A Mixed Integer Linear Program to Compress Transition Probability Matrices in Markov Chain Bootstrapping

Roy Cerqueti[1], Paolo Falbo[2], Cristian Pelizzari[2], Federica Ricca[3], Andrea Scozzari[4]

[1]Università degli Studi di Macerata, roy.cerqueti@unimc.it
[2]Università degli Studi di Brescia,
falbo@eco.unibs.it, pelizcri@eco.unibs.it
[3]Sapienza Università di Roma, federica.ricca@uniroma1.it
[4]Università degli Studi Niccolò Cusano - Telematica Roma,
andrea.scozzari@unicusano.it

Abstract

Bootstrapping time series is one of the most acknowledged tools to study the statistical properties of an evolutive phenomenon. An important class of bootstrapping methods is based on the assumption that the sampled phenomenon evolves according to a Markov chain. This assumption does not apply when the process takes values in a continuous set, as frequently happens for time series related to economic and financial phenomena. In this paper we apply Markov chain theory for bootstrapping continuous-valued processes, starting from a suitable discretization of the support that provides the state space of a Markov chain of order $k \geq 1$. Even for small k, the number of rows of the transition probability matrix is generally too large and, in many practical cases, it may incorporate much more information than it is really required to replicate the phenomenon satisfactorily. The aim of this paper is to study the problem of compressing the transition probability matrix while preserving the "law" that characterizes the process generating the observed time series, in order to obtain resampled series that maintain the typical features of the observed time series. To this purpose, we formulate a partitioning problem of the set of rows of such a matrix and propose a Mixed Integer Linear Program specifically tailored for this particular problem. We also provide an empirical analysis by applying our model to the time series of Spanish and German electricity prices in given time periods, and we show that, for these medium size real-life instances, bootstrapped time series obtained by the compressed transition probability matrix reproduce the typical features of the observed ones.

Keywords: Time series bootstrapping, Mixed Integer Linear Programming, Markov chains, Transition probability matrix compression, Continuous-valued stochastic processes.

1 Introduction

After the seminal paper by Efron (1979), several developments and applications of bootstrap methods appeared in the literature. Methods following the original idea by Efron and based on re-sampling of model errors have been largely applied in Economics and Finance. The reader is referred to Freedman (1984), Freedman and Peters (1984), Efron and Tibshirani (1993) for a methodological discussion and to Brock et al. (1992), Sullivan et al. (1999) for an application to financial markets. However, the original approach of Efron suffers, in general, of model

misspecification risk and requires observations to be time independent. To overcome such limitations, nonparametric, model-free bootstrap methods have been proposed in the literature. In Bühlmann (2002) several bootstrap methods of this type are compared, such as the *block*, the *sieve*, and the *local* methods. The advantage of nonparametric, model-free methods is that they do not require the observations to be time independent: data themselves capture the dependence structure of the time series, thus relieving the researcher of the responsibility of choosing a model. Among the nonparametric bootstrap methods, a relatively recent group is based on Markov chain theory (see, e.g., Anatolyev and Vasnev, 2002). The major issue in this research direction is the estimation of the true dimension of the transition probability matrix, which, in turn, consists of estimating the relevant states and order of the process. Even if these two estimates refer to different aspects of the process, they are not independent and they have been extensively examined in the area of Information Theory to model alphabet processes (see, for example, Bühlmann and Wyner, 1999; Bühlmann, 2002). These studies concern only the problem of estimating the true order of a Markov chain.

In some real applications, the order k of a Markov process is specified in advance. On the basis of the observed series, the value of k can be rather large. Moreover, the series might range in a wide set of values. All such occurrences may cause that the number of rows in the transition probability matrix is too large w.r.t. the information necessary to model the process (i.e., several rows are redundant). In this case, one would like to reduce the number of rows, focusing in particular on those which do not "carry" enough information.

Reducing the dimension of the transition probability matrix means, in general, to opportunely aggregate the states of the related Markov chain into "super-states", in order to get a reduced model that still well represents the original process. Therefore, model reduction in Markov chains basically consists of finding a suitable partition of the state space and the corresponding aggregation of the rows and/or the columns of the transition probability matrix.

Different reduction methodologies have been proposed in the literature to fulfil the different requirements ruled by the specific application under study.

The aggregation rules strongly depend on the order of the considered Markov chain. In the classical framework of Markov chains of order 1, the aggregation should be performed to obtain states of a new Markov chain of order 1. For example, Zhu et al. (2002), who propose an efficient link prediction procedure for the design of adaptive web sites, base their reduction of the transition probability matrix on a "similarity" criterion and apply an algorithm devised by Spears (1998).

In the literature, a well-known Markov chain reduction approach is *lumping*, in which the states of a Markov chain must be aggregated guaranteeing that the transition probabilities of the reduced Markov chain satisfy the Chapman-Kolmogorov equations. A lumping approach is followed, for example, by Deng et al. (2011), who study the problem of reducing a Markov chain of order 1 by minimizing the "Kullback-Leibler" divergence rate between the original and the reduced transition probability matrix. The Kullback-Leibler divergence rate, which measures the loss of information when replacing the original matrix with the reduced matrix (see Rached et al., 2004), is the objective function of their optimization model. Since it is a nonlinear function, their problem is hard to solve so that they must use a heuristic solution procedure.

For lumping, we refer the reader to the introductory paper of Burke and Rosenblatt (1958) and also to Kemeny and Snell (1976), Barr and Thomas (1977), Abdel-Moneim and Leysieffer (1984), and White et al. (2000). More recently, Thomas (2010) has provided an encyclopedic overview on lumping, along with the related main theoretical results. A different aggregation approach for the Markov chain reduction problem relies on *spectral graph theory*. The basic idea is searching for an optimal partition of a graph into subgraphs by implementing a procedure

grounded on the eigenvectors of a special pre-specified matrix. For excellent surveys and bibliographic reviews on spectral graph theory, see Chung (1997). In spectral theory, graphs and Markov chains can be linked in a natural way: indeed, a Markov chain can be obtained as a random walk on the vertices of a given graph, and transition probabilities are the normalized weights assigned to the edges. The partition of the graph is then translated into the partition of the state space of the Markov chain. Spectral Markov theory generalizes the lumping framework and it has been the scientific ground of some important studies. Verma and Meila (2003) compare different clustering three-stage algorithms based on matrices eigenvectors. In particular, they introduce a preprocessing phase where the matrix of interest M is normalized to $\tilde{M}$, moving *de facto* from graphs to a Markov chain with transition probability matrix $\tilde{M}$. Meila and Xu (2004) discuss the explicit relationships between graphs and Markov chains and introduce a new spectral clustering-based optimization problem, that falls into the considerably difficult class of NP-hard problems.

When dealing with an order $k > 1$, the constitutive properties regarding the Markov chains of order 1 stand no longer in force. In particular, the rows of the transition probability matrix are null when the related k-state occurs with zero probability; moreover, the Chapman-Kolmogorov Theorem cannot be properly stated. This shows that the lumping framework does not fit *any* Markov chain reduction problem.

When model reduction in Markov chain is formulated as a partitioning problem, one has to take into account that, in general, the computational complexity of the problem forces to the use of heuristic procedures. For example, a heuristic approach is adopted by Cerqueti et al. (2013), who propose a Tabu Search procedure for the problem of simultaneously partitioning the state space of a continuous-valued process and identifying the order of the process. Their approach starts by modeling a continuous-valued process as "Markov switching regimes" (see, e.g., Hamilton, 1996; Jeanne and Masson, 2000; Hamilton, 2005), and identifying the "relevant" states defined by the thresholds that mark the passage from a regime to another (see also Cerqueti et al., 2010). Their idea is to partition the set of the rows of the transition probability matrix (associated to the discretized version of the process) to reduce its size and yet minimizing the information loss that may derive from matrix compression. The rows belonging to the same class of the partition are "similar" according to a suitable distance indicator. The computational effort for solving such problem increases exponentially w.r.t. the order of the Markov chain, and follows the law of the Bell numbers w.r.t. the number of states included in the initial discretization. Even if the authors provide a pre-processing that drastically reduces the size of the solution space[1], the partitioning problem remains intractable and a heuristic procedure for its solution is required.

In this paper, we are concerned with Markov chains of order $k \geq 1$, our interest being mainly for the case $k > 1$. Aggregation involves only the rows of the transition probability matrix, i.e., the k-states of the process, while the columns of the matrix remain associated to the original (single) states. Therefore, in general, we have a Markov chain defined by a rectangular transition probability matrix, and also the aggregated matrix is rectangular. This is a main distinctive feature of our model that requires an approach different from lumping, in which the transition probability matrix of the given Markov chain is required to be square, and both its rows and columns must be aggregated producing a reduced Markov chain with a square transition probability matrix (whose size is reduced w.r.t. the size of the original matrix).

[1]The size of the solution space is reduced from $B\left(n^k\right)$ to $[B\left(n\right)]^k$, where $B\left(n\right)$ is the n-th Bell number, i.e. the number of partitions of the set of n states of a Markov chain of order k.

Hence, the grounding assumption on the order k of the Markov chain prevents us to use lumping techniques and requires an *ad hoc* aggregation procedure.

We assume k fixed and focus on the problem of optimally partitioning the set of rows of the transition probability matrix by proposing an exact approach through a Mixed Integer Linear Programming formulation (MILP). The objective function relies on a dissimilarity measure between pairs of rows, and a special constraint is introduced to guarantee that the bootstrapped series are sufficiently diversified. The constraint is introduced to fulfil the basic requirement of diversification of outputs when applying bootstrapping methods. In our application, relying on the partition provided by our MILP, we are able to generate bootstrapped series that maintain the statistical properties of the original sample without being its exact replication.

The proposed MILP can be solved via common optimization packages, like the most recent GUROBI optimization solver. This allows to efficiently solve problems where the number of rows of the matrix ranges from 40 to 60, which is the typical size of certain economic and financial phenomena. As a real application of our method we consider the problem of bootstrapping series of the Spanish and German electricity prices observed daily for 6 and 7 consecutive years, respectively. We point out that such time series are considered in the literature as "hard cases" to replicate, because of their nonlinear dependence structure.

The present paper contributes to the literature on Markov chain bootstrapping by providing a nonparametric, model-free approach, which is particularly suited for time series with nonlinear dependence of data, as those related to evolutive phenomena in electricity markets. Our work is focused on the original purpose of bootstrapping, i.e., to improve parameter estimation obtained from a sample of observed data. We point out that bootstrapping consists of resampling the original data (with replacement), and, therefore, it cannot be immediately used to simulate out-of-sample trajectories, as it generally happens for other simulation methods (e.g., Monte Carlo, neural networks, etc.). This kind of analysis is out of the scope of our application. To conclude, we point out that in the literature there exists a variety of contributions dealing with Mixed Integer Linear Programming for optimal (constrained) partitioning problems of different nature (see, e.g., Mueller and Kramer, 2010, Saglam et al., 2006, and the references therein). It must be notice that, even if in this paper we adopt standard modeling techniques (such as, for example, the one to linearize the objective function), our MILP provides a new contribution in Markov chain model reduction under different aspects. It is specifically designed for the real-life problems related to Markov chain bootstrapping. In particular, the objective function can be viewed as an innovative way for measuring information loss when the Markov chain reduction problem consists of aggregating only the rows of the transition probability matrix. In addition, it provides tools for controlling diversification of the bootstrapped series, and proposes a bi-objective approach that is able to find aggregations of rows with a good compromise between the number of "super-states" and the dissimilarity between rows aggregated in the same super-state.

