4open 2020, **3**, 9
© J.E. García et al., Published by EDP Sciences, 2020
https://doi.org/10.1051/fopen/2020010

Available online at:
www.4open-sciences.org

RESEARCH ARTICLE OPEN ACCESS

Risk of fraud classification

Jesús Enrique García[1], Verónica Andrea González-López[1], Hugo Helito da Silva[2], and Thainá Soares Silva[1,*]

[1] Department of Statistics, University of Campinas, Sergio Buarque de Holanda, 651, CEP: 13083-859, Campinas, SP, Brazil
[2] CPFL, Rod. Eng. Miguel Noel Nascentes Burnier, 1755 – Chácara Primavera, CEP: 13088-900, Campinas, SP, Brazil

Received 17 March 2020, Accepted 19 July 2020

Abstract – In this article, we define consumers' profiles of electricity who commit fraud. We also compare these profiles with users' profiles not classified as fraudsters in order to determine which of these clients should receive an inspection. We present a statistically consistent method to classify clients/users as fraudsters or not, according to the profiles of previously identified fraudsters. We show that it is possible to use several characteristics to inspect the classification of fraud; those aspects are represented by the coding performed in the observed series of clients/users. In this way, several encodings can be used, and the client risk can be constructed to integrate complementary aspects. We show that the classification method has success rates that exceed 77%, which allows us to infer confidence in the methodology.

Keywords: Bayesian Information Criterion, Partition Markov Models, Metric in Markov Processes

Introduction

This article is oriented to the solution of a real problem through stochastic processes techniques. Institutions/companies collect information from users/customers to determine their profiles on consumption practices, preferences, and socio-economic features, among other aspects. That is, in general terms, they seek to establish behavioral profiles. This knowledge can facilitate the placement of products or the rapid adaptation of an institution to meet the needs of its users. The coding of the information allows defining these profiles, which constitute representations of the behavior. Such representations provide information to institutions and companies to form teams that can dedicate themselves to optimizing the relationship with these groups characterized by specific profiles. Those profiles are defined using the knowledge about the performance of certain sequences (user history coding). The problem of determining groups and profiles can be approached from discrete stochastic processes tools, since, in this area, there are powerful tools to deal with the problem, see [1], [2] and [3]. The sequences resulting from the coding of user/customer data can be identified as samples coming from discrete stochastic processes. In this article, we develop a method to classify sequences, according to k_I previously determined profiles. Then the k_I profiles are compared with the performance of other unclassified sequences (or group of). For this purpose, it is necessary to have some tools, (A) a tool that is capable of (i) discriminating between processes by samples from them, (ii) determining whether the processes represented by their samples are from the same stochastic law, (B) a tool that allows drawing a stochastic profile of the behavior of a process based on a series of sequences (group) that are judged to come from the same process. As a consequence of addressing this issue, in this article, we address the problem of classifying clients as fraudsters. We employ a set of real data on electricity consumption. The proposal is to attribute to each classified client, a risk related to the similarity that its series of consumption shows with some group of fraudulent clients, identified through (A)–(B).

This article is organized as follows. Section 2 addresses the theoretical foundations and classification strategy. Section 3 describes the data, the coding, and the calculation of the risk to customers. Also, in this section, the notion of fraud customers and the groups found in the database are discussed. The conclusions and considerations are given in Section 4.

Theoretical background

We begin this section by introducing the notation used in the formalization of stochastic tools. Let $(Z_t)_t$ be a discrete time Markov chain of order o ($o < \infty$) with finite alphabet A. Let us call $\mathcal{S} = A^o$ the state space and denote the string $a_m a_{m+1} \ldots a_n$ by a_m^n where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$ define the transition probability $P(a|s) = \text{Prob}\,(Z_t = a | Z_{t-o}^{t-1} = s)$. In a given sample z_1^n, coming from the stochastic process, the number of occurrences of s is denoted by $N_n(s)$ and the number of occurrences of s followed by a is denoted by $N_n(s, a)$. In this way, $\frac{N_n(s,a)}{N_n(s)}$ is the maximum likelihood estimator of $P(a|s)$. Consider

*Corresponding author: thainass@outlook.com

now, two Markov chains $(Z_{1,t})_t$ and $(Z_{2,t})_t$, of order o, disposed on the finite alphabet A with state space $\mathcal{S}$. Given $s \in \mathcal{S}$ denote by $\{P(a|s)\}_{a\in A}$ and $\{Q(a|s)\}_{a\in A}$ the sets of transition probabilities of $(Z_{1,t})_t$ and $(Z_{2,t})_t$, respectively. Consider now the local metric d_s introduced by [1], note that d_s is a metric in $\mathcal{S}$ (not negative, symmetric and follows triangular inequality) and it allows defining a global notion (in $\mathcal{S}$) of similarity between sequences.

Definition 1. Consider two Markov chains $(Z_{1,t})_t$ and $(Z_{2,t})_t$ of order o, with finite alphabet A, state space $\mathcal{S} = A^o$ and independent samples $z_{1,1}^{n_1}$, $z_{2,1}^{n_2}$ respectively. Then, set

(i) for each $s \in \mathcal{S}$, $d_s(z_{1,1}^{n_1}, z_{2,1}^{n_2}) =$

$$\frac{\alpha}{(|A|-1)\ln(n_1+n_2)} \sum_{a\in A} \left\{ \sum_{k=1,2} N_{n_k}(s,a) \ln\left(\frac{N_{n_k}(s,a)}{N_{n_k}(s)}\right) - N_{n_1+n_2}(s,a) \ln\left(\frac{N_{n_1+n_2}(s,a)}{N_{n_1+n_2}(s)}\right) \right\},$$

(ii) $dmax(z_{1,1}^{n_1}, z_{2,1}^{n_2}) = max_{s\in\mathcal{S}}\{d_s(z_{1,1}^{n_1}, z_{2,1}^{n_2})\}$,

with $N_{n_1+n_2}(s,a) = N_{n_1}(s,a) + N_{n_2}(s,a)$, $N_{n_1+n_2}(s) = N_{n_1}(s) + N_{n_2}(s)$, where N_{n_1} and N_{n_2} are given as usual, computed from the samples $z_{1,1}^{n_1}$ and $z_{2,1}^{n_2}$ respectively. Moreover, α is a real and positive value.

The Definition 1 introduces two notions of proximity between sequences, *i.* is local, *ii.* is global; both are statistically consistent, since, by increasing the $\min\{n_1, n_2\}$, grows their capacity to detect discrepancies (when the underlying laws are different) and similarities (when the underlying laws are the same). To decide if the sequences follow the same law, is only necessary to check that $d_s < 1$. This threshold is derived from the *Bayesian Information Criterion* (BIC), see [1]. In the application, we use $\alpha = 2$, with this value, we recover the usual expression of the BIC, given by [4].

The next notion (Partition Markov Model-PMM) allows postulating a parsimonious model for a Markov process, aiming at the identification of states in the state space, which have in common their transition probabilities. Through this model we build the stochastic profiles.

