8.38%
		Transcription	32.68%	17.18%	3.94%	9.3%	19.72%	3.38%	13.8%
PROSODY	GAY	Copy 1	42.23%	24.45%	3.52%	9.57%	10.46%	1.41%	8.35%
		Copy 2	41.19%	24.19%	3.93%	9.86%	11.56%	1.89%	7.38%
		Fake Essay	42.54%	24.72%	4.81%	8.25%	10.84%	2.36%	6.49%
		True Essay	43.7%	26.5%	4.79%	7.25%	9.35%	3.07%	5.35%
	GUN	Copy 1	43.39%	21.76%	3.14%	8.29%	12.65%	2.09%	8.68%
		Copy 2	43.04%	23.22%	3.4%	8.87%	12.27%	1.93%	7.27%
		Fake Essay	42.66%	23.67%	5%	7.72%	12.49%	2.28%	6.18%
		True Essay	42.78%	25.73%	5.68%	8.22%	9.22%	2.57%	5.79%
	REVIEW	Copy 1	42.06%	21.48%	3.65%	9.37%	12.06%	2.41%	8.97%
		Copy 2	42.9%	22.88%	3.45%	8.85%	11.87%	1.73%	8.31%
		Fake Review	43.9%	24.35%	5.68%	7.76%	10.49%	2.17%	5.64%
		True Review	44.43%	24.89%	5.03%	7.24%	9.61%	2.21%	6.59%

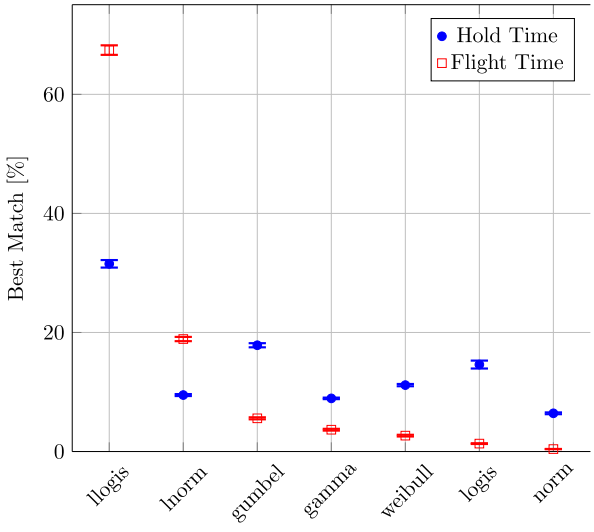


Fig. 5. Merit for two-parameter distributions. Higher is better.

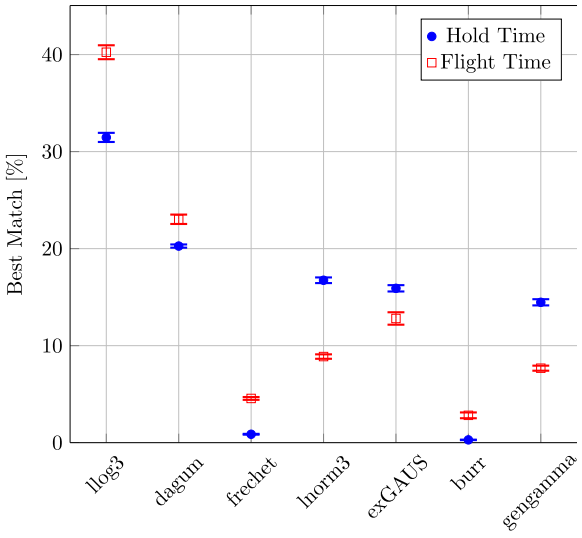


Fig. 6. Merit for three-parameter distributions. Higher is better.

Tables 6 and 7 show detailed results for hold times and flight times for three-parameter distributions, while Fig. 6 displays them graphically and with confidence intervals.

The second set of tables, spanning from Table 8 to Table 11 shows the average of one-half the AIC metric of every profile in each dataset/task pair. The layout of tables and figures, for hold times and flight

Table 8. Average log-likelihood for hold times, two-parameter distributions.

Dataset	Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
LSIA	Free text	4.393 (0.019)	4.409 (0.019)	4.416 (0.02)	4.413 (0.019)	4.478 (0.021)	4.425 (0.02)	4.454 (0.021)
KM	Free text	4.579 (0.044)	4.656 (0.039)	4.62 (0.037)	4.607 (0.04)	4.627 (0.039)	4.57 (0.04)	4.595 (0.038)
	Transcription	4.583 (0.044)	4.645 (0.04)	4.624 (0.04)	4.605 (0.041)	4.621 (0.04)	4.575 (0.04)	4.596 (0.039)
PROSODY GAY	Copy 1	4.67 (0.014)	4.713 (0.014)	4.71 (0.014)	4.724 (0.014)	4.799 (0.015)	4.729 (0.015)	4.8 (0.016)
	Copy 2	4.68 (0.014)	4.723 (0.013)	4.719 (0.014)	4.733 (0.014)	4.81 (0.016)	4.737 (0.014)	4.808 (0.015)
	Fake Essay	4.693 (0.013)	4.736 (0.013)	4.733 (0.014)	4.744 (0.014)	4.814 (0.015)	4.75 (0.014)	4.818 (0.016)
	True Essay	4.694 (0.013)	4.738 (0.013)	4.733 (0.013)	4.747 (0.013)	4.821 (0.014)	4.75 (0.013)	4.822 (0.015)
PROSODY GUN	Copy 1	4.67 (0.014)	4.717 (0.014)	4.71 (0.015)	4.722 (0.015)	4.791 (0.015)	4.721 (0.014)	4.793 (0.016)
	Copy 2	4.689 (0.015)	4.734 (0.014)	4.726 (0.015)	4.738 (0.015)	4.808 (0.015)	4.738 (0.015)	4.808 (0.016)
	Fake Essay	4.698 (0.015)	4.746 (0.014)	4.738 (0.015)	4.75 (0.015)	4.816 (0.015)	4.751 (0.015)	4.824 (0.016)
	True Essay	4.693 (0.014)	4.742 (0.013)	4.735 (0.014)	4.745 (0.014)	4.813 (0.014)	4.745 (0.014)	4.82 (0.015)
PROSODY REVIEW	Copy 1	4.699 (0.013)	4.741 (0.013)	4.735 (0.013)	4.749 (0.013)	4.817 (0.014)	4.753 (0.014)	4.82 (0.015)
	Copy 2	4.696 (0.013)	4.736 (0.013)	4.731 (0.014)	4.743 (0.013)	4.811 (0.014)	4.748 (0.014)	4.813 (0.015)
	Fake Review	4.739 (0.013)	4.782 (0.012)	4.773 (0.013)	4.784 (0.013)	4.846 (0.013)	4.786 (0.013)	4.85 (0.014)
	True Review	4.734 (0.012)	4.775 (0.012)	4.771 (0.013)	4.781 (0.012)	4.845 (0.013)	4.787 (0.013)	4.851 (0.014)

Table 9. Average log-likelihood for flight times, two-parameter distributions.

