Risk of fraud classification

Jesús Enrique García¹, Verónica Andrea González-López¹, Hugo Helito da Silva², and Thainá Soares Silva¹,*

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 fraudules; 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 the previously determined profiles. Then the k 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 \( Z_t \) be a discrete time Markov chain of order \( o \) (\( o < \infty \)) with finite alphabet \( A \). Let us call \( S = A^o \) the state space and denote the string \( a_m a_{m-1} \ldots a_1 \) by \( a^m \), where \( a_i \in A \), \( m \leq i \leq n \). For each \( a \in A \) and \( s \in S \) define the transition probability

\[ P(a|s) = \text{Prob} \{ Z_t = a | Z_{t-1} = s \} \]

In a given sample \( x^n \), coming from the stochastic process, the number of occurrences of \( s \) is denoted by \( N_s(x) \) and the number of occurrences of \( s \) followed by \( a \) is denoted by \( N_{sa}(x, a) \). This way, \( \frac{N_{sa}(x, a)}{N_s(x)} \) is the maximum likelihood estimator of \( P(a|s) \). Consider

*Corresponding author: thainass@outlook.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
now, two Markov chains \((Z_{1,t})_t\) and \((Z_{2,t})_t\) of order \(n\), disposed on the finite alphabet \(A\) with state space \(S\). Given \(s \in S\) 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 \(S\) (not negative, symmetric and follows triangular inequality) and it allows defining a global notion (in \(S\)) of similarity between sequences.

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

\[
(i) \quad d_s(z_{1,t}^{n_1}, z_{2,t}^{n_2}) = \frac{\alpha}{|A| - 1} \ln \left( \frac{N_{n_1}(s, a) N_{n_2}(s, a) - N_{n_1 + n_2}(s, a)}{N_{n_1}(s) N_{n_2}(s)} \right),
\]

\[
(ii) \quad \max_{s \in S}(z_{1,t}^{n_1}, z_{2,t}^{n_2}) = \max_{s \in S}(d_s(z_{1,t}^{n_1}, z_{2,t}^{n_2}))
\]

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

\[
(i) \quad (Z_t)\text{ is a Markov chain with partition } \mathcal{L} = \{L_1, L_2, \ldots, L_k\} \text{ if this partition is the one defined by the equivalence introduced by item } i. \\
(ii) \quad (Z_t)\text{ is a Markov chain with partition } \mathcal{L} = \{L_1, L_2, \ldots, L_k\} \text{ if this partition is the one defined by the equivalence introduced by item } i.
\]

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 \(A\), \(P(|L|) = \sum_{s \in S} P(|s|)\). Given a sample of \((Z_t)\), \(Z_t^n\), according to [2] the partition is estimated by means of \(d_L\) given by **Definition 3**.

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

\[
d_L(i, j) = \frac{\alpha}{|A| - 1} \ln \left( \frac{\sum_{a \in A} \sum_{l \in L_i} N_n(L_{ai}, a) \ln \left( \frac{N_n(L_{ai}, a)}{N_n(L_{aj}, a)} \right)}{\sum_{a \in A} \sum_{l \in L_j} N_n(L_{aj}, a) \ln \left( \frac{N_n(L_{aj}, a)}{N_n(L_{ai}, a)} \right)} \right),
\]

with \(N_n(L) = \sum_{s \in L} N_n(s, a)\), \(N_n(L, a) = \sum_{s \in L} N_n(s, a)\), for \(a \in A\), \(L_i \subseteq L_j\) and \(L_j\) are equivalent if \(\mathcal{L}\) is composed by all the separate elements, \(L_i \subseteq L_j\) and \(L_j\) must be in the same part (see properties of \(d_L\) in [2]). For each part \(L\) of \(\mathcal{L}\) (estimated partition) the transition probability is estimated by \(\hat{P}(a|L) = \frac{N_n(L_{ai}, a)}{N_n(L_{aj}, a)}\). 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 \(\mathcal{C} = \{Z_t^n\}_{i=1}^p\), under the assumptions of **Definition 1**, the notion \(\max \pi i - \pi j\) of similarity is used to define clusters in \(\mathcal{C}\). We introduce an algorithm that shows how this is done.