Definition 2. Let $(Z_t)_t$ be a discrete time Markov chain of order o on a finite alphabet A, with state space $S = A^o$,

(i) $s, r \in S$ are equivalent if $P(a|s) = P(a|r)\ \forall\ a \in A$.
(ii) $(Z_t)_t$ is a Markov chain with partition $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ if this partition is the one defined by the equivalence introduced by item *i.*

The model given by Definition 2 was introduced in reference [2] as well as the strategy for its consistent estimation that is also based on a metric defined on the state space and based on the BIC. The parameters to be estimated are (a) the partition $\mathcal{L}$, (b) the transition probabilities of each part L to any element of A, $P(\cdot|L) = \sum_{s\in S} P(\cdot|s)$. Given a sample of $(Z_t)_t$, z_1^n, according to [2] the partition is estimated by means of $d_{\mathcal{L}}$ given by Definition 3.

Definition 3. Let $(Z_t)_t$ be a Markov chain of order o, with finite alphabet A and state space $S = A^o$, z_1^n a sample of the process and let $\mathcal{L} = \{L_1, L_2, \ldots, L_{|\mathcal{L}|}\}$ be a partition of S such that for all s, $r \in L$, $P(\cdot|s) = P(\cdot|r)$. Then, set $d_{\mathcal{L}}(i,j)$ between parts L_i and L_j as

$$d_{\mathcal{L}}(i,j) = \frac{\alpha}{(|A|-1)\ln(n)} \sum_{a\in A} \left\{ \sum_{k=i,j} N_n(L_k,a) \ln\left(\frac{N_n(L_k,a)}{N_n(L_k)}\right) - N_n(L_{ij},a) \ln\left(\frac{N_n(L_{ij},a)}{N_n(L_{ij})}\right) \right\},$$

with $N_n(L) = \sum_{s\in L} N_n(s)$, $N_n(L,\ a) = \sum_{s\in L} N_n(s,\ a)$, for $a \in A$, $L \in \mathcal{L}$, $L_{ij} = L_i \cup L_j$, $N_n(L_{ij}) = N_n(L_i) + N_n(L_j)$ and $N_n(L_{ij},\ a) = N_n(L_i,\ a) + N_n(L_j,\ a)$, for $a \in A$. α a real and positive value.

The metric $d_{\mathcal{L}}$ is designed to build a structure in the state space, identifying equivalent states, it is applied for example in an initial set consisting of the entire state space S, and whenever $d_{\mathcal{L}}(i,j) < 1$ the elements L_i and L_j must be in the same part (see properties of $d_{\mathcal{L}}$ in [2]). For each part L of $\hat{\mathcal{L}}$ (estimated partition) the transition probability is estimated by $\hat{P}(a|L) = \frac{N_n(L,a)}{N_n(L)}$. Note that all equivalent states are used to estimate these probabilities, in this way, an economy is produced in the total number of probabilities to be estimated.

In the next subsection we show how integrate the tools presented here to build sequence groups (clusters) with the same stochastic law. We also explain how to define the stochastic profile of each cluster.

Clusters of sequences and partition by cluster

Given a collection of p sequences $\mathcal{C} = \{z_{i,1}^{n_i}\}_{i=1}^p$, under the assumptions of Definition 1, the notion *dmax ii* – Definition 1 is used to define clusters in $\mathcal{C}$. We introduce an algorithm that shows how this is done.

Algorithm 1

- Input $\mathcal{C} = \{z_{i,1}^{n_i}\}_{i=1}^p$
  1. $M = \mathcal{C}$
  2. $M = \{m_1, \ldots, m_{|M|}\}$,
  3. $(i*, j*) = \text{argmin}\{d\text{max}(m_i, m_j),\ i \neq j,$ $i, j \in \{1, 2, \ldots, |M|\}\}$
     * if $d\max(m_{i_*}, m_{j_*}) < 1$, $M = \{\{Mm_{i_*}\}\}m_{j_*}\}\}\cup m_{i_*j_*}$ with $m_{i_*j_*} = \{m_{i_*}, m_{j_*}\}$ and go back to 2,
     * otherwise the procedure ends.
- Output (clusters of $\mathcal{C}$) $M = \{C_1, \ldots, C_k\}$

That is, the initial M is composed by all the separate sequences and the final M corresponds to the groups of sequences or clusters. Note that given two different sequences $z_{i_*,1}^{n_{i_*}}, z_{j_*,1}^{n_{j_*}} \in \mathcal{C}$ the occurrence of each $s \in \mathcal{S}$ is recorded by $N_{n_{i_*}}(s)$ and $N_{n_{j_*}}(s)$ respectively, and the occurrence of s followed by $a \in A$ is computed by $N_{n_{i_*}}(s,a)$ and $N_{n_{j_*}}(s,a)$. Already, when defining the new unit

$m_{i_*j_*} := \left\{z_{i_*,1}^{n_{i_*}}, z_{j_*,1}^{n_{j_*}}\right\}$ (after verifying $dmax(z_{i_*,1}^{n_{i_*}}, z_{j_*,1}^{n_{j_*}}) < 1$) the count of the ocurrences of s, is given by $N_{n_{i_*}}(s) + N_{n_{j_*}}(s)$ and if $a \in A$, the ocurrences of s followed by a is $N_{n_{i_*}}(s,a) + N_{n_{j_*}}(s,a)$. That is, in the case of $m_{i_*j_*}$ both sequences, $z_{i_*,1}^{n_{i_*}}$ and $z_{j_*,1}^{n_{j_*}}$, contribute to the count attributed to $m_{i_*j_*}$.

Once the proximity between sequences is determined in order to build the clusters $\{C_1, \ldots, C_k\}$ and for each cluster we can build a PMM, representing the cluster. In addition, it is possible to quantify the dissimilarity between clusters using the notion $d$max. Suppose the cluster i is C_i and it is composed by m_i independent sequences,

$$C_i = \left\{z_{i_1,1}^{n_{i_1}}, z_{i_2,1}^{n_{i_2}}, \ldots, z_{i_{m_i},1}^{n_{i_{m_i}}}\right\} = \left\{z_{i_m,1}^{n_{i_m}}\right\}_{m=1}^{m_i},$$

the sample size related to C_i is $\sum_{m=1}^{m_i} n_{i_m}$. For each $s \in \mathcal{S}$, compute the ocurrences of s in C_i as

$$N(C_i, s) = \sum_{m=1}^{m_i} N_{n_{i_m}}(s) \tag{1}$$

and the ocurrences of s followed by a as

$$N(C_i, (s,a)) = \sum_{m=1}^{m_i} N_{n_{i_m}}(s,a), \tag{2}$$

for $N_{n_{i_m}}(\cdot)$ related to the sample $z_{i_m,1}^{n_{i_m}}$.

The following remark shows how to determine the stochastic profile of each cluster.