Dataset	Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
LSIA	Free text	6.395 (0.025)	6.432 (0.024)	6.459 (0.024)	6.463 (0.022)	6.524 (0.021)	6.603 (0.024)	6.714 (0.022)
KM	Free text	6.04 (0.034)	6.086 (0.033)	6.152 (0.034)	6.162 (0.032)	6.247 (0.031)	6.334 (0.036)	6.531 (0.033)
	Transcription	5.859 (0.037)	5.904 (0.037)	5.916 (0.038)	5.949 (0.037)	6.045 (0.036)	6.054 (0.04)	6.213 (0.041)
PROSODY GAY	Copy 1	5.707 (0.012)	5.748 (0.012)	5.8 (0.013)	5.813 (0.012)	5.891 (0.012)	5.952 (0.014)	6.125 (0.016)
	Copy 2	5.725 (0.011)	5.766 (0.012)	5.821 (0.013)	5.832 (0.012)	5.908 (0.012)	5.976 (0.014)	6.148 (0.015)
	Fake Essay	5.814 (0.011)	5.857 (0.011)	5.943 (0.011)	5.942 (0.012)	6.018 (0.011)	6.116 (0.013)	6.311 (0.015)
	True Essay	5.824 (0.01)	5.866 (0.01)	5.954 (0.01)	5.952 (0.011)	6.027 (0.01)	6.126 (0.012)	6.327 (0.014)
PROSODY GUN	Copy 1	5.73 (0.011)	5.772 (0.011)	5.825 (0.012)	5.835 (0.012)	5.911 (0.012)	5.977 (0.013)	6.152 (0.015)
	Copy 2	5.746 (0.011)	5.788 (0.011)	5.838 (0.012)	5.848 (0.011)	5.923 (0.012)	5.989 (0.013)	6.159 (0.015)
	Fake Essay	5.823 (0.011)	5.864 (0.011)	5.947 (0.011)	5.942 (0.012)	6.014 (0.011)	6.115 (0.013)	6.302 (0.014)
	True Essay	5.845 (0.011)	5.889 (0.011)	5.976 (0.011)	5.97 (0.011)	6.043 (0.011)	6.148 (0.012)	6.348 (0.013)
PROSODY REVIEW	Copy 1	5.706 (0.011)	5.747 (0.011)	5.798 (0.012)	5.81 (0.012)	5.887 (0.012)	5.95 (0.013)	6.116 (0.015)
	Copy 2	5.709 (0.011)	5.749 (0.011)	5.801 (0.012)	5.812 (0.011)	5.888 (0.011)	5.952 (0.013)	6.118 (0.015)
	Fake Review	5.83 (0.01)	5.873 (0.01)	5.957 (0.01)	5.954 (0.011)	6.027 (0.01)	6.128 (0.012)	6.318 (0.014)
	True Review	5.821 (0.01)	5.862 (0.01)	5.948 (0.01)	5.945 (0.011)	6.019 (0.01)	6.12 (0.012)	6.312 (0.013)

Table 10. Average log-likelihood for hold times, three-parameter distributions.

Dataset	Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
LSIA	Free text	4.373 (0.034)	4.369 (0.021)	4.478 (0.071)	4.372 (0.021)	4.4 (0.022)	4.461 (0.021)	4.43 (0.022)
KM	Free text	4.551 (0.04)	4.559 (0.04)	4.812 (0.04)	4.565 (0.071)	4.571 (0.039)	4.623 (0.04)	4.875 (0.077)
	Transcription	4.567 (0.041)	4.57 (0.042)	4.826 (0.1)	4.578 (0.075)	4.583 (0.041)	4.629 (0.043)	4.803 (0.042)
PROSODY GAY	Copy 1	4.632 (0.017)	4.64 (0.017)	4.734 (0.017)	4.652 (0.029)	4.667 (0.017)	4.725 (0.018)	4.858 (0.055)
	Copy 2	4.615 (0.018)	4.62 (0.018)	4.705 (0.018)	4.638 (0.03)	4.65 (0.018)	4.71 (0.017)	4.833 (0.058)
	Fake Essay	4.645 (0.017)	4.658 (0.017)	4.753 (0.016)	4.667 (0.03)	4.679 (0.017)	4.738 (0.017)	4.814 (0.058)
	True Essay	4.655 (0.016)	4.662 (0.016)	4.768 (0.015)	4.678 (0.026)	4.688 (0.016)	4.745 (0.016)	4.842 (0.052)
PROSODY GUN	Copy 1	4.647 (0.019)	4.669 (0.018)	4.747 (0.019)	4.674 (0.031)	4.687 (0.019)	4.748 (0.018)	4.85 (0.058)
	Copy 2	4.668 (0.02)	4.686 (0.019)	4.787 (0.019)	4.694 (0.032)	4.714 (0.02)	4.767 (0.019)	4.911 (0.055)
	Fake Essay	4.683 (0.019)	4.704 (0.018)	4.767 (0.019)	4.706 (0.033)	4.723 (0.02)	4.784 (0.018)	4.845 (0.06)
	True Essay	4.678 (0.018)	4.694 (0.017)	4.78 (0.017)	4.705 (0.03)	4.718 (0.017)	4.774 (0.017)	4.9 (0.05)
PROSODY REVIEW	Copy 1	4.661 (0.02)	4.673 (0.019)	4.744 (0.019)	4.684 (0.033)	4.701 (0.019)	4.761 (0.019)	4.831 (0.065)
	Copy 2	4.665 (0.021)	4.682 (0.019)	4.757 (0.02)	4.688 (0.032)	4.704 (0.02)	4.766 (0.019)	4.809 (0.066)
	Fake Review	4.705 (0.018)	4.716 (0.018)	4.794 (0.018)	4.728 (0.031)	4.747 (0.018)	4.799 (0.017)	4.848 (0.059)
	True Review	4.712 (0.017)	4.725 (0.016)	4.809 (0.016)	4.732 (0.025)	4.752 (0.017)	4.803 (0.016)	4.891 (0.053)

Table 11. Average log-likelihood for flight times, three-parameter distributions.