The paper is organized as follows. In Section 2, we discuss in detail Markov chain bootstrapping and describe the Mixed Integer Linear Program proposed for partitioning the rows of a transition probability matrix. Section 3 is dedicated to the analysis of two real-life problems: after the description of the data sets, we illustrate our results for the reduction of the transition probability matrix and then provide a comparative statistical analysis of the bootstrapped time series with respect to the observed ones. In Section 4, we collect some final thoughts and concluding remarks.

2 Markov chain bootstrapping based on Mixed Integer Linear Programming

In this section we introduce the Markov chain resampling problem and propose an optimization approach to solve it. The basic idea underlying the problem is first introduced, together with the corresponding notation and definitions.
Consider a time-varying phenomenon. Under the hypothesis that the phenomenon evolves according to a k-th order Markov chain $(X(t), t \geq 0)$, with $k \geq 1$, our aim is to resample it through a bootstrapping method. The estimation of the transition probability matrix associated to $(X(t), t \geq 0)$ relies on an available time-ordered sample of observations of the investigated phenomenon.
In resampling procedures, two aims are pursued which are somehow conflicting: on the one hand, the exact replication of the sample at each simulation should be avoided (*diversification* or *multiplicity* criterion); on the other hand, the statistical properties of the original sample should be reproduced in the replications as much as possible (*similarity* criterion). In order to generate a bootstrapped series satisfying the two above properties, we propose a method based on the aggregation of the rows of the transition probability matrix of $(X(t), t \geq 0)$. The problem is formulated as a partition problem of the set of rows of the transition probability matrix, and it is modeled as a Mixed Integer Linear Program (MILP) in order to solve it exactly.

2.1 Notation and definitions

Assume that the possible realizations of an evolutive phenomenon vary in an interval $[\alpha, \beta] \subseteq \mathbb{R}$. Consider the following series of time-ordered realizations of the phenomenon observed in the time horizon $(1, \ldots, t, \ldots, T)$:

$$e(T) = (e_1, \ldots, e_t, \ldots, e_T). \tag{1}$$

We also refer to T as the *length* of the observed series. Let $[b_0, b_1), \ldots, [b_{z-1}, b_z), \ldots, [b_{n-1}, b_n]$ be a partition of $[\alpha, \beta]$ into n intervals such that $b_0 = \alpha$ and $b_n = \beta$. In order to simplify the notation, we refer to the z-th interval $[b_{z-1}, b_z)$ as α_z. This is only a matter of notation, and α_z is not meant to be a single point replacing the whole interval $[b_{z-1}, b_z)$. By definition, one has:

$$\begin{cases} \alpha_i \cap \alpha_j = \emptyset, \ i, j = 1, \ldots, n, \ i \neq j \\ \bigcup_{z=1}^{n} \alpha_z = [\alpha, \beta]. \end{cases}$$

We collect the elements of such partition in a set $A = \{\alpha_1, \ldots, \alpha_z, \ldots, \alpha_n\}$ so that A represents the set of all possible states for the time-ordered series under analysis. States in the set A are called *theoretical*. We will refer to the above partition of $[\alpha, \beta]$ as the *initial partition*. A *state* corresponds to one of the intervals α_z, $z = 1, \ldots, n$, which are then called the *initial intervals* or *initial states*. The choice of adopting the term "state" is taken to refer explicitly to the Markov chain framework described later in this section. Relying on the above discretization, one has that, for each observed value e_t (point), there is a unique α_z (interval) in A such that $e_t \in \alpha_z$. As a consequence, a time series as (1) will be represented as a sequence of states of a Markov chain $a(T) = (a_1, \ldots, a_t, \ldots, a_T)$, where $a_t = \alpha_z$ whenever $e_t \in \alpha_z$, for $z = 1, \ldots, n$.
Given the set A of the initial states, for a fixed k the cartesian product A^k collects all the k-tuples formed by states in A (*theoretical k-states*). We denote a generic theoretical k-state by $\alpha^h(k) = (\alpha_{h_k}, \alpha_{h_{k-1}}, \ldots, \alpha_{h_1})$, with $h_w \in \{1, \ldots, |A|\}$ and $w = 1, \ldots, k$.

We denote a sequence of $k < T$ consecutive elements extracted from $e(T)$ as:

$$e(\tilde{t},k) = (e_{\tilde{t}}, \ldots, e_{\tilde{t}+j}, \ldots, e_{\tilde{t}+k-1}) \quad \text{for some } \tilde{t} \in \{1, \ldots, T-k+1\}. \tag{2}$$

Notice that in formula (2) $e(\tilde{t},k)$ depends on two parameters, namely, the *length* k of the above sequence and $\tilde{t}$, which identifies the starting time of the sequence. An *observed k-state* (or simply *k-state*) $a(\tilde{t},k)$ is the sequence of the k observed states corresponding to $e(\tilde{t},k)$. More precisely, $a(\tilde{t},k) = (a_{\tilde{t}}, \ldots, a_{\tilde{t}+j}, \ldots, a_{\tilde{t}+k-1})$ is obtained by setting $a_t = \alpha_z$ when $e_t \in \alpha_z$, for $t \in [\tilde{t}, \tilde{t}+k-1]$.

Let $\mathcal{O}_k$ denote the set of observed $a(\tilde{t},k)$, that is

$$\mathcal{O}_k = \Big\{ a(\tilde{t},k),\ \tilde{t} \in \{1, \ldots, T-k+1\} \Big\},$$

for which one has $|\mathcal{O}_k| = T-k+1$.
Assume that the evolutive phenomenon is modeled as a Markov chain of order k, $(X(t), t \geq 0)$, with state space A. Then, the k-th order transition probability matrix is estimated by using the definition of Ching et al. (2008), which involves the estimation of the empirical frequencies.

We introduce the transition probability for the process to reach state α_z immediately after the sequence of states of $\alpha^h(k)$, that is:

$$\mu^{(k)}_{\alpha^h(k),\alpha_z} = Prob(X(t) = \alpha_z | X(t-1) = \alpha_{h_1}, \ldots, X(t-k) = \alpha_{h_k}). \tag{3}$$

Let us denote by M the k-th order transition probability matrix, with $|A^k| = n^k$ rows and $|A| = n$ columns. To avoid a cumbersome notation, we set $n^k = m$. Matrix M is built starting from the observed sample $e(T)$ and its generic entry (h,z) contains the value $\mu^{(k)}_{\alpha^h(k),\alpha_z}$ corresponding to k-state $\alpha^h(k)$ in A^k and state α_z in A.

According to the approach in Ching et al. (2008), we estimate $\mu^{(k)}_{\alpha^h(k),\alpha_z}$ as follows[2]:

$$\mu^{(k)}_{\alpha^h(k),\alpha_z} = \begin{cases} \frac{\Lambda(\alpha^h(k),\alpha_z)}{\sum_{j=1}^{n} \Lambda(\alpha^h(k),\alpha_j)} & \text{if } \sum_{j=1}^{n} \Lambda(\alpha^h(k),\alpha_j) \neq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where

$$\Lambda(\alpha^h(k),\alpha_z) = \Big| \bigcup_{\tilde{t}=1}^{T-k} \{ a(\tilde{t},k+1) \in \mathcal{O}_{k+1} \ :\ a(\tilde{t},k+1) \equiv (\alpha^h(k),\alpha_z) \} \Big| \tag{5}$$

is the number of times that a sequence $(\alpha^h(k),\alpha_z)$ has been observed in the sample.
For a given k, each row of the transition probability matrix M contains the transition probabilities from a k-state $\alpha^h(k) \in A^k$ to the states $\alpha_z \in A$. Formula (4) is used to compute M on the basis of the original sample. Notice that, when a k-state is observed only once in the original series $e(T)$, the estimation of its transition probabilities is equal to 1 for the unique state to which the corresponding trajectory evolved, while it is equal to 0 for all the other states. We refer to this case as *deterministic* k-states, and the corresponding rows of the transition probability matrix are called *deterministic* rows. Each row not completely filled with zeros reports the estimated probabilities from an observed k-state to the observed states, which substantially

[2]For the sake of simplicity, we avoid introducing here a specific notation for the estimates of the probabilities.

means that only sequences $\alpha^h(k)$ such that $\alpha^h(k) \equiv a(\tilde{t}, k)$ for some $\tilde{t} \in \{1, \ldots, T-k+1\}$ and states $\alpha_z \equiv a_t$ for some $t \in \{1, \ldots, T\}$ are considered.

To introduce our MILP model, in the following we indifferently refer to $\alpha^u(k)$ and $\alpha^v(k)$ of A^k and to the corresponding rows, u and v respectively, of matrix M. Similarly, we will indifferently refer to the set A^k and to the corresponding set of rows V of matrix M.
The *dissimilarity* between two rows u and v of M can be measured as follows:

$$d_{uv} = \sum_{z=1}^{n} \left| \mu^{(k)}_{\alpha^u(k),\alpha_z} - \mu^{(k)}_{\alpha^v(k),\alpha_z} \right|. \tag{6}$$

It is, in fact, a distance indicator between rows of the matrix M which takes values in the interval $[0, 2]$ (see Cerqueti et al., 2010). As we will argue below, the dissimilarity d_{uv} can be viewed as a proxy for the "cost" of putting k-states $\alpha^u(k)$ and $\alpha^v(k)$ in the same class of a partition, and, in fact, the objective function of our optimization model is based on it.