Remark 1. If we replace in Definition 3 the sample size n by $\sum_{m=1}^{m_i} n_{i_m}$ and applying the Algorithm 1, substituting (i) $\mathcal{C} = \{z_{i,1}^{n_i}\}_{i=1}^{p}$ by $S = A^{\circ}$, (ii) $dmax$ by $d_{\mathcal{L}}$, the Output of the algorithm will be the partition $\hat{\mathcal{L}}_i$ of S, related to the cluster C_i.

The following remark shows how to measure the similarity between two clusters.

Remark 2. To establish the dissimilarity between the clusters C_1 and C_2 (since by construccion those are different) we use the Definition 1 – ii. In the calculation of i-def 2.1, we replace $N_{n_k}(s)(N_{n_k}(s,a))$ by equation (1) ((2)), with $i = k$. We replace also $N_{n_1+n_2}(s)$ by $N(C_1, s) + N(C_2, s)$ and $N_{n_1+n_2}(s,a)$ by $N(C_1, (s, a)) + N(C_2, (s, a))$. Using those ocurrences we can compute the dissimilarity between the clusters.

The next section is intended to apply the concepts detailed here as well as the strategies presented to real data.

Risk through Discretized Information

Data and structure of the analises

In Table 1, we describe the data inspected in this paper. The data correspond to serial records of energy consumption of clients of a company of power supply (CPFL) during the period: January 2011 to June 2019. We have two types of records, *Irregular* classified by specialists in fraud and *Other* which are not be classified as *Irregular*. That is, irregular cases have already been classified since they were identified by the fraud detection system of the company. The other cases appear to be normal but could have been disregarded by the system used in fraud detection.

Table 1. Biphasic clients reported by CPFL, period: January, 2011 to June, 2019.

Type	Total of clients
Other	7828
Irregular	553

Figure 1. Scheme of the organization of the ordered values $\{a_v(i)\}_{i=1}^{7828}$, on the left those that indicate greater risk, on the right those of lower risk. Cut = 1 indicates the threshold given by the BIC.

The monthly energy consumption sequence of each client i, $x_{i,1}^{q_i}$ is discretized in order to identify it with a sample of a Markov stochastic process $(Z_{i,t})_t$, of finite order o in the discrete and finite alphabet A, for $i = 1, \ldots, 8381$. The first inspection to be carried out seeks to identify clusters in *Irregular* clients, this classification could point to specific fraud practices. So we determine $\{I_1, \ldots, I_{k_I}\}$ clusters of irregular clients, applying the Algorithm 1 in the set of irregular clients. For the group *Other*, we also determine the clusters, say $\{O_1, \ldots, O_{k_O}\}$ (by applying Algorithm 1 in *Other*). So, we can classify customers into consumer practices. Once the *Irregular* clusters have been constructed, it is possible to quantify the dissimilarity between them, this is done by means of $d$max as described in the previous section (see Remark 2), computing $d\max(I_i, I_j)$, $i \neq j$, $i, j \in \{1, 2, \ldots, k_I\}$. In a second instance, we compare the behavior of the O_l, $l = 1, \ldots, k_O$ clusters with the irregular ones, computing $d\max(I_i, O_l)$, $i \in \{1, \ldots, k_I\}$, $l \in \{1, \ldots, k_O\}$. We do this comparison in order to identify which could be considered as indistinguishable from some irregular cluster, this happens when $d\max(I_i, O_l) < 1$. Such a comparison generates a risk index in the class $\{O_1, \ldots, O_{k_O}\}$, as to guide the inspection of the company in that class. For each client $v \in O_{i_v}$ we define:

$$a_v = \min_{1 \leq j \leq k_I} \{d\max(O_{i_v}, I_j)\} \tag{3}$$

In this way, the values obtained from equation (3) reported for the clients in the class *Other* (Tab. 1) are $\{a_v\}_{v=1}^{7828}$. By construction for the client v in *Other*, $\exists!\ i_v \in \{1, \ldots, k_O\}$ such that $v \in O_{i_v}$ allowing the good definition of equation (3). Denote by $a_v(1)$, $a_v(2)$, ..., $a_v(7828)$ the ordered values in an increasing way. Thus, the client that receives the value $a_v(1)$ is the one with the highest risk and the one that receives the value $a_v(7828)$ is the client with the lowest risk, taking into account that the threshold equal to 1 allows us to pay attention only in the clients whose values fall in $[0, 1)$. Figure 1 illustrates the situation.

Table 2. Description of the $k_I = 12$ irregular clusters I_i, $i = 1, \ldots, 12$. n_{I_i} : clients in I_i, d_i^* : maximal $d$max attained by I_i, values reported from left to right in increasing order. Using the process $(Z_{i,t})_t$ – see equation (4).

i	1	2	3	4	5	6	7	8	9	10	11	12
n_{I_i}	5	30	6	29	19	18	118	12	44	121	76	75
d_i^*	0.000	0.478	0.484	0.516	0.666	0.793	0.828	0.887	0.893	0.929	0.948	0.965

Table 3. $d$max between the irregular clusters $\{I_1, \ldots, I_{k_I}\}$ (see Eq. (4) and Remark 2). In bold type, the three highest values.

	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
I_1	1.355	3.557	2.294	2.581	3.409	3.141	2.047	2.739	2.045	1.859	3.063
I_2		2.024	1.397	1.362	2.393	**4.181**	1.644	3.731	1.429	2.812	3.368
I_3			1.557	2.006	1.387	1.186	1.522	1.715	1.170	1.566	1.485
I_4				2.584	2.176	1.609	1.085	**5.580**	2.336	2.985	2.859
I_5					1.511	1.863	2.332	1.632	1.176	2.295	2.623
I_6						1.164	1.599	2.260	1.238	1.563	2.006
I_7							1.442	3.525	1.901	3.324	3.726
I_8								3.509	1.638	1.696	2.357
I_9									2.092	2.299	3.579
I_{10}										2.287	2.701
I_{11}											**3.773**

Table 4. States $s \in \mathcal{S}$ and consumption behavior, coding $Z_{i,t}$ – see equation (4).