Dataset	Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
LSIA	Free text	6.311 (0.026)	6.325 (0.025)	6.313 (0.026)	6.325 (0.025)	6.324 (0.025)	6.538 (0.034)	6.566 (0.083)
KM	Free text	5.994 (0.034)	5.998 (0.034)	6 (0.034)	6.022 (0.033)	6.022 (0.033)	6.06 (0.039)	6.447 (0.147)
	Transcription	5.818 (0.039)	5.825 (0.039)	5.843 (0.038)	5.844 (0.038)	5.835 (0.038)	5.911 (0.046)	5.979 (0.203)
PROSODY GAY	Copy 1	5.637 (0.015)	5.645 (0.015)	5.67 (0.015)	5.665 (0.015)	5.684 (0.015)	5.708 (0.015)	6.1 (0.059)
	Copy 2	5.66 (0.015)	5.666 (0.014)	5.691 (0.015)	5.688 (0.014)	5.704 (0.014)	5.729 (0.015)	6.126 (0.058)
	Fake Essay	5.742 (0.013)	5.748 (0.013)	5.759 (0.013)	5.772 (0.013)	5.796 (0.013)	5.814 (0.014)	6.273 (0.054)
	True Essay	5.75 (0.012)	5.755 (0.012)	5.769 (0.012)	5.782 (0.012)	5.807 (0.012)	5.818 (0.013)	6.331 (0.045)
PROSODY GUN	Copy 1	5.655 (0.014)	5.663 (0.014)	5.686 (0.014)	5.683 (0.014)	5.702 (0.014)	5.733 (0.015)	6.07 (0.059)
	Copy 2	5.669 (0.014)	5.677 (0.014)	5.697 (0.014)	5.698 (0.014)	5.714 (0.014)	5.744 (0.015)	6.137 (0.056)
	Fake Essay	5.748 (0.014)	5.76 (0.013)	5.766 (0.014)	5.777 (0.014)	5.802 (0.013)	5.827 (0.014)	6.23 (0.054)
	True Essay	5.749 (0.012)	5.757 (0.012)	5.767 (0.012)	5.78 (0.012)	5.805 (0.012)	5.826 (0.013)	6.283 (0.047)
PROSODY REVIEW	Copy 1	5.654 (0.016)	5.662 (0.016)	5.686 (0.016)	5.682 (0.016)	5.701 (0.016)	5.719 (0.016)	6.061 (0.069)
	Copy 2	5.672 (0.015)	5.679 (0.015)	5.7 (0.015)	5.7 (0.015)	5.718 (0.015)	5.737 (0.015)	6.109 (0.063)
	Fake Review	5.786 (0.014)	5.796 (0.014)	5.807 (0.014)	5.816 (0.014)	5.843 (0.014)	5.856 (0.014)	6.342 (0.05)
	True Review	5.762 (0.013)	5.772 (0.013)	5.784 (0.013)	5.794 (0.013)	5.822 (0.013)	5.832 (0.013)	6.27 (0.053)

times, and for two-parameter and three-parameters distributions, follows the first set of tables.

Finally, the third set of tables, spanning from Table 12 to Table 15, ranks candidate distributions according to the frequency of hypothesis rejection (the lower the better). Once again, a similar layout has been followed.

The gray shading in each cell of each table is meant to convey, at a glance, the relative merit of each distribution in its row, which

Table 12. Rejection percentage for hold times, two-parameter distributions.

Dataset		Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
	LSIA	Free text	5.61%	10.09%	11.72%	27.41%	24.72%	8.88%	18.32%
KM		Free text	9.37%	31.62%	33.49%	27.71%	37.94%	11.01%	25.06%
		Transcription	5.48%	23.57%	26.67%	22.44%	30.48%	7.38%	18.81%
PROSODY	GAY	Copy 1	7.19%	14.41%	14.07%	25.13%	30.25%	12.95%	26.35%
		Copy 2	7.24%	14.85%	13.6%	26.11%	29.63%	12.41%	26.37%
		Fake Essay	6.98%	14.84%	13.9%	25.4%	28.44%	12.95%	26.01%
		True Essay	7.62%	15.58%	15.18%	26.84%	31.66%	13.64%	28.47%
	GUN	Copy 1	6.66%	15.4%	13.34%	28.06%	30.08%	10.91%	25.36%
		Copy 2	6.27%	14.99%	12.88%	25.22%	28.19%	10.59%	24.23%
		Fake Essay	6.41%	15.12%	13.26%	27.44%	28.04%	10.39%	24.26%
		True Essay	7.82%	17.27%	15.49%	29.09%	32.28%	12.55%	27.63%
	REVIEW	Copy 1	7.82%	14.02%	12.91%	26.18%	27.25%	11.22%	24.37%
		Copy 2	7.6%	14.23%	12.86%	25.45%	27.21%	11.91%	23.7%
		Fake Review	8.11%	15.32%	13.88%	27.13%	26.97%	12.27%	24.13%
		True Review	8.71%	15.9%	14.82%	27.43%	28.86%	13.02%	26.2%

Table 13. Rejection percentage for flight times, two-parameter distributions.

Dataset		Task	llogis	lnorm	gumbel	gamma	weibull	logis	norm
	LSIA	Free text	5.62%	16.22%	21.84%	31.27%	38.5%	36.52%	62.02%
KM		Free text	15.69%	43.09%	60.66%	68.77%	79.86%	77.05%	91.57%
		Transcription	4.29%	23.1%	25.71%	47.38%	64.76%	53.81%	79.05%
PROSODY	GAY	Copy 1	0.74%	8.39%	14.73%	31.85%	37.46%	28.87%	59.02%
		Copy 2	0.98%	8.38%	14.84%	31.5%	36.68%	28.37%	59.48%
		Fake Essay	1.07%	9.05%	22.02%	38.72%	42.84%	37.67%	66.81%
		True Essay	1.78%	11.89%	25.09%	42.85%	46.54%	41.75%	69.13%
	GUN	Copy 1	1.5%	9.23%	15.05%	31.77%	37.23%	28.85%	58.92%
		Copy 2	1.45%	9.26%	14.39%	30.19%	36.16%	27.78%	58.58%
		Fake Essay	1.4%	9.06%	20.43%	36.47%	39.71%	35.82%	64.56%
		True Essay	2.62%	12.33%	24.41%	42.03%	45.53%	40.59%	70.01%
	REVIEW	Copy 1	0.9%	6.86%	12.45%	28.06%	32.95%	24.58%	54.86%
		Copy 2	0.97%	6.34%	11.36%	26.81%	32.94%	23.54%	55.36%
		Fake Review	0.94%	7.19%	19.3%	35.61%	38.5%	33.6%	65.94%
		True Review	1.07%	7.85%	20.8%	37.03%	40.62%	36.28%	65.71%

Table 14. Rejection percentage for hold times, three-parameter distributions.