**Algorithm 1**

- **Input** \(\mathcal{C} = \{Z_t^n\}_{i=1}^p\)  
  1. \(M = \mathcal{C}\)  
  2. \(M = \{m_1, \ldots, m_M\}\)  
  3. \((i, j) = \arg\min_{i \neq j} \max_{m_i, m_j} \{m_i, m_j\} \)  
  * if \(d\max(m_i, m_j) < 1\), \(M = \{m_i, m_j\}\)  
  * 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 \(z_{1,t}^{n_1}, z_{2,t}^{n_2} \in C\) the occurrence of each \(a \in S\) is recorded by \(N_{n_1}(s, a)\) and \(N_{n_2}(s, a)\) respectively, and the occurrence of \(s\) followed by \(a \in A\) is computed by \(N_{n_1}(s, a)\) and \(N_{n_2}(s, a)\). Already, when defining the new unit
Once the proximity between sequences is determined in order to build the clusters \( \{C_1, \ldots, C_l\} \) and for each cluster we can build a PFM, representing the cluster. In addition, it is possible to quantify the dissimilarity between clusters using the notion \( d_{\text{max}} \). Suppose the cluster \( i \) is \( C_l \) and it is composed by \( m_i \) independent sequences,

\[
C_l = \left\{ z_{i,1}^{n_i}, z_{i,2}^{n_i}, \ldots, z_{i,m_i}^{n_i} \right\} = \left\{ z_{im_i,1}^{n_i} \right\}_i m_i,
\]

the sample size related to \( C_l \) is \( \sum_{m=1}^{m_i} n_{\text{mc}} \). For each \( s \in S \), compute the occurrences of \( s \) in \( C_l \) as

\[
N(C_l, s) = \sum_{m=1}^{m_i} N_{\text{mc}}(s)
\]

and the occurrences of \( s \) followed by \( a \) as

\[
N(C_l, (s, a)) = \sum_{m=1}^{m_i} N_{\text{ma}}(s, a),
\]

for \( N_{\text{mc}}(\cdot) \) related to the sample \( z_{im_i,1}^{n_i} \).

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 \( \sum_{m=1}^{m_i} n_{\text{mc}} \) and applying the Algorithm 1, substituting (i) \( C = \{z_{im_i,1}^{n_i}\}_m \) by \( S = A^{\infty} \), (ii) \( d_{\text{max}} \) by \( d_{\text{Ch}} \), the Output of the algorithm will be the partition \( C_l \) of \( S \), related to the cluster \( C_l \).

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

Remark 2. To establish the dissimilarity between the clusters \( C_l \) and \( C_k \) (since by construction those are different) we use the Definition 1 – ii. In the calculation of i-def \( 2.1\), we replace \( N_{\text{mc}}(s)(N_{\text{ma}}(s, a)) \) by equation (1) ((2)), with \( i = k \). We replace also \( N_{1+\text{mc}}(s) \) by \( N(C_l, s) + N(C_k, s) \) and \( N_{1+\text{ma}}(s, a) \) by \( N(C_l, (s, a)) + N(C_k, (s, a)) \). Using those occurrences 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 analyses**

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.

The monthly energy consumption sequence of each client \( i, X_{i,t} \) is discretized in order to identify it with a sample of a Markov stochastic process \( (Z_{i,t}) \), of finite order \( o \) in the discrete and finite alphabet \( A \), for \( i = 1, \ldots, 8831 \). 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 \( \{O_1, \ldots, O_{8831}\} \) clusters of irregular clients, applying the Algorithm 1 in the set of irregular clients. For the group **Other**, we also determine the clusters, say \( \{O_1, \ldots, O_{8828}\} \) (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 \( d_{\text{max}} \) as described in the previous section (see Remark 2), computing \( d_{\text{max}}(I_i, I_j) \), \( i \neq j, i, j \in \{1, 2, \ldots, k_I\} \). In a second instance, we compare the behavior of the \( O_h, l = 1, \ldots, k_O \) clusters with the irregular ones, computing \( d_{\text{max}}(I_i, O_l) \), \( i \in \{1, \ldots, k_I\} \), \( l \in \{1, \ldots, k_O\} \). We do this comparison in order to identify which could be considered as indistinguishable from some irregular cluster, this happens when \( d_{\text{max}}(I_i, O_l) < 1 \). Such a comparison generates a risk index in the class \( \{O_1, \ldots, O_{8828}\} \), as to guide the inspection of the company in that class. For each client \( v \in O_h \), we define:

\[
a_v = \min_{1 \leq j \leq k_I} \{d_{\text{max}}(O_h, I_j)\}
\]