State	Event	Trajectory ending X_{t-1} and X_t
11	$X_t < X_{t-1} < \min\{X_{t-2}, X_{t-3}\}$	Decreasing
13	$X_{t-1} \leq X_t < X_{t-2}$ & $X_{t-1} < X_{t-3}$	Increasing/maintenance
14	$X_{t-1} < X_{t-2} \leq X_t$ & $X_{t-1} < X_{t-3}$	Increasing
21	$X_t < X_{t-1} < X_{t-2}$ & $X_{t-3} \leq X_{t-1}$	Decreasing
23	$X_{t-3} \leq X_{t-1} \leq X_t < X_{t-2}$	Increasing/maintenance
24	$X_{t-3} \leq X_{t-1} < X_{t-2} \leq X_t$	Increasing
31	$X_t < X_{t-2} \leq X_{t-1} \leq X_{t-3}$	Decreasing
32	$X_{t-2} \leq X_t < X_{t-1} < X_{t-3}$	Decreasing
34	$X_{t-2} \leq X_{t-1} \leq X_t$ & $X_{t-1} < X_{t-3}$	Increasing/maintenance
41	$X_t < X_{t-2} \leq X_{t-1}$ & $X_{t-3} \leq X_{t-1}$	Decreasing
42	$X_{t-2} \leq X_t < X_{t-1}$ & $X_{t-3} \leq X_{t-1}$	Decreasing
44	$X_t \geq X_{t-1} \geq \max\{X_{t-2}, X_{t-3}\}$	Increasing/maintenance

Results

We compare the series of consumption through a discretization that considers four possible states, and reports the performance of the series in relation to the magnitude of the consumption in the last measurement (at time t) when compared with the two previous measurements (times $t-1$ and $t-2$). For each client i with $x_{i,1}^{q_i}$ consumption series define the sample $z_{i,1}^{n_i}$ of the discrete process $(Z_{i,t})_t$ as

$$Z_{i,t} = \begin{cases} 1, & x_{i,t} < x_{i,t-1}, x_{i,t} < x_{i,t-2} \\ 2, & x_{i,t} < x_{i,t-1}, x_{i,t} \geq x_{i,t-2} \\ 3, & x_{i,t} \geq x_{i,t-1}, x_{i,t} < x_{i,t-2} \\ 4, & x_{i,t} \geq x_{i,t-1}, x_{i,t} \geq x_{i,t-2}. \end{cases} \quad (4)$$

Then, $A = \{1, 2, 3, 4\}$ and $|A| = 4$. In the set of sequences, the smaller one has a sample size equal to 39, the indicated order o follows the rule: $o < \log_4(39) = 2.643$, then $o = 2$. The application of the Algorithm 1 in the *Irregular* group generates $k_I = 12$ clusters. Table 2 exposes the maximum value of $d$max reported by the algorithm, inside each cluster, denoted by d_i^*. These results measure the homogeneity within each irregular cluster. Lower values of d_i^* indicate greater homogeneity, and values of d_i^* close to 1 indicate greater heterogeneity.

Table 5. Impossible states, coding $Z_{i,t}$ – see equation (4).

State	Impossible event
12	$\{X_{t-2} \leq X_t < X_{t-2}\}$
22	$\{X_{t-2} \leq X_t < X_{t-1} < X_{t-2}\}$
33	$\{X_t < X_{t-2} \leq X_{t-1} \leq X_t\}$
43	$\{X_{t-2} \leq X_{t-1} \leq X_t < X_{t-2}\}$

After identifying 12 *Irregular* clusters, we can explore the dissimilarity between them computing the values of $d$max between the clusters (see Remark 2). Table 3 shows the results.

In a stochastic way, the table quantifies the differences in fraud practices classified by the company. With the purpose of exploring the dynamics of each irregular cluster, we fit a PMM model for each cluster. This leads us to describe the meaning of each possible state for the $(Z_{i,t})_t$ process. In Table 4, we report the relation of the states $s \in \mathcal{S}$ with the consumption. Each possible state is composed of the concatenation of a and b in A, so the state is ab. By construction, the states relate the magnitudes of the energy consumption at times $t-3$, $t-2$, $t-1$ and t, thus, for example, the state $ab = 13$ means $\{X_t \geq X_{t-1}$ & $X_t \geq X_{t-2}\}$ (associated with $b = 3$) and $\{X_{t-1} < X_{t-2}$ & $X_{t-1} < X_{t-3}\}$ (associated with $a = 1$). Note that some combinations are not allowed by construction, those are 12, 22, 33, 43 (see Tab. 5).

Table 6. PMM for the irregular clusters I_i, $i = 1, \ldots, 6$ see equation (4) and Remark 1. In bold type the highest probability for the cluster.

	I_1					I_2			
Part	1	2	3	4	Part	1	2	3	4
11 41	**0.615**	0.000	0.385	0.000	11	**0.521**	0.000	0.338	0.140
13	0.545	0.091	0.000	0.364	13 23	0.323	0.186	0.000	0.492
14 24 44	0.176	0.265	0.000	0.559	14 34 24	0.245	0.358	0.000	0.397
31 42 34	0.286	0.000	0.143	0.571	21 31 41	0.380	0.000	0.342	0.278
					3242	0.403	0.000	0.140	0.457
					44	0.420	0.283	0.000	0.297
	I_3					I_4			
Part	1	2	3	4	Part	1	2	3	4
11	0.212	0.000	0.300	0.488	11 42 41	**0.442**	0.000	0.283	0.276
13 44 14	0.243	0.181	0.000	0.576	21 31 32				
21 23	**0.810**	0.000	0.127	0.063	13 23 44	0.398	0.162	0.000	0.439
24 34	0.352	0.521	0.000	0.127	34				
31 41 42	0.433	0.000	0.369	0.197	14 24	0.134	0.429	0.000	0.437
	I_5					I_6			
Part	1	2	3	4	Part	1	2	3	4
11 21 32	0.410	0.000	0.276	0.314	11 21 32	0.457	0.000	0.164	0.379
13	0.369	0.084	0.000	**0.547**	31 42				
14 23 24	0.240	0.319	0.000	0.441	13 44 14	0.201	0.276	0.000	**0.522**
44 34					23				
31 41	0.508	0.000	0.364	0.128	24 34	0.451	0.234	0.000	0.315
42	0.489	0.000	0.060	0.451	41	0.388	0.000	0.454	0.158

Note that the states described in Table 4 indicate a decreasing/increasing trajectory ending, and this behavior is reflected in the partitions generated for each irregular cluster (see Tabs. 6 and 7), with only two possible exceptions, for states that allow consumption maintenance. Consider two large groups: those of increasing final trajectories (a) 13, 14, 23, 24, 34, 44 (including increasing/maintenance) and those of decreasing final trajectories (b) 11, 21, 31, 32, 41, 42. We see that all models (except in two situations) have separated the states into those two large classes. That is, in each part of each model we only find states of one type. For example, let's take I_{11}, it is composed by 5 parts, 2 parts composed by states with a decreasing trajectory ending: $L_1 = \{11, 31, 42, 21, 32\}$ and $L_5 = \{41\}$ and, 3 parts composed by states with increasing trajectory ending: $L_2 = \{13, 23\}$, $L_3 = \{14, 24\}$, $L_4 = \{34, 44\}$. The exceptions are for I_1, the part $L_4 = \{31, 42, 34\}$, where 31 and 42 have decreasing endings and 34 allows increasing ending, and for I_3 the part $L_3 = \{21, 23\}$ where 21 has decreasing ending and 23 allows increasing ending.