Dataset		Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
	LSIA	Free text	5.12%	5.14%	5.64%	18.52%	15.39%	17.12%	78.15%
KM		Free text	1.77%	4.81%	40.78%	12.12%	10.1%	24.49%	77.46%
		Transcription	0.84%	3.37%	25.58%	6.18%	7.58%	18.82%	74.05%
PROSODY	GAY	Copy 1	7.3%	8%	14.61%	15.62%	15.8%	21.95%	78.28%
		Copy 2	6.7%	7.9%	12.52%	16.23%	14.82%	22.52%	79.56%
		Fake Essay	7.05%	7.72%	13.07%	14.96%	14.12%	21.82%	77.68%
		True Essay	7.79%	8.33%	14.18%	16.64%	15.73%	23.39%	78.56%
	GUN	Copy 1	6.09%	7.74%	14.37%	18.4%	15.11%	21.95%	79.69%
		Copy 2	6.36%	7.4%	13.9%	16.68%	16.3%	21.03%	78.19%
		Fake Essay	5.79%	7.52%	13.86%	16.49%	15.5%	22%	78.55%
		True Essay	7.85%	9.19%	17.55%	19.33%	17.24%	24.96%	78.37%
	REVIEW	Copy 1	8.8%	9.65%	13.88%	19.68%	18.08%	21.68%	79.63%
		Copy 2	9.37%	10.3%	14.03%	18.8%	17.43%	22.97%	79.76%
		Fake Review	9.52%	10.54%	12.17%	18.64%	18.1%	22.43%	78.45%
		True Review	9.17%	9.62%	14.56%	19.78%	18.38%	21.88%	79.24%

corresponds to a dataset and a task. Lighter shading means better performance. For example, in Table 4, Gumbel is the distribution with the highest match count and thus depicted the lighter, while norm is the worst performing and thus depicted the darker.

Table 2 details the number of users and evaluated profiles with more than 20 and 40 samples in each task of every dataset, while summarizing the best performing distribution for each line of every table. A line of Table 2 includes the twelve best scoring distributions for the same dataset and task in tables ranging from Table 6 to 15. As llogis/llogis and llog3/llog3 are the most common entries for two and three parameters, values that deviate from the former have been emphasized to make them easier to spot.

Table 15. Rejection percentage for flight times, three-parameter distributions.

Dataset		Task	llog3	dagum	frechet	lnorm3	exGAUS	burr	gengamma
	LSIA	Free text	1.72%	2.92%	2.8%	6.81%	16.6%	46.89%	85.5%
	KM	Free text	8.38%	5.6%	10.46%	23.86%	36.55%	32.23%	91.12%
		Transcription	2.26%	2.83%	6.94%	11.27%	12.39%	31.27%	82.23%
PROSODY	GAY	Copy 1	0.52%	0.71%	1.4%	6.24%	16.94%	9.39%	87.52%
		Copy 2	0.44%	0.54%	1.59%	6.22%	16.46%	9.81%	88.27%
		Fake Essay	0.23%	0.27%	1.17%	5.71%	22.09%	11.47%	90.59%
		True Essay	0.86%	0.83%	2.2%	9.31%	24.56%	12.93%	91.46%
	GUN	Copy 1	1%	1.27%	2.4%	7.37%	18.05%	11.34%	87.43%
		Copy 2	0.71%	1.19%	1.8%	6.14%	16.23%	10.9%	87.9%
		Fake Essay	0.79%	1.01%	1.66%	7.54%	21%	11.18%	90.4%
		True Essay	1.33%	1.66%	2.75%	9.4%	23.55%	15.01%	90.71%
	REVIEW	Copy 1	0.51%	0.9%	1.31%	4.77%	16.49%	6.56%	86.63%
		Copy 2	0.49%	0.54%	0.68%	4.21%	13.92%	6.96%	86.83%
		Fake Review	0.46%	0.88%	1.13%	4.99%	19.41%	9.15%	91.34%
		True Review	0.65%	0.93%	1.24%	6.39%	20.67%	9.49%	90.03%

3.9. Akaike information criteria

A word of caution and an extended explanation is needed with respect to the second set of tables. The intention when building them was to consider that the distribution providing the best fit most of the time must not necessarily be the one that fits better on average. To arrive at this latter value a measure of fit is needed, and not only a binary best/not best answer for each profile and distribution. Starting with Akaike valuation for a model

$$AICc = 2k - 2\ln(\hat{L}) \tag{1}$$

where k is the number of parameters, here two or three, and $\hat{L}$ is the value of the maximum value of the likelihood function for the model. From the set of candidates, the distribution yielding the lowest value of AICc should be preferred. The term $2k$ is meant to penalize the introduction of more parameters than necessary.

As explained by Akaike in his seminal paper [56], the value of the mean log-likelihood is an estimator of the (negative of the) cross-entropy between the evaluated distribution f with parameter set θ and the "true" underlying distribution g. Thus, the former tends to the latter with probability one as N, the number of samples, is increased indefinitely.

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \log f(x_i|\theta) = \int g(x)\log f(x|\theta)dx \tag{2}$$

for almost all g, f and point sets. Using AICc a better estimator for practical purposes can be constructed thanks to the additional term $2k$ correcting the downward bias introduced by the number of estimated parameters and other terms that can account for small sample sizes [57]. However, because of AICc's multiplicative constant, it must be divided by two to recover the cross-entropy. Formally, the expression is

$$\lim_{N\to\infty} \frac{1}{2N} AICc = -\int g(x)\log f(x|\theta)dx \tag{3}$$

$$= \mathbb{E}_g[-\log f_\theta] \tag{4}$$

This is the value whose average for each dataset is reported in the second set of tables. It can be interpreted as the average cross entropy between the candidate distribution and the true statistical distribution generated by the typing process, thus the lower the value the better the candidate resembles the legitimate source.

Following an information theoretic interpretation these numbers, if weighted with key frequency, would represent the average information content in nats of a keystroke under each candidate hypothesis. This can be useful for compression of keystroke timing data, but this line will not be pursued further here.

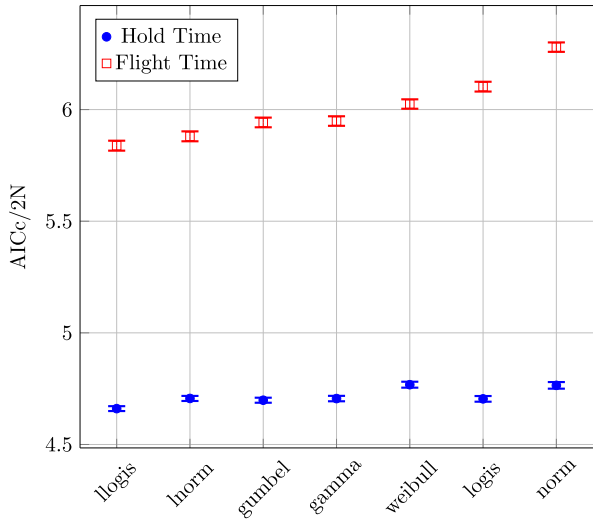


Fig. 7. Average values of AICc for two-parameter distributions. Lower is better.