In this way, the values obtained from equation (3) reported for the clients in the class **Other** (Tab. 1) are \( \{a_1, \ldots, a_{8828}\} \). By construction for the client \( v \) in **Other**, \( \exists i, i \in \{1, \ldots, k_I\} \) such that \( v \in O_h \) allowing the good definition of equation (3). Denote by \( a_1(1), a_2(2), \ldots, a_{8828}(7828) \) the ordered values in an increasing way. Thus, the client that receives the value \( a_1(1) \) is the one with the highest risk and the one that receives the value \( a_{8828}(7828) \) 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 1. Biphasic clients reported by CPFL, period: January, 2011 to June, 2019.**

| Type       | Total of clients |
|------------|------------------|
| Irregular  | 553              |
| Other      | 7828             |

**Figure 1.** Scheme of the organization of the ordered values \( \{a_i(1)\}_{i=1}^{7828} \), on the left those that indicate greater risk, on the right those of lower risk. Cut = 1 indicates the threshold given by the BIC.
Table 2. Description of the $k_l = 12$ irregular clusters $I_i$, $i = 1, \ldots, 12$. $n_i$: clients in $I_i$. $d_i^*$: maximal $\Delta$max attained by $I_i$ values reported from left to right in increasing order. Using the process $(Z_{i,l})_l$ - see equation (4).

| $i$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| $n_i$ | 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. $\Delta$max between the irregular clusters $\{I_1, \ldots, I_{12}\}$ (see Eq. (4) and Remark 2). In bold type, the three highest values.

| $I_2$ | $I_3$ | $I_4$ | $I_5$ | $I_6$ | $I_7$ | $I_8$ | $I_9$ | $I_{10}$ | $I_{11}$ | $I_{12}$ |
|-------|-------|-------|-------|-------|-------|-------|-------|---------|---------|---------|
| 1.355 | 3.557 | 2.294 | 2.581 | 3.409 | 3.141 | 2.047 | 2.739 | 2.045 | 1.859 | 3.063 |
| 2.024 | 1.397 | 1.362 | 2.393 | 1.522 | 1.715 | 1.170 | 1.566 | 1.295 | 2.623 | 1.485 |
| 1.557 | 2.006 | 1.387 | 1.186 | 1.085 | 2.336 | 2.985 | 2.859 | 1.176 | 1.238 | 2.006 |
| 2.584 | 2.176 | 1.609 | 1.863 | 2.322 | 1.639 | 1.563 | 3.279 | 3.324 | 3.372 | 2.357 |
| 1.511 | 1.164 | 1.599 | 2.620 | 1.901 | 3.525 | 3.509 | 1.638 | 1.696 | 3.773 |
| 2.092 | 2.287 | 2.209 | 2.379 | 2.209 | 2.701 | 2.287 | 2.209 | 3.379 | 3.773 |

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

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

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

| State | Impossible event |
|-------|------------------|
| 12    | $\{X_{t-2} \leq X_t < X_{t-3}\}$ |
| 22    | $\{X_{t-2} \leq X_t < X_{t-1} \leq X_{t-3}\}$ |
| 33    | $\{X_{t-2} \leq X_t < X_{t-3} \leq X_{t-1}\}$ |
| 43    | $\{X_{t-2} \leq X_t < X_{t-1} \leq X_{t-2}\}$ |

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 $X_{t-1}^i$ consumption series define the sample $Z_{i,1}^j$ of the discrete process $(Z_{i,1})_l$ as

$$Z_{i,l} = \begin{cases} 1, & x_{it} < x_{i,t-1}, x_{it} < x_{i,t-2} \\ 2, & x_{it} < x_{i,t-1}, x_{it} \geq x_{i,t-2} \\ 3, & x_{it} \geq x_{i,t-1}, x_{it} < x_{i,t-2} \\ 4, & x_{it} \geq x_{i,t-1}, x_{it} \geq x_{i,t-2} \end{cases}.$$  (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 $k_l = 12$ clusters. Table 2 exposes the maximum value of $\Delta$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.

After identifying 12 Irregular clusters, we can explore the dissimilarity between them computing the values of $\Delta$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 $(Z_{i,l})_l$ process. In Table 4, we report the relation of the states $s \in S$ 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 $\{X_t \geq X_{t-1} \text{ & } X_{t-2} \leq X_{t-3}\}$ (associated with $b = 3$) and $\{X_{t-1} < X_{t-2} \text{ & } X_{t-3} < X_{t-1}\}$ (associated with $a = 1$). Note that some combinations are not allowed by construction, those are 12, 22, 33, 43 (see Tab. 5).
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 have separated the states into those two large classes. That is, in each part of each model we only have two possible exceptions (in bold), for state 1, cases $I_1$, $I_2$, $I_3$, $I_4$, $I_5$, $I_6$, $I_7$, $I_8$, $I_9$, $I_{10}$, $I_{11}$, 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 $I_1$). 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 $Z_{ij} –$ 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 $d_{\text{max}}$ between such irregular cluster and the clusters $O_{ji}$, $j = 1, \ldots, k_O$. Table 8 summarizes the obtained values.