From the magnitudes of the estimated probabilities (Tabs. 6 and 7) we see that the clusters show two preferences (in bold), for state 1, cases I_1, I_2, I_3, I_4, I_7, I_9, I_{12} and for state 4 the remaining cases, I_5, I_6, I_8, I_{10}, I_{11}. State 1 indicates decrease in consumption at time t in relation to the other previous instances $t-1$ and $t-2$ and 4 indicates increase/maintenance of consumption at time t in relation to the other previous instances $t-1$ and $t-2$. Moreover, for all cases I_i, $i = 1, \ldots, 12$ the first two elections (the two highest probabilities) fall in states 1 or 4. Note that when the preference is the state 1, the past states (elements of the parts) end in 1 or 2, that is to say, that according to the classification given in Table 4, the process was already in a decreasing final trajectory (except I_3). When the preference is state 4, the past states (elements in the parts) end in 3 or 4, that is, according to the classification (Tab. 4), the process was in maintenance or increasing trajectory.

The group *Other* is divided by the Algorithm 1 (coding $Z_{i,t}$ – equation (4)) in 391 clusters, so $k_O = 391$. As the purpose of this paper is to identify those customers in the *Other* category that resemble an irregular cluster, we proceed to measure this similarity. For each I_i we calculate $d$max between such irregular cluster and the clusters O_j, $j = 1, \ldots, k_O$. Table 8 summarizes the obtained values.

In Table 9, we report which O_j clusters behave as irregular. The lower the value of $d$max on the right, the higher the risk of group O_j as it becomes indistinguishable from an irregular.

We note that there are 63 clients that deserve a detailed inspection, since their risks are pronounced ($d$max values below 0.7). And certainly, the priority is for the 36 with approximately zero $d$max.

As set forth in Table 4, *Irregular* processes define their minimum units, parts of the partitions, according to the types of final trajectories (a) increasing/maintenance final trajectories and (b) decreasing final trajectories, which leads us to inspect the consumption series via a representation showing that trend. The following subsection is intended for this purpose.

Table 7. PMM for the irregular clusters I_i, $i = 7, \ldots, 12$ see equation (4) and Remark 1. In bold type the highest probability for the cluster.

	I_7					I_8			
Part	1	2	3	4	Part	1	2	3	4
11	0.357	0.000	0.245	0.398	11 41	0.548	0.000	0.250	0.202
13 23 44	0.332	0.233	0.000	0.435	13 24	0.150	0.121	0.000	**0.729**
14	0.200	0.425	0.000	0.375	14	0.273	0.438	0.000	0.289
21 42	**0.514**	0.000	0.237	0.249	21 31 42	0.595	0.000	0.095	0.311
24 34	0.255	0.302	0.000	0.444	34 44	0.397	0.108	0.000	0.495
3141	0.391	0.000	0.398	0.211					
32	0.423	0.000	0.115	0.462					
	I_9					I_{10}			
Part	1	2	3	4	Part	1	2	3	4
11 32 21	0.527	0.000	0.164	0.309	11 31 21	0.458	0.000	0.275	0.266
31 42					13 34 44	0.289	0.268	0.000	0.443
13 34 24	0.176	0.305	0.000	0.520	24				
14 44					14	0.178	0.386	0.000	0.436
23	0.459	0.059	0.000	0.482	23	0.387	0.127	0.000	**0.486**
41	**0.564**	0.000	0.392	0.044	32 42	0.434	0.000	0.183	0.382
					41	0.439	0.000	0.362	0.198
	I_{11}					I_{12}			
Part	1	2	3	4	Part	1	2	3	4
11 31 42	0.496	0.000	0.253	0.251	11 21 32	**0.523**	0.000	0.167	0.311
21 32					13 23 34	0.261	0.244	0.000	0.495
13 23	0.368	0.117	0.000	0.515	14 24	0.211	0.373	0.000	0.416
14 24	0.194	0.406	0.000	0.400	31 42	0.497	0.000	0.271	0.232
34 44	0.228	0.235	0.000	**0.537**	41	0.391	0.000	0.390	0.219
41	0.516	0.000	0.336	0.148	44	0.332	0.336	0.000	0.332

Table 8. By line (for $i = 1, \ldots, k_I$) for each cluster I_i is reported the minimum, median and maximum value of $d$max computed between I_i and each group O_j, $j = 1, \ldots, k_O$, see equation (4) and Remark 2. With * we indicate when similarity was detected, for some O_j.

Cluster	Min	Median	Max
I_1	0.000*	2.865	7.470
I_2	1.024	3.230	9.780
I_3	0.673*	2.425	7.865
I_4	0.799*	2.946	10.713
I_5	0.837*	2.738	9.878
I_6	1.147	2.815	11.286
I_7	0.992*	3.008	18.388
I_8	0.917*	2.647	8.206
I_9	1.163	3.324	13.183
I_{10}	1.067	2.895	12.370
I_{11}	1.042	3.128	10.593
I_{12}	1.063	3.036	16.403

Increasing and decreasing movements

Based on the findings we introduce a complementary coding that allows us another perspective of the study. For that, we consider two movements in the consumption series. For each client i with series $x_{i,1}^{q_i}$, define the sample $y_{i,1}^{n_i}$ of the discrete process $(Y_{i,t})_t$ as

$$Y_{i,t} = \begin{cases} 0, & x_{i,t} < x_{i,t-1} \\ 1, & x_{i,t} \geq x_{i,t-1}. \end{cases} \quad (5)$$

Table 9. For I_i, $i = 1, 3, 4, 5, 7, 8$ are reported the clusters O_j with $d$max < 1, see equation (4) and Remark 2. With * we indicate the highest risk cases.

Cluster I	Cluster O	Clients in O	$d$max
I_1	O_{75}	36	0.000*
	O_{116}	9	0.960
I_3	$O127$	6	0.885
	$O167$	27	0.673*
	O_{242}	11	0.853
	O_{248}	16	0.848
	O_{297}	24	0.727*
	O_{356}	49	0.952
I_4	O_{72}	7	0.841
	O_{217}	11	0.799
	O_{285}	22	0.995
	O_{380}	60	0.958
I_5	O_{243}	13	0.837
	O_{261}	19	0.926
I_7	O_{217}	11	0.992
I_8	O_{270}	10	0.917

Table 10. States $s \in \mathcal{S}$ and consumption behavior, codification $Y_{i,t}$ – see equation (5).

State	Event
00	$X_t < X_{t-1} < X_{t-2}$
01	$X_t \geq X_{t-1}$ & $X_{t-1} < X_{t-2}$
10	$X_t < X_{t-1}$ & $X_{t-1} \geq X_{t-2}$
11	$X_t \geq X_{t-1} \geq X_{t-2}$

Table 11. Description of the $k_I = 22$ irregular clusters I_i^{0-1}, $i = 1, \ldots, 22$. $n_{I_i^{0-1}}$: clients in I_i^{0-1}, d_i^* : maximal $d$max attained by I_i^{0-1}, values reported from left to right in increasing order. Using the process $(Y_{i,t})_t$ – see equation (5).

i	1	2	3	4	5	6	7	8	9	10	11
$n_{I_i^{0-1}}$	4	19	12	7	8	19	21	14	16	31	15
d_i^*	0.001	0.175	0.196	0.241	0.247	0.300	0.368	0.370	0.408	0.412	0.443
i	12	13	14	15	16	17	18	19	20	21	22
$n_{I_i^{0-1}}$	30	9	34	18	48	48	12	42	68	35	43
d_i^*	0.451	0.587	0.598	0.639	0.705	0.717	0.721	0.814	0.863	0.893	0.987

Then, $A = \{0, 1\}$ and $|A| = 2$. We adopt the memory $o = 2$ in order to facilitate the interpretation in concordance with the previous inspection. See the meaning of the states of the process $(Y_{i,t})_t$ in Table 10.