4. Results and discussion

4.1. Validation of results

Several widely acknowledged observations can be used as test cases to validate the general correctness of the evaluations here presented. One of them is that hold times, being the result of a purely motor process, ought to contain less information than flight times, which not only includes decision delays but also external pauses and interruptions. The latter phenomenon is easy to observe; users rarely, if ever, respond to interruptions by holding a key pressed. The second set of tables (8 to 11) together with Figs. 7 and 8, roughly show values of AICc around 4.5 for hold times and around 6 for flight times; please refer to section 3.8 for the meaning of these numbers. Being disjoint from the former after considering the confidence intervals, it provides evidence for the stated observation, as well as an approximate measure of the information content of hold and flight times.

Another obvious phenomenon that can be used to validate the experiment is that distributions with three parameters provide a better fit than those with two. Once again, the second set of tables (8 to 11) shows the appreciation true. In Tables 12 and 13, as well as in Fig. 9, it is shown that normality is rejected very often both for hold and flight times. The latter histograms, being heavier tailed, have a higher percentage of rejections. What is more, the fact that hold time histograms

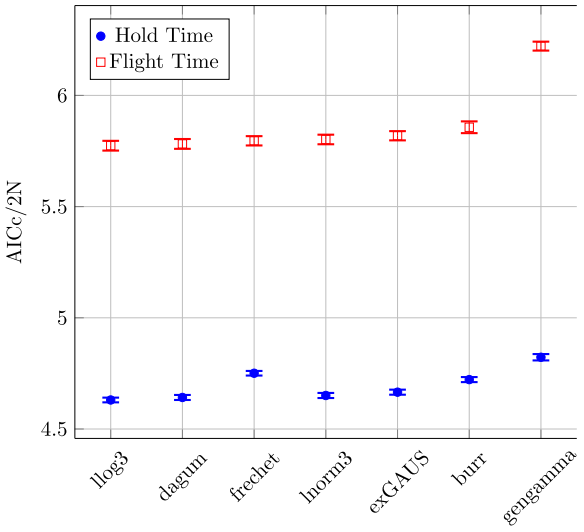


Fig. 8. Average values of AICc for three-parameter distributions. Lower is better.

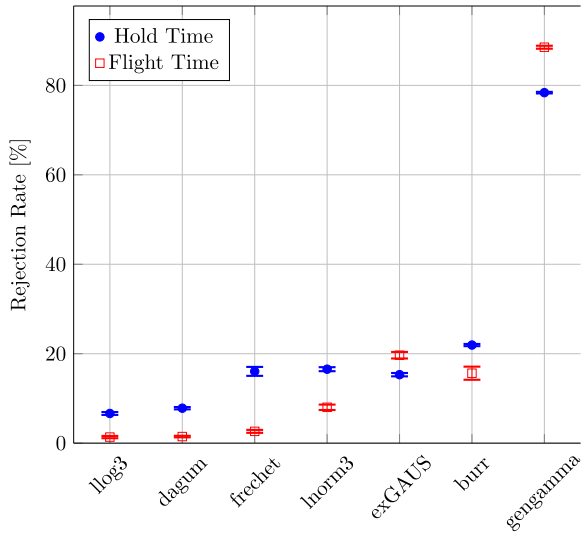


Fig. 9. Rejection rate for two-parameter distributions. Lower is better.

present more stable shapes can be seen in the fairer distribution of winner shares in their first set of tables compared with flight times.

Even though it is not a clear winner, the log-normal distribution has proved itself worthwhile in both the two- and three-parameter form. In almost every table, whenever it is not one of the top performers, it also stays far from the worst one. A curious exception appears in Table 4 for hold times of dataset KM, where even the normal distribution is chosen more often than log-normal as the best fit.

4.2. Highlights from the results

These previous observations on the shape of timing distributions in keystroke dynamics profiles, as stated in the section on previous studies, were confirmed here. However, several new and interesting facts emerge from the numerical results.

The log-logistic distribution, both in its two- and three-parameter version, is a clear winner among all candidates. Fitting flight times with two parameters, its lowest best match count for all datasets is around 50% in LSIA and above 65% in the rest, whereas the following candidate, the (two-parameter) log-normal distribution, achieves 28% in LSIA and less than 20% in the rest. For hold times it still outranks every other distribution by a margin of more than 10%, and generally around 20%, with the exception of the two KM tasks. Fitting with three parameters, the log-logistic distribution still beats all the rest by more than 10%, with a wider margin for flight times. Being consistently the least — sometimes the second to least — rejected for all datasets and almost always providing the smallest information content (followed closely by Dagum and log-normal) for both hold and flight times, the log-logistic distribution turns out to be an unexpected and unchallenged winner. To the best of our knowledge, and Google Scholar's, no previous mention of this distribution appears in keystroke dynamics research.

The best-fitting distribution does not depend much on the evaluation dataset. Whenever log-logistic, both in its two- or three-parameter version, does not achieve first place, it makes the second or third place by a negligible margin. Thus, the environmental conditions of the data capture setting do not seem to have a strong influence on the detailed shape of timing histograms.

Three-parameter distributions' merits are not that clear-cut. Information content and rejection tables do not show such a clear-cut distinction between the best scoring and the following ones, as do the tables for two-parameter distributions. Most of the time three, or even four, of them present very similar values and almost overlapping confidence intervals. Dagum's values are almost always a close call to those of the log-logistic. The average information content of log-logistic, Dagum, Frechet, and log-normal are almost identical.

The log-normal distribution is the second best choice among two-parameter candidates, even though it is slightly worse than Gumbel for hold times. It is still a good choice when using the three-parameter version, though behind Dagum and exgaussian distributions. This hardly comes as a surprise considering the attention it has received in the past. Scanning the tables it can also be seen that, together with log-logistic, its two-parameter version can compete with three-parameter distributions in information content and rejections confirming it as an excellent candidate.

The performance of the Dagum distribution comes as a surprise, as it has never been considered before in the literature about free-text keystroke dynamics. Yet, its performance is not far from log-logistic.

The exgaussian distribution is not a very good choice for modeling flight time histograms. In spite of the motivation presented in the section on past studies, the exgaussian distribution does not often provide best fit and its rejection percentage is high, especially in the PROSODY dataset. However, while the information content is relatively one of the worsts in Table 11, it is still not far from the best ones in absolute terms. External interruptions increase the noise in the histograms and the exgaussian distribution suffers its effects more than the other candidates that can absorb it seamlessly, explaining the high rejection rate and poor best match count.

Different typing tasks and topics in PROSODY dataset do not change significantly the best-fitting distribution. The three sets of tables show rather similar merit ranks for every task and topic in PROSODY dataset. Thus, broad histogram shapes do not seem to be correlated with the emotional context of the typing user except for the way in which the latter influences the parameters of the underlying distribution. This is an interesting observation because confirming it would rule out the better fitting distribution as a criterion for truth or task detection. Unluckily, being the only dataset with such a fine-grained distinction, it is impossible to know if this observation is generalizable or only applies to PROSODY dataset.

4.3. Discussion of the results

In Figs. 5 to 10, the values achieved by each candidate distribution for the three evaluation criteria have been represented together with confidence intervals. Large differences in performance for fitting the empirical histograms of keystroke timings were found when comparing them using best match count and hypothesis rejection rate, and more so when considering two-parameter distributions instead of three.