In Table 9, we report which $O_j$ clusters behave as irregular. The lower the value of $d_{\text{max}}$ 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 ($d_{\text{max}}$ values below 0.7). And certainly, the priority is for the 36 with approximately zero $d_{\text{max}}$.

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.

| $I_1$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11 41 | **0.615** | 0.000 | 0.385 | 0.000 |
| 13    | 0.545 | 0.091 | 0.000 | 0.364 |
| 14 24 44 | 0.176 | 0.265 | 0.000 | 0.559 |
| 31 42 34 | 0.286 | 0.000 | 0.143 | 0.571 |

| $I_2$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11    | **0.521** | 0.000 | 0.338 | 0.140 |
| 13 23 | 0.323 | 0.186 | 0.000 | 0.492 |
| 14 34 24 | 0.245 | 0.358 | 0.000 | 0.397 |
| 21 31 41 | 0.380 | 0.000 | 0.342 | 0.278 |
| 32 42 | 0.403 | 0.000 | 0.140 | 0.457 |
| 44    | 0.420 | 0.283 | 0.000 | 0.297 |

| $I_3$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11    | 0.212 | 0.000 | 0.300 | 0.488 |
| 13 44 14 | 0.243 | 0.181 | 0.000 | 0.576 |
| 21 23 | **0.810** | 0.000 | 0.127 | 0.063 |
| 24 34 | 0.352 | 0.521 | 0.000 | 0.127 |
| 31 41 42 | 0.433 | 0.000 | 0.369 | 0.197 |

| $I_4$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11 21 32 | 0.410 | 0.000 | 0.276 | 0.314 |
| 13    | 0.369 | 0.084 | 0.000 | 0.557 |
| 14 23 24 | 0.240 | 0.319 | 0.000 | 0.441 |
| 44 34 | 0.319 | 0.000 | 0.364 | 0.128 |
| 42    | 0.489 | 0.000 | 0.060 | 0.451 |

| $I_5$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11 21 32 | 0.410 | 0.000 | 0.276 | 0.314 |
| 13    | 0.369 | 0.084 | 0.000 | 0.557 |
| 14 23 24 | 0.240 | 0.319 | 0.000 | 0.441 |
| 44 34 | 0.319 | 0.000 | 0.364 | 0.128 |
| 42    | 0.489 | 0.000 | 0.060 | 0.451 |

| $I_6$ | Part | 1 | 2 | 3 | 4 |
|-------|------|---|---|---|---|
| 11 21 32 | 0.410 | 0.000 | 0.276 | 0.314 |
| 13    | 0.369 | 0.084 | 0.000 | 0.557 |
| 14 23 24 | 0.240 | 0.319 | 0.000 | 0.441 |
| 44 34 | 0.319 | 0.000 | 0.364 | 0.128 |
| 42    | 0.489 | 0.000 | 0.060 | 0.451 |

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.
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 $x_{qi}$ define the sample $y_{ni}$ of the discrete process $(Y_{i,t})$ as

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

### 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.

| $I_7$ | Part | 1 | 2 | 3 | 4 |
|------|------|---|---|---|---|
| 11   | 0.357| 0.000| 0.245| 0.398|
| 13 24| 0.332| 0.233| 0.000| 0.435|
| 14   | 0.200| 0.425| 0.000| 0.375|
| 21 42| 0.514| 0.000| 0.237| 0.249|
| 24 34| 0.255| 0.302| 0.000| 0.444|
| 3141 | 0.391| 0.000| 0.398| 0.211|
| 32   | 0.423| 0.000| 0.115| 0.462|

| $I_8$ | Part | 1 | 2 | 3 | 4 |
|------|------|---|---|---|---|
| 11   | 0.548| 0.000| 0.250| 0.209|
| 13 24| 0.150| 0.121| 0.000| 0.729|
| 14   | 0.273| 0.438| 0.000| 0.289|
| 21 31| 0.595| 0.000| 0.095| 0.311|
| 34 44| 0.397| 0.108| 0.000| 0.495|