2.2 An optimization approach to the aggregation of the rows of a transition probability matrix

Let us introduce our partitioning problem of the set V of the rows of the transition probability matrix M. Consider a partition of V into q classes, $\pi = \{C_1, \ldots, C_p, \ldots, C_q\}$, with $1 \leq q \leq m$, and let Π be the set of all possible such partitions.
The *diameter of a class* C_p is defined as the maximum dissimilarity between two elements in C_p, and it is denoted by $\delta(C_p)$:

$$\delta(C_p) = \max_{u,v \in C_p} d_{uv}. \tag{7}$$

The *diameter of a partition* $\pi \in \Pi$ is defined as

$$\Delta(\pi) = \max_{C_p \in \pi} \delta(C_p). \tag{8}$$

Remark 1 *According to formula (8), and since* $d_{uv} \in [0, 2]$, $\forall\, u, v \in V$, $\Delta(\pi) \in [0, 2]$ *for each partition* $\pi \in \Pi$. *In the special case of the partition* π *in which each class includes only one element, which we call singleton partition, one has* $\Delta(\pi) = 0$. *Notice that, given* π, π' *in* Π, π *is a* refinement *of* π' *when* π *has the same classes of* π' *with the exception of some classes which in* π *are further partitioned. When* π *is a refinement of* π', *one has* $\Delta(\pi) \leq \Delta(\pi')$. *The less refined partition is the one composed by only one class grouping all the elements of* V, *which therefore corresponds to the partition with the maximum diameter, call it* $\Delta_{\max}$, *among all the partitions* $\pi \in \Pi$. *Finally, notice that* $\Delta_{\max} \leq 2$ *and, depending on the data set under study, it may, or may not, reach this upper bound.*

For a given $\pi \in \Pi$, we denote by $C_{\bar{p}}$ a class for which $\delta(C_{\bar{p}}) = \Delta(\pi)$, that is:

$$C_{\bar{p}} \in \operatorname{argmax}\{\delta(C_p), C_p \in \pi\}.$$

In our partitioning problem we want to minimize both the number of classes and the diameter of the partition. These two objectives are obviously in conflict, since the diameter of a partition tends to increase when the number of classes decreases and vice versa. In fact, we have a bi-objective problem which we handle by minimizing the diameter of the partition while controlling

for the number of its classes. To this purpose, we introduce in the model a parameter $\gamma \geq 0$ that limits the value of the diameter $\Delta(\pi)$ from below, implying a constraint on the cardinality of the partition. By means of γ we implement a control in the resampling procedure to meet the multiplicity criterion, i.e., to avoid the exact replication of the original sample. For a fixed value $\gamma \geq 0$ and an integer $1 \leq q \leq m$, we formulate the following partitioning problem:

among all the possible partitions $\pi \in \Pi$ with at most q classes and such that $\delta(C_{\bar{p}})$ is at least equal to γ, find a partition π^* that minimizes $\Delta(\pi)$.

The problem can be formulated as a Mixed Integer Linear Program, as illustrated in the following.

To formalize the optimization model, let us observe that a partition $\pi = \{C_1, \ldots, C_p, \ldots, C_q\}$, with $1 \leq q \leq m$, can be equivalently written as $\pi = \{C_1, \ldots, C_p, \ldots, C_m\}$, with

$$\begin{cases} C_p \neq \emptyset, & \text{for } p = 1, \ldots, q \\ C_p = \emptyset, & \text{for } p = q+1, \ldots, m \end{cases} . \tag{9}$$

Definition 1 *Given a partition of V, $\pi = \{C_1, \ldots, C_p, \ldots, C_m\}$, the class C_p is active if $C_p \neq \emptyset$.*

Let $v \in V$ be a generic row of M. We introduce the binary variables x_{vp} and y_p, with $p = 1, \ldots, m$, such that:

$$x_{vp} = \begin{cases} 1, & \text{if row } v \text{ belongs to class } C_p \\ 0, & \text{otherwise} \end{cases} ,$$

$$y_p = \begin{cases} 1, & \text{if class } C_p \text{ is active} \\ 0, & \text{otherwise} \end{cases} .$$

Consider now $u, v \in V$. It is easy to see that the product $x_{up} \cdot x_{vp}$ is equal to 1 if and only if the rows u and v belong to the same class C_p of the partition:

$$x_{up} \cdot x_{vp} = \begin{cases} 1, & \text{if both } u \text{ and } v \text{ belong to class } C_p \\ 0, & \text{otherwise.} \end{cases} .$$

Hence, the *cost* of assigning both u and v to class C_p is given by:

$$d_{uv} \cdot x_{up} \cdot x_{vp}.$$

With this notation, the diameter of class C_p can be rewritten as $\delta(C_p) = \max_{u,v \in V} \{d_{uv} \cdot x_{up} \cdot x_{vp}\}$. The problem of finding the partition of the set of rows V into *at most* q classes that minimizes the maximum diameter of a class can be formulated as follows:

$$\min_{\pi \in \Pi} \max_{C_p \in \pi} \max_{u,v \in V} \{d_{uv} \cdot x_{up} \cdot x_{vp}\}$$

$$\begin{aligned} &(1) && \sum_{p=1}^{m} x_{vp} = 1 && \forall\, v \in V \\ &(2) && x_{vp} \leq y_p && \forall\, v \in V, \quad p = 1, \ldots, m \\ &(3) && \sum_{p=1}^{m} y_p \leq q \\ & && x_{vp}, y_p \in \{0, 1\} && \forall\, v \in V, \quad p = 1, \ldots, m. \end{aligned} \tag{10}$$

Notice that in the above formulation we bound the number of classes of the partition by q. Thus, varying q in $\{1, \ldots, m\}$ allows us to tackle the bi-objective problem as a sequence of m single-objective ones. In problem (10) the first set of constraints states that each row must belong to only one class of the partition. According to constraints (2), a row can belong to a class C_p only if C_p is active. Constraint (3) provides an upper bound on the number of classes of π. Problem (10) is an integer nonlinear program which can be linearized by introducing additional *association* variables w_{uvp}, with $0 \leq w_{uvp} \leq 1$:

$$w_{uvp} = x_{up} \cdot x_{vp} = \begin{cases} 1, & \text{if both } u \text{ and } v \text{ belong to class } C_p \\ 0, & \text{otherwise} \end{cases},$$

and the following set of constraints:

$$\begin{aligned} w_{uvp} &\leq x_{up} && \forall\, u, v \in V, \quad p = 1, \ldots, m \\ w_{uvp} &\leq x_{vp} && \forall\, u, v \in V, \quad p = 1, \ldots, m \\ w_{uvp} &\geq x_{up} + x_{vp} - 1 && \forall\, u, v \in V, \quad p = 1, \ldots, m. \end{aligned} \tag{11}$$

Thus, if both u and v belong to class C_p one has $x_{up} = x_{vp} = 1$, and therefore $w_{uvp} = 1$. On the other hand, if either $x_{up} = 0$ or $x_{vp} = 0$, one has $w_{uvp} = 0$, too. This also implies that the bounds $w_{uvp} \leq 1$ are always satisfied, and hence they need not to be included explicitly in the problem formulation. Notice that constraints (11) guarantee that variables w_{uvp} assume only values 0 or 1, and, therefore, they can be introduced in the model as real variables.
The objective function can be written in terms of the new variables as follows:

$$\min_{\pi \in \Pi} \quad \max_{C_p \in \pi} \quad \max_{u,v \in V} \{d_{uv} \cdot w_{uvp}\}.$$

Then, the objective function can be linearized introducing a new variable d that replaces $\max\limits_{u,v \in V} \{d_{uv} \cdot w_{uvp}\}$ and adding the following set of constraints:

$$d_{uv} \cdot w_{uvp} \leq d, \quad \forall\, u, v \in V, \quad p = 1, \ldots, m.$$

Let us notice that the above model does not prevent the optimal objective function value to become very small (it may even turn out to be 0). We recall that the core of the bootstrapping method is to generate resamplings of the observed series $e(T)$ which should be sufficiently "similar" to $e(T)$, in order to guarantee that they can be seen as different realizations of the same phenomenon that generated $e(T)$. However, they have to be also sufficiently "diversified" from $e(T)$, in order to have some degree of "variability" among them. The optimal partition applied to M should then be structured so as to guarantee these requirements as much as possible. If, on the one hand, similarity is pursued in the model via the minimization of the diameter, on the other hand, diversification can be controlled by imposing the diameter to be not smaller than a prefixed (positive) threshold γ. This additional constraint prevents from choosing the trivial singleton partition, where each class is formed by a single element. This, in particular, happens if we choose $\gamma = 0$. In fact, for this partition, the value of the objective function is zero, but no aggregation of the rows of M is actually performed. In order to formalize the additional condition related to the threshold γ, we introduce the binary variables t^{γ}_{uvp} defined as follows:

$$t^{\gamma}_{uvp} = \begin{cases} 1, & \text{if } d_{uv} \cdot w_{uvp} \geq \gamma \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, u, v \in V, \quad p = 1, \ldots, m.$$

Adding further suitable constraints involving t^{γ}_{uvp} and γ, we obtain the MILP representing our specific partitioning problem:

$$\min d$$

$$\begin{array}{lcl}
(1) & \sum_{p=1}^{m} x_{vp} = 1 & \forall\, v \in V \\
(2) & x_{vp} \leq y_p & \forall\, v \in V, \quad p = 1, \ldots, m \\
(3) & \sum_{p=1}^{m} y_p \leq q & \\
 & & \\
(4) & w_{uvp} \leq x_{up} & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
(5) & w_{uvp} \leq x_{vp} & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
(6) & w_{uvp} \geq x_{up} + x_{vp} - 1 & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
 & & \\
(7) & d_{uv} \cdot w_{uvp} \leq d & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
(8) & d_{uv} \cdot w_{uvp} \geq t^{\gamma}_{uvp} \cdot \gamma & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
 & & \\
(9) & \sum_{p=1}^{m} \sum_{u,v \in V} t^{\gamma}_{uvp} \geq 1 & \\
 & & \\
 & x_{vp},\ y_p \in \{0,1\},\ t^{\gamma}_{uvp} \in \{0,1\} & \forall\, u, v \in V, \quad p = 1, \ldots, m \\
 & & \\
 & d \geq 0,\ w_{uvp} \geq 0 & \forall\, u, v \in V, \quad p = 1, \ldots, m.
\end{array} \tag{12}$$

In the above model, constraint (8) describes the relation between the variables w and the new variables t^{γ}. Constraint (9) forces at least one variable t^{γ} to be equal to 1, thus guaranteeing that the diameter of the optimal partition is at least γ.

Cerqueti et al. (2010) discuss the coherence of some distance indicators to information-type criteria advanced by Kolmogorov (1965). According to Remark 1, the dissimilarity measure used in the objective function of model (12) respects the conditions of Kolmogorov (1965). Therefore the optimal solution π^* of model (12) ensures that the "loss of information" (about the evolution properties of the process) is minimized for any given level γ fixed to meet the multiplicity criterion. In other words, the information efficiency of π^* implies that, although generated through a coarser structure of the information in the aggregated transition probability matrix M^*, a resampled series maintains simultaneously a good statistical similarity w.r.t. the original sample and a satisfactory diversification from other resampled series (multiplicity). The bootstrapping method adopted in this paper has been advanced in Cerqueti et al. (2013), to which we refer the interested reader for further details. Here the procedure takes as inputs the optimal partition π^* of model (12), the compressed transition probability matrix M^* associated to π^*, as well as the observed time series $e(T) = (e_1, \ldots, e_t, \ldots, e_T)$.