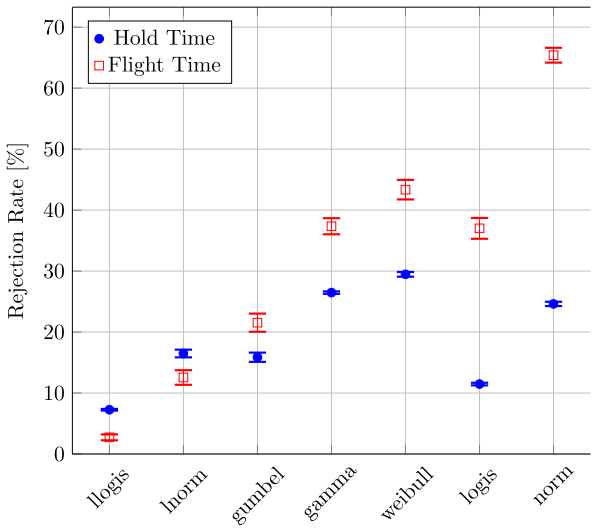


Fig. 10. Rejection rate for three-parameter distributions. Lower is better.

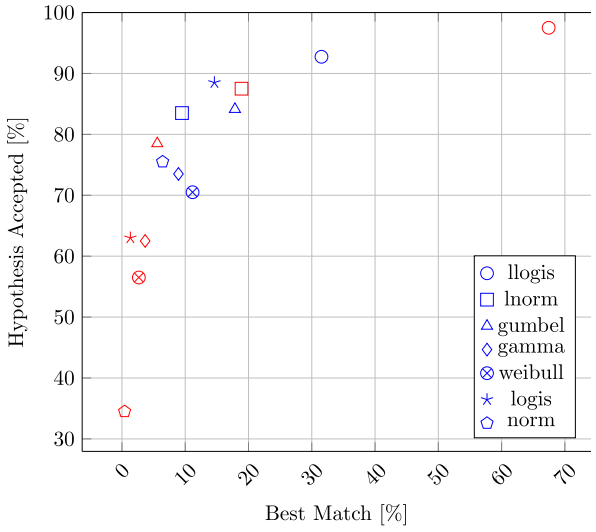


Fig. 11. Merit and hypothesis acceptance for two-parameter distributions. Hold times in blue, flight times in red.

On the other hand, the differences in AICc turned out to not be very pronounced except for the outliers norm and gengamma. Thus, to facilitate a visual comparison of the relative merits at a glance, Figs. 11 and 12 show scatter plots for the rate of hypothesis acceptance against best match count for distributions of two and three parameters; the nearer a distribution is to the top right, the better, and the nearer to the bottom left, the worse.

In the first figure a distinctive positive correlation between both performance metrics can be observed, while the pattern in the second one is much less clear. This can be expected, as an additional third parameter helps to accommodate wider variations in the shape of the empirical histograms, reducing the rate of hypothesis rejection. Yet, the log-logistic distribution in its two- and three-parameter versions manages to outclass all the other candidates. The reason why such is the case cannot be established with the existing data, and further experimentation is needed to determine the cause. The conclusion cannot be attributed to bias or an artifact of the data, as three publicly available datasets with several typing tasks, captured by different sets of non-collaborating authors in different environments, were used for the evaluation, and the log-logistic distribution achieved the best results,

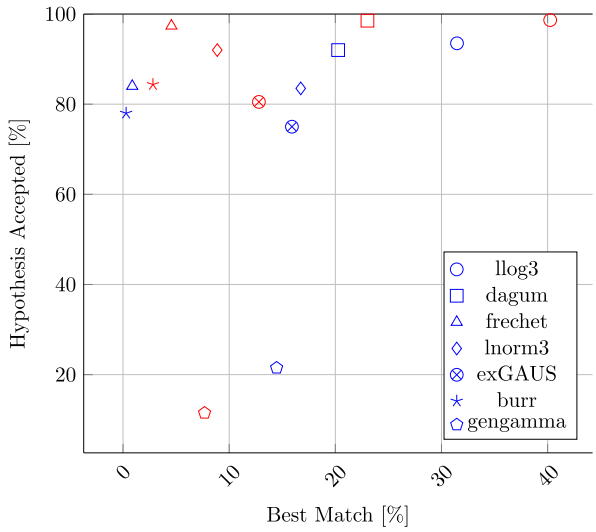


Fig. 12. Merit and hypothesis acceptance for three-parameter distributions. Hold times in blue, flight times in red.

measured with three metrics, in all but a handful of instances as Table 3 revealed.

The results have shown consistency with those of previous studies, as far as they can be compared. In particular, the log-normal distribution has proved to be a robust choice and much better for the task at hand than the normal distribution, as was first noted by [16]. Contrasting with the keystroke dynamics of passwords, where the Gumbel distribution was previously found to provide the best fit [12], it scores in third place behind the log-logistic and log-normal when evaluated with free text as was done here and as long only two parameters are allowed. Divergent results with respect to the best fitting distribution for passwords and free text confirm the initial motivation for this study.

4.4. Limitations of this study

This study has been restricted to free-text keystroke dynamics. The types of writing tasks represented in the datasets comprise composition and transcription. No claim is made about the shape of timing distributions generated by other types of writing tasks; in particular, password typing and short fixed texts were not considered. The reader interested in these cases is referred to [11, 12].

Two languages were considered in this study, English and Spanish, and no significant differences were found between them with respect to the timings distributions involved while typing. Whether these findings are generalizable to other languages or not can only be decided with further experimentation. Unluckily, the necessary datasets are lacking.

As stated in section 3.5, the criteria for selection of candidate distributions were being right tailed, with positive skew and infinite support in the positive real numbers, and bearing a general shape resemblance to the observed empirical profiles. Seven distributions with two parameters and seven more with three parameters were chosen for the comparison; it is a limitation of this study that many others were necessarily excluded, but the rather tight draw in information content for the best performing candidates, shown in Figs. 7 and 8, suggests that a significant improvement is improbable. However, the companion dataset provides the means to evaluate other distributions not evaluated here.

Candidate distributions with four or more parameters were not examined in this study. To our best knowledge, none have been previously considered in the literature of keystroke dynamics, to avoid overfitting and because of the large amount of timing observations required to estimate the parameters accurately.

4.5. Future lines of research

The main result of this paper, that the log-logistic distribution is a clear winner among all candidates for fitting keystroke dynamics timing histograms, naturally suggests two lines of inquiry. An immediate pragmatic question is whether considering the aforementioned distribution can reduce the error rates of existing algorithms beyond those achieved in [16] using the log-normal, and by how much. A tougher research question would be asking for the underlying motor and decision process that results in such a distribution for timings. Why the log-logistic and not other distributions? The candidate with the simplest explanation based on the former processes, the exgaussian, was ruled out here.