### 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|

### 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 $I_i$ | Cluster $O_j$ | Clients in $O_j$ | $d_{max}$ |
|---------------|---------------|-----------------|-----------|
| $I_1$         | $O_{75}$      | 36              | 0.000*    |
| $I_1$         | $O_{16}$      | 9               | 0.960     |
| $I_2$         | $O_{17}$      | 6               | 0.885     |
| $I_3$         | $O_{57}$      | 27              | 0.673*    |
| $I_4$         | $O_{56}$      | 49              | 0.952     |
| $I_5$         | $O_{23}$      | 7               | 0.841     |
| $I_6$         | $O_{23}$      | 7               | 0.812     |
| $I_7$         | $O_{23}$      | 7               | 0.841     |
| $I_8$         | $O_{23}$      | 7               | 0.812     |

### Table 10. States $s \in S$ 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    | $X_t \geq X_{t-1} \geq X_{t-2}$ |
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 $(Y_{i,t})$ in Table 10.

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

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 $(Y_{i,t})$. 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: $I^{i-1}$, $i = 11, 12, 13, 20$.

The group Other, under the representation $Y_{i,t}$ equation (5) is divided by Algorithm 1 in 128 clusters, $O^{i-1}_1, \ldots, O^{i-1}_{k_O}$ 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 $I^{i-1}$ we calculate the $d_{\text{max}}$ between such irregular cluster and the clusters $O^{i-1}_j$, $j = 1, \ldots, k_O$. Table 12 shows the results. According to coding 0–1, the only cluster $I^{i-1}_1$ that is not associated with any element of the class Other (see Tab. 1) is the $I^{20}_{20}$ which has 68 clients. We must not lose sight of the fact that the risk increases when obtaining values of $d_{\text{max}}$ 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 $d_{\text{max}}$ near to the threshold 1 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.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|-----|---|---|---|---|---|---|---|---|---|----|----|
| $n_{i-1}$ | 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 |

| Cluster | Min | Median | Max |
|---------|-----|--------|-----|
| $I^{1}_1$ | 0.000* | 5.528 | 29.342 |
| $I^{2}_0$ | 0.224* | 5.881 | 20.855 |
| $I^{3}_0$ | 0.132* | 5.646 | 26.392 |
| $I^{4}_0$ | 0.238* | 5.508 | 14.693 |
| $I^{5}_0$ | 0.052* | 4.670 | 17.643 |
| $I^{6}_0$ | 0.274* | 5.221 | 24.873 |
| $I^{7}_0$ | 0.353* | 8.258 | 32.664 |
| $I^{8}_0$ | 0.157* | 6.450 | 33.795 |
| $I^{9}_0$ | 0.174* | 6.887 | 29.933 |
| $I^{10}_0$ | 0.062* | 6.772 | 34.111 |
| $I^{11}_0$ | 0.228* | 5.817 | 33.809 |
| $I^{12}_0$ | 0.213* | 7.179 | 35.029 |
| $I^{13}_0$ | 0.123* | 6.060 | 23.934 |
| $I^{14}_0$ | 0.474* | 7.484 | 35.978 |
| $I^{15}_0$ | 0.807* | 6.319 | 30.182 |
| $I^{16}_0$ | 0.381* | 8.999 | 36.916 |
| $I^{17}_0$ | 0.088* | 10.210 | 44.606 |
| $I^{18}_0$ | 0.191* | 8.152 | 35.757 |
| $I^{19}_0$ | 0.580* | 9.758 | 46.892 |
| $I^{20}_0$ | 1.109 | 10.858 | 34.376 |
| $I^{21}_0$ | 0.599* | 7.012 | 36.268 |
| $I^{22}_0$ | 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 $(Z_{i,t})$ and $(Y_{i,t})$ 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 $v \in O_{i} \cap O_{k_O}^{i-1}$ we compute:

$$\text{Risk}(v) = \text{Risk}(Z_{i,t}) + \text{Risk}(Y_{i,t})$$
Table 13. For $I_{i}^{\alpha}$, $i \neq 20, i = 1, \ldots, 22$ are reported the clusters $O_{i}^{\alpha-1}$ with $d_{\text{max}} < 1$ – see equation (5) and Remark 2.