In Table 11 we show the $k_I = 22$ clusters defined by the Algorithm 1 in the *Irregular* class (Tab. 1), I_i^{0-1}, $i = 1, \ldots, 22$ clusters derived from the codification $Y_{i,t}$ – see equation (5).

We see that in relation to the irregular clusters via the $Z_{i,t}$ encoding, the $Y_{i,t}$ encoding almost doubles the irregularity modalities. While the Table 2 reports only 3 in 12 (25%) cases with $d^* < 0.5$, Table 11 reports 12 in 22 (55%) cases with $d^* < 0.5$, which explains the increase in the number of groups reported in Table 11. In the Appendix, Tables 16 and 17, we report the PMM for each *Irregular* cluster derived from the codification $(Y_{i,t})_t$. We report 14 models with only two parts, 7 with 3 parts and 1 model with 4 parts. States 00 and 11, which reiterate a trend of consecutive decrease in energy consumption or consecutive increase/maintenance are found in separate parts in all models except in four cases: I_i^{0-1}, $i = 11, 12, 13, 20$.

The group *Other*, under the representation $Y_{i,t}$ equation (5) is divided by Algorithm 1 in 128 clusters, $O_1^{0-1}, \ldots, O_{k_O}^{0-1}$ with $k_O = 128$. That is, we obtain an increase the profiles of the *Irregular* clusters, from 12 to 22, and a decrease of the clusters in the class *Other*, from 391 to 128.

For each I_i^{0-1} we calculate the $d$max between such irregular cluster and the clusters O_j^{0-1}, $j = 1, \ldots, k_O$, Table 12 shows the results. According to coding 0−1, the only cluster I_i^{0-1} that is not associated with any element of the class *Other* (see Tab. 1) is the I_{20}^{0-1} which has 68 clients. We must not lose sight of the fact that the risk increases when obtaining values of $d$max close to zero, and only those cases need to be identified. All criteria are asymptotic so, they should be considered with caution, that is to say that, cases with $d$max near to the threshold 1 can wrongly point cases that are regular one.

As we can see, from the results reported in Table 13, we see that in relation to the meaning given by the $Y_{i,t}$ coding (see Tab. 10), the total number of cases in the class *Other* that can be identified with irregular clusters, increases considerably. These results could indicate the relevance of the memory of the process, since we see that coding 1–4 reaches a greater past in comparison with coding 0−1 (compare Tabs. 4 and 10), being able to separate in a more realistic way the class *Other* of the class *Irregular*.

As the discretization caused by equations (4) and (5) lead us to simplifications of the original information, we proceed to consider both for the classification of clients.

Table 12. By line (for $i = 1, \ldots, k_I$) for each cluster I_i^{0-1} is reported the minimum, median and maximum value of $d$max computed between I_i^{0-1} and each group O_j^{0-1}, $j = 1, \ldots, k_O$, see equation (5) and Remark 2. With * we indicate when similarity was detected, for some O_j^{0-1}.

Cluster	Min	Median	Max
I_1^{0-1}	0.000*	5.528	29.342
I_2^{0-1}	0.224*	5.881	20.855
I_3^{0-1}	0.132*	5.646	26.392
I_4^{0-1}	0.238*	5.508	14.693
I_5^{0-1}	0.052*	4.670	17.643
I_6^{0-1}	0.274*	5.221	24.873
I_7^{0-1}	0.353*	8.258	32.664
I_8^{0-1}	0.157*	6.450	33.795
I_9^{0-1}	0.174*	6.887	29.933
I_{10}^{0-1}	0.062*	6.772	34.111
I_{11}^{0-1}	0.228*	5.817	33.809
I_{12}^{0-1}	0.213*	7.179	35.029
I_{13}^{0-1}	0.123*	6.060	23.934
I_{14}^{0-1}	0.474*	7.484	35.978
I_{15}^{0-1}	0.807*	6.319	30.182
I_{16}^{0-1}	0.381*	8.999	36.916
I_{17}^{0-1}	0.088*	10.210	44.606
I_{18}^{0-1}	0.191*	8.152	35.757
I_{19}^{0-1}	0.580*	9.758	46.892
I_{20}^{0-1}	1.109	10.858	34.376
I_{21}^{0-1}	0.599*	7.012	36.268
I_{22}^{0-1}	0.151*	8.214	37.610

In the next subsection we take both codifications into account and propose a strategy for the inspection of potentially fraudulent customers.

Risk of clients from two codifications

It is always wise to consider that the representations given by the $(Z_{i,t})_t$ and $(Y_{i,t})_t$ processes (see Eqs. (4) and (5)) only capture certain aspects of the original consumption series. As those reveal complementary information, in this subsection we consider both to guide the decision-making process in the search for undetected frauds. We introduce a function that allows a risk classification integrating both codifications, generated by equation (3). For each client $\upsilon \in O_{i_\upsilon} \cap O_{k_\upsilon}^{0-1}$ we compute:

Table 13. For I_i^{0-1}, $i \neq 20$, $i = 1, \ldots, 22$ are reported the clusters O_j^{0-1} with $d\max < 1$ – see equation (5) and Remark 2.