Besides the above theoretical arguments, the good statistical properties shown by our resampled series will be analyzed in detail in Section 3.4, where the values of several statistics computed

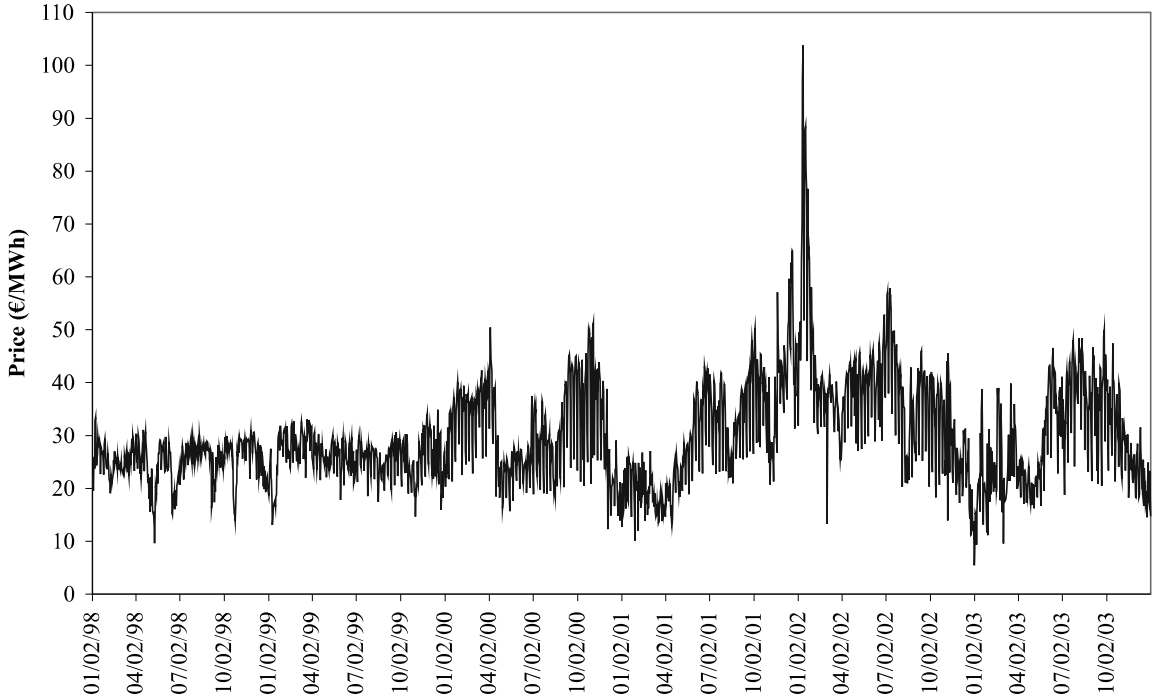

Figure 1: Spanish daily electricity prices.

on the bootstrapped series will be compared with the corresponding values calculated on the original sample.

3 Application and results

In this section we provide a case study application of the bootstrapping method we propose. It relies on data taken from the Spanish and German electricity markets as described in detail in the following. The analyzed series show several interesting features which make them difficult to treat for a bootstrapping method. Therefore they represent challenging tests for evaluating the performance of our approach.

3.1 Data

In our application we study two time series, namely the "Mibel Spanish Electric System Arithmetic Average Price" and the German "EEX Phelix Day Base Price". The series have been observed daily in the period from January 1st, 1998 to December 31st, 2003 (Spain) and from June 17th, 2000 to May 8th, 2007 (Germany). The prices are expressed in euros and refer to 1 MWh. The Spanish time series consists of $T = 2190$ observations, while for the German series we have $T = 2517$.
Figures 1 and 2 show the two time series which are characterized by the following features:

- a weekly and annual seasonality;
- a slightly positive trend;
- stochastic volatility;

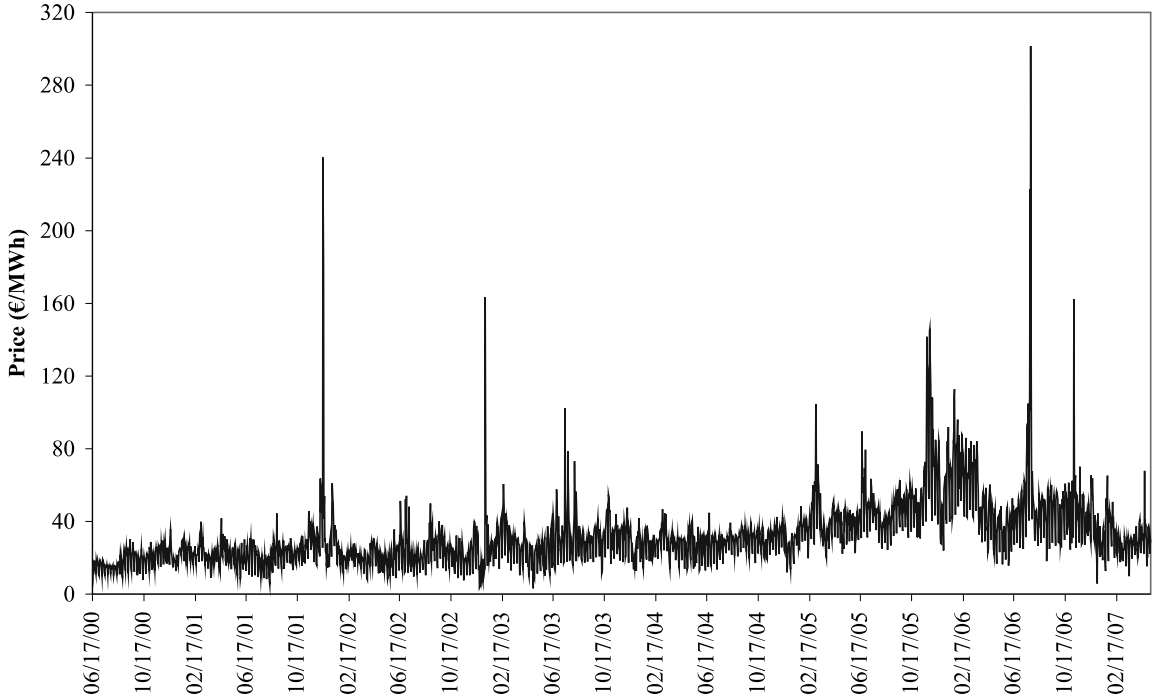

Figure 2: German daily electricity prices.

- nonlinear dependence of data;
- two clear regimes of prices: *normal trading* and *occasional spiking* periods.

Spikes are occasional, since they usually correspond to unexpected shortages on the supply side of the electricity system, or unexpected and temporary increases of the demand (e.g., sudden meteorological events driving to high consumption). Since our data consist of daily data, intra-day seasonality is neglected in this analysis. Because of the joint presence of such features, in the literature, electricity price series are usually considered as "hard to model" cases. For a review of the difficulties in modeling electricity prices and the methods developed to solve them, see, for example, Bunn (2004), Huisman and Mahieu (2003), Weron et al. (2004), and Weron (2006). Raw data prices have been removed of an (exponential) weekly seasonal component as well as of an (exponential) trend. Two main reasons justify this pre-treatment. Removing the weekly seasonality lets us free to reduce the order of the Markov chain below 7, which corresponds to a great reduction in the complexity of our problem. The removal of the trend component makes the series more stationary. Of course both components are added back to the bootstrapped series. The estimation of exponential (rather than linear) components is recommended to avoid that this removal/reintroduction process generates occasional negative prices. In Appendix A we give the details about this data treatment.

3.2 Preliminary segmentation of the support

As already explained in Subsection 2.1, to discretize our continuous-valued process, a preliminary segmentation of the support $[\alpha, \beta]$ is performed by partitioning it into n initial intervals $\alpha_1, \ldots, \alpha_z, \ldots, \alpha_n$. We notice that, if there is no theoretical evidence that the regimes of the phenomenon are a few, increasing the number of initial intervals n will help finding a segmenta-

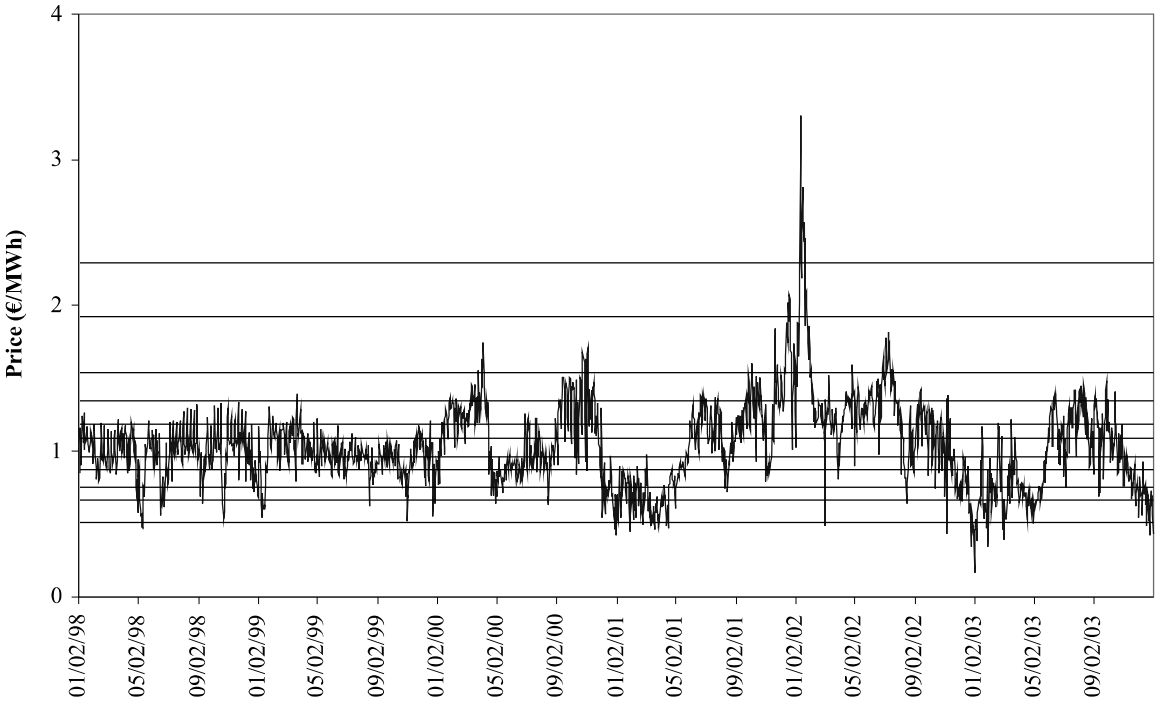

Figure 3: Spanish daily electricity prices without exponential trend and exponential weekly seasonality. Horizontal lines mark the extremes of the intervals $\alpha_1, \ldots, \alpha_{12}$.

tion whose classes approximate the relevant regimes more precisely. Unfortunately, fixing large values for n has two limitations. First, this will increase the number of rows of the transition probability matrix M, thus affecting the compression procedure on M from a complexity viewpoint. Second, this reduces the number of observations available thus affecting the statistical significance of the estimates of the transition probabilities, since this reduces the expected frequency of observations for each initial interval. Therefore, the value of n should be set in order to overcome such limitations. In our case study, we finally set $n = 12$ which, however, corresponds to an order of magnitude larger than the number of states commonly considered in the literature for the regimes of electricity prices (usually 2 or 3, see, e.g., Huisman and Mahieu, 2003).