| Cluster $I$ | Cluster $O$ | Clients in $O$ | $d_{\text{max}}$ |
|------------|-------------|----------------|-----------------|
| $I_{1}^{0-1}$ | $O_{i}^{0-1}$ | 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 |
| $O_{24}^{0-1}$ | 27 | 0.760 |
| $O_{66}^{0-1}$ | 12 | 0.863 |
| $O_{66}^{0-1}$ | 115 | 0.224 |
| $O_{61}^{0-1}$ | 96 | 0.777 |
| $O_{62}^{0-1}$ | 37 | 0.931 |
| $I_{3}^{0-1}$ | $O_{41}^{0-1}$ | 51 | 0.253 |
| $O_{79}^{0-1}$ | 72 | 0.132 |
| $O_{81}^{0-1}$ | 96 | 0.770 |
| $I_{4}^{0-1}$ | $O_{43}^{0-1}$ | 47 | 0.238 |
| $O_{43}^{0-1}$ | 92 | 0.709 |
| $O_{69}^{0-1}$ | 41 | 0.513 |
| $O_{80}^{0-1}$ | 51 | 0.303 |
| $O_{88}^{0-1}$ | 142 | 0.764 |
| $O_{111}^{0-1}$ | 143 | 0.571 |
| $I_{5}^{0-1}$ | $O_{9}^{0-1}$ | 67 | 0.864 |
| $O_{9}^{0-1}$ | 53 | 0.052 |
| $O_{11}^{0-1}$ | 37 | 0.671 |
| $O_{72}^{0-1}$ | 125 | 0.757 |
| $I_{6}^{0-1}$ | $O_{38}^{0-1}$ | 101 | 0.849 |
| $O_{65}^{0-1}$ | 35 | 0.274 |
| $I_{7}^{0-1}$ | $O_{69}^{0-1}$ | 57 | 0.353 |
| $O_{65}^{0-1}$ | 45 | 0.569 |
| $O_{91}^{0-1}$ | 86 | 0.873 |
| $O_{91}^{0-1}$ | 118 | 0.888 |
| $O_{128}^{0-1}$ | 27 | 0.420 |
| $I_{8}^{0-1}$ | $O_{24}^{0-1}$ | 43 | 0.640 |
| $O_{60}^{0-1}$ | 68 | 0.652 |
| $O_{71}^{0-1}$ | 36 | 0.157 |
| $O_{124}^{0-1}$ | 28 | 0.901 |
| $I_{9}^{0-1}$ | $O_{34}^{0-1}$ | 78 | 0.174 |
| $O_{67}^{0-1}$ | 66 | 0.297 |
| $I_{10}^{0-1}$ | $O_{25}^{0-1}$ | 25 | 0.784 |
| $O_{63}^{0-1}$ | 50 | 0.999 |
| $O_{62}^{0-1}$ | 81 | 0.062 |
| $O_{62}^{0-1}$ | 75 | 0.944 |
| $I_{11}^{0-1}$ | $O_{55}^{0-1}$ | 53 | 0.630 |
| $O_{53}^{0-1}$ | 39 | 0.997 |
| $O_{61}^{0-1}$ | 48 | 0.228 |
| $I_{12}^{0-1}$ | $O_{13}^{0-1}$ | 50 | 0.677 |
| $O_{12}^{0-1}$ | 60 | 0.213 |
| $O_{14}^{0-1}$ | 37 | 0.474 |
| $O_{60}^{0-1}$ | 34 | 0.631 |

Table 13. (Continued)

| Cluster $I$ | Cluster $O$ | Clients in $O$ | $d_{\text{max}}$ |
|------------|-------------|----------------|-----------------|
| $I_{13}^{0-1}$ | $O_{51}^{0-1}$ | 17 | 0.548 |
| $O_{25}^{0-1}$ | 76 | 0.969 |
| $O_{65}^{0-1}$ | 39 | 0.123 |
| $O_{14}^{0-1}$ | 45 | 0.495 |
| $O_{66}^{0-1}$ | 140 | 0.736 |
| $I_{14}^{0-1}$ | $O_{21}^{0-1}$ | 24 | 0.557 |
| $O_{62}^{0-1}$ | 51 | 0.550 |
| $O_{123}^{0-1}$ | 21 | 0.474 |
| $O_{26}^{0-1}$ | 202 | 0.944 |
| $I_{15}^{0-1}$ | $O_{21}^{0-1}$ | 46 | 0.807 |
| $O_{67}^{0-1}$ | 152 | 0.988 |
| $I_{16}^{0-1}$ | $O_{31}^{0-1}$ | 48 | 0.727 |
| $O_{69}^{0-1}$ | 23 | 0.381 |
| $I_{17}^{0-1}$ | $O_{34}^{0-1}$ | 41 | 0.988 |
| $O_{22}^{0-1}$ | 37 | 0.917 |
| $O_{96}^{0-1}$ | 98 | 0.735 |
| $I_{18}^{0-1}$ | $O_{26}^{0-1}$ | 18 | 0.943 |
| $O_{31}^{0-1}$ | 86 | 0.191 |
| $O_{33}^{0-1}$ | 54 | 0.637 |
| $I_{19}^{0-1}$ | $O_{24}^{0-1}$ | 43 | 0.780 |
| $O_{129}^{0-1}$ | 77 | 0.695 |
| $O_{129}^{0-1}$ | 157 | 0.580 |
| $I_{21}^{0-1}$ | $O_{127}^{0-1}$ | 27 | 0.599 |
| $O_{127}^{0-1}$ | 27 | 0.761 |
| $I_{22}^{0-1}$ | $O_{29}^{0-1}$ | 50 | 0.667 |
| $O_{62}^{0-1}$ | 41 | 0.151 |