Cluster I	Cluster O	Clients in O	$d$max
I_1^{0-1}	O_4^{0-1}	61	0.000
	O_{64}^{0-1}	45	0.632
	O_{103}^{0-1}	54	0.854
I_2^{0-1}	O_{20}^{0-1}	50	0.987
	O_{54}^{0-1}	27	0.760
	O_{66}^{0-1}	12	0.863
	O_{80}^{0-1}	115	0.224
	O_{81}^{0-1}	96	0.777
	O_{82}^{0-1}	37	0.931
I_3^{0-1}	O_{11}^{0-1}	51	0.253
	O_{79}^{0-1}	72	0.132
	O_{81}^{0-1}	96	0.770
I_4^{0-1}	O_{33}^{0-1}	47	0.238
	O_{39}^{0-1}	92	0.709
	O_{69}^{0-1}	41	0.513
	O_{90}^{0-1}	51	0.303
	O_{108}^{0-1}	142	0.764
	O_{111}^{0-1}	143	0.571
I_5^{0-1}	O_9^{0-1}	67	0.864
	O_{19}^{0-1}	53	0.052
	O_{71}^{0-1}	37	0.671
	O_{87}^{0-1}	125	0.757
I_6^{0-1}	O_{38}^{0-1}	101	0.849
	O_{56}^{0-1}	35	0.274
I_7^{0-1}	O_{49}^{0-1}	57	0.353
	O_{65}^{0-1}	45	0.569
	O_{91}^{0-1}	86	0.873
	O_{120}^{0-1}	118	0.888
	O_{124}^{0-1}	27	0.420
I_8^{0-1}	O_{24}^{0-1}	43	0.640
	O_{40}^{0-1}	68	0.652
	O_{57}^{0-1}	36	0.157
	O_{77}^{0-1}	28	0.901
I_9^{0-1}	O_{51}^{0-1}	78	0.174
	O_{67}^{0-1}	66	0.297
I_{10}^{0-1}	O_2^{0-1}	25	0.784
	O_{13}^{0-1}	50	0.999
	O_{112}^{0-1}	81	0.062
	O_{116}^{0-1}	75	0.944
I_{11}^{0-1}	O_{55}^{0-1}	53	0.630
	O_{63}^{0-1}	39	0.997
	O_{83}^{0-1}	48	0.228
I_{12}^{0-1}	O_{13}^{0-1}	50	0.677
	O_{32}^{0-1}	60	0.213
	O_{34}^{0-1}	37	0.474
	O_{60}^{0-1}	34	0.631
	O_{61}^{0-1}	17	0.548
I_{13}^{0-1}	O_{25}^{0-1}	76	0.969
	O_{63}^{0-1}	39	0.123
	O_{65}^{0-1}	45	0.495
	O_{94}^{0-1}	140	0.736
I_{14}^{0-1}	O_{14}^{0-1}	24	0.557
	O_{53}^{0-1}	51	0.550
	O_{62}^{0-1}	21	0.474
	O_{113}^{0-1}	202	0.944
I_{15}^{0-1}	O_{73}^{0-1}	46	0.807
	O_{97}^{0-1}	152	0.988
I_{16}^{0-1}	O_{31}^{0-1}	48	0.727
	O_{109}^{0-1}	23	0.381
I_{17}^{0-1}	O_{30}^{0-1}	41	0.088
	O_{82}^{0-1}	37	0.917
	O_{96}^{0-1}	98	0.735
I_{18}^{0-1}	O_{26}^{0-1}	18	0.943
	O_{91}^{0-1}	86	0.191
	O_{103}^{0-1}	54	0.637
I_{19}^{0-1}	O_{24}^{0-1}	43	0.780
	O_{76}^{0-1}	77	0.695
	O_{119}^{0-1}	157	0.580
I_{21}^{0-1}	O_{28}^{0-1}	27	0.599
	O_{127}^{0-1}	27	0.761
I_{22}^{0-1}	O_{20}^{0-1}	50	0.667
	O_{102}^{0-1}	41	0.151

$$a_\upsilon = \min_{1 \leq j \leq 12} \{d\max(O_{i_\upsilon}, I_j)\} \tag{6}$$

$$b_\upsilon = \min_{1 \leq j \leq 22} \{d\max(O_{k_\upsilon}^{0-1}, I_j^{0-1})\}. \tag{7}$$

Note that by construction, for each client υ in *Other*, $\exists!$ $i_\upsilon \in \{1, \ldots, 391\}$ such that $\upsilon \in O_{i_\upsilon}$ and $\exists!$ $k_\upsilon \in \{1, \ldots, 128\}$ such that $\upsilon \in O_{k_\upsilon}^{0-1}$, then we obtain a good definition of equations (6) and (7). a_υ and b_υ represent marginal risks of client υ, since, a_υ depends on $Z_{i,t}$ – see equation (4) and b_υ depends on $Y_{i,t}$ – see equation (5). Even more, we can include in the risk definition process other representations, according to the information provided by the inspeccion. In Table 14 we report the number of cases, in the class *Other* (see Tab. 1), by risk bands.

If we consider as low risk those clients with $d$max near to 1 or more, there are 79 risk cases that should be inspected, cases inside the set $[0, 0.9] \times [0, 0.9]$ (in bold letter – Tab. 14). As exemplified in Table 14, various representations of the original information can be integrated into the definition of a client's risk, in this case, we have adopted two, which have revealed the need to first inspect 79 clients, according to both representations. If the cases indicated

Table 14. Number of clients by interval, a_υ given by equation (6) and b_υ given by equation (7). In bold the number of cases with high risk.

$b_\upsilon \backslash a_\upsilon$	[0, 0.1]	(0.1, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1)	[1, ∞)
[0, 0.1]	0	0	0	0	0	323
(0.1, 0.2]	0	**27**	0	0	7	343
(0.2, 0.3]	0	0	**5**	0	24	518
(0.3, 0.4]	0	0	0	0	30	184
(0.4, 0.5]	0	0	0	0	0	237
(0.5, 0.6]	0	0	0	0	10	447
(0.6, 0.7]	**36**	0	0	**9**	11	507
(0.7, 0.8]	0	0	**2**	0	28	588
(0.8, 0.9]	0	0	0	0	0	211
(0.9, 1)	0	0	0	11	6	469
[1, ∞)	0	0	28	20	66	3681

Table 15. Percentage of cases classified in the most indicated cluster. N: number of simulations with $s = 15$. Up: coding $Z_{i,t}$ – equation (4), Down: coding $Y_{i,t}$ – equation (5).

	N	Minimum (%)	First Quartile	Median	Mean	Third Quartile	Maximum
	30	66.27	75.90	77.71	77.31	80.42	84.34
$(Z_{i,t})_t$	50	65.06	74.70	78.31	77.86	81.93	86.75
	100	66.27	74.70	78.31	77.84	80.72	86.75
	30	98.80	100.00	100.00	99.96	100.00	100.00
$(Y_{i,t})_t$	50	100.00	100.00	100.00	100.00	100.00	100.00
	100	98.80	100.00	100.00	99.96	100.00	100.00

for inspection are many, according to the availability of the company, customer selection criteria such as the one described in [5] may be applied. Reference [5] shows that through a robust criterion, it is possible to select a representative client of the cluster that could be first inspected.

In the following subsection, we analyze the ability to detecting fraud, of the proposed strategy (see Eq. (3)), under each type of discretization (4) and (5).

Assertiveness of classification

We reserve this subsection to identify the predictive capacity of the classification given by Algorithm 1. What interests us is the quality of classification of *Irregular* customers, as these have gone through rigorous inspection processes, being defined as fraud. The database of our inspection is given in Table 1; for this, we proceed as follows. Consider the clusters defined by Algorithm 1 in the *Irregular* class: $I_1, I_2, \ldots, I_{k_I}$, (i) randomly select s% of irregular customers, say $\upsilon_{i_1}, \ldots, \upsilon_{i_{sk_I/100}}$; (ii) apply Algorithm 1 in the *Irregular* class (without the clients selected in (i)) and denote the clusters as $I'_1, I'_2, \ldots, I'_{k'_I}$. (iii) For each element $\upsilon_{i_j} \in I_{\upsilon_{i_j}}$ find the cluster $I'_{\upsilon_{i_j}}$ such that $I'_{\upsilon_{i_j}} \cap I_{\upsilon_{i_j}} \geq I'_i \cap I_{\upsilon_{i_j}}$, $\forall i \in \{1, \ldots, k'_I\}$, (iv.a) compute $\delta(\upsilon_{i_j}) = d\max(\upsilon_{i_j}, I'_{\upsilon_{i_j}})$ and (iv.b) record $|j \in \{i_1, \ldots, i_{sk_I/100}\} : \delta(\upsilon_j) < 1|$.