The segmentation of the support into 12 initial states was performed through the minimum-variance clustering procedure provided in Ward (1963), after having removed trend and weekly seasonality from the original series (see Appendix A for further details). Figures 3 and 4 show the series of Spain and Germany, together with the initial 12 intervals $\alpha_1, \ldots, \alpha_{12}$ (separated by horizontal lines). Appendix B details this preliminary segmentation.
The transition probability matrices M_S and M_G of order $k = 2$, estimated for the Spanish (S) and German (G) markets, respectively[3], were computed using formula (4).
The set of observed 2-states, $\mathcal{O}_2$, is composed by 2188 elements for the Spanish instance and 2515 elements for the German one[4]. The percentage of deterministic 2-states over all the observed ones is about 1% in both cases (23 for Spain and 21 for Germany). The remaining observed 2-states are probabilistic (2165 for Spain and 2494 for Germany). In partitioning the rows of

[3]The matrices are available from the authors upon request.

[4]For both Spain and Germany the last observed 2-state is excluded from the computation of the cardinality of $\mathcal{O}_2$.

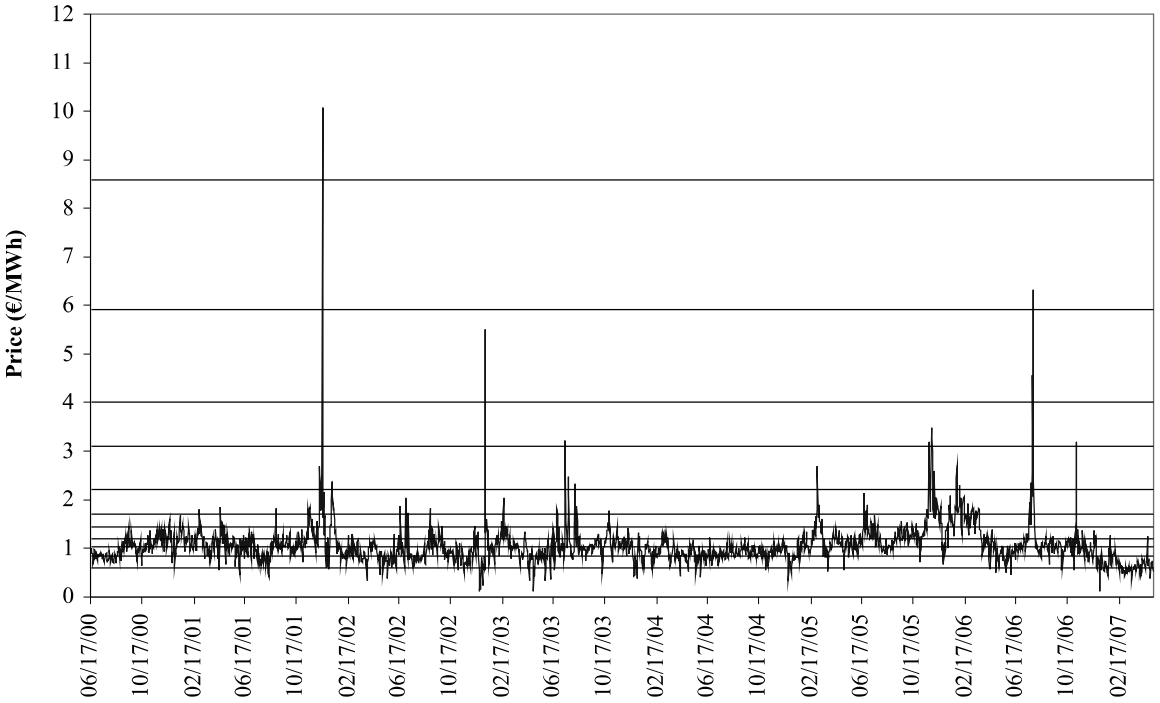

Figure 4: German daily electricity prices without exponential trend and exponential weekly seasonality. Horizontal lines mark the extremes of the intervals $\alpha_1, \ldots, \alpha_{12}$.

the transition probability matrices M_S and M_G, we actually considered only the probabilistic 2-states. On the other hand, we defined a priori a single class collecting all the non observed 2-states, while we set up as many classes as the number of observed deterministic 2-states. For $n = 12$ we obtained 63 and 44 probabilistic rows in the transition probability matrix of Spain and Germany, respectively, while the deterministic rows are 19 in both cases.

3.3 Optimization phase

The central element in model (12) is given by parameter γ. This parameter provides a lower bound to the objective function value. Fixing γ to a positive value in the model means imposing that at least one class of the partition is not formed by a single row. On the contrary, when $\gamma = 0$ no bound is imposed, so that the optimal solution always corresponds to the trivial singleton partition, provided that the maximum number of classes q in constraint (3) of model (12) is set large enough (it suffices to set $q = m$). Since the objective function value of the singleton partition is equal to 0, such partition is, obviously, the one with maximum similarity within its classes[5], and, therefore, it can be taken as a benchmark for the evaluation of the partition finally chosen to perform the bootstrapping method.
In our experimental analysis we tested a set of values for γ in model (12). After Remark 1, we know that the set of possible values for this parameter is a subset of the interval $[0, 2]$. Different values within this range lead model (12) towards different optimal solutions, since constraints (8) and (9) become stricter as the value of γ increases. We notice that, for a given γ, the same optimal value in model (12) could be obtained w.r.t. different possible values of q. In this case,

[5]Actually, in this case, no aggregation at all is performed on the rows of the transition probability matrix which remains the original one.

for a given value of γ, we always prefer the minimum number of classes which enables model (12) to satisfy constraints (8) and (9). For each fixed value of γ this provides a preferred solution characterized by two attributes: i) the optimal value of the objective function in model (12), that we denote by $d^*(\gamma)$; ii) the corresponding preferred value of q associated to such optimum, denoted by $q^*(\gamma)$.

It is clear that, a priori, we do not have any information about which, among the infinite values in $[0, 2]$, is the best choice for our aggregation purposes; we know that the extreme values of the above interval are not good choices, but, within that interval, we need to test different values of γ and evaluate the pair of attributes $(q^*(\gamma),\ d^*(\gamma))$ of the corresponding preferred solution. Indeed, different pairs of attributes correspond to different compromises between the number of classes in the final partition and the value of its diameter. In our analysis, we produce different compromises and then we compare them to choose the one that best fits the similarity and multiplicity purposes of the bootstrapping method. To this purpose, we tested a finite number of values for γ equally spaced in the interval $[0, 2]$. This is a standard practice for parameter tuning when the only available information about the parameter is the range of its possible values and there is no reason to prefer one of these values a priori.

For each tested value of γ, we applied our MILP model with increasing values for the integer parameter q, from 2 to $m - 1$. For any fixed γ, as q increases, we observed a converging process towards a stable optimal value of the objective function, i.e., the optimal value $d^*(\gamma)$ is reached for $q^*(\gamma)$, and it remains unchanged for all $q > q^*(\gamma)$, so that there is no need to choose a value $q > q^*(\gamma)$. In addition, since in the model we want to minimize the maximum diameter of a class of the partition, d, and γ is a lower bound for d, the ideal case would be attaining exactly this bound. In Figure 5 we plot the solutions obtained by applying model (12) to a data set extracted from the German time series for which the transition probability matrix has 25 rows (called German25) with different values of parameter γ. In this example, one can understand that lowering the value of γ allows the model to reach smaller values of the objective function, and that any fixed value $\gamma = 0.25, 0.5, 0.75, 1, 1.25, 1.5$ can be actually reached by the objective function from the value $q^*(\gamma)$ on. In this example we observe an ideal behavior of the model which, in all cases, manages to attain the minimum for the objective function.

As expected, for the successive values of γ, taken in their decreasing order, we have lower values of d corresponding to the same q. In particular, the figure highlights the point $(q^*(\gamma),\ d^*(\gamma))$ where each curve becomes flat. We notice that, given γ, the smallest q of the flat part corresponds to the pair $(q^*(\gamma),\ d^*(\gamma))$ that is the optimal solution value of model (12). Reducing even further q implies accepting higher values for d, that is, solutions identified by larger values of γ. As a result, the solutions obtained for a given value $\bar{\gamma}$ are contained in the envelop of the solutions corresponding to values $\gamma < \bar{\gamma}$.

The plot in Figure 5 shows that good improvements in the objective function can be obtained (for successive values of γ) by increasing q up to $q^*(\gamma) = 7$ for $\gamma = 0.5$. Going beyond this requires $q^* = 15$ for $\gamma = 0.25$. We did not analyze values of γ smaller than 0.25, since this would have implied too many classes for the optimal partition and required a too long computational time for solving model (12). In order to choose the final partition for the aggregation of the rows of the transition probability matrix, we defined an evaluation index that takes into account both the values $d^*(\gamma)$ and $q^*(\gamma)$:

$$EV(\gamma) = \frac{d^*(\gamma)}{m - q^*(\gamma)}. \tag{13}$$

The index is defined for $2 \leq q < m$, and it positively evaluates those γ that produce a small value of $d^*(\gamma)$ combined with a value of $q^*(\gamma)$ "sufficiently smaller" than its maximum ($q = m$).

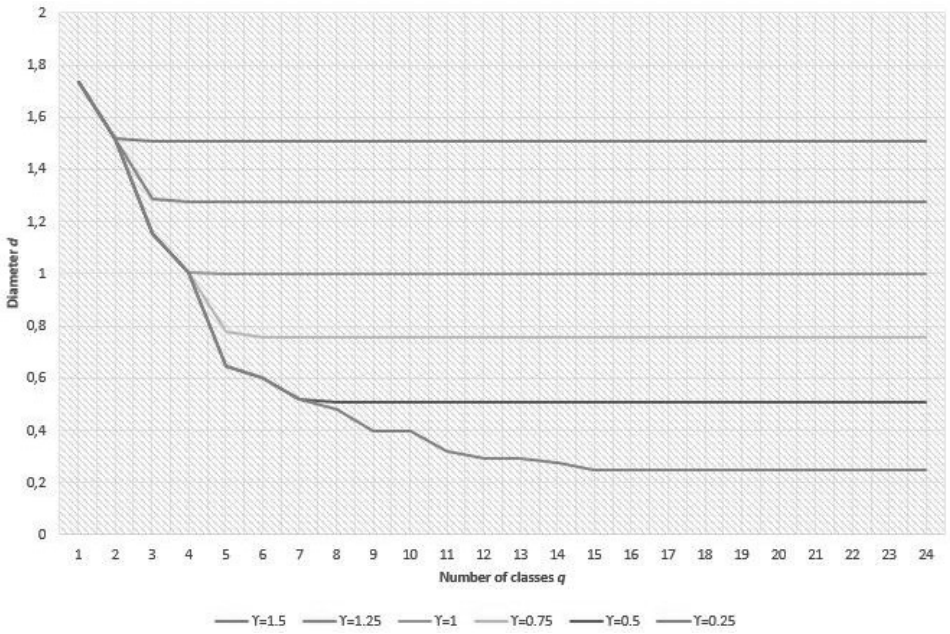

Figure 5: The solutions obtained by model (12) for $\gamma = 0.25, 0.5, 0.75, 1, 1.25, 1.5$ on the data set German25 with transition probability matrix with 25 rows.