$$a_{\alpha} = \min_{1 \leq i \leq 12} \{d_{\text{max}}(O_{\alpha}, I_{j})\}$$  \(\text{Eq. (6)}\)

$$b_{\beta} = \min_{1 \leq i \leq 22} \{d_{\text{max}}(O_{\beta}^{0-1}, I_{j}^{0-1})\}.$$  \(\text{Eq. (7)}\)

Note that by construction, for each client $v$ in Other, $\exists! \; i_{v} \in \{1, \ldots, 391\}$ such that $v \in O_{i_{v}}$ and $\exists! \; k_{v} \in \{1, \ldots, 128\}$ such that $v \in O_{k_{v}}^{0-1}$, then we obtain a good definition of equations (6) and (7). $a_{\alpha}$ and $b_{\beta}$ represent marginal risks of client $v$, since, $a_{\alpha}$ depends on $Z_{i,t}$ – see equation (4) and $b_{\beta}$ 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 inspection.

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 $d_{\text{max}}$ 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, \( a_i \) given by equation (6) and \( b_i \) given by equation (7). In bold the number of cases with high risk.

| \( b_i \) \( a_i \) | [0, 0.1] | (0.1, 0.7] | (0.7, 0.8] | (0.8, 0.9] | (0.9, 1] | [1, \( \infty \] |
|----------------|---------|----------|----------|----------|---------|---------|
| [0, 0.1]      | 0       | 0        | 0        | 0        | 0       | 323     |
| (0.1, 0.2]    | 0       | 27       | 0        | 0        | 0       | 343     |
| (0.2, 0.3]    | 0       | 0        | 5        | 0        | 0       | 24      |
| (0.3, 0.4]    | 0       | 0        | 0        | 0        | 0       | 30      |
| (0.4, 0.5]    | 0       | 0        | 0        | 0        | 0       | 0       |
| (0.5, 0.6]    | 0       | 0        | 0        | 0        | 0       | 10      |
| (0.6, 0.7]    | 36      | 0        | 0        | 0        | 9       | 11      |
| (0.7, 0.8]    | 0       | 0        | 2        | 0        | 0       | 28      |
| (0.8, 0.9]    | 0       | 0        | 0        | 0        | 0       | 6       |
| (0.9, 1]      | 0       | 0        | 0        | 0        | 0       | 0       |
| [1, \( \infty \] | 0       | 0        | 28       | 0        | 20      | 66      |

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

| \( Z_{i,t} \) | \( N \) | Minimum (%) | First Quartile | Median | Mean | Third Quartile | Maximum |
|-------------|-------|-------------|----------------|--------|------|----------------|---------|
| \( Z_{i,t} \) | 30    | 66.27       | 75.90          | 77.71  | 77.31| 80.42          | 84.34   |
| 50          | 65.06 | 74.70       | 78.31          | 77.84  | 80.72| 86.75          | 86.75   |
| 100         | 66.27 | 74.70       | 78.31          | 77.84  | 80.72| 86.75          | 86.75   |
| \( Y_{i,t} \) | 30    | 98.80       | 100.00         | 100.00 | 100.00| 100.00         | 100.00  |
| 50          | 100.00| 100.00      | 100.00         | 100.00 | 100.00| 100.00         | 100.00  |
| 100         | 98.80 | 100.00      | 100.00         | 100.00 | 100.00| 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) randomly select \( 8\% \) of irregular customers, say \( u_{i_1}, \ldots, u_{i_{8\%/\text{num}}}; \) (ii) apply Algorithm 1 in the Irregular class (without the clients selected in (i)) and denote the clusters as \( I_{i_1}', I_{i_2}', \ldots, I_{i_k}' \). (iii) For each element \( v_j \in I_{i_k}' \) find the cluster \( I_{v_j}' \) such that \( I_{v_j} \cap I_{i_k} \geq I_{v_j} \cap I_{i_k}' \), \( \forall i \in \{1, \ldots, k\} \), (iv.a) compute \( \delta(u_{i_j}) = \hat{d} \max(u_{i_j}, I_{v_j}') \) and (iv.b) record \( j \in \{i_1, \ldots, i_{8\%/\text{num}}\} : \delta(u_{i_j}) < 1 \).