Note that the cluster $I'_{\upsilon_{i_j}}$ such that (iii) is verified can be considered as the most indicated cluster for the client υ_{i_j}, since the client υ_{i_j} is a member of $I_{\upsilon_{i_j}}$ and the sets $I'_{\upsilon_{i_j}}$ and $I_{\upsilon_{i_j}}$ share the largest number of customers.

Note that under both discretizations, the average percentage of successes is greater than 77% and, the minimum percentages exceed 65%, in the three settings by discretization, see Table 15.

4 Conclusion

In practical terms, this article deals with the capacity that discretizations have to extract relevant information contained in observations in series. Such discretizations make it possible to use and adapt tools from discrete stochastic processes. Through the metric – Definition 1 (see [1]), it is possible to measure the similarity/discrepancy between samples of discrete stochastic processes. Such a metric is statistically consistent for establishing the similarity/discrepancy. Based on the metric, in this article is proposed Algorithm 1 that defines clusters of samples, where each cluster contains those sequences that respond to the same stochastic law. From the previously demonstrated properties, see [1], the clusters are then assembled consistently and represent different profiles associated with the sequences inside. To identify how these profiles operate, we use the Partition Markov Models – Definition 2 – [2], which by means of a metric – Definition 3 is consistently estimated using the sequences located in the cluster. We generate a model for each cluster, which gives the minimal representation of the state space (partition) and the transition probabilities for any element of the alphabet. Based on all these elements, we deal with a real problem where there are sequences of observations of energy consumption of two groups (i) Irregular, (ii) Other – see Table 1. We define two types of discretization ((4) and (5)) through them we proceed to identify the clusters of (i) that group similar

consumption practices, we do the same with (ii). Through Partition Markov Models, we represent the stochastic profile of each cluster of (i) – see Remark 1. We identify which clusters of (ii) are confused with the clusters of (i) – see Remark 2, which allows us to point out the cases in (ii) that deserve inspection. The classification's rates of success given by the procedure are high, as shown in the study of Section 3.5, and on average, these exceed 77%. This whole procedure allows us to establish risk indicators for (ii) and also an order that indicates the most and least serious cases. We see then that, by means of two discretizations it is possible to point cases to be reviewed, according to the magnitudes of the notion (3), our results – Table 14 – states that 79 cases should go through revision, according to (6) and (7). For additional details, see [6].

Acknowledgments

The authors Hugo Helito da Silva and T. Soares Silva gratefully acknowledge the financial support provided by ANEEL R&D Program under grant PD-00063-3037/2018, developed by CPFL Energia. Also, the authors wish to thank the Editor-led peer review process and the reviewers which generated many helpful comments and suggestions on an earlier draft of this paper.

References

1. García Jesús E, Gholizadeh R, González-López VA (2018), A BIC-based consistent metric between Markovian processes. Appl Stoch Models Bus Ind 34, 6, 868–878.
2. García Jesús E, González-López VA (2017), Consistent estimation of Partition Markov Models. Entropy 19, 4, 160.
3. Cordeiro MTA, García Jesús E, González-López VA, Londoño SLM (2019), Classification of autochthonous dengue virus type 1 strains circulating in Japan in 2014. 4open 2, 20.
4. Schwarz G (1978), Estimating the dimension of a model. Ann Stat 6, 2, 461–464.
5. Fernández M, García Jesús E, Gholizadeh R, González-López VA (2019), Sample selection procedure in daily trading volume processes. Math Meth Appl Sci 43, 7537–7549. https://doi.org/10.1002/mma.5705.
6. Soares Silva T (2020), Similaridades entre Processos de Markov (unpublished master's thesis).

Appendix

Table A1. PMM for the irregular clusters I_i^{0-1}, $i = 1, \ldots, 12$, see equation (5) and Remark 1.

	I_1^{0-1}	
Part	0	1
00	0.767	0.233
01 10 11	0.391	0.609

	I_2^{0-1}	
Part	0	1
00 10	0.356	0.644
01 11	0.570	0.430

	I_3^{0-1}	
Part	0	1
00	0.234	0.766
01 10	0.484	0.516
11	0.632	0.368

	I_4^{0-1}	
Part	0	1
00 10	0.412	0.588
01	0.549	0.451
11	0.862	0.138

	I_5^{0-1}	
Part	0	1
00 01	0.486	0.514
10 11	0.652	0.348

	I_6^{0-1}	
Part	0	1
00	0.483	0.517
01 10 11	0.578	0.422

	I_7^{0-1}	
Part	0	1
00	0.569	0.431
01	0.388	0.612
10 11	0.472	0.528

	I_8^{0-1}	
Part	0	1
00 01	0.627	0.373
10 11	0.418	0.582

	I_9^{0-1}	
Part	0	1
00 01	0.548	0.452
10	0.385	0.615
11	0.732	0.268

	I_{10}^{0-1}	
Part	0	1
00	0.554	0.446
01 10 11	0.486	0.514

	I_{11}^{0-1}	
Part	0	1
00 11 10	0.585	0.415
01	0.442	0.558

	I_{12}^{0-1}	
Part	0	1
00 11	0.463	0.537
01 10	0.560	0.440

Table A2. PMM for the irregular clusters I_i^{0-1}, $i = 13, \ldots, 22$, see equation (5) and Remark 1.

I_{13}^{0-1}

Part	0	1
00 11	0.587	0.413
01 10	0.361	0.639

I_{14}^{0-1}

Part	0	1
00	0.366	0.634
01	0.620	0.380
10 11	0.490	0.510

I_{15}^{0-1}

Part	0	1
00	0.339	0.661
01 10 11	0.590	0.410

I_{16}^{0-1}

Part	0	1
00	0.396	0.604
01 10	0.459	0.541
11	0.596	0.404

I_{17}^{0-1}

Part	0	1
00 10	0.380	0.620
01 11	0.645	0.355

I_{18}^{0-1}

Part	0	1
00 10	0.598	0.402
01 11	0.373	0.627

I_{19}^{0-1}

Part	0	1
00 10	0.503	0.497
01	0.590	0.410
11	0.411	0.589

I_{20}^{0-1}

Part	0	1
00 01 11	0.546	0.454
10	0.357	0.643

I_{21}^{0-1}

Part	0	1
00 10	0.489	0.511
01 11	0.623	0.377

I_{22}^{0-1}

Part	0	1
00	0.382	0.618
01	0.542	0.458
10	0.446	0.554
11	0.692	0.308

Cite this article as: García JE, González-López VA, da Silva HH & Silva TS 2020. Risk of fraud classification. 4open, **3**, 9.