Montalvão et al. [16] have shown in the past how histogram normalization based on a log-normal distribution can improve the performance of a user verification task. Based on the results here reported, our current work on the topic focuses on evaluating whether considering the log-logistic and other well-performing distributions can provide further refinements of the error rates.

5. Conclusions

After evaluating seven distributions with two and three parameters separately against three publicly available free-text keystroke dynamics datasets, three groups of tables were produced that showed the results according to different criteria: the number of times each distribution was chosen as the best-fitting for each profile, its average information content over all profiles, and its rejection rate after an Anderson-Darling goodness of fit test.

The results confirm the established use in the research community of the log-normal distribution, in its two- and three-parameter variations, as excellent choices for modeling the shape of timings histograms in keystroke dynamics profiles. However, the log-logistic distribution emerges as a clear winner among all two- and three-parameter candidates, consistently surpassing the log-normal and all the other candidates under the three evaluation criteria for both hold and flight times. This comes as a pleasant surprise, for this distribution has not been mentioned or evaluated before in keystroke dynamics literature. So does a rather similar performance by the Dagum distribution, another newcomer to the arena. The reason why these two distributions score so well cannot be established with the existing data, and further experimentation is needed to determine the cause.

The relative merits of three-parameter distributions are not that clear-cut, as information content and rejection tables do not show significant differences between the best scoring and the following ones. Most of the time three, or even four, distributions present very similar values and almost overlapping confidence intervals. Dagum's values are almost always a close call to those of the log-logistic. The average information content of log-logistic, Dagum, Frechet, and log-normal are almost identical. Is the distinction between three–parameter distributions too fine-grained to analyze such a noisy source of data as free-text keystroke dynamics, where pauses, distractions, and intermissions pollute the timings neatly generated by the sum of a motor process and a decision process in the brain of the typing user? If that was not the case, the exgaussian would be an excellent candidate, and it was shown it is not.

Last but not least, it was shown that tasks and topics do not influence the shape of timing histograms enough to distinguish them, even though the value of their parameters can (as can be seen in [40]). This result cannot be generalized beyond dataset PROSODY, for it is the only one to contain such a distinction. But, as was stated in the Contributions section to motivate the present study, results for passwords and free text have been shown to differ, with the best candidate for the former, the Gumbel distribution, scoring third after the log-logistic and log-normal for free text.

Declarations

Author contribution statement

Nahuel González: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Enrique P. Calot: Contributed reagents, materials, analysis tools or data.

Jorge S. Ierache, Waldo Hasperué: Conceived and designed the experiments.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data associated with this study has been deposited at Mendeley Data under the accession number DOI: https://doi.org/10.17632/sjk7kz35nh.1. Data associated with this study has been deposited at IEEE DataPort under the accession number DOI: https://doi.org/10.21227/ngv9-fa18.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to thank Susan Essex for proofreading and language editing the manuscript.

References

[1] R. Stockton Gaines, William Lisowski, S. James Press, Norman Shapiro, Authentication by keystroke timing: Some preliminary results, Technical report, DTIC Document, 1980.

[2] Paulo Henrique Pisani, Ana Carolina Lorena, A systematic review on keystroke dynamics, J. Braz. Comput. Soc. 19 (4) (2013) 573–587.

[3] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, John V. Monaco Typenet, Scaling up keystroke biometrics, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2020, pp. 1–7.

[4] Ritwik Banerjee, Song Feng, Jun Seok Kang, Yejin Choi, Keystroke patterns as prosody in digital writings: a case study with deceptive reviews and essays, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1469–1473.

[5] Aythami Morales, Alejandro Acien, Julian Fierrez, John V. Monaco, Ruben Tolosana, Ruben Vera, Javier Ortega-Garcia, Keystroke biometrics in response to fake news propagation in a global pandemic, in: 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2020, pp. 1604–1609.

[6] Clgayton Epp, Michael Lippold, Regan L. Mandryk, Identifying emotional states using keystroke dynamics, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp. 715–724.

[7] Kevin S. Killourhy, Roy A. Maxion, Comparing anomaly-detection algorithms for keystroke dynamics, in: Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference, IEEE, 2009, pp. 125–134.

[8] Stefan Deian, Danfeng Yao, Keystroke-dynamics authentication against synthetic forgeries, in: 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010), IEEE, 2010, pp. 1–8.

[9] Kevin S. Killourhy, A scientific understanding of keystroke dynamics, Technical report, DTIC Document, 2012.

[10] Freeman Dyson, A meeting with Enrico Fermi, Nature (ISSN 1476-4687) 427 (6972) (Jan 2004) 297, https://doi.org/10.1038/427297a.

[11] Denis Migdal, Christophe Rosenberger, Analysis of keystroke dynamics for the generation of synthetic datasets, in: 2018 International Conference on Cyberworlds (CW), IEEE, 2018, pp. 339–344.

[12] Denis Migdal, Christophe Rosenberger, Statistical modeling of keystroke dynamics samples for the generation of synthetic datasets, Future Gener. Comput. Syst. 100 (2019) 907–920.

[13] Fabian Monrose, Aviel Rubin, Authentication via keystroke dynamics, in: Proceedings of the 4th ACM Conference on Computer and Communications Security, ACM, 1997, pp. 48–56.

[14] F. Bergadano, D. Gunetti, C. Picardi, User authentication through keystroke dynamics, ACM Trans. Inf. Syst. Secur. (TISSEC) 5 (4) (2002) 367–397.

[15] Daniele Gunetti, Claudia Picardi, Keystroke analysis of free text, ACM Trans. Inf. Syst. Secur. (TISSEC) 8 (3) (2005) 312–347.

[16] Jugurta R. Montalvão Filho, Eduardo O. Freire, On the equalization of keystroke timing histograms, Pattern Recognit. Lett. 27 (13) (2006) 1440–1446.

[17] John V. Monaco, Charles C. Tappert, The partially observable hidden Markov model with application to keystroke biometrics, arXiv preprint, arXiv:1607.03854, 2016.

[18] John V. Monaco, Md Liakat Ali, Charles C. Tappert, Spoofing key-press latencies with a generative keystroke dynamics model, in: Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference, IEEE, 2015, pp. 1–8.

[19] Albert-Laszlo Barabasi, The origin of bursts and heavy tails in human dynamics, Nature 435 (7039) (2005) 207–211.

[20] Evgeny Chukharev-Hudilainen, Pauses in spontaneous written communication: a keystroke logging study, J. Writ. Res. 6 (1) (2014) 61–84.

[21] John Vincent Monaco, Time intervals as a Behavioral Biometric, PhD thesis, PACE University, 2015.