Figure 6 plots this index for each value of γ for the data set German25. It shows that there is a systematic improvement when γ is lowered, but it becomes less evident for the last two values. We also considered the computational time required by the solver to obtain $(q^*(\gamma), d^*(\gamma))$ and decided between the values $\gamma = 0.5$ and $\gamma = 0.25$ on the basis of both measures. Unfortunately, for $\gamma = 0.25$ the computational time was always too high to justify the small improvement in d^* that it was able to produce. For example, for German25, to obtain d^* we have 512 seconds for $\gamma = 0.5$ and 8650 seconds for $\gamma = 0.25$. A similar behavior was observed for the other data sets, thus leading to the final choice of $\gamma = 0.5$, and excluding $\gamma = 0.25$ from the analysis (see Table 1 for the computational burden of the runs performed on our medium size German and Spanish data sets).

The above considerations suggest the choice of $\gamma = 0.5$ for the aggregation of the rows of the transition probability matrix that will be used in the bootstrapping method. This choice is also supported by the statistical analysis that we performed on 5000 bootstrapped series to assess the performance of the method we advance (we illustrate this analysis in detail in the following subsection). Indeed, the 5000 series obtained with $\gamma = 0.5$ reflect in a rather satisfactory manner the statistical properties of the original sample, and a very low probability of duplication over the 5000 tested cases is observed. Therefore, for the bootstrapping method, we selected the optimal partition corresponding to $\gamma = 0.5$ for each data set. We denote them by π_S^{milp} and π_G^{milp}, for the Spanish and the German data sets, respectively, while M_S^{milp} and M_G^{milp} indicate the corresponding 2^{nd} order aggregated transition probability matrices[6]. The singleton partitions obtained for $\gamma = 0$ and $q = m$ are denoted by π_S and π_G, respectively. Partition π_S^{milp} consists of 72 classes (19 of which correspond to deterministic 2-states, and one contains all the non observed 2-states). Partition π_G^{milp} has 43 classes (19 of which correspond to deterministic 2-states, and one contains all the non observed 2-states).

[6]The matrices are available from the authors upon request.

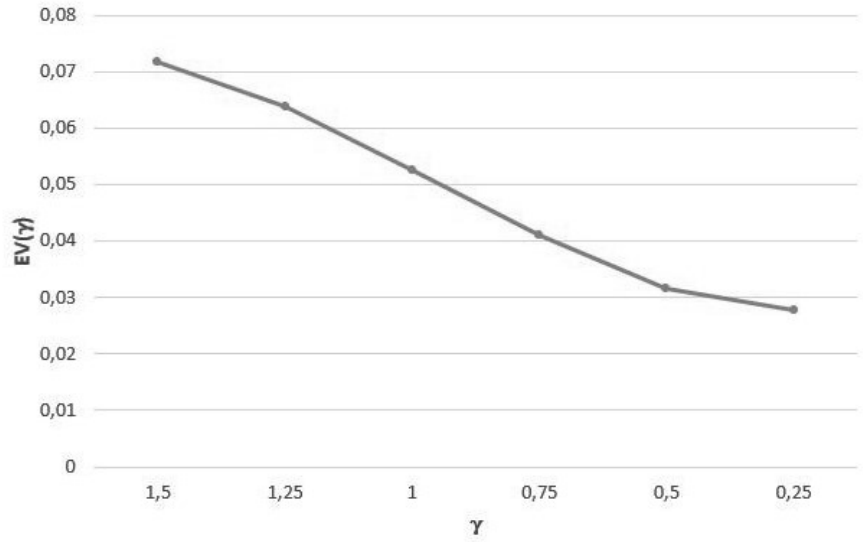

Figure 6: Index $EV(\gamma)$ for the German25 time series.

Table 1: Summary of the results of the MILP model for the two data sets. The number of classes corresponds to the smallest value of q for which the optimal objective function value can be attained. A "-" means that the time limit of 20 hours has been reached by the optimization procedure.

γ	Spain			Germany		
	$d(\gamma)$	$q(\gamma)$	Times (Sec.)	$d(\gamma)$	$q(\gamma)$	Times (Sec.)
0.50	0.50	52	-	0.50	23	-
0.75	0.75	31	16449	0.75	16	1010
1.00	1.00	12	-	1.00	9	3376
1.25	1.25	10	5128	1.25	7	1014
1.50	1.50	6	30106	1.50	5	606

Referring to our medium size data sets, Table 1 reports the values $q(\gamma)$ and $d(\gamma)$ of the solutions found for model (12) w.r.t the different values of $\gamma = 0.5, 0.75, 1, 1.25, 1.5$, together with the total time needed to find such solutions. In our tests, we imposed a limit of 20 hours on the running time: in Table 1 we put a "-" when the time limit was reached, and, in these cases, we report the best solution found so far. In all other cases, $q(\gamma)$ and $d(\gamma)$ correspond to $q^*(\gamma)$ and $d^*(\gamma)$. It must be noticed that, in spite of the premature arrest of the procedure, in all cases in which the time limit was reached the value of the best solution found was always equal to its lower bound γ. For the model solution, we used the well-known software AMPL 8, calling the solver GUROBI v6.04 for the optimization. The experiments have been carried out on a workstation equipped with an Intel Core i7-4810MQ processor with 2,28Ghz clock rate and 16 GB RAM. We also used the default GUROBIs values of $1e$-10 and $1e$-4 as absolute and relative tolerances on the gap between the lower and upper objective bound, and the gap between the MIP objective bound and incumbent solution objective, respectively.

3.4 Statistical analysis

The theoretical justification of our experiments is grounded on the definition that any trajectory generated by a stochastic process is a random sample extracted from an infinite population. Starting from this point, one should expect that the distribution of any statistics built on a population (i.e., moments, minimum, maximum, etc.) can be used to identify an acceptance range within which the same statistics, computed on a single observation ω, is expected to fall with a given probability, under the null hypothesis that ω is extracted from that population. If a large set of statistics calculated on ω falls on average into the corresponding acceptance range, then the null hypothesis cannot be rejected.

For each market (Spain and Germany), we evaluate the performance of the bootstrapping method we advance in two different scenarios:

i. a "conservative" scenario, where we consider the two partitions of singletons π_S and π_G, which represent the benchmark situation of minimum multiplicity and maximum similarity for the generation of the bootstrapped series;

ii. a "progressive" scenario, where we consider the two partitions π_S^{milp} and π_G^{milp} ($\gamma = 0.5$), which are expected to generate higher diversification and lower similarity than those of the conservative scenario, i.e., π_S and π_G.

We generated 4 sets of 5000 bootstrapped series (one set for each partition) with length $\ell = 2188$ for the Spanish case and $\ell = 2515$ for the German one. In the following, we analyze the statistical properties of the bootstrapped series in order to compare them with the ones of the original series.
Before introducing the indices, we want to give an idea of how the bootstrapped series look like. Figures 7 and 8 report two bootstrapped series, one for the Spanish market and one for the German instance, based on partitions π_S^{milp} and π_G^{milp}. Here, the bootstrapped series include the exponential trend and the exponential weekly seasonality initially removed from the original samples (see Appendix A). Each value of the series is classified as deterministic (thick mark) or probabilistic (thin mark).
We can make the following considerations:

- both the bootstrapped series in Figures 7 and 8 reproduce the spikes observed in the original series (see, Figures 1 and 2);

- also the normal trading regime appears satisfactorily reproduced. Indeed, the two series take values in ranges strongly overlapping those of the original one;

- weekly seasonality is clearly distinguishable, as well as, a slightly positive trend;

- the frequency of deterministic values is 1% for both the Spanish and German cases, similarly to the values observed in the original series. The bootstrapping method probabilistically reproduces sequences of the original series and occasionally such segments are interleaved by deterministically chosen values. Let us observe at this point the key advantage of the Markov chain bootstrapping method: depending on the different probability distributions associated to each conditioning event, the resampling method switches from deterministic (e.g., in the case of spikes) to highly unpredictable.

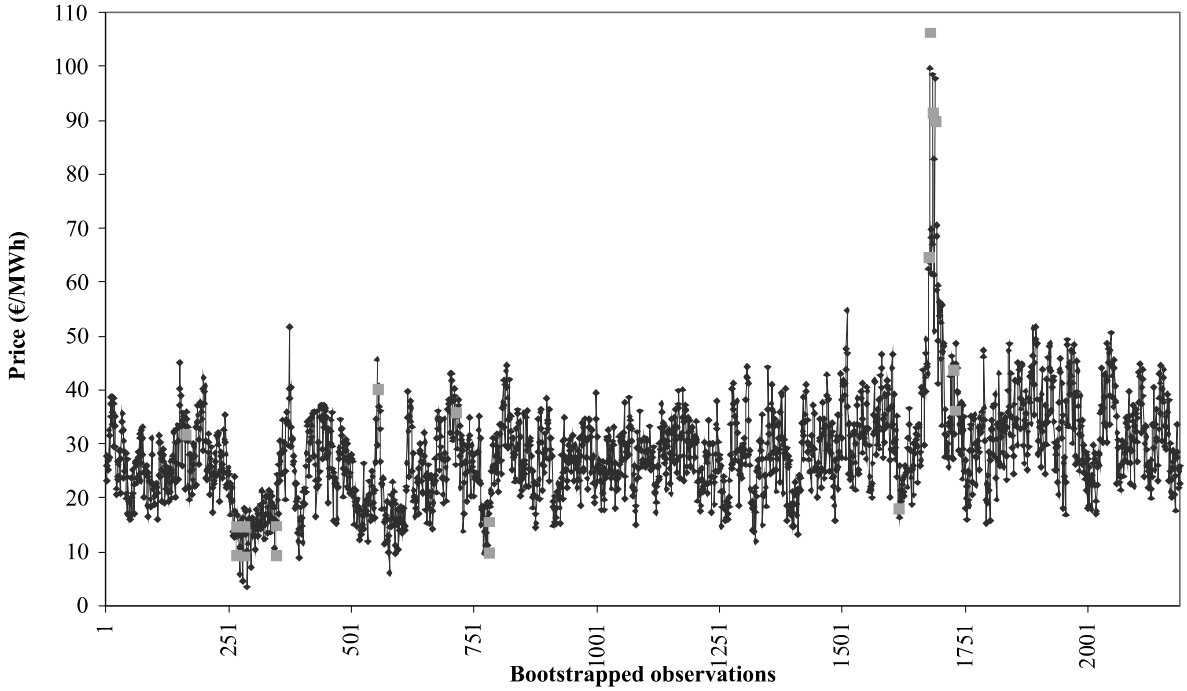

Figure 7: Bootstrapped Spanish electricity prices. The thin mark indicates that the selected value is the ending point of an observed probabilistic 2-state, the thick mark indicates that the bootstrapped values are the last points of observed deterministic 2-states.