Note that the cluster \( I_{v_j}' \) such that (iii) is verified can be considered as the most indicated cluster for the client \( u_{i_j} \), since the client \( v_j \) is a member of \( I_{v_j} \) and the sets \( I_{v_j} \) and \( I_{i_k}' \) 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 $P_{i}^{0-1}$, $i = 1, \ldots, 12$, see equation (5) and Remark 1.

| $P_{1}^{0-1}$   | $P_{2}^{0-1}$   | $P_{3}^{0-1}$   |
|-----------------|-----------------|-----------------|
| Part 0 1        | Part 0 1        | Part 0 1        |
| 00              | 0.767 0.233     | 00              | 0.356 0.644     |
| 01 10 11        | 0.391 0.609     | 01 11           | 0.570 0.430     |
| $P_{4}^{0-1}$   | $P_{5}^{0-1}$   | $P_{6}^{0-1}$   |
| Part 0 1        | Part 0 1        | Part 0 1        |
| 00              | 0.412 0.588     | 00              | 0.486 0.514     |
| 01              | 0.549 0.451     | 10 11           | 0.652 0.348     |
| 11              | 0.862 0.138     | 01 10 11        | 0.578 0.422     |
| $P_{7}^{0-1}$   | $P_{8}^{0-1}$   | $P_{9}^{0-1}$   |
| Part 0 1        | Part 0 1        | Part 0 1        |
| 00              | 0.569 0.431     | 00              | 0.627 0.373     |
| 01              | 0.388 0.612     | 10 11           | 0.418 0.582     |
| 10 11           | 0.472 0.528     | 10              | 0.385 0.615     |
| $P_{10}^{0-1}$  | $P_{11}^{0-1}$  | $P_{12}^{0-1}$  |
| Part 0 1        | Part 0 1        | Part 0 1        |
| 00              | 0.554 0.446     | 00 11 10        | 0.585 0.415     |
| 01 10 11        | 0.486 0.514     | 01              | 0.442 0.558     |

J.E. García et al.: 4open 2020, 3, 9
Table A2. PMM for the irregular clusters $I_i^{0-1}$, $i = 13, \ldots, 22$, see equation (5) and Remark 1.

| Part | $I_{13}^{0-1}$ | $I_{14}^{0-1}$ | $I_{15}^{0-1}$ |
|------|---------------|---------------|---------------|
| 00   | 0.587         | 0.366         | 0.339         |
| 01   | 0.361         | 0.620         | 0.661         |
| 10   | 0.490         | 0.510         | 0.590         |
| 11   |               |               |               |

| Part | $I_{16}^{0-1}$ | $I_{17}^{0-1}$ | $I_{18}^{0-1}$ |
|------|---------------|---------------|---------------|
| 00   | 0.396         | 0.604         | 0.598         |
| 01   | 0.459         | 0.541         | 0.454         |
| 11   | 0.596         | 0.404         | 0.380         |
| 12   |               |               |               |

| Part | $I_{19}^{0-1}$ | $I_{20}^{0-1}$ | $I_{21}^{0-1}$ |
|------|---------------|---------------|---------------|
| 00   | 0.503         | 0.497         | 0.489         |
| 01   | 0.590         | 0.410         | 0.546         |
| 11   | 0.411         | 0.589         | 0.357         |
| 12   |               |               |               |

| Part | $I_{22}^{0-1}$ | $I_{23}^{0-1}$ |               |
|------|---------------|---------------|---------------|
| 00   | 0.382         | 0.618         |               |
| 01   | 0.542         | 0.458         |               |
| 10   | 0.446         | 0.554         |               |
| 11   | 0.692         | 0.308         |               |