[22] Aamo Iorliam, Anthony T.S. Ho, Norman Poh, Santosh Tirunagari, Patrick Bours, Data forensic techniques using Benford's law and Zipf's law for keystroke dynamics, in: Biometrics and Forensics (IWBF), 2015 International Workshop on, IEEE, 2015, pp. 1–6.

[23] Andrew Heathcote, Stephen J. Popiel, D.J. Mewhort, Analysis of response time distributions: an example using the Stroop task, Psychol. Bull. 109 (2) (1991) 340.

[24] Marco Steinhauser, Ronald Hübner, Distinguishing response conflict and task conflict in the Stroop task: evidence from ex-Gaussian distribution analysis, J. Exp. Psychol. Hum. Percept. Perform. 35 (5) (2009) 1398.

[25] Romain Giot, Mohamad El-Abed, Christophe Rosenberger, Web-based benchmark for keystroke dynamics biometric systems: a statistical analysis, in: 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IEEE, 2012, pp. 11–15.

[26] A.W. Harzing, Publish or perish 7, https://harzing.com/resources/publish-or-perish, 2007.

[27] Ruben Vera-Rodriguez, Ruben Tolosana, Javier Hernandez-Ortega, Alejandro Acien, Aythami Morales, Julian Fierrez, Javier Ortega-Garcia, Modeling the complexity of signature and touch-screen biometrics using the lognormality principle, in: The Lognormality Principle and Its Applications in e-Security, e-Learning and e-Health, World Scientific, 2021, pp. 65–86.

[28] Alejandro Acien, Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, Javier Hernandez-Ortega, Active detection of age groups based on touch interaction, IET Biometrics 8 (1) (2019) 101–108.

[29] Yushi Cheng, Xiaoyu Ji, Xiaopeng Li, Tianchen Zhang, Sharaf Malebary, Xianshan Qu, Wenyuan Xu, Identifying child users via touchscreen interactions, ACM Trans. Sens. Netw. (TOSN) 16 (4) (2020) 1–25.

[30] Zi Chu, Steven Gianvecchio, Haining Wang, Bot or human? A behavior-based online bot detection system, in: From Database to Cyber Security, Springer, 2018, pp. 432–449.

[31] Galit Shmueli, et al., To explain or to predict?, Stat. Sci. 25 (3) (2010) 289–310.

[32] Preeti Khanna, M. Sasikumar, Recognising emotions from keyboard stroke pattern, Int. J. Comput. Appl. 11 (9) (2010) 1–5.

[33] Ann Gledson, Dommy Asfiandy, Joseph Mellor, Thamer Omer Faraj Ba-Dhfari, Gemma Stringer, Samuel Couth, Alistair Burns, Iracema Leroi, Xiaojun Zeng, John Keane, et al., Combining mouse and keyboard events with higher level desktop actions to detect mild cognitive impairment, in: Healthcare Informatics (ICHI), 2016 IEEE International Conference, IEEE, 2016, pp. 139–145.

[34] Kevin S. Killourhy, Roy A. Maxion, Should security researchers experiment more and draw more inferences?, in: CSET, 2011.

[35] Anil Jain, Brendan Klare, Arun Ross, Guidelines for best practices in biometrics research, in: Biometrics (ICB), 2015 International Conference, IEEE, 2015, pp. 541–545.

[36] Douglas G. Altman, Sheila M. Gore, Martin J. Gardner, Stuart J. Pocock, Statistical guidelines for contributors to medical journals, Br. Med. J. (Clinical res. ed.) 286 (6376) (1983) 1489.

[37] Nahuel González, Enrique P. Calot, Jorge S. Ierache, A replication of two free text keystroke dynamics experiments under harsher conditions, in: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE, 2016, pp. 1–6.

[38] Nahuel González, Enrique P. Calot, Finite context modeling of keystroke dynamics in free text, in: Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the, IEEE, 2015, pp. 1–5.

[39] Kevin S. Killourhy, Roy A. Maxion, Free vs. transcribed text for keystroke-dynamics evaluations, in: Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results, ACM, 2012, pp. 1–8.

[40] Ritwik Banerjee, Song Feng, Jun Seok Kang, Yejin Choi, Keystroke patterns as prosody in digital writings: a case study with deceptive reviews and essays, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1469–1473, http://www.aclweb.org/anthology/D14-1155.

[41] Antonio Torralba, Alexei A. Efros, Unbiased look at dataset bias, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference, IEEE, 2011, pp. 1521–1528.

[42] Nahuel González, Timing distributions in free text keystroke dynamics profiles, https://doi.org/10.21227/ngv9-fa18, 2021.

[43] Nahuel González, Timing distributions in free text keystroke dynamics profiles, 2021b, https://doi.org/10.17632/sjk7kz35nh.1, 2021.

[44] Christian Walck, Handbook on Statistical Distributions for Experimentalists, vol. 10, University of Stockholm, 2007.

[45] Nornadiah Mohd Razali, Yap Bee Wah, et al., Power comparisons of Shapiro-wilk, Kolmogorov-Smirnov, lilliefors and Anderson-Darling tests, J. Stat. Mod. Anal. 2 (1) (2011) 21–33.

[46] C.D. Sinclair, B.D. Spurr, M.I. Ahmad, Modified Anderson darling test, Commun. Stat., Theory Methods 19 (10) (1990) 3677–3686.

[47] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.

[48] Laure Marie, Delignette-Muller and Christophe Dutang. fitdistrplus: an R package for fitting distributions, J. Stat. Softw. 64 (4) (2015) 1–34, http://www.jstatsoft.org/v64/i04/.

[49] Carlos J. Gil Bellosta, ADGofTest: Anderson-Darling GoF test, https://CRAN.R-project.org/package=ADGofTest, 2011, R package version 0.3.

[50] Christophe Dutang, Vincent Goulet, Mathieu Pigeon actuar, An R package for actuarial science, J. Stat. Softw. 25 (7) (2008) 38, http://www.jstatsoft.org/v25/i07.

[51] Paul-Christian Bürkner brms, An R package for Bayesian multilevel models using Stan, J. Stat. Softw. 80 (1) (2017) 1–28, https://doi.org/10.18637/jss.v080.i01.

[52] P. Ruckdeschel, M. Kohl, T. Stabla, F. Camphausen, S4 classes for distributions, R News 6 (2) (May 2006) 2–6.

[53] Francois Aucoin, FAdist: distributions that are sometimes used in hydrology, https://CRAN.R-project.org/package=FAdist, 2015, R package version 2.2.

[54] Mikis Stasinopoulos, Robert Rigby, Gamlss.dist: distributions for generalized additive models for location scale and shape, https://CRAN.R-project.org/package=gamlss.dist, 2018, R package version 5.1-1.

[55] Thomas Roth, qualityTools: statistics in quality science., http://www.r-qualitytools.org, 2016, R package version 1.55.

[56] Hirotugu Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.

[57] Clifford M. Hurvich, Chih-Ling Tsai, Regression and time series model selection in small samples, Biometrika (1989) 297–307.