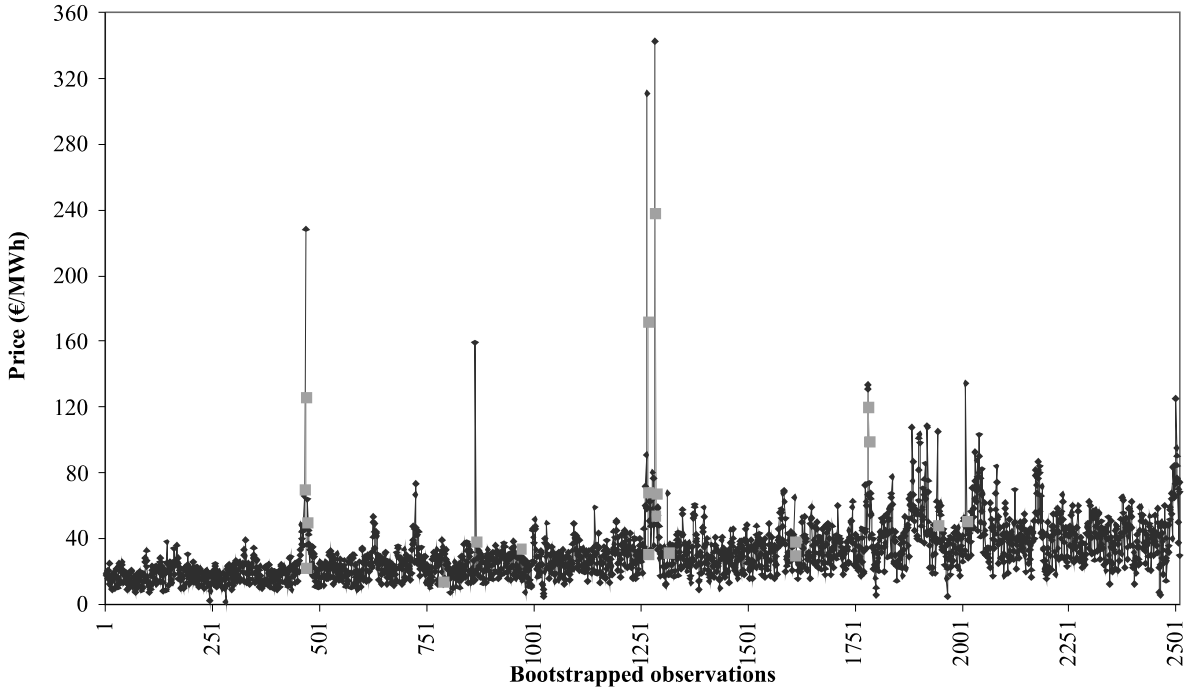

Figure 8: Bootstrapped German electricity prices. The thin mark indicates that the selected value is the ending point of an observed probabilistic 2-state, the thick mark indicates that the bootstrapped values are the last points of observed deterministic 2-states.

To evaluate more rigorously the statistical properties of the bootstrapped series with respect to their original counterparts, for each bootstrapped series, we calculated the following statistics:

1. average
2. standard deviation
3. skewness
4. kurtosis
5. minimum
6. maximum
7. autocorrelation at lag k, $k = 1, \ldots, 8$
8. slope of a linear regression model, b, with $\hat{x}_j = a + b \cdot j + \varepsilon_j$, $j = 1, \ldots, \ell$.

The statistics $1. - 6.$ are concerned with the distribution of prices, while 7. and 8. are more concerned with the dynamic structure of the series. The autocorrelations at lags $k = 3$ to $k = 8$ are observed to check if the similarity between the original and the bootstrapped series is kept beyond the order $k = 2$ used to define the driving process.
For the distribution of each of the above statistics $1. - 8.$, Tables 2 and 3 report (for Spain and Germany, respectively) the 5^{th} and the 95^{th} percentiles, together with the corresponding value for the original series. To evaluate these distributions, we also report the *percentile rank*, i.e., the percentage of cases in which the statistic is smaller than or equal to the original observed one.

We can make the following observations:

i. For all the four partitions π_S, π_S^{milp}, π_G, and π_G^{milp}, all the statistics computed for the original series take values in between the two above percentiles. This is true also for the autocorrelations at lags $k = 3$ to $k = 8$, with the exception of the autocorrelations at lag $k = 7$ for the Spanish case.

ii. The percentile ranks are more fluctuating in the Spanish scenarios than in the German cases: the lowest percentage for Spain is 46% (see, row "average" of partition π_S in Table 2), while it is 51% for Germany (see, row "average" of partition π_G^{milp} in Table 3). The highest percentages in the Spanish and German cases are 99% and 94%, respectively. In general, though, the differences can be considered negligible.

iii. There seems to be no remarkable difference of values between the 5^{th} and the 95^{th} percentiles generated with the two partitions π_S and π_S^{milp}, therefore suggesting that the bootstrapped series obtained with the two partitions are rather similar. The same observation applies to the German case.

iv. Not all the 5000 series generated in each setting showed a spike. This feature reflects the desirable property that a rare event, like a spike, does not appear regularly.

Table 2: Percentiles of the distributions of some indices computed over the 5000 bootstrapped series of Spain for two scenarios: the conservative scenario (partition π_S) and the progressive scenario (partition π_S^{milp}).

Index	Spain - π_S				Spain - π_S^{milp}			
	5^{th} percentile	95^{th} percentile	Value of original series	Percentile rank of original series value	5^{th} percentile	95^{th} percentile	Value of original series	Percentile rank of original series value
average	28.318	31.975	29.688	46	28.080	31.541	29.688	56
standard dev.	7.796	12.538	9.562	62	7.788	12.318	9.562	65
skewness	0.211	2.126	1.383	67	0.214	2.099	1.383	67
kurtosis	−0.127	8.590	5.092	70	−0.150	8.630	5.092	69
minimum	3.599	9.466	5.469	65	3.573	9.303	5.469	69
maximum	58.261	110.729	103.758	80	58.020	110.819	103.758	81
aut. at lag 1	0.748	0.868	0.818	66	0.752	0.868	0.818	65
aut. at lag 2	0.613	0.795	0.705	59	0.617	0.794	0.705	59
aut. at lag 3	0.539	0.764	0.680	70	0.543	0.762	0.680	72
aut. at lag 4	0.485	0.736	0.667	78	0.487	0.733	0.667	80
aut. at lag 5	0.442	0.715	0.661	84	0.444	0.714	0.661	85
aut. at lag 6	0.472	0.729	0.721	93	0.470	0.727	0.721	94
aut. at lag 7	0.526	0.755	0.802	99	0.523	0.753	0.802	99
aut. at lag 8	0.391	0.685	0.683	94	0.387	0.680	0.683	95
lin. regr. slope	0.001	0.006	0.004	69	0.001	0.006	0.004	69

Table 3: Percentiles of the distributions of some indices computed out of the 5000 bootstrapped series of Germany for two scenarios: the conservative scenario (partition π_G) and the progressive scenario (partition π_G^{milp}).

Index	Germany - π_G				Germany - π_G^{milp}			
	5^{th} percentile	95^{th} percentile	Value of original series	Percentile rank of original series value	5^{th} percentile	95^{th} percentile	Value of original series	Percentile rank of original series value
average	30.683	33.497	32.230	60	30.940	33.717	32.230	51
standard dev.	14.017	21.117	18.396	75	14.141	20.925	18.396	77
skewness	1.308	8.187	3.973	62	1.364	7.803	3.973	66
kurtosis	3.457	145.269	35.562	56	3.751	135.671	35.562	61
minimum	1.451	4.051	3.117	81	1.570	4.533	3.117	71
maximum	132.782	493.512	301.542	60	136.361	486.727	301.542	65
aut. at lag 1	0.635	0.772	0.717	60	0.633	0.772	0.717	62
aut. at lag 2	0.428	0.628	0.572	67	0.430	0.623	0.572	70
aut. at lag 3	0.370	0.550	0.507	76	0.368	0.545	0.507	80
aut. at lag 4	0.335	0.516	0.477	79	0.333	0.511	0.477	83
aut. at lag 5	0.339	0.522	0.487	81	0.338	0.515	0.487	85
aut. at lag 6	0.394	0.597	0.586	92	0.395	0.589	0.586	94
aut. at lag 7	0.442	0.673	0.644	84	0.439	0.664	0.644	89
aut. at lag 8	0.355	0.558	0.544	91	0.353	0.549	0.544	93
lin. regr. slope	0.010	0.014	0.013	62	0.010	0.014	0.013	61

The above results suggest that, both in the Spanish and in the German case, it cannot be confuted that the original series was generated by the same Markov process that produced the 5000 bootstrapped series. In addition, the statistical properties of the series generated by the bootstrapping method based on π_S^{milp} and π_G^{milp} do not significantly differ from those of the series generated by the procedure based on π_S and π_G, respectively.

To conclude, our empirical analysis provides evidence that the use of the aggregated transition probability matrix M^* does not significantly alter the "information" contained in the original matrix M. Thus M^* can be adopted in the bootstrapping method to generate resampled series with the same characteristic features of the original one.

4 Conclusions

In this paper we study economic and financial phenomena that evolves according to a Markov chain, and we apply bootstrapping to improve parameter estimation obtained from a sample of data of the observed series. This nonparametric, model-free approach is particularly suited for time series with nonlinear dependence in the data, like electricity prices, and its performance is improved here by adopting a Markov chain reduction that maintains the typical features of the observed time series. The key aspect of our work is that a Mixed Integer Linear Programming model, specifically tailored for our application, is adopted to formulate and solve the reduction problem. Even if solving the proposed model is computationally hard from a theoretical viewpoint, in our application on medium size real-life cases, the problem was solved exactly within reasonable computational times. With this approach, also the purpose of simplifying the original transition probability matrix, and yet preserving the law driving the process, can be considered satisfactorily fulfilled, also in the case of electricity prices that are particularly hard to replicate.

Since for large scale partitioning problems (i.e., when the transition probability matrix has a very large number of rows) a heuristic approach is still advised, the development of *ad hoc* heuristics for this kind of problems is one of our future lines of research on this subject.
Another interesting development could be the introduction in the proposed MILP of new constraints for modeling specific relations that characterize the states of the Markov chain under study. This could help in reducing the number of feasible partitions and, therefore, it may be a viable strategy to reduce the computational effort for solving our reduction problem in the general case.

Appendix A - Trend and weekly seasonality removal

The estimation of the exponential trend and weekly seasonality is based on the following model:

$$e_t^{(c)} = \exp(rt + \eta_1 \mathbb{I}_1(t) + \eta_2 \mathbb{I}_2(t) + \eta_3 \mathbb{I}_3(t) + \eta_4 \mathbb{I}_4(t) + \eta_5 \mathbb{I}_5(t) + \eta_6 \mathbb{I}_6(t) + \eta_7 \mathbb{I}_7(t) + \varepsilon_t), \quad (14)$$

where $e_t^{(c)}$ are the raw original prices, $\mathbb{I}_j(t)$ is the dummy variable signalling whether t is the jth day of the week, with $j = 1, \ldots, 7$, r is the growth rate, η_j is the coefficient of dummy variable $\mathbb{I}_j(t)$, with $j = 1, \ldots, 7$, and ε_t are the errors. If we take the natural logarithm on both sides of Formula (14), we obtain the following formula:

$$z_t = rt + \eta_1 \mathbb{I}_1(t) + \eta_2 \mathbb{I}_2(t) + \eta_3 \mathbb{I}_3(t) + \eta_4 \mathbb{I}_4(t) + \eta_5 \mathbb{I}_5(t) + \eta_6 \mathbb{I}_6(t) + \eta_7 \mathbb{I}_7(t) + \varepsilon_t,$$

where $z_t = \ln e_t^{(c)}$.
For estimation purposes, we assume that the usual hypotheses of linear regression on the errors

ε_t hold. We obtain the OLS estimates of r and η_j, $j = 1, \ldots, 7$, and they are significant at a level of 5% (see Table 4).

Table 4: Coefficients estimates of an exponential regression model of trend and weekly seasonality applied to the series of electricity prices of Spain and Germany.

Coefficient estimate	Spain	Germany
$\hat{r}$	0.0001161214	0.0003865792
$\hat{\eta}_1$	2.9922407785	2.4365264186
$\hat{\eta}_2$	3.2673051542	2.9583694626
$\hat{\eta}_3$	3.2697769413	3.0280264215
$\hat{\eta}_4$	3.2754546141	3.0209612161
$\hat{\eta}_5$	3.2888136542	3.0002419013
$\hat{\eta}_6$	3.2762719043	2.9386885451
$\hat{\eta}_7$	3.1382584293	2.7011989900

To the purpose of removing the exponential trend and the exponential weekly seasonality from our original series, we define the series of prices $e(T) = (e_1, \ldots, e_t, \ldots, e_T)$, where:

$$e_t = \exp[z_t - (\hat{r}t + \hat{\eta}_1\mathbb{I}_1(t) + \hat{\eta}_2\mathbb{I}_2(t) + \hat{\eta}_3\mathbb{I}_3(t) + \hat{\eta}_4\mathbb{I}_4(t) + \hat{\eta}_5\mathbb{I}_5(t) + \hat{\eta}_6\mathbb{I}_6(t) + \hat{\eta}_7\mathbb{I}_7(t))],\ t = 1, \ldots, T.$$

Set $e(T)$ is an input of the bootstrapping method, while the output is the bootstrapped series $x(\ell) = (x_1, \ldots, x_\ell)$. To re-introduce the exponential trend and the exponential weekly seasonality in $x(\ell)$, we multiply each point x_j by $e^{(\hat{r}j + \hat{\eta}_1\mathbb{I}_1(j) + \hat{\eta}_2\mathbb{I}_2(j) + \hat{\eta}_3\mathbb{I}_3(j) + \hat{\eta}_4\mathbb{I}_4(j) + \hat{\eta}_5\mathbb{I}_5(j) + \hat{\eta}_6\mathbb{I}_6(j) + \hat{\eta}_7\mathbb{I}_7(j))}$, $j = 1, \ldots, \ell$.

Appendix B - Initial states, or intervals

Table 5 reports the 12 intervals of the initial partition of the support $[\alpha, \beta]$ of the series of Spain and Germany after removal of trend and weekly seasonality.

Table 5: Elements of the initial partition of the support of the detrended series of electricity prices of Spain and Germany. In both cases, the upper limit of the interval 12, i.e., β, represents a "high enough" value such that no price can be reasonably thought to be greater than it. For example, in our experiment we take $\beta = 1,000,000$.

Interval label	Interval of prices	
	Spain	Germany
α_1	$[0, 0.51)$	$[0, 0.59)$
α_2	$[0.51, 0.66)$	$[0.59, 0.83)$
α_3	$[0.66, 0.76)$	$[0.83, 1.04)$
α_4	$[0.76, 0.87)$	$[1.04, 1.20)$
α_5	$[0.87, 0.96)$	$[1.20, 1.43)$
α_6	$[0.96, 1.09)$	$[1.43, 1.70)$
α_7	$[1.09, 1.18)$	$[1.70, 2.20)$
α_8	$[1.18, 1.35)$	$[2.20, 3.09)$
α_9	$[1.35, 1.54)$	$[3.09, 4.02)$
α_{10}	$[1.54, 1.92)$	$[4.02, 5.93)$
α_{11}	$[1.92, 2.29)$	$[5.93, 8.58)$
α_{12}	$[2.29, \beta]$	$[8.58, \beta]$

References

[1] A. M. Abdel-Moneim and F. W. Leysieffer, Lumpability for Non-Irreducible Finite Markov Chains, *Journal of Applied Probability*, 21(3) 1984, 567-574.

[2] S. Anatolyev and A. Vasnev, Markov Chain Approximation in Bootstrapping Autoregressions, *Economics Bulletin*, 3(19) 2002, 1-8.

[3] D. R. Barr and M. U. Thomas, An Eigenvector Condition for Markov Chain Lumpability, *Operations Research*, 25(6) 1977, 1028-1031.

[4] W. Brock, J. Lakonishok and B. LeBaron, Simple Technical Trading Rules and the Stochastic Properties of Stock Returns, *The Journal of Finance*, 47(5) 1992, 1731-1764.

[5] P. Bühlmann and A. J. Wyner, Variable Length Markov Chains, *The Annals of Statistics*, 27(2) 1999, 480-513.

[6] P. Bühlmann, Bootstraps for Time Series, *Statistical Science*, 17(1) 2002, 52-72.

[7] D. W. Bunn, *Modelling Prices in Competitive Electricity Markets*, John Wiley & Sons, Chichester, 2004.

[8] C. J. Burke and M. A. Rosenblatt, A Markovian Function of a Markov Chain, *The Annals of Mathematical Statistics*, 29(4) 1958, 1112-1122.

[9] R. Cerqueti, P. Falbo, G. Guastaroba and C. Pelizzari, A Tabu Search Heuristic Procedure in Markov Chain Bootstrapping, *European Journal of Operational Research*, 227(2) 2013, 367-384.

[10] R. Cerqueti, P. Falbo and C. Pelizzari, Relevant States and Memory in Markov Chain Bootstrapping and Simulation, Munich Personal RePEc Archive, 2010, http://mpra.ub.uni-muenchen.de/46254/1/MPRApaper46250.pdf.

[11] Wai-Ki Ching, M. K. Ng and E. S. Fung, Higher-Order Multivariate Markov Chains and Their Applications, *Linear Algebra and Its Applications*, 428(2-3) 2008, 492-507.

[12] F.K.R. Chung, *Spectral Graph Theory*, American Mathematical Society, Providence, RI, 92, 1997.

[13] K. Deng, P. G. Mehta and S. P. Meyn, Optimal Kullback-Leibler Aggregation Via Spectral Theory of Markov Chains, *IEEE Transactions on Automatic Control*, 56(12) 2011, 2793-2808.

[14] B. Efron, Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7(1) 1979, 1-26.

[15] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, 1993.

[16] D. Freedman, On Bootstrapping Two-Stage Least-Squares Estimates in Stationary Linear Models, *The Annals of Statistics*, 12(3) 1984, 827-842.

[17] D. A. Freedman and S. C. Peters, Bootstrapping a Regression Equation: Some Empirical Results, *Journal of the American Statistical Association*, 79(385) 1984, 97-106.

[18] O. Jeanne and P. Masson, Currency Crises, Sunspots and Markov-Switching Regimes, *Journal of International Economics*, 50(2) 2000, 327-350.

[19] J. D. Hamilton, Specification Testing in Markov-Switching Time-Series Models, *Journal of Econometrics*, 70(1) 1996, 127-157.

[20] J. D. Hamilton, What's Real About the Business Cycle?, *Federal Reserve Bank of St. Louis Review*, 87(4) 2005, 435-452.

[21] R. Huisman and R. Mahieu, Regime Jumps in Electricity Prices, *Energy Economics*, 25(5) 2003, 425-434.

[22] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Springer, Berlin, 1976.

[23] A. N. Kolmogorov, Three Approaches to the Quantitative Definition of Information, *Problemy Peredachi Informatsii*, 1(1) 1965, 3-11.

[24] M. Meila and L. Xu, Multiway Cuts and Spectral Clustering, University of Washington - Department of Statistics, 442 2004, www.stat.washington.edu/research/reports/2004/tr442.pdf.

[25] M. Mueller and S. Kramer, Integer Linear Programming Models for Constrained Clustering; Discovery Science. B. Pfahringer, G. Holmes and A. Hoffman (Eds.), Springer, Berlin/Heidelberg, 6332 2010, 159-173.

[26] Z. Rached, F. Alalaji and L. L. Campbell, The Kullback-Leibler Divergence Rate Between Markov Sources, *IEEE Transactions on Information Theory*, 50(5) 2004, 917-921.

[27] B. Sağlam, F. S. Salman, S. Sayin and M. Türkay, A Mixed-Integer Programming Approach to the Clustering Problem with an Application in Customer Segmentation, *European Journal of Operational Research*, 173(3) 2006, 866-879.

[28] W. M. Spears, A Compression Algorithm for Probability Transition Matrices, *SIAM Journal on Matrix Analysis and Applications*, 20 1998, 60-77.

[29] R. Sullivan, A. Timmermann and H. White, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *The Journal of Finance*, 54(5) 1999, 1647-1691.

[30] M. U. Thomas, Aggregation and Lumping of DTMCs, Wiley Encyclopedia of Operations Research and Management Science. J. J. Cochran, L. A. Cox, Jr., P. Kesikinocak, J. P. Kharoufeh and J. C. Smith (Eds.), John Wiley & Sons, Hoboken, NJ, 2010.

[31] D. Verma and M. Meila, Comparison of Spectral Clustering Methods, *Advances in Neural Information Processing Systems*, 15, 2003, www.stat.washington.edu/spectral/papers/nips03-comparison.ps.

[32] L. B. White, R. Mahony and G. D. Brushe, Lumpable Hidden Markov Models-Model Reduction and Reduced Complexity Filtering, *IEEE Transactions on Automatic Control*, 43(12) 2000, 2297-2306.

[33] J. H. Ward, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58(301) 1963, 236-244.

[34] R. Weron, M. Bierbrauer and S. Trueck, Modeling Electricity Prices: Jump Diffusion and Regime Switching, *Physica A: Statistical Mechanics and its Applications*, 336(1-2) 2004, 39-48.

[35] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*, John Wiley & Sons, Chichester, 2006.

[36] J. Zhu, J. Hong and J. G. Hughes, Using Markov Chains for Link Prediction in Adaptive Web Sites, SoftWare 2002: Computing in an Imperfect World. D. Bustard, W. Liu and R. Sterritt (Eds.), Springer-Verlag, Berlin/Heidelberg, 2311 2002, 55-66.